Term Level Search Result Diversification

Van Dang and W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{vdang, croft}@cs.umass.edu

ABSTRACT

Current approaches for search result diversification have been categorized as either *implicit* or *explicit*. The implicit approach assumes each document represents its own topic, and promotes diversity by selecting documents for different topics based on the difference of their vocabulary. On the other hand, the explicit approach models the set of query topics, or aspects. While the former approach is generally less effective, the latter usually depends on a manually created description of the query aspects, the automatic construction of which has proven difficult. This paper introduces a new approach: term-level diversification. Instead of modeling the set of query aspects, which are typically represented as coherent groups of terms, our approach uses terms without the grouping. Our results on the ClueWeb collection show that the grouping of topic terms provides very little benefit to diversification compared to simply using the terms themselves. Consequently, we demonstrate that term-level diversification, with topic terms identified automatically from the search results using a simple greedy algorithm, significantly outperforms methods that attempt to create a full topic structure for diversification.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

Search result diversification, term level, topic level.

1. INTRODUCTION

Search result diversification has been studied as a task of re-ordering an initial ranking of documents retrieved for a query. The goal is to produce a more diverse ranked list with respect to some set of topics or aspects associated with this query. Existing approaches to diversification have been classified as either *implicit* or *explicit* [30]. The implicit approach includes MMR [4] and its probabilistic variants [31]. These techniques do not assume any explicit representation of the underlying topics for a query. Instead, they assume each document represents its own topic. As a result, diversity is achieved by iterating over the input ranking and selecting documents based on the difference of their vocabulary, as measured by document similarity. These methods are generally less effective [1, 30, 18] as there are no guarantees that the topics covered by the resulting documents correspond to query aspects.

The explicit approach, on the other hand, models the set of query aspects and select documents for each of them. This includes algorithms such as IA-Select [1], xQuAD [30] and Proportionality Model [18]. The success of these methods, however, has been observed mostly with descriptions of query aspects that have been created manually, either as a concise list of topics [30, 18], a larger taxonomy from which the query topics can be inferred [1], or a list of topics obtained directly from commercial search engines [30, 18].

Generating the query aspects or topic descriptions automatically, on the other hand, is not as well understood. Although there have been a number of attempts to do this [5, 27, 17], only more recent techniques that build topic descriptions by combining information from several sources have been shown to be effective on web copora [19, 20].

In the literature, a query topic or aspect is usually identified as a single phrase or unit. More generally, a topic is a coherent group of what we call topic terms. Fig. 1 shows an example TREC query: *joints* (topic 82) with two topics: *treat joint pain* and *woodwork joint type*. These topics contain five topic terms: *treat*, *joint*, *pain*, *woodwork* and *type*. The question that we address in this paper is whether diversification with respect to these topics benefits from the additional structure or grouping of terms or would diversification using the topic terms directly be just as effective?

We investigate the problem of *term level diversification* empirically. Instead of modeling the set of query aspects, each of which is a coherent group of terms, this approach directly models these terms without their topical grouping. Thus, it still explicitly models the user intents, but it uses a weaker representation of them. Our experiments on the ClueWeb collection using two existing diversification frameworks [30, 18] confirm that discarding the topic structure does not result in any significant loss in diversification effec-

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

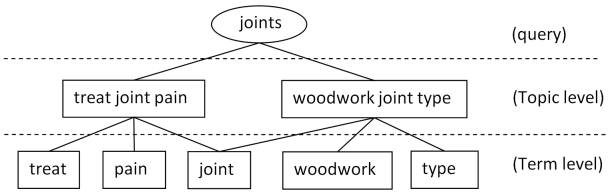

Figure 1: Two different levels for diversification: topic level and term level.

tiveness. Therefore, instead of trying to recover the topics for a query, we only need to identify a set of terms that cover most of the query topics. This is, in fact, the main task for multi-document summarization (e.g., [29, 23, 24]).

Consequently, we propose to use a simple greedy algorithm from the document summarization literature for identifying topic terms for diversification from the initial ranking of documents [23, 24]. Our results show that this simple method significantly outperforms many existing approaches for estimating the full topic structure from the same data on a wide ranges of both relevance and diversity measures. To the best of our knowledge, our method is the first one that can provide statistically significant improvement over standard relevance-based retrieval models in both relevance and diversity measures, without relying on any external data source or manually created topic set.

In summary, the main contribution of this paper is term level diversification. It simplifies the current topic level approach by taking as input a set of terms as opposed to a set of topic descriptions. This is important since automatic topic generation has proven challenging. We show that our approach with terms generated automatically, using a summarization technique [23, 24], significantly outperforms its topic level counterpart with topics generated using existing methods. When ground-truth query topics are available, our approach remains comparable to the topic level alternative.

In the next section, we briefly mention related work. Section 3 presents the current topic level diversification frameworks, which will also be used for term diversification. Section 4 describes in more detail the notion of term level diversification as well as our algorithm for identifying topic terms. Section 5 and 6 contains the experimental setup and results, as well as analysis and discussions. Finally, Section 7 concludes.

2. RELATED WORK

Search result diversification has been studied as the task of retrieving documents covering multiple possible topics, aspects, or interpretations of a query. Existing work can be categorized using two orthogonal criteria: their representation of these topics and their notion of diversity.

Query Topic Representation. Proposed techniques are usually classified as either *implicit* or *explicit*. The implicit approach, in fact, does not assume any of such representation. Instead, it assumes each document has its own topic. It promotes diversity by selecting documents that are different to one another in terms of vocabulary, as captured by document similarity such as cosine [4] or Pearson's correlation [28] between the document vectors and KL divergence between their language models [31]. As the selected documents do not necessary cover any of the query topics, this approach often fails to provide consistent improvement over standard relevance-based retrieval model on large web corpora [19, 18].

The explicit approach, on the other hand, models the set of query aspects and returns documents for each of them [5, 1, 30, 17]. Our term level diversification scheme belongs to this second category. The difference is, instead of modeling the set of query topics, each of which is a group of terms, we model these terms directly without their grouping structure. As we will show later on, the grouping provides very little benefit to diversification compared to the presence of the topic terms. This effectively reduces the task of finding a set of topics into finding a simple set of terms.

The success of the explicit approach, in fact, has been observed primarily with query topics that are either created manually (e.g. TREC subtopic descriptions [30, 18] or a larger predefined taxonomy [1]) or obtained directly from related queries provided by commercial search engines [30, 18]. Generating these aspects automatically, on the other hand, is not as well understood. For example, while clustering queries from logs [27] or anchor text and ngrams from the web [17] can produce interesting looking clusters of text, their effectiveness for diversification has yet been confirmed. Topics extracted from clustered documents, either deterministically or probabilistically via topic modeling [5], were only evaluated on a very small collection. In addition, their effectiveness is concluded to be only comparable to MMR [4], the canonical technique from the implicit approach [5]. Only more recent work [19, 20] has achieved some success, but they generally build topic descriptions by combining informations from several sources of data.

Instead of trying to generate a set of topics for a query, we apply a simple greedy algorithm [23, 24] to extract a diverse set of topic terms automatically from the input ranking. We then evaluate and compare their effectiveness for diversification (term level) with topics generated using some of the subtopic mining techniques mentioned above that utilize the same data (topic level) [5].

Notion of Diversity. There are two notions of diversity in the current literature: diversity by *redundancy* and by *proportionality*. The concepts of redundancy and novelty are based on the cascade user model which assumes users will scan the result list from top to bottom [14]. Therefore, documents at any position in the result list that provide the same information as those at earlier ranks are considered redundant. Similarly, novel documents are those that provide new information. A ranking is more diverse if it contains less redundancy, or equivalently, more novelty. Common to these techniques [4, 31, 7, 1, 5, 28, 30, 32, 18] are the greedy framework which sequentially selects documents with minimal redundancy, the measure of which is where they differ. For example, MMR [4] (implicit) measures redundancy of a document by its cosine similarity to the documents selected previously. IA-Select [1] and xQuAD [30] (explicit) measures how much it covers the query topics that have not been well covered by those chosen earlier.

On the other hand, a proportional ranking of documents with respect to a topic popularity distribution is a ranking in which the number of documents on each topic is proportional to its popularity [18]. By this definition, perfectly proportional search results would naturally be diverse. The main algorithm in this class is PM-2 [18], which selects doc-

uments in a similar greedy fashion, except that it maximizes proportionality using the Sainte-Laguë formula.

In this paper, we compare term level diversification to the topic level counterpart using both frameworks. In particular, we choose xQuAD (redundancy-based) and PM-2 (proportionality-based) simply because they have been demonstrated to be effective on the ClueWeb collection, which we also use to conduct experiments.

3. DIVERSIFICATION FRAMEWORK

In this section, we first formally describe the problem of diversification at the topic level. Then we will present the two frameworks for diversification in the current literature: redundancy-based and proportionality-based diversification. These frameworks will later be used for term diversification.

3.1 Topic Level Diversification

Let q indicate a user query and $T = \{t_1, t_2, ..., t_n\}$ indicate the set of topics for q. Let $W = \{w_1, w_2, ..., w_n\}$ denote the weights for each of the topics $t_i \in T$. These weights can be interpreted as the importance [30] or popularity [18] depending on the diversification techniques. In addition, let $R = \{d_1, d_2, ..., d_m\}$ indicate a ranked list of documents initially retrieved for q and $P(d|t)$ denote some probabilistic estimate of d's relevance to a topic t. The task of topic level diversification is to select a subset of R using $\{T, W, P(d|t)\}$ to form a diverse ranked list S of size k.

It is worth noting that the type of topics $T = \{t_1, t_2, ..., t_n\}$ will determine the relevance measure $P(d|t)$. For example, if T is a set of short textual descriptions (e.g. queries), $P(d|t)$ is often the relevance score of d to t given by some retrieval models [30, 18].

3.2 Framework

3.2.1 Diversity by Redundancy

This framework promotes diverse rankings of documents by penalizing redundancy at every rank. It does so by greedily selecting documents in R to put into S. At each step, it selects the document that is most different to those previously selected (thus minimizing redundancy), while remains relevant to the query q:

$$d^* = \arg\max_{d_j \in R} \ (1-\lambda) \times P(d_j|q) + \lambda \times D(d_j, S) \quad (1)$$

where $D(d_j, S)$ is a measure of novelty, which indicates the difference between the candidate document d_j and each of the documents in S. Different choices of $D(d_j, S)$ correspond to different instantiations of this framework [1, 5, 30]. In this paper, we choose xQuAD [30] simply because it has proven effective on several TREC Web Track query sets [30, 18], which we use to carry out our evaluation. Our findings, nevertheless, should apply to all techniques within this framework.

xQuAD measures the difference between documents by the topics they cover. It defines p_i to be the "portion" of the topic t_i that has not been covered by documents in S:

$$p_i = \prod_{d_j \in S} (1 - P(d_j|t_i)) \quad (2)$$

Higher p_i indicates that most of the documents in S are not relevant to t_i. As such, t_i is less substantially covered and it should have higher "priority" in getting more documents. With this, $D(d_j, S)$ is calculated as follows:

$$D(d_j, S) = \sum_{t_i \in T} w_i \times P(d_j|t_i) \times p_i \quad (3)$$

which means the novelty of a document is its ability to cover the topics that need covering (i.e. higher p_i) weighted by the importance of the topics w_i.

3.2.2 Diversity by Proportionality

The main algorithm in this class is the proportionality model PM-2 [18]. It is a probabilistic adaptation of the Sainte-Laguë method for assigning seats to members of competing political parties such that the number of seats for each party is proportional to the votes they receive. PM-2 starts with a ranked list S with k empty seats. For each of these seats, it computes the quotient qt_i for each topic t_i following the Sainte-Laguë formula:

$$qt_i = \frac{w_i}{2s_i + 1}$$

According the the Sainte-Laguë method, this seat should be awarded to the topic with the largest quotient in order to best maintain the proportionality of the list. Therefore, PM-2 assigns the current seat to the topic t_{i^*} with the largest quotient. The document to fill this seat is the one that is not only relevant to t_{i^*} but to other topics as well:

$$d^* = \arg\max_{d_j \in R} \ \lambda \times qt_{i^*} \times P(d_j|t_{i^*}) + (1-\lambda) \sum_{i \neq i^*} qt_i \times P(d_j|t_i) \quad (4)$$

After the document d^* is selected, PM-2 increases the "portion" of seats occupied by each of the topics t_i by its normalized relevance to d^*:

$$s_i = s_i + \frac{P(d^*|t_i)}{\sum_{t_j \in T} P(d^*|t_j)}$$

This process repeats until we get k documents for S or we are out of candidate documents. The order in which each document is put into S determines its ranking.

4. TERM LEVEL DIVERSIFICATION

4.1 Problem Statement

Diversification at the term level is very similar to the topic level. Let $t_i = \{t_i^1, t_i^2, ..., t_i^{|t_i|}\}$ be the set of terms for topic t_i. Instead of diversifying R using the set of topics $T = \{t_1, t_2, ..., t_n\}$, we propose to perform diversification using $T' = \{t_i^1, t_i^2, ..., t_i^{|t_i|}, ..., t_n^1, t_n^2, ..., t_n^{|t_n|}\}$, in effect treating each t_i^j as a topic.

Let us reuse the example query provided in Fig. 1 earlier to illustrate this. Instead of diversifying the initial ranking for the query *joints* with respect to two of its topics: *treat joint pain* and *woodwork joint type*, we propose to perform diversification with respect to its topic terms: *treat*, *joint*, *pain*, *woodwork* and *type*.

We will now compare diversification at the topic level to its term level counterpart at an intuitive level, using both frameworks, to provide some justification for why one can expect similar performance from these two paradigms. This is based on the assumption that if a document is more relevant to one topic than another, it is also more relevant to the terms associated with this topic than any of the terms

from the other topic. In other words, if a document is more relevant to *treat joint pain* than it is to *woodwork joint type*, we assume that it is also more relevant to *treat* and *pain* than it is to *woordwork* and *type*. Since both topics have the term *joint*, we will ignore it for the ease of explanation.

Let us first explain using the xQuAD's framework. At either the topic or term level, it follows from Eq. (1) that the first selected document d_1 is the one that is most relevant to the user query. Let us assume this document d_1 is more relevant to *treat joint pain* than it is to *woodwork joint type*. At the second step at the topic level, *woodwork joint type* will have higher "priority" (higher value for p) to get documents and thus, xQuAD will favor documents on this topic. At the same time at the term level, d_1 should be also more relevant to *treat* and *pain* than it is to *woodwork* and *type* (because of our assumption). Therefore, *woodwork* and *type* will have higher "priority" than either *treat* or *pain*. It follows that, if the topic level system is able to find a document d^* for its relevance to *woodwork joint type*, this same document should also emerge at the term level. Once this document is selected, the "priority" of *woodwork* and *type* decreases as does that of *woodwork joint type*. As the algorithm proceeds, the two approaches may select different documents due to the different numbers of "topics", each of which is down-weighted by a different amount (by Eq.(2)). Nevertheless, the general idea still applies, that if a document is selected for its relevance to any particular topic, that same document should be at least a highly potential candidate at the term level due to its relevance to the corresponding terms.

Similarly, in the framework of PM-2, let us also assume the first document d_1 is more relevant to *treat joint pain* than it is to *woodwork joint type* at both levels. As a result, the "portion" of seats occupied by *treat* and *pain* is also higher than that of both *woodwork* and *type*, just as *treat joint pain* will have a higher portion than *woodwork joint type*. At the second step, *woodwork joint type* should be assigned a higher quotient by the Sainte-Laguë formula, which again indicates *woodwork joint type* has higher "priority" similarly to what happens in the xQuAD framework.

4.2 Choice of $P(d|t)$

As mentioned earlier, diversification frameworks assume $\{T, W, P(d|t)\}$ as inputs and the choice of T will determine $P(d|t)$. An obvious choice for $P(d|t)$ for term level diversification is $P(t_i^k|d)$, the probability that the document d generates the topic term t_i^k. This is, however, highly problematic. At the term level, in addition to those true query topics which have become latent, there are also "false" latent topics formed by the wrong combinations of terms. In the context where we identify topic terms for a query automatically, some of them might be generic and ineffective. As the number of bad terms increases, the number of "false" topics will grow exponentially. Combined with the fact that there are likely many non-relevant documents in the baseline ranking, term diversification under the effects of these "false" topics might end up promoting non-relevant documents.

Assuming any document that is relevant to a true query topic should be relevant to the query itself, we propose to expand each topic term with the query. Let $\{q_1, q_2, ..., q_n\}$ be the set of terms of the query q. $P(d|t_i^k)$ is estimated as follows:

$$P(d|t_i^k) = (P(t_i^k|d)P(q|d))^{\frac{1}{|t_i^k|+|q|}} = (P(t_i^k|d)\prod_{q_j \in q} P(q_j|d))^{\frac{1}{|t_i^k|+|q|}}$$

which is essentially the query likelihood model for ranking d with respect to the query $\{t_i^k, q_1, q_2, ..., q_n\}$ [15] normalized by the query length to avoid biased towards shorter terms (i.e. terms can include both unigrams or phrases). In the case where all terms have the same length, the normalization is certainly not necessary.

Note that the inclusion of the query is not the only mechanism to keep non-relevant documents under control. Interpolating $P(t_i^k|d)$ with $P(q|d)$ as has been done in the redundancy-based framework (Eq. (1)) is another possibility. We do not do so because we do not want to introduce more parameters into the framework. In principle, any combination of $P(t_i^k|d)$ and $P(q|d)$ should be applicable. Since $P(q|d)$ can be obtained directly from the baseline rankings, this does not increase computational complexity.

4.3 Automatic Identification of Topic Terms

We now present `DSPApprox`, the *topic term extraction* algorithm proposed by Lawrie and Croft [23, 24] for hierarchical multi-document summarization. The goal of the algorithm is to select from a collection of documents a small set of highly representative terms that best summarize them. This algorithm is applied hierarchically, resulting in an hierarchical topic structure.

Since we only need a single diverse set of topic terms, we only apply the algorithm once on the initial ranking of documents $R = \{d_1, d_2, ..., d_m\}$ retrieved for the query q. The algorithm first identifies a set of *vocabulary* from these documents, from which it forms a set of more specific *topic terms*. It then measures for these terms their *topicality* and how well they predict the occurrences of other terms. Finally, it greedily selects a subset of topic terms, aiming to maximize both their topicality and their coverage of the vocabulary.

Vocabulary Identification. We consider as vocabulary all terms that (1) appear in at least two documents, (2) have at least two characters and (3) are not numbers. In our experiments, we test two types of terms: **unigrams** and **phrases**. We use a very simple method for phrase extraction. We scan through terms in each document and at each position, we select the longest sequence of terms that matches a wikipedia title as a phrase.

Topic Terms Identification. All vocabulary terms that co-occur with any of the query terms within a proximity window of size w is selected as topic terms.

Topicality and Predictiveness. Topicality of a term measures how informative it is at describing the set of documents. To compute topicality, a relevance model $P_R(t|q)$ [25] is first estimated from the initial set of documents R:

$$P_R(t|q) = \sum_{d_i \in R} P(t|d_i)P(d_i|q)$$

where $P(t|d)$ is the probability that d_i generates the term t and $P(d_i|q)$ is relevance of d_i to the query. The topicality $TP(t)$ of a term t is estimated as its contribution to the KL divergence between this relevance model and the language model for the entire retrieval collection:

$$TP(t) = P_R(t|q) log_2 \frac{P_R(t|q)}{P_c(t)}$$

It is equivalently t's contribution to the clarity score of the query q [16].

Predictiveness, on the other hand, measures how much the occurrence of a term predicts the occurrences of others. Let $P_w(t|v)$ indicate the probability that a term t occurs within a window of size w of another term v and C_t indicate the set all such v. The predictiveness of t is estimated as follows:

$$PR(t) = \frac{1}{Z} \sum_{v \in C_t} P_w(t|v)$$

where Z is the hierarchy level specific normalization factor. In our case, we set it to the size of the vocabulary.

Greedy Algorithm. Pseudo-code for this algorithm is presented as Algorithm 1. It iteratively selects terms from the candidate topic term set T. The utility of each term is the product of its topicality and predictiveness. At each step, the algorithm selects the topic term $t^* \in T$ with maximum utility. Then, it decreases the predictiveness of other topic terms that predict the same vocabulary. This makes sure topic terms that cover the uncovered part of the vocabulary will emerge for selection in the next iteration. The algorithm stops once the utility of all candidate topic terms reaches 0, indicating that all vocabulary has been covered. Some example topic terms (both unigrams and phrases) generated by `DSPApprox` for the query *joints* are provided in Table 1.

Algorithm 1 `DSPApprox` for identifying topic terms.

```
V = {v_1, v_2, ..., v_n}: the set of vocabulary
T = {t_1, t_2, ..., t_m}: the set of candidate topic terms
C_{t_i}: set of terms occurring within a window to t_i
P_w(t_i|v): co-occurrence (within window of size w) statistics
Compute topicality TP(t_i), ∀t_i ∈ T
Compute predictiveness PR(t_i), ∀t_i ∈ T
DTT : the output diverse set of topic terms
PREDV : vocabulary that has been predicted by DTT
DTT ← ∅
PREDV ← ∅
while PREDV ⊂ V and |T| > 0 do
    t* ← arg max_{t_i ∈ T} TP(t_i) × PR(t_i)
    DTT ← DTT ∪ t*
    T ← T \ {t*}
    pred ← C_{t*}
    for all v ∈ pred \ PREDV do
        for all t_i ∈ T do
            PR(t_i) ← PR(t_i) − P_w(t_i|v)
        end for
    end for
    PREDV = PREDV ∪ pred
end while
```

5. EXPERIMENTAL SETUP

Query and Retrieval Collection. Our query set consists of the 147 queries with relevance judgments from three years of the TREC Web Track's diversity task (2009 [10], 2010 [11] and 2011 [12]). Our evaluation is done on the ClueWeb09 Category B retrieval collection[1], which contains roughly 50 million web pages in English. This collection is stemmed using the Krovetz stemmer [22]. Stopword removal is only performed on the query using a small stopword list.

Baseline Retrieval Model. We use the standard query-likelihood model [15] implemented in Indri[2] to conduct the initial retrieval run. This run serves not only as a means to provide a set of documents for the diversification systems but also as a baseline to verify their usefulness.

Spam filtering is known to be an important component of web retrieval [2]. In addition, documents with too few stopwords are found to have poor readability [21, 26]. Therefore, we incorporate both of these into our baseline ranking. We use the spam filtering technique described by Cormack et al. [13], which assigns a "spamminess" percentile $S(d)$ to each document d in the collection. Let $\sigma(d)$ be the stopword to non-stopwords ratio in d and $p(d|q)$ indicate the score the retrieval model assigns to the document d. Following Bendersky et al. [2], the final score of d is given by:

$$P(d|q) = \begin{cases} p(d|q) & \text{if } S(d) \geq 60 \text{ and } \sigma(d) \geq 0.1 \\ -\infty & \text{otherwise} \end{cases}$$

Diversification Frameworks. We compare term level diversification to topic level diversification using both xQuAD [30] and Proportionality Model (PM-2) [18], which we have described in Section 3. While xQuAD obtains diversity by penalizing redundancy at every position in the ranked list, PM-2 does so by promoting proportionality at every rank.

Evaluation Metric. We report our evaluation results using several standard metrics that have been used in the official evaluation of the diversity tasks at TREC [11, 12]: α-NDCG [8], ERR-IA (a variant of ERR [6]) and NRBP [9]. These metrics penalize redundancy at each position in the ranked list based on how much of that information the user has already seen from documents at earlier ranks. In addition, we also report our results using Precision-IA [1] and subtopic recall, which indicate respectively the precision across all topics of the query and how many of those topics are covered in the search results. All of these measures are computed using the top 20 documents retrieved by each model to be consistent with official TREC evaluation. Statistically significant differences are measured using two-tailed t-test with p-value < 0.05.

Most diversification mechanisms are evaluated using only diversity measures [1, 30, 18]. It is unclear if diversity is achieved at a cost to relevance. Therefore, in addition to all diversity measures above, we also report our results using two standard relevance-based metrics for web retrieval: NDCG and ERR, which are also evaluated at the top 20 documents.

Parameter Settings. All of the diversification approaches under evaluation are applied on the top K retrieved documents. We set $K = 50$ to be consistent with existing research which found that both xQuAD and PM-2 achieve their highest performance at $K = 50$ [18]. Consequently, all topic and term extraction techniques will also operate on these top 50 documents.

Each topic and term extraction technique, as we will show later, has several *free* parameters that require tuning. xQuAD and PM-2 also have one parameter λ to tune. To enforce fair comparison, all parameters are selected via 3-fold cross validation.

We consider for λ values in the range of $[0.05, 1.0]$ with a increment of 0.05. Value ranges for parameters of the topic and term extraction methods will be presented in their respective sections.

[1] http://boston.lti.cs.cmu.edu/Data/clueweb09/
[2] http://www.lemurproject.org

Table 1: Some example outputs of DSPApprox for the query "joints" (topic number 82). Important terms from the original TREC subtopics for this query are also provided for easy references.

TREC Sub-topic	DSPApprox[Unigram]	DSPApprox[Phrase]
1) joints in human body	spine	elbow joint
	articulate	knee joint
2) woodworking joints types	miter	mitter joint
	planter	mitter box
3) treat joint pain	symptom	joint pain
	grease	joint anti inflamory

6. EVALUATION

6.1 Term Level Diversification: Effectiveness

We first compare the term level diversification approach to the topic level approach using the set of *true topics* associated with each query (TREC "sub-topics"). A topic is a coherent group of terms. These topics represent the oracle grouping of the oracle topic terms. By comparing the diversification effectiveness of this set of topics (topic level diversification) with that of the corresponding set of *unigram* topic terms (term level diversification), we can separate the benefit diversification algorithms get from the grouping with the benefit they get from the presence of topic terms.

In addition, related queries provided by commercial search engines have been demonstrated to be very effective for diversification [30, 18]. These queries too can be considered good underlying topics for the original query. As a result, we also compare the two diversification paradigms using this topic set. It is worth noting that the search engine provides no suggestions for *three* of the queries in our set. The query set for this experiment only contains 144 queries (out of 147).

Similar to existing work [18], the document-topic relevance function $P(d|t)$ for topic level diversification is implemented as the query-likelihood score for d with respect to t (each topic t is treated as a query). In particular, let $t_i = \{t_i^1, t_i^2, .., t_i^n\}$ indicates the set of terms for the topic t_i. $P(d|t_i)$ is computed using the geometric mean to avoid biased towards shorter topics:

$$P(d|t_i) = \left(\prod_{t_i{}^k \in t_i} P(t_i^k|d) \right)^{\frac{1}{|t_i|}}$$

For term level diversification, $P(d|t)$ is calculated as described in Section 4.2.

Table 2 compares term diversification to topic diversification using both topic sets and both diversification frameworks. The first thing to notice is that both topic and term diversification, using both PM-2 and xQuAD, significantly outperform the baseline in all metrics. This is certainly unsurprising since we are using the oracle data. Nevertheless, it confirms the effectiveness of both of these frameworks at providing relevant and diverse results.

What is more interesting from Table 2 is that the set of topic terms maintains a highly comparable level of performance to the topic structures. There are no statistically significant differences in all cases. These results are consistent across different diversification techniques and topic sets. This suggests that existing diversification frameworks are capable of returning relevant documents for topics without the explicit topical grouping.

We notice, however, that some of the query topics are different to the query itself by only one term. For example, topics for the query *"south africa"* include *"history of south africa"* and *"maps of south africa"*. Both of these topics have only one key term, which is *"history"* and *"maps"* respectively. It is possible that term level diversification is competitive with the topic level alternative because of queries like this.

To investigate this issue, we use the notion of *key term* to indicate the number of non-stopword terms in a query topic that are different to the query text. To quantify the impact the number of key terms has on our approach, we plot the number of topics where each approach is able to provide at least one relevant document against the number of key terms for these topics. In addition, we also plot the actual number of relevant documents retrieved for each topic (on log scale) against the number of key terms it contains. These plots are presented by Fig. 2 (a) and (b) respectively. Note that we only show the plots for PM-2 because the analysis with xQuAD is very similar.

Fig. 2 reveals that not only is our approach comparable with its topic counterpart on topics with a single key term, it also remains competitive consistently across different numbers of key terms. In particular, term level diversification has a slight advantage with topics that have only one single key term. With topics that have two and three key terms, although the topic level systems perform better, the difference is very small. The two approaches become comparable with larger numbers of key terms. Given that the term level systems do not need the topical structure, this very slight performance loss seems reasonable.

In summary, our experiments with the "oracle topics" show that the benefits for diversification of grouping topic terms are minimal compared to just using the terms themselves. The existing frameworks, PM-2 and xQuAD to be specific, can perform topical diversification at the term level. Together, these findings indicate that term level diversification is worth pursuing.

6.2 Automatically Generated Topics vs. Terms

We now evaluate the effectiveness of DSPApprox for automatically extracting topic terms, considering unigrams and phrases separately, from the initial ranking of documents. The total number of unigrams and phrases the algorithm returns are approximately 100 and 500 respectively. Since using too many terms is inefficient and unlikely to be effective, we use a parameter T to control the number of terms used for diversification. The second parameter is w, which determines the size of the window in which (1) a term has to co-occur with at least one query term in order to be considered a candidate topic term, and (2) prediction boundary: a term cannot predict terms that are more than w words away. We consider $w \in \{20, 30, 40, 50\}$ and $T \in \{5, 10, 20, 40, 60, 80, 100\}$.

Table 2: Performance comparison between term level (marked as [Term]) diversification, topic level ([Topic]) and the non-diversification baseline query-likelihood (QL) using two frameworks: PM-2 and xQuAD. They are evaluated on both the oracle topic set (TREC) and the set obtained from a commercial search engine (C.S.E) using a wide range of diversity and relevance measures. Win/Loss (W/L) is with respect to α-NDCG. Q indicates statistically significant differences to QL.

			Diversity						Relevance	
			α-NDCG	W/L	ERR-IA	Prec-IA	S-Recall	NRBP	NDCG	ERR
TREC		QL	0.3927		0.2807	0.1829	0.5824	0.2446	0.2742	0.1371
	PM-2	[Topic]	0.4742^{Q}	84/41	0.3536^{Q}	0.2021^{Q}	0.6341^{Q}	0.3212^{Q}	0.2979	0.1468
		[Term]	0.4635^{Q}	92/30	0.344^{Q}	0.2064^{Q}	0.6268^{Q}	0.3094^{Q}	0.3048^{Q}	0.1535^{Q}
	xQuAD	[Topic]	0.4447^{Q}	85/40	0.3207^{Q}	0.2035^{Q}	0.6259^{Q}	0.2856^{Q}	0.3002^{Q}	0.1490^{Q}
		[Term]	0.4366^{Q}	89/33	0.3143^{Q}	0.2067^{Q}	0.6171^{Q}	0.2777^{Q}	0.3054^{Q}	0.1515^{Q}
C.S.E.		QL	0.3884		0.2783	0.1789	0.5735	0.2428	0.2739	0.1394
	PM-2	[Topic]	0.4308^{Q}	62/60	0.3137^{Q}	0.1815	0.6084^{Q}	0.2784^{Q}	0.2790	0.1415
		[Term]	0.4305^{Q}	69/52	0.3162^{Q}	0.1874	0.6084^{Q}	0.2804^{Q}	0.2872	0.1471
	xQuAD	[Topic]	0.4024^{Q}	63/51	0.2888	0.183	0.5884	0.252	0.2797	0.1423
		[Term]	0.4118^{Q}	70/48	0.2965^{Q}	0.1883	0.5954^{Q}	0.2603^{Q}	0.2906	0.1464

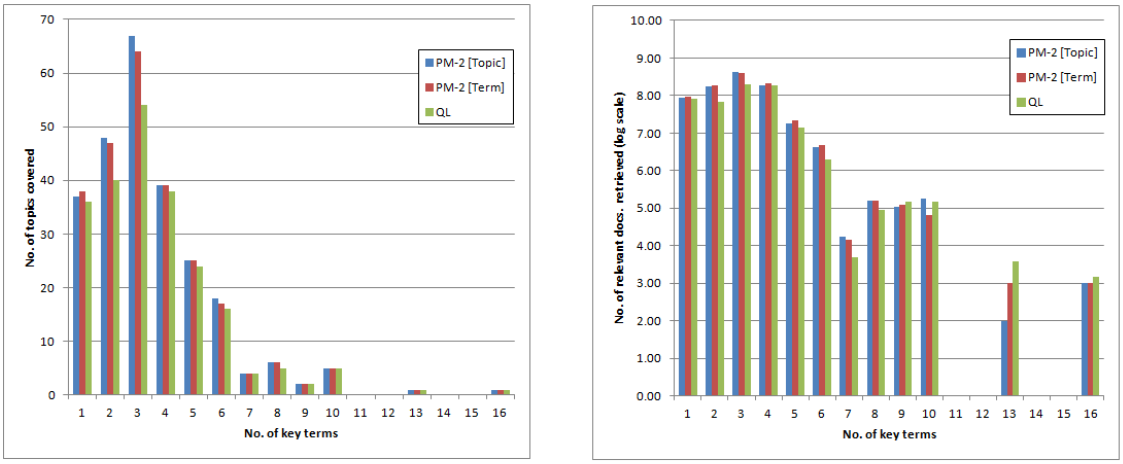

Figure 2: Figure (a) provides the total number of topics (across all queries) the term level the topic level diversification systems (under the PM-2's framework) as well as query-likelihood cover with respect to their number of key terms. Figure (b) shows the number of relevant documents (on log scale) retrieved for each of these topics.

6.2.1 *Baselines*

Baseline 1. Our first baseline for comparison, first proposed by Carterette and Chandar [5], estimates topic models using LDA [3] from the documents and uses the resulting clusters for diversification. This model only has one parameter, which is the number of latent topics $c \in [2..10]$. The relevance between a document to a topic is provided by the LDA framework. We use the multi-threaded implementation of LDA that is publicly available [3].

Baseline 2. Our second baseline technique, also proposed by Carterette and Chandar [5], applies k-nearest neighbor (KNN) first to cluster the documents. After that, it estimates a relevance model [25] from each of the clusters and use it as a topic model. Its parameters include $k \in [2, 10]$ and $T \in \{5, 10, 20\}$, which are the number of neighbors and the number of top terms from the relevance model to be used as topic description respectively. The topic description is treated as an Indri weighted query. The relevance between a document and this topic is obtained directly via Indri's output relevance score.

Baseline 3. MMR [4] has become a canonical baseline in the diversity literature [31, 30, 18]. Though it does not explicitly model topics, it fits into the class of algorithms that relies solely on the set of documents. The framework of MMR is very similar to the one presented in Eq. (1). The novelty component $D(d, S)$, which indicates the different between a document d and those previously selected in S, is aggregated over its difference to each of the document $d_j \in S$. The difference between two documents is implemented based on the cosine similarity. We experimented with three aggregation functions: *max*, *min* and *average* and report results with *max* since it is the most effective.

6.2.2 *Results*

Table 3 presents the results comparing the systems mentioned above. All of their parameters are determined using 3-fold cross validation. The letters Q, M, L, K indicate sta-

[3]https://sites.google.com/site/rameshnallapati/software

Table 3: Performance comparison among (1) topic terms (both unigrams and phrases) generated by DSPApprox (abbreviated as DSP), (2) topics generated by LDA and KNN, (3) MMR which does not explicit model query topics, and (4) the non-diversification baseline query-likelihood (QL) using two frameworks: PM-2 and xQuAD. Evaluation is done using a wide range of diversity and relevance measures. Win/Loss (W/L) is with respect to α-NDCG. *Q, M, L, K* indicate statistically significant differences (p-value < 0.05) to QL, MMR, LDA and KNN respectively. Bold face indicates the best performance in each group.

		Diversity						Relevance	
		α-NDCG	W/L	ERR-IA	Prec-IA	S-Recall	NRBP	NDCG	ERR
	QL	0.3927		0.2807	0.1829	0.5824	0.2446	0.2742	0.1371
	MMR	0.393	30/39	0.2804	0.1829	0.5855	0.2445	0.2700^{Q}	0.1365
PM-2	LDA[Topic]	0.3762	56/70	0.2592^{Q}	0.1579	**0.5977**	0.2192^{Q}	0.2395^{Q}	0.1226^{Q}
	KNN[Topic]	0.3991	41/62	0.2882^{Q}	0.1808	0.5825	0.2536^{Q}	0.2711	0.1373
	DSP[Unigram]	$\mathbf{0.4161}_{L}^{Q,M}$	**71/54**	$0.3085_{L,K}^{Q,M}$	$\mathbf{0.1953}_{L,K}$	0.5789	$0.2788_{L,K}^{Q,M}$	$0.2981_{L,K}^{Q,M}$	0.1440_{L}
	DSP[Phrase]	$0.4159_{L}^{Q,M}$	68/57	$\mathbf{0.3131}_{L,K}^{Q,M}$	$\mathbf{0.1953}_{L,K}$	0.5684	$\mathbf{0.2867}_{L,K}^{Q,M}$	$\mathbf{0.3011}_{L,K}^{Q,M}$	$\mathbf{0.1480}_{L}$
xQuAD	LDA[Topic]	0.3905	55/59	0.2798	0.154^{Q}	**0.5884**	0.2453	0.2350^{Q}	0.1288^{Q}
	KNN[Topic]	0.3897	46/42	0.2786	0.1846	0.5824	0.2426	0.2752	0.1369
	DSP[Unigram]	0.3906	54/57	0.2837	0.1844_{L}	$0.5594_{L,K}^{Q,M}$	$0.252_{K}^{Q,M}$	0.2780_{L}^{M}	0.1386_{L}
	DSP[Phrase]	**0.3943**	56/63	**0.2888**	$\mathbf{0.1923}_{L}$	0.561^{M}	$\mathbf{0.2587}_{L,K}^{Q,M}$	$\mathbf{0.2889}_{L}$	$\mathbf{0.1408}_{L}$

tistically significant differences (p-value < 0.05) to query-likelihood, MMR, LDA and KNN respectively. Among the three baseline techniques for topic generation, MMR's performance is the most similar to the baseline. It is interesting to see that while LDA has the best results in S-Recall, it performs poorly on all other measures, regardless of the diversification techniques. This can be explained by the fact that an LDA topic is a distribution over the entire vocabulary of the initial set of documents. Thus, each topic has a higher chance of matching its documents, but at the same time it also matches several non-relevant documents. Overall, KNN is the only technique among the three baselines that can provide some improvement over query-likelihood (with PM-2). Nevertheless, KNN has a trade-off between relevance and diversity: while KNN topics used by PM-2 help most of the diversity measures, it hurts relevance. On the other hand, the topics it generates, when used by xQuAD, helps relevance but hurt most of the diversity measures. Overall, the difference between these baselines and query-likelihood is mostly not statistically significant.

In contrast, both the unigrams and phrases generated using DSPApprox when used by PM-2 substantially outperform all other systems under comparison on many measures. Statistically significant differences are observed in many cases. In fact, it is the only system that optimizes for diversity measures yet outperforms query-likelihood in both relevance measures. Between unigrams and phrases, the former appears to be slightly more robust by improving more queries and hurting fewer, but the latter manages to retrieve more relevant results. In addition, their performance with xQuAD is still slightly higher than all three baselines on most precision-based measures.

The limited performance of our method when using xQuAD can be explained by xQuAD's vulnerability to large numbers of topics. Let us revisit Eq. (3) and assume that at the k-th step, the topic t_i has the highest "priority" p_i. We can rewrite Eq. (3) with respect to t_i as follows:

$$D(d,S) = w_i \times P(d|t_i) \times p_i + \sum_{\substack{t_j \in T \\ t_j \neq t_i}} w_j \times P(d|t_j) \times p_j$$

There is an implicit *uncontrolled* trade-off here between the relevance of a document to t_i, the topic with the highest priority, and its relevance to other lower priority topics. As the size of T increases, it becomes possible that a document d^* that is relevant to many t_j will be selected even though xQuAD should be selecting documents for t_i. This is certainly not a big problem for topic level diversification since the number of topics is relatively small. At the term level, however, our algorithm generates hundreds of terms, many of which can be very generic. As such, some non-relevant documents can appear randomly relevant to many of such terms, dominating the topic term with the highest priority.

Note that PM-2 has the same trade-off as xQuAD (Eq. (4)). The difference is that it is *controlled* by the parameter λ. If a topic term set is too noisy, cross-validation should be able to specify a larger value for λ to put more emphasis on the topic with the highest priority .

6.2.3 Improvement and Failure Analysis

We focus our analysis of DSPApprox results using PM-2. As can be seen from Table 3, although DSPApprox has slightly lower subtopic recall compared to query-likelihood (QL), the difference is not significant. Our investigation suggests that not only do DSPApprox and QL cover about the same number of topics, they cover almost the same set of topics (97% overlap). This high percentage of overlap suggests that the terms generated by DSPApprox are biased towards topics covered by the top ranked documents in the initial ranking.

We believe the cause of this bias is the way DSPApprox computes topicality. We observe that the topicality of a term is relatively proportional to the probability that it is given by the relevance model [25] estimated from the top 50 documents. This model usually assigns higher probabilities to frequent terms from higher ranked documents since they are assumed more relevant. If a document at a very low position covers topics that are different from those at early ranks, chances are their topic terms do not appear in these documents with high frequency. Therefore, their chance to be included in the resulting set of terms is relatively small, causing these topics to be excluded from the coverage of the final set. This is the main reason why subtopic recall was not improved.

Table 4: Contribution of (1) better topic coverage and (2) within topic coverage and ranking of relevant documents to the overall improvement on α-NDCG. WIN and LOSS indicate the sets of queries whose α-NDCG DSPApprox (DSP for short) improves and hurts respectively. $S.Rec \uparrow$ is the subset of WIN on which subtopic recall is also improved and $REST$ is its complement. $S.Rec \downarrow$ is the subset of LOSS on which subtopic recall is also lowered and $REST$ is its complement. ΔP is the difference of α-NDCG compared to query-likelihood. [U] and [P] indicate terms and phrases respectively.

		ΔP		#q	ΔP
DSP[U]	WIN	+0.2682	$S.Rec \uparrow$	10	+0.0652
			$REST$	61	**+0.2030**
	LOSS	−0.1400	$S.Rec \downarrow$	10	−0.0616
			$REST$	44	−0.0784
DSP[P]	WIN	+0.3285	$S.Rec \uparrow$	11	+0.0749
			$REST$	57	**+0.2536**
	LOSS	−0.2034	$S.Rec \downarrow$	14	−0.1170
			$REST$	43	−0.0864

Regardless, DSPApprox still manages to outperform QL in both α-NDCG and Precision-IA. This indicates that while both of them have the same topic coverage, DSPApprox retrieves more relevant documents for these topics as well as provides better ranking for them. More quantitative analysis on this is provided in Table 4. WIN and LOSS indicate the set of queries where DSPApprox helps and hurts α-NDCG compared to QL. ΔP denotes the performance difference in α-NDCG. $S.Rec \uparrow$ indicates the subset of WIN where S-Recall is also improved and $REST$ indicates the remaining of the set. Similarly, $S.Rec \downarrow$ indicates the subset of LOSS where S-Recall is also lower and $REST$ indicates the remaining. It can be seen that the increase of S-Recall contributes very little to the overall improvement on α-NDCG (the $S.Rec \uparrow$ sets). At the same time, they are also responsible for some performance loss (the $S.Rec \downarrow$ sets). On the other hand, a significant chunk of improvement is observed on the sets of queries that cover no more topics than QL (the $REST$ sets).

The analyses above suggest that the terms provided by DSPApprox, though unable to recover additional topics due to the bias issue, correctly represent most of those covered by QL. Consequently, they help surface more documents on theses topics, significantly improving α-NDCG.

It is worth noting that, diversification with both unigrams and phrases provided by DSPApprox also significantly improves the relevance of the results (NDCG). Our approach, in fact, turns out to be very similar to pseudo-relevance feedback. The difference is that traditional relevance feedback uses the extracted terms to update the query model to retrieve new documents. Our approach, on the other hand, only attempts to re-order the input ranking, pushing more relevant documents to earlier ranks. As such, diversification can be considered a precision-driven framework for relevance feedback.

6.2.4 Parameter Sensitivity

As mentioned earlier, our topic term identification algorithm has two parameters. T controls how many of the top output terms to use for diversification and w determines how many words away can a term predict as well as how far can a topic term be from the query term. The best parameter values selected for DSPApprox (using 3-fold cross-validation) with unigrams is $\{T = 40, w = 20\}$ and the best values for phrases are $\{T = 80, w = 40\}$.

We first vary T from 10 to 200 and and keep $w = 20$ for unigrams and $w = 40$ for phrases. Note that the line for unigrams stops at $T = 100$ since our algorithms generates at most 100 unigram terms. Fig. 3 shows that regardless of the value of T, their is always some improvement. The set of unigrams, in particular, is very robust: it provides substantial improvement for most of T's values. As for phrases, two few (e.g. less than 50) or too many (more than 100) terms result in very minor improvement.

We then vary w from 20 to 50 and keep T constant. Fig. 3 shows the sensativity of this parameter. Similarly, improvement is observed at every value.

7. CONCLUSIONS AND FUTURE WORK

This paper introduces a new approach to topical diversification: diversification at the term level. Existing work models a set of aspects for a query, where each aspect is a coherent group of terms [30, 18]. Instead, we propose to model the topic terms directly. Our experiments, using both TREC subtopics and related queries provided by a commercial search engine, show that the two approaches achieve highly comparable results in all diversity and relevance measures. It indicates that the topical grouping provides little benefit to diversification compared to the presence of the terms themselves. The reason for this is that if a document is selected by the topic level system for its relevance to some particular topics, it is often relevant to the corresponding topic terms as well. Thus, this document also appears as a highly potential candidate to term level system. Term level diversification, in fact, works in the same principles as the topic counterpart.

This effectively reduces the task of finding a set of query topics, which has proven difficult, into finding a simple set of terms. Consequently, we propose to use a simple greedy algorithm from the literature of multi-document summarization [23, 24] to identify a diverse set of topic terms (unigrams and phrases). Our results demonstrate that, diversification using these terms significantly outperforms its topic level alternative with automatically extracted topics, as well as the standard relevance-based retrieval models on various diversity and relevance measures.

For future work, we will consider applying DSPApprox on not only the initial retrieved documents but also on external data such as Wikipedia and anchor text collections. We believe this will help alleviate the current bias issue, improving sub-topic recall. In addition, note that DSPApprox itself is a term diversification algorithm: it selects a set of terms that best cover the vocabulary. It is worth examining the possibility of replacing it with techniques such as PM-2.

8. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part under subcontract #19-000208 from SRI International, prime contractor to DARPA contract #HR0011-12-C-0016. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

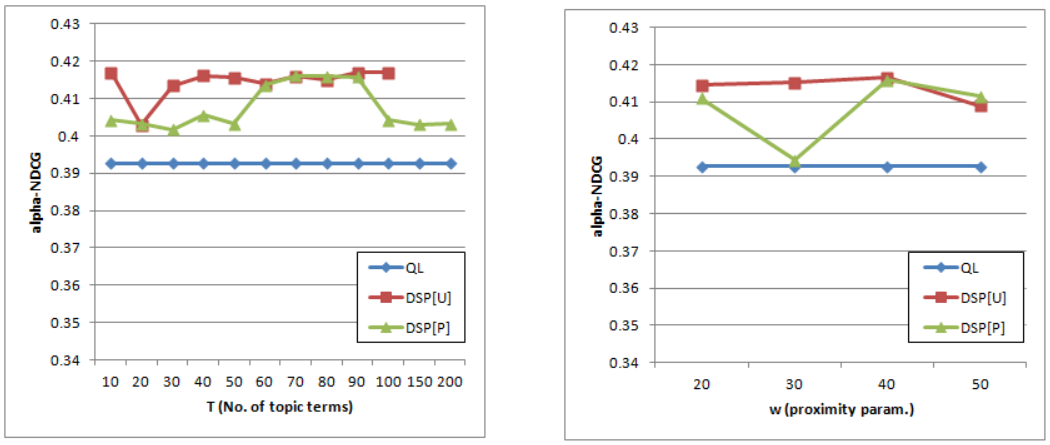

Figure 3: Sensitivity of the two parameters of `DSPApprox` (DSP for short): T (the number of topic terms used for diversification) and w (the proximity parameter). [U] and [P] indicate unigrams and phrases respectively.

9. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM*, pages 5-14, 2009.

[2] M. Bendersky, D. Fisher, and W.B. Croft. UMass at TREC 2010 Web Track: Term dependence, spam filtering and quality bias. In *Proceedings of TREC*, 2010.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. In *JMLR*, (3):993–1022, 2003.

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

[5] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of CIKM*, pages 1287-1296, 2009.

[6] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of CIKM*, pages 621-630, 2009.

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

[8] C.L.A. Clarke, M. Kolla, G.V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR*, pages 659-666, 2008.

[9] C.L.A. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Proceedings of ICTIR*, pages 188-199, 2009.

[10] C.L.A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *TREC*, 2009.

[11] C.L.A. Clarke, N. Craswell, I. Soboroff, and G.V. Cormack. Overview of the TREC 2009 Web track. In *TREC*, 2010.

[12] C.L.A. Clarke, N. Craswell, I. Soboroff, and E.M. Voorhees. In *TREC*, 2011.

[13] G.V. Cormack, M.D. Smucker, and C.L.A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. Apr 2010.

[14] N. Craswell, O. Zoeter, M.J. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of WSDM*, pages 87-94, 2008.

[15] W.B. Croft, D. Metzler, and T. Strohman. Search Engines: Information Retrieval in Practice. *Addison-Wesley*, 2009.

[16] Predicting query performance. S. Cronen-Townsend, Y. Zhou and W.B. Croft. In *Proceedings of SIGIR*, pages 299–306, 2002.

[17] V. Dang, X. Xue, and W.B. Croft. Inferring query aspects from reformulations using clustering. In *Proceedings of CIKM*, pages 2117-2120, 2011.

[18] V. Dang and W.B. Croft. Diversity by proportionality: An election-based approach to search result diversification. In *Proceedings of SIGIR*, pages 65-74, 2012.

[19] Z. Dou, S. Hu, K. Chen, R. Song, and J.R. Wen. Multidimensional search result diversification. In *Proceedings of WSDM*, pages 475-484, 2011.

[20] J. He, V. Hollink, and A. de Vries. Combining implicit and explicit topic representations for result diversification. In *Proceedings of SIGIR*, pages 851-860, 2012.

[21] T. Kanungo and D. Orr. Predicting the readability of short web summaries. In *Proceedings of WSDM*, pages 202–211, 2009.

[22] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of SIGIR*, pages 191âĂŞ202, 1993.

[23] D. Lawrie, W.B. Croft, and A. Rosenberg. Finding topic words for hierarchical summarization. In *Proceedings of SIGIR*, pages 349–357 , 2001.

[24] D. Lawrie and W.B. Croft. Generating hierarchical summaries for web searches. In *Proceedings of SIGIR*, pages 457–458 , 2001.

[25] V. Lavrenko and W.B. Croft. Relevance-Based Language Models. In *Proceedings of SIGIR*, pages 120–127, 2001.

[26] A. Ntoulas and M. Manasse. Detecting spam web pages through content analysis. In *Proceedings of WWW*, pages 83-92, 2006.

[27] F. Radlinski, M. Szummer, and N. Craswell. Inferring query intent from reformulations and clicks. In *Proceedings of WWW*, pages 1171-1172, 2010.

[28] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of SIGIR*, pages 115-122, 2009.

[29] M. Sanderson and W.B. Croft. Deriving concept hierarchies from text. In *Proceedings of SIGIR*, pages 206-213 1999.

[30] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW*, pages 881-890, 2010.

[31] C. Zhai, W.W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR*, pages 10-17, 2003.

[32] W. Zheng, X. Wang, H. Fang and H. Cheng. Coverage-based Search Result Diversification. In *Information Retrieval*, 15(5): 433-457, 2012.