Selecting Effective Expansion Terms for Diversity

Saúl Vargas
Escuela Politécnica Superior
Universidad Autónoma de Madrid
28049 Madrid, Spain
saul.vargas@uam.es

Rodrygo L.T. Santos, Craig Macdonald and Iadh Ounis
School of Computing Science
University of Glasgow
Glasgow, G12 8QQ, UK
{rodrygo,craigm,ounis}@dcs.gla.ac.uk

ABSTRACT

Query expansion has been successfully applied in Information Retrieval, mostly for adhoc search tasks. On the other hand, query expansion can also fail, particularly in light of query ambiguity. For an ambiguous query, an effective strategy is to diversify the search results, in the hope of retrieving at least one relevant result for each of the possible information needs underlying the query. In this paper, we propose to tailor query expansion to diversify the search results in order to tackle query ambiguity. In particular, we introduce a novel approach to select diverse expansion terms given a suitable partition of the feedback provided by the search users. Thorough experiments in the context of the TREC 2009, 2010 and 2011 Web tracks examine the effectiveness of our approach at improving the diversification performance of state-of-the-art query expansion techniques.

1. INTRODUCTION

Web search users typically submit short queries to represent their information needs [19]. Frequently, such queries are also ambiguous, in that they may lead to more than one *interpretation* (this could occur, for example, with acronyms or polysemic words [29]). Consider the query "apple", which could refer to the fruit, the computer company, a record label, and other less common interpretations. A user interested in one interpretation would not usually be interested in documents relevant to the others. Even when the query identifies a unique concept or entity, it may still be underspecified, in the sense that it may have different *facets*. Consider the example query "Mallorca": while it clearly refers to an island in the Mediterranean Sea, still involves uncertainty about the specific user interest, which might relate to general information about the island, touristic deals, the football team, etc. In this case, these aspects do not need to be mutually exclusive, that is, users may be interested in two or more of them [9]. Throughout this paper, unless otherwise noted, we refer to interpretations and aspects indistinctly as *subtopics* of the initial query.

OAIR '13, May 22-24, Lisbon, Portugal.

Search result diversification is an effective strategy to deal with query ambiguity. In general, diversification approaches try to present the user with documents that cover the multiple information needs underlying a query as far as possible, and as early in the ranking as possible [33]. To this end, the state-of-the-art diversification approaches in the literature employ some explicit representation of these multiple information needs [1, 30]. In particular, these approaches can be thought of as improving the representation of the query itself, by expanding it with a representation of the multiple information needs that underlie it [33].

A technique that has been traditionally used to improve the representation of queries for search is query expansion (QE). Given a set of *feedback* documents, this technique adds new terms to the original query and, additionally, assigns new weights to both the original and the new query terms [27]. Feedback documents can come from user relevance judgements (relevance feedback, RF), the top retrieved documents for the initial query (pseudo-relevance feedback or local feedback, PRF), or even the whole collection of documents (global feedback) [4].

On the one hand, query expansion has been successfully applied to improve recall in adhoc retrieval tasks [2]. On the other hand, query expansion can also fail, particularly for difficult queries [3]. One of the main causes of query difficulty is ambiguity. In particular, ambiguous queries often result in an *incoherent* feedback set [13], which in turn may lead query expansion techniques to drift away from the user's original query topic [18, 25]. Conversely, as we will show in this paper, the feedback set for an ambiguous query can be sometimes *biased* towards a single subtopic, in which case the multitude of information needs underlying the query is poorly captured by the existing query expansion techniques.

In this paper, we provide the first empirical analysis of the limitations of state-of-the-art query expansion techniques for search result diversification. Motivated by this analysis, to tackle incoherence and bias, we propose a novel term selection approach for query expansion. Our approach leverages an explicit partitioning of the feedback set produced for an ambiguous query, in order to identify effective expansion terms for promoting diversity. We thoroughly evaluate our proposed approach using the standard experimentation paradigm provided by the diversity task of the TREC 2009, 2010, and 2011 Web tracks [10, 11, 12]. The results of this investigation evaluate the effectiveness of our term selection approach at significantly improving the diversification performance of query expansion for ambiguous queries, with no significant decreases in adhoc retrieval performance.

The major contributions of this paper are:

- We analyse the limitations of current techniques for expanding ambiguous queries, in terms of both the incoherence and the bias of their feedback set.
- We propose a novel term selection approach to identify diverse expansion terms in order to overcome incoherent and biased feedback sets.
- We thoroughly evaluate our proposed term selection approach at improving the effectiveness of state-of-the-art query expansion techniques.

The remainder of this paper is organised as follows. Section 2 overviews related work on query expansion and search result diversification. Section 3 analyses the challenges posed by incoherent and biased feedback sets when expanding ambiguous queries. Section 4 introduces our diversity-oriented term selection approach, aimed to overcome these challenges. Sections 5 and 6 describe the experimental setup and the results of our thorough evaluation. Finally, Section 7 presents our concluding remarks.

2. RELATED WORK

Query expansion has been in use since Rocchio's algorithm [27], which aims at iteratively improving the representation of the user's original query towards resembling the representation of relevant documents. More recently, notable developments have been made in the language modelling framework, with the introduction of relevance models [21] and the model-based feedback approach [38], which essentially aims at improving the query language model in the presence of feedback documents. Lately, query expansion has also been tackled within the divergence from randomness (DFR) framework, with the introduction of a series of information-theoretic, non-parametric term weighting models [2]. Other developments for improved query expansion in the literature include collection enrichment approaches, which aims to identify feedback documents from alternative, higher-quality corpora, such as Wikipedia [16, 23] or the query logs of commercial web search engines [14]. Lastly, with the rise of supervised methods applied to IR, proposals to find good feedback documents [17] and good expansion terms [6, 22, 34] using machine learning have also been made for improving query expansion.

Despite its successful application for adhoc search tasks [2], query expansion can also fail, particularly when used for difficult queries [3, 13, 36]. Such queries typically result in an incoherent feedback set, which may lead to the selection of irrelevant terms [18, 25]. To overcome this problem, several approaches have been proposed to quantify query difficulty. For instance, Cronen-Townsend et al. [13] introduced a clarity score to quantify query difficulty proportionally to the divergence between the query and the collection language models. Analogously, Amati et al. [3] proposed an information-theoretic model within the DFR framework in order to predict the potential improvement brought by query expansion. In the same vein, Yom-Tov et al. [36] proposed a simpler predictor of the effectiveness of query expansion, by estimating the overlap between the top documents retrieved for the original query and those retrieved for each of the individual query terms. According to this predictor, very low and very high overlaps denoted difficult queries.

Query difficulty is often caused by ambiguity [13]. In particular, an ambiguous query may result in an incoherent feedback set, covering multiple query subtopics. While some of these subtopics may be less plausible and hence lead to topic drift, biasing the query expansion towards a single subtopic may also be detrimental, as we will show in Section 3. A sensible approach for dealing with query ambiguity in a principled manner is to diversify the retrieved documents, so as to improve the chance that different users will find at least one relevant document for their particular information need. In general, diversification algorithms aim to maximise the coverage of multiple query subtopics in the ranking, while minimising redundancy with respect to the covered subtopics. Existing diversification approaches can be generally classified as either implicit or explicit [31]. Implicit diversification approaches rely on characteristics of the retrieved documents in order to identify novel documents, such as the terms contained by these documents [7], their language models [37], or their relevance scores [35]. Alternatively, explicit approaches seek to promote documents with maximum coverage of some characteristic of the query itself, such as the categories it belongs to from a taxonomy [1] or its most frequent reformulations in a query log [30, 32].

In Section 4, we will adapt a state-of-the-art search result diversification approach in order to select diverse terms for query expansion. In particular, the xQuAD framework [30] leverages multiple reformulations of the user's original query as *sub-queries*, with each sub-query conveying a suitable representation of one of the multiple possible information needs underlying the query [33]. In our adaptation of xQuAD for query expansion, instead of diversifying the *documents* retrieved for a given query, we diversify candidate expansion *terms* selected for this query by any existing query expansion technique. As a result, we overcome incoherent and biased feedback sets by appropriately selecting expansion terms with a high coverage and low redundancy with respect to the query subtopics. In practice, as we will show in Section 6, such a diverse-oriented term selection mechanism results in a significantly improved effectiveness in terms of diversification performance, with no significant decreases in terms of standard adhoc retrieval performance.

Table 1: Subtopics for the TREC 2010 Web track query 79, "voyager". Other subtopics of this query are ommited as they have no relevants in our corpus.

subtopic	description	relevants
2	Find information about the NASA Voyager spacecraft and missions.	70
3	Find information about the television series, "Star Trek: Voyager"	9

3. INCOHERENCE AND BIAS

Existing query expansion techniques typically select expansion terms from a set of feedback documents and add these terms to the original query with appropriate weights. As discussed in Section 2, for adhoc search tasks, these techniques tend to underperform, or even harm the retrieval performance when expanding an ambiguous query, primarily due to the incoherence of the produced feedback set [3, 13, 36]. In the context of search result diversification, the

Table 2: Expansion terms for different partitions of the relevance feedback set for TREC topic 79 "voyager".

$\mathbf{q}^*_{\mathbf{all}}$	voyager	spacecraft	saturn	jupiter	solar	interstellar	nasa	uranus	probe	neptune
	1.551	0.129	0.065	0.064	0.062	0.059	0.059	0.058	0.056	0.054
$\mathbf{q}^*_{\mathbf{2}}$	voyager	spacecraft	saturn	jupiter	solar	interstellar	nasa	uranus	probe	neptune
	1.551	0.129	0.065	0.064	0.062	0.059	0.059	0.058	0.056	0.054
$\mathbf{q}^*_{\mathbf{3}}$	voyager	trek	maqui	borg	janeway	star	uss	quadrant	starfleet	officer
	1.450	0.166	0.102	0.101	0.082	0.077	0.074	0.063	0.051	0.049

incoherence problem is also present, as the existing query expansion techniques tend to select terms that are meaningful to the feedback set as a whole [6, 22, 34]. In particular, assuming that different subtopics are represented by different sets of important terms, expanding the query based on a single, incoherent feedback set may lead to the selection of excessively general terms as opposed to terms that are only important for each individual subtopic.

Besides incoherence, a second problem may also play a role. In particular, the feedback set may be *biased* towards documents covering a single, dominant subtopic. In this case, by using the feedback set as a single source of evidence for query expansion, the terms important to marginal subtopics may never be selected. While the retrieval performance with respect to the dominant subtopic may be improved, the overall coverage of the multiple subtopics underlying the original query—and hence, the diversity of the documents retrieved for the expanded query—may be degraded. For instance, consider the query numbered 79 from the TREC 2010 Web track [11]: $q =$ "voyager". This query has two distinct subtopics with relevant documents in the ClueWeb09 category B dataset.[1]—details can be found in Table 1. For this query, the relevant documents for subtopic 2 outnumber those that are relevant to subtopic 3, and there is only one document covering both subtopics.

To investigate how a standard query expansion technique would behave for such an ambiguous query, we apply the Bo1 query expansion model from the DFR framework [2], using the relevance judgements provided by TREC as feedback documents. In particular, we select the top 10 expanded terms for this query, considering three different partitions of the feedback documents: all 79 documents relevant to the query (q^*_{all}), only those 70 relevant to subtopic 2 (q^*_2), and only those 9 relevant to subtopic 3 (q^*_3). The expanded queries from these feedback sets are shown in Table 2.

From Table 2, we observe that the expansion terms obtained when considering the feedback documents covering each individual subtopic—q^*_2 and q^*_3—are highly related with the thematic of the corresponding subtopic. However, the expansion terms produced in q^*_{all} are identical to those produced in q^*_2, including their weights. In contrast, q^*_3 has nothing in common with q^*_{all}. As a result, while q^*_{all} has a high coverage of subtopic 2, it poorly covers subtopic 3.[2] Clearly, the reason for the dominance of terms related to subtopic 2 in q^*_{all} lies in the disproportionate number of relevant documents for each subtopic: 70 to 9.

[1] `http://www.lemurproject.org/clueweb09/`

[2] Arguably, the terms "spacecraft", "interstellar", and "probe" in q^*_{all} could be considered related to subtopic 3, although in a much weaker way than those of q^*_3.

Table 3: Number of relevant documents retrieved in the top 20 results for the original query and its expansions considering different feedback sets.

	nrel@20 (subtopic 2)	**nrel@20** (subtopic 3)
$\mathbf{q}$	2	0
$\mathbf{q}^*_{\mathbf{all}}, \mathbf{q}^*_{\mathbf{2}}$	17	0
$\mathbf{q}^*_{\mathbf{3}}$	1	7

As Table 3 shows, the expanded query q^*_{all}, although improving notably the global recall for subtopic 2 compared to the original query q, does not retrieve any relevant documents for subtopic 3. However, the query q^*_3 increases notably the recall for subtopic 3, but it decreases the number of relevant documents for subtopic 2. In this case, query expansion clearly improves the global recall—more relevant documents retrieved—but shows no positive effect in terms of the diversity of the retrieved documents—subtopic 3 is not covered at all. These results suggest that there is room for the improvement of the recall for both subtopics if an adequate query expansion is performed.

4. SELECTION OF EXPANSION TERMS

Given the previous example and the intuitive arguments of why standard query expansion can impede the diversity of search results, the following question arises:

> *Is it possible to adapt or modify current query expansion models to promote the diversity of search results?*

In this section, we analyse possible ways to proceed and explain in detail an adaptation of state-of-the-art query expansion models for the selection of terms from diverse, unambiguous sets of feedback documents.

As pointed out in the Section 3, we believe that one key step is to explicitly identify the subtopics each feedback document is covering and therefore limit the effects of *incoherence* or *bias* that standard query expansion models suffer in terms of diversity. Once we determine which subtopic is being covered by each document, we propose to identify and select "good" terms (as in [6, 22, 34]): given a set of candidate expansion terms, the aim is to select some of those terms in order to have an expanded query that could retrieve more diverse documents than the original query.

We propose a query expansion strategy that selects terms from the expanded queries from different sets of feedback documents. The idea is that, having identified groups of documents covering the same subtopic, a local expanded

Table 4: Expanded query after combining terms from subtopic-specific expansion terms. The subtopic each term came from is indicated between the term and the weight.

$\mathbf{q}^*_{\text{xQuAD}}$	voyager	trek	spacecraft	maqui	borg	janeway	saturn	jupiter	solar	interstellar
		3	2	3	3	3	2	2	2	2
	2.000	0.167	0.142	0.103	0.102	0.102	0.082	0.077	0.074	0.072

query is generated for each feedback group. Afterwards, a selection of the terms of those local expanded queries is made so that it maximises how well each underlying subtopic is covered with minimum redundancy. For this purpose, we adapt the xQuAD algorithm [30], originally used to re-rank search results for enhancing diversification by modelling the information needs of an ambiguous query as a set of sub-queries, as described in Section 2.

Algorithm 1 The ts_{xQuAD} algorithm for expansion terms selection. The parameter τ defines the length of the resulting expanded query.

$Q = \emptyset$
$T = \bigcup_i \{t \in q_i^*\}$
while $|Q| < \tau$ **do**
  $t^* = \arg\max_{t \in T \setminus Q} \ (1-\lambda)\ \text{P}(t|q) + \lambda\, \text{P}(t, \bar{Q}|q)$
  $T = T \setminus \{t^*\}$
  $Q = Q \cup \{t^*\}$
end while
return q^*_{xQuAD} using the terms in Q

Our proposed term selection approach, called ts_{xQuAD}, takes a set of local expanded queries $\{q_i^*\}$ as input, with each expanded query generated from feedback documents covering the same subtopic. It then performs a selection of the terms $T = \bigcup_i \{t \in q_i^*\}$ to obtain a new expanded query q^*_{xQuAD}. This selection is based on a greedy algorithm (see Algorithm 1) whose objective function

$$(1-\lambda)\ \text{P}(t|q) + \lambda\, \text{P}(t, \bar{Q}|q) \tag{1}$$

represents a trade-off—parametrised by $\lambda \in [0,1]$—between the probability $\text{P}(t|q)$ of the term t being included in the expansion that considers all feedback documents (a "relevance" component) and the probability $\text{P}(t, \bar{Q}|q)$ of the term t being drafted for the expansion of the original query q but not the previously selected terms in Q (a "diversity" component). The latter probability can be expanded by marginalising over each local expanded query q_i^*, as follows:

$$\text{P}(t, \bar{Q}|q) = \sum_{q_i^*} \text{P}(q_i^*|q)\, \text{P}(t|q_i^*) \prod_{t' \in Q} \left(1 - \text{P}(t'|q_i^*)\right) \tag{2}$$

where $\text{P}(q_i^*|q)$ takes a uniform distribution over the considered expanded queries.

The result of applying the previously described ts_{xQuAD} algorithm to the example query of Section 3 can be found in Table 4. In this case, the resulting expanded query q^*_{xQuAD} contains a selection of the expansion terms of q_2^* and q_3^*. As Table 5 shows, this new query helps retrieving a balanced number of documents covering both subtopics—6 and 4 relevant documents to subtopics 2 and 3, respectively—with some loss in terms of global recall with respect to q^*_{all}—17 relevant documents for q^*_{all} but 10 for q^*_{xQuAD}.

Table 5: Number of relevant documents of the top 20 results for the expanded query with ts_{xQuAD}.

	nrel@20 (subtopic 2)	**nrel@20** (subtopic 3)
$\mathbf{q}^*_{\text{xQuAD}}$	6	4

5. EXPERIMENTAL SETUP

In this section, we describe the experimental setup to analyse the effect of query expansion in terms of adhoc retrieval and the diversity of search results and compare some state-of-the-art query expansion techniques with our term selection strategy. In particular, we aim to answer the following questions:

- **RQ1**: What is the effect of state-of-art query expansion from PRF in terms of diversity metrics?
- **RQ2**: How does ts_{xQuAD} perform in terms of adhoc retrieval and diversity compared to existing query expansion approaches?

The experiments are carried out in the context of the diversity task of the TREC 2009, 2010 and 2011 Web tracks [10, 11, 12]. The goal of this task is to produce a ranking of documents for a given query that maximises the coverage of the possible aspects underlying this query, while reducing its overall redundancy with respect to the covered aspects.

The test collection used in this task is the ClueWeb09 Category B dataset. A total of 150 topics are available for this task. Each topic includes from 3 to 8 subtopics, as identified by TREC assessors, with relevance judgements provided at the subtopic level. The Terrier IR platform [24][3] is used for both indexing and retrieval, with Porter's stemmer and standard English stopwords removal.

To support the generality of our results for both research questions, we use different term weighting models for retrieval and query expansion. In particular, we test the following retrieval models to obtain search results for the original and the expanded queries: BM25 [26] and the DPH, TF-IDF and PL2 Divergence From Randomness (DFR) models [2]. For query expansion, the models we use for selecting and weighting expansion terms are also based in DFR, namely Bo1, Bo2 and KL [2].

The evaluation results in the diversity task of the TREC 2009, 2010 and 2011 Web tracks are reported according to five official metrics: MAP, nDCG, α-nDCG, ERR-IA and S-recall. We select MAP and nDCG [4] to measure the adhoc performance, which is the original objective of standard query expansion methods. The other three metrics evaluate diversity in different ways. The α-nDCG [9] metric, a generalisation of nDCG, balances relevance and diversity through the tuning of a redundancy parameter α, set in our experiments to 0.5. The ERR-IA metric [1], which generalises

[3] `http://www.terrier.org`

Table 6: Results for different query expansion models using PRF with all queries from TREC Web tracks 2009, 2010 and 2011. Triangles mark significant differences (Paired T-test $p < 0.05$) between query expansion variants and their non-expanded baselines: better/worse as upwards/downwards.

		MAP×10		nDCG		α-nDCG		ERR-IA		S-recall	
		10	**20**	**10**	**20**	**10**	**20**	**10**	**20**	**10**	**20**
DPH		0.123	0.200	0.144	0.141	0.193	0.221	0.139	0.147	0.331	0.402
+Bo1	**5**	0.142▲	0.231	0.145▲	0.143▲	0.183▼	0.209	0.141	0.149	0.268▼	0.347▼
	10	0.145▲	0.238	0.152▲	0.149	0.197	0.219	0.154	0.160	0.290▼	0.347▼
+Bo2	**5**	0.123▲	0.201	0.129▼	0.132▼	0.169▼	0.196▼	0.123▼	0.131▼	0.275▼	0.353▼
	10	0.129▲	0.209	0.145▲	0.141	0.181▼	0.207▼	0.133▼	0.141▼	0.299▼	0.368▼
+KL	**5**	0.140▲	0.226	0.144▲	0.141	0.184▼	0.208	0.141▲	0.148	0.284▼	0.347▼
	10	0.155	0.243	0.155▲	0.151	0.192▼	0.214	0.148▲	0.154	0.294▼	0.344▼
BM25		0.092	0.163	0.119	0.122	0.159	0.190	0.115	0.123	0.271	0.367
+Bo1	**5**	0.102▲	0.180	0.115▼	0.123▲	0.159▼	0.188	0.121	0.129	0.253▼	0.340▼
	10	0.122▲	0.208	0.129▲	0.133	0.158▼	0.188	0.115	0.123	0.263▼	0.342▼
+Bo2	**5**	0.111▲	0.196	0.115▼	0.123	0.155▼	0.177▼	0.111▼	0.117▼	0.265▼	0.320▼
	10	0.115▲	0.198	0.117▼	0.122	0.144▼	0.170▼	0.101▼	0.108▼	0.244▼	0.326▼
+KL	**5**	0.110▲	0.191	0.118▼	0.127	0.157	0.186	0.114	0.122	0.261▼	0.348▼
	10	0.124▲	0.202	0.131▲	0.133	0.159▼	0.187	0.114▼	0.122	0.268▼	0.342▼
TF-IDF		0.092	0.166	0.121	0.124	0.164	0.191	0.117	0.125	0.279	0.356
+Bo1	**5**	0.105▲	0.179▲	0.117▼	0.123▼	0.158▼	0.186	0.118	0.125	0.258▼	0.335▼
	10	0.118▲	0.201	0.131▲	0.134	0.161	0.189	0.116	0.125	0.261▼	0.342▼
+Bo2	**5**	0.109▲	0.195	0.114▼	0.122	0.156▼	0.179▼	0.113▼	0.119▼	0.257▼	0.321▼
	10	0.114▲	0.193	0.122▲	0.125	0.150▼	0.175▼	0.108▼	0.115▼	0.240▼	0.322▼
+KL	**5**	0.114▲	0.189	0.119▼	0.125	0.157	0.185	0.111	0.119	0.269▼	0.348▼
	10	0.124▲	0.208	0.133▲	0.135	0.163	0.189	0.117	0.125	0.269▼	0.334▼
PL2		0.082	0.146	0.115	0.117	0.148	0.183	0.103	0.112	0.267	0.380
+Bo1	**5**	0.084▲	0.152▲	0.111▼	0.114▼	0.138▼	0.166▼	0.096▼	0.103▼	0.250▼	0.333▼
	10	0.087▲	0.159▲	0.113▼	0.118	0.144▼	0.173▼	0.102▼	0.110▼	0.246▼	0.336▼
+Bo2	**5**	0.086▲	0.158▲	0.106▼	0.111▼	0.130▼	0.158▼	0.089▼	0.096▼	0.233▼	0.317▼
	10	0.090▲	0.161▲	0.109▼	0.115▼	0.134▼	0.164▼	0.093▼	0.101▼	0.240▼	0.325▼
+KL	**5**	0.081▼	0.151▲	0.107▼	0.113▼	0.134▼	0.163▼	0.092▼	0.100▼	0.240▼	0.329▼
	10	0.089▲	0.157▲	0.113▼	0.117▼	0.141▼	0.169▼	0.097▼	0.105▼	0.256▼	0.333▼

ERR [8], rewards the diversity of a ranking by marginalising the gain for each of the possible interpretations of a query and considering their relative importance. Finally, we compute S-recall [37], which measures the ratio of covered subtopics for a given query. In our evaluation, all five metrics are reported at two different rank cutoffs: 10 and 20. These cutoffs focus on the evaluation at ranks that are typically present in a Web search context [20].

6. EXPERIMENTAL EVALUATION

In this section we describe the experiments performed for aforementioned research questions, discuss and analyse the results for the experiments. Finally, we answer both research questions. Section 6.1 and Section 6.2 cover the experiments aimed to answer RQ1 and RQ2, respectively.

6.1 Diversity in Pseudo-Relevance Feedback

In order to answer RQ1—What is the effect of state-of-art query expansion from PRF in terms of diversity metrics?, we perform an extensive evaluation of query expansion in a PRF setting using the aforementioned retrieval and query expansion models. In particular, we evaluate the effect of using expanded queries using the first 5 and 10 retrieved results from the original, non-expanded queries. The results of this evaluation are summarised in Table 6. These differ notably in terms of adhoc and diversity metrics.

In terms of MAP at cutoff 10, almost all results from query expansion show statistically significant (Paired T-test $p < 0.05$) improvements from their respective non-expanded baselines. However, when measuring MAP at cutoff 20, although the results from using query expansion are slightly higher, they are not significant. In terms of nDCG, query expansion models do not seem to be as effective. In particular, PL2 is, almost everywhere, better than its query expansion variants. For the rest of the non-expanded baselines, the results show that using 10 documents for feedback offers some improvements, while using 5 hurts the performance with respect to the baseline. In general, query expansion models show improvements over the original queries, although these are not substantial. We postulate that this is caused by the low precision of the baselines.

In terms of diversity metrics, the results show that the original queries perform better than their expanded alternatives or, at least, not significantly worse. Particularly interesting is the case of S-recall, which accounts for the number of covered subtopics for each topic: query expansion models obtain worse results in all considered scenarios.

The previous observations support that state-of-the-art query expansion models in PRF, although being able to give some improvement in terms of MAP or nDCG, are suboptimal for diversity purposes.

Table 7: Results for xQuAD, standard (s) and diversified (d) query expansion from RF with selected queries from TREC Web tracks 2009, 2010 and 2011. Triangles mark significant differences (Paired T-test $p < 0.05$) between d and s: better/worse as upwards/downwards.

		MAP		nDCG		α-nDCG		ERR-IA		S-recall	
		10	**20**	**10**	**20**	**10**	**20**	**10**	**20**	**10**	**20**
DPH		0.057	0.084	0.269	0.256	0.328	0.365	0.257	0.268	0.566	0.658
+xQuAD		0.071	0.104	0.314	0.295	0.382	0.420	0.308	0.319	0.610	0.711
+Bo1	**s**	0.115	0.161	0.438	0.398	0.479	0.514	0.400	0.410	0.673	0.759
	d	0.117	0.160	0.432	0.387▼	0.506▲	0.538▲	0.424▲	0.433▲	0.714▲	0.792▲
+Bo2	**s**	0.107	0.148	0.406	0.367	0.456	0.489	0.381	0.391	0.646	0.732
	d	0.105	0.142▼	0.390▼	0.352▼	0.478▲	0.514▲	0.400▲	0.411▲	0.688▲	0.780▲
+KL	**s**	0.115	0.159	0.433	0.392	0.476	0.509	0.397	0.407	0.672	0.760
	d	0.111	0.152▼	0.408▼	0.371▼	0.490▲	0.525▲	0.408	0.418▲	0.714▲	0.795▲
BM25		0.040	0.061	0.195	0.195	0.233	0.279	0.177	0.190	0.427	0.577
+xQuAD		0.054	0.079	0.233	0.232	0.298	0.343	0.236	0.249	0.515	0.637
+Bo1	**s**	0.098	0.138	0.378	0.350	0.423	0.457	0.346	0.356	0.616	0.703
	d	0.095	0.129▼	0.355▼	0.324▼	0.439▲	0.471▲	0.363▲	0.372▲	0.657▲	0.740▲
+Bo2	**s**	0.098	0.135	0.371	0.338	0.421	0.453	0.351	0.360	0.598	0.689
	d	0.098	0.132	0.359	0.327▼	0.447▲	0.479▲	0.370▲	0.379▲	0.661▲	0.742▲
+KL	**s**	0.097	0.138	0.375	0.349	0.417	0.452	0.336	0.346	0.625	0.710
	d	0.095	0.130▼	0.353▼	0.323▼	0.434▲	0.467▲	0.354▲	0.364▲	0.659▲	0.742▲
TF-IDF		0.051	0.076	0.228	0.226	0.275	0.319	0.210	0.223	0.494	0.633
+xQuAD		0.062	0.092	0.261	0.260	0.330	0.377	0.261	0.275	0.571	0.693
+Bo1	**s**	0.105	0.147	0.393	0.368	0.452	0.489	0.374	0.385	0.655	0.750
	d	0.096▼	0.131▼	0.360▼	0.328▼	0.446	0.478	0.368	0.377	0.667	0.753
+Bo2	**s**	0.106	0.145	0.394	0.357	0.448	0.481	0.374	0.384	0.634	0.723
	d	0.100▼	0.136▼	0.367▼	0.334▼	0.460	0.493▲	0.383	0.393	0.671▲	0.758▲
+KL	**s**	0.103	0.145	0.392	0.365	0.446	0.481	0.362	0.372	0.673	0.752
	d	0.098▼	0.135▼	0.365▼	0.335▼	0.452	0.487	0.370	0.380	0.685	0.771▲
PL2		0.035	0.052	0.175	0.177	0.201	0.243	0.146	0.158	0.396	0.537
+xQuAD		0.042	0.064	0.197	0.203	0.240	0.289	0.182	0.195	0.449	0.586
+Bo1	**s**	0.067	0.100	0.279	0.276	0.314	0.360	0.240	0.253	0.545	0.661
	d	0.075▲	0.108▲	0.305▲	0.294▲	0.364▲	0.410▲	0.284▲	0.298▲	0.612▲	0.724▲
+Bo2	**s**	0.067	0.097	0.282	0.274	0.321	0.366	0.249	0.262	0.540	0.667
	d	0.070	0.100	0.285	0.276	0.344▲	0.388▲	0.266▲	0.279▲	0.583▲	0.701▲
+KL	**s**	0.067	0.097	0.272	0.269	0.306	0.355	0.235	0.249	0.527	0.658
	d	0.070	0.100	0.282▲	0.275▲	0.337▲	0.384▲	0.257▲	0.271▲	0.587▲	0.709▲

6.2 Diversified Selection of Expansion Terms

In order to answer RQ2—How does ts_{xQuAD} perform in terms of adhoc relevance and diversity compared to existing query expansion approaches?, we compare our proposed ts_{xQuAD} built on top of different retrieval and query expansion models with their standard variants. We test ts_{xQuAD} with a fixed $\lambda = 1.0$, that is, only considering its diversity component, because we find positive enough results without the need of tuning this parameter. Additionally, a classic xQuAD diversification algorithm using sub-queries extracted from Bing suggestions is added as a reference.

For this experiment, we consider a RF setting where feedback from the assessors for a given query is used to generate an expanded query—as in the TREC Relevance Feedback track [5]. This setting simulates a situation where users provide feedback for their interpretation of the query. Thus, the problem lies in the combination of different sources referring to possibly more than one subtopic, therefore being also subjected to possible *incoherence* or *bias* when selecting the expansion terms. Further, we assume that there is complete information about the subtopics each document covers. We want to test our term selection strategy in a scenario where no other issues involved in query expansion, apart from *incoherence* and *bias*, can significantly distort our results.

As mentioned, feedback documents are extracted from the relevance judgements provided by TREC assessors. This imposes the following constraints to our experiments:

- Feedback documents need to be removed from the relevance judgements. This evaluation procedure is known as the *residual evaluation method* [28]. This method ensures an unbiased and more realistic evaluation of feedback.

- There must be, approximately, the same number of relevant documents for each subtopic for RF and evaluation. This way, the distribution of documents for each subtopic is maintained, so it keeps stable the importance of each subtopic for evaluation.

- Each subtopic must have, at least, six relevant documents, making a total of three for RF and three for evaluation. This reference number of feedback documents for subtopic is taken from the query expansion

experiments from [2], where three or more feedback documents were used to obtain improvements in term of adhoc metrics.

The previous requirements eliminate 34 topics, leaving 116 of the 150 TREC 2009, 2010 and 2011 Web track topics for our experiment, each one with 2.95 subtopics in average. With these topics, four different splits with the previous constraints are made—to ensure independence from a particular split, hence in total there are 8 RF sets for each query that are evaluated using the rest of the documents from the relevant judgements. Finally we average the results of all the splits to see the general trend.

Table 7 shows the results of this experiment. Note that the values of the baseline retrieval models differ from those of Table 6, since we took the residuals from the relevance judgements provided by TREC. Interestingly, query expansion with RF shows clear improvements over the baselines and the xQuAD variants in all metrics. Since our query expansion models in this setting use exclusively relevant documents, they present an advantage that makes them incomparable with PRF.

On the one hand, focusing on the performance of the diversified query expansion models, the results, in general, are worse when compared to non-diversified variants in terms of adhoc metrics (MAP and nDCG), although they remain better than the baseline and the xQuAD re-ranking. PL2 is an exception, since the diversified query expansion variants show notable improvements. On the other hand, our diversified variants of query expansion obtain better results in terms of diversity metrics. The exception here is the TF-IDF baseline, where the diversified query expansion obtains worse results that the standard query expansion alternatives, although most of them are not significant.

This results conclude the following answer to RQ2: the ts_{xQuAD} algorithm improves substantially over existing query expansion alternatives in terms of diversity at a small cost in terms of adhoc relevance.

7. CONCLUSIONS

In this paper, we have analysed the suitability of query expansion techniques for search result diversification. We have shown how a partition of the feedback documents—with respect to the query subtopics covered by each document—and the selection of terms added to the expanded query positively affect the final results in terms of diversity. Specifically, we have proposed an adaptation of the state-of-the-art xQuAD framework [30] for search result diversification in order to diversify candidate expansion terms. In particular, our ts_{xQuAD} term selection mechanism chooses expansion terms from several expanded queries generated from each partition of the feedback set. A thorough evaluation of ts_{xQuAD} shows that it improves the diversity of the search results at a negligible cost in terms of adhoc relevance.

In the future, we plan to expand our term selection strategy for RF and PRF scenarios where there is no a priori information about the subtopics covered by each document—specially in PRF, where not even the relevance of the documents with respect to the query is guaranteed. We believe that this could be achieved either by using strong retrieval models that improve the effectiveness of the originally retrieved documents or by finding good feedback documents in the retrieval list [17]. In the same way, we would like to be able to automatically cluster the feedback documents to serve as input to our approach. A supervised clustering approach like that of [15] could be considered. Finally, other techniques for term selection apart from our xQuAD adaptation will also be considered.

8. ACKNOWLEDGEMENTS

The first author wants to thank the FPI-UAM program of the Universidad Autónoma de Madrid for funding his 3-month internship at the University of Glasgow. We would also like to thank M-Dyaa Albakour for his feedback on this paper.

9. REFERENCES

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

[2] G. Amati. *Probability Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, University of Glasgow, 2003.

[3] G. Amati, C. Carpineto, G. Romano, and F. U. Bordoni. Query difficulty, robustness and selective application of query expansion. In *Proc. of ECIR*, pages 127–137, 2004.

[4] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Pearson Education Ltd., 2 edition, 2011.

[5] S. Buckley, C. Robertson. Relevance feedback track overview: TREC 2008. In *Proc. of TREC*, 2008.

[6] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proc. of SIGIR*, pages 243–250, 2008.

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

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

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

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

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

[12] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the TREC 2011 Web track. In *Proc. of TREC*, 2011.

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

[14] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Query expansion by mining user logs. *IEEE Trans. on Knowl. and Data Eng.*, 15(4):829–839, July 2003.

[15] T. Finley and T. Joachims. Supervised clustering with support vector machines. In *Proc. of ICML*, pages

217–224, 2005.

[16] B. He and I. Ounis. Combining fields for query expansion and adaptive query expansion. *Inf. Process. Manage.*, 43(5):1294–1307, 2007.

[17] B. He and I. Ounis. Finding good feedback documents. In *Proc. of CIKM*, pages 2011–2014, 2009.

[18] B. He and I. Ounis. Studying query expansion effectiveness. In *Proc. of ECIR*, pages 611–619, 2009.

[19] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the Web. *SIGIR Forum*, 32(1):5–17, 1998.

[20] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: a study of user queries on the web. *SIGIR Forum*, 32(1):5–17, Apr. 1998.

[21] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. of SIGIR*, pages 120–127, 2001.

[22] C.-J. Lee, Y.-C. Lin, R.-C. Chen, and P.-J. Cheng. Selecting effective terms for query formulation. In *Proc. of AIRS*, pages 168–180, 2009.

[23] Y. Li, W. P. R. Luk, K. S. E. Ho, and F. L. K. Chung. Improving weak ad-hoc queries using wikipedia as external corpus. In *Proc. of SIGIR*, pages 797–798, 2007.

[24] C. Macdonald, R. McCreadie, R. L. T. Santos, and I. Ounis. From puppy to maturity: Experiences in developing Terrier. In *Proc. of OSIR at SIGIR*, pages 60–63, 2012.

[25] C. Macdonald and I. Ounis. Expertise drift and query expansion in expert search. In *Proc. of CIKM*, pages 341–350, 2007.

[26] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. of TREC*, pages 109–126, 1994.

[27] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. 1971.

[28] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *JASIS*, 41:288–297, 1990.

[29] M. Sanderson. Ambiguous queries: test collections need more sense. In *Proc. of SIGIR*, pages 499–506, 2008.

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

[31] R. L. T. Santos, C. Macdonald, and I. Ounis. On the role of novelty for search result diversification. *Inf. Retr.*, 15(5):478–502, 2012.

[32] R. L. T. Santos, C. Macdonald, and I. Ounis. Learning to rank query suggestions for adhoc and diversity search. *Inf. Retr.*, 2013.

[33] R. L. T. Santos and I. Ounis. Diversifying for multiple information needs. In *Proc. DDR at ECIR*, pages 37–41, 2011.

[34] R. Udupa, A. Bhole, and P. Bhattacharyya. "a term is known by the company it keeps": On selecting a good expansion set in pseudo-relevance feedback. In *Proc. of ICTIR*, pages 104–115, 2009.

[35] J. Wang. Mean-variance analysis: A new document ranking theory in information retrieval. In *Proc. of ECIR*, pages 4–16, 2009.

[36] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: Including applications to missing content detection and distributed information retrieval. In *Proc. of SIGIR*, pages 512–519, 2005.

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

[38] C. X. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proc. of CIKM*, pages 403–410, 2001.