Looking at things differently: Exploring perspective recall for informal text retrieval

Weerkamp, W.; de Rijke, M.

Publication date
2008

Published in
Proceedings of the 8th Dutch-Belgian Information Retrieval Workshop (DIR 2008)

Looking at Things Differently

Exploring Perspective Recall for Informal Text Retrieval

Wouter Weerkamp
weerkamp@science.uva.nl

Maarten de Rijke
mdr@science.uva.nl

ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam

ABSTRACT

When retrieving informal text such as blogs, comments, contributions to discussion forums, users often want to uncover different perspectives on a given issue. To help uncover perspectives, we examine the use of query expansion against multiple external corpora. We consider two informal text retrieval tasks: blog post finding and blog finding. We operationalize the idea of uncovering multiple perspectives by query expansion against multiple corpora from different genres. We use two approaches to incorporate these perspectives: as a rank-based combination of runs and a mixture of query models.

The use of external sources does indeed generate different views on a topic as becomes clear from the unique relevant results identified by the expanded runs compared to the baseline run. Even after combining the expanded run with the original run, unique relevant documents are found by both of the perspectives. As to the combination methods, the mixture of query models outperforms the rank combination, and leads to significant improvements in MAP score over the baseline.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Algorithms, Experimentation, Measurement, Performance

Keywords

Blog distillation, Post retrieval, Query expansion, Perspective recall

1. INTRODUCTION

In this paper we report on preliminary work concerning perspective recall in the setting of user generated content. We take a perspective on a topic to refer to an attitude, opinion or set of beliefs regarding the topic. Perspective recall should be distinguished from *aspect* (or subtopic) *recall*. The latter has been identified as an important cause of error or even failure of retrieval engines [3, 8]. Retrieval systems often fail because they fail to identify and return important aspects of a given topic, zooming in on a single aspect only. This effect is often exacerbated by query expansion techniques as they tend to emphasize the dominant angle on a topic. In this work we focus on perspectives rather than aspects: given a topic there are various ways to look at, and write about, it.

In the setting of user generated content (blogs, discussion fora, etc), the issue of multiple perspectives on a topic is particularly relevant. The ability to successfully uncover perspectives is an important ingredient of systems that aim to effectively address information needs such as "What do people think about X?" There are many motivations why people blog. Some blogs or posts may be aimed at sharing original insights or knowledge on a particular topic, others reflect on a topic that is currently popular in the news media, and still others share stories that happened in the blogger's personal life [2]. We assume that a blog post relevant to a given topic will, to some degree, reflect the blogger's motivation and perspective. That is, an "insight blogger" will use language that is commonly associated with the topic being blogged about, while a "piggyback blogger" who follows the news will share much of her vocabulary with the news reports on which she piggybacks. Our working hypothesis, then, is that by observing the language usage around a topic in these respective sources (i.e., an authoritative source, or a news source), we may be able to uncover additional perspectives of a topic.

Specifically, in order to uncover perspectives on a topic, we propose to perform query expansion using terms that are sampled from multiple corpora from different genres. We assume that the effect of selecting expansion terms the target corpus (from which documents need to be retrieved) is to strengthen perspectives that are already present or even dominant in the top ranked results. Instead, we expand queries against external corpora for revealing additional perspectives on the topic. Below, we compare the impact on retrieval effectiveness of performing query expansion against multiple external corpora—where we assume that different corpora capture different angles.

In this paper our target corpus is a blog corpus, and for capturing additional perspectives we consider corpora from two aditional genres: a news corpus and Wikipedia as external corpora, the former capturing the zeitgeist at the time a topic is being discussed in the blogosphere, and the latter capturing the established vocabulary around the topic. We compare different ways of merging the "news perspective" and the "Wikipedia perspective" with the original query, using (i) rank combinations, and (ii) a mixture of query

DIR-2008 April 14–15, 2008, Maastricht, the Netherlands.

models. Next, we analyze the impact on the retrieval performance, both in terms of traditional measures such as MAP and MRR, and in terms of the relevant documents identified (only) by these new perspectives and how they are ranked.

The remainder of the paper is organized as follows. We discuss related work in Section 2, detail our methods and experimental setup in Section 3, report on our results in Section 4, which we then analyze in Section 5, before concluding in Section 6.

2. RELATED WORK

Related work comes in several kinds. Since the launch of the TREC Blog track in 2006 [15], blog post retrieval has been one of the tasks assessed at the track; the full task being examined at TREC is to retrieve opinionated blog posts in response to a query, i.e., posts that are about the topic and in which an opinion (positive or negative) about the query topic is being expressed. Here we only consider the "topical sub task": to identify posts about a given topic. An important finding from the 2006 and 2007 editions of the TREC Blog track is that good performance on the opinionated post finding task is strongly dominated by good performance on the underlying topic-relevance task [14, 15], which motivates the present work.

Most participants in the blog post retrieval task use a two-stage retrieval approach, in which the first step identifies possible relevant posts, and a second step is used to determine opinionatedness of the post [14]. The first, relevance identifying step is usually done with standard approaches like language modeling [6], vector-space models [24], BM25 [22, 25], and Divergence From Randomness [7]. On top of these basic models, extra features are used to improve topical retrieval; in [24] blind relevance feedback on the top results is followed by a re-weighting step using an SVM classifier. Query expansion using an external source is used in [6, 25] with encouraging results. Best topical performance is achieved by [25] who apply an extensive retrieval system, based on concept identification and query expansion, use these to issue the query and apply a post-retrieval filtering step to remove spam.

As part of the TREC 2007 Blog track [14] a new task was introduced: blog distillation. The aim of this task is to rank blogs rather than individual blog posts given a topic; this is summarized as *find me a blog with a principle, recurring interest in X*. The scenario underlying this task is that of a user wanting to add feeds of blogs about a certain topic to his or her RSS reader. This task is different from a filtering task [17] in which a user issues a repeating search on posts, constructing a feed from the results. TREC 2007 witnessed a broad range of approaches to the task of identifying key blogs. These approaches are usually different in the documents they use to construct the retrieval index: either blog posts, or full blogs (i.e., concatenated text from posts). The former is examined by various participants [5, 6, 20], but seems to perform worse then its blog counterpart (when comparisons are made [5, 20]). Besides these separate uses of posts or blogs, several approaches are introduced that use some kind of combination of the two [19, 20]. Results are mixed: in [19] the initial run on a blog index performs better than the final combination with post characteristics, whereas in [20] the combination of post and blog relevance performs better than any of the single approaches. Finally some approaches were proposed that incorporate external knowledge or information in the blog distillation task: Lee and Lommatzsch [12] uses Web 2.0 applications (like Wikipedia, WordNet, and Dmoz) to construct topic maps and uses a classifier to determine relevance of blogs. In [5] the best performing run of all participants is constructed using query expansion on Wikipedia on a blog index.

Query modeling, i.e., transformations of simple keyword queries into more detailed representations of the user's information need (e.g., by assigning (different) weights to terms, expanding the query, or using phrases), is often used to bridge the vocabulary gap between the query and the document collection. Many query expansion techniques have been proposed, and they mostly fall into two categories, i.e., global analysis and local analysis. The idea of *global* analysis is to expand the query using global collection statistics based, for instance, on a co-occurrence analysis of the entire collection. Thesaurus- and dictionary-based expansion as, e.g., in [16], also provide examples of the global approach.

Our focus in this paper is on *local* approaches to query expansion, that use the top retrieved documents as examples from which to select terms to improve the retrieval performance [18]. In the setting of language modeling approaches to query expansion, the local analysis idea has been instantiated by estimating additional query language models [10, 21] or relevance models [11] from a set of feedback documents. Yan and Hauptmann [23] explore query expansion in the setting of multimedia retrieval.

Kurland et al. [9] provide an iterative "pseudo-query" generation technique to uncover multiple aspects of a query, using cluster-based language models; however, they do not bring in external corpora to uncover additional perspectives, like we do for perspectives. Finally, Diaz and Metzler [4] were the first to give a systematic account of query expansion using an external corpus in a language modeling setting, although their motivation was not to uncover additional perspectives of a topic but to improve the estimation of relevance models.

3. EXPERIMENTAL SETUP

Following the scenario sketched in Section 1 and the previous work in this area, we perform several experiments regarding perspective recall for informal text. This section elaborates on the research questions we address, the various experiments we deploy to answer these question, and the resources and measures we use in the process.

3.1 Research questions

We address the following five questions in this paper:

1. Can we improve both blog post retrieval and blog distillation by introducing new perspectives in the results? In our experiments we use two external sources to introduce new perspectives through query expansion and we compare the results of these expanded runs to our baseline runs.

2. How does expansion on the target corpus perform compared to expansion on external sources? Besides using the external sources, we also try to incorporate perspectives from the blog corpus itself; we then compare performance of these runs to the baseline and the other expanded runs.

3. What are effective ways to combine perspectives from different sources? Combination of perspectives can be done in various ways: we experiment with and compare two ways of combining.

4. Is there a difference between the tasks of blog post retrieval and blog distillation when it comes to perspective recall? We do not only focus on blog post retrieval, but also explore the impact of perspective recall on blog distillation. An analysis of the results for both tasks should yield insights in the influence of task on perspective recall.

5. To what extent do the expanded runs retrieve different perspectives? To be sure that we are actually retrieving different

perspectives with our external sources, we perform a qualitative comparison of the resulting runs in terms of uniquely retrieved documents and using rank correlation measures.

3.2 Retrieval approach

Our retrieval system of choice is an out-of-the-box implementation of Indri.[1] The reasons for choosing Indri as retrieval system are twofold: First, it shows excellent out-of-the-box performance [6] and second, the Indri query language allows us to use weighted queries. The latter means that a query model with n terms with different probabilities (see Sections 3.3 and 3.4) can easily be translated to a query like *#weight(* $w_1 term_1 \ldots w_n term_n$ *)*.

3.3 Relevance models

One way of expanding the original query is by using blind relevance feedback: assume the top M documents to be relevant given a query. From these documents we sample terms that are then used to form the expanded query model $\hat{Q}$. Lavrenko and Croft [11] suggest a reasonable way of obtaining $\hat{Q}$, by assuming that $p(t|\hat{Q})$ can be approximated by the probability of term t given the (original) query Q. We can then estimate $p(t|Q)$ using the joint probability of observing the term t together with the query terms $q_1, \ldots, q_k \in Q$, and dividing by the joint probability of the query terms:

$$p(t|\hat{Q}) \approx p(t|Q) = \frac{p(t, q_1, \ldots, q_k)}{p(q_1, \ldots, q_k)} = \frac{p(t, q_1, \ldots, q_k)}{\sum_{t'} p(t', q_1, \ldots, q_k)}. \quad (1)$$

In order to estimate the joint probability $p(t, q_1, \ldots, q_k)$, Lavrenko and Croft [11] propose two methods. The two methods differ in the independence assumptions that are being made. Based on previous experiments we assume that query words $q_1, \ldots, q_k$ are independent of each other, but we keep their dependence on t:

$$p(t, q_1 \ldots q_k) = p(t) \prod_{i=1}^{k} \sum_{D \in M} p(D|t) p(q_i|D). \quad (2)$$

That is, the value $p(t)$ is fixed according to some prior, then the following process is performed k times: a document $D \in M$ is selected with probability $p(D|t)$, then the query word q_i is sampled from D with probability $p(q_i|D)$.

A final decision we need to make is to determine the number of documents we use in relevance feedback, and the number of terms we extract from these documents. Based on previous experiments and as reported by [5] we choose to use the top 40 documents, and extract the top 20 terms from these.

3.4 Combination methods

Combining perspectives can be done in various ways. In this paper we experiment with two obvious combination methods: run combination based on ranking, and a so-called mixture of query models (MoQM). The former uses two result lists (original and expanded), takes the weighted sum of the ranks of each result, and orders them by the inverse of the resulting:

$$Score(d) = \lambda Rank_{ori}(d) + (1 - \lambda) Rank_{exp}(d) \quad (3)$$

The second approach adds the original query terms to the newly extracted terms with a certain mixture weight and normalizes all weights to sum to 1. This mixed query is then used to retrieve the final result list. Eq. 4 formalizes this approach:

$$p(t|\theta_Q) = \lambda \cdot p(t|Q) + \frac{(1 - \lambda) \cdot p(t|\hat{Q})}{\sum_{t' \in \hat{Q}} p(t'|\hat{Q})}, \quad (4)$$

where $p(t|Q) = n(t, Q) \cdot |Q|^{-1}$, $n(t, Q)$ is the number of occurrences of term t in query Q and $|Q|$ is the length of query Q. $p(t|\hat{Q})$ is defined in Eq. 1. The resulting query model contains at least 20 and at most $20 + |Q|$ terms and is translated into an Indri query, using $p(t|\theta_Q)$ as weights, as explained in Section 3.2.

3.5 Collections

We perform our experiments on the TRECBlog06 collection [13]. This collection was constructed as part of the TREC 2006 Blog track [15] and consists of 3.2 million blog posts from 100,000 different blogs, collected during an 11 week period (December 2005–February 2006). We use the permalinks of the posts (the blog post in HTML format), and ignore other data in the collection (e.g. syndicated content and homepages). For the construction of the blog index we aggregate all blog posts from one blog into one document, and use these documents to construct the index.

The external news source we use is the AQUAINT-2 corpus [1]. The AQUAINT-2 collection consists of newswire articles that are roughly contemporaneous with the TRECBlog06 collection. The collection comprises approximately 2.5 GB of text (about 907K documents) spanning the time period of October 2004–March 2006. Articles are in English and come from a variety of sources. We select the articles covering the same period as the the blog corpus, resulting in a final index of 149,500 documents.

Finally, we chose Wikipedia as the external qualitative source in our experiments. The Wikipedia collection we use is a crawl of the English Wikipedia from August 2007 and contains 3.1 million articles.

3.6 Tasks

The two tasks we compare are blog post retrieval and blog distillation. The former has been part of TREC Blog track for two consecutive years (2006 and 2007) which results in a total number of 100 topics and relevance assessments for this task.

The blog distillation task was introduced at TREC 2007. A total of 45 assessed topics is available.

For both tasks we use the title field (T) only to construct queries, and ignore additional information like description and narrative.

3.7 Evaluation measures

We evaluate our runs using the usual IR measures: average precision (AP) for each topic, the mean of the average precision over all topics (MAP), early precision (p@10), and the mean reciprocal rank (MRR). Besides these obvious measures, we also evaluate our runs using measures that aim to directly compare two runs in terms of the rankings they assign to documents and in terms of the number of unique relevant documents they manage to identify. First, we look at the Spearman's rank correlation coefficient. This coefficient determines the correlation between two ranked lists and reports on the ρ value (where $\rho = 1$ and $\rho = -1$ indicate perfect correlation).

Second, we introduce the mean uniqueness of a run at different cut-off levels n. This metric looks at the relevant results within the top n of run A that are *not* present in the full result list of run B. We divide the number of unique results by n to get the uniqueness@n score for a query. Averaging these values over all test queries leaves us with the mean uniqueness@n (MU@n). The formal definition

[1] `http://www.lemurproject.com/indri`

of MU@n is given in Eq. 5.

$$U@n(q; A, B) = \frac{\sum_n R(D_n, q, A, B)}{n}$$
$$MU@n(A, B) = \frac{\sum_{q \in Q} U@n(q; A, B)}{|Q|}, \quad (5)$$

where $R(d, q, A, B)$ is a binary function indicating whether or not document d is relevant given query q *and* is present in result list A and not in result list B.

4. RESULTS

In this section we describe the results of the experiments performed to answer our research questions. We start by establishing a baseline, and continue with incorporating alternative perspectives into the results.

4.1 Baselines

We start by reporting on our baseline scores; these scores are obtained by issuing the original query to the post and blog indexes. The results of the baseline runs are listed in Table 1.

Table 1: Baseline results for both tasks, and all years.

task	MAP	p@10	MRR
post'06	.3213	.6720	.7236
post'07	.4327	.6820	.7558
blog	.3224	.4378	.6141

The baseline scores reported in Table 1 are competitive, and well above the median performance achieved by participants at TREC 2006 and 2007.

4.2 Expanded runs

To increase perspective recall and combine multiple perspectives on a topic, we issue the 145 queries to two external corpora as detailed in Section 3.5. From the top results of each initial query we select the terms with the highest weight, as explained in Section 3.3. These newly selected terms are then used as a query on the original blog (or post) index to retrieve the final result list. We first examine the performance of these expanded queries, without mixing them with the original query or results (i.e., the original query terms might not be present anymore, or are 'overpowered' by new terms). Table 2 reports on the performances of these purely expanded runs.

Table 2: Results of the expanded runs (without mixing the original queries); best scores per year/track and metric in boldface.

task	expansion corpus	MAP	p@10	MRR
post'06	news	.2168	.5400	.6555
	wikip.	.2957	**.6340**	**.7510**
	posts	**.2986**	.6100	.6729
post'07	news	.1801	.4140	.4749
	wikip.	**.3743**	**.6200**	.6668
	posts	.3389	.5940	**.6679**
blog	news	.2813	**.4867**	.6547
	wikip.	**.3434**	.4778	**.7251**
	blogs	.2629	.3311	.5014

Table 3: Perspective difference of the expanded runs with baseline results

task	exp. corpus	MU @10	MU @100	MU @1000	$avg.\rho$	$\lvert\rho\rvert > .5$
post'06	news	.0060	.0242	.0266	.5003	48%
	wikip.	.0000	.0070	.0204	.6632	76%
	posts	.0020	.0050	.0224	.6076	80%
post'07	news	.0220	.0190	.0143	.4989	35%
	wikip.	.0040	.0028	.0072	.6425	67%
	posts	.0020	.0038	.0080	.5636	70%
blog	news	.0444	.0571	–	.5054	36%
	wikip.	.0177	.0531	–	.5917	80%
	blogs	.0133	.0338	–	.5777	84%

As expected, most scores are lower than the baseline scores, because the original query is not yet mixed back in. An exception is the blog retrieval run using Wikipedia: this run improves over the baseline, but only the improvement on MRR is significant. Since we are not so much interested in these runs alone, but in the combination of perspectives (i.e., mixing back the original query), we look at how different these runs are from the baseline runs. Do they really retrieve a different perspective, or is it merely reranking results that causes the differences in scores?

We list the results of the run comparisons in Table 3. We report the mean uniqueness values for $n = 10$, 100, 1000; note that for the blog retrieval task we only retrieve the top 100 (hence the missing MU@1000 values). Additional scores reported are Spearman's ρ (averaged over all topics, and computed by comparing the expanded run against the baseline run) as well as the fraction of topics that have an absolute ρ value above .5: these are topics for which high correlation values are found. Again, correlation is computed between the run whose results are reported and the baseline.

When looking at the results we see that, in general, expansion against the news corpus seems to generate more unique relevant results than expansion against the Wikipedia corpus. This is supported by the results of the rank correlation: the average correlation between the baseline run and runs based on expansion against the news corpus is lower than the average correlation between the baseline and runs based on expansion against Wikipedia, while the fraction of highly correlated topics ($|\rho| > .5$) is lower for the news-based runs than for the Wikipedia-based ones. We hypothesize that expansions with higher MU values and lower average ρ and $|\rho| > .5$ values will lead to bigger improvements in retrieval effectiveness when mixed in with the original query.

4.3 Combined perspectives

Next, we consider the retrieval performance on the post and blog finding tasks when we combine the baseline query ("the original perspective") with an expanded query ("an alternative persective").

As detailed in the previous section, both combination methods that we consider in this paper—rank-based and based on a mixture of query models—have a parameter λ that determines the weight of the original perspective in the final results; we use sweeps to determine the optimal value for λ on each task and corpus. The plots in Figure 1 show the values of λ compared to the MAP of the combined runs.

Similarly the plots in Figure 2 show which λ settings for the MoQM method lead to the best performance on MAP for each task.

Finally, we list the results of the optimal settings in Table 4. We use a two-tailed paired t-test to test for significant differences between the mixed runs and the baseline: $^{\blacktriangle}$ or $^{\blacktriangledown}$ indicate a significant

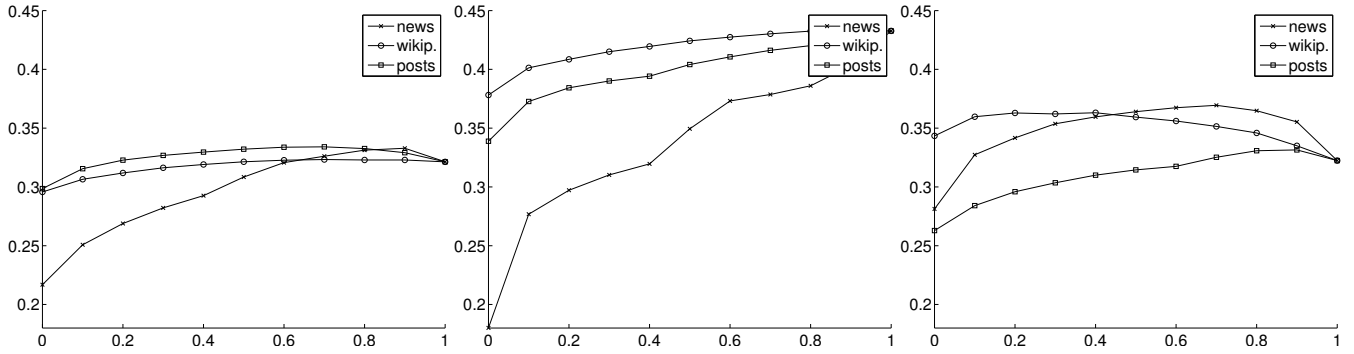

Figure 1: MAP against λ on rank combination; (Left): post 2006, (Center): post 2007, (Right) blog.

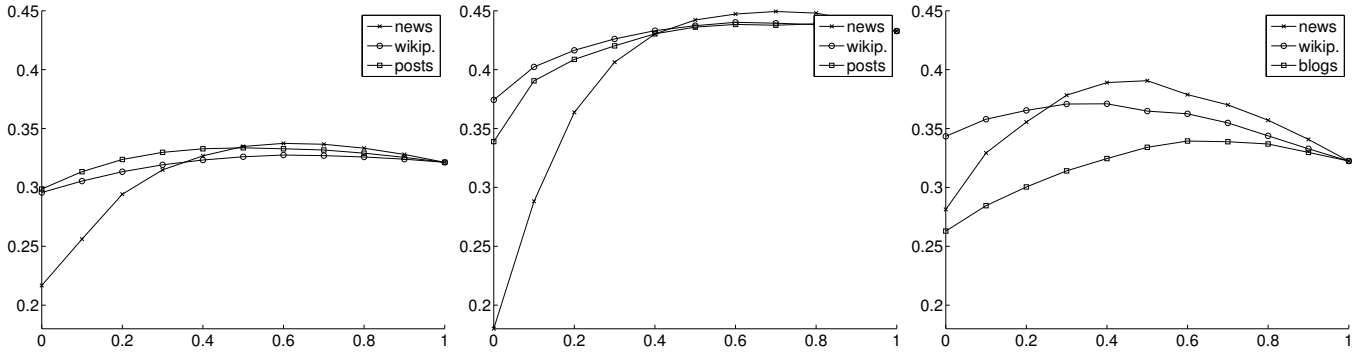

Figure 2: MAP against λ on MoQM; (Left): post 2006, (Center): post 2007, (Right) blog.

improvement or drop on $\alpha = .01$, $^{\triangle}$ and $^{\triangledown}$ indicate the same on $\alpha = .05$.

A few observations are in order. In terms of MAP, the combination runs nearly always improve over the baseline, and combinations based on the mixture of query models tend to improve more than the simple rank-based improvements. In absolute terms, the MoQM combination runs outperform the highest MAP scores reported in the literature for the 2006 post finding task as well as for the blog distillation task. Finally, the MAP-optimized combinations on which we report in Table 4 do not always lead to improvements on other, more precision-oriented meaures over the baseline; for instance, the Wikipedia-based MoQM expansion for 2006 performs less than the baseline.

Finally, we see that mixing in the originial query with expansions that have higher MU values and lower average ρ and $|\rho| > .5$ values tends to lead to bigger improvements in retrieval effectiveness over the baseline: the expansion with terms from the news generally scores highest, for each of the tasks/years.

5. ANALYSIS

We analyze our results in two ways: first, we compare the best MoQM news, Wikipedia, and blog (post) perspective runs on each task with respect to the number of (unique) relevant documents retrieved. Second, we perform a per-topic analysis to determine to which extent scores show different behavior depending on perspective.

Table 5 lists the results of the number of relevant documents retrieved per task by the different perspectives. Compared to the baseline runs we identify the number of relevant documents that are added to the result list by each perspective, and also the number of documents lost by using this perspective.

Looking at the differences between documents added and lost, we conclude that the news perspectives gives best results, which is translated to the best overall performance in Table 4. The Wikipedia perspective follows at close range, while the target index (the blog perspective) shows only modest improvements, as can be expected.

We take the analysis of the number of retrieved relevant documents a step further and compare per task all runs: three perspectives and the baseline. For each run we determine the number of unique relevant documents within the top 1000 results. The results of this comparison are displayed in Table 6. These results show us that even when comparing all these runs together, every single run (be it baseline or perspective) still retrieves unique relevant documents. In line with our earlier findings, the news perspective returns most unique relevant documents.

To compare the performance of both perspectives relative to the baseline we plot the difference in AP per topic between the baseline and each of the runs. Next, we sort these by decreasing difference for the news runs, resulting in the plots shown in Figure 3. The general trend on the 2006 post retrieval topics is that the news perspective run performs better than the Wikipedia perspective on most topics; it improves over the baseline for more topics than it hurts, and also the absolute increases (in terms of AP) are bigger than the decreases: 31 out of 50 are helped by the news perspective, while 29 and 33 out of 50 are helped by the Wikipedia and post perspective. A topic displaying interesting behavior is topic 858, "super bowl ads." Using a news or post perspective, the topic improves over the baseline, but using the Wikipedia perspective we get the biggest drop. Table 7 lists the top weighted terms for all three query models. The table suggests that it is probable that the original query itself is already quite specific, making it hard to add good terms to the query. The Wikipedia expansion adds some

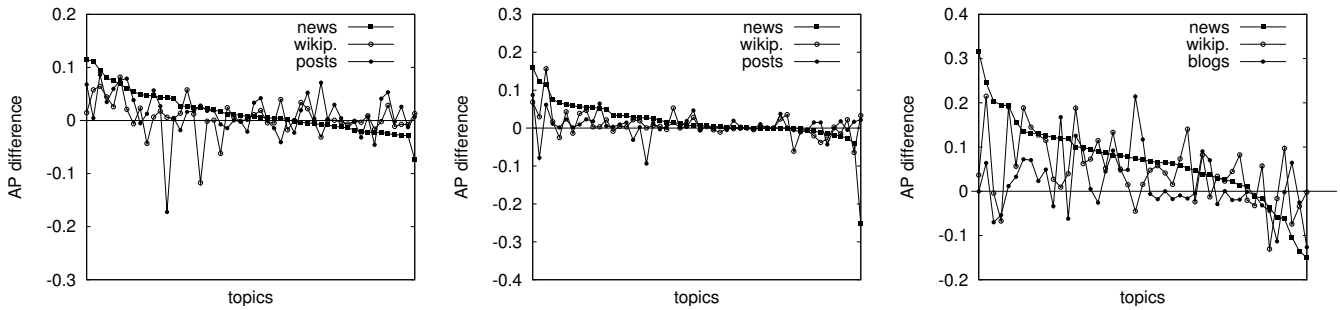

Figure 3: AP differences between baseline and news, Wikipedia and blogs runs; (Left): post 2006, (Center): post 2007, (Right) blog.

Table 4: Results of the mixed runs. Best scores per task in boldface.

task	comb. method	exp. corpus	λ	MAP	p@10	MRR
post'06				.3213	.6720	.7236
	MoQM	news	0.6	**.3373**$^{\blacktriangle}$	**.7020**	**.7700**
		wikip.	0.6	.3275	.6900	.7025
		posts	0.5	.3336$^{\triangle}$	.6420	.6944
	Rank	news	0.9	.3329$^{\blacktriangle}$	.6860	.7606
		wikip.	0.7	.3233	.6980	.7269
		posts	0.7	.3341$^{\triangle}$	.6520	.7346
post'07				.4327	.6820	.7558
	MoQM	news	0.7	**.4494**$^{\triangle}$	**.7400**$^{\blacktriangle}$	**.8140**
		wikip.	0.6	.4402	.7040	.8099
		posts	0.8	.4389	.6720	.7873
	Rank	news	1.0	.4327	.6820	.7558
		wikip.	0.9	.4346	.6960	.7678
		posts	1.0	.4327	.6820	.7558
blog				.3224	.4378	.6141
	MoQM	news	0.5	**.3906**$^{\blacktriangle}$	**.5333**$^{\blacktriangle}$	.7800$^{\blacktriangle}$
		wikip.	0.4	.3710$^{\blacktriangle}$	.4933$^{\blacktriangle}$	.6830
		blogs	0.6	.3394	.4377	.5651
	Rank	news	0.7	.3694$^{\blacktriangle}$	.5311$^{\blacktriangle}$	**.7932**$^{\blacktriangle}$
		wikip.	0.4	.3631$^{\blacktriangle}$	.4888$^{\triangle}$	.6891
		blogs	0.9	.3315	.4400	.5939

uninformative or even off-topic terms like *2005*, *tokyo* and *categories*. The news and post perspectives, on the other hand, add terms like *ad*, *commercials*, *advertising*, and *advertisers*, that seem much more revealing in this case.

A similar analysis is possible for the post retrieval runs on the 2007 topics. First, 33 out of 50 are helped by the news perspective, while 30 out of 50 are helped by the Wikipedia perspective. The post perspective also helps 33 out of 50 topics. From the plot (Figure 3) we observe (at the far right-hand side of the plot) a topic has a huge drop in performance from the news perspective; topic 902. This topic concerns "lactose gas," something that proves to be difficult for the news perspective. Table 8 lists the top 10 terms for all three perspectives. A typical example of topic shift: the news perspective is completely off and shifts to natural gas, and the gas crisis between the Ukraine and Russia (with Gazprom as a major player). The Wikipedia and post perspectives stick to the original topic, but do not really add new insights, making them perform only slightly above the baseline. In the plots in Figure 3 we see topic 902 all the way at the right, and it is the only very bad performing topic

Table 5: Comparison of the impact on the number of relevant documents retrieved per task

	Relevant retrieved			
task	baseline	added/lost by target	added/lost by news	added/lost by Wikip.
post'06	11,951	610/534	685/441	470/346
post'07	8,229	109/64	245/151	158/94
blog	1,072	95/84	195/82	188/91

Table 6: Comparison of the impact on the number of unique relevant documents retrieved per task

	Unique relevant retrieved			
task	baseline	target	news	Wikip.
post'06	125	170	234	102
post'07	23	22	125	72
blog	18	23	62	57

in the news perspective.

The final task, blog distillation, shows a very noisy graph. In general, the news perspective improves more over the baseline than the Wikipedia perspective, but the latter also improves on most topics (37 out of 45 for the news, 33 out of 45 for Wikipedia). The blog perspective hurts in 23 out of 45 topics and shows thus the worst performance of the three perspectives, although still slightly above the baseline. An interesting topic is visible at the beginning of the plot: topic 994 ("formula f1") shows a huge improvement for the news perspective (+.2 MAP) and a small decrease on MAP for the Wikipedia (−.004) and blogs (−.07) perspective. What causes this difference? Table 9 again lists the most important terms for all three perspectives. Although not completely wrong, and an interesting additional *aspect*, Wikipedia produces mostly terms that have to do with a Formula 1 game for the Playstation console. The news perspective reveals an angle more closely related to the actual competition and returns informative terms like *driver*, *super aguri*, and *championship*. The blogs perspective introduces too much noise to be helpful, e.g., *dvd*, *saf1*, and *net*.

6. CONCLUSIONS

In this paper we focused on "perspective recall" and tried to use query expansion against external corpora to uncover additional perspectives on a given topic, thereby hoping to improve retrieval performance. By combining perspectives from the blog corpus with either the Wikipedia corpus or a news corpus, we tried to improve

Table 7: Term weights for the query 858 "super bowl ads" for (Left): Wikipedia, (Center): news, and (Right): post perspective.

$p(t\|\theta_Q)$	t	$p(t\|\theta_Q)$	t	$p(t\|\theta_Q)$	t
.3000	bowl	.2634	super	.2727	bowl
.2991	super	.2620	bowl	.2710	super
.2000	ads	.2398	ads	.2343	ads
.0231	nfl	.0261	year	.0272	commercials
.0219	game	.0250	ad	.0261	game
.0141	tokyo	.0208	game	.0213	ad
.0132	football	.0182	commercials	.0210	2006
.0128	team	.0175	million	.0127	nfl
.0116	2005	.0138	time	.0118	advertising
.0115	categories	.0132	advertisers	.0104	30

Table 8: Term weights for query 902 "lactose gas" for (Left): Wikipedia, (Center): news, and (Right): post perspective

$p(t\|\theta_Q)$	t	$p(t\|\theta_Q)$	t	$p(t\|\theta_Q)$	t
.3979	lactose	.4353	gas	.4342	lactose
.3196	gas	.3500	lactose	.4297	gas
.0308	milk	.0279	ukraine	.0215	intolerance
.0248	intolerance	.0201	russia	.0140	foods
.0217	categories	.0152	natural	.0136	milk
.0216	lactase	.0132	energy	.0135	people
.0211	products	.0130	cubic	.0071	intestine
.0201	0	.0128	russian	.0065	symptoms
.0182	bacteria	.0117	supplies	.0061	digestive
.0167	flatulence	.0107	gazprom	.0056	sugar

retrieval effectiveness. We use two approaches to incorporate these perspectives: rank-based combination of two runs, and a mixture of query models (MoQM). We test these methods on two tasks: blog post retrieval and blog distillation.

Our results show that the use of external sources does indeed generate a different view on a topic; this becomes clear from the uniqueness of the expanded runs compared to the baseline run. Even after combining the expanded run with the original run, unique relevant documents are found by both of the perspectives. As to the two methods of combination, the MoQM performs better than the rank combination, and leads to mostly significant improvements in MAP score over the baseline.

Ongoing and future work is focused on several options: to explore different ways of combining perspectives; to combine more than one perspective with the baseline; to find per-topic weights of different perspectives; to try different ways of term selection; and to explore perspective recall in other settings, not just informal text. Finally, we want to determine the relation between aspects of, and perspectives on a given topic.

7. ACKNOWLEDGMENTS

We would like to thank our reviewers for their valuable comments. Both authors were supported by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104. Maarten de Rijke was also supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13-001, 612.000.106, 612.066.302, 612.-069.006, 640.001.501, and 640.002.501.

8. REFERENCES

[1] AQUAINT-2, 2007. URL: `http://trec.nist.gov/data/qa/2007_qadata/qa.07.guidelines.html#documents`.

[2] R. Bhargava. The 25 basic styles of blogging and when to use each one, 2007. `http://www.slideshare.net/rohitbhargava/`.

[3] C. Buckley. Why current IR engines fail. In *SIGIR '04*, pages 584–585, 2004.

[4] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA, 2006. ACM.

[5] J. Elsas, J. Arguello, J. Callan, and J. Carbonell. Retrieval and feedback models for blog distillation. In *TREC 2007 Working Notes*, 2007.

[6] B. J. Ernsting, W. Weerkamp, and M. de Rijke. The University of Amsterdam at the TREC 2007 Blog Track. In *TREC 2007 Working Notes*, 2007.

[7] D. Hannah, C. Macdonald, J. Peng, B. He, and I. Ounis. University of Glasgow at TREC 2007: Experiments in Blog and Enterprise Tracks with Terrier. In *TREC 2007 Working Notes*, 2007.

[8] D. Harman and C. Buckley. The NRRC reliable information access (RIA) workshop. In *SIGIR '04*, pages 528–529, 2004.

[9] O. Kurland, L. Lee, and C. Domshlak. Better than the real thing?: Iterative pseudo-query processing using cluster-based language models. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, New York, NY, USA, 2005. Acm.

[10] J. Lafferty and C. Zhai. Probabilistic relevance models based on document and query generation. In *Language Modeling for Information Retrieval*, Kluwer International Series on Information Retrieval. Springer, 2003.

[11] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01*, pages 120–127, 2001.

[12] W.-L. Lee and A. Lommatzsch. Feed distillation using adaboost and topic maps. In *TREC 2007 Working Notes*, 2007.

Table 9: Term weights for query 994 "formula f1" for (Left): Wikipedia, (Center): news, and (Right): blog perspective

$p(t\|\theta_Q)$	t	$p(t\|\theta_Q)$	t	$p(t\|\theta_Q)$	t
.3122	f1	.3078	f1	.4305	f1
.2905	formula	.2988	formula	.3133	formula
.0472	ps1	.0530	team	.0358	net
.0405	ps2	.0449	year	.0297	2005
.0378	game	.0306	aguri	.0219	2006
.0370	games	.0287	season	.0189	takuma
.0280	racing	.0263	driver	.0175	12
.0240	playstation	.0238	super	.0132	gp
.0224	pc	.0191	2006	.0126	dvd
.0204	series	.0173	championship	.0110	saf1

[13] C. Macdonald and I. Ounis. The TREC Blogs06 collection: Creating and analyzing a blog test collection. Technical Report TR-2006-224, Department of Computer Science, University of Glasgow, 2006.

[14] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 Blog Track. In *TREC 2007 Working Notes*, pages 31–43, 2007.

[15] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the TREC-2006 Blog Track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, 2007.

[16] Y. Qiu and H.-P. Frei. Concept based query expansion. In *SIGIR '93*, pages 160–169, 1993.

[17] S. Robertson and J. Callan. Routing and filtering. In *TREC Experiment and Evaluation in Information Retrieval*, pages 99–122. MIT, 2005.

[18] J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall, 1971.

[19] K. Seki, Y. Kino, and S. Sato. TREC 2007 Blog Track Experiments at Kobe University. In *TREC 2007 Working Notes*, 2007.

[20] J. Seo and W. B. Croft. UMass at TREC 2007 Blog Distillation Task. In *TREC 2007 Working Notes*, 2007.

[21] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169, New York, NY, USA, 2006. Acm.

[22] O. Vechtomova. Using subjective adjectives in opinion retrieval from blogs. In *TREC 2007 Working Notes*, 2007.

[23] R. Yan and A. Hauptmann. Query expansion using probabilistic local feedback with application to multimedia retrieval. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 361–370, New York, NY, USA, 2007. ACM.

[24] Q. Zhang, B. Wang, L. Wu, and X. Huang. FDU at TREC 2007: Opinion Retrieval of Blog Track. In *TREC 2007 Working Notes*, 2007.

[25] W. Zhang and C. Yu. UIC at TREC 2007 Blog Track. In *TREC 2007 Working Notes*, 2007.