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

Weerkamp, W.; de Rijke, M.

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

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Looking at Things Differently

Exploring Perspective Recall for Informal Text Retrieval

Wouter Weerkamp
twerpkm@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 [1, 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 additional genres: a news corpus and Wikipedia as external corpora, the former capturing the zeitgeist at the time a topic is being discussed (or subtopic) regarding the 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 additional 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 models.
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. We detail our methods and experimental setup in Section 3. We 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, 23], 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 than 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 [4] 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[6]. 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 is 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 \( p(t|\hat{Q}) \) 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:

\[
p(t|\hat{Q}) \approx p(t|Q) = \frac{p(t,q_1,\ldots,q_k)}{p(t,q_1,\ldots,q_k) + \sum_{t' \notin \hat{Q}} p(t',q_1,\ldots,q_k)} 
\]

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 \( t \):

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

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:

\[
Score(d) = \lambda \text{Rank}_{ori}(d) + (1 - \lambda) \text{Rank}_{exp}(d) 
\]

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. 3 formalizes this approach:

\[
p(t|\theta_Q) = \lambda \cdot p(t|Q) + \frac{(1 - \lambda) \cdot p(t|\hat{Q})}{\sum_{t' \in \hat{Q}} p(t'|Q)} 
\]

where \( p(t|\hat{Q}) = n(t, Q) \cdot |Q|^{-1} \) and \( n(t, Q) \) 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 [11]. 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 \( \rho \) 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

www.lemurproject.com/indri
The results of the baseline runs are listed in Table 1. Obtained by issuing the original query to the post and blog indexes.

4.1 Baselines

We start by establishing a baseline, and continue with incorporating alternative perspectives into the results.

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.

4.2 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 3.3. 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.

<table>
<thead>
<tr>
<th>task</th>
<th>corpus</th>
<th>MAP</th>
<th>p@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>post’06</td>
<td>news</td>
<td>.2168</td>
<td>.5400</td>
<td>.6555</td>
</tr>
<tr>
<td></td>
<td>wikip.</td>
<td>.2957</td>
<td><strong>.6340</strong></td>
<td><strong>.7510</strong></td>
</tr>
<tr>
<td></td>
<td>posts</td>
<td>.2986</td>
<td>.6100</td>
<td>.6729</td>
</tr>
<tr>
<td>post’07</td>
<td>news</td>
<td>.1801</td>
<td>.4190</td>
<td>.4749</td>
</tr>
<tr>
<td></td>
<td>wikip.</td>
<td>.3743</td>
<td>.6200</td>
<td>.6668</td>
</tr>
<tr>
<td></td>
<td>posts</td>
<td>.3389</td>
<td>.5940</td>
<td>.6679</td>
</tr>
<tr>
<td>blog</td>
<td>news</td>
<td>.2513</td>
<td><strong>.4867</strong></td>
<td>.6547</td>
</tr>
<tr>
<td></td>
<td>wikip.</td>
<td><strong>.3434</strong></td>
<td>.4778</td>
<td><strong>.7251</strong></td>
</tr>
<tr>
<td></td>
<td>blogs</td>
<td>.2629</td>
<td>.3311</td>
<td>.5014</td>
</tr>
</tbody>
</table>

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 3.3. 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 3: Perspective difference of the expanded runs with baseline results

<table>
<thead>
<tr>
<th>task</th>
<th>exp. MU</th>
<th>@10</th>
<th>@100</th>
<th>@1000</th>
<th>avg.ρ</th>
<th>ρ &gt; .5</th>
</tr>
</thead>
<tbody>
<tr>
<td>post’06</td>
<td>news</td>
<td>.0060</td>
<td>.0242</td>
<td>.0266</td>
<td>.5003</td>
<td>48%</td>
</tr>
<tr>
<td></td>
<td>wikip.</td>
<td>.0000</td>
<td>.0070</td>
<td>.0204</td>
<td>.6632</td>
<td>76%</td>
</tr>
<tr>
<td></td>
<td>posts</td>
<td>.0020</td>
<td>.0050</td>
<td>.0224</td>
<td>.6076</td>
<td>80%</td>
</tr>
<tr>
<td>post’07</td>
<td>news</td>
<td>.0220</td>
<td>.0190</td>
<td>.0143</td>
<td>.4989</td>
<td>35%</td>
</tr>
<tr>
<td></td>
<td>wikip.</td>
<td>.0040</td>
<td>.0028</td>
<td>.0072</td>
<td>.6425</td>
<td>67%</td>
</tr>
<tr>
<td></td>
<td>posts</td>
<td>.0020</td>
<td>.0038</td>
<td>.0080</td>
<td>.5636</td>
<td>70%</td>
</tr>
<tr>
<td>blog</td>
<td>wikip.</td>
<td>.0444</td>
<td>.0571</td>
<td>–</td>
<td>.5054</td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td>blogs</td>
<td>.0133</td>
<td>.0338</td>
<td>–</td>
<td>.5777</td>
<td>84%</td>
</tr>
</tbody>
</table>

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 (|ρ| > .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 |ρ| > .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 perspective”).

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 MoQMM 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: * or † indicate a significant
of documents lost by using this perspective.

are added to the result list by each perspective, and also the number
baseline runs we identify the number of relevant documents that
retrieved per task by the different perspectives. Compared to the
MoQM news, Wikipedia, and blog (post) perspective runs on each
5. ANALYSIS

tive.

Second, we perform a per-topic analysis to determine to
MoQM news, Wikipedia, and blog (post) perspective runs on each
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 re-
trieved. Second, we perform a per-topic analysis to determine to
which extent scores show different behavior depending on perspec-
tive.

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
documents lost by using this perspective.

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 combi-
nation runs nearly always improve over the baseline, and combina-
tions 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 re-
ported in the literature for the 2006 post finding task as well as for
the blog distillation task. Finally, the MAP-optimized combina-
tions on which we report in Table 6 do not always lead to improve-
ments on other, more precision-oriented measures over the baseline;
for instance, the Wikipedia-based MoQM expansion for 2006 per-
forms less than the baseline.

Finally, we see that mixing in the original query with expan-
sions that have higher MU values and lower average \( \rho \) and \( |\rho| > .5 \)
values tends to lead to bigger improvements in retrieval effective-
ness 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 re-
trieved. Second, we perform a per-topic analysis to determine to
which extent scores show different behavior depending on perspec-
tive.

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
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 docu-
ments a step further and compare per task all runs: three perspec-
tives 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 docu-
ments. 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 base-
line and each of the runs. Next, we sort these by decreasing dif-
fERENCE 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 perspec-
tive, 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
each 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
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 902. This topic concerns “lactose gas,” something that proves to be much more revealing in this case.

A different 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

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.

<table>
<thead>
<tr>
<th>rank</th>
<th>title</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>super aguri</td>
<td>championship</td>
</tr>
<tr>
<td>2</td>
<td>driver</td>
<td>super aguri</td>
</tr>
<tr>
<td>3</td>
<td>dvd</td>
<td>super aguri</td>
</tr>
<tr>
<td>4</td>
<td>saf1</td>
<td>super aguri</td>
</tr>
<tr>
<td>5</td>
<td>net</td>
<td>super aguri</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>task</th>
<th>comb. method</th>
<th>exp. corpus</th>
<th>λ</th>
<th>MAP</th>
<th>p@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>post’06</td>
<td>news</td>
<td>0.6</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
</tr>
<tr>
<td></td>
<td>MoQM</td>
<td>wikip.</td>
<td>0.6</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
</tr>
<tr>
<td></td>
<td>posts</td>
<td>0.5</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
</tr>
<tr>
<td>post’07</td>
<td>news</td>
<td>0.7</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
</tr>
<tr>
<td></td>
<td>MoQM</td>
<td>wikip.</td>
<td>0.7</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
</tr>
<tr>
<td></td>
<td>posts</td>
<td>0.7</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
</tr>
<tr>
<td>blog</td>
<td>news</td>
<td>0.7</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
</tr>
<tr>
<td></td>
<td>MoQM</td>
<td>wikip.</td>
<td>0.7</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
</tr>
<tr>
<td></td>
<td>posts</td>
<td>0.7</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
<td>.327</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>task</th>
<th>baseline</th>
<th>added/lost</th>
<th>by target</th>
<th>added/lost</th>
<th>by news</th>
<th>added/lost</th>
<th>by Wikip.</th>
</tr>
</thead>
<tbody>
<tr>
<td>post’06</td>
<td>11,951</td>
<td>610/534</td>
<td>685/441</td>
<td>470/346</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>post’07</td>
<td>8,229</td>
<td>109/64</td>
<td>245/151</td>
<td>158/94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blog</td>
<td>1,072</td>
<td>95/84</td>
<td>195/82</td>
<td>188/91</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>task</th>
<th>baseline</th>
<th>added/lost</th>
<th>by target</th>
<th>added/lost</th>
<th>by news</th>
<th>added/lost</th>
<th>by Wikip.</th>
</tr>
</thead>
<tbody>
<tr>
<td>post’06</td>
<td>125</td>
<td>170</td>
<td>234</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>post’07</td>
<td>23</td>
<td>22</td>
<td>125</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blog</td>
<td>18</td>
<td>23</td>
<td>62</td>
<td>57</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 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 Wikipedia perspective (+.2 MAP) and a small decrease on MAP for the news perspective (−.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*.
Table 7: Term weights for the query 888 “super bowl ads” for (Left): Wikipedia, (Center): news, and (Right): post perspective.

<table>
<thead>
<tr>
<th>Term</th>
<th>Wikipedia</th>
<th>News</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>bowl</td>
<td>.2634</td>
<td>.2710</td>
<td>.2727</td>
</tr>
<tr>
<td>super</td>
<td>.2991</td>
<td>.2343</td>
<td>.2398</td>
</tr>
<tr>
<td>capsules</td>
<td>.3500</td>
<td>.0173</td>
<td>.0118</td>
</tr>
<tr>
<td>lactose</td>
<td></td>
<td>.0126</td>
<td>.0297</td>
</tr>
<tr>
<td>product</td>
<td></td>
<td>.0167</td>
<td>.0204</td>
</tr>
<tr>
<td>safety</td>
<td></td>
<td>.0182</td>
<td>.0167</td>
</tr>
<tr>
<td>ads</td>
<td>.2000</td>
<td>.2398</td>
<td>.2219</td>
</tr>
<tr>
<td>mixture</td>
<td></td>
<td>.0219</td>
<td>.0141</td>
</tr>
<tr>
<td>rinsing</td>
<td></td>
<td>.0128</td>
<td>.0116</td>
</tr>
<tr>
<td>categories</td>
<td></td>
<td>.0115</td>
<td>.0116</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Term</th>
<th>Wikipedia</th>
<th>News</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>lactose</td>
<td>.4533</td>
<td>.4342</td>
<td>.4342</td>
</tr>
<tr>
<td>gas</td>
<td>.3916</td>
<td>.4297</td>
<td>.4297</td>
</tr>
<tr>
<td>milk</td>
<td>.0308</td>
<td>.0279</td>
<td>.0217</td>
</tr>
<tr>
<td>intolerance</td>
<td></td>
<td>.0215</td>
<td>.0121</td>
</tr>
<tr>
<td>foods</td>
<td></td>
<td>.0140</td>
<td>.0121</td>
</tr>
<tr>
<td>categories</td>
<td>.0217</td>
<td>.0152</td>
<td>.0121</td>
</tr>
<tr>
<td></td>
<td>.0216</td>
<td>.0132</td>
<td>.0121</td>
</tr>
<tr>
<td>energy</td>
<td>.0211</td>
<td>.0130</td>
<td>.0121</td>
</tr>
<tr>
<td>cubic</td>
<td></td>
<td>.0010</td>
<td>.0010</td>
</tr>
<tr>
<td>symptoms</td>
<td></td>
<td>.0065</td>
<td>.0065</td>
</tr>
<tr>
<td>digestive</td>
<td></td>
<td>.0061</td>
<td>.0061</td>
</tr>
<tr>
<td>flatulence</td>
<td>.0167</td>
<td>.0107</td>
<td>.0107</td>
</tr>
<tr>
<td>gasprom</td>
<td></td>
<td>.0056</td>
<td>.0056</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Term</th>
<th>Wikipedia</th>
<th>News</th>
<th>Blog</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>.3122</td>
<td>.2978</td>
<td>.4305</td>
</tr>
<tr>
<td>formula</td>
<td>.2905</td>
<td>.3078</td>
<td>.4305</td>
</tr>
<tr>
<td>ps</td>
<td></td>
<td>.0405</td>
<td>.0115</td>
</tr>
<tr>
<td>ps2</td>
<td></td>
<td>.0297</td>
<td>.0297</td>
</tr>
<tr>
<td>games</td>
<td></td>
<td>.0378</td>
<td>.0238</td>
</tr>
<tr>
<td>season</td>
<td></td>
<td>.0370</td>
<td>.0370</td>
</tr>
<tr>
<td>driver</td>
<td>.0280</td>
<td>.0263</td>
<td>.0219</td>
</tr>
<tr>
<td>player</td>
<td></td>
<td>.0240</td>
<td>.0240</td>
</tr>
<tr>
<td></td>
<td>.0224</td>
<td>.0191</td>
<td>.0173</td>
</tr>
<tr>
<td></td>
<td>.0204</td>
<td>.0173</td>
<td>.0173</td>
</tr>
</tbody>
</table>

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


