Diversity in Blog Feed Retrieval

Mostafa Keikha*, Fabio Crestani*, W. Bruce Croft‡
* University of Lugano, Lugano, Switzerland
‡ CIIR, University of Massachusetts Amherst, Amherst, MA
mostafa.keikha@usi.ch, fabio.crestani@usi.ch
croft@cs.umass.edu

ABSTRACT

Blog distillation (blog feed retrieval) is a task in blog retrieval where the goal is to rank blogs according to their recurrent relevance to a query topic. One of the main properties of blog feed retrieval is that the unit of retrieval is a collection of documents as opposed to a single document as in other IR tasks. This collection retrieval nature of blog distillation introduces new challenges and requires new investigations specific to this problem.

Researchers have addressed this problem by considering a wide range of evidence and information resources. However, previous work has not studied the effect of on-topic diversity of blog posts in blog relevance. By on-topic diversity of blog posts we mean that those posts that are about the query topic need to have high diversity and cover different subtopics of the query.

In this study, we investigate three types of on-topic diversity and their effect on retrieval performance: topical diversity, temporal diversity and hybrid diversity. Our experiments over different blog collections and different baseline methods show that on-topic diversity can improve the performance of the retrieval system. Among the three types of diversity, hybrid diversity, that considers both topical and temporal diversities, achieves the best performance.

Categories and Subject Descriptors

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

Keywords

Blog Retrieval, Diversity, Novelty

1. INTRODUCTION

Amongst all the information resources, the web is becoming the main source for providing new and useful information to users. Everyday, people receive up-to-date information via web-based services such as news web sites, social networks or blogs. User generated content, like blogs, plays an important role in this phenomenon. Millions of people write about their experiences and express their opinions in blogs providing a rich source of information.

Considering this huge amount of user-generated data and its specific properties, designing new retrieval methods is necessary to facilitate addressing the different types of information needs that blog users may have. Studies show that one of the main categories of blog-related queries is about finding blogs that deal with a specific topic of interest [12]. These queries, which are called concept queries, are the focus of a blog distillation system [9]. The blog distillation task (also known as blog feed retrieval)[1] is concerned with ranking blogs according to their recurring central interest to the topic of a user's query. In other words, our aim is to discover relevant blogs for each topic[2] that a user can add to his reader and refer to in the future [10].

A relevant blog is expected to regularly publish posts about the topic. When it comes to blog relevance estimation, the usual emphasis is on the number of topic-related posts that a blog publishes. Nevertheless, the amount of new information that each of these posts add to the blog is another influential factor. If a blog publishes repetitive information with low novelty, it would be less interesting for the user to follow. In our diversity-based methods, we leverage this property and penalize those blogs that have low diversity in their posts related to the topic.

There are different scenarios where considering the diversity might help the retrieval performance. Table 1 shows some real examples of blogs that were affected by one of our diversity-based methods. The examples are part of our experiments on two selected topics of TREC09 and are compared to the best performing baseline method. We can see that by penalizing those blogs that have low diversity, we can filter out non-relevant blogs such as spam blogs or blogs that publish similar posts multiple times. On the other hand, it can help us to retrieve those relevant blogs that have high diversity among their posts even though they might have low initial score for the query.

In this paper, we discuss how to add a measure of diversity to the existing blog retrieval methods and study the effect on retrieval performance. Our main aim is to answer the following questions::

[1]We use words "feed" and "blog" interchangeably
[2]We use words "topic" and "query" interchangeably

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

Table 1: Examples where measuring diversity improves blog retrieval

Topic ID: 1107
Topic title: Mountain Climbing
Topic description: Looking for blogs about equipment, clubs and good paths for climbing mountains as a beginner
Example 1
Feed no: BLOG08-feed-1167752
Blog URL: http://hockeycrew.livejournal.com
Blog title: "I Smile Because I'm Confused"
Explanation: This is a non-relevant blog that is retrieved by the best baseline method. The blog has multiple posts about Saturday hikes of the blogger. Those posts are written time to time with very general vocabulary and description that makes them very similar.
Its rank in the baseline method: 10
Its rank after applying diversity: above 100 (not retrieved)
Example 2
Feed no: BLOG08-feed-1218342
Blog URL: http://playwintersports.blogspot.com
Blog title: "INFORMATION ABOUT DIFFERENT TYPES OF WINTER SPORTS AND WHERE TO FIND EQUIPMENT"
Explanation: This is another retrieved non-relevant blog that has four posts about *"Mount McKinley"* with identical content. Those posts are probably published mistakenly and considering all of them for scoring the blog gives a high score to the blog.
Its rank in the baseline method: 61
Its rank after applying diversity: above 100 (not retrieved)
Example 3
Feed no: BLOG08-feed-070681
Blog URL: http://himadventures.blogspot.com
Blog title: "Trekking, Camping, Climbing And Traveling In Himalayas"
Explanation: This is a relevant blog that belongs to a professional climber who describes their trips to Himalayas. He describes the climbs and conditions of the mountains in very detail. The detailed description and in-depth vocabulary usage makes its posts very long with low scores in response to the query. Beside, these properties makes the posts less similar to each other while each one adds new information to the blog with respect to the query.
Its rank in the baseline method: above 100 (not retrieved)
Its rank after applying diversity: 81
Topic ID: 1122
Topic title: Skiing
Topic description: looking for blogs with information and advice on skiing, ski resorts and ski organizations.
Example 4
Feed no: BLOG08-feed-616929
Blog URL: http://justjetskis.blogspot.com
Blog title: "JUST SKI RESOURCES"
Explanation: This is a spam blog advertising "jet ski". Most of the posts in the blog are advertisement for jet ski equipment with almost identical content.
Its rank in the baseline method: 1
Its rank after applying diversity: above 100 (not retrieved)

- How important is the diversity of blog posts to blog relevance?
- What types of diversity can we define over blog posts?
- How can we capture the diversity of blog posts and integrate it into a blog feed retrieval method?
- For what type of queries can we expect diversity-based methods to be more effective?

The rest of the paper is organized as follows. In section 2 we review state of the art methods of blog retrieval. Section 3 describes the methods that rely on post level evidence aggregation for calculating blog relevance. Those methods are the baselines for our experiments and our diversity methods are developed based on them. Section 4 explains our proposed diversity-based methods. Experimental results over different data sets are discussed in section 5. Finally, we conclude the paper and describe future work in section 6.

2. RELATED WORK

Research in blog distillation started mostly after 2007, when the TREC organizers proposed the task as part of the blog track. Researchers have employed different approaches from related areas such as ad-hoc search, expert search, and resource selection in distributed information retrieval.

The simplest models use ad-hoc search methods for finding blog relevant to a specific topic. They treat each blog as one long document created by concatenating all of its posts [3, 4, 13]. These methods ignore any specific property of blogs and usually use standard IR techniques to rank blogs. Despite their simplicity, these methods perform fairly well in blog retrieval.

Some other approaches have been applied to blog retrieval based on expert search methods. Expert search is a task in the TREC Enterprise Track where systems are asked to rank candidate experts with respect to their predicted expertise about a query, using documentary evidence of the expertise found in the collection [14]. Based on the similarity between

blog distillation and expert search, some researchers have adapted expert retrieval methods for blog retrieval [1, 9]. In these models, each post in a blog is seen as evidence of blog interest in the query topic. Balog *et al.* adapt two language modeling approaches of expert finding and show their effectiveness in blog distillation [1]. MacDonald *et al.* use data fusion models to combine post-based evidence to compute a final relevance score of the blog [9].

Other researchers have employed resource selection methods from distributed information retrieval for blog retrieval. In distributed information retrieval, the cost of searching all servers for each query is considered prohibitively expensive, so server selection algorithms are used [5]. Queries are then routed only to servers that are likely to have many relevant documents for the query. Elsas *et al.* deal with blog distillation as a resource selection problem [4]. They model each blog as a collection of posts and use a language modeling approach to select the best collection. Similar work is described by Seo *et al.*, which they call Pseudo Cluster Selection [13].

None of the previous work paid attention to the diversity of blog posts in their retrieval methods. Previously, the cohesiveness of a blog was assumed to be a positive sign of blog relevance [4, 13, 9, 6]. However, diversity, as a negatively correlated measure to cohesiveness, has never been studied. In the following, we consider the diversity of blog posts, its effect on the retrieval systems, and its relation with the cohesiveness of the blog.

3. AGGREGATION-BASED BLOG RETRIEVAL METHODS

In general, most of the blog distillation methods can be seen as an aggregation of the post-level relevance evidence to calculate the blog-level relevance score. The following are the main existing aggregation methods that we use as baselines in our experiments:

- *CombSum*: Summation of the post-level relevance scores is one of the simplest aggregation methods that can be applied in blog retrieval [9]:

$$score_{\texttt{CombSum}}(b_i, q) = \sum_{j=1}^{|b_i \cap R(q)|} score(p_{ji}, q)$$

  where b_i is a blog feed and q is the given query. Here $R(q)$ denotes the set of initially retrieved posts for query q and the intersection $b_i \cap R(q)$ denotes only those retrieved posts that belong to blog b_i. p_{ji} shows the j-th retrieved post from blog i and $score(p_{ji}, q)$ denotes the relevance score assigned to the post p_{ji} with respect to the query. MacDonald and Ounis use the exponential of the scores in the aggregation [9]. However our experiments show that with the language model scores the exponential function does not have any effect on the retrieval performance and thus we use the original scores.

- *PCS*: In the Pseudo Cluster Selection (PCS) method, Seo and Croft use the geometric mean of the top k retrieved posts of a blog as its score [13]:

$$score_{\texttt{PCS}}(b_i, q) = (\prod_{j=1}^{k} score(p_{ji}, q))^{\frac{1}{k}}$$

  here p_{ji} is the j-th top retrieved post from blog b_i. In this method, if the blog has more than k posts retrieved for the query, the rest of the posts are ignored. However, if the blog has m posts where $m < k$, we need to find a way to give a non-zero score to the $(k - m)$ missing operands. In the original PCS method, they use the minimum score among all the retrieved posts as the score of the missing posts. We use a slightly different choice and smooth the score of posts with the collection likelihood for the query:

$$score(p, q) = (1 - \gamma) score_{ini}(p, q) + \gamma score(C, q) \quad (1)$$

  where C is the collection. Since in all our experiments we use the language model query likelihood as the score, we can rewrite the smoothing as follows:

$$\begin{aligned} score(p, q) &= P(q|p) = (1 - \gamma) P_{init}(q|p) + \gamma P(q|C) \\ &= (1 - \gamma) P_{ini}(q|p) + \gamma \prod_{w \in q} P(w|C)^{P(w|q)} \end{aligned}$$

  where $P(w|C)$ is the probability of a query term in the collection, $P(w|q)$ is the probability of the term in the query, and $P_{ini}(q|p)$ is the query likelihood of the post. Using the smoothed scores, we do not need to use the minimum score of posts for blogs with less than k posts. We simply assume that the missing posts have initial score of zero and use their smoothed score. This can be seen as an equivalent of the Jelinek-Mercer smoothing of language model, however we mainly use it to give score to posts that have not been observed. Employing the smoothed scores, PCS performs better compared to the original method. Also this variation simplifies the calculation in our diversity-based method which we describe later.

- *SDM*: Small Document Model (SDM) is one of the best-performing and best-justified methods for blog feed retrieval [4]. Following the Language Modeling approach to IR, blogs are ranked according to their likelihood given the query:

$$P(b_i|q) = \frac{P(b_i)P(q|b_i)}{P(q)} \stackrel{rank}{=} \underbrace{P(b_i)}_{\substack{\text{Blog} \\ \text{Prior}}} \underbrace{P(q|b_i)}_{\substack{\text{Query} \\ \text{Likelihood}}}$$

  Blog prior, $P(b_i)$ can be set to a logarithmic function of the number of posts in the blog, so as to favor longer blogs, since they are more likely to contain useful information [4]. The query likelihood for a blog is calculated by summing the query likelihoods for each post in the blog scaled according to the probability (centrality) of the post within the blog:

$$P(q|b_i) = \sum_{j=1}^{|b_i \cap R(q)|} P(q|p_{ji}) P(p_{ji}|b_i) \quad (2)$$

  Here p_{ji} is a post in the blog b_i, and $P(q|p_{ji})$ is the query likelihood for each post which is computed over

query terms using Dirichlet smoothing [16]. Our experiments show that uniform estimations of $P(p_{ji}|b_i)$ over posts of each blog performs better and is easier to calculate than the original centrality score of the posts proposed by Elsas *et al.* [4]. Thus the final relevance score of a blog is calculated as follows:

$$score_{\texttt{SDM}}(q, b_i) = \frac{log(N_{b_i})}{N_{b_i}} \sum_{j=1}^{|b_i \cap R(q)|} P(q|p_{ji}) \quad (3)$$

where N_b is the number of posts in the blog b. We assume that the query likelihood of those posts that are not retrieved in $R(q)$ is equal to zero and thus we do not consider them in the summation. Similar estimations are used by Balog *et al.* [1].

All the mentioned methods have one parameter as the size of $R(q)$ which we call n. This parameter defines the size of the initially retrieved set of posts for the query that will be inputs of the aggregation methods. The value of n is learnt using a training set which will be described in more detail later. The PCS method has two more parameters that are assigned as follows:

- k: number of the top posts from each blog used in the calculations. Without any further tuning, we set this parameter to 5 in our experiments which is proved to be an optimal choice in the original work [13].
- Smoothing parameter γ: we set this parameter in all the experiments to 0.01 which was found to be an optimal choice for both baseline and respective diversity-based methods.

4. DIVERSITY IN BLOG RETRIEVAL

The relation between posts in each blog can give us useful information about the blog. Previous studies considered cohesiveness as a way to capture how posts in the blog are similar to the blog in general [4, 13, 9]. In these studies, a blog with higher cohesiveness and thus less diversity is considered a better blog to retrieve. Elsas *et al.* define the centrality of a post as its similarity to the blog as a whole. They assume that if the retrieved posts from a blog have high centrality, that blog is a better candidate for retrieval [4]. Seo and Croft penalize those blogs that have high diversity among their posts [13]. Defining the goal of blog retrieval system to retrieve topic-centric blogs, they assume that blogs with high diversity are not topic-centric and thus should be penalized. MacDonald and Ounis define similar measures to retrieve blogs with focused interests [9]. They also consider temporal distribution of posts to retrieve blogs with recurring interests.

In contrast to previous work where all the posts of the blog are used for estimations, we focus only on those posts that are about the query topic. While diversity of a blog as a whole might be negative evidence for blog relevance, we assume that on-topic diversity is an asset. In other words, we assume that a relevant blog should have a high coverage over sub-topics and thus should have a high diversity among posts that it publishes about the topic.

We define three types of diversity over the posts and investigate their effectiveness on performance of a blog retrieval system:

- Topical Diversity
- Temporal Diversity
- Hybrid Diversity

In the following we discuss each of the methods in more detail.

4.1 Topical Diversity of Blog Posts

First we study topical diversity among the posts of each blog. Blog distillation queries tend to be more general than normal web queries and thus have a wider range of sub-topics [4]. We assume that posts that are retrieved from each blog for the query should have high diversity in order to cover different sub-topics. We want to investigate how this on-topic diversity of posts can affect the relevance of a blog to the query.

In order to test if the diversity of blog posts is an important factor in the relevance of blogs, we carry out preliminary experiments on the Blog08 collection. In these experiments, we assume that higher similarity of blog posts indicates less diversity among them. This assumption is consistent with previous work on diversification [2].

For each query, we first retrieve the top n posts using a traditional retrieval method. Then for each blog that has more than one post in the retrieved set, we calculate the average similarity between its retrieved posts, which we call *On-topic Intra-feed Similarity* (OIS):

$$OIS(b_i, q) = \frac{\sum_{j=1}^{|b_i \cap R(q)|} \sum_{l=j+1}^{|b_i \cap R(q)|} sim(p_{ji}, p_{li})}{\binom{|b_i \cap R(q)|}{2}} \quad (4)$$

The higher the *OIS* value, the less diverse is the feed with respect to the query. We use cosine similarity as the similarity measure between two posts.

Figure 1 shows the distribution of *OIS* for relevant and non-relevant blogs. It can be seen that non-relevant blogs are more likely to have high *OIS* values. The mean of *OIS* for relevant blogs is `0.43` compared to `0.50` for non-relevant blogs. Based on the Welch two sample t-test, the difference between the mean values is statistically significant with a p-values equal to `2.6e-16`. This shows the possibility of using the diversity of posts for discriminating between relevant and non-relevant blogs and consequently having a better retrieval system.

In figure 1, we use the top 2000 posts for each topic. Similar behavior has been observed for a smaller number of posts. However, it is possible that with an increasing number of retrieved posts, we retrieve more posts for each blog and we end up with less diverse relevant blogs or more diverse non-relevant blogs. In order to test this possibility, we run the same analysis with the top 15000 posts for each query. Figure 2 shows the outcome of this experiment. The mean of *OIS* values for relevant blogs is `0.34` compared to `0.42` for non-relevant blogs. The difference between the mean values is statistically significant with the p-value less than `2.2e-16`.

As we can see, the *OIS* values decrease with increasing numbers of posts. However, the difference between relevant and non-relevant blogs still exists. This is quite surprising, since we expect that by retrieving more posts, we increase the chance that non-relevant blogs will have a more diverse set of retrieved posts.

The per-topic analysis in figure 3 shows the difference more clearly. This figure shows the average of *OIS* values

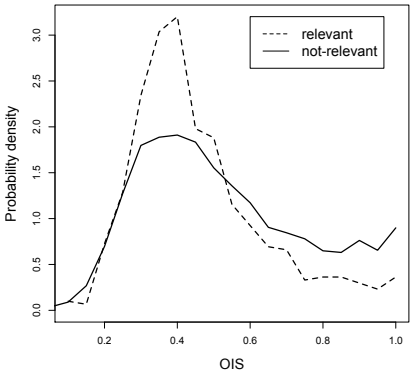

Figure 1: Distribution of On-topic Intra-feed Similarity(*OIS*) using top 2000 retrieved posts.

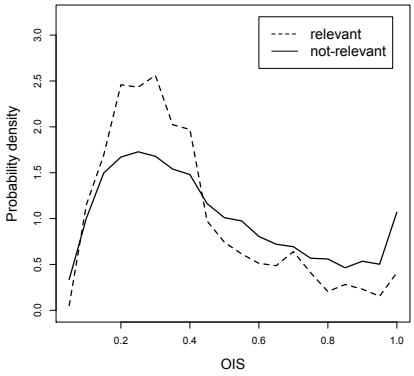

Figure 2: Distribution of On-topic Intra-feed Similarity(*OIS*) using top 15000 retrieved posts..

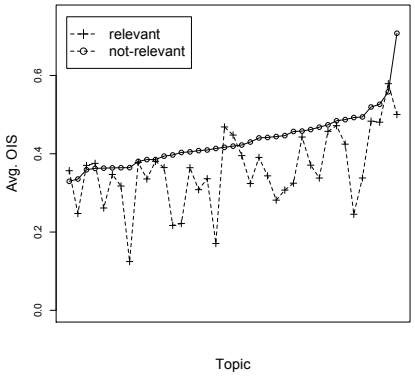

Figure 3: Average *OIS* values for each topic using top 15000 retrieved posts.

for relevant and non-relevant blogs for each query topic. For clarity, topics are sorted by the average *OIS* of their non-relevant feeds. As can be seen for most topics, non-relevant feeds have higher similarity among their posts compared to the relevant ones.

To capture the diversity of blog posts, we adapt a variation of Maximal Marginal Relevance (MMR) [2]. The goal of MMR is to maximize diversity of retrieved set of documents. It selects documents that are more similar to the query and less similar to already retrieved documents:

$$MMR \stackrel{def}{=}$$

$$Arg\ max_{d_i \in R \setminus S} \left[\lambda sim(d_i, q) - (1-\lambda)\ max_{d_j \in S}\ sim(d_i, d_j)\right]$$

where R is an initial set of documents and S is the set of already retrieved documents.

Similar to the MMR method, our diversity detection method exploits the fact that similar documents are less diverse, thus it penalizes posts that are similar to other posts in the blog. In other words, only information of a post that does not appear in other blog posts can contribute to blog relevance:

$$\begin{aligned} &score_{div-topical}(p_i, q) = \\ &score_{ini}(p_i, q)(1 - \lambda\ max_{p_j \in S}\ sim_{topical}(p_i, p_j)) \end{aligned} \tag{5}$$

where $score_{ini}(p_i, q)$ is the initial score of post p_i for the query q. S is the set of posts that belong to the same blog and have higher scores than p_i. This method assumes that $sim(p_i, p_j)$ is in $[0, 1]$ and thus it decreases the post score based on its similarity to other posts. The parameter λ controls the importance of the post novelty in its score, which can vary for different similarity methods or different queries. $sim_{topical}$ captures the content similarity between the two posts and can be replaced by any similarity measure. We use *cosine* similarity, since it always has a value in $[0, 1]$ and does not need an extra normalization step:

$$sim_{topical}(p_i, p_j) = \frac{\sum_w tf(w, p_i) \times tf(w, p_j)}{\sqrt{\sum_w tf(w, p_i)^2 \times \sum_w tf(w, p_j)^2}} \tag{6}$$

where $tf(w, p)$ is the term frequency of the term w in the post p. When there is no similarity between a post and other blog posts that have higher score, the maximum of similarities will be zero. In this case, the diversity score of the post will be the same as its initial score and therefore all the relevance evidence of the post can contribute to the blog relevance score. In contrast, if there is another post very similar to the current post, then the cosine similarity will be close to one. As a result, depending on the value of λ, the diversity score of the post can be close to zero. Therefore, the blog will not gain any relevance from that post even if the post has high query likelihood.

After obtaining new scores for posts, any existing aggregation method can be used to aggregate the new scores and calculate the score of blogs. Applying the diversity-based scores introduces a new parameter λ that needs to be estimated.

4.2 Temporal Diversity of Blog Posts

In addition to the content of blog posts, temporal information is another important source of information that can be used by the retrieval system [8, 9]. Analogous to topical diversity, we can assume that a relevant blog should have a high temporal diversity among its published posts for the query. In other words, we expect that blog posts have high coverage over the temporal space and not be concentrated on specific time windows.

We employ a similar approach to the topical diversity where we penalize the posts that have high temporal similarity to other posts. To this end, we define a temporal similarity function between two posts as an un-normalized Gaussian function:

$$sim_{temporal}(p_i, p_j) = e^{\frac{(t_i - t_j)^2}{2\sigma^2}} \tag{7}$$

where t_i is the time-stamp of p_i given in days. We can see that the temporal similarity has a value between zero and one. If the two posts are published in the same day their similarity will be one and the similarity will decrease when the temporal distance increases. Finally, we penalize post scores based on their temporal similarity:

$$\begin{aligned} &score_{div-temporal}(p_i, q) = \\ &score_{ini}(p_i, q)(1 - \lambda\ max_{p_j \in S}\ sim_{temporal}(p_i, p_j)) \end{aligned} \tag{8}$$

Using this method, if the post is published around the

same date as some other posts, it will contribute less to the blog relevance than a post which is published at a distant time. In other words, this captures the property of whether the blog posts are published all over the temporal space or if they are mostly published in some specific small time windows.

The temporal similarity function adds a new parameter σ to the model which is the standard deviation of the Gaussian function.

4.3 Hybrid Diversity

Finally, we can consider hybrid measure of diversity that takes both topical and temporal diversities into account. Two possible cases in which temporal and topical diversities can complement each other can be described as follows:

- Retrieve a relevant blog that writes about similar subtopics but in different time windows. This might be a subtopic that is discussed periodically over time. For example any seasonal query that repeats over time can have such a property. In this case the topical diversity would be low and the temporal diversity would be high. Thus a combination of the two diversities can better handle the situation than considering only topical diversity which would over-penalize the blog.

- Retrieve a relevant blog that writes about different subtopics in the same time window. A blog can heavily discuss the topic in some time periods and if the published posts are about different sub topics, they might be valuable for the blog relevance. In this case, we would have low temporal diversity and high topical diversity and may need a combination of the two measures for better representation.

To consider these cases in our method, we define an hybrid similarity function as follows:

$$sim_{hybrid}(p_i, p_j) = sim_{topical}(p_i, p_j) \ . \ sim_{temporal}(p_i, p_j)$$

where $sim_{topical}$ and $sim_{temporal}$ are calculated using formula 6 and formula 7 respectively. Similar to topical and temporal diversity scores in formulas 5 and 8, we penalize those posts that have high hybrid similarity with other posts.

We can see that in this method, we mainly penalize those posts that are published around the same date as other posts and also have very similar content to them. Thus, in any of the two mentioned scenarios where one of the similarities is high and the other one is low, the post will not be highly penalized and can still contribute to the blog relevance.

5. EXPERIMENTAL RESULTS

We conduct our experiments over four years worth of TREC blog track data from the blog distillation task, including TREC'07, TREC'08, TREC'09 and TREC'10 data sets. The TREC'07 and TREC'08 data sets include 45 and 50 queries respectively and use the Blog06 collection. The TREC'09 and TREC'10 data sets use Blog08, a new collection of blogs, and have 39 and 46 queries respectively. We use only the title of the topics as the queries.

The Blogs06 collection is a crawl of about one hundred thousand blogs over an 11-week period [10], and includes blog posts (permalinks), feed, and homepage for each blog. Blog08 is a collection of about one million blogs crawled over a year with the same structure as the Blog06 collection [11]. In our experiments we only use the permalinks component of the collection, which consist of approximately 3.2 million documents for Blog06 and about 28.4 million documents for Blog08.

We use the Terrier Information Retrieval system[3] to index the collection with the default stemming and stop-words removal. In all the methods, the language modeling approach using the Dirichlet smoothing has been used to score the posts and retrieve top posts for each query. Without further tuning, the Dirichlet smoothing parameter is set to 5000 [16].

We use the three blog retrieval methods discussed in section 2 as our baseline methods. We apply each of these methods on the language model scores and also on the three diversified scores calculated by our proposed methods.

Since TREC'07 and TREC'08 query sets share the same collection, we use one to tune the parameters for the other. We do the same for the TREC'09 and TREC'10 query sets that share the Blog08 collection. By fixing the parameters k and σ in the PCS method and the Dirichlet smoothing parameter across all the methods, the remaining parameters to be tuned are the following:

- The number of the top retrieved posts, n, that are initially retrieved for each query. The examined values for this parameter are 500, 1000, 2000, 4000, 8000 and 15000.

- The diversity parameter, λ, in each of the diversity methods. We tested different values for this parameter as follows: 0.01, 0.05, 0.1, 0.2, ..., 0.9, 0.95 and 0.99.

- The standard deviation of the gaussian function, σ, in the temporal and hybrid diversity methods. For the Blog06 collection we tried the values 3, 5, 7, 15, 30, 45, 60, 90. Since the Blog08 collection has a wider time span, we also tried the values 120, 250 and 370 for the TREC'09 and TREC'10 experiments.

Tables 2, 3, 4 and 5 show the performance evaluation of the methods over TREC'07, TREC'08, TREC'09 and TREC'10 respectively. The first three rows in each table represent the performance of baseline methods. The second three rows show the performance of the corresponding methods based on the topical diversity scores. The third three rows show the performance of the methods using the temporal diversity scores followed by the last rows that show the performance of the hybrid diversity scores.

Statistical significance tests are performed using the paired T-test at 0.05 level of significance. The symbol $\uparrow$ indicates that a diversity method has a statistically significant improvement over its corresponding baseline. The symbols $\triangle$ and $\blacktriangle$ show that the hybrid diversity method has a statistically significant improvement over the temporal diversity and topical diversity methods respectively. The bold values in each column indicate the best performance for the corresponding evaluation measure.

As we can see in the tables, diversity-based methods generally improve their corresponding baseline methods. In most cases, these improvements are at a statistically significant level. The temporal diversity methods are not as effective as their topical diversity counterpart. However, when combined together, the resulting hybrid diversity methods

[3]http://ir.dcs.gla.ac.uk/terrier/

Table 2: Evaluation results for the implemented models over TREC'07 data set.

Model	MAP	P@10	Bpref
CombSum	0.2259	0.3844	0.2721
PCS	0.2695	0.4378	0.3115
SDM	0.2867	0.4444	0.3439
CombSum-topical	0.2506 ↑	0.4244 ↑	0.2859 ↑
PCS-topical	0.2901 ↑	0.4422	0.3184 ↑
SDM-topical	0.3168 ↑	0.4756 ↑	**0.3667** ↑
CombSum-temporal	0.2612 ↑	0.4267	0.2840 ↑
PCS-temporal	0.2844 ↑	0.4911 ↑	0.3134
SDM-temporal	0.3023 ↑	0.4756 ↑	0.3539
CombSum-hybrid	0.2466 ↑	0.4333 ↑	0.2866 ↑△
PCS-hybrid	0.2961 ↑△	0.4622	0.3197 ↑△
SDM-hybrid	**0.3191** ↑△	**0.5156** ↑△ ▲	0.3622 ↑△

Table 3: Evaluation results for the implemented models over TREC'08 data set.

Model	MAP	P@10	Bpref
CombSum	0.1719	0.3000	0.2380
PCS	0.2016	0.3400	0.2709
SDM	0.2096	0.3740	0.2699
CombSum-topical	0.1847	0.3340	0.2446
PCS-topical	0.2086	0.3460	0.2640
SDM-topical	0.2284 ↑	0.3820	0.2729
CombSum-temporal	0.1852 ↑	0.3380 ↑	0.2430
PCS-temporal	0.2073	0.3660 ↑	0.2540
SDM-temporal	0.2289 ↑	0.3800	0.2769
CombSum-hybrid	0.1918 ↑ ▲	0.3460 ↑	0.2422
PCS-hybrid	0.2179 ↑△	0.3500	0.2661
SDM-hybrid	**0.2329** ↑	**0.3840**	**0.2777**

Table 4: Evaluation results for the implemented models over TREC'09 data set.

Model	MAP	P@10	Bpref
CombSum	0.1889	0.3154	0.2148
PCS	0.2093	0.3103	0.2279
SDM	0.2636	0.3821	0.2858
CombSum-topical	0.2180 ↑	0.3282	0.2297 ↑
PCS-topical	0.2196 ↑	0.3282	0.2374 ↑
SDM-topical	0.2888 ↑	**0.4333** ↑	0.3091 ↑
CombSum-temporal	0.2087 ↑	0.3051	0.2187
PCS-temporal	0.2082	0.3103	0.2250
SDM-temporal	0.2759	0.3846	0.2909
CombSum-hybrid	0.2176 ↑△	0.3282 △	0.2334 ↑△
PCS-hybrid	0.2224 ↑△	0.3462 ↑△	0.2387 ↑△
SDM-hybrid	**0.2961** ↑△ ▲	0.4308 ↑△	**0.3125** ↑△

Table 5: Evaluation results for the implemented models over TREC'10 data set.

Model	MAP	P@10	Bpref
CombSum	0.1631	0.2696	0.1631
PCS	0.1758	0.2870	0.1771
SDM	0.2273	0.3022	0.2273
CombSum-topical	0.1856 ↑	0.2435	0.1856 ↑
PCS-topical	0.1843	0.2913	0.1843
SDM-topical	0.2420 ↑	0.3435 ↑	0.2559 ↑
CombSum-temporal	0.1623	0.2196	0.1623
PCS-temporal	0.1759	0.2500	0.1759
SDM-temporal	0.2151	0.2891	0.2231
CombSum-hybrid	0.1738 △	0.2239	0.1831 ↑△
PCS-hybrid	0.1842	0.3000 ↑	0.1842
SDM-hybrid	**0.2571** ↑△ ▲	**0.3500** ↑△	**0.2649** ↑△

usually produce the best results. In some cases, combining temporal diversity with topical diversity results in a statistically significant improvement over each one of them.

An interesting observation is that the *SDM* method, as the strongest baseline, benefits the most from the diversification. As a result, the *SDM-hybrid* method significantly outperforms the *SDM* method in most of the collections and evaluation metrics.

We previously mentioned that spam blogs are one category of non-relevant blogs that would be filtered using diversity-based methods. It is interesting to see what portion of blogs that are affected by diversity methods are in fact spam blogs. To this end, we manually checked non-relevant blogs that were initially retrieved by SDM and were removed from the ranked list after applying hybrid diversity measure. We chose ten random queries from the TREC'09 query set and compared their ranked lists before and after considering diversity. Among 106 non-relevant blogs that were removed from the ranked list, 27 of them (25%) were spam blogs and the rest were non-spam. This shows that considering diversity does not just filter out spam blogs, but also removes other non-relevant blogs from the ranked list. On the other hand, our examination shows that the method promotes relevant blogs in the ranking. For the examined queries, there were 30 new relevant blogs retrieved after applying diversity and only 6 relevant blogs were removed from the ranked list.

In order to test the robustness of proposed approaches, we analyze the sensitivity of their retrieval performance to the parameters. We analyze the results over the TREC'09 query set, as an example of a large collection, and the TREC'07 query set, as an example of a small collection. For simplicity, we only consider the *SDM* method and the corresponding diversity-based methods in this analysis. To study each of the parameters, we fix the other parameters with values that are learnt from the training set in the previous experiments.

Figures 4 and 7 show the effect of n, the number of initially retrieved posts, on the performance of retrieval systems. The diversity-based methods outperform the baselines at all the values of n. This confirms our analysis in section 4 that the difference between on-topic diversity of relevant blogs and non-relevant blogs does not change by increasing n. We can see that by increasing the number of posts, the performance increases and the difference of performances for values higher than 5000 are insignificant.

Figures 5 and 8 show the effect of λ when we fix the values of n and σ. We can see that the performance of the hybrid and topical diversity methods generally increase by increasing λ. In both collections, the best performance of these methods is achieved when λ has a value close to one. This shows that the maximum penalty for those blogs that publish repetitive information produces the best results. On the other hand, temporal diversity is less robust and the performance decreases for λ values higher than 0.5.

Finally, figures 6 and 9 show the effect of σ on the perfor-

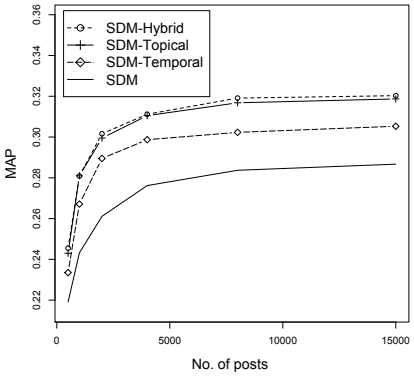

Figure 4: Effect of the number of the top retrieved posts in the performance over TREC'07 data set

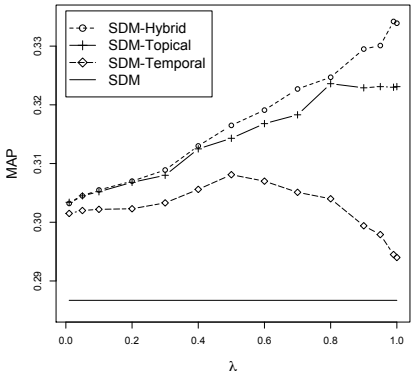

Figure 5: Effect of the diversity parameter λ in the performance over TREC'07 data set

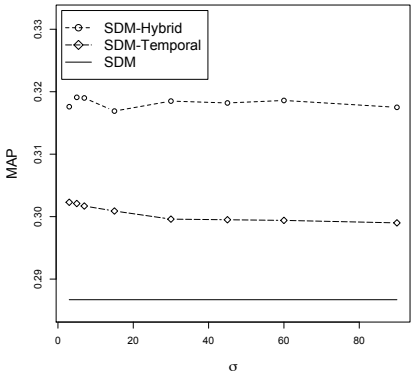

Figure 6: Effect of the parameter σ in the performance over TREC'07 data set

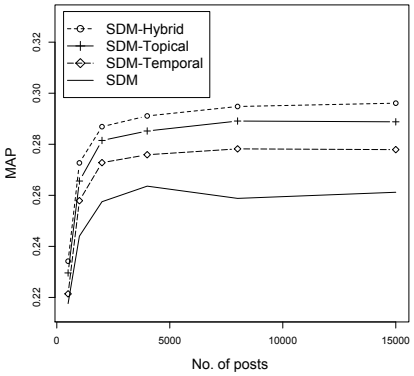

Figure 7: Effect of the number of the top retrieved posts in the performance over TREC'09 data set

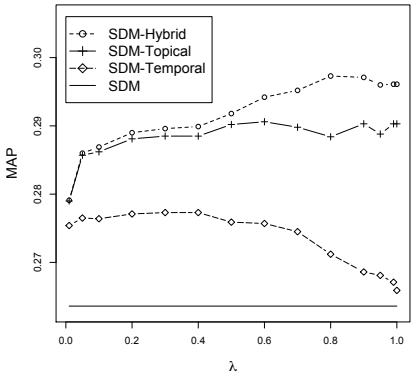

Figure 8: Effect of the diversity parameter λ in the performance over TREC'09 data set

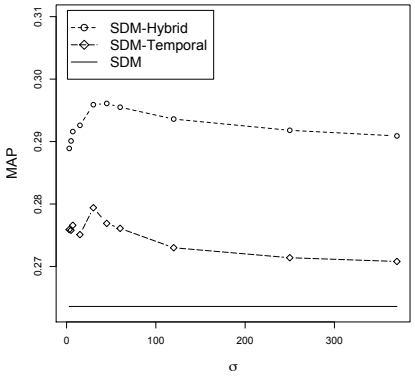

Figure 9: Effect of the parameter σ in the performance over TREC'09 data set

mance of temporal and hybrid diversity methods. It can be seen that σ is less influential on the performance and it has little negative effect when its value increases. It is interesting to see the relation between the size of the collection and the best value of σ. While for the TREC'07 collection the best performance is achieved when σ is around 5, this value for the TREC'09 collection is around 40.

5.1 Diversity vs. Cohesiveness of a blog

So far, we showed that blog post diversity is a discriminative feature in blog retrieval that can improve retrieval performance. In this section, we investigate a comparison between blog post diversity and the previously proposed blog cohesiveness measures. The cohesiveness of a blog is assumed to show whether the blog focuses on a specific topic or not. While it is a reasonable assumption, the proposed methods in previous work did not show any positive effect on retrieval effectiveness:

- MacDonald and Ounis suggest that high cohesiveness of a blog shows that most of the blog posts are about similar topics and thus we should give higher score to that blog. Adding their cohesiveness measure to the weighting model not only did not improve their results but also decreased the results in some cases [9].

- Seo and Croft define a penalty factor based on a clarity score [13]. They assume that a blog that covers many different topics has a language model similar to that of the collection and thus should be penalized. They add the clarity-based penalty to two weighting models and in both cases it decreased the performance of the system.

- Elsas *et al.* propose a centrality measure for each post to calculate $P(p_{ji}|b_i)$ in equation 2 [4]. The centrality captures how similar the post is to the blog as a whole. The assumption is that if the blog is cohesive, then the centrality of the posts is high and thus the overall score of the blog should be higher. Our experiments show that replacing the centrality measure with a simple uniform distribution over blog posts improves the retrieval performance [4].

[4]For clarity of the paper we do not report this comparison here.

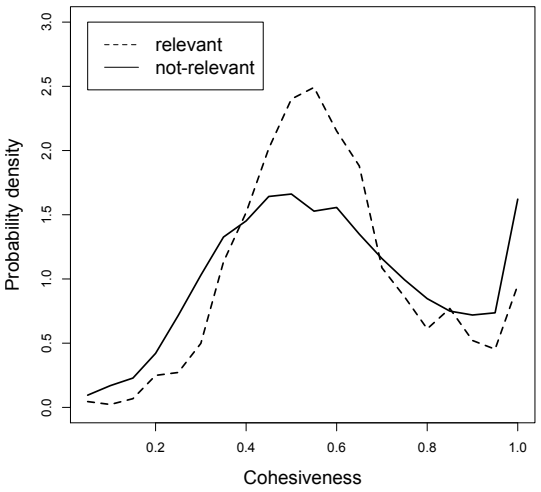

Figure 10: Cohesiveness distribution of relevant and non-relevant blogs

All these works assumed cohesiveness to be a positive feature of blog relevance and did not investigate if this is a valid assumption or not. In order to verify the validity of this assumption, we use the following cohesiveness measure to compare the relevant and non-relevant blogs:

$$cohesiveness(b_i, q) = \frac{\sum_{j=1}^{|b_i \cap R(q)|} sim(p_{ji}, VD(b_i))}{|b_i \cap R(q)|}$$

where $VD(b_i)$ is a virtual document created by concatenating all the posts of b_i. We use the cosine similarity measure for calculating $sim(p_{ji}, b_i)$, as the similarity between a post and a blog. This measure of cohesiveness compares the initially retrieved posts of a blog to the blog as a whole. In other words, it captures the cohesiveness of the blog with respect to the topic. It is similar to those measures used in other works mentioned earlier [9, 4].

Figure 10 shows the distribution of cohesiveness values for relevant and non-relevant blogs in TREC'09 data set. The value of n, the size of $R(q)$, is set to 15000 for this analysis. The mean value of cohesiveness for relevant blogs is `0.56` and for non-relevant blogs, it is `0.55`. The difference between the mean values is not statistically significant and the p-value of the Welch t-test is `0.66`. This shows that the cohesiveness is not a strong discriminative feature and explains why adding cohesiveness did not improve the performance of the retrieval systems in the previous works.

It is interesting to verify if cohesiveness is correlated with *IOS*, as defined in equation 4. If a blog has high cohesiveness, it is not surprising that it would also have high similarity among its posts and thus high *IOS* value. However if the top posts of a blog are very similar (high *IOS*), one can not directly conclude that the blog has high cohesiveness. In order to examine such a correlation we calculate the Pearson correlation between the cohesiveness and the *OIS* values. The correlation is statistically significant with a value of `0.83`. The high correlation shows that blogs with high *OIS* are very likely to also have high cohesiveness.

5.2 Diversity in Facet Detection

In 2009, a more complex and refined version of the blog distillation task was introduced in TREC, named "Faceted Blog Distillation" [11]. The new task, which was first introduced by Hearst *et al.* [7], aims to consider not only the blog relevance to the topic but also the "quality aspects" (facets) of the blog. For each query, a specific facet is determined and only blogs that satisfy that facet are considered to be relevant. In this section, we discuss the effect of diversity on the performance of the system for different facets. Different facets can be seen as different categories of blogs that users might search for. It is important to be able to decide for what type of information needs we can use diversity to get the maximum overall performance.

We use the introduced facets in the TREC'09 and TREC'10 data collections which include Opinionated vs. Factual, Personal vs. Company and In-depth vs. Shallow facets [11]. Tables 6 and 7 compare the performance of the diversity method with the baseline method for each facet. Without further tuning, we use the same parameter values as the previous section.

The statistically significant improvements are shown by ↑. Since each facet has very few queries, the statistical significance tests will be very sensitive and do not always show a difference. Thus we report also the percentage of improvement over the baseline method. The maximum improvement for each measure is shown in bold.

As we can see, diversity improves the performance of the system for almost all of the facets and evaluation measures. This shows that the effectiveness of diversity is not restricted to specific categories of topics. As a result, one can expect to have improvements by applying diversity for any type of information needs in blog search.

It is interesting to notice the improvement of P@10 for the in-depth queries which is the maximum among all the facets. For the in-depth queries, a relevant blog is expected to have in-depth thoughts and analyses about the topic [11]. As we previously saw in example 3 in table 1, one of the scenarios where diversity is effective is in retrieving blogs with such an in-depth property. The obtained improvements for in-depth queries confirm our previous observation. The results shows that a diversity-based method retrieves in-depth blogs at the top ranks and significantly improves the P@10 measure.

6. CONCLUSION AND FUTURE WORK

In this paper, we studied the effect of employing different diversity measures on blog retrieval. We showed that diversity is an important feature that can help in distinguishing relevant blogs from non-relevant ones. We introduced three types of diversity and investigated their effect on the performance of the retrieval systems. Our experiments on standard blog retrieval collections showed improvements over different baseline methods.

We further showed that the common assumption of cohesiveness is not an indicative feature of blog relevance and that is in fact the reason that there is no improvement observed in the previous work.

Finally, we investigated the effect of diversity on different types of queries and showed that diversity can be beneficial to all query types.

Table 6: Effect of diversity on different facets over TREC'09 data set.

Facet	Model	MAP	P@10	Bpref
Opinionated (13 queries)	SDM	0.1153	0.1615	0.0998
	SDM-hybrid	0.1236 (+7%)	0.1923 (+19%)	0.1132 (+13%)
Factual (13 queries)	SDM	0.1749	0.1692	0.1570
	SDM-hybrid	0.1788 (+2%)	0.1538 (-9%)	0.1661 (+5%)
Personal (8 queries)	SDM	0.1840	0.2000	0.1384
	SDM-hybrid	0.2201 (+19%) ↑	0.2375 (+18%)	0.1756 (**+26%**) ↑
Official (8 queries)	SDM	0.1876	0.1750	0.1537
	SDM-hybrid	0.2319 (**+23%**)	0.1750 (0.0%)	0.1819 (+18%)
Indepth (18 queries)	SDM	0.2723	0.2222	0.2624
	SDM-hybrid	0.2872 (+5%)	0.2667 (**20%**) ↑	0.2526 (-3%)
Shallow (18 queries)	SDM	0.1348	0.0889	0.1118
	SDM-hybrid	0.1540 (+14%) ↑	0.1056 (+18%)	0.1378 (+23%)

Table 7: Effect of diversity on different facets over TREC'09 data set.

Facet	Model	MAP	P@10	Bpref
Opinionated (15 queries)	SDM	0.1061	0.1800	0.1257
	SDM-hybrid	0.1260 (+18%) ↑	0.2000 (+11%)	0.1369 (+8%)
Factual (15 queries)	SDM	0.1028	0.1000	0.0820
	SDM-hybrid	0.1190 (+15%)	0.1067 (+6%)	0.0985 (**+20%**)
Personal (15 queries)	SDM	0.1290	0.1533	0.1095
	SDM-hybrid	0.1426 (+10%)	0.1467 (-4%)	0.1060 (-3%)
Official (15 queries)	SDM	0.2057	0.1267	0.1697
	SDM-hybrid	0.2274 (+10%)	0.1400 (+10%)	0.1928 (+13%)
Indepth (16 queries)	SDM	0.2396	0.1250	0.2079
	SDM-hybrid	0.2953 (**+23%**) ↑	0.1625 (**+30%**) ↑	0.2430 (+16%)
Shallow (16 queries)	SDM	0.0833	0.1250	0.0818
	SDM-hybrid	0.1003 (+20%)	0.1313 (+5%)	0.0838 (+2%)

One possible extension for future work is to apply the proposed approach to similar problems. It would be interesting to see if the diversity assumption holds for collection selection in distributed IR, expert search or user search in microblogs.

We are also interested in a more general framework for considering blog diversity that can be used in any existing blog retrieval method. So far, we have only applied diversity measures to the methods that are an aggregation of post-level scores. However, there are some methods that do not just aggregate post-level scores, such as the Blogger Model method [1] or two stage retrieval model [15]. It would be interesting to develop a framework to adapt diversity measures for those methods as well.

7. ACKNOWLEDGMENTS

This work was supported partly by the Center for Intelligent Information Retrieval and partly by the Swiss National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

We thank Jangwon Seo for his collaboration and comments while he was in CIIR. We also like to thank Armita Kaboli for her helpful discussions and comments.

8. REFERENCES

[1] K. Balog, M. de Rijke, and W. Weerkamp. Bloggers as experts: feed distillation using expert retrieval models. In *Proceedings of SIGIR'08*, 2008.

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

[3] M. Efron, D. Turnbull, and C. Ovalle. University of Texas School of Information at TREC 2007. In *Proceedings of TREC'07*, 2007.

[4] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog feed search. In *Proceedings of SIGIR'08*, pages 347–354, 2008.

[5] D. Hawking and P. Thomas. Server selection methods in hybrid portal search. In *Proceedings of SIGIR'05*, pages 75–82, 2005.

[6] J. He, W. Weerkamp, M. Larson, and M. de Rijke. An effective coherence measure to determine topical consistency in user-generated content. *IJDAR*, 12(3):185–203, 2009.

[7] M. A. Hearst, M. Hurst, and S. T. Dumais. What should blog search look like? In *Proceedings of the 2008 ACM workshop on Search in social media*, pages 95–98, 2008.

[8] M. Keikha, S. Gerani, and F. Crestani. Temper: A temporal relevance feedback method. In *Proceedings of ECIR'11*, pages 436–447, 2011.

[9] C. Macdonald and I. Ounis. Key blog distillation: ranking aggregates. In *Proceeding of CIKM'08*, pages 1043–1052, 2008.

[10] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 Blog Track. In *Proceedings of TREC'07*, 2007.

[11] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC-2009 Blog Track. In *Proceedings of TREC'09*, 2009.

[12] G. Mishne and M. de Rijke. A study of blog search. In *Proceedings of ECIR'06*, pages 289–301, 2006.

[13] J. Seo and W. B. Croft. Blog site search using resource selection. In *Proceedings of CIKM'08*, pages 1053–1062, 2008.

[14] I. Soboroff, A. de Vries, and N. Craswell. Overview of the TREC 2006 Enterprise Track. In *Proceedings of TREC'06*, 2006.

[15] W. Weerkamp, K. Balog, and M. de Rijke. Blog feed search with a post index. *Information Retrieval*, 14(5):515–545, 2011.

[16] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.