Blog Site Search Using Resource Selection

Jangwon Seo
jangwon@cs.umass.edu

W. Bruce Croft
croft@cs.umass.edu

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003

ABSTRACT

A blog site consists of many individual blog postings. Current blog search services focus on retrieving postings but there is also a need to identify relevant blog sites. Blog site search is similar to resource selection in distributed information retrieval, in that the target is to find relevant collections of documents. We introduce resource selection techniques for blog site search and evaluate their performance. Further, we propose a "diversity factor" that measures the topic diversity of each blog site. Our results show that the appropriate combination of the resource selection techniques and the diversity factor can achieve significant improvements in retrieval performance compared to baselines. We also report results using these techniques on the TREC blog distillation task.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Experimentation

Keywords

blog site, site search, resource selection

1. INTRODUCTION

Web logs or blogs are an increasingly popular method of recording and communicating personal opinions and views, and the scale of the "blogosphere" has grown dramatically. While blogs share some similar features with traditional Web pages, they also have distinct characteristics in that they have structural features to help users generate content and contain mostly subjective content with little editing. Search techniques customized for blogs are needed to identify relevant material amongst the enormous amount of blog "noise". The creation of the TREC Blog track represents an effort in that direction.

Most blog search services focus on blog postings (although Google's blog search[1] does provide links to "related blogs"). One reason for this is that most researchers and service providers view blog searches and general Web page searches as being essentially the same thing.

On the other hand, as blog subscription methods such as RSS and ATOM have become more prevalent, it is important to be able to identify relevant blogs as well as blog postings. When selecting blogs to subscribe through RSS or ATOM, it should be more effective to find blogs which cover mostly the topic of interest than to find blogs which contain a few relevant postings. Further, many blogs address a small number of specific topics rather than being completely general. If there is a relevant blog related to a specific topic, then that blog can be expected to consistently generate good quality postings about the topic. The creation of Blog Distillation Task of the TREC 2007 Blog track [15] whose goal is finding a feed with a "principle, recurring interest in a topic", reflects the interest in this type of search.

In this paper, we focus on search techniques for complete blogs rather than postings. Since the term "blog search" often means "posting search" we instead use the term "blog site search", where a blog site refers to the collection of postings in the blog.

As an example of the difference between blog site and blog posting searches, consider the following two queries:

Q1: "Nikon D3 review"
Q2: "digital camera reviews"

In the case of Q1, the user has specified a product name and probably expects to retrieve postings reviewing that product. Generally, blog sites containing reviews about only one product are rare and such reviews are scattered over many review sites. Therefore, Q1 would be better handled using posting search. On the other hand, Q2 is more general. Although it would be difficult for a single posting to include all the content relevant to Q2, a set of postings, i.e. a blog site, can address a general topic. Q2 is appropriate for blog site search, and is more likely to lead to a subscription to a feed.

In the general context of information retrieval research, blog site search is very similar to resource selection in distributed information retrieval. Finding relevant blog sites can be regarded as selecting relevant collections from a number of collections, in that each blog site can be considered

[1] http://blogsearch.google.com

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.

as a collection of postings. Thus, in this paper, we study how to apply resource selection techniques to blog site search and further suggest customized methods to improve retrieval performance.

The rest of this paper is organized as follows. In Section 2, we introduce resource selection techniques for blog site search. In Section 3, we present the experimental setup and the results. In Section 4, we discuss types of blog sites and propose diversity penalties that improve retrieval performance. In Section 5, we compare the posting search technique to the blog site search technique. In Section 6, we apply our blog site search techniques to the TREC blog distillation task. In Section 7, we briefly survey related work on blog search. Finally, we conclude with a discussion of the results in Section 8.

2. RESOURCE SELECTION TECHNIQUES FOR BLOG SITE SEARCH

Resource selection in distributed information retrieval is used to select the most relevant collections from a large number of possible collections. Since a blog site is a collection of postings and our target is finding relevant blog sites, we can employ existing resource selection techniques for blog site search.

The goal of resource selection used for blog site search is, however, somewhat different from that of distributed information retrieval. In distributed information retrieval, resource selection is used as a means for finding relevant documents in each collection. That is, if a system can effectively find relevant documents in the distributed environment, the performance of the resource selection technique used in the system does not matter. On the other hand, our goal is to find relevant collections, i.e. blog sites, rather than relevant documents. Of course, we could use blog site search as a technique for improving posting search. The work reported here, however, focuses on retrieving relevant blog sites without considering posting search. Hence, some care must be taken when using resource selection methods developed for distributed information retrieval for blog site search.

In this section, we introduce three resource selection techniques for blog site search. Two of them are natural adaptations of traditional resource selection techniques and we consider them as our baselines. The other is our new approach. Each method uses retrieval based on language models [5].

2.1 Global Representation

One of the simplest approaches to resource selection treats a collection as a single, large document [3, 25]. For a blog site search, we can generate a virtual document for a blog site by concatenating all postings in a blog. This virtual document D_i for a blog site c_i can then be represented using a language model (probability distribution of words) and the query likelihood of the document for a query Q is used as a ranking function.

$$\begin{aligned}\lambda_{GR}(Q, c_i) &= P(Q|D_i) \\ &= \prod_{q \in Q} P(q|D_i) \\ &= \prod_{q \in Q} \frac{tf_{q,D_i} + \mu \cdot cf_q/|C|}{|D_i| + \mu}\end{aligned}$$

where q is a query term of query Q, tf_{q,D_i} is the number of times term q occurs in virtual document D_i, $|D_i|$ is the length of virtual document D_i, cf_q is the number of times term q occurs in the entire collection, $|C|$ is the length of the collection, and μ is a Dirichlet smoothing parameter [27].

This simple, intuitive method was effective in TREC 2007 blog distillation task without any help from advanced techniques [7, 21]. Since the blog distillation task is very similar to blog site search, this method can be considered as a strong baseline. However, this technique has some problems. One of the problems is that the virtual document might be a mixture of various topics. In this case, it is hard for a single language model to accurately reflect the content of the blog site. Further, the content of the virtual document can be skewed by a few large postings.

We call this technique "global representation" and use it as the first baseline for our experiments.

2.2 Query Generation Maximization

Si and Callan introduced a state-of-the-art technique for resource selection based on estimating the probabilities of relevance of documents in the distributed environment [22]. This method, which is referred to as "unified utility maximization", does resource selection to maximize a utility function.

The utility function can be defined as a solution of two types of maximization problems. One is for high-recall and the other is for high-precision. Since our goal is finding relevant collections rather than relevant postings, we consider the high-recall case. The utility function for the high-recall problem is defined as follows:

$$U(\vec{\sigma}) = \sum_{i=1}^{N_C} I(c_i) \sum_{j=1}^{\tilde{n}_i} \tilde{R}(d_{ij})$$

where c_i is a collection, i.e. $\{d_{i1}, d_{i2}, \cdots\}$, N_C is the number of total collections, $\tilde{n}_i$ is the number of the returned documents from the collection c_i and $I(c_i)$ is an indicator function which is 1 if c_i is selected and 0 otherwise. $\vec{\sigma}$ is a selection vector, i.e. $[I(c_1), I(c_2), \cdots, I(c_{N_C})]$ and $\tilde{R}(d_{ij})$ is an estimated probability of relevance of the returned document d_{ij}. As mentioned above, our goal is finding a selection vector to maximize the utility function with the limited number of selection; thus, the problem is described as follows:

$$\vec{\sigma}^* = \arg\max_{\vec{\sigma}} \sum_{i=1}^{N_C} I(c_i) \sum_{j=1}^{\tilde{n}_i} \tilde{R}(d_{ij}) \quad (1)$$

$$subject\ to: \sum_{i=1}^{N_C} I(c_i) = N_{\vec{\sigma}} \quad (2)$$

where $N_{\vec{\sigma}}$ is the predetermined number for selection. The optimized solution of this problem is selecting $N_{\vec{\sigma}}$ collections with the largest expected number of the relevant documents, i.e. $\sum_{j=1}^{\tilde{n}_i} \tilde{R}(d_{ij})$.

In order to apply this method to blog site search, we simplify the process as follows. We build an index of postings ignoring which blog site the postings are from. Since we already know statistics of each collection, we can directly translate the query likelihood score to the probability of relevance of the document for a given query without any estimation process. Therefore, by substituting a query likelihood score for the probability of relevance, $R(d_{ij})$, we can

rewrite Equation (1) as follows:

$$\vec{\sigma}^* = \arg\max_{\vec{\sigma}} \sum_{i=1}^{N_C} I(c_i) \sum_{j=1}^{\tilde{n}_i} P(Q|d_{ij})$$

where $P(Q|d_{ij})$ is the query likelihood of the document d_{ij} for the query Q as follows.

$$\begin{aligned} P(Q|d_{ij}) &= \prod_{q \in Q} P(q|d_{ij}) \\ &= \prod_{q \in Q} \frac{tf_{q,d_{ij}} + \mu \cdot cf_q/|C|}{|d_{ij}| + \mu} \end{aligned}$$

In this case, the optimized solution is selecting $N_{\vec{\sigma}}$ collections with the highest expected generation of the query, i.e. $\sum_{j=1}^{\tilde{n}_i} P(Q|d_{ij})$.

We induce a ranking function based on the maximization.

$$\lambda_{QGM}(Q, c_i) = \sum_{j=1}^{\tilde{n}_i} P(Q|d_{ij})$$

Therefore, what we need to do is simply sum the query likelihood scores of postings from the same blog site in the ranked list which is returned from the index. Next, we can obtain a final ranked list in decreasing order of the sum value. It means that this method can be easily implemented by a simple post-processing after posting search.

We call this modified method "query generation maximization" and use it as the second baseline for our experiments.

2.3 Pseudo-Cluster based Selection

Xu and Croft [26] showed that distributed information retrieval using clustering is very effective because clustering redistributes documents in collections and makes topic-based sub-collections. There are two methods to use clustering for distributed information retrieval. One is the global clustering method. It makes clusters using all documents regardless of the collection. The other is the local clustering method. It makes clusters using documents within a collection. After clustering, both of the methods build an index for each cluster and retrieve documents from relevant clusters.

However, since our goal is not to find relevant documents using resource selection but to find resources themselves, redistribution of documents of each collection using clustering is not likely to be effective. Instead, we create "pseudo-clusters" by ranking blog postings and then grouping highly-ranked postings from the same blog. To represent the pseudo-clusters, we borrow a method from cluster-based retrieval.

Liu and Croft has reviewed various representation methods for cluster-based retrieval [12]. However, most of the representation methods have problems. One of the biggest problems is that the representation of a cluster can be biased by some documents in the cluster. For example, in case of the centroid vector representation, the representation depends on term frequencies of terms in each document. If a document has a specific term which has very high term frequency, the weight of the term might be quite high in the centroid vector even when other documents in the cluster do not have the term. This problem might be significant because our pseudo-clusters usually have only a small number of documents.

To avoid such a problem, we customize a new representation method suggested by Liu and Croft [13]. This method expresses probability distribution of words over clusters using a geometric mean as follows:

$$P(w|g) = \left(\prod_{j=1}^{N_g} P(w|d_j) \right)^{\frac{1}{N_g}}$$

where w is a word, g is a cluster, d_j is a document in cluster g, and N_g is the number of documents in cluster g. The geometric mean is relatively robust against the situation where the influence of some documents overwhelms that of the others.

If we apply the representation method to our pseudo-cluster, then we can easily compute a query likelihood of blog site c_i by a geometric mean of query likelihoods of postings of blog site c_i in the ranked list (under a unigram assumption) as follows.

$$\begin{aligned} P(Q|c_i) &= \prod_{q \in Q} P(q|c_i) \\ &= \prod_{q \in Q} \left(\prod_{j=1}^{\tilde{n}_i} P(q|d_{ij}) \right)^{\frac{1}{\tilde{n}_i}} \\ &= \left(\prod_{j=1}^{\tilde{n}_i} \left(\prod_{q \in Q} P(q|d_{ij}) \right) \right)^{\frac{1}{\tilde{n}_i}} \\ &= \left(\prod_{j=1}^{\tilde{n}_i} P(Q|d_{ij}) \right)^{\frac{1}{\tilde{n}_i}} \end{aligned}$$

Note that the number of documents from each blog site in the ranked list is different in contrast to Liu and Croft's original method using actual clustering. Although query generation maximization also assumes different numbers of documents for blog sites, it looks reasonable that blog sites having more relevant postings, i.e. more documents in the ranked list get good scores. On the other hand, in case of representation by a geometric mean, this causes a problem. For example, a blog site p has a single document in a ranked list and the document is ranked at the second place, whereas a blog site q has three documents in the ranked list, which are ranked at the first, the third and the fourth places. In this case, blog site p might have a higher geometric mean than q. This seems unfair. To resolve this, we use K documents with high ranks in the ranked list regardless of the number of documents of each blog site, where K is a parameter independent of clusters. Then, our ranking function is defined as follows.

$$\lambda_{PCS}(Q, c_i) = \left(\prod_{j=1}^{K} P(Q|d_{ij}) \right)^{\frac{1}{K}}$$

If a blog site has less than K documents in the ranked list, then we can estimate the upper bound of the geometric mean of the blog site using the minimum query likelihood score in the list as follows.

$$\begin{aligned} d_{\min} &= \arg\min_{d_{ij}} P(Q|d_{ij}) \\ \lambda_{PCS}(Q, c_i) &= \left(P(Q|d_{\min})^{K - \tilde{n}_i} \prod_{j=1}^{\tilde{n}_i} P(Q|d_{ij}) \right)^{\frac{1}{K}} \end{aligned}$$

This can be also simply computed from the ranked list of postings. We refer to this method as "pseudo-cluster selection".

3. EXPERIMENTS

3.1 Data

We used the TREC Blogs06 Collection [14] for experiments. The collection was crawled by the University of Glasgow from December 6, 2005 to February 21, 2006 and contains 3,215,171 postings and 100,641 unique blog sites. Since our approaches are based on the postings, we used only posting components in the collection. The postings were stemmed by the Porter stemmer after HTML tags were removed.

We made new relevance judgments for blog site search for ourselves. We selected 50 queries from queries of the topic distillation task of the TREC 2002 Web Track and the TREC 2003 Web Track. The queries of the topic distillation task are a mixture of abstract queries and explicit queries, and we felt that they fit well with the experiments.

To make the relevance judgments for each query, we used a pooling method [23]. Three techniques introduced in Section 2, relevance feedback [2] and dependence models [16] contributed to the pools. As a result, we made judgments for about 2,500 blog sites. The criteria used for relevance is as shown in Table 1.

In the second set of experiments, we used the data for the TREC 2007 blog distillation task.

3.2 Experimental Design

We do experiments for three resource selection techniques, i.e. global representation, query generation maximization and pseudo-cluster selection.

For global representation, we built an index of each blog site after concatenating each posting from the same blog site. We used the query likelihood retrieval model as the ranking method for the global representation. Query generation maximization and pseudo-cluster selection require an initial retrieval. We built an index from all postings and used the query likelihood retrieval model for the initial run. To get the result, we post-processed the results of the initial run by using the respective technique.

For our experiments, we used Indri [24] as the retrieval system. Indri is a search engine based on both the language modeling and the inference network frameworks.

3.3 Training

We performed exhaustive grid search to find optimal parameters for each technique. In case of the global representation, we have one parameter to be trained, i.e. the μ parameter for Dirichlet smoothing. The query generation maximization requires training for two parameters, i.e. the smoothing parameter and the number of the documents to be used for the post-process of the results of the initial retrieval, N_R. For the pseudo-cluster selection, the parameter for the cluster size restriction, K, is additionally required. We used the normalized discounted cumulative gain (NDCG) [10], the mean average precision (MAP) and the precision at the rank 10 (P@10) as the evaluation measures. For binary relevance judgment-based metrics such as MAP and P@10, we regarded a blog site having a grade of Table 1 equal to or greater than 1 as a relevant blog site. The parameter trainings were also done for each measure.

3.4 Evaluation

We performed 10-fold cross validation in order to evaluate performance. 50 queries were randomly partitioned. For one partition, the parameters were trained with all the other partitions and performance for the partition is evaluated with the trained parameters. We concatenated the ranked lists from each partition and evaluated them.

3.5 Retrieval Performance

Table 2 presents the performance of each resource selection method. Two baselines, global representation and query generation maximization showed similar performance. Pseudo-cluster selection significantly outperformed the other techniques.

In a practical sense, query generation maximization and pseudo-cluster selection have an advantage over global representation. Nowadays, most of the blog publishing or blog search service providers have already provided posting search services. Since the two techniques use the results of posting search, they can be easily implemented by reusing the index for posting search.

Note that ranking methods other than query likelihood could have been used for the initial run. Although we will not explore it in detail here, advanced retrieval techniques such as relevance models [11] or the dependence model [16] are likely to improve performance. We leave this as future work.

4. CUSTOMIZING THE SEARCH

Blog site search involves somewhat different strategies compared to resource selection due to specific features of blog sites. For better resource selection, it is desirable to choose collections which include a greater number of relevant documents. This might not be always true for blog site search. We discuss which customizations may be appropriate by first introducing several types of blog sites in the next subsection.

4.1 Types of Blog Sites

In order to better understand the problem of blog site retrieval, we classified blog sites into three types based on how they are managed and the degree of diversity of the topics covered.

Type I is the diary type of blog. In this type, a blogger usually posts descriptions of their daily life. In many cases, the postings are related to personal issues such as relationships, appointments or concerns. Some postings can be about a person's interests or opinions about a specific issue or object. However, it is rare that other postings about similar topics are regularly updated in the blog site.

Type II is the news blog. Documents covering a large number of topics are posted, and many of these blogs are managed by an organization or a company. Another common situation is when most of the postings are not composed by the blogger but are collected by the blogger. For example, if a blogger finds some good articles while surfing news sites, they may copy and paste them into their own blog. In this case, the blog functions like a scrapbook, which causes many duplicate documents over the whole web collection. In sum, even though this type contains relatively good quality documents, it often lacks originality and is not topic-centric.

Table 1: The criteria for the relevance judgments.

Grade	Criterion
0	The blog site does not consistently create postings relevant to the topic.
1	More than 25% of the postings in the blog deal with the topic.
2	More than 50% of the postings in the blog deal with the topic.
3	More than 75% of the postings in the blog deal with the topic.

Table 2: Resource selection performance. α and β in a cell indicate statistically significant improvement ($p < 0.1$) over the baselines, global representation and query generation maximization, respectively.

	NDCG	MAP	P@10
Global Representation	0.5448	0.3708	0.2780
Query Generation Maximization	0.5422	0.3785	0.2920
Pseudo-cluster Selection	0.5632	$0.4091^{\alpha\beta}$	$0.3300^{\alpha\beta}$

Furthermore, this second type is related to an important issue of blog search. Blogs are a subset of general Web pages. When blog search services crawl the Web to find blog postings, they typically identify them by checking whether the Web page contains a feed link for RSS or ATOM. Many general Web news sites also contain feed links for their subscribers, and this can cause these sites to be included in the blog collection. Since such sites have not only a large number of good quality documents but also relevant documents for all kinds of topics, they may often be retrieved. To prevent this requires some type of penalty factor.

Type III is the topic-focused type of blog. This is managed by one or a few individuals and concentrates on a small number of topics. The quality of postings varies on the blogger, but often is good. This type of blog site with a topic specialty exists for many topics. The typical examples that are frequently seen are product review blogs or political advocate blogs. It is probable that documents related to the specific topic are regularly posted in this type of blog site. The success of our retrieval methods will depend on how well we are able to find this type of blog site for a given query. Table 3 summarizes the properties of each type.

To verify the validity of our categories, we manually classified 100 blog sites randomly selected from the pools for relevance judgments described in Section 3.1. Of course, it is not easy to simply classify a blog site into a single category because diary postings, news postings and topic-focused postings might coexist in a blog site. For this reason, we classified them by observing what type of postings mainly exists in the blog site. There were some cases that we could not decide which category a blog site is in because it did not match any category. Most of such blog sites were spam sites, e.g., sites which do not contain real contents but instead are mostly advertisement links. We tagged such sites as "Unclassifiable".

Three annotators independently labeled the blog sites. By majority voting, we assigned the label which more than two annotators agreed with to each blog site. If all annotators had different labels for a blog site, then we tagged the site as "Unclassifiable". Table 4 presents the result. Most blog sites were mapped onto our categories. As we expected, the majority of relevant blog sites were in the topic-focused category. To measure inter-annotator agreement, Fleiss' κ [9] was computed. The coefficient was 0.76 and this indicates a substantial agreement.

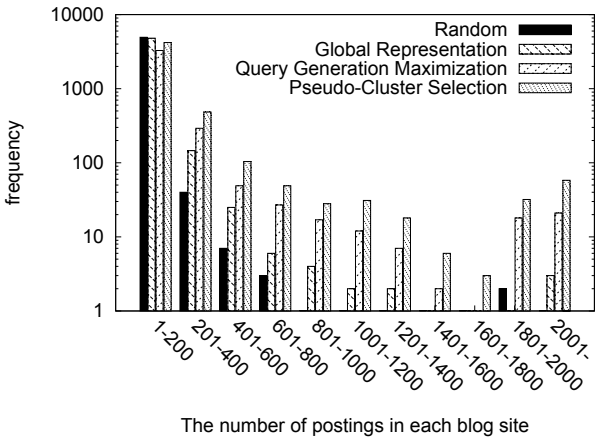

Figure 1: The distribution of the number of postings in the blog sites returned by each resource selection technique

4.2 Diversity Penalty

Based on the previous subsection, we need to penalize Type I and Type II blog sites. To do this, we focus on the fact that they are not topic-centric. Accordingly, we considered a method for penalizing blog sites with diverse topics.

We have to decide whether or not the blog site is topic-centric at the global level, i.e. the blog site level. Therefore, the penalty should be able to be used at the global level. Further, it will be more helpful if the penalty can reflect the relevance for the topic.

4.2.1 Diversity Penalty by Global Representation

We already have seen a component that could be used as a diversity penalty. It is the query likelihood score from the global representation used as a baseline in Section 2. We compute the score at the global level. Further, if the blog site deals with the diverse topics, then the distribution of the words in the blog site are probably widely scattered. As a result, the occurrence of the words closely related to a specific topic might be relatively low compared to the topic-centric blog sites. It causes a low query likelihood score in the language modeling-based retrieval.

Figure 1 shows indirect evidence supporting this claim.

Table 3: Type classification of blog sites

Type	Topic-centric	Document Quality	Originality
Type I (Diary)	Low	Low	High
Type II (Newspaper)	Low	High	Mid
Type III (Topic-focused)	High	Mid	High

Table 4: Manual classification result with 100 blog sites

Type	#Blog Sites	#Relevant Blog Sites
Unclassifiable	7	0
Type I (Diary)	26	2
Type II (News)	25	1
Type III (Topic-focused)	42	11

We obtained the result ranked list for 50 queries by using each resource selection technique. We analyzed the distribution of the number of postings in the returned blog sites according to the above mentioned techniques. Further, we provide the distribution of the number of postings of blog sites in the entire collection by randomly selecting the same number of blog sites as those in the ranked lists ("Random" in Figure 1). As we can see from the histogram in Figure 1, the global representation definitely returned much fewer blog sites which have a large number of postings. Although it is an over-generalization to assume that a blog site having many documents is diverse, there is such a tendency. For example, it is apparent that the news sites where thousands of articles are posted daily have much more documents than the topic-focused blogs where at most several postings a week are registered.

In summary, the query likelihood score can be useful as a measure of diversity of blog sites. Furthermore, this score reflects the relevance of the blog site for the given topic. Accordingly, to supplement the other two resource selection techniques, we can use this score as a penalty factor for diversity by multiplying it by the previous ranking function as follows.

For query generation maximization,

$$\lambda_{QGM\text{-}GR}(Q, c_i) = \lambda_{QGM}(Q, c_i) \cdot \lambda_{GR}(Q, c_i)^{\pi}$$

For pseudo-cluster selection,

$$\lambda_{PCS\text{-}GR}(Q, c_i) = \lambda_{PCS}(Q, c_i) \cdot \lambda_{GR}(Q, c_i)^{\pi}$$

where π is a weight parameter. The multiplication is used to prevent from being biased as in Section 2.3. Further, it can be interpreted as a linear combination of the log probabilities.

4.2.2 Clarity Score as a Penalty Factor

Another candidate which we can consider as a penalty factor for diversity is a clarity score. Cronen-Townsend et al. [6] showed that query performance can be predicted using the relative entropy between a query language model and the corresponding collection language model as a clarity score. That is, since the query which has the similar language model to that of the collection seems somewhat ambiguous, we do not expect good retrieval performance with that query.

However, in our work, we want to know the difference between a blog site and the whole collection rather than between a query and a collection. We assume that if a blog site covers many general topics, then the language model of the blog site is similar to that of the whole collection. On the other hand, in a blog site which addresses a few specific topics, some terms related to the topics occur relatively frequently and accordingly, the language model is expected to be different from that of the whole collection. Thus, we compute the clarity score by using the relative entropy, or Kullback-Leibler divergence [4] between a blog site and the whole collection as follows.

$$\text{Clarity}(c_i) = \sum_{w} P(w|c_i) \log \frac{P(w|c_i)}{P(w|\text{Coll})}$$

We also use this score as a penalty factor for diversity by multiplying it by the previous ranking function as follows.

For query generation maximization,

$$\lambda_{QGM\text{-}Clarity}(Q, c_i) = \lambda_{QGM}(Q, c_i) \cdot \text{Clarity}(c_i)^{\pi}$$

For pseudo-cluster selection,

$$\lambda_{PCS\text{-}Clarity}(Q, c_i) = \lambda_{PCS}(Q, c_i) \cdot \text{Clarity}(c_i)^{\pi}$$

4.2.3 Diversity Penalty by Random Sampling

We need to keep additional information like the index for global representation in order to use two penalty factors introduced above because they depend on the statistics of a whole blog site. This requirement might be a considerable burden for most blog service providers. Further, the penalty factors ignore boundaries of postings, and accordingly, there can be bias problems. As seen in Figure 1, the global representation is biased toward small size blog sites. Both penalty factors favor collections which have long postings because such long postings dominate the whole blog site, regardless of the number of them, and the blog sites are considered topic-centric.

To address these problems, we suggest a randomized approach. In pseudo-cluster selection, we use postings in the ranked list to get postings relevant to a given topic. On the other hand, we randomly sample M postings from a blog site to obtain postings independent of any topic. Note that the randomly sampled postings might or might not be in the ranked list. We compute the query likelihoods for the sampled postings with the given query. If the blog site is topic-centric and relevant to the topic, then the postings are likely to relevant to the topic and the query likelihoods have high values. Otherwise, postings about various topics are picked and the query likelihoods have small values. Therefore, the query likelihoods can be used for estimating diversity of a blog site. Further, this approach is free from bias

problems in that postings are directly used, and additional information is not required.

We make a diversity penalty factor with the query likelihoods of the randomly sampled postings in the same way as used in pseudo-cluster selection. In other words, we compute a geometric mean of the query likelihoods. This diversity penalty factor can be used by multiplying it by the previous ranking function as follows.

For query generation maximization,

$$\lambda_{QGM\text{-}Random}(Q, c_i) = \lambda_{QGM}(Q, c_i) \cdot \left(\prod_{j=1}^{M} P(Q|r_{ij}) \right)^{\frac{\pi}{M}}$$

For pseudo-cluster selection,

$$\lambda_{PCS\text{-}Random}(Q, c_i) = \lambda_{PCS}(Q, c_i) \cdot \left(\prod_{j=1}^{M} P(Q|r_{ij}) \right)^{\frac{\pi}{M}}$$

where r_{ij} is the j^{th} randomly selected posting of blog site c_i.

A problem of this random sample-based approach is that the retrieval result is changed every time even when there is not any change in the target collection. Such unstable search results might frustrate users. Therefore, a specific (pseudo-random) sampling may be more desirable than purely random sampling. The choice of a sampling method depends on the goals of blog site search services. If a blog site search service favor blog sites that have a more recent focus on a specific topic, then using M recent postings in each blog site instead of randomly sampled postings can be a good choice. We provide experimental results in cases of using recent postings as well as randomly sampled postings in the next section.

4.3 Experimental Results

We did experiments to study the effectiveness of each suggested penalty factor. Table 5 shows the experimental results after applying the penalty factors.

The results show that there is the improvement in performance for both of the methods in case of using the global representation score as the penalty factor. In the experiment for the query generation maximization, the effectiveness according to MAP and P@10 became better, but the improvement was not still statistically significant except for P@10. In case of pseudo-cluster selection, the performance only for MAP was improved, whereas the performance for P@10 and NDCG was similar or lower compared to the original method. Nevertheless, the performance with respect to the baselines for both of the measures was statistically significantly improved.

In contrast, using the clarity score as a penalty factor hurt the overall performance. Although the degradation of the performance with respect to the baselines for all the measures was not statistically significant, the performance became consistently worse. The reason is that the clarity score is independent of the queries and does not reflect the relevance for the topics.

Since the results by a penalty factor by randomly sampled postings are different every time, we did the same run 10 times and used the average of evaluation values for each query. Figure 2 shows the change of the MAP score for each run of pseudo-cluster selection with a penalty factor by random postings. Note that the scores have similar values to an

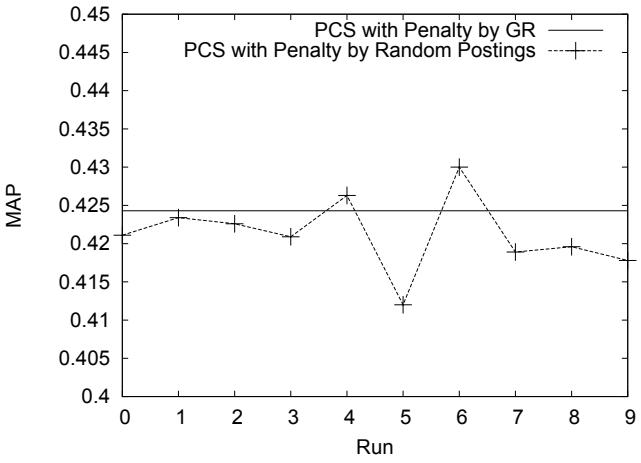

Figure 2: MAP scores for each run of pseudo-cluster selection with a peanlty factor by random postings. GR and PCS stand for global representation and pseudo-cluster selection, respectively.

MAP score of pseudo-cluster selection with a penalty factor by global representation. There was no statistically significant difference ($p < 0.1$) between the performance of pseudo-cluster selection with a penalty factor by global representation and the performance of each run of pseudo-cluster selection with a penalty factor by random postings.

Penalty factors by random sampling were very effective for both query generation maximization and pseudo-cluster selection. The methods consistently showed substantial performance gain for NDCG and MAP (and a small loss for P@10) compared to the original method. The improvement is statistically significant over baselines. In particular, a penalty factor by recent postings showed the best performance in our experiments. However, we cannot rule out the possibility that our relevance judgments are unconsciously biased toward recent postings of each blog site.

5. POSTING SEARCH TECHNIQUE VS. BLOG SITE SEARCH TECHNIQUE

One of the persistent questions throughout our work is whether there is an actual difference between posting search and blog site search. If the posting search technique is effective enough for blog site search, then we do not need to distinguish blog site search from posting search.

To answer this question, we compare the effectiveness of using a posting search technique for blog site search with our baseline techniques. After running a posting search, we need to convert the posting search results to a blog site search result. We keep the order of the postings returned and convert each posting ID to a blog site ID that it belongs to. If there is more than one posting from a blog site in the ranked list, we use only the best posting of the blog site. This method is widely used for the topic distillation tasks or the named page finding tasks. The ranking function is expressed as follows.

$$\lambda_P(Q, c_i) = \max_j P(Q|d_{ij})$$

When this ranking function is used for web site search, it is usually combined with various web features such as anchor

Table 5: Retrieval performance for resource selection techniques combined with each penalty factor. GR, QGM and PCS stand for global representation, query generation maximization and pseudo-cluster selection, respectively. α and β in a cell indicate statistically significant improvement ($p < 0.1$) over the baselines, global representation and query generation maximization, respectively.

	NDCG	MAP	P@10
QGM with Penalty by GR	0.5344	0.3957	$0.3040^{\alpha\beta}$
PCS with Penalty by GR	0.5631^{α}	$0.4217^{\alpha\beta}$	$0.3240^{\alpha\beta}$
QGM with Penalty by Clarity	0.5286	0.3610	0.2760
PCS with Penalty by Clarity	0.5207	0.3444	0.2720
QGM with Penalty by Random Postings	0.5579^{β}	$0.4011^{\alpha\beta}$	0.3012
PCS with Penalty by Random Postings	$0.5782^{\alpha\beta}$	$0.4213^{\alpha\beta}$	$0.3252^{\alpha\beta}$
QGM with Penalty by Recent Postings	0.5705^{β}	$0.4134^{\alpha\beta}$	$0.3080^{\alpha\beta}$
PCS with Penalty by Recent Postgins	$0.5841^{\alpha\beta}$	$0.4323^{\alpha\beta}$	$0.3280^{\alpha\beta}$

text, PageRank or HTML structures. However, we used the basic query likelihood for our methods and our methods are likely to be also improved by such features. Therefore, for now, it seems appropriate and fair to use query likelihood for the posting search without other features. We leave studies about the way to exploit web features for blog site search for future work.

Table 6 shows the results of experiments with the posting search technique. As we can see, the technique performed significantly worse than the baselines. This result shows that posting search techniques are not desirable for blog site search because the relevance of the posting is not necessarily related to the relevance of the blog site which contains the relevant posting.

6. BLOG DISTILLATION TASK

The blog distillation task which was defined in TREC 2007 [15], identifies feeds relevant to a specific topic. The task is almost the same as blog site search in that a blog site generally has a feed and the feed is a summary of the blog site. Therefore, we can apply our blog site search techniques to the task.

The judgment set of TREC 2007 blog distillation contains 17,411 judged feeds for 45 topics. Although the distillation task is finding relevant feeds in the feed components in the TREC Blogs06 collection, we use only the posting collections as done before. Thus, we have to convert result blog site IDs to the feed IDs.

We applied the same baselines and the techniques that showed good performance in the previous experiments, i.e., global representation, pseudo-cluster selection and pseudo-cluster selection with a penalty factor by global representation, random postings and recent postings for penalizing diversity. We used parameters learned by our relevance judgments. Table 7 shows the results of experiments.

Surprisingly, global representation performed better than pseudo-cluster selection. We suspected that the reason is that pseudo-cluster selection is sensitive to query lengths. To confirm our assumption, we computed the correlation between the query length and the following performance differences of global representation and pseudo-cluster selection.

$$MD = \frac{MAP_{GR} - MAP_{PCS}}{MAP_{PCS}} \quad (3)$$

where MAP_{GR} and MAP_{PCS} are the Mean Average Precision (MAP) of the global representation and the pseudo-cluster selection, respectively.

Kendall's τ was computed with MD and the number of terms in each query where p-value < 0.1. The correlation coefficient value was about 0.2 and the result was statistically significant. Since the value is somewhat small, we cannot say that they are tightly correlated. Nevertheless, there is some relationship between them. That is, for the longer queries, pseudo-cluster selection can be better than global representation. This is not unreasonable. Since global representation uses a greater amount of text, other terms closely related to the topic but not in the query as well as the query terms can be often used in the relevant blog. That is, the effect of the terms in the query is diluted by the large amount of text. Consequently, if the query is long or it contains terms which are not generally used, then even a relevant blog might be determined to be irrelevant to the query. On the other hand, pseudo-cluster selection is a technique that represents a cluster with a relatively small number of documents (In our experiments, $K = 5$). Here, the documents are directly selected by the initial search using the given query. Therefore, when the query is clear, pseudo-cluster selection works well. But, when the query is somewhat general, ambiguous or short, the initial search result is likely to be unreliable. Consequently, pseudo-cluster selection can perform poorly in these situations. While the average number of terms of queries in our relevance judgment set is 2.6, the average number for the queries in the blog distillation judgment set is 1.9. This difference might be critical for pseudo-cluster selection.

On the other hand, the combination of global representation and pseudo-cluster selection significantly outperformed the baselines. In fact, the MAP score is as good as the best reported in the TREC 20007 blog distillation task [7, 19]. While the best run achieved the performance by a novel query expansion technique, our method uses a simple post-processing of query likelihoods, which does not require any other information but a posting index. Furthermore, our technique is likely to be integrated with various techniques like query expansion to gain the better performance. That is, this approach is very effective for the blog distillation task as well as for the blog site search.

Penalty factors by random sampling were still effective but not as much as that they showed on our dataset. In particular, the method using recent postings as a penalty factor, which showed the best performance on our dataset, was worse than the method using random postings. This presents that the current blog distillation task does not pursue recency and weighing on recent postings is an inappro-

Table 6: Retrieval performance for posting search. α and β in a cell indicate statistically significant degradation ($p < 0.1$) with respect to the baselines, global representation and query generation maximization, respectively.

	NDCG	MAP	P@10
Posting Search Technique	$0.4801^{\alpha\beta}$	$0.2778^{\alpha\beta}$	$0.2040^{\alpha\beta}$

Table 7: Retrieval performance for the blog distillation task. GR, QGM and PCS stand for global representation, query generation maximization and pseudo-cluster selection, respectively. α and β in a cell indicate statistically significant improvement ($p < 0.1$) over the baselines, global representation and query generation maximization, respectively.

	MAP	P@10
GR	0.3454	0.4889
QGM	0.2709	0.4311
PCS	0.3171	0.4622
PCS with Penalty by GR	$0.3725^{\alpha\beta}$	$0.5356^{\alpha\beta}$
PCS with Penalty by Random Postings	0.3542^{β}	$0.5289^{\alpha\beta}$
PCS with Penalty by Recent Postings	0.3480^{β}	$0.5356^{\alpha\beta}$

priate strategy for the blog distillation task. However, topics addressed by blogs often change. Considering that the blog distillation task is a filtering task for future postings, the importance of recent topics of blogs might be improperly overlooked in the judgment process for the blog distillation task.

Although the method using the global representation score as a penalty factor outperformed the random sampling approaches, the differences are not statistically significant. Considering the practical advantage of the random sampling methods which do not require additional indexes, the methods should be taken into account for the blog distillation task.

7. RELATED WORK

We are not aware of any previous research on blog site search. However, the TREC blog distillation task [15] is similar to blog site search. Arguello et al. [1] and Elsas et al. [8] introduced various blog representations for retrieval and suggested a novel query expansion method using Wikipedia for blog feed search. Besides, there has been some work focusing on blog posting search. Mishne and de Rijke [17] showed that blog posting searches have different goals than general web searches by analyzing blog search engine query logs. The TREC blog opinion task [19] tries to locate blog postings that express an opinion about a given target. Qu et al. [20] tried to automatically categorize blogs into four topics, i.e. personal diary, news, politics, and sports based on tf-idf approaches.

There is much previous work on resource selection in the context of distributed information retrieval. The most common technique for resource selection has been handling each collection as a virtual document. This approach has been explored by some researches [3, 25]. Callan [3] introduced the CORI algorithm based on variants of tf-idf approaches on the virtual document. Xu and Croft showed that the topic-based retrieval using clustering is effective for resource selection [26]. Nottelman and Fuhr [18] introduced a decision-theoretic framework (DTF) to minimize the expected overall cost of distributed retrieval on the basis of the estimation of relevance of the documents. Si and Callan [22] introduced the unified utility maximization (UUM) algorithm which achieves high recall or high precision based on the probabilities of relevance of the document estimated from the centralized index.

8. CONCLUSION AND FUTURE WORK

In this work, we defined the properties of blog sites and the goal of blog site search. Based on this goal, we introduced various resource selection algorithms for site search in blog collections. Furthermore, we classified the types of blog sites and claimed that an appropriate penalty factor reflecting the diversity of the topics of each blog site is required. Our experiments presented that the score of the global representation method can be a good candidate for the factor. Our experiments demonstrated that pseudo-cluster selection combined with a global representation penalty outperformed the other methods, both on our data and for the TREC Blog Distillation task.

In addition, we compared a posting search technique with our customized technique for blog site search. The fact that our blog site search technique is definitely superior to the posting search technique shows that blog site search is different than posting search and that separate techniques for blog site search are necessary.

There are a variety of directions for future work. We used simple language modeling-based retrieval techniques in order to compare the performance of each resource selection method. It would be interesting to investigate other advanced retrieval techniques for blog site search. Further, we would like to further exploit the unique features of blogs. For example, the feed for a blog could provide us with information given that it is generally well formatted and includes the posting time or the summary of the posting. Furthermore, it is a critical issue to study how the results of blog site search could be used to improve the effectiveness of posting search.

9. ACKNOWLEDGMENTS

We thank Jiwoon Jeon and Kyung-Soon Lee for their advice on the pseudo-cluster algorithm. This work was supported in part by the Center for Intelligent Information Retrieval, in part by NHN Corp. and in part by NSF grant

#IIS-0534383. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

10. REFERENCES

[1] J. Arguello, J. Elsas, J. Callan, and J. Carbonell. Document representation and query expansion models for blog recommendation. In *Proceedings of the Second International Conference on Weblogs and Social Media (ICWSM 2008)*, 2008.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[3] J. Callan. Distributed information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, Norwell, MA, USA, 2000.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, USA, second edition, 2006.

[5] W. B. Croft and J. Lafferty. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, USA, 2003.

[6] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306, 2002.

[7] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog distillation. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, 2008.

[8] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog feed search. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 347–354, 2008.

[9] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical Methods for Rates and Proportions*. Wiley, New York, third edition, 2003.

[10] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, 2000.

[11] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, 2001.

[12] X. Liu and W. B. Croft. Representing clusters for retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 671–672, 2001.

[13] X. Liu and W. B. Croft. Evaluating text representations for retrieval of the best group of documents. In *Proceedings of 30th European Conference on IR Research, (ECIR 2008)*, pages 454–462, 2008.

[14] C. Macdonald and I. Ounis. The trec blogs06 collection : Creating and analysing a blog test collection. Technical Report TR-2006-224, University of Glasgow, Department of Computing Science, 2006.

[15] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the trec-2007 blog track. In *TREC 2007 Notebook*, 2007.

[16] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479, 2005.

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

[18] H. Nottelmann and N. Fuhr. Evaluating different methods of estimating retrieval quality for resource selection. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 290–297, 2003.

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

[20] H. Qu, A. L. Pietra, and S. Poon. Automated blog classification: Challenges and pitfalls. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 184–185, 2006.

[21] J. Seo and W. B. Croft. Umass at trec 2007 blog distillation task. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, 2008.

[22] L. Si and J. Callan. Unified utility maximization framework for resource selection. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 32–41, 2004.

[23] K. Sparck Jones and C. J. van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32(1):59–75, 1976.

[24] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[25] J. Xu and J. Callan. Effective retrieval of distributed collections. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 254–261, 1998.

[26] J. Xu and W. B. Croft. Topic-based language models for distributed retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*, pages 151–172. Kluwer Academic Publishers, Norwell, MA, USA, 2000.

[27] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342, 2001.