UMass at TREC 2008 Blog Distillation Task

Jangwon Seo and W. Bruce Croft
Center for Intelligent Information Retrieval
University of Massachusetts, Amherst

Abstract

This paper presents the work done for the TREC 2008 blog distillation task. We introduce two new methods based on blog site search using resource selection which was the framework we used for the TREC 2007 blog distillation task. One is a new factor that penalizes the topical diversity of a blog. The other is a query expansion technique. We compare the methods to strong baselines.

1 Introduction

The TREC 2008 Blog Track is composed of four tasks: the baseline adhoc retrieval task, the opinion finding task, the polarized opinion task and the blog distillation task. We participated in the blog distillation task. The goal of the blog distillation task is *'Find feeds that are principally devoted to topic X'*. The blog distillation task is different from the other tasks because it finds feeds of relevant blogs whereas the other tasks find relevant postings satisfying a special need, e.g., postings with positive opinions. A blog usually has a feed, and finding a feed can be interpreted as finding a blog or a blog site owning the feed.

The TREC 2008 blog distillation task is very similar to the TREC 2007 blog distillation task in terms of the goal and the dataset. Therefore, we use the same framework that we used last year, which is '*blog site search using resource selection*' [6]. The framework consists of two factors: a basis factor for resource selection and a supplementary factor handling topicality of blogs. This year, we suggest a new supplementary factor. Further, we investigate a query expansion technique for our framework.

2 Dataset and processing

We used the permalink collection (posting collection) instead of the feed collection because of the following two reasons. First, the goal of feed search is to find a blog to be subscribed to through the feed link rather than directly getting information from the feed. In other words, the feed search task can be considered as a blog search task. Therefore, we do not need to insist on using only the feed collection. Second, a feed often contains only a small amount of text because it is a short summary of a posting. If we use only the feed collection, retrieval performance may suffer from the sparsity of words that causes a word mismatch problem.

The permalink collection was processed as follows. We stripped HTML tags and stemmed text by the Krovetz stemmer. Stopwords were used only at query time. We did not perform any pre-processing such as filtering out splogs and non-English blogs because our framework was shown to be able to deal with such issues in the TREC 2007 blog distillation task.

Our algorithms retrieve relevant postings or blogs whereas the blog distillation task requires relevant feed IDs (FEEDNO) for the submitted runs. Therefore, we had to convert the posting IDs or the blog IDs in our search results to feed IDs (FEEDNO).

3 Blog Site Search Using Resource Selection

A blog is a collection of its postings. That is, finding relevant blog sites is similar to finding relevant collections. This is a typical cases where resource selection in distributed information retrieval is used. Therefore, we exploit resource selection techniques for this task. We call this approach *"blog site search using resource selection"*. Note that all described techniques are based on language modeling retrieval techniques [1].

3.1 Baseline

We review two resource selection techniques for the blog distillation task: Global Representation and Pseudo Cluster Selection [5, 6].

Global Representation handles a blog as a big document where all postings of the blog are concatenated into a virtual document. We can rank the blogs by a language model learned from the virtual document as follows:

$$\phi_{GR}(Q, c_i) = P(Q|D_{c_i})$$

where $\phi_{GR}(Q, c_i)$ is a ranking function for Global Representation, Q is a query, and D_{c_i} is a virtual document for blog c_i. For Global Representation, we need to make an index for the virtual documents. We refer to the virtual document collection and the index as a blog collection and a blog index, respectively.

Pseudo-cluster Selection is inspired by a clustering-based resource selection technique which is known for being effective [7]. We assume that if some postings of a blog are highly ranked in the posting search result for a query (a topic), then the postings virtually form a topic-based cluster which we call a topic-dependent pseudo-cluster. We can rank the pseudo-clusters by using cluster-based retrieval techniques. Since many cluster representation techniques have bias problems from long documents or high frequency terms in the cluster, we use a geometric-mean cluster representation technique which is known to be relatively free from such biases [4] as follows:

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

where $\phi_{PCS}(Q, c_i)$ is a ranking function for Pseudo-cluster Selection and d_{ij} is the j^{th} posting of blog c_i in the top N posting search result.

Pseudo-cluster Selection requires K postings for each cluster. However, a blog cannot always have more than K postings in the top N posting search result. Therefore, if the number of postings, $\tilde{n}_i$ from blog site c_i in the ranked list is less than K, we manipulate the representation so that the upper bound is used instead.

$$d_{\min} = \arg\min_{d_{ij}} P(Q|d_{ij})$$

$$\phi_{PCS}(Q, c_i) = \left(P(Q|d_{\min})^{K-\tilde{n}_i} \prod_{j=1}^{\tilde{n}_i} P(Q|d_{ij}) \right)^{\frac{1}{K}}$$

This technique uses an index for postings, i.e., documents in the permalink collection. In this paper, we call the permalink collection and the index a posting collection and a posting index, respectively.

Global Representation and Pseudo-cluster Selection have their pros and cons. Global Representation is easily dominated by long postings although it is very simple and it handles topicality of a blog well. Pseudo-cluster Selection behaves like sampling several relevant postings from a blog. As long as the blog addresses a small number of topics, Pseudo-cluster Selection works well. However, if a blog addresses too many topics (for example, the blog contains thousands of postings related to other topics and only ten relevant postings), then this method may not work.

To tackle the weakness of both methods, we consider combinations of the two methods. In the combination, a superior method becomes a basis factor and the other becomes a supplementary factor. The basis factor is fixed whereas the supplementary factor can be sometimes replaced by other techniques depending on tasks. In practice, Pseudo-cluster Selection only uses an index for posting search that

most blog publishing services already have. This benefit in the system implementation gives Pseudo-cluster Selection some superiority. Furthermore, Seo and Croft showed Pseudo-cluster Selection outperformed Global Representation in blog site search [5]. Therefore, Pseudo-cluster Selection is considered as our basis factor. Accordingly, Global Representation becomes the supplementary factor.

The combination is computed by multiplication of the ranking function of each method as follows.

$$\phi_{baseline}(Q, c_i) = \phi_{PCS}(Q, c_i) \cdot \phi_{GR}(Q, c_i) \qquad (1)$$

This combination showed good results in the TREC 2007 blog distillation task [6]. We use it as our baseline here.

3.2 New Supplementary Factor

The combination of Global Representation and Pseudo-cluster Selection is very effective for both feed search and blog site search [5]. However, a blog index is additionally required for Global Representation as we mentioned in Section 3.1. In practice, operating with two indexes for a system requires considerable extra effort. To avoid such a burden, we consider a new supplementary factor which can substitute for Global Representation.

A major role of Global Representation in the combination is penalizing the topical diversity of a blog. A goal of the blog distillation task is locating blogs where users can consistently get relevant information through feed subscriptions. If a user subscribes to a blog which addresses too many topics, then feeds of various topics will be delivered from the blog. Therefore, penalizing topic diversity is necessary for effective feed search. In Global Representation, if a blog addresses many topics, the effect of terms related to one topic tends to be diluted by terms for other topics in the language model. Accordingly, it penalizes topical diversity. Our new supplementary factor must be able to play this role.

Let's consider randomly sampling several postings from a blog without regard to relevance. We can compute the query-likelihood scores of the sampled postings for a given query. If the blog is topic-focused, then many of the sampled postings are likely to address similar topics and have high query-likelihood scores. On the other hand, if the blog addresses various topics, the sampled postings may address different topics and have relatively low query-likelihood scores. Therefore, we can estimate the topical diversity of a blog by these randomly sampled postings. We call a set of the postings sampled from a blog a topic-independent pseudo-cluster. We can rank the pseudo-clusters using the same geometric-mean representation as Pseudo-Cluster Selection. We replace the ranking function of Global Representation in Equation 1 with the ranking function of this new pseudo-cluster as follows:

$$\phi_{random}(Q, c_i) = \phi_{PCS}(Q, c_i) \cdot \left(\prod_{j=1}^{K} P(Q|r_{ij}) \right)^{\frac{1}{K}}$$

where r_{ij} is the j^{th} randomly selected posting of blog site c_i. Seo and Croft showed this supplementary factor is comparable to Global Representation [5].

This supplementary factor, however, produces time-variant results because it is based on random samples. This time-variant property is not usually preferred in information retrieval system in that it may confuse users. Therefore, we suggest a strategic sampling approach, i.e. sampling only recently updated postings instead of sampling postings randomly as follows:

$$\phi_{recency}(Q, c_i) = \phi_{PCS}(Q, c_i) \cdot \left(\prod_{j=1}^{K} P(Q|r'_{ij}) \right)^{\frac{1}{K}}$$

where r'_{ij} is the j^{th} recent posting of blog site c_i.

We believe that this sampling is capable of penalizing diversity because it can still form a topic-independent pseudo-cluster. This sampling is sensitive to temporal aspects of blogs because it prefers blogs which have recently updated relevant postings. This property may be beneficial depending on the type of queries and the goal of tasks. We will look into how suitable this sampling is for the blog distillation task.

3.3 Query Expansion

Query expansion effectively deals with the word mismatch problem caused by short queries. Since queries for the blog distillation task are usually short (each title query of the TREC 2008 blog distillation task topics consists of about two words), we expect that query expansion could play an important role for achieving accurate retrieval. To expand queries, we use the relevance model by Lavrenko and Croft [3].

Our framework uses the combination of factors based on two different collections, i.e. the posting collection and the blog collection. This is somewhat different from the usual environment of query expansion. Diaz and Metzler [2] showed that the relevance model can be improved by using any external collection that contains more relevant documents than the original (target) collection. They expanded queries using the external collection and ran the expanded queries against the target collection. This approach seems appropriate in that we also have two indexes.

Our two collections have different characteristics although they are originally from the same document collection, i.e. the permalink collection. Particularly, when we look at a posting or blog in each collection as a topic unit, a posting is relatively topic-oriented compared to a blog because a blog usually addresses multiple topics. If we retrieve the top N documents from each collection, then topical terms tend to be more densely distributed in the postings than the blogs. Therefore, we set the posting collection as the external collection which contains more relevant documents in terms of the mixture of relevance models. On the other hand, our target collection is the blog collection since our goal is finding relevant blogs or their feeds. We apply the mixture of relevance models to this setting.

4 Experiments

We used the Indri[1] search engine for experiments. We built two indexes. We concatenated documents with the same feed ID (FEEDNO) in the permalink collection into a virtual document and made a blog index with them. We also made a posting index with the permalink collection.

We submitted four runs. The first run (UMassBlog1) used the baseline, i.e. the combination of Global Representation and Pseudo-cluster Selection. This run used titles of the TREC topic 1051-1100 as queries. The second run (UMassBlog2) replaced the ranking function of Global Representation of the baseline with the ranking function based on the topic-independent pseudo-cluster composed of recent postings in each blog as described in Section 3.2. The second run also used the titles of the topics as queries. The third run used the same approach as the first run except that both the titles and the descriptions of the topics were used as queries. The fourth run used the query expansion technique suggested in Section 3.3. The titles of the topics were used as queries for this run. We performed the third run in order to compare our query expansion to manual query expansion because including terms in the description as query terms can simulate an effect of manual query expansion.

Our systems have several parameters. Global Representation has a Dirichlet smoothing parameter for the language models, i.e. μ. Pseudo Cluster Selection has two parameters, i.e. K and μ. Query expansion has two more parameters, i.e. the number of the expanded terms and the number of documents used to estimate the relevance model. To learn the parameters, we used the relevance judgments of the TREC 2007 blog distillation task. The evaluation measure for the training was MAP (mean average precision).

5 Results

The results from our runs are given in Table 1. The relevance judgments of the TREC 2008 blog distillation task use 3 grades, i.e. non-relevant (0), relevant (1) and highly relevant (2). Accordingly, we present scores of MAP and P@10 (precision at 10) of two cases: the case when considering a document whose grade ≥ 1 relevant and the case when considering a document whose grade ≥ 2 relevant. We also display NDCG (normalized discounted cumulative gain) scores.

[1] http://www.lemurproject.org/indri/

	rel≥ 1		rel≥ 2		
Run	MAP	P@10	MAP	P@10	NDCG
UMassBlog1 (Title)	0.2520	0.3880	0.2561	0.2400	0.4777
UMassBlog2 (Title)	0.2587	0.4080 *	0.2515	0.2400	0.4751
UMassBlog3 (Title + Description)	0.2711 *	0.4240 *	0.2772 *	0.2620 *	0.4969 *
UMassBlog4 (Title)	0.2423 !	0.3900 !	0.2531 !	0.2420 !	0.4629 !

Table 1: Performance of our submitted runs. A * indicates a significant improvement over the baseline (UMassBlog1). A ! indicates a significant degradation with respect to the UMassBlog3 (UMassBlog4 only).

Our new supplementary factor, the topic-independent pseudo-cluster (UMassBlog2) showed comparable effectiveness to the baseline (UMassBlog1). When the relevance level threshold is 1, the score of P@10 of the new factor is better than that of the baseline. It shows that the relevance judgments may be based on the recency of relevant postings in some degree. Not surprisingly, the run using descriptions of topics (UMassBlog3) significantly outperformed the other runs. However, the performance of our query expansion technique (UMassBlog4) is somewhat disappointing. It did not show any improvement over the baseline, and further it was significantly worse than the manual query expansion (UMassBlog3). The question "What are the proper query expansion techniques for our framework?" remains unsolved.

6 Conclusions

We performed follow-up study for our successful framework used in TREC 2007, blog site search using resource selection. We introduced a new supplementary factor to substitute a role of Global Representation. The factor is better suited to real world implementations. Further, the performance of this factor was comparable to or even better than the baseline in our experiment. We also suggested a query expansion approach for our framework. However, it did not demonstrate any improvement on the baseline. We plan to explore more advanced information retrieval techniques including other query expansion techniques for our framework in the future.

7 Acknowledgments

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-0711348. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] W. Bruce Croft and John Lafferty. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, USA, 2003.

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

[3] Victor Lavrenko and W. Bruce 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.

[4] Xiaoyong Liu and W. Bruce 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.

[5] Jangwon Seo and W. Bruce Croft. Blog site search using resource selection. to appear in ACM 17th

Conference on Information and Knowledge Management (CIKM 2008).

[6] Jangwon Seo and W. Bruce Croft. UMass at TREC 2007 blog distillation task. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, 2008.

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