UMass at TREC 2007 Blog Distillation Task

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

Abstract

The focus of the blog distillation task is finding blogs with a principle, recurring interest in a specific topic. For this task, we considered a blog as a collection of postings and used resource selection approaches. Further, we investigated techniques that penalized general blogs and combined resource selection techniques. This combination demonstrated significant improvements over baselines.

1 Introduction

We participated in the blog distillation task in the blog track in TREC 2007. This task was to find relevant blogs for any specific topic. A blog can be considered as a collection composed of its own postings. From this point of view, our intuition was that finding relevant blogs is similar to finding relevant collections in a distributed search environment. Therefore, we employed resource selection techniques normally used for distributed information retrieval. Further, in order to take advantage of the topical characteristics of blogs, we suggested a supplementary factor to penalize general blogs.

2 Data Processing

Although this task is often referred as the feed distillation task, directly using the feed collection might not be effective. RSS, the subscription method that is currently most prevalent, does not have any requirement that feeds have to have a summary of the corresponding posting content, contrary to ATOM. In many cases, a RSS feed has no other meaningful information than the posted date and the title. Moreover, the title in the feed is sometimes not the title of the corresponding posting but the title of the blog. In this case, we cannot infer what the blog or postings are about from the feed. Therefore, we decided not to use the feed collection for now. Of course, exploiting feeds is still possible to augment retrieval performance later.

Consequently, our target collection was the permalink collection. This collection contains a considerable amount of splogs and non-english blogs which are intentionally added. We did not get rid of this "noise" from the permalink collection because we expected that our techniques should be robust enough to deal with it. Instead, we removed only HTML tags in each posting. We used only blogs which have valid blog IDs (`BLOGHPNO`). Further, we used the blog IDs instead of feed IDs (`FEEDNO`) as keys to identify what blog each posting belongs to because a blog might have more than one feed link.

3 Resource Selection Techniques

Given that a blog is a collection of postings, finding relevant blogs is similar to resource selection for finding relevant collections. In that sense, using resource selection approaches for this task looks desirable. A variety of resource selection techniques have been explored in distributed information retrieval. Here, we introduce three techniques for blog distillation task. Note that all described techniques are based on language modeling techniques [2].

3.1 Baseline: Global Representation

We can handle a blog as a document. That is, all postings are concatenated into a virtual document. After that, we can build a language model from each virtual document. This is one of the simplest and most widely used methods [1, 4].

A ranking function for Global Representation is the same as query likelihood:

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

where Q and D_{c_i} are a query and a virtual document with a blog id c_i, respectively.

However, this technique has a critical weakness. If a posting is longer than the others, then this model can be biased by the long posting regardless of its relevance.

We refer to this method as "Global Representation" and use it as a baseline.

3.2 Pseudo Cluster Selection

Clustering is an effective approach for distributed information retrieval [5]. That is why a topic-based document set from collections which are not generally topic-centric can be gathered by clustering. But, although there are quite a few exceptions, a blog typically addresses a small number of topics. By exploiting this property, we propose how to construct a pseudo clustering without actually clustering documents. We can get a ranked list by searching a posting collection index with a topic. We assume that the highly ranked documents from the same blog address similar topics. That is, we can consider a set of these documents as a pseudo-cluster. To retrieve such pseudo-clusters, we use a new cluster representation method introduced by Liu and Croft [3]. They showed that a cluster representation using a geometric mean of language models can be a good alternative to other representations. Our ranking function is formed as follows.

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

This method uses a fixed parameter K independent of clusters. However, it is possible that each blog does not have enough documents equal to or greater than K from the blog in the top N ranked list. To compensate for the original method, we estimate the upper bound of the geometric mean using the minimum query likelihood score in the list as follows.

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

With this value, we can compensate our ranking function as follows.

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

We call this method "Pseudo Cluster Selection".

3.3 Combination of Global Representation and Pseudo Cluster Selection

In blog search, we cannot say that blogs which have more relevant postings are necessarily more relevant than other blogs which do not. For example, blog A is daily updated by adding one or two new postings. About 60% of them are relevant. We may decide that the blog is relevant. On the other hand, blog B is hourly updated by adding hundreds of postings. About 5% of them are relevant. It is likely that blog B has more relevant documents than blog A. But, is blog B more relevant? To simplify this case, let's recall the goal of the blog (feed) distillation task. The task can be interpreted as finding blogs where we can get relevant information when we subscribe to feeds from the blog. In this sense, is blog B still relevant? If we subscribe blog B, then a great number of feeds will be delivered. We have to struggle to find a few relevant documents among them. After all, blog B seems irrelevant in this task.

Therefore, we need to penalize blogs which address diverse topics. We use the Pseudo Cluster Selection score for the base score and the Global Representation score for the penalizing factor, and we multiply them as follows.

$$\hat{\phi}(Q, c_i) = \phi_{PCS}(Q, c_i) \cdot \phi_{GR}(Q, c_i)$$

If a blog is not topic-centric, then its virtual document in the global representation has word distributions which are not concentrated on any specific topic keywords but widely scattered. Further, the global representation is originally designed for estimating relevance. Therefore, the global representation can be an appropriate factor to reflect the topic-centric characteristic and relevance of the blog.

4 Experiments

We used the Indri[1] search engine as our foundation for experiments. We built two indexes. We concatenated documents with the same blog ID (`BLOGHPNO`) in the permalink collection into a virtual document and made a new collection with them. The first index was built on this collection for Global Representation. The second index is for Pseudo Cluster Selection and was built on the permalink collection. We used the Krovetz stemmer and the standard stopwords for all indexes. To implement the above mentioned techniques, we post-processed the initial search results from Indri and converted Blog IDs (`BLOGHPNO`) in the search result to Feed IDs (`FEEDNO`).

We performed four runs. For three of these runs, we used only titles of topics as queries. The three runs were for Global Representation, Pseudo Cluster Selection and the combination thereof, respectively. For the fourth run, we used both titles and descriptions as queries. We applied the combination of Global Representation and Pseudo Cluster Selection to this run.

Our systems have some 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 mu. To learn the parameters, we used relevance judgments which were made by ourselves. The relevance judgment contains about 2500 judgments for 50 topics. The topics were chosen from topics of the TREC 2003 web distillation task and the TREC 2004 web distillation task. We performed 10-fold cross validation by randomly partitioning the training data. The evaluation measure for the training was the mean average precision (MAP). Finally, we used the mean of learned parameters for each partition.

Run	MAP	P@10
UMaTiGR	0.2381	0.4822
UMaTiPCS	0.2169	0.4644
UMaTiPCSwGR	0.2529	0.5111
UMaTDPCSwGR	0.2741	0.5356

Table 1: Summary of the submitted runs. UMaTiGR, UMaTiPCS and UMaTiPCSwGR are referred to the title-only runs using Global Representation, Pseudo Cluster Selection and the combination thereof, respectively. UMaTDPCSwGR is referred to the run by the combination of Pseudo Cluster Selection and Global Representation with the titles and the description. We compared the results by using mean average precision (MAP) and precision at 10 (P@10). We performed the paired t-tests with p-value < 0.05. The differences between all pairs are statistically significant except the difference between UMaTiGR and UMaTiPCS for P@10.

5 Results

The results from our runs are given in Table 1. For title-only runs, the combination of Global Representation and Pseudo Cluster Selection (UMaTiPCSwGR) outperformed others as we expected. It is somewhat surprising that the simple and naive Global Representation (UMaTiGR) shows the better performance than does Pseudo Cluster Representation (UMaTiPCS). The run using the titles and the descriptions (UMaTDPCSwGR) achieved the best performance of our runs. This shows that descriptions are helpful.

6 Post-submission Experiment

We did another experiment using the collection preprocessed in a different way as a post-submission experiment. We used a feed ID (`FEEDNO`) instead of a blog ID (`BLOGHPNO`) as a key to identify which blog each posting belongs to because we found that some

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

Run	MAP	P@10
UMaTiGR	0.3454	0.4889
UMaTiPCS	0.3155	0.4600
UMaTiPCSwGR	0.3725	0.5356
UMaTDPCSwGR	0.4051	0.5733

Table 2: Summary of the post submission runs. UMaTiGR, UMaTiPCS and UMaTiPCSwGR are referred to the title-only runs using Global Representation, Pseudo Cluster Selection and the combination thereof, respectively. UMaTDPCSwGR is referred to the run by the combination of Pseudo Cluster Selection and Global Representation with the titles and the description. We compared the results by using mean average precision (MAP) and precision at 10 (P@10). We performed the paired t-tests with p-value < 0.05. The differences between all pairs are statistically significant except the difference between UMaTiGR and UMaTiPCS for P@10.

relevant documents in the relevance judgment set for TREC 2007 Blog Distillation Task do not have valid blog IDs. Accordingly, the collection for Global Representation was newly created by concatenating postings with the same feed ID. Table 2 shows the result.

The post-submission runs show better performance than our submitted runs using the only documents with valid blog IDs. It shows that runs submitted by many other groups, which contributed to the relevance judgment pool, contain many documents with invalid Blog IDs. Nevertheless, the post-submission result shows the same aspect as the submitted result. Global Representation (UMaTiGR) is better than Pseudo Cluster Representation (UMaTiPCS) and the combination of Global Representation and Pseudo Cluster Selection (UMaTiPCSwGR / UMaTDPCSwGR) is still the most effective method.

7 Conclusions

We applied resource selection techniques to this task and showed that they work well. Further, we showed that the effectiveness can be increased by using an advanced technique, which combines the features in order to penalize diverse blogs. Therefore, we conclude that resource selection techniques can be a good approach to this task; accordingly, we plan to explore more advanced resource selection techniques later.

8 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NHN Corp. 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] Jamie Callan. Distributed information retrieval. In W. Bruce Croft, editor, *Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, Norwell, MA, USA, 2000.

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

[3] Xiaoyong Liu and W. Bruce Croft. Evaluating text representations for retrieving the best group of documents. CIIR Technical Report. University of Massachusetts.

[4] Jinxi Xu and Jamie 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.

[5] 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.