Generalizing Diversity Detection in Blog Feed Retrieval

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

ABSTRACT

The goal of a blog retrieval system is to retrieve and rank blogs, as collections of documents, in response to a given query. Previous studies have shown that diversity among the top retrieved posts from a blog is a positive feature for indicating relevance of the blog to the query. However, existing methods capture the diversity of a blog using post-level properties that limits their application to a specific category of retrieval methods. In this paper, we propose a blog-level diversity measure where there is no assumption made about the underlying blog-ranking technique. The proposed measure enables us to integrate diversity in any existing blog retrieval method. Our experimental results show that the proposed method, while being more general, produces comparable results to the post-level diversity detection methods.

Categories and Subject Descriptors

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

Keywords

Blog Retrieval, Blog Distillation, Diversity, Novelty

1. INTRODUCTION

User generated data is becoming one of the main sources of information on the web. The size and diversity of this kind of data on the web requires new search applications specially tailored to the properties of this kind of data. Blogs are one of the main categories of user generated data. Different search applications on blog data have attracted researchers' attention lately including post retrieval, opinion detection and blog feed retrieval.

This paper investigates blog feed retrieval (also known as blog distillation) that is concerned with ranking blogs according to their recurring central theme relative to the topic of a user's query. In other words, the goal is to find the most relevant blogs for each topic that a user can add to his reader and refer to in the future [6].

There are properties that differentiate blog distillation from other retrieval tasks. The main property of blog distillation is that the unit of retrieval is a blog, as a collection of documents, as opposed to a single document in other retrieval tasks. Exploiting this property is the focus of the most existing blog retrieval methods. Most of the methods emphasize the number of topic-related posts that a blog publishes when estimating the relevance of the blog. However, the amount of new information that each of these posts add to the blog is an influential factor. In other words, existing methods assume that posts (documents) in each blog (collection) are independent from each other and each of the posts independently contributes to the final relevance score of the blog.

Posts in a blog are not, however, independent and if the blog publishes repetitive information with low novelty, it will be less interesting for the user to follow. In our diversity-based methods, we leverage this property and penalize blogs with low diversity among their top retrieved posts.

A recent study by Keikha *et al.* shows that taking dependency of blog posts into account and penalizing the blogs with low diversity can improve the performance of blog retrieval systems [4]. They propose two types of diversity, namely topical and temporal, and show that blogs with topically diverse posts that are published over diversified time intervals are more likely to be relevant to the queries. While this method integrates diversity of blog posts into some blog ranking techniques, it is not applicable to all existing retrieval methods. In this paper, we explain the limitation of this previous diversity-based method and propose a more general method that removes that limitation and is applicable to a broader set of blog retrieval methods.

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 blogs relevant to a specific topic. They treat each blog as one long document created by concatenating all of its posts [2, 3, 8]. 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 evaluations.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2507855 .

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 [9]. Based on the similarity between blog distillation and expert search, some researchers have adapted expert retrieval methods for blog retrieval [1, 5]. In these models, each post in a blog is counted as evidence of the blog's "interest" in the query topic. Balog *et al.* adapt two language modeling approaches for 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 for the blog [5].

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. 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 [3]. They model each blog as a collection of posts and use a language modeling approach to select the best collection. Seo *et al* proposed a similar approach called Pseudo Cluster Selection [8].

Most of the existing methods aggregate post-level scores in order to estimate the final relevance score of the blogs. For example, Small Document (SD) model is one of the strongest existing methods which uses the summation of post level relevance scores for ranking relevant blogs:

$$\begin{aligned} P_{SD}(b_i|q) &\stackrel{rank}{=} P(b_i)P(q|b_i) \\ &= P(b_i) \sum_{j=1}^{|b_i \cap R(q)|} P(q|p_{ji})P(p_{ji}|b_i) \\ &\propto \frac{log(N_{b_i})}{N_{b_i}} \sum_{j=1}^{|b_i \cap R(q)|} P(q|p_{ji}) \end{aligned} \quad (1)$$

where q is the query, b_i is a blog, N_{b_i} is the number of posts in the blog, p_{ji} is a post in blog b_i and $R(q)$ is the set of the initially retrieved posts in response to the query. In the last line of above formula we set the blog prior to be logarithmically proportional to its size. $P(q|p_{ji})$ is the post-level query likelihood and $P(p_{ji}|b_i)$ is the similarity of a post to the blog as a whole and is set to be uniform among all blog posts. As we can see, SD model is an aggregation (summation) of the post-level query likelihood scores. Other methods like data fusion models [5] or Pseudo Cluster Selection [8] are also post-level score aggregations.

There are other methods that do not use post-level score aggregations. For example, Blogger Model assumes a different generative process and directly estimates term probabilities for each blog without calculating post-level query likelihood [1]:

$$\begin{aligned} p(q|b_i) &= \prod_{t \in q} p(t|b_i) \\ &= \prod_{t \in q} \sum_{j=1}^{|b_i \cap R(q)|} P(t|p_{ji})P(p_{ji}|b_i) \\ &= \frac{1}{N_{b_i}} \prod_{t \in q} \sum_{j=1}^{|b_i \cap R(q)|} P(t|p_{ji}) \end{aligned} \quad (2)$$

where $p(t|b_i)$ is estimated based on the term probabilities in the associated posts and $P(p_{ji}|b_i)$ is assumed to be uniform.

Keikha *et al.* proposed methods to capture diversity among topic-related posts in a blog and integrate it in the existing blog retrieval methods. They penalize the score of the blog posts based on their similarity to other posts in the blog:

$$\begin{aligned} &score_{diversity}(p_{ji}, q) = \\ &score_{ini}(p_{ji}, q)(1 - \lambda \; max_{p_{ki} \in S} \; sim(p_{ji}, p_{ki})) \end{aligned} \quad (3)$$

where $score_{ini}(p_{ji}, q)$ is the initial score of post p_{ji} for the query q, S is a subset of posts that belong to the same blog and have higher scores than p_{ji} and λ is the parameter that controls the amount of penalty. They use the new scores of the posts in different aggregation-based blog retrieval methods and show that considering diversity can improve the retrieval performance. Since this approach modifies the existing post scores, it is not applicable to non-aggregation methods such as the Blogger Model. In the next section, we modify this approach and define diversity as a blog-level measure that could be integrated into any blog retrieval method.

3. DIVERSITY AS AN INDEPENDENT COMPONENT

As was mentioned in the previous section, some of blog retrieval methods, e.g, Blogger Model, do not employ post-level scores aggregation and thus the existing diversity-detection measure cannot be applied to them. Therefore we need a more general method for integrating diversity into those blog retrieval methods.

One option is to penalize the blog score as a whole based on the repetitive information that it contains. To this end, we penalize blog relevance scores based on the average similarity of their posts. The new score of a blog is calculated by:

$$Score_{diversity}(b_i, q) = score_{ini}(b_i, q)(1 - \gamma AvgSim(b_i, q)) \quad (4)$$

where γ is a control parameter and average similarity is calculated between blog posts that are initially retrieved in response to the query:

$$AvgSim(b_i, q) = \frac{\sum_{j=1}^{|b_i \cap R(q)|} \sum_{l=j+1}^{|b_i \cap R(q)|} sim(p_{ji}, p_{li})}{\binom{|b_i \cap R(q)|}{2}} \quad (5)$$

The higher the *AvgSim* value, the less diverse is the blog feed with respect to the query. Similar to previous work by Keikha *et al.*, we assume similarity values are in [0,1] and the similarity function can be based on topical information, temporal information or any other type of information.

Model	MAP		P@10		Bpref	
SD	0.2867		0.4444		0.3439	
SD-PostLevel	0.3191	↑	0.5156	↑	0.3622	↑
SD-BlogLevel	0.3233	↑	0.4933	↑	0.3693	↑
BloggerModel	0.3441		0.5333		0.3903	
BloggerModel-BlogLevel	**0.3581**	↑	**0.5689**	↑	**0.4001**	↑

Table 1: Evaluation results for blog-level penalty over TREC'07 data set.

Model	MAP		P@10		Bpref	
SD	0.2636		0.3821		0.2858	
SD-PostLevel	**0.2961**	↑	0.4308	↑	**0.3125**	↑
SD-BlogLevel	0.2909	↑	**0.4385**	↑	0.3066	
BloggerModel	0.2764		0.4154		0.2899	
BloggerModel-BlogLevel	0.2896	↑	0.4205		0.3003	↑

Table 2: Evaluation results for blog-level penalty over TREC'09 data set.

As opposed to previous methods, our proposed method does not explicitly model the novelty of each post. However, it gives us a measure for evaluating the average novelty of posts in the blog. This method can be seen similar to the application of portfolio theory in information retrieval where the goal is to model the risk of retrieving a ranked list of documents [10]. Analogous to those risk-aware ranking methods, we can assume that the initial score of a blog is the expected gain of retrieving that blog and the $AvgSim$ value is the risk of retrieving a non-informative blog. Thus the method tries to retrieve blogs with higher gains and lower risks.

4. EXPERIMENTS

We conduct our experiments over TREC'07 and TREC'09 blog track data sets from the blog distillation task (Similar results were obtained for TREC'08 and TREC'10 data sets, but for the sake of clarity and lack of space we do not report them here. However TREC'08 and TREC'10 data sets are employed for parameter tuning). 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 more recent 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 [6], 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 [7]. 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[1] to index the collection with the default stemming and stop-words removal. In all the methods, the language modeling approach using Dirichlet smoothing has been used to score the posts and retrieve the top posts for each query. Without further tuning, the Dirichlet smoothing parameter is set to 5000 [11].

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

Since TREC'07 and TREC'08 query sets share the same collection, we use TREC'08 data for tuning the parameters for the TREC'07 data set. We do the same for the TREC'09 and TREC'10 query sets that share the Blog08 collection. These parameters include λ in formula 3 and γ in formula 4. We fix the number of the initially retrieved posts for each query to 15000 that is shown to be the best choice in previous experiments [4].

We get our initial results using Small Document Model and Blogger Model as discussed in section 2 and used them as our baseline methods. In our diversity detection method, the initial blog scores are penalized as was described in formula 4. We select Small Document Model and Blogger Model as examples of aggregation-based techniques and non-aggregation-based techniques respectively. In order to have a comprehensive comparison, we also apply post-level penalty on the initially retrieved set of posts and calculate the new scores using SD model as proposed in previous work [4].

As the similarity between two posts in formula 5, we use a hybrid similarity measure as defined in [4]. Hybrid similarity considers both temporal and topical similarity between two posts and is shown to be effective in previous diversity-based blog retrieval experiments.

Tables 1 and 2 shows the performance evaluation of the proposed methods on the TREC'07 and TREC'09 data sets respectively. In these tables, SD and Blogger models are our baseline techniques. SD-PostLevel is the previously proposed method that applies a post-level penalty to the initial post scores. SD-BlogLevel and BloggerModel-BlogLevel are our proposed methods that apply blog-level penalty to the baseline scores.

Statistical significance tests are performed using the paired T-test at 0.05 level of significance. The symbol ↑ indicates that a diversity-based method has a statistically significant improvement over its non-diversified version. The bold values in each column indicate the best performance for the corresponding evaluation measure.

As we can see the blog-level penalty can improve baseline methods in both data sets and in all the evaluation measures. In most of the cases, the improvements are at a statistically significant level. In the case of SD model, we can

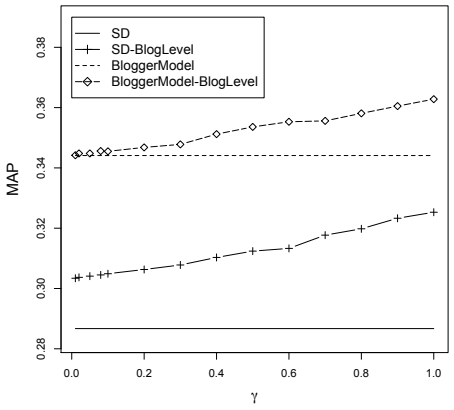

Figure 1: Effect of the diversity parameter γ in the performance over TREC'07 data set.

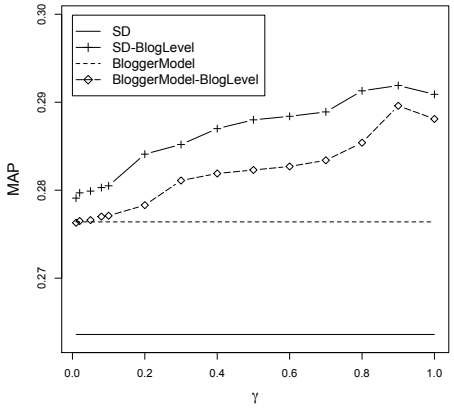

Figure 2: Effect of the diversity parameter γ in the performance over TREC'09 data set

see that blog-level penalty has comparable results to post-level penalty. This shows the value of our proposed method that is more general without losing any performance.

As we mentioned before, the γ parameter is tuned using training data. Since this is an important parameter of the method, we show the sensitivity of the system to this parameter in figures 1 and 2. We can see that the proposed method can improve the baseline techniques for almost all values of γ while the best results are obtained by γ values close to one.

5. CONCLUSION AND FUTURE WORK

In this paper, we studied the effect of diversity among blog posts in blog relevance. We proposed a new approach for integrating diversity into existing blog retrieval methods that is more general than previously proposed methods but produces comparable results. Our proposed method defines diversity as a blog-level feature that is captured using average similarity between blog posts. This is in contrast with previous methods that define diversity as a post-level feature that is captured by maximum similarity of the post to other posts in the blog. Our experimental results over standard blog retrieval collections and state-of-the-art baseline methods showed that our methods can result in statistically significant improvements over different baselines and different test collections.

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. Also it is interesting to employ more advanced diversification techniques and study their effect on the system.

6. ACKNOWLEDGMENTS

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

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

[3] 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.

[4] M. Keikha, F. Crestani, and W. B. Croft. Diversity in blog feed retrieval. In *Proceeding of CIKM'12*, pages 525–534, 2012.

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

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

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

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

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

[10] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of SIGIR'09*, pages 115–122, 2009.

[11] 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.