Bloggers as Experts

Feed Distillation using Expert Retrieval Models

Krisztian Balog kbalog@science.uva.nl
Maarten de Rijke mdr@science.uva.nl
Wouter Weerkamp weerkamp@science.uva.nl

ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam

ABSTRACT

We address the task of (blog) feed distillation: to find blogs that are principally devoted to a given topic. The task may be viewed as an association finding task, between topics and bloggers. Under this view, it resembles the expert finding task, for which a range of models have been proposed. We adopt two language modeling-based approaches to expert finding, and determine their effectiveness as feed distillation strategies. The two models capture the idea that a human will often search for key blogs by spotting highly relevant posts (the Posting model) or by taking global aspects of the blog into account (the Blogger model). Results show the Blogger model outperforms the Posting model and delivers state-of-the art performance, out-of-the-box.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [**Information Systems Applications**]: H.4.m Miscellaneous

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Feed distillation, people-topic associations, language modeling

1. INTRODUCTION

Information needs in the blogosphere come in many flavors. The task on which we focus is *blog distillation*, to identify key blogs with a recurring interest in the topic, that provide credible information about the topic. The blog distillation task is an interesting one, since it addresses a real information need, shared by professional and non-professional searchers of the blogosphere. Given this task, how, then, should we model it? From a modeling point of view, blog distillation bears a strong resemblance to tasks considered in the area of expertise retrieval [2, 13]. In expert finding, systems return a ranked list of names of people that are knowledgeable about a given topic. Most approaches to this task view it as an *association finding* task, and rank people by the degree to which they are associated with the topic, as determined by examining the documents in which the two—people and the topic—co-occur. Can these expert finding ideas be used to address the blog distillation task?

Below, we report on an experiment in which we apply two state-of-the-art expert finding models, both based on generative language modeling, to the feed distillation task. The first model, called Blogger model, explicitly models bloggers and examines the themes they are interested in. The second model, called Posting model, identifies key posts on a given topic and then determines the bloggers from whose blogs these originate. For expert finding a counterpart of the Posting model has been found to outperform (the expert finding analogue) of the Blogger model. We find that for feed distillation the situation is the other way around: the Blogger model outperforms the Posting model, suggesting that the two tasks (expert finding and feed distillation) are essentially different and best addressed with different approaches.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. RELATED WORK

Responding to the emerging interest in the blogosphere, TREC launched a blog track in 2006 [9]. The initial focus was on finding relevant blog *posts*, with a special interest in their opinionatedness, resulting in many insights in blog post retrieval (see, e.g., [7–9]). The task of finding relevant *blogs* was considered at the 2007 edition of the track [7]. Specifically, the aim of the feed distillation task is to rank blogs (not individual posts) given a topic. TREC 2007 witnessed a broad range of approaches to this new task. Approaches differed in the indexing units they considered: either individual posts [3, 4, 12], or full blogs (i.e., concatenated posts) [3, 12]. The best performing TREC run uses a blog index (and expands queries using Wikipedia) [3]. Several approaches used a combination of post and blog level evidence [11, 12]. Results are mixed with the combination performing worse than a blog run in [11], but better than either blog or post approaches in [12].

As to expert finding, the task has been formulated in terms of people-topic associations, for which two main language modeling based approaches have been proposed [1]. Balog et al.'s first model directly models the knowledge of an expert from associated documents (and is analogous with our Blogger model), while their second model first locates documents on the topic and then finds the associated experts (and corresponds to our Posting model). Most systems that took part in the 2005 and 2006 editions of the Expert Finding task at TREC implemented (variations on) one of these two models; see [2, 13]. Macdonald and Ounis [6] propose a different approach for ranking people based on data fusion techniques, without using collection-specific heuristics. And Petkova and Croft [10] propose yet another approach, based on a combination of the two models just described, while explicitly modeling topics.

3. MODELING FEED DISTILLATION

In modeling feed distillation we consider the Blogger model and the Posting model. Since blogs are much longer than queries, we obtain a more accurate estimate by invoking Bayes' Theorem and estimating $p(blog|q) = (p(q|blog)p(blog))/p(q)$. For the purpose

Model	MAP	R-prec	bpref	P5	P10	MRR
Blogger	.3272	.4023	.3192	.4844	.4844	.6892
Posting	.2325	.3360	.2751	.3822	.3733	.4850

Table 1: Blogger vs Posting model for feed distillation.

of ranking blogs, we can drop $p(q)$. Our task, then, is to estimate $p(q|blog)$ (see Sections 3.1–3.2) and $p(blog)$ (see Section 3.3).

3.1 Blogger Model

The Blogger model estimates the probability of a query given a blogger (or blog) by representing the blog as a multinomial probability distribution over the vocabulary terms:

$$p(q|\theta_{blog}) = \prod_{t \in q} p(t|\theta_{blog})^{n(t,q)}. \quad (1)$$

Next, we smooth the probability of a term given a blog with the background probabilities:

$$p(t|\theta_{blog}) = (1 - \lambda_{blog}) \cdot p(t|blog) + \lambda_{blog} \cdot p(t). \quad (2)$$

Finally, we estimate $p(t|blog)$ as follows:

$$p(t|blog) = \sum_{post \in blog} p(t|post, blog) \cdot p(post|blog). \quad (3)$$

We assume that the post and the blog are conditionally independent, thus $p(t|post, blog) = p(t|post)$, and approximate $p(t|post)$ with the standard maximum likelihood estimate; for the conditional probability $p(post|blog)$, see below.

3.2 Posting Model

In the Posting model individual posts are queried and then the blogs to which these posts belong are considered:

$$p(q|blog) = \sum_{post \in blog} \prod_{t \in q} p(t|\theta_{post})^{n(t,q)} \cdot p(post|blog), \quad (4)$$

where the probability of a term t given the post is estimated by inferring a post model $p(t|\theta_{post})$ for each post following:

$$p(t|\theta_{post}) = (1 - \lambda_{post}) \cdot p(t|post) + \lambda_{post} \cdot p(t). \quad (5)$$

3.3 Blog and Post Importance

Our Blogger and Posting models both offer the possibility of expressing the prior importance of a blog (i.e., $p(blog)$) and the importance of a post given a blog (i.e., $p(post|blog)$). In our experiments we assume both probabilities to be uniform. In other words, all posts within a blog are equally important, as are all blogs.

4. EXPERIMENTAL EVALUATION

We set up experiments to answer our main question: can we successfully apply out-of-the-box expert retrieval models to blog feed distillation? And if so, which of the two models, Blogger or Posting, displays the best performance on this task?

As our test collection we use the TRECBlog06 corpus [5]; we index only the HTML permalinks of the posts and ignore other collection contents like syndicated content and home pages. The TREC 2007 Blog track offers 45 feed distillation topics and assessments [7]. We only use the topic field of the topics.

For the smoothing parameter λ_x in Eq. 2 and 5 (with $x \in \{blog, post\}$), we set λ_x equal to $n(x)/(\beta + n(x))$, where $n(x)$ is the length of the blog (i.e., summarizing the length of all posts of the blog) or the post. We set β to be the average number of terms in the document: $\beta = 17,400$ for blogs and $\beta = 515$ for posts.

The results of our experiments are listed in Table 1. We see that the Blogger model significantly[1] outperforms the Posting model on all metrics. A few comments are in order. First, the results of the Blogger model would be ranked second in TREC 2007 (best run: MAP .3695, second best: MAP .2923). Second, while the Blogger model outperforms the Posting model for the feed distillation task, for the expert finding task, the relative ranking is the other way around [1]. What does that tell us about the difference between the two tasks? For expert finding, for a candidate expert to be ranked highly for a given topic it suffices for him or her to be one of (relatively) few people mentioned in the context of the topic; it is not important whether the candidate expert wrote a lot about the topic or whether he or she is also associated with other topics. In contrast, for feed distillation, it appears we need to identify people that write mainly about the topic at hand. Hence, it makes sense that we explicitly model individual bloggers (as in the Blogger model) and take a close look at the main themes that occupy them individually.

[1]Significance is tested using a two-tailed paired t-test with $\alpha = .01$

5. CONCLUSION

We applied two state-of-the-art expert finding models to the feed distillation task, and arrived at two main findings. First, out-of-the-box expert finding methods can achieve competitive scores on the feed distillation task. Second, there is a qualitative difference between the expert finding and feed distillation tasks, as a result of which an effective strategy for identifying key bloggers is to explicitly model them and the main themes that occupy them.

6. ACKNOWLEDGEMENTS

Balog and De Rijke were supported by the Netherlands Organisation for Scientific Research (NWO) under project number 220-80-001. Weerkamp and De Rijke were supported by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104. De Rijke was also supported by NWO under project numbers 017.001.190, 640.001.501, 640.002.501, STE-07-012.

7. REFERENCES

[1] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proc. SIGIR'06*, pages 43–50, New York, NY, USA, 2006. ACM Press.
[2] N. Craswell, A. de Vries, and I. Soboroff. Overview of the TREC-2005 Enterprise Track. In *Proc. TREC 2005*, 2006.
[3] J. Elsas, J. Arguello, J. Callan, and J. Carbonell. Retrieval and feedback models for blog distillation. In *TREC 2007 Working Notes*, 2007.
[4] B. J. Ernsting, W. Weerkamp, and M. de Rijke. The University of Amsterdam at the TREC 2007 Blog Track. In *TREC 2007 Working Notes*, 2007.
[5] C. Macdonald and I. Ounis. The TREC Blogs06 collection: Creating and analyzing a blog test collection. Technical Report TR-2006-224, Department of Computer Science, University of Glasgow, 2006.
[6] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *Proc. CIKM '06*, pages 387–396, 2006.
[7] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 Blog Track. In *TREC 2007 Working Notes*, pages 31–43, 2007.
[8] G. Mishne. *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, 2007.
[9] I. Ounis, C. Macdonald, M. de Rijke, G. Mishne, and I. Soboroff. Overview of the TREC 2006 Blog Track. In *Proc. TREC 2006*. NIST, 2007.
[10] D. Petkova and W. B. Croft. Hierarchical language models for expert finding in enterprise corpora. In *Proc. ICTAI 2006*, pages 599–608, 2006.
[11] K. Seki, Y. Kino, and S. Sato. TREC 2007 Blog Track Experiments at Kobe University. In *TREC 2007 Working Notes*, 2007.
[12] J. Seo and W. B. Croft. UMass at TREC 2007 Blog Distillation Task. In *TREC 2007 Working Notes*, 2007.
[13] I. Soboroff, A. de Vries, and N. Craswell. Overview of the TREC 2006 Enterprise Track. In *Proc. TREC 2006*, 2007.