A Two-Stage Model for Blog Feed Search

Wouter Weerkamp
w.weerkamp@uva.nl

Krisztian Balog
k.balog@uva.nl

Maarten de Rijke
derijke@uva.nl

ISLA, University of Amsterdam, Science Park 107
1098 XG Amsterdam

ABSTRACT

We consider blog feed search: identifying relevant blogs for a given topic. An individual's search behavior often involves a combination of exploratory behavior triggered by salient features of the information objects being examined plus goal-directed in-depth information seeking behavior. We present a two-stage blog feed search model that directly builds on this insight. We first rank blog posts for a given topic, and use their parent blogs as selection of blogs that we rank using a blog-based model.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms: Algorithms, Measurement, Performance, Experimentation

Keywords: Blog feed search, two-stage model

1. INTRODUCTION

We focus on blogs: the unedited, unregulated voice of an individual [5], as published on a web page containing time-stamped entries. The blogosphere has shown a huge increase in volume in recent years, and is now a major source of information online. To allow for end users to follow a blog that regularly covers a given topic, we can provide them with a ranking of blogs that are likely to show a *recurring interest* in the topic. This task of identifying topically relevant blogs is referred to as *blog feed search*. Even though the unit of retrieval is blogs, the indexing unit should be blog posts, as this allows for easy incremental indexing, and the use of a single index for both blog feed search and blog post retrieval. Indeed, all current approaches to this task use a post index [1, 2, 4, 6].

We propose a two-stage model to blog feed search. The model exploits the following observation about human strategies for identifying complex information objects such as blogs (or people, for that matter). Prior to in-depth examination of complex information objects, humans display exploratory search behavior triggered by salient features of such objects [3]. This insight gives rise to the following two-stage model for blog feed search. In stage 1, we take individual utterances (i.e., *posts*) to play the role of "attention triggers" and select an initial sample of blogs based on the most interesting posts given the query, using a post-based approach. Here, we define "interesting" as topically relevant, but more elaborate techniques can also be applied (e.g., credibility, novelty, etc. [8]). Then, in stage 2, we only consider these most interesting blogs, which we then examine more in-depth by considering all their posts to determine the likelihood of the topic being a central theme of the blog, using a blog-based approach.

We hypothesize that by pruning the list of posts taken into account in stage 1, and thus focusing on the most interesting utterances, we can achieve improvements over a blog-based model baseline on effectiveness and, as a result of blog selection, also on efficiency. Furthermore, we expect to see a considerable improvement on precision when applying pruning, and more so, when using a lean, title-only document representation in stage 1.

2. RELATED WORK

Previous work in blog feed search centered around two families of approaches: blog-based models and post-based models. Blog-based models use posts to construct a model for each blog. Ranking is done by matching the query against these blog models. Examples of blog-based models include the Blogger model [1] and the Long Document model [2]. The other type of approaches, post-based models, start from a ranking of blog posts and scores of individual posts are aggregated in order to infer a ranking of blogs [1, 2, 4]; these models boil down to estimating the relevance of blog posts and associating these posts to their parent blogs, using some weighted association value. Various ways of aggregating scores [4] and association values [2, 9] have previously been discussed. Both families of approaches have problems: post-based models ignore the *recurring interest* requirement; just a few relevant posts in a blog can be enough for it to be ranked high, even though other posts are non-relevant. On the other hand, blog-based models can become quite inefficient when applied to large numbers of frequently updated blogs, where each consists of several posts.

3. A TWO-STAGE MODEL

We are working in a language modeling setting and rank documents (blogs, posts) based on their likelihood of being generated from the query. We use Bayes' Theorem to rewrite this probability: $P(D|Q) = P(D)P(Q|D)/P(Q)$. As $P(Q)$ does not influence the ranking, we drop this term; we assume $P(D)$, i.e., the a priori belief of a document D being relevant, to be uniform. This term can therefore be ignored. We are left with $P(D|Q) \propto P(Q|D)$. In our specific case we rank blogs, and try to estimate the probability $P(Q|blog)$.

In stage 1, we are looking for salient utterances on the topic in blog posts. Therefore, we rank posts for a given query using:

$$P(Q|\theta_{post}) = \prod_{t \in Q} P(t|\theta_{post})^{n(t,Q)}, \quad (1)$$

where $P(t|\theta_{post}) = (1-\lambda_{post})P(t|post)+\lambda P(t)$ (i.e., smoothing with the background collection) and $n(t,Q)$ is the number of times

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

term t is present in the query. We use the top N most relevant utterances (posts) to identify the set of possibly interesting blogs:

$$\mathcal{B} = \{blog | \sum_{post \in N} P(Q|\theta_{post}) P(post|blog) > 0\}, \tag{2}$$

where $P(post|blog)$ denotes the importance of a post given a blog, which is assumed to be uniform. Note that the summation part in Eq. 2 corresponds to a post-based model for ranking blogs, however, in our approach it is only used for identifying blogs that deserve to be ranked for the topic.

Having identified the set of possibly interesting blogs, we now estimate the probability of each $blog \in \mathcal{B}$ having generated the query, i.e., displaying a recurring interest in the topic:

$$P(Q|blog) \propto \prod_{t \in Q} P(t|\theta_{blog})^{n(t,Q)}. \tag{3}$$

We represent blogs as a multinomial probability distributions over the vocabulary terms, and infer a blog model θ_{blog}, such that the probability of a term given the blog model is $P(t|\theta_{blog})$. To construct such a representation, we first aggregate all terms from posts of the blog to estimate an empirical model:

$$P(t|blog) = \sum_{posts} P(t|post) \cdot P(post|blog). \tag{4}$$

Then, this probability $P(t|blog)$ is smoothed using the background collection probability $P(t)$, to arrive at $P(t|\theta_{blog})$ (smoothing blogs is done analogously to smoothing applied to posts).

4. EXPERIMENTS AND RESULTS

We test whether our two-stage model is capable of effective blog feed search. We perform two series of experiments. First, we investigate the amount of pruning applied, i.e., the value of N in Eq. 2. We consider three settings: no pruning, topic-independent pruning (train on 2007 topics, test on 2008, and vice versa), and topic-dependent pruning (best empirically found values per topic). The “no pruning” condition corresponds to the blog-based model (Blogger model in [1]). Our second set of experiments concerns the representation of posts for stage 1 of our approach. We compare the results of the two-stage model to a blog-based model only, since blog-based models usually outperform post-based models [1].

The blog feed search task ran at TREC 2007 and 2008 [7] and uses the TRECBlog06 corpus. We use the English blog posts, and ignore blogs that only have 1 post. We have a total of 95 topics and relevance judgments, and we only use the title field of the topics. In our experiments, we optimize for MAP. Testing for significance is done using a two-tailed paired t-test; significant differences are indicated using $^\blacktriangle$ and $^\blacktriangledown$ ($\alpha = 0.01$), and $^\vartriangle$ and $^\triangledown$ ($\alpha = 0.05$).

Table 1 lists the results on the 2007 and 2008 topics for the blog-based model, and the various settings for stage 1 in our two-stage model. Results show that our model is at least as effective as the blog-based model, while being considerably more efficient: The blog-based model examines all 2.5M blog-post associations, while our two-stage model considers just 1% (23,700) in stage 2. Topic independent pruning results in a slight improvement in effectiveness over the blog-based model, while topic dependent pruning results in significant improvements. The use of a lean document representation, with an average document length of just 12 words, results in very good overall precision scores.

5. CONCLUSIONS

We have proposed a two-stage model for blog feed search. The model only tries to rank bloggers that stand out because of salient posts and then determines whether the topic is a central concern. Experiments show that the two-stage model can improve over a blog-based model. Topic dependent pruning of the post list in stage 1 helps, and we can combine this with a lean document representation to improve early precision even further. Future work is aimed at learning the optimal pruning level per topic.

2007 topics		MAP	P@5	MRR
Blog-based model		0.3260	0.5422	0.7193
Two-stage model				
Representation	Pruning			
full content	1,700	0.3348$^\blacktriangle$	0.5422	0.7213
full content	topic-dep.	0.3611$^\blacktriangle$	0.5689$^\vartriangle$	0.7243
title-only	-	0.3549$^\vartriangle$	0.6444$^\blacktriangle$	0.8476$^\blacktriangle$
title-only	7,000	0.3577$^\blacktriangle$	0.6622$^\blacktriangle$	0.8587$^\blacktriangle$
title-only	topic-dep.	**0.3813**$^\blacktriangle$	**0.6889**$^\blacktriangle$	**0.8604**$^\blacktriangle$
2008 topics				
Blog-based model		0.2521	0.4880	0.7447
Two-stage model				
Representation	Pruning			
full content	1,700	0.2551	0.4960	0.7483
full content	topic-dep.	**0.2747**$^\blacktriangle$	**0.5080**	0.7504
title-only	-	0.2363	0.4880	0.7524
title-only	7,000	0.2368	0.4840	0.7524
title-only	topic-dep.	0.2571	**0.5080**	**0.7591**

Table 1: Results of the various instances of the two-stage model compared to the blog-based model. Significance tested against blog-based model.

Acknowledgements This research was supported by the European Union’s ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.061.814, 612.061.815, 640.004.802.

REFERENCES

[1] K. Balog, M. de Rijke, and W. Weerkamp. Bloggers as experts. In *SIGIR 2008*, pages 753–754, 2008.

[2] J. Elsas, J. Arguello, J. Callan, and J. Carbonell. Retrieval and feedback models for blog feed search. In *SIGIR 2008*, 2008.

[3] C. Kuhlthau. *Seeking Meaning: A Process Approach to Library and Information Services*. Libraries Unlimited, 2003.

[4] C. Macdonald and I. Ounis. Key blog distillation: Ranking aggregates. In *CIKM 2008*, pages 1043–1052, 2008.

[5] G. Mishne. *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, 2007.

[6] J. Seo and W. B. Croft. Blog site search using resource selection. In *CIKM 2008*, 2008.

[7] TREC Blog track wiki. http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG, 2010.

[8] W. Weerkamp and M. de Rijke. Credibility Improves Topical Blog Post Retrieval. In *ACL-08: HLT*, pages 923–931, 2008.

[9] W. Weerkamp, K. Balog, and M. de Rijke. Finding key bloggers, one post at a time. In *ECAI 2008*, pages 318–322, 2008.