Finding key bloggers, one post at a time

Weerkamp, W.; Balog, K.; de Rijke, M.

Published in: Frontiers in Artificial Intelligence and Applications

Citation for published version (APA):
Finding Key Bloggers, One Post At A Time

Wouter Weerkamp and Kristian Balog and Maarten de Rijke

Abstract. User generated content in general, and blogs in particu-
lar, form an interesting and relatively little explored domain for min-
ing knowledge. We address the task of blog distillation: to find blogs
that are principally devoted to a given topic, as opposed to blogs that
merely happen to discuss the topic in passing. Working in the setting
of statistical language modeling, we model the task by aggregating a
blogger’s blog posts to collect evidence of relevance to the topic and
persistence of interest in the topic. This approach achieves state-of-
the-art performance. On top of this baseline, we extend our model by
incorporating a number of blog-specific features, concerning docu-
ment structure, social structure, and temporal structure. These blog-
specific features yield further improvements.

1 Introduction

With the growth of the blogosphere comes the need to provide effec-
tive access to the knowledge and experience contained in the many
tens of millions of blogs out there. Information needs in the blogo-
sphere come in many flavors. E.g., Mishne and de Rijke [9] consider
both ad hoc and filtering queries, and argue that blog searches have
different intents than typical web searches, suggesting that the pri-
mary targets of blog searches are tracking references to named enti-
ties and identifying blogs or posts which focus on a certain concept.
The Blogranger system [3] offers several types of search facilities; in
addition to post retrieval facilities, it also offers a blog search engine,
i.e., an engine aimed at identifying blogs about a given topic which a
user can then add to an RSS reader.

The task on which we focus in this paper is the blog distillation
 task: to find blogs that are principally devoted to a given topic. That
is, instead of identifying individual “utterances” (posts) by bloggers,
we want to identify key blogs with a recurring interest in the topic,
that provide credible information about the topic. The blog distil-
lation task is an interesting one, for a number of reasons. First, it
addresses a real information need, shared by professional and non-
professional searchers of the blogosphere. Second, a retrieval model
for this task seems to require multiple types of evidence: “local” evi-
dence derived from (a small number of) blog posts of a given blogger
plus more “global” evidence derived from a blog as a whole.

Successful approaches at the new feed (blog) distillation task at
TREC 2007 Blog track, take the entire blog as indexing unit, that
is, the contents of individual posts belonging to the same blog are
concatenated into a single document: the blog. Even though this ap-
proach performs well in TREC, we want to use individual posts as
indexing unit for three (practical) reasons: (i) to allow for easy in-
cremental indexing, (ii) for presentation of retrieval results, posts are
natural and coherent units, and (iii) the most important reason, to
allow the use of one index for both blog post and blog retrieval.

Given this constraint, how, then, should we model the blog distil-
lation task? In this paper we view blog distillation as an association
finding task, more specifically, as a blogger-topic association finding
task: which blogger is most closely associated with a given topic?
Given our choice of working with posts (as opposed to entire blogs)
as indexing units, we need effective ways of estimating blogger-topic
associations from blog posts. As we will see below, we also need to
be able to incorporate the importance of an individual blog post given
the blog from which it originates.

Given this choice of model, we explore a number of dimensions.
First, how does our (post-based) model perform compared to other
known solutions to blog distillation? Second, and assuming that blog
distillation is a precision-oriented task (like so many search tasks on
the web), can we use the document structure that blogs come with
to favor relatively rare but high quality matches; i.e., if we represent
blog posts using their titles only (as opposed to title-plus-body) do we
observe a strong precision-enhancing effect (perhaps at the expense
of recall)? And what if we combine the title-only representation with
the title-plus-body representations? Third, how can we make use of
blog specific features to improve blog distillation?

Our main finding is that a post-based approach to feed distillation
can be as effective as state-of-the-art blog-based approaches. Addi-
tionally, on top of this competitive baseline we explore blog-specific
features such as document structure, social structure and temporal
structure, and find that (i) the lean title-only content representation
has a clear precision-enhancing effect when compared to a title+body
representation, while a combination of the two representations out-
performs both; (ii) post priors based on comments and post length
have a positive influence on performance; (iii) temporal ordering of
posts can be used as an indication of the importance of a post given a
blog; and (iv) a combination of all blog specific features on top of our
baseline shows significant improvements over the baseline on most
metrics.

The remainder of the paper is organized as follows. In Section 2
we discuss related work. Section 3 details our experimental setup.
In Section 4 we introduce our blog distillation model and assess its
performance. Next, in Section 5 we discuss the implementation of
the blog specific features; this is followed by an experimental eval-
uation and discussion of our findings in Section 6. We conclude in
Section 7.

2 Related Work

Responding to the emerging interest in blogging and in informa-
tion access tasks for the blogosphere, TREC launched a blog track
in 2006 [10]. The first round of this track focused mainly on find-
ing relevant blog posts, with a special interest in their opinionat-
edness. Many insights in blog post retrieval have been gained (see,
e.g., [6, 8, 10]), but the task of finding relevant blogs has received
3 Experimental Setup

Before we introduce our modeling of blog distillation in Section 4, we first describe our test collection, the metrics we use, and the smoothing settings we employed.

3.1 Test Collection

As our test collection we use the TRECBlog06 corpus [5]. This corpus has been constructed by monitoring feeds for a period of 11 weeks and downloading all permalinks. For each permalink (or blog post or document) the feed number is registered. Besides the permalinks (HTML documents) we also have syndicated content to our availability; we only used the HTML documents.

For our experiments we construct two indices: a title-only index (T), and a title-and-body index (TB). The former consists of the \(<title>\) field of the documents; the latter combines this field with the content of the \(<body>\) part of the documents. Table 1 lists the characteristics of both indices.

<table>
<thead>
<tr>
<th>index</th>
<th>size</th>
<th>terms</th>
<th>unique terms</th>
<th>avg. length</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>674MB</td>
<td>17.4M</td>
<td>439,747</td>
<td>5</td>
</tr>
<tr>
<td>TB</td>
<td>16GB</td>
<td>1,656.3M</td>
<td>9,106,161</td>
<td>515</td>
</tr>
</tbody>
</table>

Table 1. Characteristics of T and TB indices

The TREC 2007 Blog track offers 45 feed distillation topics and assessments [6]. Both topic development and assessments are done by the participants. Assessors were asked to check a substantial number of blog posts of a retrieved feed to determine the relevance of the entire feed.

For all our runs we use the topic field (T) of the topics and ignore all other information available (e.g., description (D) or narrative (N)).

3.2 Metrics

We use the following standard metrics to determine the effectiveness of our retrieval methods: mean average precision (MAP), R-precision (R-prec), as well as three precision-oriented measures: precision at ranks 5 and 10 (P@5 and P@10) and mean reciprocal rank (MRR).

4 Modeling Blog Distillation

To tackle the problem of identifying key blogs given a query, we take a probabilistic approach and formulate the task as follows: what is the probability of a blog (feed) being a key source given the query topic \(q\)? That is, we determine \(p(\text{blog}|q)\), and rank blogs according to this probability. Since the query is likely to consist of very few terms to describe the underlying information need, a more accurate estimate can be obtained by applying Bayes’ Theorem, and estimating:

\[
p(\text{blog}|q) = \frac{p(q|\text{blog}) \cdot p(\text{blog})}{p(q)},
\]

where \(p(\text{blog})\) is the probability of a blog and \(p(q)\) is the probability of a query. Since \(p(q)\) is a constant (for a given query), it can be ignored for the purpose of ranking. Thus, the probability of a blog being a key source given the query \(q\) is proportional to the probability of a query given the blog \(p(q|\text{blog})\), weighted by the \textit{a priori} belief that a blog is a key source, \(p(\text{blog})\):

\[
p(\text{blog}|q) \propto p(q|\text{blog}) \cdot p(\text{blog}).
\]
of a blog, based on posts that belong the blog. From this representation we estimate the probability of the query topic given the blog’s model. The language modeling setting allows us to use blog posts to build associations between queries and blogs in a transparent and principled manner.

4.1 A baseline model

Our baseline model for estimating the probability of a query given a blog, \( p(q|\text{blog}) \), represents the blog by a multinomial probability distribution over the vocabulary of terms. Therefore, a blog model \( \theta_{\text{blog}} \) is inferred for each blog, such that the probability of a term given the blog model is \( p(t|\theta_{\text{blog}}) \). The model is then used to predict how likely a blog would produce a query \( q \). Each query term is assumed to be sampled identically and independently. Thus, the query likelihood is obtained by taking the product across all terms in the query:

\[
p(q|\theta_{\text{blog}}) = \prod_{t \in q} p(t|\theta_{\text{blog}})^{n(t,q)},
\]

where \( n(t,q) \) denotes the number of times term \( t \) is present in query \( q \).

To ensure that there are no zero probabilities due to data sparseness, it is standard to employ smoothing. That is, we first obtain an empirical estimate of the probability of a term given a blog \( p(t|\text{blog}) \), which is then smoothed with the background collection probabilities \( p(t) \):

\[
p(t|\theta_{\text{blog}}) = (1 - \lambda_{\text{blog}}) \cdot p(t|\text{blog}) + \lambda_{\text{blog}} \cdot p(t).
\]

In Eq. 4, \( p(t) \) is the probability of a term in the document repository. In this context, smoothing adds probability mass to the blog model according to how likely it is to be generated (i.e., published) by any blog.

To approximate \( p(t|\text{blog}) \) we use the posts as a bridge to connect the term \( t \) and the blog in the following way:

\[
p(t|\text{blog}) = \sum_{post \in \text{blog}} p(t|post, \text{blog}) \cdot p(post|\text{blog}),
\]

We assume the post and blog to be conditionally independent, and therefore set \( p(t|post, \text{blog}) = p(t|post) \). The probability \( p(post|\text{blog}) \) expresses the importance of a given post within the blog. The simplest approach is to set this distribution to be uniform, i.e., all posts of a blog are equally important. This results in \( p(post|\text{blog}) = \text{posts}(\text{blog})^{-1} \), where \( \text{posts}(\text{blog}) \) is the number of posts in the blog. In Section 5 we shall see an alternative way of setting this probability, based on blog specific features. Next, we discuss our estimation of the smoothing parameter \( \lambda_{\text{blog}} \). It is followed by an experimental evaluation of our baseline model.

4.2 Smoothing parameters

It is well-known that smoothing can have a significant impact on the overall performance of language modeling-based retrieval methods [15]. For the smoothing parameter \( \lambda_{\text{blog}} \) in Eq. 4, we set \( \lambda_{\text{blog}} \) equal to \( \frac{n(t, \text{blog})}{\text{posts}(\text{blog})} \), where \( n(t, \text{blog}) \) is the length of the blog (i.e., summarizing the length of all posts of the blog). Essentially, the amount of smoothing is proportional to the length of the blog (and is like Bayes smoothing with a Dirichlet prior [7]). So if there are very few posts in the blog then the model of the blog is more uncertain, leading to a greater reliance on the background probabilities. We set \( \beta \) to be the average blog length. That is \( \beta = 170 \) for the title-only index and \( \beta = 17,400 \) for the title-and-body index.

4.3 Assessing the baseline model

We report on the retrieval scores achieved by our baseline model; see Table 2 for the results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fields</th>
<th>MAP</th>
<th>R-prec</th>
<th>P@5</th>
<th>P@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline TB</td>
<td>.3272</td>
<td>.4023</td>
<td>.4844</td>
<td>.4844</td>
<td>.6892</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Results of our baseline retrieval model.

The scores obtained by our baseline model would have been ranked second (according to most measures) if submitted to the TREC 2007 Blog distillation task; cf. [6].

Ensuring our baseline model achieves state-of-the-art performance on the blog distillation task using posts as indexing units allows us to look for improvements over this baseline using blog specific features. The next section details on these features and how we implement these in our model.

5 Beyond the baseline

The document collection at hand, blog posts, has several specific features that could be used to improve blog retrieval effectiveness. We distinguish three types of feature: (i) document structure, (ii) temporal structure, and (iii) social structure. Below we discuss each of these types and show how they can be implemented in our blog distillation model. Results obtained using the various implementations are discussed in Section 6.

5.1 Document structure

There are (at least) two blog specific document structure characteristics that can be incorporated in the retrieval model: (i) document length and (ii) representation. The former is implemented using a prior probability of a post being relevant, the latter is achieved using a linear combination of two content representations.

Blog posts are characterized by relatively short document lengths. For a blog post to be considered relevant to a given topic though, it should contain “enough” information. Short posts, containing mostly uninformative utterances of a blogger, have a smaller a priori probability of being relevant than longer posts. We model this feature as the prior probability of a post \( p(post) \). We rewrite the \( p(post|\text{blog}) \) component from Eq. 5 using Bayes’ Theorem:

\[
p(post|\text{blog}) = \frac{p(\text{blog}|post) \cdot p(post)}{p(\text{blog})},
\]

where \( p(\text{blog}) \) is constant for all posts in a blog, and \( p(\text{blog}|post) \) is set to 1. We therefore set \( p(post|\text{blog}) \propto p(post) \). In the case of post length we set \( p(post) \propto \log(|post|) \), where |post| is the length of the post in words.

The second document structure characteristic we use is title (T) versus title+body (TB) representation. A blog post consists of a title and a body part. Since the title usually is a very clean, yet short description of the post content, we expect a precision-enhancing effect from using the T index instead of the TB index. On the other hand, due to the very short documents, we believe a T only run cannot achieve state-of-the-art performance on recall-based metrics (e.g., MAP). We therefore explore the possibilities of a mixture of two representations: The rationale behind mixing two content representations is to mimic a user’s search behavior: after being presented with a relevant blog post, a user might look at the titles of other posts.
within the same blog to come to a final relevance judgement concerning the entire blog. We mimic this behavior by combining the T and TB representations in a linear way:

\[ p(q|\theta_{blog}) = \lambda_{TB} \cdot p_{TB}(q|\theta_{blog}) + (1 - \lambda_{TB}) \cdot p_T(q|\theta_{blog}), \]

where both \( p_{TB}(q|\theta_{blog}) \) and \( p_T(q|\theta_{blog}) \) are defined in Eq. 3. In the final combination of this linear combination run with the other features, we apply the non-uniform probabilities \( p(post|blog) \) to the TB run only, and keep this probability uniform in the T run (i.e. \( p(post|blog) = posts(blog)^{-1} \)).

### 5.2 Temporal structure

Blogs have a very specific temporal structure in that posts within a blog are time-stamped and reverse chronologically ordered (most recent post first). When searching for blogs devoted to a given topic, one could assume the most recent posts to be of more importance than much older posts: even a blogger’s interest can shift over time, one could assume the most recent posts to be of more importance over the baseline on all metrics. This result does show that using the leaner content representation of the title only index helps improving blog distillation performance when combined with the full post index.

#### 5.3 Social structure

Probably the most eye-catching feature of blogs is their social structure: this structure displays itself in more than one way (e.g., blog rolls), but we focus on the commenting feature of blogs. Readers of a post can usually respond to that post by leaving a comment; the more comments readers leave with a certain post, the more important we consider this post to be. As with the document length, we implement this feature using a prior probability \( p(post) \) (following Eq. 6). We set \( p(post) \propto 1+\log(\text{comments}(post)) \), where \( \text{comments}(post) \) is the number of comments for a post.

### 6 Results

In this section we present the outcomes of our experiments, followed by the main observations from these outcomes and a more detailed analysis of these outcomes.

All results are summarized in Table 3: significance is tested using a two-tailed paired t-test, with significant differences compared to the baseline TB run reported with ** for \( \alpha = .01 \), and * for \( \alpha = .05 \). For runs (A) and (F) we need an estimation of \( \lambda_{TB} \), the weight of the title+body index (see Eq. 7). We estimate this parameter empirically by performing a sweep over possible \( \lambda_{TB} \) values. We obtained best performance using \( \lambda_{TB} = .7 \) for run (A) and \( \lambda_{TB} = .8 \) for run (F).

#### 6.1 Observations

We formulate the main observations based on our experiments. First, our baseline model displays state-of-the-art performance using a blog post index and no special features (Table 2). Further main observations include the influence of the title index, the blog specific features, and the combination of all previously presented components. Finally, we pick three example topics that are discussed in more detail.

##### Title-only index

We see that by using a title only index we can improve on early precision (P@5, P@10) and mean reciprocal rank, where the difference on P@5 is significant. This gain in precision has the side-effect of a loss in recall, as shown by significant drops in MAP and R-precision scores. Still, this title only run would have been ranked third in TREC 2007 (ordered by MAP). When we combine the T and TB runs we improve over the baseline on most metrics. This result does show that using the leaner content representation of the title only index helps improving blog distillation performance when combined with the full post index.

##### Blog specific features

As we can see from Table 3 all three blog features (post length, recency, and comments) help to improve over the baseline on most metrics. Especially post length seems a good indicator of post importance, whereas the number of comments shows only marginal improvements. When we combine all three features by averaging over their probabilities, we see an increase over the baseline on all metrics except P@10. From the fact that improvements of the combined run on MAP and R-precision are more significant than their single run counterparts, we can conclude that the combination has a recall-enhancing effect.

##### Combination of features

The combination of the TB run with all features and the T run shows best overall performance. It achieves highest scores on 3 of 5 metrics and improves significantly over the baseline on MAP, R-precision, P@10, and MRR. The performance in terms of MAP is close to the best TREC 2007 run (MAP .3695).

#### 6.2 Topic analysis

Zooming in on the results of the final run (F) compared to the baseline, we observe AP gains for most topics (see Figure 1: 32 out of 45 topics gain over the baseline, while 13 drop). Also, the improvements in AP are in general stronger than the AP drops. We select the two topics with the largest improvement (990 and 982), and the topic with the biggest AP drop (967):
Post-based approach to blog distillation with techniques that have significant improvements over the baseline on most metrics. A combination of all blog specific features on top of our baseline shows used as an indication of importance of a post given a blog; and (iv) a feature to improve over our baseline model.

Our main focus concerned an approach that uses blog posts as indexing units, instead of the (aggregated posts of) blogs that have so far mostly been used. There are a number of pragmatic reasons that account for this choice: (i) to allow for easy incremental indexing, (ii) for presentation of retrieval results posts are natural and coherent, (iii) to allow the use of one index for both blog post and blog retrieval. At the same time we aimed to achieve state-of-the-art performance. On top of this, we wanted to deploy blog specific units, and (iii) to allow the use of one index for both blog post and blog retrieval.

Our main finding is that the we can indeed achieve state-of-the-art performance. On the other hand, all other features, but especially the title only run can outperform the combination.

7 Conclusion

In this paper we described a language modeling approach to the task of identifying blogs that are principally devoted to a given topic. Our main focus concerned an approach that uses blog posts as indexing units, instead of the (aggregated posts of) blogs that have so far mostly been used. There are a number of pragmatic reasons that account for this choice: (i) to allow for easy incremental indexing, (ii) for presentation of retrieval results posts are natural and coherent units, and (iii) to allow the use of one index for both blog post and blog retrieval. At the same time we aimed to achieve state-of-the-art performance. On top of this, we wanted to deploy blog specific features to improve over our baseline model.

Our main finding is that the we can indeed achieve state-of-the-art performance using a blog post index. Additionally, we find that (i) the lean title-only content representation has a clear precision-enhancing effect when compared to a title-body representation, while a combination of the two representations outperforms both; (ii) post priors based on post length and number of comments have a positive influence on retrieval performance; (iii) temporal ordering of posts can be used as an indication of importance of a post given a blog; and (iv) a combination of all blog specific features on top of our baseline shows significant improvements over the baseline on most metrics.

As to future work, we are interested in combining our proposed post-based approach to blog distillation with techniques that have been shown to improve retrieval effectiveness for the task at hand, including query enrichment, link analysis, credibility scoring [14], and spam filtering.

Acknowledgements

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

REFERENCES