Relevance Stability in Blog Retrieval

Mostafa Keikha  
Faculty of Informatics  
University of Lugano  
Switzerland  
mostafa.keikha@lu.unisi.ch

Shima Gerani  
Faculty of Informatics  
University of Lugano  
Switzerland  
shima.gerani@lu.unisi.ch

Fabio Crestani  
Faculty of Informatics  
University of Lugano  
Switzerland  
fabio.crestani@unisi.ch

ABSTRACT
This paper investigates blog distillation where the goal is to rank blogs according to their recurrent relevance to the topic of the query. One of the main features of blogs is their relation to time but this important feature is under-utilized in the current blog retrieval methods. We propose a probabilistic framework to measure the stability of blogs relevance over time. We then study the effect of the proposed stability measure in the blog retrieval performance. We evaluate the proposed framework on the standard TREC Blog08 collection. The results show statistically significant improvements over state of the art models.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

Keywords
Blog search, Temporal Analysis, Relevance Stability

1. INTRODUCTION
User generated content is growing very fast and becoming one of the most important sources of information on the Web. Blogs are one of the main sources of information in this category. Millions of people write about their experiences and express their opinions in blogs everyday.

Considering this huge amount of user generated data and its specific properties, designing new retrieval methods is necessary to facilitate addressing different types of information needs that users may have. Users information needs in blogsphere are different from those of general Web users. Mishne and de Rijke [10] analyzed a blog query log and accordingly they divided blog queries into two broad categories called context and concept queries. In the context queries, users are looking for Named Entities with special interest in new events. While in concept queries they are looking for information related to one of their topic of interests. In this paper we focus on the blog distillation task where the goal is to answer topics from the second category [7].

Blog distillation is concerned with ranking blogs according to their recurring central interest to the topic of a user query [8]. According to the definition, a relevant blog for the distillation task should have recurring relevant post to the user’s query [7]. In other words, we are not interested in blogs that have many relevant posts in a short period of time and then stop posting about the topic. In order to capture this time dependent property in the retrieval model, we score each blog by its temporal relevance stability. Through this probability score, we aim to give more importance to blogs that have highly relevant posts in more time intervals and therefore show more relevance stability over time.

An important parameter in the proposed method is the size of the time interval (window size). We have tried both query independent and query dependent variations of the window size estimation. In the query independent version, we learn the size of the window using a training data whereas in the query dependent version, we propose a method to estimate the size of the time interval automatically for each query. We will show in the experimental part that this estimation can lead to comparable result to the fixed-size window selection while have the advantage of not needing any training.

Our aim in this paper can be summarized as answering the following questions:

- Do relevant blogs exhibit more stability in publishing relevant posts over time than non-relevant ones?
- How can we model the temporal stability in the probabilistic framework?
- How can we estimate the best time interval (window size) for each query?

The rest of the paper is organized as follows. In section 2 we review state of the art methods in blog retrieval. Section 3 describes our baseline in detail. Sections 4 and 5 explain our motivations and approach for using relevance stability in blog retrieval. Experimental results over the TREC 2009 blog data set are discussed in section 6. Finally, we conclude the paper and describe future works in section 7.

2. RELATED WORK
The main research on the blog distillation started after 2007, when the TREC organizers proposed this task in the blog track [8]. Researchers have applied different methods from areas that are similar to blog distillation, like ad-hoc
search, expert search and resource selection in distributed information retrieval.

The most simple models used ad-hoc search methods for finding relevant blogs to a specific topic. They treated each blog as one long document created by concatenating all of its posts together [3, 4, 13]. These methods ignored any specific property of blogs and mostly used standard IR techniques to rank blogs. Despite their simplicity, these methods performed fairly well in blog retrieval.

Some other approaches have been applied from expert search methods in blog retrieval [2, 7]. In these models, each post in a blog was seen as evidence that the blog has an interest in the query topic. In [7], MacDonald et al. used data fusion models to combine this evidence to compute a final relevance score of the blog, while Balog et al. adapted two language modeling approaches of expert finding and showed their effectiveness in blog distillation [2].

Some other researchers have employed resource selection methods from distributed information retrieval for blog retrieval [4, 1, 13]. Elsaas et al. dealt with blog distillation as a recourse selection problem [4, 1]. They modeled each blog as a collection of posts and used a Language Modeling approach to select the best collection. Similar approach was proposed by Seo and Croft [13], which they called Pseudo Cluster Selection. The idea was to create topic-based clusters of posts in each blog, and select the blogs which have the most similar clusters to the query. One of these methods, Small Document Model (SDM) [4], is a baseline for our experiments and will be discussed in more details in the next section. We combine the stability scores of the blogs with the scores calculated by SDM for the final ranking of the blogs.

Temporal properties of posts have been considered in different ways in blog retrieval. Nunes et al. [11] linearly combined the initial BM25 rank of the blog with the two different temporal ranks. The temporal ranks were based on two measures of temporal span and temporal dispersion. Although they did not show any statistically significance improvement over a content-only baseline.

Some other approaches, e.g., [5, 14] gave higher scores to more recent posts before aggregating them; they showed some improvement over the baseline which used only the contextual information of the blog. The most related approach to the one proposed in this paper is the work by Macdonald and Ounis [7] in which they use a heuristic measure to capture the recurring interests of blogs over time. Following the intuition that a relevant blog will continue to publish relevant posts throughout the timescale of the collection, they divided the collection into a series of equal time intervals. They then weight the relevance score of each blog by the total relative number of its relevant posts over different time intervals. Although they try to capture the recurring interest of bloggers to the topic by this measure, it fails in some cases. For instance, the measure does not discriminate between different distributions of relevant posts over time properly. In our model we capture these situations and give higher score to a blog that its relevant posts are distributed more uniformly over time.

3. LANGUAGE MODEL BASED BLOG RETRIEVAL

Resource selection in distributed information retrieval is a similar problem to the blog distillation task, where the goal is to rank collections of documents. The Small Document Model (SDM) is an adapted method from resource selections methods that deals with blogs as collections of posts and use a Language Modeling approach to select the best collection [4, 1]. This method is well justified from a probabilistic perspective and has been shown to perform well in practice. We have thus chose to use this methods as a baseline for our analysis and therefore describe it in more detail below.

Following the Language Modeling approach to IR [15], SDM ranks blogs according to their likelihood given the query:

\[
P(B|Q) = \frac{P(B)P(Q|B)}{P(Q)} = \min \frac{P(B)}{P(B)} \frac{P(Q|B)}{P(Q|B)}
\]

(1)

Where the query likelihood for a blog is calculated by summing over the query likelihoods for each post in the blog scaled by the probability (centrality) of the post within the blog:

\[
P_{SDM}(Q|B) = \sum_{p \in B} P(Q|p)P(p|B)
\]

(2)

Here \(p\) is the post in the blog, and \(P(Q|p)\) is the query likelihood for each post which is computed over query terms using Jelinek-Mercer smoothing. Our experiments show that uniform estimations perform fairly well as the blog prior and post centrality in the blog. In these estimations we consider a uniform prior over the blogs and as the post centrality we use a uniform distribution over posts in each blog. The query likelihood for a blog is then calculated by:

\[
P_{SDM}(Q|B) \sim \frac{1}{N_B} \sum_{p \in B} P(Q|p)
\]

(3)

4. IMPORTANCE OF STABILITY IN BLOG RELEVANCE

As mentioned earlier, from the blog distillation definition we expect a relevant blog to write about the topic regularly over time. It means that the blog relevance to the topic should be temporally stable. In order to capture this property, we propose an approach to calculate the relevance stability of the blog over time.

![Figure 1: Distribution of blogs with respect to the Query Supportive Windows.](image-url)
We first introduce the following terminology which we will use in the rest of the paper to explain our proposed method.

**Query Supportive Window:** A window in which a blog has at least one relevant post to the query.

Our proposed approach is based on the following assumption:

**Assumption:** The are more Query Supportive Windows for a relevant blog compared to a non-relevant blog.

In order to test the validity of this assumption on the data, we do some analysis using the relevance judgement of TREC09. We first partition the collection based on a time-interval (window) of 14 days. We then use the blogs that have been retrieved by our baseline, SDM, and calculate the percentage of blogs that have a post in more than the specified number of windows. Figure 1 shows distribution of the relevant and non-relevant blogs with respect to the number of Query Supportive Windows. For example considering value 5 on the X axis and the corresponding values on the Y axis for the relevant and non-relevant blogs in the graph, we can see that 32% of the relevant blogs have at least five Query Supportive Windows, while this value is 19% for the non-relevant ones. It is worth noting that all the examined blogs have been retrieved by SDM and have similar relevance scores for the topics based on that model. In fact if we just consider the content-based relevance score of blogs by the SDM, we are not able to capture this difference.

In the next section we describe how we can model the stability measure and integrate it into the blog retrieval system.

## 5. RELEVANCE STABILITY MEASURE

We can measure the relevance stability of a blog by estimating its relevance in different time windows. We then give more importance to blogs with higher number of relevant post in more number of Query Supportive Windows.

The first step is to divide the time span of the collection into equal sized time intervals (windows). We then use the following formula to calculate the relevance stability of a blog over time:

\[
P_{\text{stability}}(B|Q) = \sum_{w} P(B, w|Q)
\]

\[
= \sum_{w} P(B|Q, w)P(w|Q)
\]

This probability estimation is based on two components: \(P(B|Q, w)\) and \(P(w|Q)\), where \(P(w|Q)\) indicates the importance of the window for the given query and \(P(B|Q, w)\) indicates the query likelihood of the blog in the specified window for the given query. We will call these probability components **window importance** and **window-based query likelihood** respectively.

Further to take advantage of both stability and SDM methods, we interpolate the stability score of the blog with its SDM score:

\[
P(B|Q) = \alpha P_{\text{stability}}(B|Q) + (1 - \alpha) P_{\text{sm}}(B|Q)
\]

where \(\alpha\) is a parameter that controls the amount of stability that is considered.

In the rest of this section we explain the estimation of the two component of the stability probability in more detail.

**Window Importance:** The window importance component provides a way to integrate the popularity of a query in the window into our framework. We can assume the same importance for different windows independently of the query and use a uniform probability function for the importance, \(P(w|Q) \propto 1\).

Another way to estimate the window importance can be based on the popularity of the query in that window. The topic can be more popular in some intervals and those intervals should have more importance for the query. We can estimate the popularity of a query in a window, \(w\), by the amount of relevant information written about the query in that window, \(P(w|Q) \propto \sum_{p \in w} P(p|Q)\). This is similar to the query profile approach that is proposed by Jones and Diaz [6].

**Window-based Query Likelihood:** We consider the most relevant post of a blog in each window, as the representative of that blog in that window (we can use also other evidence like sum of the relevance scores of the blog posts in each window as the representative of the blog, but our experiments show that choosing the most relevant post is more effective). Let \(p_{i, j, t}\) show the post \(j\) of the blog \(i\) which has the time stamp \(t\). We define the score of the blog \(B_i\) in the window \(w\) as follows:

\[
score(B_i, w, Q) = \max\{score(p_{i, j, t})|t \in w\}
\]

where \(score(p, Q)\) indicates the relevance score of post \(p\) to the query \(Q\) and can be obtained from any retrieval model.

We use the Dirichlet-smoothed language model probability as the score of the post [15]:

\[
score(p, Q) \approx P(Q|p) = \prod_{t \in Q} P(t|p) = \prod_{t \in Q} \left(\frac{tf(t, p) + \mu P_{\text{ml}}(t|Q)}{\sum_t tf(t, p) + \mu}\right)
\]

We normalize the scores in each window to get the query likelihood of the blogs in that window:

\[
P(B_i|Q, w) = \frac{score(B_i, w, Q)}{\sum_j score(B_j, w, Q)}
\]

With this framework, we give higher score to a blog which its relevant posts are distributed in more time intervals than the one that has its posts in few intervals.

**Window Size Estimation:** The size of the window is an important parameter in the stability evaluation. We assume the window size to be the expected temporal distance between two relevant posts of a relevant blog. In other word, it indicated the dynamicity of the topic with more dynamic topic having shorter window size.

A simple way of defining the size of the window is to learn it from the training data. But since the training data is not always available and the training process is expensive, learning is not the most efficient solution for estimating the window size. Moreover, the window size indicates how regular we expect a relevant blog to write about the query topic. Thus, the window size highly depends on the query topic and the size that is learnt by the training topics is not necessarily the best estimation for a new topic.

Another way of defining the window size is to estimate it for each query separately based on the retrieved blogs for the query. We assume the top retrieved blogs for each topic to be relevant. Posts in each retrieved blog are used as samples of the distribution and the average distance between two consequent posts can be used as the expected value of the
window size for the query. The average temporal distance between the relevant posts of blog \( B_i \) to the query \( Q \), can be estimated as follows:

\[
E_{iQ} = \frac{\sum_{d_k \in R_{iQ}} \text{dist}(d_k, d_{k+1})}{|R_{iQ}| - 1}
\]

where \( R_{iQ} \) is the set of retrieved posts of the blog \( B_i \) for the query \( Q \) and \( \text{dist} \) function gives the temporal distance between two consequent posts. Thus, by having a set of retrieved blogs for the query, we can estimate a different expected value for each blog.

By assuming that blogs publish their relevant posts independently, each of the calculated estimations will be an independent estimator. We can then combine these independent estimators to have the final expected value of the window size [12]:

\[
E_Q = \frac{\sum_i E_{iQ}/(\sigma_i^2 + \epsilon)}{\sum_i 1/(\sigma_i^2 + \epsilon)}
\]

where \( E_{iQ} \) is the average temporal distance between two relevant posts in the blog \( B_i \) and \( \sigma_i^2 \) is the variance of those distances. The \( \epsilon \) is a constant value to overcome the situation with blogs having zero value of variance. A zero variance means the distribution is uniform and with small value of \( \epsilon \) we give high importance to blogs with uniform temporal distance between their posts. The \( E_Q \) in the equation 9 estimates the temporal distance between two consequent relevant posts in a relevant blog for the given query.

6. EXPERIMENTS

To evaluate our methods we use the TREC Blog08 data collection that is a collection of about 1 million blogs crawled over a year [9]. The collection consists of blog posts (permalink), feed, and homepage for each blog. We use only the permalinks component of the data which consist of approximately 28.4 million documents in the collection. We use 39 officially reported topics of the TREC 2009 and consider their titles as the queries [9].

Another available standard data set is the Blog06 collection that is used for blog track in TREC 2007 and 2008 [8]. Our experiment did not show any improvement on this collection and so we do not report the results here. The failure of the methods is mainly because this collection covers only 11 weeks of crawled blogs and does not include enough temporal information for the proposed method.

We use the Terrier Information Retrieval system\(^1\) to index the collection. We select the top 15000 relevant posts for each query as the working set. The posts in the working set are used for the windows size estimation as described in Section 5. In the estimation process, we simply consider blogs that have at least three posts in the working set.

We use 10 fold cross validation to tune the parameters of the model, included the \( \alpha \) in equation 5 and the size of the window, when it is not estimated by the retrieved blogs. Parameters are optimized to maximize the Mean Average Precision (MAP) metric.

In order to test for statistical significant improvements, we use the Wilcoxon signed-rank test on scores for each query at different levels.

Table 1 shows the results of the proposed framework in comparison with the baseline methods. The first two rows show the results of baselines methods. SDM indicates the Language Model based method as described in Section 3. SDM-Date is another baseline that uses the recurrent measure as introduced by MacDonald and Ounis to weight the SDM score [7].

The last six rows of the Table 1 show the evaluation of the stability framework with different settings. Statistically significant improvements in comparison with SDM and SDM-Date are shown by \( * \) and \( ** \) respectively. The symbols \( * \) and \( ** \) show the statistically significance at level 0.05, while the symbols \( *** \) and \( **** \) show the significance at level 0.01.

The first step of the experiments with the stability framework (the third and fourth rows of Table 1), uses the uniform importance for different windows. In this part of the experiments, the query likelihood of each blog is calculated using Equation 7 and finally blogs are ranked by Equation 5. Two possibilities for estimating the window size are explored. In the first approach, the window size is estimated using a cross fold validation method. In the second approach, it is estimated using the retrieved blogs for each query as described in the Section 5. As we can see, the stability framework outperforms both the baselines in MAP and Bpref metrics. There are small improvements in Precision at 10 in comparison with SDM, however they are not significant.

In the next set of experiments (the fifth and sixth rows of Table 1), we keep all the setting as before except the window importance. In this set of experiments, the window importance is calculated based on the popularity of the topic in each window as described in section 5. As with the previous step, we have statistically significant improvements over the baselines. However, we can see that adding the window importance to the stability framework does not change the performance significantly. Our further analysis showed that the popularity of the topics of the data set does not change between windows and this is the reason for windows importance not being effective.

As perviously discussed, the size of the window is an important parameter in the stability analysis. Figure 2 shows the sensitivity of the system to the size of the window. We can see that, it can highly affect the performance (MAP) of the system.

Because of the effect of the window size on the performance, it is necessary to have a proper estimation for it. In Section 5, we proposed an approach to estimate this parameter by top retrieved blogs for the query. However, number of the blogs to be used in this estimation is a new parameter of the model. Figure 3 shows the effect of this new parameter on the performance of the system. We can see that, the performance is robust against different values of the parameter. Figure 3 and 2 together show that we replaced a sensitive parameter of the system with an insensitive parameter that is very important in the performance of the system.

7. CONCLUSIONS AND FUTURE WORK

This paper investigated blog distillation where the goal is to rank blogs according to their recurrent relevance to the topic of the query. We focused on the temporal properties of blogs specifically on the relevance stability of blogs over time. We showed that relevant blogs are more stable over time regarding their relevance to the topic. We proposed a probabilistic framework to capture the relevance stability and used it in the ranking of the blogs. Since the time unit is an important parameter of the model, we introduced a
Table 1: Evaluation results over TREC09 query set. Statistically significant improvements over SDM and SDM-Dates are indicated by * and • respectively. **/•• show statistically significance at the level 0.05 and 0.01 respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>Window size</th>
<th>Window importance</th>
<th>MAP</th>
<th>P@10</th>
<th>Bpref</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDM</td>
<td>-</td>
<td>-</td>
<td>0.2046</td>
<td>0.3410</td>
<td>0.2365</td>
</tr>
<tr>
<td>SDM-Dates</td>
<td>-</td>
<td>-</td>
<td>0.2100</td>
<td>0.3462</td>
<td>0.2409</td>
</tr>
<tr>
<td>Stability Learnt from the training data</td>
<td>Uniform</td>
<td>0.2335 ** ••</td>
<td>0.3435 ** •</td>
<td>0.2584 ** •</td>
<td></td>
</tr>
<tr>
<td>Stability Estimated for each query</td>
<td>Uniform</td>
<td>0.2301 ** ••</td>
<td>0.3461 ** •</td>
<td>0.2548 ** •</td>
<td></td>
</tr>
<tr>
<td>Stability Learnt from the training data</td>
<td>Popularity</td>
<td>0.2331 ** ••</td>
<td>0.3384 ** •</td>
<td>0.2570 ** •</td>
<td></td>
</tr>
<tr>
<td>Stability Estimated for each query</td>
<td>Popularity</td>
<td>0.2357 ** ••</td>
<td>0.3564 ** •</td>
<td>0.2589 ** •</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Effect of the fixed window size on the performance

Figure 3: Effect of the number of employed feeds for window estimation on the performance

method to estimate this parameter using the retrieved posts for the query. Our experiments showed statistically significant improvements over state of the art methods in different settings of the framework.

Future work will involve more analysis on temporal properties of blogs and topics. Measuring the evolution of the topics over time can help us to better estimate the relevance models of them. The temporal relevance model is an unexplored and interesting area in blog retrieval. Also defining time-based centrality measures of blogs for a specific topic might be useful and needs further investigation.

8. REFERENCES