FEUP at TREC 2008 Blog Track: Using Temporal Evidence for Ranking and Feed Distillation

Sérgio Nunes, Cristina Ribeiro, Gabriel David
Departamento de Engenharia Informática
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
{ssn,mcr,gtd}@fe.up.pt

Abstract

This paper presents the participation of FEUP, from University of Porto, in the TREC 2008 Blog Track. FEUP participated in two tasks, the baseline adhoc retrieval task and the blog finding distillation task. Our approach was focused on the use of the temporal information available in the TREC Blog06 collection. For the baseline adhoc retrieval task a simple temporal sort was evaluated. In the blog finding distillation task we tested three alternative scoring functions based on temporal evidence. All features were combined with a BM25 baseline run using a standard rank aggregation approach. We observed small, but statistically significant, improvements in several evaluation measures when temporal information is used.

1 Introduction

In this paper we describe the first participation of *Faculdade de Engenharia da Universidade do Porto* (**FEUP**) in the TREC 2008 Blog Track. FEUP's participation was focused on the exploration of temporal evidence in two of the defined tasks — the baseline adhoc retrieval task and the blog finding distillation task.

The Blog Track uses the TREC Blog06 [8] collection, a large sample of documents crawled from the blogosphere. The collection contains more than 100,000 feeds of blogs and over 3.2 million permalink documents, both crawled over an eleven week period. This resource is one of the few standard collections of web documents containing temporal information.

Time is an intrinsic property of blogs, where each post is typically attached with a timestamp representing its publish date. In the TREC Blog06 collection, from the 3.2 million permalink documents available, almost 2 million have a date within the period of the collection (∼60%) [8]. This is an encouraging figure for further research based on the temporal properties of these documents.

The TREC 2008 Blog Track edition was structured in four distinct tasks: baseline adhoc retrieval, opinion finding, polarized opinion finding and blog finding distillation. We did not participate in any of the opinion retrieval tasks. Thus, considering the adhoc and distillation tasks and the number of submissions allowed, we defined the goals for our participation as follows.

For the adhoc retrieval task the unit of retrieval is the post, while for the blog distillation it is the feed. In the adhoc retrieval task we simply judge if the ordering of posts by publication date is a relevant criteria. For the blog distillation task we evaluated if the temporal dispersion of posts within a given feed is a positive criteria for retrieval. These ideas can be summarized in the following research questions:

- *Is the temporal order of documents a relevant criterion for adhoc retrieval?*
- *Is the temporal dispersion of relevant posts*

within a single feed a relevant criterion for feed distillation?

In the following sections we describe FEUP's participation in TREC 2008 Blog Track. In Section 2 we present a brief overview of the system used to conduct the experiments. In Section 3 we present our approach to the adhoc retrieval task. Then, we describe our experiments in the blog distillation task in Section 4. A short overview of related work is included in Section 5. The overall conclusions and a brief discussion of the results are presented in Section 6.

2 System Overview

We used the Terrier information retrieval platform [10] for all our experiments. Terrier is a state-of-the-art system with a very simple installation procedure and advanced customization features. Also, it supports the indexing of standard TREC collections natively, which greatly reduced the initial setup overhead.

The TREC Blog06 collection is structured in feeds (∼39GB), permalinks (∼89GB) and homepages (∼29GB). We used Terrier to build an index based only on the permalink documents, excluding the following TREC tags: `DOCHDR`, `EEDNO`, `BLOGHPNO`, `BLOGHPURL`, `PERMALINK`.

For the topics defined in each task, we retrieved a set of documents using Terrier's implementation of the BM25 model [11]. We used the default parameters defined in Terrier: $k_1 = 1.2d$, $k_3 = 8d$ and $b = 0.75d$. This set of results was then used in the adhoc and distillation tasks as described in sections 3 and 4.

All our retrieval experiments were made using this setup. It is worth noting that we did not made any effort to remove SPAM from the collection. The TREC Blog06 has been intentionally injected with SPAM to better reflect a realistic setting [8]. However, we opted to ignore SPAM since we believed that the presence of SPAM would not prevent us from addressing our initial research questions.

The temporal information associated with each permalink document is included in the collection. We ignored high granularity information like hours, minutes and seconds.

3 Baseline Adhoc Retrieval

The baseline adhoc retrieval task is a classic retrieval task, where the goal is to find all relevant information (blog posts) about a given topic. No opinion-finding techniques should be used in this task. We submitted two runs for this task. The first run is a standard automatic title-only run using Terrier with the BM25 model (see previous section). The second run is a combination of the BM25 run and a temporal score based on the post's publish date.

Typically, each blog post has a timestamp representing the date when the text was published. We defined the temporal score as a value between 0 and 1 computed by a linear transformation of each timestamp. To determine if older or newer posts are more relevant for adhoc retrieval, we built two different temporal ranks (ranked by newest and ranked by oldest). For instance, given that the collection spans from December 6th 2005 to February 21st 2006, a post published on the 1st of January 2006 would have a temporal score of $\frac{26days}{77days} = 0.34$ (ranked by newest) or a score of $\frac{51days}{77days} = 0.66$ (ranked by oldest).

As mentioned previously, in the TREC Blog06 collection approximately 60% of the posts have a valid date. We have discarded from the temporal ranks all posts not containing a valid timestamp (either missing or out of bounds).

To test if the timestamp of posts is a valid feature for adhoc ranking, we combined the initial BM25 rank with the two temporal ranks using a standard rank aggregation formula (Equation 1). We discarded all score information in this approach.

$$\alpha \times rank_{bm25} + (1 - \alpha) \times rank_{temporal} \qquad (1)$$

The α parameter was determined using data from the 2007 edition of the Blog Track. With $\alpha = 0.99$ and a temporal rank ordered by newest first, both P@20 and R-prec exhibited small improvement with 2007 topics and qrels. We submitted both the BM25

Run	MAP	R-prec	b-Pref	P@20
BM25	0.2482	0.3214	0.3454	0.5243
BM25T	**0.2483**	**0.3225**	**0.3455**	**0.5280**

Table 1: Adhoc Retrieval Task runs.

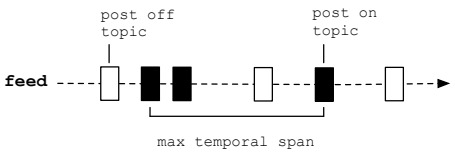

Figure 1: Temporal Span.

baseline rank and this combined rank to the baseline task.

Unfortunately, after submission, we detected a flaw in the combined run. Thus, this approach was not directly evaluated in TREC's assessment procedures. Nevertheless, after correcting the initial bug, we used the results produced by TREC assessors to conduct local evaluations and recreate the flawed run (named **BM25T**). The results are summarized in Table 1.

In MAP and b-Pref we observed only a very small improvement (0.03% in both) not statistically significant. The improvement observed in P@20 (0.7%) is statistically significant at a level of 0.1 ($p = 0.093$). Finally, the improvement verified in R-prec (0.34%) is statistically significant at a level of 0.05 ($p = 0.046$).

Despite the small improvements, these results confirm our initial expectations that basic temporal information is a positive criterion for adhoc retrieval. In Section 6 we discuss these results in more detail and propose ideas for future research.

4 Blog Finding Distillation

The blog distillation (or feed search) task consists in identifying blogs with a principle, recurring interest in a given topic. We defined a baseline run for this task by combining each post score into a feed score. Using the initial BM25 score for each blog post and topic, we calculated a feed score by adding all post scores (from the feed) and dividing by the number of posts available in the collection for that same feed. This approach was submitted as a run named **feupbase**.

Given the limit of four runs by team for this task, we opted to test two different strategies using temporal features, both individually and combined. In the following sections we describe these approaches, presenting the ideas for the sources of temporal evidence and how they have been combined. In sections 4.1 and 4.2 we present both ideas that make use of temporal evidence. In Section 4.3 we show how we combined these two approaches in a mixed setting. Finally, in Section 4.4 a brief overview of the results obtained in this task is presented.

4.1 Using Temporal Span

A blog (or feed) can be seen as a sequence of temporally ordered texts. Some texts will be relevant for a given topic, while others won't. One of our first ideas was to evaluate if the maximum temporal span covered by the relevant posts is a positive criteria in the task of blog distillation.

In Figure 1 we illustrate this simple idea. The feed depicted in the figure has six posts, three relevant to the topic (in black) and three not relevant (in white). The **temporal span** of a topic in a feed corresponds to the period between the newest relevant post and the oldest relevant post. As noted, this proposed feature is topic-dependent.

Having a list of feeds ranked by temporal span for each topic, we combine this rank with the baseline rank based on BM25. We used a standard rank aggregation approach as defined by Equation 1. To choose α we used data from the TREC 2007 Blog Track edition (topics and qrels) and optimized for b-Pref. The run combining the BM25 baseline and the temporal span was submitted with reference **feupfs**.

4.2 Using Temporal Dispersion

Still focused on the temporal properties of a feed, we investigated how the dispersion of relevant posts in a feed would impact the feed distillation task. Consider the two feeds depicted in Figure 2, where only

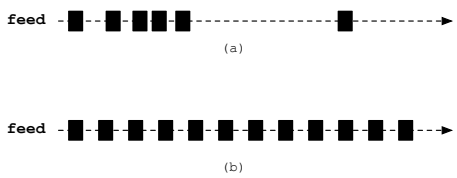

Figure 2: Examples of Temporal Dispersion.

relevant posts are included. In the top feed (*feed a*) the relevant posts are less dispersed than the posts in *feed b*. Which pattern is more relevant for the feed distillation task? Is the dispersion of relevant posts over time a property of relevant blogs?

We opted to use **negentropy** (or negative entropy) as a feature to represent the temporal dispersion of relevant posts in a feed. Negentropy is a measure used in information theory to represent the distance to normality. Negentropy is always positive and reaches its minimum for a gaussian random variable. We tested several alternative measures using data from previous editions of this track (e.g. Kurtosis, Skewness). The best results were obtained using the negentropy-based measure.

First, each post's publish date was converted to a relative global scale between 0 and 1. Each date was converted to the number of days since the beginning of the collection (6th of December 2005) and then divided by the total number of days in the collection (77 days). For each feed we obtained an ordered set of values between 0 and 1, corresponding to the publish dates of the relevant posts. Consider $p(i)$ to be the intervals between the subsequent values in this set. The negentropy of the relevant posts of a feed is given by Equation 2, where N is the total number of intervals between relevant posts (equal to the total number of relevant posts minus one).

$$Negen = -1 \times \frac{\sum_{i=1}^{N} p(i) \times ln(p(i))}{ln(N)} \qquad (2)$$

As an example, consider a feed containing four relevant posts published at the following (normalized) times: 0.3, 0.4, 0.5, 0.5. The intervals between the values are, respectively: $p(1) = 0.1$, $p(2) = 0.1$, $p(3) = 0$. The negentropy value for this set of posts is shown below [1].

$$-1 \times \frac{0.1 \times ln(0.1) + 0.1 \times ln(0.1) + 0 \times ln(0)}{ln(3)} = 0.42$$

In the example shown in Figure 2, the negentropy (i.e. dispersion) of *feed b* would be greater than the negentropy of *feed a*.

This feature was combined with the base BM25 rank using the same approach defined previously (see Equation 1). The α parameter was tuned using data from the previous track edition, and optimized for b-Pref. This run was submitted with reference **feupne**.

4.3 Mixed Approach

Our last run submitted for this task combined both the temporal span score and the temporal dispersion score, with the base BM25 feed scores. This run was derived by combining the two best runs including temporal span and temporal dispersion. These two runs were combined using Equation 1. The label for this run is **feupfsne**.

4.4 Results

The FEUP team submitted four runs to the blog distillation task. Our goal was to evaluate the impact of temporal features in a specific information retrieval task. We submitted one baseline run (feupbase) that completely discards all temporal information. Then, we tested two time-dependent features in runs feupfs (temporal span) and feupne (temporal dispersion). Finally, we mixed these two features in run feupfsne.

Unfortunately, an implementation bug was found in one of our basic functions. This resulted in erroneous values in all submitted runs. After identifying the bug we reconstructed the runs and results using the official qrels. All results presented in Table 2 were reconstructed offline.

Both proposed features improve over the temporally agnostic baseline in most metrics. The best results were achieved with the run combining the BM25 baseline and feed dispersion. The improvements for

[1] Note that we have considered that $0 \times ln(0) = 0$.

Run	α	MAP	b-Pref	P@10
feupbase		0.1993	0.2436	0.3280
feupfs	0.90	0.1999	0.2530	**0.3400**
feupne	0.85	0.2007	**0.2547**	**0.3420**
feupfsne	0.10	0.2008	**0.2547**	**0.3420**

Table 2: Results for reconstructed runs for the Distillation Task.

this run in b-Pref (+4.55%) and P@10 (+4.27%) are statistically significant at a level of 0.1. All statistically significant improvements are highlighted in bold ($p < 0.1$). Again, these results support our initial hypotheses about the value of temporal information for the distillation task. We briefly discuss these results in Section 6.

5 Related Work

The temporal information contained in the TREC Blog06 collection has been explored previously in the task of SPAM detection. Lin et al. [7] have shown that SPAM blogs (splogs) have a very distinct temporal dynamics pattern, typically due to the use of automated publishing mechanisms (e.g. bulk submissions at a given time). Their approach has proven to be very successful in splog detection. Our work also investigates the temporal features of the same collection but to address different tasks.

In a related work, Ernsting et al. [3] used a language modeling approach in the tasks of blog post and feed finding. In this work, a time-based probability of the document being considered was defined. More recent documents were considered to be more relevant (i.e. better reflect the current interests of a blogger). Results showed that, using this time-dependent prior, only a slight statistically significant improvement in MAP was observed. The authors also note an improvement in P@30. Our work is distinct since we propose and test different temporal features. Also, we used a probabilistic model as a framework for experimentation.

6 Conclusions

The general research question that we are tackling can be summarized as follows: *"Is temporal information a valuable evidence for web information retrieval tasks?"*. Addressing this problem has always been problematic due to the lack of standard test collections containing temporal information [9]. The TREC Blog06 collection emerged as an exception in this landscape of static (snapshot-like) corpora.

Although an important portion of the blog posts in the collection do not contain temporal information (∼40%), we tried to make use of this evidence in two of the Blog Track tasks: adhoc retrieval and blog distillation. We tested three time-dependent features: temporal order, temporal span and temporal dispersion. In each task we combined these with a BM25-based baseline.

We used a standard rank aggregation approach to combine the features. The weights used in the aggregation equation were tuned using data from last year's Blog Track edition and optimizing for b-Pref [6]. It is important to note that the rank aggregation approach is based solely on the rank of each result, thus discarding all the information contained in the scores.

Overall, results were positive and support our initial hypothesis — temporal information can be used as a source of valuable features for standard information retrieval tasks. We found statistically significant improvements in standard IR measures using simple time-dependent features like temporal order of posts.

This was a first approach to the use of temporal features for traditional information retrieval tasks based on a real web collection. Results were positive and encourage further research. We identify two main directions for future research: the definition of further time-dependent features, and the development of better feature combination formulas. Since we only used rank order in the aggregation formula, it should be possible to improve these results given that scores contain more information [2].

7 Acknowledgments

This work was supported in part by Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE - III Quadro Comunitário de Apoio), under grant SFRH/BD/31043/2006.

References

[1] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pages 200–209, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[2] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 416–423, New York, NY, USA, 2005. ACM Press.

[3] B. Ernsting, W. Weerkamp, and M. de Rijke. Language modeling approaches to blog post and feed finding. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, 2007.

[4] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of web pages. *Softw. Pract. Exper.*, 34(2):213–237, February 2004.

[5] C. Grimes. Microscale evolution of web pages. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1149–1150, New York, NY, USA, 2008. ACM.

[6] B. He, C. Macdonald, and I. Ounis. Retrieval sensitivity under training using different measures. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 67–74, New York, NY, USA, 2008. ACM.

[7] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng. Splog detection using self-similarity analysis on blog temporal dynamics. In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 1–8, New York, NY, USA, 2007. ACM Press.

[8] C. Macdonald and I. Ounis. The TREC Blog06 collection: Creating and analysing a blog test collection. Technical report, Department of Computing Science, University of Glasgow, Scotland, United Kingdom, 2006.

[9] S. Nunes. Exploring temporal evidence in web information retrieval. In A. Macfarlane, L. Azzopardi, and I. Ounis, editors, *BCS IRSG Symposium Future Directions in Information Access (FDIA 2007)*, pages 44–50. BCS IRSG, BCS IRSG, August 2007.

[10] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[11] S. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*, pages 73–96, 1995.