FEUP at TREC 2009 Blog Track: Temporal Evidence in the Faceted Blog Distillation Task

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 describes the participation of FEUP, from the University of Porto, in the TREC 2009 Blog Track. FEUP participated in the faceted blog distillation task with work focused on the use of temporal features available in the new TREC Blogs08 collection. The approach presented in this paper uses the temporal information available in most individual posts to amplify (or reduce) each post's score. Blog scores, and subsequent ranks, are obtained by combining individual posts' scores. While preparing the runs, no endeavors were made to identify a priori any temporal differences between the three distinct facets.

1 Introduction

In this paper we describe the participation of a group from *Faculdade de Engenharia da Universidade do Porto* (**FEUP**) in the TREC 2009 Blog Track. FEUP's participation was focused on the exploration of temporal evidence for the faceted blog distillation task.

In this year's edition of the TREC Blog Track, two significant changes were introduced. First, several new tasks were initiated, most notably a faceted blog distillation task. Second, a new base collection was used. This new collection is significantly larger than the previous Blogs06 collection and covers a much broader period of time. Given our interest in temporal properties, this is particularly relevant.

Our previous participation in the Blog Track [5] has shown that temporal information can have a positive impact in blog search. We continue this line of research by considering new approaches for incorporating time-sensitive features in the new faceted distillation task.

2 Blogs08 Collection Overview

The Blogs08 test collection was released in early April 2009 and is the official collection for the 2009 edition of the TREC Blog Track. For preparing this collection, a total of 1,303,520 feeds were polled once a week from January 14th, 2008 to February, 10th 2009 (394 days). The polled feeds, associated permalinks and homepage documents were stored, resulting in collection with a total compressed size of 453 GB.

Figure 1 presents an overview of the total number of posts per day found in the collection. The date information is obtained directly from the `DATE_XML` field available in the collection. As reported in the Blogs08 specification, *"DATE_XML is the date of issue of the permalink, as stated in the RSS or Atom feed. As such tags are optional in the feeds, this information is not always present. Should you choose to use this information, you should make your own decision on how to supplement it when it is not present for a document."*

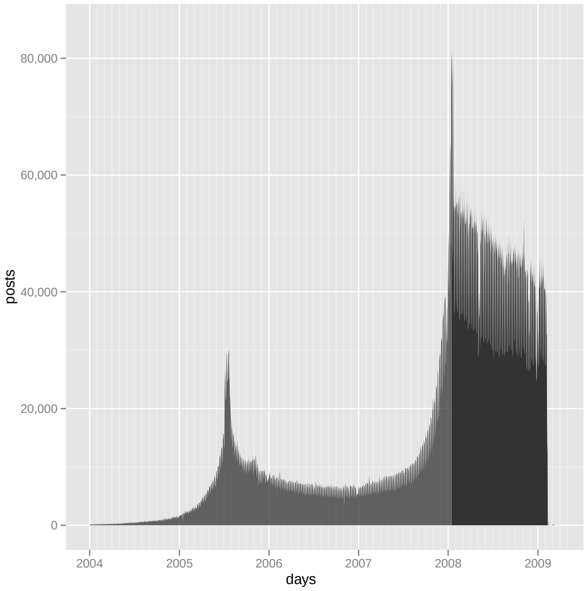

Figure 1: Blogs08 collection overview.

From the total number of permalink documents (posts) available in the collection (28,488,766), 97.5% had date information, while only 707,991 (2.5%) had an empty DATE_XML field. When considering only the posts with date information, 60.4% reported a date within the official crawling period. These posts are considered to have *valid dates*, while posts that report dates outside the crawling period are considered to have *invalid dates*. In Figure 1, valid and invalid posts were identified by using different colors. Table 1 summarizes the distribution of posts over a selected number of periods. There is a significant amount of documents published before the crawling period. This result is similar to that observed in the Blogs06 collection [4].

3 System Overview

The Terrier information retrieval platform [6] was used to index the permalink documents included in the Blogs08 collection, with the following TREC tags excluded: DOCHDR, DATE_XML, FEEDNO, BLOGHPNO, BLOGHPURL, PERMALINK. Documents were retrieved using Terrier's implementation of the BM25 model [7], maintaining the default parameters: $k_1 = 1.2d$, $k_3 = 8d$ and $b = 0.75d$.

Table 1: Summary of distribution of post's dates.

Period	**Total Posts**
Crawling Period	16,787,445 (58.9%)
After Crawling Period	4,987 (0.02%)
Before Crawling Period	10,988,343 (38.6%)
in 2007	4,406,209 (15.5%)
in 2006	2,386,807 (8.4%)
in 2005	3,283,107 (11.5%)
in 2004	205,590 (0.7%)
Without Date	707,991 (2.5%)

Document retrieval was done in two steps. First, a *phrase query* was used to retrieve documents, i.e. all terms needed to appear in the same phrase (e.g.: "term1 term2"). However, for some topics, this approach returned zero results. Thus, for these few topics, a more relax query was used – terms only needed to appear in the same document (e.g. "+term1 + term2"). It is worth noting that no effort was made to identify or remove SPAM from the collection.

4 Faceted Blog Distillation

The Faceted Blog Distillation is a new task introduced in this year's edition of the TREC Blog Track. This task is a refinement of the previous *blog distillation task*, where quality aspects of the retrieved blogs were not evaluated. These quality aspects were introduced by considering *facets* as proposed by Hearst et al. [2]. Three (binary) facets were considered in this first edition: *opinionated/factual*, *personal/official* and *in-depth/shallow*.

At FEUP we are focused on the study of the temporal properties available in blogs and their value for ranking tasks. We have prepared and submitted several runs that use temporal information to rank blogs. We have not tried to identify a priori how temporal properties could influence each facet property. Instead, we have adopted an exploratory attitude by submitting the same run for each topic and all facet options. In other words, each submitted run has the same ranking for the three facet options: *no facet applied*, *facet on for 1st value* and *facet on for 2nd value*. In a nutshell, for a given topic, the same rank of blogs was submitted for all *facet options*. This permits a detailed analysis of the impact of our approach in each topic and with each facet option.

4.1 Baseline

Given the BM25 post-based ranks (see Section 3), we prepared a baseline blog-based run by simply adding each feed's posts' scores and then dividing by the total number of posts available in the collection for that same feed. This run was submitted with the reference **FEUPirlab1**.

Due to hardware limitations, the original post-based runs used for preparing the blog-based runs were limited to 1,000 items. Permalink results with a rank higher than 1,000 were discarded. For topics that generate a lot of results from the same blogs, this might have a negative impact in the final blog distillation task rank.

4.2 Boost Invalid Dates

In a first approach, temporal information was introduced in the ranking formula by distinguishing between posts with valid dates (i.e. dates within the crawling period) and posts with invalid dates. The initial idea was to revise each post's original score based on its publish date being valid or not.

We implemented this approach using a simple formula: $score_{new} = score_{original} \times (1+\alpha)$. The value of α was determined using data from the 2008 edition of the Blog Track, and conducting a linear search with $\alpha \in [0, 2]$, using 0.01 increments as shown in Figure 2. Boosting posts with invalid dates results in consistent improvements in b-Pref. The highest boosting observed was of 3.8%. We prepared a run using $\alpha = 1$, i.e. posts with invalid dates had their scores doubled (a 100% boost) before computing the aggregated feed scores. This run was submitted with reference **FEUPirlab2**.

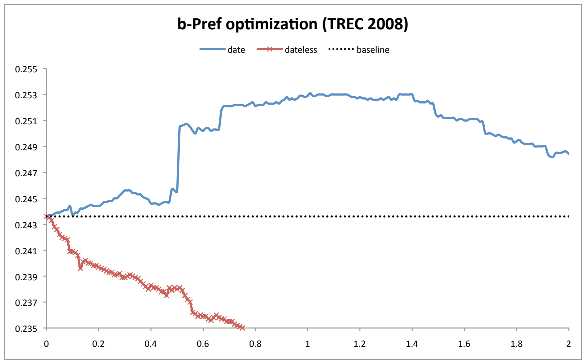

Figure 2: b-Pref values using TREC 2008 data for posts with valid and invalid dates.

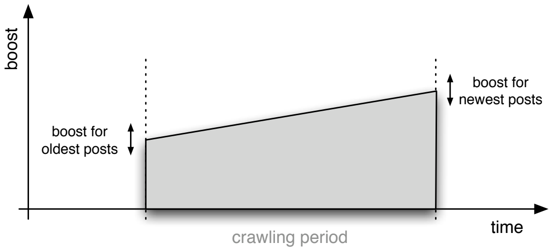

Figure 3: Boosting valid dates.

4.3 Boost Valid Dates

For posts with valid dates, two different scenarios were considered – boost newer posts and boost older posts. A simple linear scale was used as illustrated in Figure 3. Given two boost values for the limits of the crawling period (start and end), the posts in between were augmented following a linear scale. For example, if the starting boost parameter is higher than the ending boost parameter, older post are valued more than newer posts. In the figure, newer posts' scores have a greater boost than older posts.

We did not perform an exhaustive search over all possible values to identify the optimum parameters due to time limitations. Instead, we conducted a manual exploratory search testing values between 0 and 2, with 0.01 increments. The best b-Pref improvement was achieved using a starting boost of 0.1 and an ending boost of 0.43. A run named **FE-**

Table 2: Results of the faceted blog distillation task with facets off.

Run	MAP	b-Pref	R-prec	P@10
FEUPirlab1	0.1694	0.1911	0.2294	0.3179
FEUPirlab2	0.1752	0.1986	0.2447	0.3282
FEUPirlab3	0.1691	0.1950	0.2388	0.3179
FEUPirlab4	0.1662	0.1881	0.2248	0.3103

UPirlab3 was prepared with the previously defined boost applied to posts with invalid dates and a linear boost between 0.1 and 0.43 applied to posts with valid dates, i.e. newer posts have an higher boost. Additionally, a final run named **FEUPirlab4** was prepared with an inverse boost applied to posts with valid dates, from 0.43 to 0.1 (i.e. boost older posts). This last run was submitted to evaluate the impact of older posts in each facet. It is important to note that these runs were prepared using the previous **FEUPirlab2** run as a starting point.

5 Results

Our team at FEUP submitted four runs to the faceted blog distillation task as described in the previous section. The first run is *temporally agnostic* (i.e. all temporal information was discarded), and is used as a baseline to observe the impact of temporal features. Table 2 presents a summary of the official results for each run when facets are off and considering the official 39 topics. All statistically significant improvements are underlined ($p < 0.1$). Boosting posts with invalid dates resulted in an improvement of 3.42% in MAP and 3.24% in P@10. On the other hand, the refinements applied to posts with valid dates were inconclusive. Although some isolated improvements are observed, we cannot state that boosting newer posts or older posts produces better results.

A detailed analysis for each facet option is presented in Table 3. Again, all statistically significant improvements are underlined ($p < 0.1$). The most consistent improvements are observed in the *opinionated* and *official* facet options. For the opinionated facet, boosting the posts without dates combined with boosting the newer posts, resulted in an improvement of 12.21% in MAP. A similar result was observed in the official facet, with an improvement of 6% in MAP. Boosting older posts had a positive (although sporadic) impact in the *in-depth* and *personal* facets. Overall, the worst results were found in the *shallow* facet, with very low MAP values. As pointed in the results tables, few improvements are statistically significant. This can be partially explained by the small number of cases for each facet option. For instance, in many facet options, we have fewer than 10 paired cases.

6 Related Work

In the context of the TREC Blog Track, previous work has shown that temporal information available in posts can have a positive effect in both document ranking [1, 5] and SPAM detection [3]. This work differs from our previous approach [5] in two distinctive aspects. First, it uses a new collection spanning a significantly larger time period (slightly more than a year). This aspect is particularly relevant for our work, given that it depends directly on features derived from dates. A broader collection has a higher number of potential *temporal bins* to discriminate results. Second, in our previous approach we combined a starting BM25-based rank with temporally biased ranks. Rank combination discards the finer-grained score values. In this work we have added a temporal bias to the original scores. Also, given the faceted nature of the task, we have made an initial investigation about the impact of time in these three facets.

7 Conclusions

The TREC Blogs08 collection is a new resource released in early 2009. In comparison with the Blogs06 collection, a larger number of blogs was crawled over a significantly larger time period (394 *vs* 77 days). Given our interest in temporal properties, this new collection is specially valuable. Our main goal was to conduct a first exploration of the Blogs08 collection focused on the temporal properties of blog posts. In

Table 3: Results of the faceted blog distillation task for each facet option.

Run	MAP	R-prec	P@10
— in-depth (N=18) —			
FEUPirlab1	0.1490	0.1441	0.2167
FEUPirlab2	0.1489	0.1625	0.2111
FEUPirlab3	0.1412	0.1523	0.1889
FEUPirlab4	0.1494	0.1385	0.2111
— opinionated (N=13) —			
FEUPirlab1	0.0999	0.1360	0.1462
FEUPirlab2	0.1068	0.1458	0.1692
FEUPirlab3	0.1121	0.1466	0.1846
FEUPirlab4	0.0934	0.1360	0.1538
— personal (N=8) —			
FEUPirlab1	0.1764	0.1975	0.1750
FEUPirlab2	0.1791	0.2464	0.2000
FEUPirlab3	0.1203	0.1749	0.1625
FEUPirlab4	0.1749	0.2168	0.1625
— shallow (N=18) —			
FEUPirlab1	0.0506	0.0731	0.0667
FEUPirlab2	0.0491	0.0564	0.0611
FEUPirlab3	0.0497	0.0638	0.0722
FEUPirlab4	0.0500	0.0759	0.0611
— factual (N=13) —			
FEUPirlab1	0.1369	0.1184	0.1308
FEUPirlab2	0.1339	0.1107	0.1308
FEUPirlab3	0.1370	0.1258	0.1231
FEUPirlab4	0.1347	0.1143	0.1308
— official (N=8) —			
FEUPirlab1	0.1499	0.1078	0.1250
FEUPirlab2	0.1523	0.1126	0.1250
FEUPirlab3	0.1589	0.1126	0.1500
FEUPirlab4	0.1470	0.0989	0.1125

contrast with our previous approach, we integrated the temporal features in the document scores before producing the final ranks. Our previous strategy was based on combining two ranks, discarding finer-grained score information.

From these experiments, we can draw some preliminary conclusions. Favoring posts with invalid dates produces consistent improvements. To a certain extent, this can be explained by the fact that the large majority of posts with invalid dates are prior to the crawling period. Thus, by enhancing posts with invalid dates we are giving a higher score to older posts, which tend to be associated with greater authority. In some facets, such as opinionated and official, the positive impact of this strategy was clear. However, from this data, it is not possible to draw a straightforward conclusion relatively to the impact of older and newer posts in the blog distillation task. It is worth highlighting that the calibration of the scoring parameters was done using data and assessments from a different collection, thus it is expected to have limitations.

It is our conviction that time is a complex dimension that cannot be treated in a linear, one-sided fashion. For instance, on one hand, recent posts tend to be valued because they are more up-to-date and focused on the current subjects. On the other hand, old posts have an intrinsic value derived from their longevity and established nature.

8 Acknowledgments

We would like to thank the organizers of the TREC Blog Track, particularly Iadh Ounis, for the prompt feedback provided on several occasions.

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] 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.

[2] M. A. Hearst, M. Hurst, and S. T. Dumais. What should blog search look like? In I. Soboroff, E. Agichtein, and R. Kumar, editors, *Proceeding of the 2008 ACM workshop on Search in Social Media (SSM'08)*, pages 95–98, New York, NY, USA, 2008. ACM, ACM.

[3] 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.

[4] 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.

[5] S. Nunes, C. Ribeiro, and G. David. FEUP at TREC 2008 Blog Track: Using temporal evidence for ranking and feed distillation. In E. M. Voorhees and L. P. Buckland, editors, *17th Text REtrieval Conference (TREC 2008)*. NIST, November 2008.

[6] 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.

[7] 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.