Using Temporal Evidence in Blog Search

Sérgio Nunes[1] Cristina Ribeiro[1,2] Gabriel David[1,2]

[1]Departamento de Engenharia Informática
Faculdade de Engenharia da Universidade do Porto
[2]INESC-Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal

{ssn,mcr,gtd}@fe.up.pt

ABSTRACT

In this paper we present a study on the relevance of web documents over time and the use of temporal evidence in blog search tasks. Time is an intrinsic property of social media, most notably in blogs where each post is typically attached with a timestamp representing its publish date. However, due to the challenges in obtaining document collections containing temporal information, research on this field has been scarce. We base our study on the Blog06 collection and the relevance assessments produced in the context of the TREC Blog Track, to investigate the relevance of time-based features in standard retrieval tasks. We observe small, but statistically significant improvements over a BM25 baseline when temporal information is used. Also, we find a direct connection between recency and relevance of documents for ad-hoc retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; I.7.5 [**Document and Text Processing**]: Document Capture—*Document analysis*

General Terms

Algorithms, Experimentation, Measurement

Keywords

Web Information Retrieval, Temporal Features, Blog Search, Web Dynamics

1. INTRODUCTION

Research in the field of Web Information Retrieval has been mostly focused on time-independent features. Most standard collections available are snapshots of a given moment, despite the fact that the World Wide Web is a very dynamic information system. Moreover, research has shown that previous revisions of a current web document contain relevant information related to the document's content [13]. The challenges in crawling and storing a web collection containing historic information have hindered the study of temporal features in web information retrieval.

Time is an intrinsic property of social media, most notably in blogs where each post is typically attached with a timestamp representing its publish date. In this context, blogs emerge as an interesting opportunity for research on the temporal features of web documents. We use the Blog06 blog collection together with relevance assessments produced in the context of the TREC Blog Track to study the relation between document relevance and time, and also to propose and evaluate temporal features for standard blog search tasks.

The Blog06 corpus [11] is a large sample of documents crawled from the blogosphere. This 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. We used Terrier [14], an information retrieval platform, for all our experiments. The Blog06 collection is structured in feeds, permalinks and homepages. We used Terrier to build an index based only on the permalink documents, ignoring all other content.

This paper is structured in three main sections, in Section 2 we present a study of relevance over time in the context of blog post retrieval, in Section 3 we address the problem of ad-hoc retrieval in blog search and in Section 4 we work on the task of blog distillation. For the ad-hoc retrieval task the unit of retrieval is the post, while for the blog distillation it is the feed. In the ad-hoc 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 distribution of posts within a given feed is a positive criteria for retrieval.

2. RELEVANCE OVER TIME

In the Blog06 collection, from the 3.2 million permalink documents available, almost 2 million have a date within the period of the collection ($\sim$60%) [11]. This is an encouraging figure for research based on the temporal properties of these documents. The temporal information associated with each permalink document is included in the collection. Dates

were derived directly from the feeds and validated using the crawl date. We ignored high granularity information like hours, minutes and seconds, and only consider days as a unit of time.

Using the qrels from the TREC Blog Track, we first observed how the absence of date information was reflected in the distribution of both relevant and non-relevant posts. We discarded polarity information available on the qrels (i.e. positive, negative opinion) since we are only interested in the relevance of documents. In Figure 1 we present the distribution of missing temporal information over relevance in all editions of the track. For instance, in the 2006 qrels, 23% of all relevant documents have no date information attached to them, in contrast 40% of all non-relevant documents have no date information. Non-relevant blog posts tend to have missing or invalid date information.

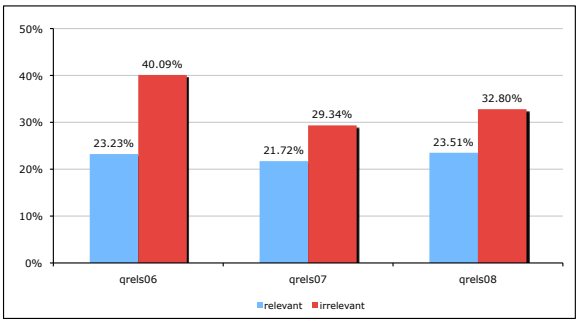

Figure 1: Documents without temporal information.

Considering the judged documents containing temporal information, we extracted the percentage of both relevant and non-relevant documents for each day of the collection. For instance, in the 57th day of the collection (January 31st 2006) we found 2.79% of all relevant documents from the 2006 qrels. In contrast, we found 1.28% of all non-relevant documents in that same day. This results in a difference of 1.51%. Figure 2 shows this difference for each one of the 77 days of the collection. A trend line is also included in the plot showing the growth over time. We found similar trends using qrels from 2007 and 2008. This observation supports the idea that the recency of documents is a positive factor for document relevance.

To further support this claim, we created runs using topics from all editions of the TREC Blog Track and computed MAP values based on the official qrels. These runs are exclusively based on temporal information. For each track edition we computed two runs, one ordered by oldest posts first and another ordered by newest posts first. Results from this experiment are listed in Table 1. All improvements observed in runs ordered by newest first are statistically significant at a level of 0.05 using a paired sample t-test.

3. BLOG POST RANKING

To evaluate the impact of temporal information in ad-hoc retrieval tasks, we combined document temporal ordering with a BM25 ranking. We used topics and qrels from the Baseline Ad-hoc Retrieval Task in the TREC Blog Track from 2008. First we retrieved a ranked set of documents using Terrier's implementation of the BM25 model [15]. We

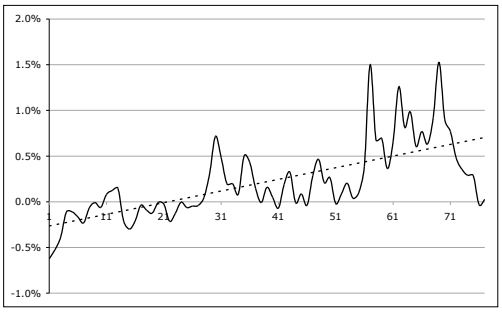

Figure 2: Difference between relevant and non-relevant posts over time. Oldest days are on the left side of the plot.

	MAP		
	2006	2007	2008
newest first	**0.1321**	**0.1145**	**0.0985**
oldest first	0.1162	0.0841	0.0823

Table 1: MAP values for temporally-ordered ranks.

used the default parameters defined in Terrier: $k_1 = 1.2d$, $k_3 = 8d$ and $b = 0.75d$.

We defined the temporal score as a value between 0 and 1 computed by a linear transformation of each timestamp. 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 $\frac{26days}{77days} = 0.34$. All posts without a valid date, either missing or out of bounds (∼40%), were discarded from this rank. The two ranks were combined using a simple rank aggregation approach as defined by Equation 1.

$$\alpha \times BM25_{rank} + (1 - \alpha) \times Temporal_{rank} \quad (1)$$

The parameter α was set based on data from the 2007 edition of the Blog Track using a linear search with 0.01 increments. The best tuning ($\alpha = 0.99$), resulted in a small improvement in R-prec and P@20. Results with 2008 data are summarized in Table 2. 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 ($p = 0.093$). Also, 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 ad-hoc retrieval.

4. BLOG FEED DISTILLATION

In the context of blog search, the distillation of blogs (or feed search) is an important task. This task is also included in the Blog Track at TREC and the goal is to identify blogs

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

Table 2: Results with temporal evidence in ad-hoc retrieval.

with a principle, recurring interest in a given topic. The unit of retrieval for this task is the blog (or feed), contrary to post retrieval task discussed previously. Since blogs combine the individual temporal information from each post, we can derive new temporal features for this specific task.

We defined a baseline run for this task by combining each post's 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 and dividing by the number of posts available in the collection for each feed. The results for this run are labeled **"BM25"** and presented in Table 3.

Since we found evidence of improvements when combining individual posts with a temporal rank (see previous section), we prepared a run using the temporal biased posts. This approach exhibited small, not statistically significant, improvements. Results are detailed in Table 3 and identified as **"BM25 + Post Time"**.

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. This led us to evaluate if the maximum temporal span covered by the relevant posts is a positive criteria for blog distillation.

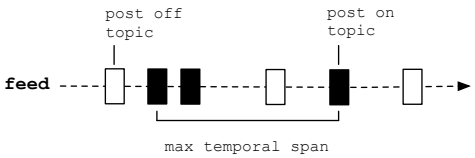

Figure 3: Temporal span.

In Figure 3 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 the same linear rank aggregation approach defined previously (see Equation 1). The parameter α was fixed using data from the TREC 2007 Blog Track edition and optimizing for b-Pref using a linear search with 0.01 increments. He et al [8] have shown that optimizing for b-Pref produces better results with incomplete relevance judgements. Results for $\alpha = 0.9$ are presented in Table 3 and labeled **"BM25 + Span"**.

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 4, where only 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?

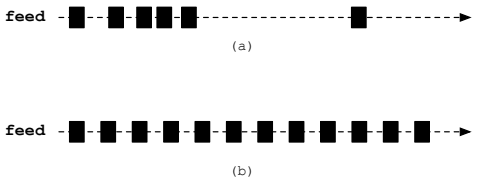

Figure 4: Examples of temporal dispersion.

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, all relevant posts from each feed were converted to a relative scale between 0 and 1. Each date was converted to the number of days since the first relevant post in the feed and then divided by the total number of days until the last relevant post (i.e. temporal span of relevant posts). For each feed we obtained an ordered set of values from 0 to 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)[1]. Given that the relevant posts in each feed were normalized, we can work with $p(i)$ as being a probability distribution, since $\sum_i p(i) = 1$.

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

As an example, consider a feed containing four relevant posts published at the following dates: 2005-12-15, 2006-01-18, 2006-01-20 and 2006-01-30. After normalization, we have the following set: [0, 0.74, 0.78, 1]. The intervals between these values are, respectively: $p(1) = 0.74$, $p(2) = 0.04$, $p(3) = 0.22$. The negentropy value for this set of posts is shown below.

$$-\frac{0.74 \times log(0.74) + 0.04 \times log(0.04)}{log(3)} -$$

$$-\frac{0.22 \times log(0.22)}{log(3)} = 0.62$$

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

In the example shown in Figure 4, 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 adopted previously. The parameter α was tuned using data from the previous track edition, using 0.01 increments and optimizing for b-Pref. The best b-Pref value was observed with $\alpha = 0.88$. Results for the 2008 data using this value are shown in Table 3 and labeled **"BM25 + Dispersion"**.

	α	MAP	b-Pref	P@10
BM25	—	0.1993	0.2436	0.3280
BM25 + Post Time	—	0.2005	0.2555	0.3340
BM25 + Span	0.9	0.1999	0.2530	**0.3400**
BM25 + Dispersion	0.88	0.1970	0.2511	0.3220

Table 3: Results with temporal evidence in feed distillation.

Overall, the *post date* and *temporal span* features result in improvements over the temporally agnostic baseline as shown in Table 3. However, not all improvements are statistically significant. In bold we highlight the improvement that is statistically significant at a level of 0.01 using a paired sample t-test. Contrary to our expectations, the *temporal dispersion* feature results in no improvement over the baseline.

Despite some marginal evidence that supports our initial hypotheses about the value of temporal information in the blog distillation task, more work is needed to fully understand its impact in this task. A more detailed discussion about these results, including possible avenues for further research, is presented in the conclusions.

5. RELATED WORK

We are not aware of any previous work about the relevance of temporal features for ad-hoc retrieval conducted over a large realistic collection containing historic web documents, and using independent relevance assessments. Previous work on the temporal nature of the web has been almost exclusively focused on the characterization of changes with the underlying goal of optimizing web crawling [1, 6, 7]. An exception is the work by Li et al. [9] where the authors explore the relationship between time and relevance using TREC ad-hoc queries. They found that time is a positive signal but only in some queries. Our work differs from this since we use a different TREC collection based on data extracted from blogs (TREC Blog06) and we propose and evaluate different time-based features (besides temporal order or recency).

The temporal information contained in the Blog06 collection has been explored previously in the task of SPAM detection. Lin et al. [10] 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, namely ad-hoc blog post retrieval and ad-hoc feed distillation.

In a related work, Ernsting et al. [5] 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. Although these results are in line with our findings, our work is distinct since we propose and test different temporal features. Also, we used a probabilistic model as a framework for experimentation.

The work by Elsas et al. [4] presents a good overview of current research in blog feed search. Our work presents original contributions to this field using *time* as a source of evidence.

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 [12]. The 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 valid temporal information (∼40%), we tried to make use of this evidence in two distinct tasks — ad-hoc blog post retrieval and ad-hoc blog distillation. We proposed and tested three time-dependent features: temporal order, temporal span and temporal dispersion. Each feature was compared against BM25-based baselines.

We used a standard rank aggregation approach to combine the features. The weights used in the aggregation equation were tuned using data from previous editions of the TREC Blog Track. 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. In this paper we present evidence that time can be used as a relevant source of information for relevance assessment. We found statistically significant improvements in standard IR measures using simple time-dependent features like temporal order of posts. Also, we studied the changes in document relevance over time, using human judgments from three editions of TREC. This clearly showed that recency is related to relevance (i.e. more recent documents tend to be considered more relevant by assessors). Topic bias can be discarded since the queries used in the assessments were collected in three distinct years.

This was a first approach to the use of temporal features for traditional information retrieval tasks based on a real web collection. Despite the short temporal span of the collection (77 days), we observed positive results that encourage further research on this topic. We identify three main direc-

tions for future research: the definition of additional time-dependent features, the development of better feature combination formulas, and the clustering of queries by classes (e.g. containing temporal expressions). 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]. Also, we plan to conduct experiments combining all signals studied (i.e. BM25, Publish Time, Temporal Span and Temporal Dispersion). Finally, in Section 2 we found evidence that non-relevant documents tend to have no temporal information associated. Although this finding was not explored in the methods proposed in this paper, we plan to explore this idea in the future.

Given the temporally rich nature of social media (e.g. blogs, wikis, micro-blogging) we consider time to be a very promising feature for future research in this field. The preliminary results presented in this paper vouch for this line of research. The importance of temporal features is further stressed by the fact that time is seen by users as a critical dimension when organizing data in the context of personal information management systems [3].

7. ACKNOWLEDGMENTS

This work was partially supported 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. The authors would also like to thank the anonymous reviewers for their useful feedback and comments.

8. 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] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 72–79. ACM Press, 2003.

[4] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog feed search. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 347–354, New York, NY, USA, 2008. ACM.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] X. Li and B. W. Croft. Time-based language models. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 469–475, New York, NY, USA, 2003. ACM Press.

[10] 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.

[11] 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.

[12] 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.

[13] S. Nunes, C. Ribeiro, and G. David. Wikichanges - exposing wikipedia revision activity. In *WikiSym'08: Proceedings of the 2008 international symposium on Wikis*, New York, NY, USA, September 2008. ACM.

[14] 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.

[15] 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.