Using Blog Properties to Improve Retrieval

Gilad Mishne
ISLA, University of Amsterdam
gilad@science.uva.nl

Abstract

This paper describes three simple heuristics which improve opinion retrieval effectiveness by using blog-specific properties. Blog timestamps are used to increase the retrieval scores of blog posts published near the time of a significant event related to a query; an inexpensive approach to comment amount estimation is used to identify the level of opinion expressed in a post; and query-specific weights are used to change the importance of spam filtering for different types of queries. Overall, these methods, combined with non-blog-specific retrieval approaches, result in substantial improvements over state-of-the-art.

Keywords

Blog retrieval, opinion retrieval, TREC

1. Introduction

The annual Text Retrieval Conference (TREC) is organized around a set of separate tracks, each investigating a particular retrieval domain, and each including one or more tasks in this domain. In 2006, TREC featured, for the first time, a track dedicated to blog retrieval: the TREC Blog Track. In particular, the track included an opinion retrieval task, where participants were requested to locate blog posts expressing an opinion about a topic in a large collection of posts. The polarity of the sentiment in a post was not required to be identified: rather, any post answering the question "What do people think about [the entity in the query]" was considered relevant. Queries included mostly person names, products, and brand names, taken from a query log of a blog search engine. More details about the opinion retrieval task, the data used for it, the queries, and the assessments carried out are found in [10].

Our approach to the opinion retrieval task identified three aspects involved in locating opinionated blog posts: *topical relevance*, *opinion expression*, and *post quality*. The first, topical relevance, is the degree to which a post deals with the given topic; this is similar to relevance as defined for ad-hoc retrieval tasks, such as many of the traditional TREC tasks. The second aspect, opinion expression, involves identifying whether a post contains an opinion: the degree to which it contains subjective information about a topic. Finally, the post quality is an estimation of the (query-independent) quality of a blog post, under the assumption that higher-quality posts are more likely to contain meaningful opinions and are preferred by users. In this last category of quality we also include detection of spam in blogs, defining a spam blog post as a low-quality one.

We addressed each of these three aspects independently of the rest, using a wide range of techniques: some of those were blog-specific, and some general methods used in various retrieval settings. Each technique resulted in a separate relevance score for each blog post: standard information retrieval approaches resulted in a ranking of posts by their topical relevance to a query; sentiment analysis was used to rank all posts by the amount of sentiment contained in them; spam filtering was used to rank all posts by their estimated spam level; and so on. The final ranking of a blog post was obtained by combining the partial scores assigned to it by the different approaches using a linear combination. Overall, this method proved as one of the top performers at TREC; more information about it is found in [7].

Of the different methods we used, in this paper we describe three, one from each of the high-level aspects we investigated; all three use properties which are specific to the blogspace, and all three are based on a straightforward, inexpensive approach. We show that each of these techniques improve over a baseline, and that, combined with other techniques we use, they improve also over state-of-the-art.

2. Improving retrieval using blog properties

We now describe in more details the three approaches; evaluation of each follows in the next Section. The first approach we discuss uses the timelined nature of blogs to identify periods of increased possible relevance. The second relates the amount of comments in a blog posts and the likelihood of an opinion being present in the post. The last of the methods we describe uses query-dependent spam filtering to reduce noise in the collection.

2.1 Temporal relevance feedback

The blogspace is a dynamic medium, quickly responding to ongoing events; as a result, a substantial number of blog search queries are related to specific events, in many cases news-oriented ones [8]. The distribution of dates in relevant documents for these queries is not uniform, but concentrated around a short period during which the event took place. For example, Figure 1 shows the distribution of dates in relevant documents for the query "state of the union," which seeks opinions about the presidential state of the union address, delivered on the evening of January 31st, 2006: clearly, relevant documents are found mostly in the few days following

the event.

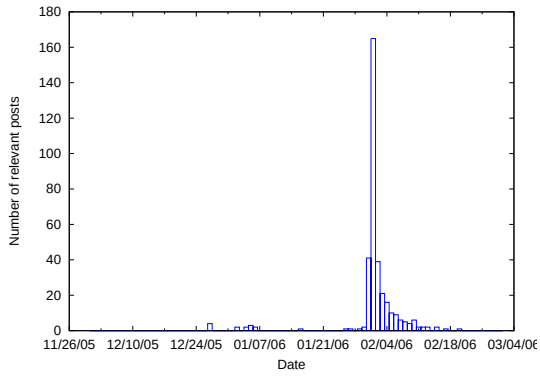

Fig. 1: *Distribution of relevant posts for the query "state of the union" over time*

A method assigning higher relevance to recent documents by incorporating it into the language modeling retrieval framework, testing the temporal distribution of query terms and using a decay function from detected peaks in the aggregated distribution, has been proposed by Li and Croft [5]. We propose a simpler, somewhat more intuitive method, which also does not require training a decay function.

As with blind relevance feedback methods, we assume that highly-ranked documents (according to topical similarity) are relevant. The distribution of dates in these highly-ranked documents serves as an estimation to the distribution of dates in relevant documents, and is simply treated as a prior probability of the relevance of a document. For example, the distribution of dates in the top-500 retrieved results for the same query as used in Figure 1, "state of the union," is shown in Figure 2. Clearly, this distribution is more noisy than that of the actual relevant posts; but the peak around the time of the event is preserved.

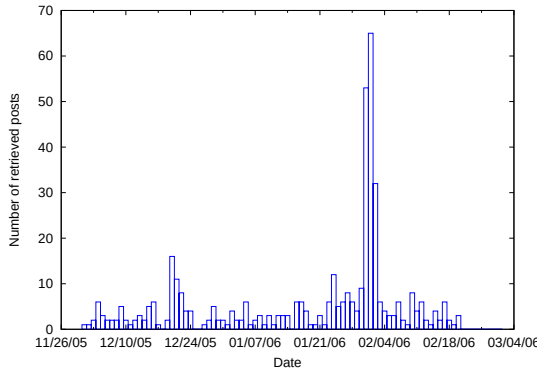

Fig. 2: *Distribution of the top-500 retrieved posts for the query "state of the union" over time*

We assume that blog posts published near the time of an event are more likely to be relevant to it, regardless of their topical similarity. As with content-based blind relevance feedback, this allows us to identify relevance also for documents which do not contain the query words. This temporal prior likelihood is then combined with its topical retrieval relevance using a linear combination to rerank the topical-retrieval only results – an approach often used for recency queries [5]. Our approach is similar to that of Diaz and Jones [2], but instead of using it to estimate query performance we use it to improve retrieval results.

2.2 Using comment information

The presence of comments in a blog post indicates that some discussion is taking place regarding the contents of the post; in previous work, we have shown this to be related to the level of dispute in a post [9]. As blogs comments are currently, to a large degree, not syndicated, extracting them is a laborious task, involving parsing the HTML contents and identifying the comments in it. However, since we estimate dispute by the presence of comments rather than by using their content, we can limit ourselves to identifying the number of comments instead of actually extracting them. To do this, we again opt for a simple approach; in this case, we actually take advantage of the fact that comments are not syndicated.

In the collection used at TREC, both the HTML content and the syndicated (RSS or Atom) content of each post are provided separately; this replicates the sources of data available to existing blog search engines. The difference between the length of the content in HTML format and the length of the syndicated content in XML format gives us the total amount of overhead found in the HTML version – layout information and content which is not directly related to the post: archive links, profile information, and so on. One of the types of content present in HTML but not in the syndication is the comments and trackbacks of the post. Generally, the amount of non-comment overhead involved is fixed over multiple posts in the same blog: the same archive links, blogger profile and so on appear in every permalink, and remain constant; the only component that changes substantially is the comment data. To exploit this, we first compute the average HTML to XML ratio for a blog; given a post, comparing its specific HTML to XML ratio to the average one gives us an indication of the amount of post-specific overhead associated with this post – which we view as reflecting the relative volume of comments of the post. This ratio, which we refer to as the "discussion level" of a post, is about 1 for posts which have the same amount of comments as other posts in the same blog, lower than 1 for posts which attract less discussion, and higher than 1 for posts with an above-average volume of responses. An additional benefit of this approach is that it assigns higher importance to longer comments – which typically indicate a more involved discussion.

More formaly, let p_{XML} be the syndicated content of a post p from blog B, and p_{HTML} the corresponding HTML content. Then the discussion level we calculate for a given post p' is

$$\text{discussion\_level} = \frac{\frac{|p'_{HTML}|}{|p'_{XML}|}}{\frac{1}{|B|}\sum_{p \in B}\frac{|p_{HTML}|}{|p_{XML}|}} \quad .$$

Table 1 compares the discussion level calculated this way with the manually extracted number of comments of a few posts in two different blogs.

We use the discussion level as a static prior for each post, assuming that posts which attract many responses are more likely to express an opinion. Again, this query-independent indicator for each post is combined, as our other partial scores, with the other indicators we use for the opinion retrieval task.

Post ID	Feed ID	Disc. level	Comments
20051221-003-0015014854	006792	0.6	0
20051207-022-0022756510	006792	1.5	1
20051207-022-0023018509	006792	2.1	2
20051207-029-0008800181	007168	0.8	3
20060125-004-0015081881	007168	1.0	7
20060215-005-0024698640	007168	2.1	16

Table 1: *Examples of discussion level values*

2.3 Query-dependent spam filtering

Link-based search engine spam has penetrated many domains on the web, and the blogspace is no exception. Spam blogs, or splogs, account for as much as 20% of all blogs, and significantly degrade the performance of blog search and analytics [12]. Currently, blog spam is in its relatively early stages; applying methods which have been developed for web spam to blogs achieves high success rates in identifying and filtering splogs [4]. We implemented a state-of-the-art spam filter for blogs based on support vector machine classification of the contents of the blogs as well as other features such as compressibility, and used it to rerank retrieval results. Sure enough, we experienced significant improvements, of about 9% in mean average precision, and up to 20% for early precision scores. However, we observed that the contribution of the spam filter to different queries varied substantially. Out of a total of 50 queries, spam filtering increased retrieval accuracy for 39 queries and decreased it for 10 (for the remaining query there was no change); Table 2 shows the TREC topics for which improvement was most substantial, and the topics where performance degraded the most. As expected, most topics where spam filtering is highly beneficial are commercially-oriented; topics where performance is degraded are assorted sports, politics, and miscellaneous ones.

	Average Precision		
Topic	Baseline	Spam filtered	Change
Highest performance gains			
883. heineken	0.1924	0.4055	(+110%)
893. zyrtec	0.0413	0.0826	(+100%)
885. Oprah	0.1241	0.2445	(+97%)
877. sonic food industry	0.0234	0.0459	(+96%)
Highest performance drops			
882. seahawks	0.0428	0.0373	(-13%)
871. cindy sheehan	0.4511	0.4014	(-11%)
892. "Jim Moran"	0.6152	0.5499	(-11%)
880. "natalie portman"	0.2332	0.2106	(-10%)

Table 2: *Performance changes when using spam filtering*

Given the difference in performance between queries, we hypothesized that if we could predict the degree to which a query is commercially-oriented, we may be able to improve average performance by increasing the importance of spam filtering for those highly commercial queries, and decreasing it for the non-commercial ones.

Sophisticated methods for deriving commercial intent of queries require manually annotated data as well as an analysis of the landing pages of queries (e.g., [1]); we choose instead a much simpler approach, which requires no training. To estimate the level of commercial intent behind a query, we measure the likelihood of observing the query terms in a commercial context; for this, we simply count the number of documents in which the query terms co-occur with a term highly correlated with commercial activity (we use the word "shop"), out of the total number of documents containing the query. This number, which we refer to as the query commercial intent value, is then used as a query-specific weight for spam reranking, instead of using a fixed, query-independent weight.

Table 3 lists the query commercial intent values of the most commercially-oriented of the TREC queries according to this method, and the least commercially-oriented ones. Some correlation with the most and least benefiting queries from Table 2 is already visible, indicating that this approach may indeed improve over treating all queries as equal for spam purposes.

Topic	Commercial intent
885. shimano	0.3675
877. sonic food industry	0.2353
859. "letting india into the club?"	0.1945
895. Oprah	0.1878
874. coretta scott king	0.0290
871. cindy sheehan	0.0295
892. "Jim Moran"	0.0310
897. ariel sharon	0.0322

Table 3: *Query commercial intent values*

3. Evaluation

We now describe the experiments used to evaluate the three proposed methods.

3.1 Experimental settings

All our experiments are based on the data used at the TREC 2006 Blog Track; the document collection, the TREC Blog06 corpus, is a crawl of more than 100,000 syndicated feeds over a period of 11 weeks, containing 3.2 million posts for a total size of 148GB (including both HTML and syndicated content). More information about this corpus is given in [6]. 50 queries were used at the track, with 63,103 blog posts being manually assessed for relevance. The evaluation metrics used are the standard TREC measures, i.e., mean average precision (MAP), R-Precision, and precision at 10 (P@10).

As a baseline model, we use the language modeling-based retrieval approach proposed by Hiemstra [3]; our experiments show that this approach outperforms, for this task, other state-of-the-art ranking models such as Okapi BM25 and Divergence from Randomness. This is a robust baseline: the median MAP score at TREC 2006 was 0.1371, while our baseline achieves a MAP of 0.1797; this places it, as-is and with no extensions to handle identification of opinionated content in the posts, among the top performers at TREC 2006 (all of which did use specialized opinion-detection modules [10]).

3.2 Results

Table 4 shows the improvements gained by each of the components we described. As noted earlier, each component is part of a larger set of heuristics aimed at addressing the three aspects of opinion retrieval we identify: topical relevance, opinion expression, and post quality. We list separately the contribution made by the component alone, by other components in the same group of techniques, and by the combination of this component with the rest of the components from the group. We also list the overall gain in performance achieved when using all components, along with the rest of

Method	MAP	R-Prec	P@10
Baseline	0.1797	0.2452	0.3560
Topical retrieval components			
Temporal relevance feedback only	0.1883 (+5%)	0.2587 (+6%)	0.3880 (+9%)
Other components	0.1999 (+11%)	0.2791 (+14%)	0.3980 (+11%)
Combined	**0.2056** (+14%)	**0.2835** (+16%)	**0.4020** (+13%)
Opinion expression components			
Discussion level only	0.1828 (+2%)	0.2498 (+2%)	0.3660 (+3%)
Other components	0.2198 (+22%)	0.2955 (+20%)	0.4300 (+20%)
Combined	**0.2271** (+26%)	**0.3011** (+23%)	**0.4400** (+24%)
Post quality components			
Spam filtering, fixed weights	0.1961 (+9%)	0.2633 (+7%)	0.4140 (+16%)
Spam filtering by commercial intent	**0.2019** (+12%)	**0.2703** (+10%)	**0.4200** (+18%)
All components	**0.2411** (+34%)	**0.3122** (+27%)	**0.4900** (+38%)

Table 4: *Contributions of separate components and their combination*

the heuristics we implemented. All results are statistically significant ($p < 0.05$) using the t-test, but it should be noted that retrieval improvements of less than 5% are considered unstable even when statistical significance is established [11].

Clearly, all of our approaches increase performance; in particular, the temporal relevance feedback and, to a lesser extent, the weighted spam filter, provide important contributions. While every approach described in this paper improves performance by a few percent only, their combination shows substantial improvements – indicating that, to a large extent, the approaches are orthogonal, and improve the ranking of different types of documents.

Along with additional, non blog-specific methods which address the opinion retrieval task (such as traditional sentiment analysis methods), we achieve a substantial gain over the baseline, as well as over state-of-the-art: the best-performing system at TREC 2006 obtained a MAP score of 0.2052.

4. Conclusions

We discussed three techniques for improving opinion retrieval in blogs, each based on a property of the blogspace and each implemented with a simple, inexpensive mechanism. These techniques correspond to different aspects of opinionated retrieval: to improve topical retrieval, we incorporated temporal information into the retrieval model via a blind relevance feedback approach. To improve identification of opinionated content, we estimated the level of discussion in a blog post using the ratio of HTML and XML content in it. Finally, we identified that some queries require more rigorous spam filtering and used a simple, but effective approach to assign varying levels of spam detection according to a query's commercial context.

Combined with other, more traditional techniques for the task of opinion retrieval, these approaches substantially improve over a baseline and over state-of-the-art. In future work, we intend to incorporate additional properties of the blogspace into the retrieval model; early experiments with linking traditional sentiment analysis methods with language properties typical to blogs (e.g., high concentration of pronouns) show promising results.

Acknowledgments

This work was supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001.

References

[1] H. K. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang, and Y. Li. Detecting online commercial intention (oci). In *WWW 2006*, 2006.

[2] F. Diaz and R. Jones. Using temporal profiles of queries for precision prediction. In *SIGIR*, 2004.

[3] D. Hiemstra. *Using Language Models for Information Retrieval*. Phd thesis, Enschede, 2001.

[4] P. Kolari, T. Finin, and A. Joshi. SVMs for the Blogosphere: Blog Identification and Splog Detection. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[5] X. Li and W. B. Croft. Time-based language models. In *CIKM'03*, 2003.

[6] C. Macdonald and I. Ounis. The trec blogs06 collection : Creating and analysing a blog test collection. Technical Report TR-2006-224, Department of Computing Science, University of Glasgow, 2006.

[7] G. Mishne. Multiple ranking strategies for opinion retrieval in blogs. In *TREC 2006*, 2006.

[8] G. Mishne and M. de Rijke. A study of blog search. In *ECIR 2006*, 2006.

[9] G. Mishne and N. Glance. Leave a Reply: An Analysis of Weblog Comments. In *WWE 2006 (WWW 2006 Workshop on Weblogging Ecosystem)*, 2006.

[10] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the trec-2006 blog track. In *TREC 2006*, 2006.

[11] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *SIGIR '05*, 2005.

[12] Umbria Inc. SPAM in the Blogosphere. white paper, 2006.