From Blogs to News: Identifying Hot Topics in the Blogosphere

Wouter Weerkamp Manos Tsagkias Maarten de Rijke

ISLA, University of Amsterdam
http://ilps.science.uva.nl/

Abstract: We describe the participation of the University of Amsterdam's ILPS group in the blog track at TREC 2009. We focus on the top stories identification task, and take an approach that does not require the headlines of top stories to be known beforehand. We explore the feasibility of a so-called blogs to news approach: given a date and a set of blog posts, identify the main topics for that date. This approach is more general than just finding top stories, but it can still be applied to the task of headline ranking. Results show that this general approach, applied to the task at hand, is among the top performing approaches in this year's TREC.

1 Introduction

This year's Blog track consisted of two tasks: *top stories identification* and *faceted blog distillation*. The latter task is very similar to the "regular" blog distillation task that ran during the previous two TREC years (2007 and 2008), except for the addition of "facets" (e.g. in-depth, opinionated, or personal). Because of the similarity between the faceted blog distillation task and previous tasks, we felt that our focus should be on the new top stories task, and we therefore dedicated most of our time and effort to submitting to the top stories identification task.

The second task, *top stories identification*, is new; the goal is to identify top stories for a given day using information from the blogosphere, and provide a listing of blog posts that support the selection of a top story. The underlying scenario is one of a news provider (in possession of news headlines) trying to rank these headlines based on what people write about news stories in their blogs. For the identification part, it calls for an approach described in the following steps:

1. construct a "query" from headline;
2. limit results to the given date;
3. count the number of relevant posts;
4. rank headlines based on these counts.

The steps above reveal two limitations: (i) headlines are needed in advance, and (ii) topics from the blogosphere can only emerge when they are about news events reported by mainstream media. In an effort to alleviate these limitations, we take on the task from a different angle:

1. observe posts from the given date;
2. see what differentiates these posts from previous posts;
3. display the emerging topics;
4. rank headlines by their similarity to the emerging topics.

Although the algorithm can stop one step short, the last step is designated to provide compatibility with the task at hand. In our participation we investigate the potential of both approaches and report on initial evaluation of the results. For the second step of the top stories identification task, namely, to provide evidence for the importance of a headline, we chose to select the top blog posts ranked by the number of their respective comments.

In the remainder of this paper we first describe the data and preprocessing for both tasks (Section 2), then, we introduce our top stories identification approaches (Section 3). We report on the performance of the submitted runs, and perform some additional analysis in Sections 4 and 5. Finally, we report on some initial conclusions for this year's Blog track participation in Section 6.

2 Data and Preprocessing

The dataset provided by TREC is the new Blogs08 collection; the collection consists of a crawl of feeds, permalinks, and homepages of 1,303,520 blogs during early 2008–early 2009. This crawl results in a total of 28,488,766 blogs posts (or permalinks). In our experiments we only used feed data, that is, the textual content of blog posts distributed by feeds (e.g. RSS) and ignored the permalinks. Two main reasons underly this decision: (i) the tasks (and especially the top stories task) are precision-oriented and benefit from a very clean collection; and (ii) using feed data requires almost no preprocessing of the data (e.g. no html-removal, etc.). Extracting posts from the feed data gave us a coverage of 97.7% (27,833,965 posts extracted). As a second preprocessing step we perform language detection and remove all

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
From Blogs to News: Identifying Hot Topics in the Blogosphere	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Amsterdam,Intelligent Systems Lab Amsterdam (ISLA),Science Park 107,1098 XG Amsterdam, The Netherlands,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009) held in Gaithersburg, Maryland, November 17-20, 2009. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**4**	

non-English blog posts from our corpus , leaving us with 16,869,555 blogs posts . We construct two indexes, one based on the full content of blog posts and one on only the blog post titles. Additional to the indexing, we extract features that can prove useful in both tasks. Extracted features are: number of comments, post length, number of spelling errors, number of shouted words, number of emoticons, and ratio of first person pronouns.

Part of the top stories task is a collection of 102,812 news headlines from the New York Times. We created a separate index of this collection, resulting in an average news headline length of 11 words. Finally, we have 55 dates designating the topics for the top stories task.

2.1 Significance testing

Throughout the paper we use a two-tailed paired t-test to test for significant differences between runs. We report on significant increases (or drops) for $p < 0.01$ using $^{\blacktriangle}$ (and $^{\blacktriangledown}$) and for $p < 0.05$ using $^{\triangle}$ (and $^{\triangledown}$).

3 Approaches

As explained in the introduction, we contrast two main approaches in identifying top stories: (i) starting from the news headlines, or (ii) starting from the blog posts. In our participation we explore the potential of both approaches and compare their results.

3.1 News to Blogs

Given the scenario where news headlines are known beforehand, they can be used as starting points for identifying top stories on a given date. As explained before, this scenario is limited, but definitely worth investigating. Our is approach is as follows: we want to estimate the probability of a news headline given a date, and rank news headlines based on this probability. We use an expert finding model from Balog et al. (2006) (more specific Model 2) and modify it to fit the data at hand. Although the model allows us to explicitly define a post's importance for a given date, we assume all posts to be equi-important (i.e., the probability of the post given the date is uniform).

We run the approach on both a post index (run **IlpsTSHlP**) and a title-only index (run **IlpsTSHlT**). The reason for using the title-only index is that we expect bloggers to use important (news) terms in their post titles, so that matching the headline to the title would result in acceptable rankings as well.

3.2 Blogs to News

Following the second scenario where the news headlines are unknown, we need to extract information from the blog posts without any prior knowledge of what is in the news. To this end, we take the top 5,000 blog posts from a given date, ordered by their respective number of comments. We then combine these posts and identify distinguishing terms between them and a background corpus. The background corpus consists of the remainder of the blog posts. These steps (covering steps 1 and 2 from the introduction) result in a set of weighted terms, where the weight indicates a term's "distinctiveness" for the given date. Based on co-occurrence statistics, the terms are clustered, leaving us with the topics that emerge from the blog posts. Up to here our approach is general as no news headlines are required. An example of the generated output, is shown in Table 1.

Terms	**News event**
ledger heath actor	Actor Heath Ledger dies
roe abortion	Roe v. Wade case on abortion; March for Life 2008
romney mitt huckabee thompson gop ...	Republican primaries 2008
luther martin king african dr	Martin Luther King Day 2008

Table 1: Example of top emerging terms (left) and related news events (right) for January 22, 2008.

To use this approach for the task at hand, we need a way of matching the extracted information to the news headlines. We index the news headlines and use the extracted term clusters to query the headline index. Headlines are ranked based on the distinctiveness of the terms, and if more than one headline matches a query, we select a maximum of 10 headlines for this "topic".

As with the previous approach, we run this on both a post index (**IlpsTSExP**) and a title-only index (**IlpsTSExT**). Here, we expect the title-only representation to contain less noise (less indistinctive terms) and therefore be able to better get the important terms on top.

4 Results

The results of our submitted runs are displayed in Table 2.

The top two lines represent the two approaches on the post index and the lower two lines on the title-only index. The first observation is that the blogs to news approach significantly outperforms the news to blogs approach on all metrics and for both indexes. Looking at each approach individually, we see that for the news to blogs approach the difference between the two indexes is not significant. For the blogs to news approach the performance of the post index is significantly better than the title-only index for MAP and MRR.

Comparing results to other participants in the blog track,

Approach	MAP	P5	MRR	RunID
b-to-n (p)	**0.1354**▲	**0.2655**▲	**0.4271**▲	IlpsTSExP
n-to-b (p)	0.0083	0.0291	0.1119	IlpsTSHlP
b-to-n (t)	0.0756▲	0.2036▲	0.2670▲	IlpsTSExT
n-to-b (t)	0.0085	0.0545	0.0958	IlpsTSHlT

Table 2: Results of our submitted runs of top stories identification task for the blogs to news (b-to-n) and news to blogs (n-to-b) approaches on a (p)ost index or (t)itle index.

we are in the top 3 performing runs on all reported metrics (MAP, P5, and MRR). Table 3 shows the results of our best run compared to the best run at TREC 2009, and the median of all participants. Interesting to see is that the two other top performing runs at TREC used our news to blog approach, which indicates that our run of this type could still improve. Nevertheless, our more general approach is reasonably close to the best news to blogs runs, while maintaining its general nature.

Run	MAP	P5	MRR
Best (uogTr)	0.1862	0.3236	0.5390
Median	0.0445	-	-
b-to-n (p)	0.1354	0.2655	0.4271

Table 3: Results of our best run, the best run at TREC 2009, and the median of all participants.

5 Analysis

We perform some initial analysis on our submissions, and do so in three ways: (i) looking at the ordering before sampling blog posts, (ii) exploring how large our sample of blog posts should be, and (iii) detailing two dates for which our approach either works well or does not.

As detailed in Section 3, the blogs to news approach depends on a sample of blog posts for a given date. Sampling these posts from the set of posts for a day can be done in various ways; in our baseline we order posts by the number of comments they receive, and take the top 5,000 posts for a day. We now look at the influence of other orderings (length of a post, and random sampling), and the influence of different sample sizes. To start with the first question, Table 4 shows the results of the three ways of ordering posts before selecting them.

The results show that ordering by post length can lead to improved early precision, but does hurt MAP slightly. Choosing a (reasonable) way of ordering improves over selecting just random posts as sample. Now we shift to the size of the sample: Here we experiment with three different sample sizes to show how this influences the results. Note that we are not looking for an optimal value, but merely want to see the behavior of our approach. Table 5 shows the results for using ordering on comments, and taking the top 500, 5,000, and 50,000 posts.

Ordering	MAP	P5	MRR
Comments	**0.1354**	0.2655	0.4271
Post length	0.1145	**0.2982**	**0.4959**
Random	0.1041	0.2327	0.3892

Table 4: Results of various ways of ordering posts before selecting them.

Sample size	MAP	P5	MRR
500	0.0418	0.1709	0.3057
5,000	**0.1354**	**0.2655**	**0.4271**
50,000	0.1050	0.2145	0.4057

Table 5: Results of various sample sizes.

We can tell from the results that it is probably better to take a larger sample than a smaller one: Results drop significantly going from 5,000 to only 500 posts. If we take more posts into account, we also observe a drop in performance, but not as big as when going down in sample size. Further research is required on how we can determine the optimal number of posts in our sample.

Finally, we look at two dates for which our approach showed either good or bad performance: First is topic 30, August 8, 2008. Table 6 shows the extracted topics on the left, and the headlines that where matched against this topic (and judged relevant) on the right. Similarly, we list the data for topic 19 (May 12, 2008) in Table 7.

From these two topics we can conclude three things: (i) our approach works well for major events in the news (Georgia-Russia conflict, opening of Olympics), but also for smaller, more local events (Brett Favre joining the New York Jets), (ii) the output generated by the approach is usable in more settings than just headline ranking, and (iii) topics that come up are not always news related (Mother's Day).

6 Conclusions

This year we focused on the new top stories identification task: use the blogosphere to rank news headlines. We follow two general approaches: news to blogs, and blogs to news. The former starts from the news headlines, uses them as queries, and ranks these queries according to the headline likelihood. The latter is more general and tries to identify topics that emerge from the blogosphere. It is only in the final step that this approach tries to link these topics to news headline (by using the topics as a query against a headline index). In our experiments, the blogs to news approach that is independent of a given day's news headlines outperforms

Terms	Relevant headlines
ossetia georgian georgia russia russian conflict region troops	Russia and Georgia clash over breakaway region Georgian troops enter breakaway enclave in region's fiercest fighting in years
favre jets	From Green Bay to Broadway: Favre is a Jet Sports of the times; In Favre the Jets have a tiger by the tail Jets 24, Browns 20; As Favre sits, another Brett is impressive
edwards affair hunter abc elizabeth	Edwards admits to affair in 2006
olympics beijing olympic athletes ceremony china opening sports chinese games coverage	Games open in a new China, dazzling an age of new media Olympic message to some in Beijing is "please leave"

Table 6: Example of top emerging terms (left) and judged relevant headlines (right) for August 8, 2008.

Terms	Relevant headlines
mothers moms mother mommy recipe flowers dish egg garden sauce mom ...	-
nba	-
hezbollah	Hezbollah begins to withdraw gunmen in Beirut after 4 days of street battles
earthquake	-

Table 7: Example of top emerging terms (left) and judged relevant headlines (right) for May 12, 2008.

the news to blogs approach on all reported metrics and the difference in performance is measured statistically significant. Other participants in the blog track however, showed that the news to blogs approach can be successful as well.

Further exploration of the system parameters revealed that using a title-only representation of blog posts does not lead to improvement, neither on recall-based metrics (as expected) nor on precision-based metrics. Both, the ordering of blog posts before selection and the sample size, make a difference in performance of the blogs to news approach. Finally, the examples show that our approach can (i) identify major and minor news events using blog posts, (ii) construct output that is more general than just headlines, and (iii) identify hot topics that are not news related.

Future work focuses on exploring the optimal sample size of blog posts, applying more elaborate models to the top stories identification task and see how we can use additional (external) information to identify emerging topics, or use explicit links and references to news events for this task.

7 Acknowledgments

This research was supported by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (`http://www.stevin-tst.org`) under project number STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, and 640.004.802,

8 References

Balog, K., Azzopardi, L., and de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In *SIGIR'06*.