News Article Ranking: Leveraging the Wisdom of Bloggers

Richard M. C. McCreadie
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ
richardm@dcs.gla.ac.uk

Craig Macdonald
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ
craigm@dcs.gla.ac.uk

Iadh Ounis
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ
ounis@dcs.gla.ac.uk

ABSTRACT

Every day, editors rank news articles for placement within their newspapers. In this paper, we investigate how news article ranking can be performed automatically. In particular, we investigate the blogosphere as a prime source of evidence, on the intuition that bloggers, and by extension their blog posts, can indicate interest in one news article or another. Moreover, we propose to model this automatic news article ranking task as a voting process, where each relevant blog post acts as a vote for one or more news articles. We evaluate this approach using the TREC 2009 Blog track top news story identification task judgments, showing strong performance in comparison to TREC systems, as well as two alternative baseline rankings. Furthermore, to increase the accuracy of the proposed approach, we examine temporal re-ranking techniques, corpus cleaning of inappropriate articles and article expansion to counter vocabulary mismatch. We conclude that, overall, blog post evidence can be a useful indicator to a news editor as to the importance of various news stories, and that our approaches for extracting this evidence are effective.

General Terms: Ranking, Information Retrieval, Social networks

Keywords: News, Blogs, Top, Stories, Identification

1. INTRODUCTION

Over the last few years, news reporting has undergone a profound shift from paper-based media, to free online publications [27]. This shift has brought with it unprecedented volume and diversity in news articles, as the barriers to publication are much lower for e-newspapers [20].

From an editorial perspective, the number of possible stories to report on has increased dramatically, while the increased competition and ease of access to alternate providers has given emphasis to the task of identifying the news stories that are important enough to be placed on a given content page (e.g. the front page) of the news website. Similarly, at a meta-level, news aggregators [15, 21] face an even greater challenge. News aggregators give users access to broad perspectives on the important news stories being reported, by grouping articles into coherent news events. However, deciding automatically on which important stories to show is an important problem with little research literature.

We investigate how news articles can be automatically ranked by readership interest for proper placement within a newspaper or any other news source. In particular, we address the following task: given a list of news articles and a day of interest, rank these articles such that they are ordered by importance for that day.

Importantly, this can be interpreted as either a real-time or a retrospective task [32], i.e. whether ranking is done in real-time on a given day using only historical evidence, or is done retrospectively using additional (future) evidence from after the given day. From an editorial point of view, the real-time task is the most relevant to the ranking of current news articles. However, the only currently available test collection for identifying top news is a TREC 2009 dataset with a retrospective nature, and hence, in this paper, we only experiment with the retrospective case.

The blogosphere is well known to respond to various newsworthy events [30]. For instance, bloggers may discuss a day's news as it breaks, or may even break news themselves [17]. We propose to leverage the blogosphere to gain insights into the most important news stories of a given day. In doing so, we build upon the assumption that the blogosphere represents a realistic sample of a readership's interest into the most important news stories. Hence, by measuring the response of the blogosphere to a real world event, a newspaper or website can gain an automatic insight into the most important news articles.

The contributions of this work are four-fold. Firstly, we show how a ranking of the most important news articles can be derived using the blogosphere. Furthermore, we show how historical or future evidence (before or after the event) can be used to improve our initial article ranking. Moreover, we perform a thorough investigation using a test collection for top news story identification, developed as part of the TREC 2009 Blog track [13]. Lastly, we examine how modification to the TREC 2009 news article corpus can facilitate the creation of more accurate rankings. In particular, we examine corpus cleaning of inappropriate articles, in addition to article expansion as a counter to vocabulary mismatch.

The remainder of the paper is structured as follows. Section 2 discusses the editorial role of news article placement, while Section 3 explores the blogosphere's response to real-world events. Section 4 proposes approaches to rank news articles, by modelling the blogosphere's reaction to news, while Section 5 describes our research questions and experimental setup. In Section 6, we evaluate our model for top news stories identification, while Section 7 explores the application of historical and future evidence to our model. Section 8 evaluates our article corpus refinement techniques, while in Section 9 we combine the best of the aforementioned techniques. Lastly, in Section 10 we provide concluding remarks, along with directions for future work.

RIAO'2010, April 28-30th, 2010, Paris, France

2. IMPORTANT NEWS

Newspapers, as a form of distributing information about current events, have been around since the advent of the mechanical printing press, in the early 18th century [3]. Newspapers typically employ a person in an *editorial* role, to oversee the organisation and selection of news stories written by their journalists or obtained from newswire services, into the final layout most likely to interest their readership. Usually, an editor will place on the front page of their newspaper the most important stories of the day based on some pre-defined criteria. For example, some classical news criteria [31] often considered are:

- Timing - important topics are usually new, or at least current.
- Significance - the number of people affected by a news story will have a bearing on its newsworthiness.
- Proximity - geographically, the nearer a news story to the readership, the more important its bearing.
- Prominence - events happening to celebrities, politicians or other famous people are more newsworthy.

Indeed, in [5], these and other criteria were studied from a story selection perspective, showing that such factors greatly impact the chances of a story being reported. However, despite different readership demographics and interpretations of these criteria, newspaper editors still choose similar stories to grace their front-pages on any given day [31]. This indicates that there are some stories which need to be reported on regardless of newspaper orientation. Indeed, such stories may cover significant *predictable events*, like elections, or prominent *unpredictable events*, such as the death of a celebrity. Naturally, the automatic detection of these events would be useful for editors.

The advent of the Internet has dramatically changed the face of news distribution. Traditional (physical) newspaper circulations are now falling [10], with many newspapers creating online presences, carrying the same stories, but supported by electronic advertising revenues. Moreover, other news websites are Web-only, either by the elimination of their physical newspapers, or due to being Web-only from the outset.

However, in such *e-newspapers*, the role of the editor has changed little. Indeed, they must still select appropriate news stories to place on the front-page or category pages (sport, technology, etc.), that address the most important issues of the day. Indeed, to maintain their readership, all the important stories need to be covered.

Furthermore, various news aggregator websites have been developed, which aim to display summaries of the main news of the day. Some news aggregators may link to a news source for a given story that highlights their favoured viewpoint (e.g. liberal or conservative). Other news aggregators might instead provide links to multiple, diverse news sources for a single story. News aggregators can be manually controlled, with an editor selecting the stories and links to appear. Here, the Drudge Report[1] and the Huffington Post[2] are classical examples. In contrast, Google News[3], NewsBlaster [15], NewsInEssence [21] and NewsExplorer [28], are examples of news aggregators which automatically group news articles into stories, and algorithmically select the most important news stories to display on their front page. However, the algorithms that drive commercial automatic news aggregators have not been the subject of much research dissemination.

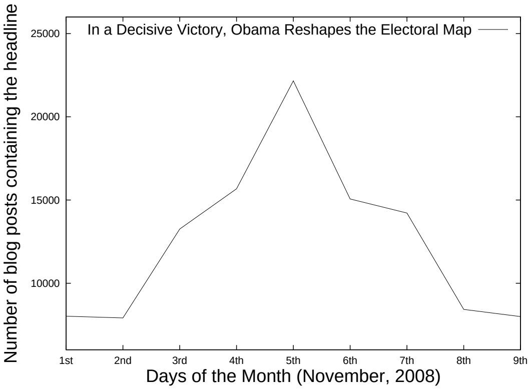

Figure 1: Distribution of blog posts retrieved for the headline 'In a Decisive Victory, Obama Reshapes the Electoral Map' for 9 days in November 2008.

This paper proposes automatic methods for determining the most important news articles on a day of interest, from a given set of news articles. In Section 3, we describe the blogosphere, and how it responds to newsworthy events. Later in Section 4, we propose our model for ranking news articles by their importance on a given day, making use of *user-generated content* (UGC).

3. NEWS ON THE BLOGOSPHERE

The blogosphere as a whole is a prime example of user-generated content. Specifically, the term *blogosphere* refers to all of the blogs on the World Wide Web (Web). The term blog is a contraction of the word 'weblog', which describes the act of someone using the Web to record their thoughts on a particular subject. A single blog contains one or more blog posts in chronological order, where each blog post is normally a statement of opinion or viewpoint on a given subject by the blogger. The popularity of blogs has increased exponentially [26] in recent years. Indeed, Technorati[4] reported tracking over 112.8 million English blogs in 2008.

With such a large volume of blogs being updated, it is intuitive that some proportion of these are news-related. Recently, a poll by Technorati has shown that 30% of their respondants blogged on news related topics [29], while work by Mishne and de Rijke [16] showed a strong link between blog searches and recent news - indeed almost 20% of searches for blogs were topical news-related, indicating that topical news is popular in the blogsphere. Moreover, Thelwall explored how bloggers reacted to the London bombings [30], showing that bloggers respond quickly to news as it happens. Furthermore, both König et al. [8] and Sayyadi et al. [25] have exploited the blogosphere for event analysis and detection, showing that news events can be detected within the blogosphere.

As a further illustration of the blogosphere's response to news events, Figures 1 and 2 show the number of blog posts retrieved using the TF_IDF weighting model for two news headlines from the Blogs08 [13] corpus over time. In particular, in Figure 1, it can be seen that the number of blog posts discussing Obama's victory in the U.S. election peaks on the day of the victory announcement (5th). Although the blog post distribution does not 'peak' for the financial story in Figure 2, the trend is centred around the 14th-20th of September (the news article was published on the 17th).

In the following, we assume that as a whole, bloggers have an interest in the current events, and therefore blog about important news stories. We hypothesise that by taking the pulse of the blogosphere into account, we can accurately predict which are the

[1] `http://www.drudgereport.com/`

[2] `http://www.huffingtonpost.com/`

[3] `http://news.google.com/`

[4] `http://technorati.com/`

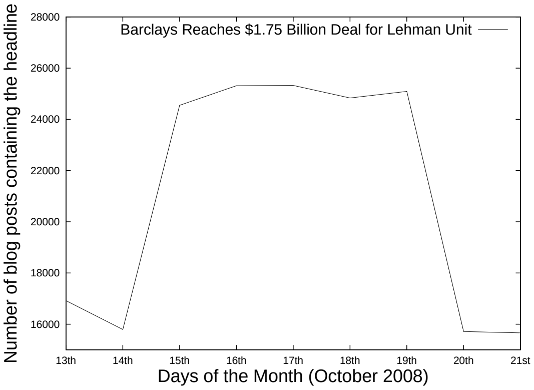

Figure 2: Distribution of blog posts retrieved for the headline 'Barclays Reaches $1.75 Billion Deal for Lehman Unit' for 9 days in October 2008.

important news articles on a given day. Importantly, we assume that the blogging population as a whole is a representative sample of newspaper readership, and hence we can treat bloggers' interest in a news article as an indication of readership interest. To support this assumption, we offer a comparison between current blogger demographics and that of the traditional (physical) newspaper readership[5].

We compare and contrast the demographics of bloggers and newspaper readers, to show that the overall response of the blogosphere to news stories are likely to be similar to that of a classical news audience. In particular, we compare both the educational and age demographics for traditional news and blogs. In terms of educational attainment, [29] shows that approximately 75% of bloggers have attained a college/university degree, which compares favourably with 61% of newspaper readers [18]. From an age perspective, the majority of bloggers (70%) are between the age of 25 to 54 [29], while in 2005, the average age of a newspaper reader was 53. While this suggests that the newspaper readership is significantly older than the blogging population, the fact that 35% of a sample of bloggers were reported to have worked with traditional news media suggests that many bloggers are likely to have an understanding of what makes interesting news [29].

Overall, we suggest that the blogosphere represents a viable evidence source for predictions on the most important news stories for a given day. In the next section, we propose novel models that use historical and future blog post volume evidence to rank the news articles of a day with respect to their predicted importance. Later, we provide experiments to measure the effectiveness of these predictions, when compared to an editorial viewpoint of news story importance.

4. MODELS FOR NEWSWORTHINESS FROM USER-GENERATED CONTENT

In this work, we tackle the problem of ranking news articles with respect to their predicted importance on a given day. Such important news articles are those which describe the main newsworthy stories of the day. In particular, for a given day of interest (which we call a "query" day and denote d_Q), we wish to score each news article a by its predicted importance, $score(a, d_Q)$. In Section 4.1, we describe our Votes approach to this problem, while in Section 4.2 we propose an enhancement which takes into account evidence on days other than d_Q.

Table 1: A sample assignment of votes for two articles (a_1 and a_2) over 3 days (d_1 to d_3).

	d_1	d_2	d_3
a_1	4	4	2
a_2	1	8	1

4.1 The Votes Approach

Our proposed approach is based on the intuition that, on any day, bloggers will create posts pertaining to prominent news stories for that day. We wish to use this evidence to rank news articles, such that those articles describing the most important news stories of the day are ranked highest. Specifically, we hypothesise that the volume of blog posts sharing content with a given article is indicative of the importance of the story that article covers according to the blogosphere, and as such can be leveraged for ranking.

To measure the blogosphere's response to a news article on any given day d, we count the number of related blog posts to that article which were published on day d. In particular, we use some textual representation of the news article (e.g. headline), denoted a, as a query. Then, an information retrieval (IR) system selects some blog posts $R(a)$ which are topically related to news article a. Let $R(a, d) \subset R(a)$ denote the subset of blog posts in $R(a)$ that were published on day d. We believe that $R(a, d)$ can be seen as a set of votes for article a to be important on day d. Hence, by counting the number of votes for article a (i.e. $|R(a, d)|$) we can estimate the article's importance on day d. Consequently, using this voting-like approach [11], which we refer to as *Votes*, the score for each article a on day d ($score(a, d)$) is calculated as:

$$score_{Votes}(a, d) = |R(a, d)| \tag{1}$$

Hence, to build the final ranking of articles for the query day d_Q, we compare the number of votes for all articles published on day d_Q, i.e. we rank by $score_{Votes}(a, d_Q)$.

Note that this is a three-stage process. Firstly, for each news article a, we score every day d by the number of blog posts voting for day d. In particular, the set of blog posts $R(a)$ are obtained by issuing a representation of article a as a query to an IR system. Next, we count the blog posts voting for each article's importance on each day d. Lastly, we rank all articles by the number of votes they received on the day of interest d_Q.

To illustrate our approach, we give a short example of Votes for 3 days (d_1 to d_3) and two news articles (a_1 and a_2). For each article, a ranking of the top 10 blog posts is analysed, and then votes are assigned to the various days. Table 1 shows, for each article (row), the number of votes for each of the 3 days. From this example, we can see that a_1 received 4 votes on days d_1 and d_2, in contrast to only 2 votes on d_3. Article a_2 obtained 8 votes on d_2 but only 1 on days d_1 and d_3. Hence, to find the most important news article on $d_Q = d_1$ (first column), a_1 would be ranked higher than a_2, since a_1 received more votes. Similarly, for $d_Q = d_2$ (second column), a_2 would be the most important news story.

Interestingly, *Votes* can be applied for both the real-time or retrospective task. In the retrospective case, the blog posts in $R(a)$ are retrieved from the entire blog post corpus, while for real-time article ranking, only the blog posts published on day d_Q or before are available. As discussed in the introduction, we evaluate *Votes* only in a retrospective mode of operation, due to the retrospective nature of our TREC 2009 evaluation dataset, leaving a full analysis of *Votes* in real-time news article ranking for future work. However, in the following section we propose the application of historical and/or future evidence for refinement of our initial *Votes* ranking, of which refinement using only historical evidence is applicable to the real-time ranking task.

[5] Note that we use traditional newspaper demographics, as to our knowledge, there is no freely available demographic data for e-newspaper readership.

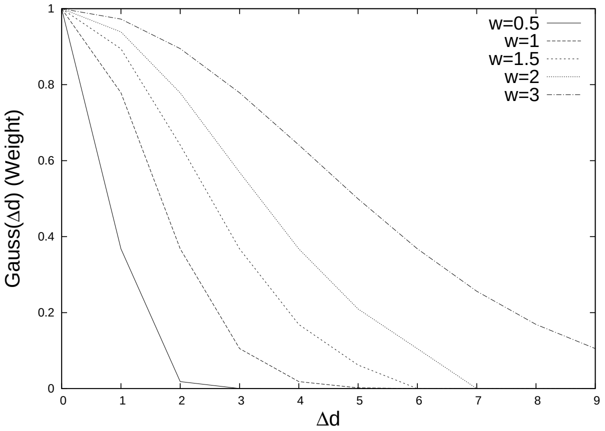

Figure 3: Example of Gaussian curves with varying values of w (the Gaussian curve width). Note that for illustration clarity, weights have been normalised so that day d_Q will always be assigned a weight of 1.

4.2 Temporal Promotion

Thus far, we have made use of blog post evidence from only the day of interest. However, historical evidence, or future blog postings may also help to improve the accuracy of our *Votes* approach. Our intuition is that news stories will often be discussed beforehand for predictable events, e.g. speculation about election results, and/or discussed afterwards for long running, controversial or important unpredictable stories, e.g. the aftermath of a terrorist bombing. Indeed, by taking this evidence into account, we can identify those stories which maintain their interest over time, and as such can be deemed more important. In particular, [7] suggested that bursts in term distributions could last for a period of time. Hence, in the following, we define two alternative techniques for calculating $score(a, d_Q)$, which leverage the *temporal distribution* of each article a over time. In particular, these techniques accumulate vote evidence from the days preceding or following d_Q, to 'boost' the score of articles which retain their importance over multiple days.

In our first proposed temporal distribution boosting technique, $NDayBoost$, we linearly combine the scores for the following n days before or after day d_Q, as:

$$score_{NDayBoost}(a, d_Q) = \sum_{d=d_Q}^{d_Q+n} |R(a, d)| \quad (2)$$

where $|R(a, d)|$ measures the importance of article a on day d_Q. n is a parameter controlling the number of days before ($n < 0$) or after ($n > 0$) d_Q to take into account, while d represents any single day. Note that this technique places equal emphasis on all days d - we expect that the distribution of $|R(a, d)|$ to peak around day d_Q.

Importantly, this approach can incorporate evidence from multiple days. However, due to the linear nature of the score aggregation, all days are treated equally, when it is intuitive to think that days more distant from d_Q will provide poorer evidence.

To address this, we propose a second temporal distribution boosting technique. In particular, $GaussBoost$ is similarly based upon the intuition that important stories will run for multiple days. However, instead of judging each subsequent day equally, we weight based on the time elapsed from the day of interest d_Q, using a Gaussian curve to define the magnitude of emphasis. In this way, we state a preference for stories that were important on days close to d_Q, rather than stories which peaked some time before/after d_Q:

$$score_{GaussBoost}(a, d_Q) = \sum_{d=d_Q}^{d_Q+m} Gauss(d - d_Q) \cdot |R(a, d)| \quad (3)$$

where m is the maximum number of days before d_Q (i.e. $m < 0$) or after d_Q ($m > 0$) to take into account and $d - d_Q$ is the number of days elapsed since the day of interest d_Q, denoted Δd. $Gauss(\Delta d)$ is the Gaussian curve value for a difference of days Δd, as given by:

$$Gauss(\Delta d) = \frac{1}{w.\sqrt{2\pi}} \cdot exp\frac{-(\Delta d)^2}{(2w)^2} \quad (4)$$

where the parameter w defines the width of the Gaussian curve. A small w will emphasise stories that were important on days very close to d_Q, while a larger w will take into account stories on more distant days, up to the maximum m days. In general, the larger the distance from d_Q (Δd), the lower the weight assigned to the evidence from that day. Figure 3 shows five sample Gaussian curves with different values for w. As we can see, when we increase w, evidence from days further from the day d_Q are taken into account. In particular, if $w = 0.5$ then only stories from the first day after d_Q receive additional importance, while a w value of 3 promotes stories from the 9 following days in a diminishing fashion.

The advantage of this approach over $NDayBoost$ is that we can control the weight placed on days other than d_Q, thereby avoiding over-emphasising stories on days which are unlikely to be useful. However, the disadvantage is that we are assuming that the Gaussian distribution is a good model of the way the evidence will diminish, which may not be the case for all stories. Indeed, we believe that this will be a promising area for future research.

5. EXPERIMENTAL SETTING

In this section, we detail the research questions to be investigated, as well as describe our experimental setup and define the baselines approaches we compare to. Specifically, in Section 5.1 we define three research questions, which will be later addressed. In Section 5.2, we describe the test collection employed, in addition to the weighting models we use, while Section 5.3 presents the baselines we compare our *Votes* approach to.

5.1 Research Questions

In the following experiments, we test three main research questions:

- Can the volume of relevant blog posts published on day d_Q provide an effective indicator of an article's interest to the e-newspaper readership on d_Q (Section 6)?
- Can further evidence on an article's importance be gained through analysis of the temporal distribution of the blog post volume both before d_Q and after d_Q (Section 7)?
- Can we improve our article rankings by refining the textual representation of each news article, in addition to the cleaning of inappropriate news articles from the corpus (Section 8)?

5.2 Experimental Setup

To address the above research questions, we experiment within the context of the Blog track at TREC 2009. In particular, the Blog track introduced a new top news stories identification task (*news task*), with the aim of ranking the most important news stories for a given day. Participants were asked to rank a set of news articles

Table 2: Salient statistics for the Blogs08 and New York Times news article corpora used during the TREC 2009 Blog track, top news stories identification task.

Corpus	Quantity	Value
Blogs08	Number of blog posts	28,488,766
	Blog post corpus timespan	14/01/08 to 10/02/09
NYT News Articles	Number of articles	102,853
	Mean articles per day	264.3
	Article corpus timespan	01/01/08 to 28/02/09

provided by the New York Times (NYT), by using evidence from the Blogs08 corpus of blog posts [13]. Evaluation was performed for 55 query days, where NYT articles for each query day have been manually judged from an editor's perspective as important or not - i.e. would have been placed on a 'front page'. Table 2 lists the salient statistics for the Blogs08 and NYT headline corpora. Importantly, the news task is retrospective in nature, hence, to use this test collection, we similarly evaluate our Votes approach in a retrospective mode of operation (see Section 4.1). Furthermore, we note that for the TREC 2009 news task, only headlines were provided as textual representations of each article.

In terms of experimental settings, we use the Terrier [19] information retrieval platform to index the Blogs08 corpus of blog posts, removing standard stopwords, and applying Porter's English stemmer. To generate the ranking of blog posts with respect to a news article $R(a)$, we test two effective weighting models to determine the suitability of each. In particular, we test with the probabilistic BM25 [23] and with DPH from the Divergence from Randomness framework [2]. Note that for the following experiments, we fix the size of $|R(a)|$ (the number of blog posts to return) to 1000, based upon recommendations in [11].

Finally, to make our results comparable with the systems participating in the TREC news task, we use the default parameters for the weighting models employed, as no training data was available for Blogs08 at the time the task was run. In particular, we use the default parameters for BM25 of $k1 = 1.2, k_3 = 1000, b = 0.75$ [22]. DPH on the other-hand, is a 'parameter-free' model, where all parameter values are derived from the collection statistics.

5.3 Baselines

In this paper, we compare our results to several baselines. Firstly, we validate our approach against the per-topic median of the systems participating in the news task. Next, we devise two simple baselines. In the first, we rank each news headline by the total number of blog posts that link to its corresponding news article on the query day (d_Q) – we denote this approach as Inlinks. Notably, such an approach was deployed in TREC 2009 [13]. Importantly, Inlinks differs from Votes in the way that the 'voting' blog posts are selected. Instead of leveraging the textual content to determine the relevant blog posts to the headline, Inlinks uses hyperlink evidence, on the assumption that the explicit inclusion of a link is a strong indicator of relevance. However, we suspect that this approach may be compromised due to sparsity of hyperlink evidence, i.e bloggers often will not link back to the story that they discuss. Secondly, we create a random permutation of the news articles on each day of our evaluation (a permutation of 100 articles are randomly selected for each day). Note that to avoid the possibility of outlier results, we report the mean over 10 attempts. This baseline is denoted Random. Lastly, we compare to the best systems participating in TREC 2009. For comparison to various baselines, statistical significance is measured using the Wilcoxon Signed-Rank test ($p < 0.01$).

Table 3: Effectiveness of Votes, in comparison to the TREC 2009 median, as well as Inlinks, Random and the best TREC systems. Symbols *, † and α denote significant improvements over the TREC median, Inlinks and IlpsTSExP, respectively (Random is a mean of multiple runs). No significant improvements over uogTrTStimes and KLEClusPrior were observed.

Run	MAP	P@10
TREC median	0.0450	0.1164
Inlinks	0.0601	0.1073
Random	0.0539	0.1867
uogTrTStimes	0.1862	0.3236
KLEClusPrior	0.1605	0.2804
IlpsTSExP	0.1354	0.2655
BM25+Votes	0.1731*†α	0.3145*†
DPH+Votes	0.1742*†α	0.2945*†

6. EXPERIMENTS USING BLOG POST VOTES

Initially, to evaluate the effectiveness of our voting approach for the automatic news article ranking problem, we experiment and compare to the described baselines. Results are reported in Table 3. For both mean average precision (MAP) and precision at 10 (P@10) measures, we report the performance achieved by the median of the TREC 2009 participating systems, as well as the performance of the Inlinks and Random baselines, and that of the 3 highest performing TREC 2009 submitted systems. From the reported performances, we note that the TREC median for this task is lower than both Random and Inlinks.

Table 3 also reports the performance of our proposed voting approaches, using both the BM25 and DPH weighting models. We note that combining the voting approach with both BM25 and DPH results in a large and statistically significant improvement over the TREC median and Inlinks baselines in terms of MAP and P@10. This disparity in performance between Inlinks and Votes indicates that many more blogger's will discuss a news story than will link back to the appropriate NYTimes article. Furthermore, we also note that our approaches outperform Random by a large margin.

Moreover, Table 3 shows how our approaches compare to the three best TREC participating systems. Firstly, we note that the uogTrTStimes system is based upon similar voting approach [14]. Secondly, we observe that our initial DPH+Votes and BM25+Votes approaches outperform the other best TREC 2009 systems, sometimes by a statistically significant margin (e.g. IlpsTSExP).

The fact that our proposed approach markedly outperforms the baselines, allows us to ascertain that using blog post volume is a useful indicator of news article importance from an editorial perspective. Additionally, this validates our initial assumption that a bloggers interests are similar to that of an newspaper's readership. Furthermore, the results show that our proposed voting approach can reasonably identify important news articles. In the next section, we examine how taking temporal distributions into account can improve the performance of our Votes approach. Note that as the DPH weighting model produces the highest MAP performance, we will use DPH+Votes as our new baseline in the remainder of this paper.

7. TEMPORAL DISTRIBUTION TECHNIQUES

In this section, we investigate the promotion of evidence from days other than the day of interest d_Q to improve news article ranking performance. Recall that our voting approach selects each news

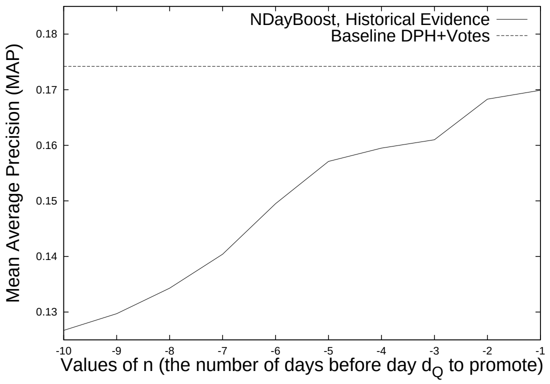

Figure 4: *Votes* MAP performance for NDayBoost using up to 10 days of historical blog post evidence.

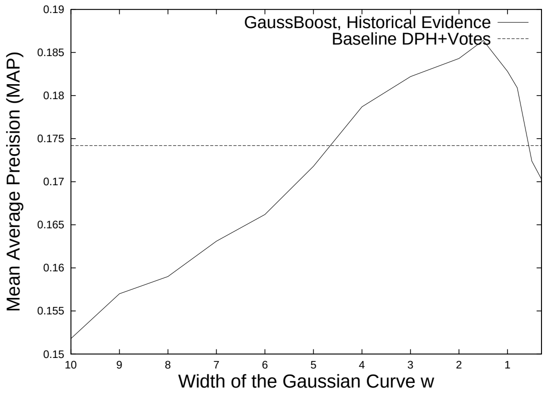

Figure 5: *Votes* MAP performance for GaussBoost when varying w using historical evidence. Note that for consistency, this graph is reflected to show that we are using evidence before d_Q.

article based on the number of blog posts voting for that news article to be important on that day. However, we believe that important news will be discussed in the blogosphere before the event, or will continue to be discussed after the event. Indeed, we hypothesise that useful evidence on a news article's importance can be gained from the blog post volume on the days before and after d_Q, and moreover, that this can be used to improve our baseline (DPH+Votes) ranking. To investigate this, we experiment using our two temporal distribution techniques, NDayBoost and GaussBoost, as described in Section 4. Both are tested using historical blog post evidence (Section 7.1) and future blog post evidence (Section 7.2).

7.1 Historical Temporal Evidence

Firstly, for historical temporal promotion, we promote news articles which appeared to be important on days before d_Q. Initially, we test the NDayBoost, for various values of n, $-10 \leq n < 0$, i.e. using up to 10 days of evidence before day d_Q. Figure 4 shows the MAP performance of NDayBoost, for different values of n. We note that as n decreases (to the left), performance also decreases. Moreover, performance stays consistently beneath our baseline as shown by the horizontal dashed line. This initially suggests that there is either no useful historical evidence, or that this simple approach is not sufficiently sophisticated to effectively make use of this evidence.

Next, we test the effectiveness of our GaussBoost technique on historical blog post evidence. Note that, for this technique, instead of varying m, the number of days of historical evidence to use, we instead vary the w parameter, which has a resulting effect on the number of days of evidence utilised (see Figure 3). Figure 5 shows the MAP performance of GaussBoost for various w values. In contrast to NDayBoost, we note that for values of $0.5 < w < 5$, effectiveness is enhanced over the performance of the baseline alone. Indeed, for $w = 1.5$, this represents a statistically significant increase in MAP of 6% ($p < 0.01$). Recall that w is not measured in days, e.g. a w value of 1.5 actually takes evidence into account from the 4-5 days before day d_Q as shown in Figure 3.

Overall, the promising performance of GaussBoost shows that historical blog post evidence can be of use to enhance the accuracy of our voting-based news ranking approach. This suggests that important predictable events are discussed beforehand within the blogosphere and can provide valuable evidence. However, this seems only to be the case when the historical information is carefully weighted such that distant evidence does not gain too much influence, i.e. looking too far into the past can take too much misleading noise into account.

7.2 Future Temporal Evidence

As shown previously, historical blog post evidence can provide an insight into the importance of predictable newsworthy events. However, after the day of interest (d_Q), both predictable and unpredictable events which occurred on d_Q can be discussed. In this section, we examine how blog post volume after the event can aid in the identification of top news stories. In particular, we investigate the performance achieved through the promotion of temporal evidence for our two techniques NDayBoost and GaussBoost.

For NDayBoost, we examine performance in terms of MAP over various values of n, $0 \leq n \leq 10$, i.e the 10 days after the query day d_Q. Figure 6 shows the MAP performance as we vary n. We note that, in contrast to NDayBoost using historical evidence, performance quickly increases above our baseline. Indeed, when evidence from the following 6 days is employed ($n = 6$), performance peaks at a statistically significant +10% MAP ($p < 0.01$).

Next, we evaluate the more fine-grained GaussBoost technique for future evidence. Recall that GaussBoost states a preference for news articles which persist over a few days close to d_Q. Hence, news articles that become important a few days after d_Q do not receive as much emphasis as when applying the NDayBoost technique. MAP performance for $0 < w < 10$ is shown in Figure 7. From the figure, we observe that the performance increases to a peak for low w values, i.e. when we focus on days close to d_Q, which indicates that looking too far into the future adds noise. However, for $w < 1$ (less than 3 days after d_Q, see Figure 3), performance is less than the maximal, suggesting that we need at least 3 or more days of additional evidence. However, overall, the MAP performance curve mirrors historical GaussBoost (Figure 5). Nevertheless, in contrast, we note that the maximal MAP achieved using future evidence is higher, which shows that more useful evidence can be garnered after the events than before. Indeed, upon a detailed comparison over the 55 query days, we observe that Gaussian boosting with future evidence resulted in a marked MAP increase for six more query days than with historical evidence. In turn, this suggests that many of the judged important news articles in our corpus were discussed in the blogosphere after the event, and that these events may have been less predictable in nature.

Lastly, we compare the overall performance of NDayBoost and GaussBoost for the Blogs08 corpus. For historical evidence, GaussBoost is clearly a more effective technique to integrate the blog post evidence, as it focused on blog post discussion closest to the event. However, for future evidence, GaussBoost and NDayBoost perform similarly, suggesting that future blog post evidence extracted from Blogs08 is more easily interpreted when identifying

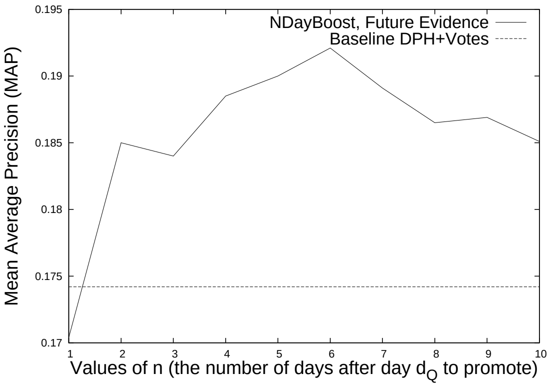

Figure 6: *Votes* MAP performance for NDayBoosting using up to 1 week of future blog post evidence.

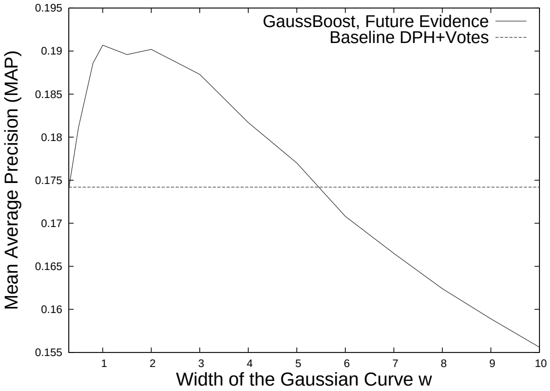

Figure 7: *Votes* MAP performance for GaussBoosting when varying w using future evidence.

important news stories. Therefore, from a practical perspective, we can conclude that for Blogs08, GaussBoost should be applied when accounting for historical evidence, while either approach can be employed for future evidence.

8. HEADLINE SELECTION & REFINEMENT

In this section, we investigate improving the NYT headline corpus, with a view to increasing the news article ranking effectiveness of our DPH+Votes baseline. In particular, from the NYT corpus, we identified various inappropriate news articles, which were unlikely to be important, noting that these news stories could be automatically removed apriori. Indeed, we hypothesise that the automatic removal of such articles beforehand will increase ranking performance. Furthermore, as only the headline of each article was available, we examine ways to *expand* and refine these headlines in such a way that further related blog posts can be identified, thus increasing ranking effectiveness. In the following, Section 8.1 describes some heuristics to automatically eliminate headlines unlikely to be important news stories on any given day. Section 8.2 describes an enrichment process, where each headline is expanded and refined.

8.1 Headline Removal

In this section, we apply NYT corpus-specific techniques to remove headlines. Figure 8 shows a sample of NYT headlines for the 6th of November 2008. From this, we note various headlines which are of dubious news value. For instance, arts reviews (NYTimes-20081106-0017) and text corrections (NYTimes-20081106-0141) are unlikely to be worthy of valuable front-page space.

Table 4: Indicator patterns of non-newsworthiness.

Paid Notice	Arts Briefly
Corrections for the Record	The Listings
Comments of the Week	Dance Review
Inside the Times	Whats On Today
Best Sellers	Critics Choice
The Week Ahead	Books of the Times
Movie Review	Music Review

Table 5: Performance when using heuristics to reduce the initial headline set, in comparison to our proposed DPH+Votes baseline. Significant improvements ($p < 0.01$) over DPH+Votes are denoted using *.

Heuristic	MAP	P@10
DPH+Votes	0.1742	0.2945
+ Patterns	0.1956*	0.3291*
+ Dates	0.1741	0.2945
+ UpperCase	0.1742	0.2945
+ All_Heuristics	**0.1996***	**0.3364***

From an inspection of the NYT headline corpus, we develop several heuristics which are likely to improve the performance of our DPH+Votes baseline. In particular, we eliminate entire sets of headlines which are unlikely to be newsworthy, and the parts of headlines which are likely to cause off-topic blog posts to enter $R(a)$. We propose the following three heuristics:

- *Patterns*: Some headlines which follow editorially defined patterns can never be newsworthy. We eliminate headlines containing patterns such as "Paid Notice", "Corrections for the record", etc. Table 4 lists the patterns used.

- *Dates*: The presence of dates in headlines may mislead the blog post retrieval system when blogs contain the publication date within the text body, e.g. NYTimes-20081106-0011 (Figure 8). We therefore remove dates during headline tokenisation.

- *UpperCase*: NYT uses uppercase prefixes to denote category information, e.g. ARTS, BRIEFLY (NYTimes-20081106-0017) and N.F.L. ROUNDUP (NYTimes-20081106-0134) in Figure 8. We remove terms all in capitals during headline tokenisation.

To test these few proposed heuristics, we compare their performance to our DPH+Votes baseline. The results are presented in Table 5. From the results in the table, we note that the Patterns heuristic is the most effective of the three and can be improved further when combined with the Dates and UpperCase techniques (+ All_Heuristics). This results in a statistically significant performance increase of 14.5% over the DPH+Votes baseline. Note that the Dates and Uppercase heuristics do not exhibit performance improvements alone, but when both are combined with Patterns, performance is enhanced.

8.2 Headline Enrichment

On inspection of Figure 8, it is evident that many news articles contain only a few information bearing terms. In addition, given that within the context of the TREC 2009 news task, only the article headlines are available to identify on-topic blog posts, there might be a vocabulary mismatch problem between the headlines and the blog posts. Therefore, in this section, we investigate ways in which the headlines can be expanded and refined, such that more on-topic blog posts can be identified for use as evidence.

```
NYTimes-20081106-0017 : ARTS, BRIEFLY; A Tale of Woe: 'Two Cities' to Close
NYTimes-20081106-0011 : Inside the Times, November 6, 2008
NYTimes-20081106-0121 : McNabb Says He Can Relate To Obama
NYTimes-20081106-0134 : N.F.L. ROUNDUP; Giants Shut Down Tyree for Season; Raiders Cut Hall
NYTimes-20081106-0141 : Corrections: For the Record
```

Figure 8: Sample of NYT headlines for the day of the 6th of November 2008.

A classical IR technique for improving adhoc retrieval performance is pseudo-relevance feedback (PRF) [24]. In PRF, from an initial ranking of documents, the top returned documents are assumed to be relevant, and information from these 'pseudo-relevant documents' is used to refine the query. Typically, this automatic process takes the form of query expansion, where some of the most informative terms from the pseudo-relevant documents are used to expand the initial query.

We experiment with two applications of query expansion (QE). In the first application, we expand each headline using the top-ranked blog posts for that headline. The expanded headline is then used to generate a further refined ranking of blog posts.

However, prior experiments for other retrieval tasks have found that QE using a corpus of blog posts is not very effective, primarily due to the varying quality of the blog posts [6]. Instead, we propose the use of collection enrichment [4, 9, 12] to expand and refine the headlines. In collection enrichment (CE), query expansion is performed using an external, higher quality corpus. The expanded query is then used to retrieve the final ranking of documents from the target corpus. In the following, we compare and contrast the retrieval performances of traditional query expansion and collection enrichment for the purposes of headline expansion, and whether these can enhance the effectiveness of our voting approach for top news identification.

In terms of experimental setting, for collection enrichment, we use an English Wikipedia crawl from early 2009, that forms a subset of the TREC ClueWeb09[6] corpus of Web documents. For both QE and CE, we identify top terms from the feedback documents using the Bo1 term weighting model from the Divergence from Randomness framework [1]. In particular, we expand the query with 10 terms from the top 3 ranked feedback documents. Note that existing terms in the headline are also re-weighted as part of the expansion process.

We compare the effectiveness of our DPH+Votes baseline for news article ranking, when using either CE or QE. The results are reported in Table 6. From the results, we observe that while both QE and CE improve performance, only CE does so by a statistically significant margin. These results suggest that the vocabulary mismatch between the headlines and blog posts is indeed an issue, as each headline only provides a limited representation of the corresponding news story. Indeed, by applying CE to mitigate this, we significantly increase performance over our DPH+Votes baseline in terms of MAP. However, as indicated by [6], query expansion is less useful for blog post corpora due to noise, and indeed exhibits little benefit on Blogs08.

Table 7 illustrates the QE and CE stemmed expansion terms for a sample news headline describing the trial of a US marine staff sergeant, F. D. Wuterich, whose men killed 24 Iraqis during a raid in the town of Haditha. From this, we can see that for QE, relevant terms are being selected, e.g. case, charg (charge), court, etc. However, these terms are fairly generic across court cases, which make them less useful for uniquely identifying this particular story. In contrast, CE selects more story specific terms, like the town Haditha and the marine's surname Wuterich. We suspect that the rise in effectiveness from CE is due to two important points. Firstly, Wikipedia is a high quality and topically focused corpus in comparison to Blogs08, meaning that selection of useful expansion terms should be easier. Secondly, we suspect that Wikipedia can be a strong source of news related information, i.e. that Wikipedia's contributors update pages with current news. Indeed, on the day of Michael Jackson's death, his Wikipedia page was updated 104 times, with a further 641 updates the day after[7].

Table 6: Performance when using Pseudo-Relevance-Feedback in comparison to our basic *Votes* approach. Significant improvements over DPH+Votes are denoted using *.

Technique	MAP	P@10
DPH+Votes	0.1742	0.2945
+ Query Expansion	0.1747	0.2945
+ Collection Enrichment	**0.1899***	**0.3145**

Table 8: Performance when using the combination of Pseudo-Relevance-Feedback, headline removal and Gaussian boosting (retrospective) in comparison to our basic DPH+Votes approach. Significant improvements over DPH+Votes are denoted using *, while significance over the best individual improvement (All_Heuristics) is demoted †.

Technique	MAP	P@10
DPH+Votes	0.1742	0.2945
+ GaussBoost(w=1,R)	0.1907*	0.3236*
+ All_Heuristics	0.1996*	0.3364*
+ CE	0.1899*	0.3145
+ GaussBoost(w=1,R) + All_Heuristics + CE	**0.2210*†**	**0.3691***

9. COMBINING EVIDENCE

Having evaluated our DPH+Votes baseline to news article ranking, as well as examined ways to improve upon that initial ranking, we now try combining our ranking improvements to determine the extent to which they are additive. Table 8 shows the effectiveness of our baseline approach in comparison to the combination of DPH+Votes and collection enrichment, headline removal and Gaussian boosting (retrospective) in terms of MAP and P@10. As can be observed from the results, performance is higher than our baseline approach by a large and statistically significant margin (+26.7% MAP). Moreover, the combination of techniques shows a statistically significant increase (in terms of MAP) of almost 11% over our best single improvement (All_Heuristics). Furthermore, this represents a large improvement over all of our baselines (TREC median, Inlinks and Random) and the best TREC 2009 top news stories identification systems.

10. CONCLUSION

In this paper, we investigated the problem of automatically ranking news articles based upon evidence from the blogosphere. We proposed modelling this task as a voting process, where each related blog post returned for a news article acts as a vote for that article and the volume of votes received is used for ranking. Within the context of the new (retrospective) top stories identification task

[6] http://boston.lti.cs.cmu.edu/Data/clueweb09/

[7] See `http://en.wikipedia.org/wiki/Michael_Jackson?action=history`

Table 7: Samples of QE and CE expansion terms for the news headline "2 more marine trials in killings of 17 iraqis". Note that the expansion terms have been stemmed using Porter's stemmer.

Expansion Technique	Expanded Headline
Query Expansion	marin (1.2799), trial (1.0452), defend (0.0870), charg (0.0578), counti (0.0551) case (0.0502), court (0.0434), north (0.0490), martial (0.0404), testifi (0.0370)
Collection Enrichment	marin (1.2229), iraqi (1.1301), haditha (0.6530), wuterich (0.1937), investig (0.1236) charg (0.1171), kill (1.1117), massacr (0.1025), murtha (0.0975), sharratt (0.0909)

within the TREC 2009 Blog track, we thoroughly evaluated our Votes approach against four baselines: the TREC median, Inlinks, Random and the best TREC systems. Our results show large statistically significant improvements over all baselines and state-of-the-art TREC systems (excluding that based on a similar voting approach), showing that blog post volume is a useful indicator for news article importance.

Furthermore, taking our initial Votes approach when paired with the DPH weighting model, we investigated various techniques with a view to improving performance. In particular, we evaluated the usefulness of historical and future evidence for news article re-ranking, heuristics to clean the New York Times corpus of non-newsworthy articles as well as news article expansion (query expansion and collection enrichment) to counter vocabulary mismatch. Of these, all except query expansion appear to be effective, indeed exhibiting statistically significant improvements in terms of MAP over our highly-performing DPH+Votes approach.

Overall, we conclude that evidence from the blogosphere can be a useful indicator as to the importance of various news stories, and that this can be successfully leveraged using our Votes approach to automatically rank headlines for a news editor. In the future, we wish to further investigate the effect of the news article representation on performance, for example, using the article body or anchor text. Additionally, as noted earlier, we are interested in investigating the application of our model to a real-time article ranking task, using both the blogosphere and other UGC corpora (e.g. Twitter).

11. REFERENCES

[1] G. Amati. Probabilistic models for information retrieval based on divergence from randomness. PhD, University of Glasgow, 2003.

[2] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog track. In *Proceedings of TREC 2007*.

[3] British Library Board. Concise history of the British newspaper in the eighteenth century. http://www.bl.uk/reshelp/findhelprestype/news/concisehistbritnews/britnews18th/, 2009.

[4] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of SIGIR 2006*.

[5] J. Galtung and M. H. Ruge. The structure of foreign news: the presentation of the Congo, Cuba and Cypris crises in four Norwegian newspapers. *The Journal of Peace Research*, 2(1):64–90, 1965.

[6] B. He, C. Macdonald, I. Ounis, V. Plachouras, and R. Santos. University of Glasgow at TREC 2008: Experiments in Blog, Enterprise, and Relevance Feedback tracks with Terrier. In *Proceedings of TREC 2008*.

[7] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of KDD 2002*.

[8] A. König, M. Gamon and Q. Wu. Click-through prediction for news queries. In *Proceedings of SIGIR 2009*.

[9] K. L. Kwok and M. S. Chan. Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of SIGIR 1998*.

[10] J. Leibowitz. "Creative destruction" or just "destruction", how will journalism survive the Internet age? In the *FTC Public Workshop: From town crier to bloggers: how will journalism survive the Internet age?* Washington, DC, USA, 2009.

[11] C. Macdonald. The voting model for people search. PhD thesis, University of Glasgow, 2009.

[12] C. Macdonald, B. He, V. Plachouras, and I. Ounis. University of Glasgow at TREC 2005: Experiments in Terabyte and Enterprise tracks with Terrier. In *Proceedings of TREC 2005*.

[13] C. Macdonald, I. Ounis and I. Soboroff. Overview of TREC-2009 Blog track. In *Proceedings of TREC 2009*.

[14] R. McCreadie, C. Macdonald, I. Ounis, J. Peng and R. Santos. University of Glasgow at TREC 2009: Experiments with Terrier. In *Proceedings of TREC 2009*.

[15] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with Columbia's NewsBlaster. In *Proceedings of LTC 2002*.

[16] G. Mishne and M. Rijke. A study of blog search. In *Proceedings of ECIR 2006*.

[17] S. Munk. Michael Jackson death covered first by Twitter crashes servers. http://www.electricpig.co.uk/2009/06/26/michael-jackson-death-covered-first-by-twitter-crashes-servers/, 2009.

[18] Newspaper Association of America. Readership trends. http://www.naa.org/TrendsandNumbers/Readership.aspx, 2009.

[19] 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 the 2nd International Workshop on Open Source Information Retrieval, SIGIR 2006*, Seattle, USA.

[20] R. G. Picard. Commercialism and newspaper quality. *Newspaper Research Journal*, 25(1):54–65, 2004.

[21] D. R. Radev, J. Otterbacher, A. Winkel, and S. Blair-Goldensohn. NewsInEssence: summarizing online news topics. *Commun. ACM*, 48(10):95–98, 2005.

[22] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu and M. Gatford. Okapi at TREC-3. In *Proceedings of TREC 1995*.

[23] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and L. Payne. Okapi at TREC-1. In *Proceedings of TREC 1992*.

[24] G. Salton. The SMART retrieval system—Experiments in automatic document processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[25] H. Sayyadi, M. Hurst and Al Maykov. Event detection and tracking in social streams. In *Proceedings of ICWSM 2009*.

[26] D. Sifry. The state of the live web, April 2007. http://www.sifry.com/alerts/archives/000493.html, 2007.

[27] J. Sigmund. Online newspaper viewership reaches record in 2007. http://www.naa.org/PressCenter/SearchPressReleases/2008/Online-Newspaper-Viewership.aspx, 2008.

[28] R. Steinberger. Highly Multilingual News Analysis Applications. In *Proceedings of ECML PKDD 2009*.

[29] M. Sussman. The state of the blogosphere 2009. http://technorati.com/blogging/article/state-of-the-blogosphere-2009-introduction/, 2009.

[30] M. Thelwall. Bloggers during the London attacks: Top information sources and topics. In *Proceedings of the 3rd annual workshop on the Weblogging Ecosystem, WWW 2006*, Edinburgh, Scotland.

[31] Wavelength Media. What makes a story newsworthy? http://www.mediacollege.com/journalism/news/newsworthy.html, 2009.

[32] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proceedings of SIGIR 1998*.