Mining the Blogosphere for Top News Stories Identification

Yeha Lee   Hun-young Jung   Woosang Song   Jong-Hyeok Lee

Division of Electrical and Computer Engineering
Pohang University of Science and Technology
Pohang, Gyungbuk, Republic of Korea
{sion, blesshy, woosang, jhlee}@postech.ac.kr

ABSTRACT

The analysis of query logs from blog search engines show that news-related queries occupy a significant portion of the logs. This raises a interesting research question on whether the blogosphere can be used to identify important news stories. In this paper, we present novel approaches to identify important news story headlines from the blogosphere for a given day. The proposed system consists of two components based on the language model framework, the query likelihood and the news headline prior. For the query likelihood, we propose several approaches to estimate the query language model and the news headline language model. We also suggest several criteria to evaluate the news headline prior that is the prior belief about the importance or newsworthiness of the news headline for a given day. Experimental results show that our system significantly outperforms a baseline system. Specifically, the proposed approach gives 2.62% and 10.19% further increases in MAP and P@5 over the best performing result of the TREC'09 Top Stories Identification Task.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval – *Retrieval models*

General Terms

Algorithms, Experimentation, Performance

Keywords

Blog Retrieval, Blogosphere, Top News Stories Identification

1. INTRODUCTION

A blog, "web log", is a special type of website in which users (individuals or groups) express their opinions or thoughts on several subjects. Blog posts consist of a wide variety of topics. As the number of blog users increase, the popularity and the importance of blogs are growing, and several commercial search engines such as Google[1] and Technorati[2] have provided blog search services.

Users' information needs for blog search differ from those for general web search. A large portion of the query logs from blog search engines are news-related queries [21, 22]. In other words, many users find information about news stories in the blogosphere. This implies that the blogosphere may be helpful when locating news stories.

A large number of news stories from various news channels are generated and updated day after day. However, a relatively few news among huge number of them receive attention from users. Therefore, it is one of the most important issues to evaluate the importance of news stories and rank them.

We investigate how to take advantage of the blogosphere for identifying top news stories. To this end, given a certain day, we retrieve and rank news headlines according to their importance or newsworthiness, using the blogosphere. Furthermore, this task is worthwhile in that it identifies the top news stories from blog users' point of view, instead of the news providers. The task is also called Top Stories Identification Task (TSIT) which was first introduced at the TREC 2009 Blog Track [21].

TSIT is a new pilot task that aims to "address the news dimension in the blogosphere" [21]. The task uses a date (day) as a query. For a date query, the system for the task ranks the news headlines in the order of their importance. Furthermore, for each news headline, the task requires a certain number of blog posts that capture diverse aspects relevant to the news headline.

TSIT has some characteristics that distinguish it from previous news-related studies such as Topic Detection and Tracking (TDT). First, the data given for TSIT contains only news headlines but no news contents. Therefore, the system for the task should rank news headlines utilizing the blogosphere (i.e. Blog08 corpus) instead of the contents of news articles. Second, unlike the corpus of news stories, blog posts are generally neither well-written articles nor topically coherent. They also include a lot of non-topical contents such as spam blogs and blog comment spam that advertise commercial products and services [16], making the task difficult.

In this paper, we present novel approaches to identify the top news stories in the blogosphere. The proposed approaches are based on the language model framework, which is widely used in information retrieval tasks. We propose a

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
Copyright 2010 ACM 978-1-60558-896-4/10/07 ...$10.00.

[1] http://blogsearch.google.com/

[2] http://www.technorati.com/

series of approaches to estimate a query language model and a news headline language model based on the blogosphere, and to rank the news headlines according to the distance between two language models. We also suggest several criteria to evaluate the prior probability that a news headline will be a top news story for a given day, and verify that these criteria are useful to identify top news stories. The experimental results show that our approach significantly improves the best performance submitted in the TREC 2009 Top Stories Identification Task.

The rest of the paper is organized as follows. In section 2, we briefly survey related work on new event detection. In section 3, we address the framework of our system, and propose several approaches to identify the top news headlines. In section 4, we conduct several experiments to evaluate the performances of our approach. Finally, we conclude the paper and discuss future work in section 5.

2. RELATED WORK

As a new pilot task, TSIT aims to identify top news stories in the blogosphere, and to provide a ranked list of news headlines. There are few researches for identifying and ranking top news stories in the blogosphere. One of the research directions closely related to TSIT may be the New Event Detection.

New Event Detection, one of the five tasks in TDT, aims to detect whether a given news story is concerned with already known events to a system or not. For the event detection problem, many approaches have been based on clustering or classification to estimate the similarity between the events and documents (e.g. the news stories); these approaches differ in the ways by which they evaluate the similarity [3, 5, 17, 25, 28, 29]. All of them compare each document with existing events. If the similarity between the document and the events is lower than some predefined criteria, the document is considered to address a new event. Otherwise, the document is assigned to the event to which it is most similar.

Various features have been proposed, including timeline analysis, burstiness and named entities. Chen *et al* proposed an aging theory to capture the life cycle of a news event, and improved the performance for event detection [7]. Chen *et al* used an aging theory and a sentence modeling to extract hot topics from news documents [8]. They analyzed the timeline to identify the key terms. The burstiness of terms was used by many researchers for the event detection [9, 12, 15, 24]. Kleinberg proposed an approach to identify the bursty features for the event detection from e-mail streams [15]. They used the infinite-state automaton to model the stream. He *et al* identified bursts of (a)periodic features using a Gaussian distribution, and then used them to detect (a)periodic events [12]. Kumaran and Allan used named entities for the event detection [17]. They showed that the usefulness of named entities can change according to certain situations. Kuo *et al* classified terms within news stories based on named entity type and parts-of-speech tags, and assigned a different weight to each term according to the type and class of news story [28].

The main difference between previous work for the event detection and our approach stems from the difference in source data for identifying events or news stories. In contrast with previous work, we identify the top news headlines using only the unorganized blogosphere, not the well-defined contents of news articles.

3. TOP STORIES RANKING MODEL

To identify the top news stories, we rank them according to their importance or newsworthiness on a specific day. The newsworthiness of a news story can be decided by several criteria[3] as follows:

- **Timing** News stories that is happening now are often more newsworthy than those that happened a week ago.
- **Significance** The number of people involved in a news story is important.
- **Proximity** News stories that occur near us are more important than distant ones.
- **Prominence** News stories about famous people are more newsworthy than stories about ordinary people.
- **Human-Interest** Human-interest stories are generally soft news. They appeal to emotions.

We assume that a news story mentioned in more blog posts or comments is more important or newsworthy on a specific day, because a top news story satisfying the above criteria may receive attention from many blog users, who express their thoughts or opinions about the news story in their blogs.

To measure the importance of a news story using the blogosphere, we adopt the language model framework, which is widely used in information retrieval tasks. Motivated by our assumption, we evaluate the importance of a news headline using the probability that blog posts published on a query day generate the headline. Let H be a news headline and let Q_d and Q_p be a given (date) query and a set of blog posts published on the query day Q_d, respectively.

$$\underbrace{Score(Q_d, H)}_{\substack{\text{Importance score}\\ \text{of a news headline}}} \propto P(H|Q_p) \propto \underbrace{P(Q_p|H)}_{\substack{\text{Query}\\ \text{Likelihood}}} \quad \underbrace{P(H)}_{\substack{\text{Headline}\\ \text{Prior}}} \tag{1}$$

3.1 The Query Likelihood

In the language model framework, the query likelihood means the probability that a document generates a given query. TSIT uses a date (day) as a query. Therefore, we regard the query likelihood as the probability that a news headline generates blog posts published on the query day (i.e. Q_p).

To this end, we should estimate two language models, the Query Language Model (QLM) and the News Headline Language Model (NHLM). Both of the language models are estimated, based on blog posts.

3.1.1 Query Language Model

For a query day Q_d, we estimate the QLM using blog posts Q_p. However, the blog posts may discuss various topics from individual daily affairs to important events recently happened. If we model the blog posts using a single language model, the language model cannot correctly capture

[3] http://www.mediacollege.com/journalism/news/newsworthy.html

the contents of the blog posts. As a result, we will get the wrong QLM.

To solve this problem, we divide the documents into K clusters. We assume that each cluster can accurately reflect one of the various topics mixed in the blog posts. To estimate the QLM, we first gather blog posts which are published on a query day. Then, we cluster them using the K-means algorithm. We represent each document using the term vector $\vec{d}: w_i = tf_i \times idf_i$, where tf_i indicates the frequency of term w_i within a document d, and $idf_i = \log(\frac{|TD|}{df_i})$ means inverse document frequency: $|TD|$ is the total number of documents in a collection and df_i is document frequency of a term w_i. We use the cosine similarity as the distance function between two documents.

$$Similarity\left(\vec{d_i}, \vec{d_j}\right) = \frac{\vec{d_i} \cdot \vec{d_j}}{|\vec{d_i}| \times |\vec{d_j}|} \tag{2}$$

where $|\vec{d_i}|$ and $|\vec{d_j}|$ indicate the length of $\vec{d_i}$ and $\vec{d_j}$, respectively.

After clustering, we can generate the QLMs from the K document sets. Let $D_k = \{d_{k_1}, d_{k_2}, \cdots, d_{k_n}\}$ be the kth document set, θ_{QLM_k} be the kth QLM. The document set D_k contains information relevant to a topic of the document set, but also contains background information. We assume that the documents are generated by a mixture model of θ_{QLM_k} and the collection language model θ_C that reflects the background information.

$$P(D_k) = \prod_i \prod_w \{(1-\lambda)P(w|\theta_{QLM_k}) + \lambda P(w|\theta_C)\}^{c(w;d_{k_i})} \tag{3}$$

where $c(w; d_{k_i})$ is the number of times term w occurred in a document d_{k_i}, $P(w|\theta_C) = \frac{ctf_w}{|C|}$: ctf_w is the number of times term w occurred in the entire collection, $|C|$ is the length of the collection, and λ is a weighting parameter. In our experiments, we set λ as 0.8.

Then, we can estimate θ_{QLM_k} using the EM algorithm [11]. The EM updates for $p(w|\theta_{QLM_k})$ are as follows:

$$t_w^n = \frac{(1-\lambda)P^n(w|\theta_{QLM_k})}{(1-\lambda)P^n(w|\theta_{QLM_k}) + \lambda P^n(w|\theta_C)} \tag{4}$$

$$P^{n+1}(w|\theta_{QLM_k}) = \frac{\sum_{i=1}^{n} c(w; d_{k_i}) t_w^n}{\sum_{w'} \sum_{i=1}^{n} c(w'; d_{k_i}) t_{w'}^n} \tag{5}$$

3.1.2 News Headline Language Model

To estimate the NHLM, for each news headline, we first retrieve blog posts relevant to its topic using the news headline itself as query. To this end, we evaluate the relevance between a news headline H and a blog post d using the the KL-divergence language model [18] with Dirichlet smoothing [27].

$$Score(H, d) \propto \sum_w P(w|H) \log P(w|d) \tag{6}$$

where $P(w|H)$ is the maximum likelihood estimates of the news headline, and $P(w|d) = \frac{c(w;d) + \mu_d P(w|\theta_C)}{|d| + \mu_d}$: $c(w; d)$ is the number of times a term w occurred in a blog post d, and μ_d is a smoothing parameter, and $|d|$ is the length of the document d.

Among the search results, we use only the blog posts whose issued date is within a certain period from a query day, because the time gap between the issued day of a blog post and a query day often means that the blog post is mentioning an event different from those that happened on that day [25]. In other words, the blog post is likely to be relevant to a topically similar, but different news headline.

We gather only the blog posts between -3 and +28 days from a query day. Then, we choose 10 blog posts that can provide as diverse aspects about the news headline as possible. We call the 10 blog posts the supporting relevant posts of the news headline.

We propose two approaches to make the supporting relevant posts reflect the diverse aspects relevant to the news headline: Relevance-Based Selection (RBS), Feed-Based Selection (FBS).

RBS is an intuitive but naive approach to choose the supporting relevant posts. This approach selects the supporting relevant posts according to a relevance score of each blog post obtained from Eq. 6. We define this approach as a baseline for our experiments.

FBS chooses the supporting relevant posts based on blog feeds which belong to each of them. Individual blog users may have different interests and tendencies for the same events, and these differences can be represented through their blog feed [19]. That is, for a given news headline, blog posts from different blog feeds can provide different aspects, even if they address information on the same news story. Therefore, to increase the diversity of the supporting relevant posts, we select them from as wide a range of blog feeds as possible. In a similar way to RBS, FBS also chooses blog posts according to their relevance score, but FBS selects only one blog post from one blog feed.

We estimate the NHLM using the maximum likelihood estimate of the 10 supporting relevant posts and the Dirichlet smoothing [27]. Let θ_{NHLM} and θ_C be the NHLM and the collection language model, respectively, and let SRP be a set of the 10 supporting relevant posts.

$$P(w|\theta_{NHLM}) = \frac{c(w; SRP) + \mu_h P(w|\theta_C)}{|SRP| + \mu_h} \tag{7}$$

where $c(w; SRP)$ is the number of times a term w occurred in SRP and μ_h is a smoothing parameter.

3.1.3 Score Function

To evaluate the query likelihood, we use the KL-divergence language model [18], one of the-state-of-the-art information retrieval models, to rank news headlines in response to a given query. We use the maximum value among scores between the QLMs and the NHLM as the relevance score of a news headline.

Let $Score_{QLH}(Q_d, H)$ be the relevance score of a news headline H with respect to a given query Q_d. We define $Score_{QLH}(Q_d, H)$ as follows:

$$Score_{QLH}(Q_d, H) = \max_k \left\{ \sum_w P(w|\theta_{QLM_k}) \log P(w|\theta_{NHLM}) \right\} \tag{8}$$

3.2 The News Headline Piror

We suggest two criteria to estimate the news headline prior $P(H)$ that is the prior belief about the importance or newsworthiness of a news headline for a given day: Temporal Profiling and Term Importance. Although the proposed approaches depend on a date such as the query day, we regard them as the priors of a news headline in that they are independent of the query language model.

3.2.1 Temporal Profiling

The Temporal Profiling criterion uses the temporal information of blog posts relevant to a news headline. We assume that if a news headline is important for a query day, many blog posts relevant to its topic will be posted on that day.

To generate the temporal profile of each news headline, we use a temporal profiling approach proposed in [14] with some modifications. The temporal profile of a news headline H is defined as follows:

$$P(t|H) = \sum_{d \in R} P(t|d) \frac{Score(H,d)}{\sum_{d' \in R} Score(H,d')} \qquad (9)$$

where t is a date (day), and R is a document set that consists of 500 blog posts selected by an order of a relevance score $Score(H,d)$ from Eq. 6, and

$$P(t|d) = \begin{cases} 1 & \text{if } t \text{ is equal to the document date} \\ 0 & \text{otherwise} \end{cases}$$

We then smoothed the temporal profile $P(t|H)$ using the background model as follows:

$$P(t|H) = (1-\alpha)P(t|H) + \alpha P(t|C) \qquad (10)$$

where $P(t|C) = \frac{1}{|TD|}\sum_{d \in C} P(t|d)$: $|TD|$ is the total number of documents in the collection, and α is a smoothing parameter. In our experiments, we set $\alpha = 0.5$.

This temporal profile is defined on each single day. However, if a news story is important for a query day Q_d, the blog posts relevant to it may be published over a certain period following the day due to the bursty nature [15]. Therefore, we smooth the temporal profile model with the model for adjacent days. Let $Score_{TP}(Q_d, H)$ be a score of a news headline estimated using the temporal profile of the news headline.

$$Score_{TP}(Q_d, H) = \frac{1}{Z_w} \sum_{t \in \phi} w(t, Q_d) P(t|H) \qquad (11)$$

where ϕ indicates a period from Q_d, and $Z_w = \sum_{t \in \phi} w(t, Q_d)$. We define a weight function for $w(t, Q_d)$ using the Cosine (Hamming) kernel function [20] as follows:

$$w(t, Q_d) = \begin{cases} \frac{1}{2}\left\{1 + cos\left(\frac{|t-Q_d| \times \pi}{\sigma}\right)\right\} & t \in \phi \\ 0 & \text{otherwise} \end{cases} \qquad (12)$$

3.2.2 Term Importance

The Term Importance criterion uses term information of a news headline. We believe that each term has a different importance for a given day. If a news headline consists of important terms, it is likely to be a top news story and vice-versa. For example, a news headline that consists of common words or stopwords may not be a top news story.

We only consider named entities, not all terms in a news headline. Named entities were used by many event detection systems, improving the performance of the systems [17, 26, 28]. We extract named entities from each news headline using the Stanford Named Entity Recognizer[4]. Then, we gather all n-gram ($n \leq 3$) from the named entities.

We evaluate the importance of the n-gram terms based on the $TF \cdot IDF$ approach that is widely used for term weighting in many information retrieval tasks.

Let nt be the extracted n-gram term and $TF(nt, Q_d)$ be a term frequency of the term nt for a query day Q_d. Intuitively, if a term nt occurs frequently within news headlines issued on a query day, it is likely to be important. In a similar way to the bursty nature of blog posts, news headlines relevant to important events that happen on the query day may be published over several subsequent days. Therefore, we define the term frequency $TF(nt, Q_d)$ as the number of a term nt within news headlines that are issued during a certain interval containing a query day Q_d.

$$TF(nt, Q_d) = \sum_{t \in \phi} c(nt; t) \qquad (13)$$

where ϕ indicates the period, and $c(nt; t)$ means the number of a term nt occurring in news headlines issued on day t.

Let $IDF(nt)$ be the inverse "date" frequency, and TND be the total number of days that the news headline corpus spans. We define $IDF(nt)$ as follows:

$$IDF(nt) = \frac{TND}{DF(nt) + \delta} \qquad (14)$$

where $DF(nt)$ indicates the number of days on which nt occurs in news headlines, and δ is a constant which controls the influence of $DF(nt)$ on $IDF(nt)$.

The inverse date frequency $IDF(nt)$ corresponds to the inverse document frequency. In other words, a term nt with a high $IDF(nt)$ value may be a keyword that distinguishes important events that happened on a query day from those that happened on other days.

Let $Score_{TI}(Q_d, H)$ be the importance of a news headline H evaluated using the term importance.

$$Score_{TI}(Q_d, H) = \max_{nt \in H} \left(TF(nt, Q_d) \times IDF(nt)\right) \qquad (15)$$

3.3 Integration of Query Likelihood and News Headline Prior

We proposed several approaches for the query likelihood and the news headline prior in section 3.1 and 3.2. They capture the different characteristics of important news headlines. For the query likelihood, we analyze the contents of the blog posts, and model the dominant topics buried in them. Then, we rank the news headlines according to the probability that each headline generates one of the topics. For the news headline prior, we proposed two criteria to reflect the properties of important news headlines, Temporal Profiling and Term Importance.

To identify the top news headline, we integrate the query likelihood with the news headline prior. To achieve this, we first adjust each score from 0 to 1.

$$Score_i^{'}(H) = \frac{Score_i(H) - min_i}{max_i - min_i} \qquad (16)$$

$$min_i = \min_{H'} Score_i(H^{'}) \quad \text{and} \quad max_i = \max_{H'} Score_i(H^{'})$$

where $Score_i(H)$ indicates one score of $Score_{QLH}(Q_d, H)$, $Score_{TP}(Q_d, H)$ and $Score_{TI}(Q_d, H)$.

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

Finally, we define the ranking function as follows:

$$Score(Q_d,H)=(1-\beta_1)Score'_{QLH}(Q_d,H) + \beta_1\left\{(1-\beta_2)Score'_{TI}(Q_d,H)+\beta_2 Score'_{TP}(Q_d,H)\right\} \quad (17)$$

where β_1 is the weighting parameter that adjusts the importance between the query likelihood and the news headline prior, and β_2 is the parameter that controls the weights between two criteria for the news headline prior.

4. EXPERIMENTS

We conducted several experiments to evaluate our system for TSIT. We measured the performance of the query likelihood and the news headline prior, respectively. We also investigated the influence of the combination of two components on the performance of TSIT, with a varying weight parameter β_1.

4.1 Setup

4.1.1 Data Set

The Blogs08 corpus and the news headline corpus from the New York Times (NYT) [21] were used for experiments. The Blogs08 corpus was created by monitoring 1 million blogs from January 14, 2008 to February 10, 2009, and consisted 808GB of feeds, 1445GB of permalink documents and 56GB of homepages. The news headline corpus consisted of headlines of articles published by NYT during the interval covered by the Blogs08 corpus.

Our experiments were performed using only Blog08 and the news headline corpus without resorting to any other resources. For the evaluation, we used the 55 topics and relevance judgments from the TREC 2009 Top Stories Identification Task.

We only used the permalinks (blog post) for the experiments. We discarded the HTML tags of each blog post, and applied the DiffPost algorithm [23] to remove non-relevant contents[5] of each blog post. Each blog post was also processed by stemming using the Porter stemmer and eliminating stopwords using the INQUERY words stoplist [2].

4.1.2 Evaluation Method

In response to each query, we retrieved 100 news headlines according to their importance on that day, and provide 10 supporting relevant posts for each news headline, as in TSIT.

The evaluation consists of two phases. In the first phase, we assess the performances of the proposed approaches for identifying the top news headlines for a query day. For each query, we considered only the news headlines corresponding to $Q_d \pm 1$ days as ranking candidates, because of the time discrepancy between the day on which the headline was and the day Q_d on which the news story actually happened [21]. We used the mean average precision (MAP) and the precision at rank 5 and 10 (P@5 and P@10) as the evaluation measures.

In the second phase, the supporting relevant posts are evaluated. The posts should provide diverse aspects relevant to their news headline. To assess the diversity of the supporting relevant posts, we used the α-nDCG [10] and IA-Precision [1] measures.

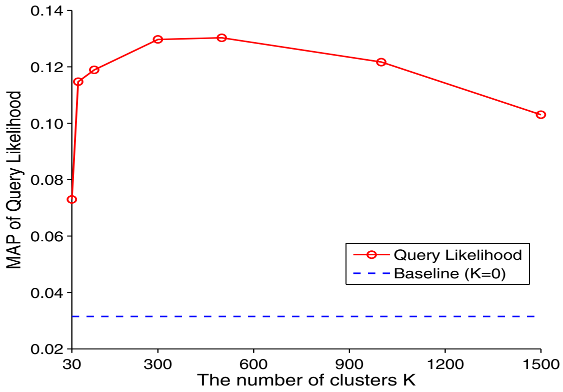

Figure 1: The MAP scores of the query likelihood according to varying the number of clusters K. The NHLM is estimated using the RBS.

Table 1: The performances of the query likelihood estimated using the RBS and the FBS approaches for the NHLM. The number of clusters K is set to 500 for the QLM estimation.

Model	MAP	P@5	P@10
QLH_{RBS}	0.1303	0.2291	0.2255
QLH_{FBS}	0.1315	0.2473	0.2355

4.2 Results and Discussion

4.2.1 The Query Likelihood

We conducted several experiments to evaluate the performance of the query likelihood for TSIT. The aim of the experiments is to determine (1) how correctly the QLMs reflect the various topics buried in blog posts; and (2) how the proposed approaches for NHLM estimation affect the performance of the query likelihood.

The query likelihood has a few parameters, the number of clusters K for QLMs and the smoothing parameter μ_h for NHLM. For our experiments, the smoothing parameter μ_h is set to 2000 without further parameter tuning.

To determine how correctly the QLMs the reflect various topics buried in blog posts, we evaluated the performance of the query likelihood according to varying K values.

$$K : 1, 30, 50, 100, 300, 500, 1000, 1500$$

For the NHLM estimation, the supporting relevant posts were chosen using the RBS approach.

Figure 1 shows the MAP scores according to varying K values. We set the baseline using $K = 1$. This means that blog posts are modeled using a single QLM.

Compared with the baseline, the performances for all $K > 1$ were significantly improved, and the best performance was obtained when using $K = 500$. From these results, we can confirm that a single QLM cannot correctly capture the contents of the blog posts, because of the topical diversity of the blog posts. This weakness reduced its ability for identifying the top news headlines.

[5] In [23], the non-relevant contents of a blog post means the useless contents for the blog search, such as menu, banner and site description

Table 2: The performances of the news headline prior estimated using Temporal Profiling, Term Importance and their combination ($\beta_2 = 0.8$). The best performances are shown in bold.

Model	Period	MAP	P@5	P@10
PRI_{TP}	ϕ_1	0.1410	0.2545	0.2691
	ϕ_2	0.1745	0.2982	0.3109
	ϕ_3	**0.1800**	**0.3055**	**0.3273**
PRI_{TI}	ϕ_1	0.0263	0.0400	0.0509
	ϕ_2	0.0448	0.1127	0.0873
	ϕ_3	**0.0458**	**0.1273**	**0.0891**
PRI_{TP+TI}	ϕ_3	**0.1957**	**0.3673**	**0.3364**

As the number of clusters K increased to 500, the respective topics buried in the blog post were captured by the K clusters. The QLM estimated using each cluster led to the improved performance of the query likelihood. When $K \geq 500$, the clusters have been overfitted, and did not provide enough information relevant to each topic. As a result, the performance decreased.

To investigate how the proposed approaches for NHLM estimation affect the performance of the query likelihood, we measured the performances with two approaches to select the supporting relevant posts. For these experiments, we set $K = 500$ to estimate the QLMs.

Let QLH_{RBS} and QLH_{FBS} be the query likelihood using the NHLM estimated using RBS and FBS approaches, respectively. Table 1 shows the performances of QLH_{RBS} and QLH_{FBS}. The performances of QLH_{FBS} are better than those of QLH_{FBS} for all measures. These results means that the supporting relevant posts selected using FBS provide more diverse aspects of a news headline than those chosen using RBS. As a result, the performance of the query likelihood increased.

4.2.2 The News Headline Prior

We proposed two criteria to estimate the news headline prior: Temporal Profiling and Term Importance. We experimentally confirmed the usefulness of the proposed criteria to estimate the news headline prior.

First, we evaluated the performance of each approach according to varying the period ϕ. The approaches consider a certain period from a query day to gather evidence for the news headline prior. Generally, blog posts and news headlines related to events that happened on a query day are published on that day or in the following days. However, they can be published on preceding days, because of the time discrepancy described in section 4.1.2.

We defined several periods as follows:

- ϕ_1 : ϕ is set between -1 and +1 days from Q_d.
- ϕ_2 : ϕ is set between -3 and +7 days from Q_d.
- ϕ_3 : ϕ is set between -3 and +14 days from Q_d.

Let PRI_{TP} and PRI_{TI} be the news headline prior estimated using the Temporal Profiling and Term Importance, respectively. Table 2 shows the performances of each approach according to varying periods and those obtained from the combination of the two approaches. For the experiments, we set the parameters, σ in Eq.12 and δ in Eq.14, by maximizing the MAP using an exhaustive search in the following values.

$$\sigma, \delta : 30, 40, 50, 60, 70$$

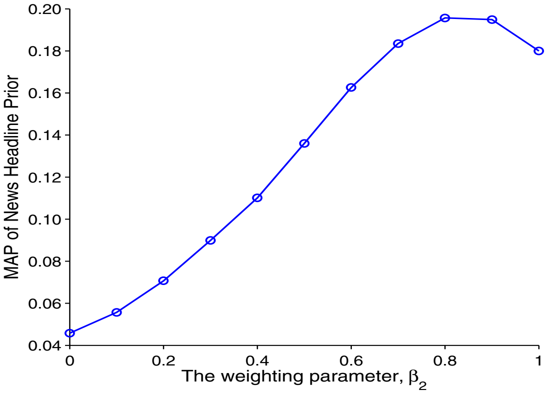

Figure 2: The Map scores of the news headline prior according to varying the parameter β_2 ($\beta_1 = 1$).

Table 3: The performances of systems integrating the query likelihood and the news headline prior, QLH_{RBS}+PRI_{TP+TI} and QLH_{FBS}+PRI_{TP+TI} ($\beta_1 = 0.8$ and $\beta_2 = 0.8$). uogTrTStimes: The best performance in TREC'09 Top Stories Identification Task, QLH_{FBS}: the best performance of the query likelihood, PRI_{TP+TI}: the best performance of the news headline prior. The best performances are shown in bold. Statistical significance at the 0.05 and 0.01 level is indicated by † and ‡ for improvement from the query likelihood, respectively, § and ¶ for improvement from the news headline prior, respectively.

Model	MAP	P@5	P@10
uogTrTStimes	0.1862	0.3236	0.3127
QLH_{FBS}	0.1315	0.2473	0.2255
PRI_{TP+TI}	0.1957	0.3673	0.3364
QLH_{RBS}+PRI_{TP+TI}	0.2081‡¶	0.4145‡¶	0.3455‡
QLH_{FBS}+PRI_{TP+TI}	**0.2124‡¶**	**0.4255‡¶**	**0.3527‡§**

For both approaches, as we consider a longer period, we obtain better results. This observation confirms our assumption that if a news story is important for a given day, blog posts and news headlines relevant to it will be posted during several days. Although Temporal Profiling resulted in good performance, the performances of the Term Importance were relatively low. For event detection, the usefulness of the named entities can change depending on which circumstances they are used in [17]. We think that the poor performance of Term Importance is because we used the named entities without considering the circumstances. However, the combination of the two approaches led to the best performance ($\sigma = 50$ and $\delta = 40$). Compared with the best performance of Temporal Profiling, we achieved 1.57% and 6.18% improvement in MAP and P@5, respectively. These results verify the usefulness of the named entities for identifying important news stories.

To explore the influence of two criteria when identifying the top news story, we measured the MAP score according

Table 4: The performances of the supporting relevant posts chosen using RBS and FBS.

Model	α-nDCG@5	α-nDCG@10	IA-P@5	IA-P@10
QLH_{RBS}+PRI_{TP+TI}	0.479	0.486	0.169	0.145
QLH_{FBS}+PRI_{TP+TI}	0.485	0.492	0.171	0.147

to varying the weighting parameter β_2 ($\beta_1 = 1$, i.e. we are utilizing only the news headline prior to evaluate the performance), in Figure 2. The weight parameter β_2 controls the relative importance of Temporal Profiling and Term Importance as the news headline prior. The best performance was obtained when $\beta_2 = 0.8$. From these results, we can again confirm that Temporal Profiling and Term Importance should be considered together to improve the performance.

4.2.3 *Integration of the Query Likelihood and the News Headline Prior*

Finally, we measured the performance of our system that integrates the query likelihood and the new headline prior. To integrate these components, we used QLH_{RBS} and QLH_{FBS} for the query likelihood, and PRI_{TP+TI} for the news headline prior.

Table 3 shows the performances of our systems, QLH_{RBS}+ PRI_{TP+TI} and $QLH_{FBS} + PRI_{TP+TI}$. In addition, for comparison purposes, we reported the best performing results of the TREC-2009 Top Stories Identification Task [21], and the best performances of the query likelihood and the news headline prior. We performed the Wilconxon signed rank test to examine whether the improvement of the performance over that of each component was statistically significant.

Integrating the two components significantly improved the performance of TSIT. For the MAP, the best performance was 8.09% and 1.67% higher than those of the query likelihood and the news headline prior, respectively. Specifically, our system achieved 2.62%, 10.19% and 4.00% further increases in MAP, P@5 and P@10 over the best performance of TREC'09 TSIT.

This result implies that the performance can be improved by combining the query likelihood and the news headline prior. They reflect different characteristics that important news headlines should be satisfying. The query likelihood identifies the important news headlines based on modeling the dominant topics in blog posts, but the news headline prior identifies them using various features such as the number of relevant posts, and the importance of terms within news headlines.

Figure 3 shows the MAP scores of $QLH_{FBS} + PRI_{TP+TI}$ according to varying the parameter β_1 ($\beta_2 = 0.8$). The weight parameter β_1 controls the relative importance of the query likelihood and the news headline prior. The best performance was obtained when $\beta_1 = 0.8$. From these results, we can again verify that the performance for TSIT can be improved when integrating the query likelihood and the news headline prior. We do not show the graph for QLH_{RBS}+PRI_{TP+TI}, because it was almost identical to that of $QLH_{FBS} + PRI_{TP+TI}$.

4.2.4 *Diversity of Supporting Relevant Posts*

The supporting relevant posts should provide the diverse aspects relevant to a news headline. We proposed two approaches to choose the supporting relevant posts, the RBS and the FBS. We evaluated the diversity of the supporting relevant posts selected by each approach and displayed the results in Table 4.

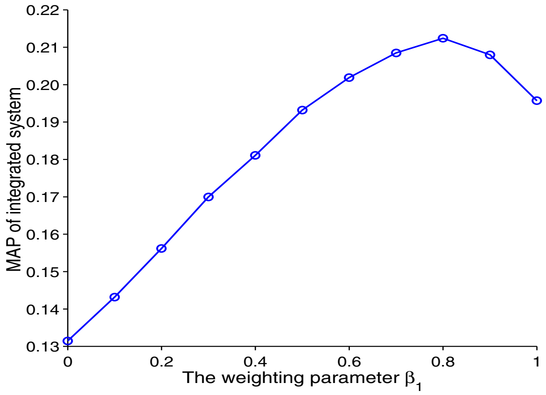

Figure 3: The Map scores of the integrated system according to varying the parameter β_1 ($\beta_2 = 0.8$)

FBS performed better than RBS. We can verify that the supporting relevant posts of FBS provided more diverse aspects of a news headline than those of RBS. These results confirm that the use of blog feeds can improve the diversity of the supporting relevant posts. Furthermore, these results agree with the results from identifying the top news headlines in Table 3. That is, compared with RBS, FBS chose the supporting relevant posts that provide more correct and diverse aspects of a news headline. As a result, $QLH_{FBS} + PRI_{TP+TI}$ outperformed $QLH_{RBS} + PRI_{TP+TI}$.

5. CONCLUSION AND FUTURE WORK

In this study, we presented several approaches for identifying top news stories in the blogosphere. Our system utilizes the query likelihood and the news headline prior, based on the language model framework. For the query likelihood, we proposed several approaches to estimate the QLM and the NHLM. The QLM can be estimated using blog posts issued on the query day. We divided the blog posts into K clusters so that each cluster can accurately contain one of the various topics buried in blog posts. Then, we estimated the K number of QLMs respective to their clusters. We also proposed two approaches to choose the supporting relevant posts for a news headline. The posts were also able to cover many different aspects of the news headline.

Furthermore, for the news headline prior, we suggested two criteria, Temporal Profiling and Term Importance. They measure the importance of a news headline in two different ways. Temporal profiling measures it using the temporal information of blog posts relevant to the news headline. Term Importance measures it using the meaningfulness of terms in the news headline.

We obtained the best performance for TSIT by considering the query likelihood and the news headline prior at the same time. From experimental results, we can verify the the proposed approaches are effective in identifying top news headlines.

Many studies remain for future work. We used K-means clustering to model various topics buried in blog posts. It would be interesting to utilize several approaches such as PLSA [13] and LDA [4] to capture topics of blog posts. To improve the diversity of the supporting relevant posts, various ways such as MMR [6] are also worthy of research. Furthermore, we believe that various features such as comments or tags can be used to improve the performance when identifying top news stories.

6. ACKNOWLEDGMENTS

This work was supported in part by MKE & IITA through IT Leading R&D Support Project and also in part by the BK 21 Project in 2010.

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM 2009*, pages 5–14. ACM, 2009.

[2] J. Allan, M. E. Connell, W. B. Croft, F.-F. Feng, D. Fisher, and X. Li. Inquery and trec-9. In *Proceedings of TREC-9*, pages 551–562, 2000.

[3] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of SIGIR 1998*, pages 37–45. ACM, 1998.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[5] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In *Proceedings of SIGIR 2003*, pages 330–337. ACM, 2003.

[6] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR 1998*, pages 335–336. ACM, 1998.

[7] C. C. Chen, Y.-T. Chen, Y. Sun, and M. C. Chen. Life cycle modeling of news events using aging theory. In *Proceedings of ECML 2003*, pages 47–59, 2003.

[8] K.-Y. Chen, L. Luesukprasert, and S.-c. T. Chou. Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Trans. on Knowl. and Data Eng.*, 19(8):1016–1025, 2007.

[9] H. L. Chieu and Y. K. Lee. Query based event extraction along a timeline. In *Proceedings of SIGIR 2004*, pages 425–432. ACM, 2004.

[10] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR 2008*, pages 659–666, New York, NY, USA, 2008. ACM.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[12] Q. He, K. Chang, and E.-P. Lim. Analyzing feature trajectories for event detection. In *Proceedings of SIGIR 2007*, pages 207–214. ACM, 2007.

[13] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of SIGIR 1999*, pages 50–57. ACM, 1999.

[14] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3):14, 2007.

[15] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of SIGKDD 2002*, pages 91–101. ACM, 2002.

[16] P. Kolari, A. Java, and T. Finin. Characterizing the splogosphere. In *Proceedings of 3rd Annl. Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th Word Wide Web Conf.*, 2006.

[17] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *Proceedings of SIGIR 2004*, pages 297–304. ACM, 2004.

[18] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR 2001*, pages 111–119. ACM, 2001.

[19] Y. Lee, S.-H. Na, and J.-H. Lee. An improved feedback approach using relevant local posts for blog feed retrieval. In *Proceeding of CIKM 2009*, pages 1971–1974. ACM, 2009.

[20] Y. Lv and C. Zhai. Positional language models for information retrieval. In *Proceedings of SIGIR 2009*, pages 299–306. ACM, 2009.

[21] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC-2009 Blog Track. In *Proceedings of TREC 2009*, 2010.

[22] G. Mishne and M. de Rijke. A study of blog search. In *Proceedings of ECIR 2006*, pages 289–301. Springer, 2006.

[23] S.-H. Nam, S.-H. Na, Y. Lee, and J.-H. Lee. Diffpost: Filtering non-relevant content based on content difference between two consecutive blog posts. In *Proceedings of ECIR 2009*, pages 791–795. Springer-Verlag, 2009.

[24] C. Wang, M. Zhang, L. Ru, and S. Ma. Automatic online news topic ranking using media focus and user attention based on aging theory. In *Proceeding of CIKM 2008*, pages 1033–1042. ACM, 2008.

[25] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proceedings of SIGIR 1998*, pages 28–36. ACM, 1998.

[26] Y. Yang, J. Zhang, J. Carbonell, and C. Jin. Topic-conditioned novelty detection. In *Proceedings of SIGKDD 2002*, pages 688–693. ACM, 2002.

[27] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

[28] K. Zhang, J. Zi, and L. G. Wu. New event detection based on indexing-tree and named entity. In *Proceedings of SIGIR 2007*, pages 215–222. ACM, 2007.

[29] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *Proceedings of SIGIR 2002*, pages 81–88. ACM, 2002.