Event Intensity Tracking in Weblog Collections

Viet Ha-Thuc[1], Yelena Mejova[1], Christopher Harris[2], Padmini Srinivasan[1, 2]

The University of Iowa
[1] Computer Science Department
[2] Informatics Program
MacLean Hall, Iowa City, IA 52242
{hathuc-viet, yelena-mejova, christopher-harris, padmini-srinivasan}@uiowa.edu

Abstract

Event tracking is the task of discovering temporal patterns of events from text streams. Existing approaches for event tracking have two limitations: scalability and their inability to rule out non-relevant portions within texts in the stream 'relevant' to the event of interest. In this study, we propose a novel approach to tackle these limitations. To demonstrate our approach, we track news events across a collection of weblogs spanning a two-month time period. In particular we track variations in the intensity of discussion on a given event over time. We demonstrate that our model is capable of tracking both events and sub-events at a finer granularity. We also present in this paper our analysis of the blog dataset distributed by the conference organizers.

Introduction

The goal in event tracking is that of discovering and following postings, discussions or more generally, tangible expressions of interest about an event in a temporal text stream. Examples of text streams are newswire, less formal media such as weblogs, more formal publication streams such as those involving journals and books. A related goal is event intensity tracking that involves discovering variations in the extent to which an event is discussed over time, again as discovered from the text streams. Two distinct events may both be discussed over the same time duration and yet the underlying community(ies) may express varying levels of interest in the two events.

Our goal is to use the weblog dataset provided by the workshop organizers (Burton et.al, 2009) to develop methods for news event intensity tracking. Discovering such intensities of events in text streams such as weblogs or newswires may reveal useful insights about the evolution and shifts of interests of the underlying community. Collective interest in an event may wax and wane or even shift at some point to a different topic. Working with blog data in this paper, our methods are designed to show which news events are interesting to bloggers and which ones are not. They can also indicate starting and ending points of events (or at least their blog discussions), as well as prime times when the events are intensively discussed. Moreover, discovering the temporal intensities of events is an important beginning for various further analyses, such as extracting relationships among events and blog summarization.

There are two challenges of the tracking problem that our methods address. The first is regarding scalability. Recent tracking research (Mei and Zhai, 2005), (Zhou et.al. 2006) use variants of probabilistic topic models (Blei, Ng, Jordan 2003) to infer intensity of each event or topic. However, inference algorithms of topic models that are based on Gibbs sampling or Expectation Maximization techniques often require hundreds of scans over the dataset. This creates a computational challenge. When the dataset cannot fit into internal memory, often the case in practice, each scan takes many external memory accesses (e.g. disk accesses). Therefore, the inference algorithms do not scale well to large datasets (as for example the blog dataset provided by the workshop organizers). A second challenge is that even when weblog documents (posts) discuss an event of interest, they often contain substantial portions that are non-relevant to the event being tracked. For example the posts could include personal stories. Indeed, this type of problem; namely that only portions and not the entirety of a document may be relevant to a current topic of interest, may be observed in almost all varieties of texts and collections. But, given the somewhat informal nature of blogs we believe this problem is particularly present in our weblog data.

In response to these challenges and the opportunities offered by the workshop, we propose a scalable framework for tracking a set of user specified events. The framework includes two phases. In the first phase, for each given event e_k, we build a probabilistic relevance model $p(w|e_k)$, which determines probability of observing a word w in documents relevant to the event e_k. The model represents language used to write about the event e_k. For instance, a relevance model for the event *US Presidential Election* will likely assign high probabilities to words like *Obama, McCain, presidential, election, campaign*. In the second phase, we scan through documents in the temporal dataset. We use the relevance models of the first phase to extract relevant terms from the documents, and then compute intensities of

the events at different time stamps or intervals. Our approach properly handles documents that contain some portions that are relevant and others that are non relevant to an event. Moreover, it requires only a single scan over the dataset to track events and hence scales well with dataset size.

The rest of the paper is organized as follows. First, we describe our approach. Then, we present our experimental results and findings. In this experimental section, we first present our analysis of the workshop blog dataset. We then present our results with our event intensity tracking methods for sample events and sub-events. In the next section we review related work. Finally, we present our conclusions and plans for further work.

Methodology

As mentioned earlier, our framework includes two phases. In the first phase, for each given event e_k, we build a relevance model $p(w|e_k)$, which determines probability of observing a word w in documents relevant to the event e_k. This model represents language used by bloggers to write about the event e_k. In the second phase, we use the relevance models estimated in the first phase to track and compare the event intensities.

Phase 1: Estimating Relevance Models

In order to estimate the event relevance models, we first obtain a training set of documents from the text stream being analyzed using a search engine. For this we indexed the blog dataset using the search engine, Lucene (http://lucene.apache.org). For each event, we manually design a query of several (10 on average) keywords describing the event. Next, we use the query to retrieve top 100 documents, assume these to be relevant, and use these pseudo-relevant documents as training data to build the event's relevance model. The challenge in estimating relevance models from these training documents is that they could contain portions that are non-relevant to the event. As mentioned earlier, previous work strictly assumes that if a document d is relevant to a topic or an event then the entire document is deemed relevant to it. Although this assumption is convenient, it rarely holds in practice.

We address this challenge by estimating relevance models using the multiple-topic framework of probabilistic topic models. Here, although a document d may be relevant to a given event e_k, it could still have non-relevant portions. Some portions could pertain to background information shared by a large number of documents. Other non relevant portions, while specific to d, may be on themes other than event e_k. Specifically, each document d is hypothesized to be generated by a combination of three topics: the topic (event) e_k to which it is relevant, a background topic b representing the general language used in the document set, and a third topic $t\_o(d)$ responsible for generating themes that, though specific to d, are neither b nor e_k. Because our model considers this mixture of three topics, it is able to identify exactly those portions of a document that are truly relevant to the event e_k. In our work, these selected portions are the ones that contribute to the estimation of the relevance model $p(w|e_k)$. The other portions generated by topics b or $t\_o(d)$ are automatically ruled out.

For example in weblog domain, suppose that the training set for the event *US Presidential Election* includes d_1 a document (posting) about roles of media in the election and d_2 a document about the Iraq war as an issue in the election. The general background topic would be responsible for common words in English and common words in the weblog domain such as "blog", "entry", and "date." The distribution for the *US Presidential Election* event would likely give high probabilities to words like "election", "Obama", "McCain", and "Palin." Topic $t\_o(d_1)$ responsible for other themes in d_1 would likely generate words relating to media mentioned in d_1 like "media", "news" and "TV" while for d_2, $t\_o(d_2)$ would likely generate words such as "mosque", and "Islam." Observe that $t\_o(d)$ is specific to d.

Model Description. Formally, the proposed relevance-based topic model is a generative model describing the process of generating training sets of pseudo-relevant documents for a given set of K events $\{e_1, e_2 \ldots e_K\}$. It is described in Figure 1 using plate notation, where the numbers in the right-lower corner of the plates (boxes) indicate the number of repetitions in our inference algorithm (Figure 2) for each of the corresponding plates. One way to view the plates in the Figure is that the inner-most plate corresponds to a training document. The central plate corresponds to the set of training documents for a given event. The outer-most plate corresponds to the union of the training sets of all K given events. In Figure 1, $|D_{ek}|$ is the number of training documents of event e_k; N_d the number of tokens in document d; w_i the word identity of token i (note that a token is a specific occurrence of a word in a document). So, each token in document d in the training set for e_k could be generated by a latent topic z which spans over three topics: b, e_k, $t\_o(d)$. That is, each token in d could be generated by one of the above three topics. The total number of topics handled by the model includes the background topic b, K topics corresponding to the K given events, and document-level topics for all documents. For instance, suppose there are two events and each event has three training documents, the total number of topics T = 1 (for b) + 2 (for K) + 2*3 (for the 6 documents) = 9.

Given observable training documents for each of the K given events, we will infer the latent topic z_i generating token i ($1 \le i \le N$, where N is the total number of tokens in the union of all training sets). Intuitively, tokens from words frequently appearing in most training sets are likely generated by the background topic, tokens from words frequently appearing in the training set of an event e_k but not the other training sets are likely generated by this e_k, and tokens from words frequently appearing in only a

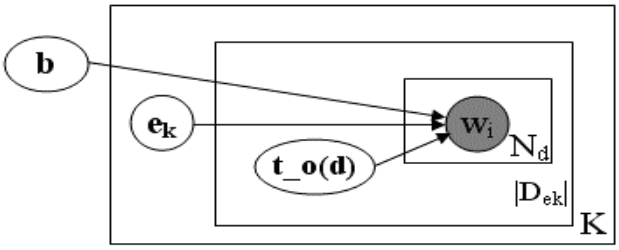

Figure 1. Relevance-based Topic Model

particular document *d* in the training set of event e_k but not in the other documents of this training set are likely generated by *t_o(d)*. Note again that *t_o(d)* is specific to document *d*. The inference algorithm is formally presented in the next section.

Inference. We need to infer the latent topic z_i generating token *i* ($1 \le i \le N$); the word-topic distributions $p(w|b)$, $p(w|e_k)$ for each event e_k and $p(w|t\_o(d))$ for each document *d*; topic-document distributions $p(z|d)$ where $z \in \{b, e_k, t\_o(d)\}$, the three topics generating the document.

The inference algorithm is presented in Figure 2. In Step 1, $p(w|\ b)$ is initialized by term frequencies across all training sets of the events. For each event e_k, $p(w|e_k)$ is initialized by term frequencies in the training set of this event. For each document *d*, $p(w|d)$ is initialized by term frequencies in *d* and $p(z|d)$ is uniformly distributed.

In Step 2, we iteratively estimate the parameters. In 2.1, z_i and w_i are the latent topic and word identities of token *i*, respectively; d_i is the index of the document in which the token appears. The latent topic of a token is estimated by taking into account the word identity of the token, document in which the token appears (context information) and the meanings of topics represented by their topic-word distributions. Instead of deterministically assigning the token to the latent topic, we probabilistically sample the latent topic with respect to the posterior distribution to avoid getting stuck in local optima. Given latent topics for all tokens, we update distributions by maximum-likelihood principle (2.2 and 2.3). In 2.2, the numerator is the number of times word *w* is assigned to topic *z* in the iteration $(s+1)$, and the denominator is the number of times topic *z* appears in this iteration. In 2.3, the numerator is the number of times topic *z* is assigned in document *d*, the denominator is document length. α and β are smoothing factors.

Phase II: Tracking Event Intensity

Given the relevance model for each event e_k, $p(w|\ e_k)$ we compute the intensity of this event at each time *t* with window size *w* using (1)

$$Intensity\ (e_i, t) = \sum_{d \in [t, t+w]} \log[p(d|e_i)] \quad (1)$$

The intensity is essentially the log-likelihood of the event in documents in the time period [t, t+w]. For each document, we use a threshold to rule out word tokens unrelated to the event, and sum over the remaining relevant word tokens in the document (equation 2). We include the term *log(h)* where *h* is *threshold* to normalize the log-likelihood to yield positive values.

$$\log[p(d|e_i)] = \sum_{w \in d: p(w|e_i) \ge h} \{\log[p(w|e_i)] - \log[h]\} \quad (2)$$

1. Initializing:
   - $p^{(0)}(w|\ b) = freq\ (w, b)$
   - $p^{(0)}(w|\ e_k) = freq\ (w, e_k)$
   - $p^{(0)}(w|\ t\_o(d)) = freq\ (w, d)$, $\forall d$
   - $p^{(0)}(z|\ d) = p^{(0)}(e_k|\ d) = p^{(0)}(t\_o(d)|\ d) = 1/3$, $\forall d$
2. For s = 0 to (S-1): (S is the number of iterations)
   - 2.1 Sample latent topic for each token
     - For i = 1 to N: (*N* is the number of tokens)
       - Sample $z_i^{(s+1)}$ from:
         $p^{(s)}(z_i=z|\ w_i, d_i) \sim p^{(s)}(w_i|\ z_i=z)\ p^{(s)}(z_i=z|\ d_i)$
         where $z \in \{b, e_k, t\_o(d_i)\}$
       - End for
   - 2.2 Estimate topic-word distributions
     - For z = 1 to T: (T is the total number of topics)
       - For w = 1 to W: (*W* is the number of words)
         $$p^{(s+1)}(w|\ z) = \frac{m_{z,w}^{(s+1)} + \beta}{\sum_{w'=1}^{W} (m_{z,w'}^{(s+1)} + \beta)}$$
       - End for
     - End for
   - 2.3 Estimate document-topic distributions
     - For d = 1 to |D|: ($|D|$ is the number of documents)
       - For each *z* in $\boldsymbol{X_d} = \{b, e_k, t\_o(d)\}$:
         $$p^{(s+1)}(z|\ d) = \frac{n_{d,z}^{(s+1)} + \alpha}{\sum_{z' \in X_d} (n_{d,z'}^{(s+1)} + \alpha)}$$
       - End for
     - End for
3. End for

Figure 2. Inference Algorithm

Complexity Analysis. In the first phase, we estimate relevance models for specific events using only training sets for these events. The total training set size (we use 100 documents/event) is significantly smaller than the whole dataset and easily fits into internal memory. In the second phase, in order to compute event intensities (simultaneously for all events), the dataset needs to be traversed only once. So, in total, it takes only one scan over the target dataset for tracking. Therefore, our approach scales well to the sizes of the datasets on which we would like to track events.

Experiments

We begin by presenting our preliminary analysis of the dataset provided. Our goal in this was to gain a better understanding of the nature (e.g. explore the presence of any specific patterns) of the data. We then track major news events across the two month period the dataset spans. Some of these events are then broken down to sub-events, which are also tracked. In particular we focus on event intensity tracking.

Dataset Characteristics

The data provided for ICWSM 2009 came from a weblog indexing service Spinn3r (http://spinn3r.com). This included 60 million postings spanned over August and September 2008. Some meta-data is provided by Spinn3r.

Each post comes with Spinn3r's pre-determined language tag. Around 24 million posts are in English, 20 million more are labeled as 'U', and the remaining 16 million are comprised of 27 other languages (Fig. 3). The languages are encoded in ISO 639 two-letter codes (ISO 639 Codes, 2009). Other popular languages include Japanese (2.9 million), Chinese/Japanese/Korean (2.7 million) and Russian (2.5 million). The second largest label is U - unknown. This data could potentially hold posts in languages not yet seen or posts in several languages. Our present work, including additional dataset analysis presented next, is limited to the English posts unless otherwise specified. In future work we plan to also consider other languages represented in the dataset.

In Fig. 4 the x axis represents document length in number of terms and y axis the number of documents with that length in the corpus. Visually the distribution reflects a logarithmic decrease except for the initial dip in blog postings with a length of less than 25 terms. The distribution peaks in the lengths between 30 and 40 terms. Counted were all the terms in the document, without the HTML tags, but with stopwords. It was noted by some reviewers of this dataset that a number of postings appear truncated; that is only a short preview is included in the content of the post. Since it is difficult to automatically determine if this is the case with a specific post or if it's just a short posting, the matter should be investigated further to determine the true extent of this problem. We plan to address this in future work.

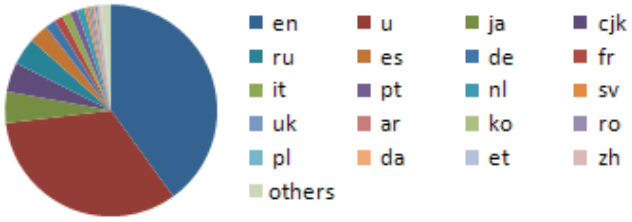

Figure 3. Language distribution

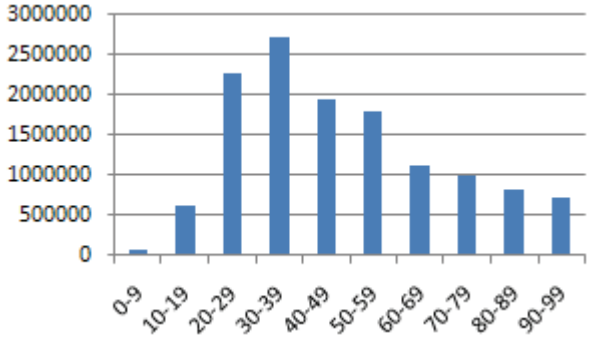

Figure 4. Document length distribution

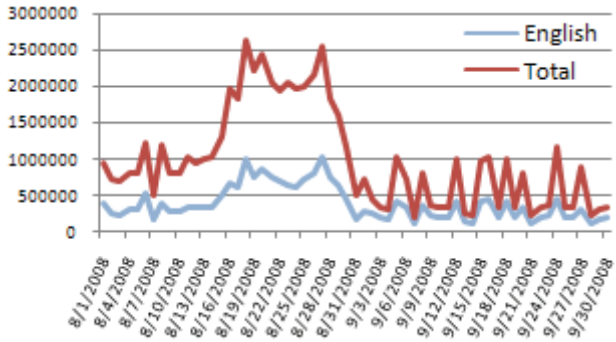

Figure 5. Publishing date distribution

None	9499357	Writing and poetry	21564
Life	94707	Politics	20818
Technology	75175	News and Politics	14650
News	61547	Blogging	13997
Technical Info	39575	Podcasts	5694
Music	28530	General	2748

Table 1. Popular categories

Interestingly blog activity is not uniform over the span of two months. Peak intensity of user posting occurs between August 16 and August 30. The distribution is non-uniform even if September is examined by itself.

One of the most intriguing metadata the posts came with is the Categories label. These are self-assigned by users, and collectively represent the collection's underlying folksonomy (Wal 2007). A substantial number of the posts

– 39% – do not list a category. Among the most popular categories are "life", "technology" and "news". Some popular categories do not provide much information, such as "general", and some are unusual, such as the category "caenorhabditis elegans" – referring to a type of roundworm. The difficulty with user-assigned categories is that they are non-standard and often overlap in coverage. For example, "news" may appear in other categories, such as "news and politics" or "sports news articles". The appearance of highly specialized categories in the list of the most popular suggests that this dataset may have interesting topical subsets reflecting some underlying hierarchical structure.

Another useful tag provided by Spinn3r is the Indegree tag, which is defined on their web site as "The raw number of inbound links to the blog since this blog has been part of our index" (spinn3r.com). A third of the data (32%) have an Indegree of 0, and 47% have Indegree of 5 or less. If a post is new and hasn't yet received much attention, it is likely to have a lower Indegree than an older post. Thus, there may be interest in determining how the Indegrees were calculated by Spinn3r. Complementing the Indegree tag, Spinn3r's Iranking tag represents the 'influence ranking' of the weblog. This is defined as a "function of how successful it is at appearing on Tailrank" (a weblog ranking system) (Tailrank Official Blog 2009).

In summary, collecting these statistics about the dataset has allowed us to gain better understanding, yet it brought up some questions about Spinn3r's collection policy and Iranking and Indegree calculations that would further illuminate the nature of this data.

Event Intensity Tracking

We demonstrate our model as described in the second section by tracking a set of popular news events across three subsets of the dataset provided. The first subset contains 8000 randomly selected documents; the second subset contains 75000 documents and the third 1 million documents. The second and third subsets were designed to determine the more "central" documents. We then examined the inlink data provided for each document and selected those above a threshold. Thresholds were 5,000 and 400 for subsets 2 and 3 respectively. Interestingly, results on all three subsets are consistent in terms of peaks and trends observed. In the interest of space, results presented in this paper are based on the third subset.

Ten news events are chosen to track over the dataset's two-month span (August and September 2008). These events are handpicked from a list of news events occurring during this period as provided by Wikipedia (wikipedia.org).

After running the inference algorithm of the relevance-based topic model, we examined a set of most interesting distributions for the distinct events. In the interest of space, only eight are shown in Fig. 6. Table 2 shows the top 10 words and the associated probabilities, which are estimated by the inference algorithm described in Fig. 2 for two events: the *US Presidential Election* and *Beijing Olympics*. We observe that proper nouns involved with a specific event have a higher probability overall.

Table 2. Relevance Models

US Presidential Election		***Beijing Olympics***	
word	*p(w| PE)*	*word*	*p(w| BO)*
obama	0.064	olymp	0.075
mccain	0.050	beij	0.071
palin	0.041	phelp	0.043
democrat	0.034	china	0.041
republican	0.030	game	0.040
clinton	0.019	gold	0.023
biden	0.018	august	0.021
convent	0.017	Michael	0.021
voter	0.015	medal	0.020
poll	0.014	ceremony	0.019

As anticipated, each topic peaks in intensity sometime within the actual duration of the event. The Beijing Olympics ran from August 8-24. Examination of Figure 6 shows a peak in the *Beijing Olympics* event on August 18, the day after the closely watched Michael Phelps won his record eighth gold medal.

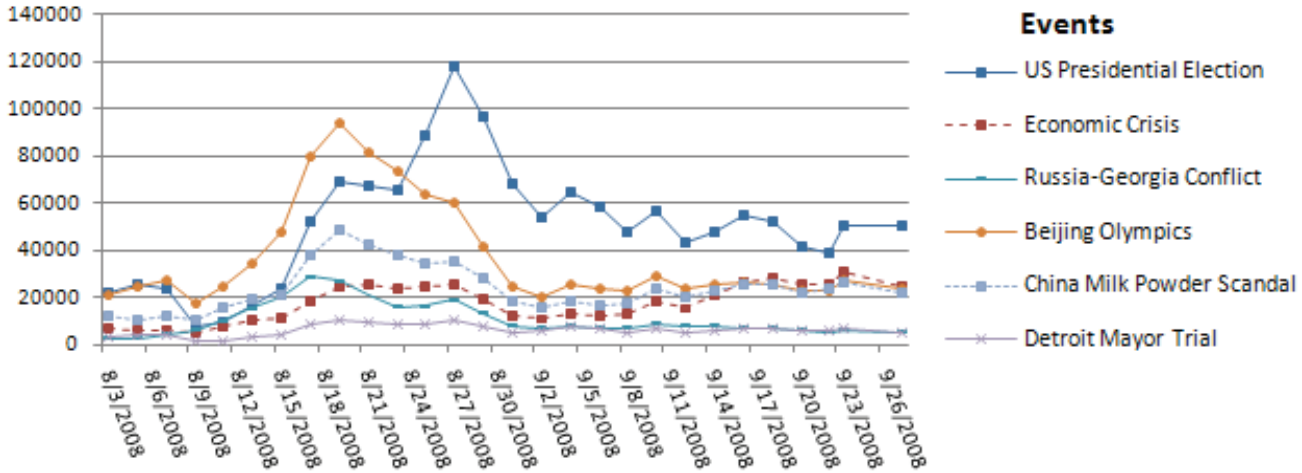

Figure 6. Temporal event intensities

Some events cannot be anticipated in advance. This gives us the ability to determine the latency time between an event's occurrence and the peak of its popularity. Four major Atlantic tropical storms and hurricanes struck over a two-week period beginning in late August (Fay – August 18, Gustav – September 1, Hannah – September 6, and Ike – September 13), but were not assigned names prior to their formation. For example, Ike, the most destructive of the three, was assigned its proper name only after it formally became a tropical storm on September 1(National Hurricane Center 2009). This allows us to track latency between the commencement of an event and the intensity of related blog posts. Some latency may be attributed to our use of a moving time window instead of examining each day in isolation. In our experiment, we use a window size of 5 days.

Other events may also reflect meaningful trends in relative interest accorded as tracked over a defined time period. The announcements of Presidential running mates of the two major political parties and their National Conventions occurred from August 23 through September 4. A marked increase in intensity of the *US Presidential Election* topic occurred during this time.

Intensity can be tracked before and after a major event to determine the event's overall effects in the blogosphere. Interest in the *Pakistan Impeachment* event increases after impeachment proceedings were launched against the President Musharraf of Pakistan on August 7, peaked on August 17, the day following the filing of formal impeachment charges, and decreased on August 20, two days after Musharraf's formal resignation. One may conjecture that this likely indicates a return to stability; however, if there was a plateau of increased activity after a negative event, one may conjecture outrage in the blogging community. These conjectures require further analysis into the nature of the posts; a goal suitable for follow up research.

Likewise, the *Russia-Georgia Conflict* event is nearly immeasurable until August 8, the day following the launch of a military attack in South Ossetia by Georgia (wikipedia.org). The peak of blogging intensity occurred on August 16, as ceasefire agreements were being signed by the two countries in conflict. The intensity of this topic then begins a slow decline. Following the formal recognition of the independence of South Ossetia and Abkhazia on August 26, the topic intensity sharply fades from the blogosphere.

Sub-Event Intensity Tracking

Sub-event tracking provides a more detailed look at what has happened within a particular event. For instance, given the temporal pattern of the *US Presidential Election* event in Fig. 6, users might be interested in teasing out the temporal trends of each party's convention. To achieve this goal, we apply our relevance-based topic model (see section Phase I) on just these two sub-events and their corresponding training sets identified through appropriate searches. Table 3 shows top representative words identified by our relevance model for each sub-event. It is worth emphasizing that at the sub-event level, the highly frequent words about the event like *convention*, *poll*, *voter*, (see last 3 rows of Table 2) and *campaign*… no longer play significant roles in representing sub-events. Our model achieves this by automatically inferring that these are background words (i.e. generated by the background topic b) since they frequently appear in the training documents of both sub-events. The relevance model of each sub-event focuses on unique word features distinctive to the sub-event. Hence, by varying the specificity of the background our approach allows us to track topics hierarchically.

Democratic Convention (DNC)		*Republican Convention (RNC)*	
word	*p(w| DNC)*	*word*	*p(w| RNC)*
obama	0.041	palin	0.073
dnc	0.040	republican	0.063
democrat	0.038	mccain	0.050
clinton	0.034	sarah	0.029
biden	0.034	rnc	0.025
denver	0.027	song	0.009
barack	0.021	paul	0.009
hillari	0.012	gop	0.009
bill	0.011	alaska	0.008
joe	0.011	hurrican	0.008

Table 3. Top 10 terms in sub-topic distributions for the *US Presidential Election* topic

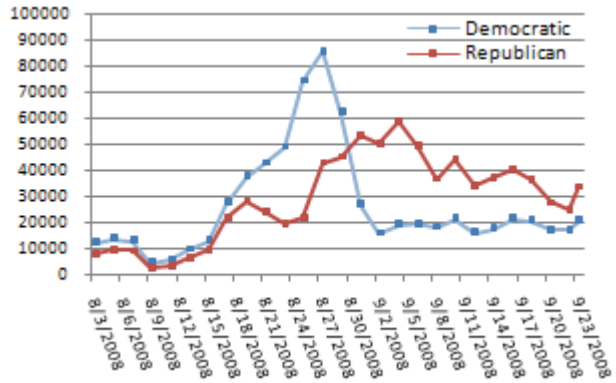

Figure 7. *US Presidential Election* Sub-Events

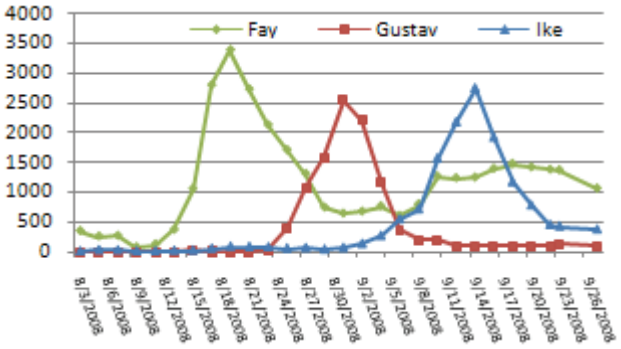

Figure 8. *2008 Hurricanes* Sub-Events

Given relevance models for the two sub-events, we use the tracking method described in Phase II section of Methodology to track these sub-events. Results are shown in Fig. 7. We see that the peaks of the lines correspond to the dates of the conventions (August 25–28 for Democratic National Convention and September 1–4 for the Republican National Convention) allowing clear separation of the more general US Presidential Election event into sub-events on each party's convention.

A parallel examination takes the general 2008 Hurricanes event and shows the intensity of three separate hurricane sub-events: Fay, Gustav and Ike. Fig. 8 presents their individual intensities. These three were chosen as the most destructive (in dollar value terms) of the four storms that occurred during August and September (National Hurricane Center 2009), making them the most likely to attract the attention of blog posters.

Tropical Storm Fay was named on August 15 and initially made landfall on August 18; likewise, Gustav was named on August 25, and made landfall on September 1; finally, Ike was named on September 1, and made landfall in Galveston, Texas on September 13 -- each of these dates closely correspond to the initial rise and peak values, respectively, as shown in the sub-topic intensity model.

Related Work

Our work in this paper is related to several existing directions below.

Topic Evolution Extraction (Mei and Zhai 2005), (Zhou et al 2006) encompasses two-fold aims: extract topics mentioned in a corpus and track the temporal intensities of these topics. Probabilistic topic models developed by other researchers (Blei, Ng, Jordan 2003) are often used for extracting latent topics in the corpus. Each topic is modeled by a probability distribution over words. The key advantage of this framework is it supports the fact that a document could be relevant to more than one topic and also that different tokens in the same document could be generated by different latent topics. We take advantage of these features in our work. Given the latent topic for each token, we compute intensities of topics by counting the frequencies of these topics in each time interval. However, probabilistic topic models have traditionally been used to *discover* topics. The problem has been that the topics discovered by topic models are *synthetic* i.e., they do not necessarily correspond to topics in reality as perceived by users. Another problem of this approach is that the inference algorithm for topic models takes hundreds of scans over the corpus, which is extremely expensive. So, this approach does not scale well with the sizes of corpora (Note that Zhou et al. use a corpus of about 100K documents, Mei et al. use two corpora containing 7,468 news articles and 496 scientific papers respectively). We base our approach on probabilistic topic models as it supports the multiple-topic framework, but we separate topic modeling and tracking into two phases. We track *given* topics and hence avoid the problem of dealing with synthetic and hard to interpret topics. In this process our approach explicitly takes into account relevance (or pseudo relevance) relationships between documents and the *given* topics. Moreover, our proposed approach requires only one scan over the corpus.

Topic tracking task defined by **Topic Detection and Tracking** (TDT) organizers (Allan 2002), (Allan et al 1998) is also related to our work in the sense of tracking *given* topics, which are specified by their relevant documents. (Note that this research effort predates the more recent topic evolution and extraction research described earlier). Given such input, the goal is to identify which unseen documents in the text stream belong to the topics. Solutions offered for this task typically consider documents as either wholly relevant or non-relevant to a particular topic. On the other hand, our probabilistic approach aims to extract the portions in unseen documents relevant to given events/topics to discover the temporal trends of these events.

In terms of the provided dataset, another direction related to our work is **Blog Mining**. The ever-increasing number of blogs online provides researchers with ample data for various types of analysis. Blogs have been used in opinion mining (Atardi and Simi 2006), corporate setting mining (Aschenbrenner and Miksch 2005), and more. Since blog posts are time-stamped, they are particularly suitable for temporal mining. Trend discovery is also a popular research area for tracking news events, people, and themes. One such trend discovery service is BlogPulse.com, which provides the most interesting posts, videos, and topics of the day (Glance, Hurst, Tomokiyo 2004). Social aspects of the blogosphere have also been studied through community evolution analysis (Kumar et al 2003). Topic tracking in blogs has also been explored (Me et al 2006) using a probabilistic approach similar to ours to extract common themes from blogs by including spatiotemporal theme model in their document representation. Once again, the previous probabilistic approaches make unrealistic assumptions, such as entire documents being relevant to a given set of topics.

In terms of **relevance modeling** (the first phase of our approach), several formal methods have been proposed such as the Robertson & Sparck-Jones probabilistic model (Robertson and Sparck-Jones 1988) and more recently, relevance–based language models (Lavrenko and Croft 2001). These approaches also estimate a relevance model for each topic by using a set of relevant or pseudo-relevant documents as training data. However, these relevance models do not exclude non-relevant parts in training documents as our model does.

Finally, we analyze the text from both local and global perspectives (topic vs. subtopic). It has been shown that local analysis is more useful to tasks such as query expansion than corpus-wide statistics (Jinxi and Croft, 1996).

Conclusions

In this paper, we propose a novel approach for event tracking that overcomes both issues of scalability and the inability to exclude non-relevant portions in documents. To demonstrate the efficiency of the approach, we hierarchically track popular news events with different levels of granularity over one million weblog documents spanning two months. The reported results indicate our approach is able to extract important temporal patterns about news events. Specifically, many of the trends depicted can be meaningfully explained in terms of the actual progression of an event over time.

One limitation of our approach is that in the tracking phase (the second phase of our approach), relevant words in unseen documents are identified individually without taking into account the relationship among words in the same documents. For instance, the word "China" in a posting about *Beijing Olympics* may also be considered relevant to the *China Milk Powder Scandal* event. This issue could be tackled if we take context of the whole document into account (similar to what we do in the first phase). This is among future directions for this work.

We would also like to explore the evolution of the interactions of the blogging community associated with each event. This will allow us to explore inter-relationships and overlap between blogging communities as they react to specific events.

Furthermore, we will look at the topic evolution by constructing the language models using a temporal window instead of the whole dataset. Then the topic would be represented as a set of language models each spanning a period of time. Other directions for future work have been identified in the paper such as extensions to consider blogs in languages other than English.

Acknowledgements

We would like to express our thanks to Robert Arens and Brian Almquist for their useful comments on this work.

References

Allan J. *Topic Detection and Tracking: Event-Based Information Organization*, Springer, 2002.

Allan J., Carbonell J., Doddington G., Yamron J., Yang Y. *Topic Detection and Pilot Study: Final Report*, In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

Apache Lucene. http://lucene.apache.org/java/docs/

Aschenbrenner A., Miksch S. *Blog Mining in a Corporate Environment*, Technical Report ASGAARD-TR-2005-11, Technical University Vienna, 2005.

Attardi G., Simi M. *Blog Mining through Opinionated Words,* In Proceedings of the 15th Text Retrieval Conference, 2006.

Blei, M., Ng, A., Jordan, M., *Latent Dirichlet Allocation*, Journal of Machine Learning Research, Vol. 3, 2003.

Burton K., Java A., and Soboroff I. *The ICWSM 2009 Spinn3r Dataset.* In Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), San Jose, CA, May 2009.

Feng A., Allan J. Hierarchical *Topic Detection in TDT-2004*. CIIR Technical Report, 2004.

Jinxi Xu, W. Bruce Croft. *Query Expansion Using Local and Global Document Analysis.* In Proceedings of SIGIR'1996. pp.4-11.

Glance N., Hurst M., Tomokiyo T. *BlogPulse: Automated Trend Discovery for Weblogs*, In Proceedings of the International World Wide Web Conference, 2004.

ISO 639 Language Codes. http://ftp.ics.uci.edu/pub/ietf/http/related/iso639.txt, February 2009.

Kumar R., Novak J., Raghavan P., Tomkins A. *On the Bursty Evolution of Blogspace*, In Proceedings of the International World Wide Web Conference, 2003.

Lavrenko, V., Croft W. B., *Relevance-based Language Models*, In Proceedings of the 24th ACM SIG International Conference on Research and Development in Information Retrieval (SIGIR), 2001.

Makkonen J., Ahonen-Myka H., Salmenkivi M. *Simple Semantics in Topic Detection and Tracking*, Information Retrieval Vol. 7 (3-4), 2004.

Me Q., Liu C., Su H., Zhai C. *A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs*. International World Wide Web Conference, 2006.

Mei, Q., Zhai, C., *Discovering Evolutionary Theme Patterns from Text – An Exploration of Temporal Text Mining,* In the Proceedings the 11th ACM SIGKDD, 2005.

National Hurricane Center. http://www.nhc.noaa.gov

Robertson, S., Sparck-Jones, K., *Relevance Weighting of Search Terms*, Journal of American Society for Information Science, 27, 1988.

Spinn3r. http://spinn3r.com/ Accessed February 2009.

Tailrank Official Blog. http://blog.tailrank.com/, February 2009.

Wal. V. Folksonomy. http://vanderwal.net/folksonomy.html, July 2007.

Wikipedia. http://www.wikipedia.org/

Zhou, D., Ji, X., Zha, H., Giles, L., *Topic Evolution and Social Interactions: How Authors Effect Research*, In the Proceedings of the 15th ACM CKIM, 2006.