A Relevance-based Topic Model for News Event Tracking

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

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

ABSTRACT
Event tracking is the task of discovering temporal patterns of popular events from text streams. Existing approaches for event tracking have two limitations: scalability and inability to rule out non-relevant portions in text streams. In this study, we propose a novel approach to tackle these limitations. To demonstrate the approach, we track news events across a collection of weblogs spanning a two-month time period.

Categories and Subject Descriptors
H.1.0 [**Models and Principles**]: General

General Terms
Algorithms, Performance, Design, Experimentation, Theory

Keywords
LDA, topic models, relevance models, event tracking

1. INTRODUCTION
The task of Event Tracking is about discovering temporal intensities of events in text streams such as weblogs or newswires. The discovered temporal patterns reveal useful information about the behavior of the various topics in the data sets. For instance, they could show which popular news events are interesting to bloggers, and which ones are not. They could also indicate starting and ending points of events (or at least their 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 news summarization.

One of the issues of the tracking problem is scalability. Recent work on this direction [2] use variants of probabilistic topic models [1] to infer intensity of each event or topic documents. However, inference algorithms of these topic models often require hundreds of scans over the dataset. Tracking specific events is also hindered by the fact that documents in the text stream (e.g. blog posts) often contain information that is non-relevant to the events of interest, such as personal stories.

In this study, we propose a two-phase framework for tracking a set of given popular events. 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. In the second phase, we scan through documents in the dataset. We use the relevance models estimated in the first phase to extract relevant terms from the documents, and then compute intensities of the events at different time stamps. So, our approach rules out non-relevant portions in the documents and takes only one scan over the dataset to track the events.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

2. PROPOSED APPROACH
2.1 Estimating Relevance Models
In order to estimate the event relevance models, we first obtain a training set of documents using a text search engine Lucene [4]. For each event, we make a query of several (10 on average) keywords describing the event. Then, we use the queries to retrieve top 100 documents, and use these pseudo-relevant documents as training data to build the relevance models for the events. Unlike previous relevance models, our model supports a crucial fact that a document relevant to an event e_k could still talk about topics other than e_k.

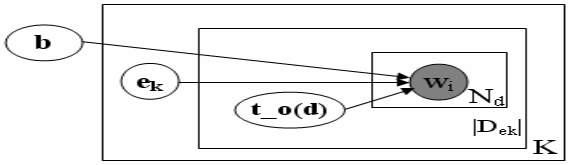

Figure 1. Relevance-based Topic Model

Specifically, each token in document d in the training set of event e_k is hypothesized to be generated by one of three topics: the e_k itself to which the document is relevant, a background topic b representing the general language, and a third topic $t\_o(d)$ responsible for generating themes other than e_k also mentioned in d (Fig. 1). 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 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 $t\_o(d)$. So, our model is able to extract the really relevant portions and rule out non-relevant portions generated by b or $t\_o(d)$ in training documents. Only the really relevant portions contribute to the estimation of relevance model $p(w|e_k)$. The three components above are inferred by Gibbs sampling, which is formally described in [1, 3].

2.2 Tracking
Given the relevance model for each event e_k, we compute the intensity of this event at each time t with window size s as in (1).

$$Intensity(e_k, t) = \sum_{d \in [t, t+s]} \log[p(d \mid e_k)] \quad (1)$$

The intensity is essentially the log-likehood of the event in documents in the time period [t, t+s]. For each document, we use a threshold to rule out word tokens non-relevant to the event, and sum over the relevant word tokens in the document (See formula 2). We add the term *log(threshold)* is to normalize the log-likelihood to yield positive values.

$$\log[p(d|e_k)] = \sum_{w \in d: p(w|e_k) \geq threshold} \{[\log(p(w|e_k) - \log(threshold)]\} \quad (2)$$

3. EXPERIMENTAL RESULTS

3.1 News Event Tracking

We demonstrate our model on a data set provided by weblog indexing service Spinn3r [5]. Ten news events are chosen to track over a two-month span (Aug and Sep 2008). These events are handpicked from the list of popular news events on Wikipedia [6]. Using the inlink counts provided by Spinn3r, we collect a subset of one million of the most popular blog posts. The results are displayed in Fig. 2.

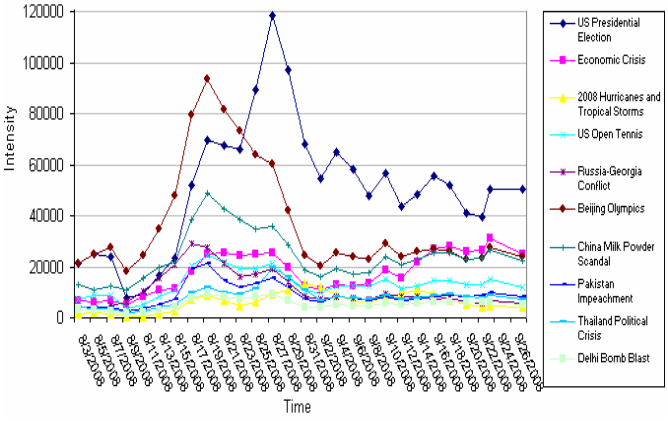

Figure 2. Temporal Event Intensities

Each topic peaks in intensity around the time of the event. For example, the Beijing Olympics started on Aug 8, and increased in popularity over the next few weeks. Three major Atlantic hurricanes struck over a two-week period in late August (Gustav – Aug 25, Ike – Sep 1, and Hannah – Sep 7). Additionally, the announcements of Presidential running mates of the two major political parties and their National Conventions occurred from August 23 through September 4, demonstrating a relative increase in intensity of US Presidential Election topic during those times.

3.2 Sub-Event 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 event *US presidential election* in Fig. 2, users might be interested in seeing temporal trends of each party's convention. To achieve this goal, we apply relevance-based topic model (Section 2.1) on training sets for the two sub-events. Table 2 shows top representative words for each sub-event. It is worth emphasizing that at sub-event level, the highly frequent words about the event like *convention, poll, vote, campaign…* should no longer play significant roles in representing sub-events. In our model, the background topic *b* will cover those words and the relevance model of each sub-event focuses on unique word features of the sub-events. So, by varying specificity of background topics our approach allows us to track topics hierarchically.

Table 1. Top 10 terms in sub-topic distributions

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

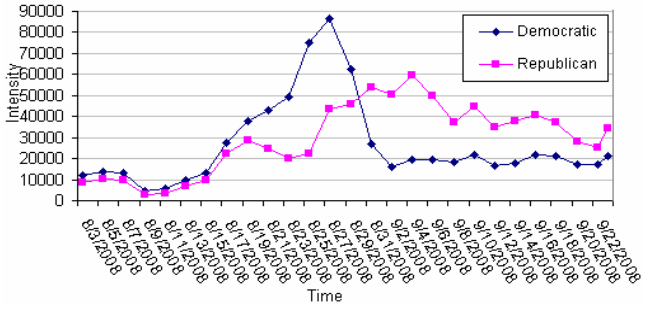

Figure 3. Temporal Sub-Event Intensities

Given relevance models for the two sub-events, we use tracking method in Section 2.2 to track the sub-events. The results are shown in Fig. 3. We see that the peaks of the lines correspond to the dates of the conventions (Aug 25–28 for DNC and Sep 1–4 for the RNC) allowing clear separation of the Presidential Election topic into subtopics on each party's convention.

4. CONCLUSIONS

In this paper, we propose a novel approach for event tracking that overcomes both issues of scalability and 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 confirm that our approach could be able to extract important temporal patterns about the news events.

5. REFERENCES

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

[2] Mei, Q., Zhai, C., *Discovering Evolutionary Theme Patterns from Text – An Exploration of Temporal Text Mining,* The 11th ACM SIGKDD, 2005.

[3] Ha-Thuc, V., Srinivasan, P. *Topic Models and a Revisit of Text-Related Applications*, The 2nd ACM PIKM, 2008.

[4] Apache Lucene. http://lucene.apache.org/

[5] Spinn3r. http://spinn3r.com/

[6] Wikipedia. http://www.wikipedia.org/