LeadLine: Interactive Visual Analysis of Text Data through Event Identification and Exploration

Wenwen Dou*[1], Xiaoyu Wang[1], Drew Skau[1], William Ribarsky[1], and Michelle X. Zhou[2]

[1]University of North Carolina at Charlotte
[2]IBM Almaden Research Center

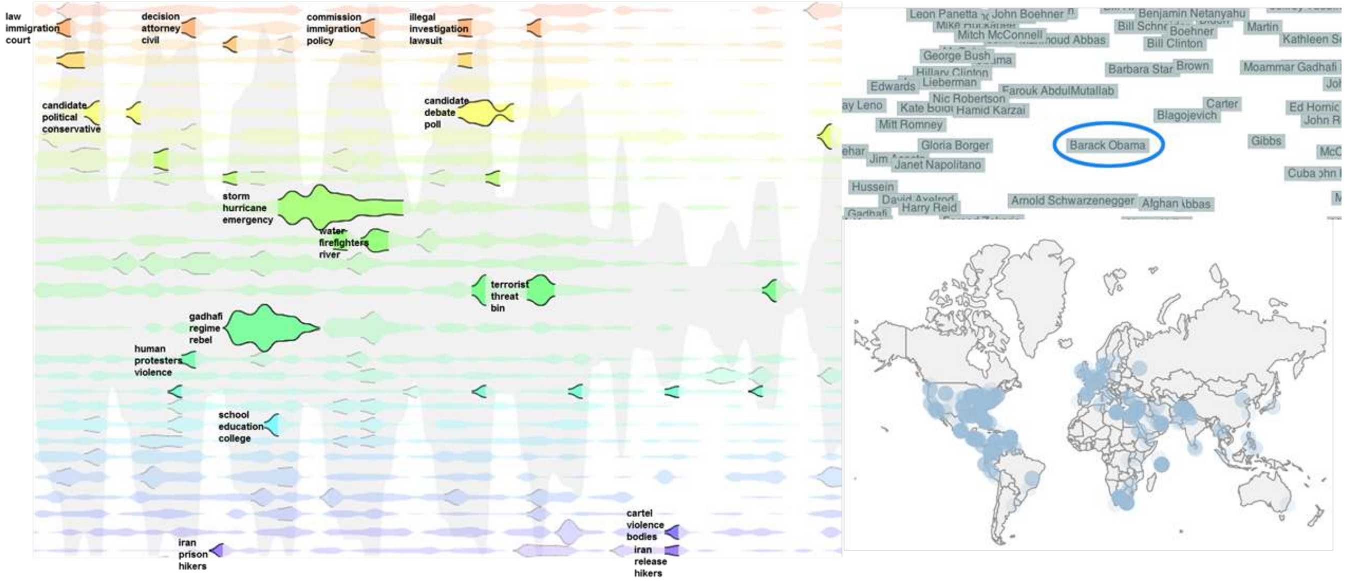


Figure 1: Overview of Leadline. Top right: people and entities related to President Obama (selected) are shown in the graph. Bottom right: locations mentioned in news articles related to the president. Left view: highlighted bursts indicate events that are related to President Obama.

Abstract

Text data such as online news and microblogs bear valuable insights regarding important events and responses to such events. Events are inherently temporal, evolving over time. Existing visual text analysis systems have provided temporal views of changes based on topical themes extracted from text data. But few have associated topical themes with events that cause the changes. In this paper, we propose an interactive visual analytics system, LeadLine, to automatically identify meaningful events in news and social media data and support exploration of the events. To characterize events, LeadLine integrates topic modeling, event detection, and named entity recognition techniques to automatically extract information regarding the investigative 4 Ws: who, what, when, and where for each event. To further support analysis of the text corpora through events, LeadLine allows users to interactively examine meaningful events using the 4 Ws to develop an understanding of how and why. Through representing large-scale text corpora in the form of meaningful events, LeadLine provides a concise summary of the corpora. LeadLine also supports the construction of simple narratives through the exploration of events. To demonstrate the efficacy of LeadLine in identifying events and supporting exploration, two case studies were conducted using news and social media data.

1 Introduction

Text data such as online news stories and microblog messages contain rich real-time information about worldwide events and social phenomena. In particular, news stories report ongoing development of events; microblogs capture people's comments and reactions to these events especially from a social aspect. Overarching patterns are lost in the here and now of the constant feed. Valuable information regarding major social and news events is hidden by the details. Therefore, methods to distill text data into not only meaningful overarching topics, but more importantly into triggering events are of great help to assemble the details into summarized information.

While summarizing large text corpora based on topical themes has received much attention, few have approached the problem from an event-driven perspective. Much work in the visualization community has been devoted to summarizing text data through representing topic evolution over time. Similar to these visual text analysis systems with a temporal focus such as ThemeRiver [22], TIARA [41, 38] and ParallelTopics [16], our approach organizes

*wdou1@uncc.edu

text corpora based on meaningful overarching topics. Yet unlike these tools, our focus is not visually presenting topical trends over time, but rather revealing indicators of events that trigger the major changes in temporal trends.

1.1 Formulating Events

Several questions are critical to identifying events from text corpora. How does one identify meaningful events given a text collection? What are the attributes that characterize an event? How does one automatically and systematically discover attributes that characterize an event from text collections? To address these questions, we first determine what comprises an event:

Merriam-Webster provides a general definition of an event as "a noteworthy happening and a social occasion or activity" [17]. In the Topic Detection and Tracking (TDT) community and event detection related research [8, 28], an event is defined based on its attributes as "something that has a specific topic, time, and location associated with it". From a storytelling standpoint, McKee refers to a story event as something that "creates meaningful change in the life situation of a character" [30].

Combining these definitions with our perspective on analyzing text corpora, we define an event as:

"*An occurrence causing **change** in the volume of text data that discusses the **associated** topic at a specific time. This occurrence is characterized by **topic** and **time**, and often associated with entities such as **people** and **location**.*"

For simplicity, we refer to < ***Topic, Time, People, Location*** > as four attributes of an event. These four attributes address common questions in investigative analysis: *When* did an event start and end? *What* is the event about? *Who* was involved? And finally *where* did the event occur?

1.2 Introducing LeadLine

Given our notion of events, we identify techniques and models that allow us to use computational methods to automatically extract events from text corpora. To discover events, we extract information regarding who, what, when and where through integrating topic modeling, event detection, and named entity recognition techniques. More specifically, we first organize text data such as news stories and microblog messages based on topical themes using Latent Dirichlet Allocation (LDA) [11]. This step provides topical information for events. To identify the temporal scale for events, we applied an Early Event Detection algorithm to automatically determine the length and "bursty-ness" of events. This step provides a beginning and an end to each event in time. To discover people and locations associated with each event, we first perform named entity recognition on the text corpora and associate the entities with events. With all four attributes explicitly modeled, our approach supports identification and exploration of events at the topical, temporal, and entity level.

To effectively communicate the event identification results, we developed a visual interface. The interface allows users to interactively explore events and, more importantly, to steer the event identification process to adjust the granularity of the detected events. In addition, organizing text corpora based on events provides basics for building narratives. This is a constructive format that describes a sequence of events [17] [30]. We extended LeadLine with the capability to examine narratives, which allows users to easily access and revisit their explorative findings.

Our approach provides an event-driven summarization of text data with exploratory capability for investigating the events based on who, what, when, and where. Specifically, our approach presents three contributions:

- A general process that couples topic modeling, named entity recognition, and early event detection techniques to identify meaningful events from text corpora.
- An interactive visual interface for exploring events based on the aspects of who, what, when, where, and further allowing interactive adjustment of event granularities.
- A narrative examination interface that allows users to report and revisit their findings.

2 Related Work

Three areas of research, namely event detection, topic analysis, and text visualization techniques, are the main inspiration for the design of LeadLine. Another thread of research on event structure in cognition provides background for using events as a summary.

2.1 Event Structure in Perception and Cognition

As Zacks and Tversky noted, the world presents nothing but continuity and flux, yet our mind has a gift to perceive activity as consisting of discrete events that have some orderly relations [43]. An event is a segment of time at a given location that is perceived by an observer to have a beginning and an end. People make sense of continuous streams of observed behavior in part by segmenting them into events.

People seem to segment observed physical activities into events effortlessly and simultaneously at multiple timescales [25]. However, little research indicates that the same skill applies to abstract continuous streams, such as topical streams derived from text corpora. Since an event is considered the unit for making sense of continuous activities, we argue that accurately identified and conveyed events may serve as a more natural representation for making sense of the activities.

2.2 Event Detection

The Biosurveillance community has long been investigating ways to improve clinical preparedness for bioterrorism [23]. Early surveillance tasks required continuous monitoring of massive quantities of multivariate data in order to identify emerging patterns [32]. Considering a particular type of health data, such as over-the-counter (OTC) medication sales, as a source for detecting events indicating disease outbreaks, Goldenberg et al. described a statistical system designed for timely detection of anthrax epidemics. This approach falls into the category of univariate methods which focus on detecting events from time series [20]. As a more general approach, Guralnik et al. [21] presented algorithms to dynamically determine the change points in time series data without prior knowledge of the temporal distributions.

Other disease surveillance systems take into account both temporal and spatial information. The system described by Sabhnani et al. [35] monitors daily data feeds from over 20,000 hospitals and pharmacies nationwide, including emergency department visits and OTC drug sales, to identify early events. Specifically, an expectation-based scan statistic approach is proposed to search for space-time regions for disease outbreaks. As an extension, Neil at al. further developed a "multivariate Bayesian scan statistic" (MBSS) [32] method for faster and more accurate event detection.

As efficient as the proposed event detection algorithms are, they lack the ability to handle text corpora, which may contain rich information about when the symptoms emerge and how they evolve over time. In this paper, our formulation allows us to transform textual data into multiple meaningful time series so that we can apply ideas from the Biosurveillance community for early event detection on text corpora.

More recently, researchers have shown interests in performing event detection on Twitter data. Petrovic et al. [34] presented a method to detect new events from a stream of tweets. The proposed algorithm, which is based on locality-sensitive hashing, enables first story detection on streaming data. However, the significance of the resulting events is not measured in the proposed method.

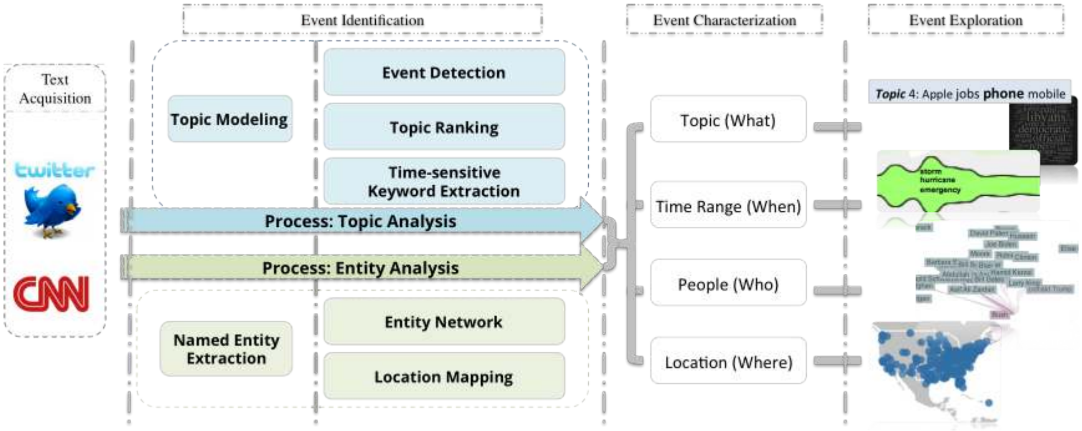


Figure 2: System Architecture of LeadLine.

To detect critical natural disaster events such as earthquakes in a timely fashion, Sakaki et al. [36] considered the event detection as a classification problem. One presumption of the approach is that users have to know what events to look for. In addition, to detect a new event, a new classifier needs to be built based on labeled examples of the new event. Another thread of event detection utilizes signal-processing techniques such as Wavelet transformation [42] to analyze word-specific signals in the frequency domain. Although the results are compelling based on manual examination, the detected events are describe using a few keywords (2 to 3) without any notion of "topics", therefore the individual event is difficult to interpret because of the lack of contextual information.

2.3 Event Extraction from News Sources

Topic Detection and Tracking (TDT) is a body of research and an evaluation paradigm that addresses event-based organization of broadcast news [8].The motivation for research in TDT is to address analysts' need to monitor new events from the high volume of broadcast news. The TDT has conducted several full-scale evaluations with research groups around the world by creating news corpora with the ground truth of events and topics. In his book that summarizes the TDT research program from 1997 to 2002, James Allan noted that much of the research on TDT has been tuning parameters to make certain information retrieval approaches more appropriate for events [8]. However, since an event is considered as something that happens at a particular time and location, Allan pointed out that existing approaches lack explicit modeling of temporal and geospatial dimension of the news streams. One exception is that Feng et al. [18] proposed incident threading as a news organizational infrastructure. The incident threading which takes into consideration of where, when and who is implemented in two algorithms, with one aching higher accuracy in identifying incidents and the other generating better links between incidents. Our approach also takes into account temporal aspects of events, as well as explicitly extracts named entities including people and location, and further allows end users to interactively make use of these aspects for inferring relationship between events.

2.4 Visualization of Events and Temporal Topical Trends

Multiple visualizations have been developed to visualize events from news sources. For example, in the time-series visualization CloudLines, Krstajic et al. proposed an incremental visualization technique that allows detection of visual clusters in a compressed view of multiple time series [24]. Luo et al. described a visual analytics approach which presents events in a river-like metaphor based on event-based text analysis [26]. To explore events reflected in microblogs, Marcus et al. presented TwitInfo [29], which is a system for summarizing events on Twitter. As opposed to a summarization approach, TwitInfo requires users to specify a Twitter keyword query to start exploring events related to the query. Different from the above systems, our approach provides investigative information regarding not only what (the events are about) and when, but also who is involved and where the events possibly happened.

More recently, Cui et al. [15] presented TextFlow for analyzing topical evolution patterns. In particular, TextFlow identifies critical events as topic birth/disappearance and topic merging/splitting. Our approach differs from their perspective of annotating topic evolution with critical events by centering on the identification of events, with topics as one aspect to characterize events.

Since topic and time are two essential attributes that characterize an event, previous work on providing temporal summaries of topics provided inspiration for our approach. There have been multiple examples of presenting topics along the time dimension. Wei et al. [41] introduced TIARA, a time-based interactive visualization system that presents topical themes in a time-sensitive manner. Similarly, in ParallelTopics [16], temporal evolvement of topical themes are also presented in a ThemeRiver visualization [22]. In our approach, we use a topic-based summarization method (topic models) to organize events over time. In addition, we allow users to interactively explore events based on who and where, as well as generate a narrative as a results of the exploration.

3 Analytics Architecture For Events Characterization

To extract information regarding events from text corpora, we integrate several techniques to identify < ***Topic, Time, People, Location*** > for each event. To extract meaningful topics and timespan for events, we leverage topic models for topical themes and an Early Event Detection method to identify a start and an end for each event (section). To extract information regarding who (people) and where (location), we perform named entity recognition and further analyze relationships between extracted entities (Section 6).

As shown in Figure 2, we categorize the identification of topic themes and timespan as topic-based analytics, in which we first extract topics from the input textual collection using Latent Dirichlet Allocation (LDA). We then apply a 1) topical-level event detection algorithm to automatically identify "bursts" as indicators of events labeled by a timespan; 2) topic ranking to facilitate the discovery of event relationships by placing bursts with similar topics nearby; and finally 3) time-sensitive keyword extraction that provides information regarding an event with a set of succinct keywords.

Complementing the topic-based analytics, our architecture also focuses on entity-based analytics by applying named entity recognition technique to identify people and location related with each event. Specifically, this process is centered on extracting key entities from the text data regarding who and where. More importantly, this process also reveals the relationship between the entities, leading to further characterization and associations of events.

As illustrated in Figure 2, visual representations are designed and tailored to both stages of topic-based and entity-based analytics processes. The interactive visual interface is an integral part of both analytics processes to bridge users and the complex analytics results. With the interactive visualization interface, LeadLine supports interactive exploration of events from various aspects (who, what, when, where), as well as allows users to interact with the underlying analytics algorithms to partially steer the process of detecting events from text streams.

In the following sections, we will detail the algorithms and processes used in both topic and entity analysis, alongside the visual designs that are tailored for identification and analysis of the events.

4 Data Acquisition and Preparation

To demonstrate the generalizability of our algorithms and their applicable domains, we have applied our analytics architecture on two types of text data: CNN news and microblogs from Twitter. While both sources contain rich information reflecting major real-world events, the primary reason for selecting the two data sources is because of their different editorial styles and the latency for responding to an event. In particular, content from news media (CNN and others) are edited by professional journalists, usually centered on a specific topic with some background. Individual tweets, on the contrary, contain mostly the unedited commentaries without much contextual information [9]. These different text sources provide various benchmarks to help us evaluate our analytic architecture.

While both sources are in the public domain, there are no datasets readily available or that can be shared due to privacy policies. Therefore, we have extended our existing scalable data architecture to acquire news and tweets using customized crawling strategies. The details of our data architecture can be reviewed in previous work [40].

News Data Acquisition: In particular, our current work extended on the previous architecture by adding the news article crawler. Both historical news as well as up-to-the-hour news are acquired by our customized webpage crawlers and our RSS daemons, respectively. Both methods employ universal programs that attempt to crawl an entire web domain, download all the webpages, extract all textual articles, parse article time information, and finally normalize articles by removing HTML formatting and noise (e.g. advertisement and external website links). More specifically, our news crawler is built with python, and Apache Nutch [5] implementation. The data is stored into our HBase data structure for fast access and MapReduce [4] based data cleaning and processing. Using these crawlers, we could retrieve and clean news articles dated back to 2004 ($\sim$102573 articles) within several hours of the crawling process.

Twitter Data Crawling: Microblogs from Twitter are also collected from dual crawling processes. The primary process utilizes our MapReduce parallel data crawler, which is interfaced with the Internet through multiple independent crawlers. Each of the crawlers constantly collects social media data from various public domains and dumps it into HBase. Specifically, we have created crawlers to connect to Twitter's public "Garden-hose" API to collect 1% of tweets, which is a statistically significant sample of all content from Twitter [6]. In addition, we have also experimented with an Online Social Network (OSN) graph-based crawling mechanism to extend our practice using Apache Nutch. Inspired by Catanese et. al [13], the concept for this crawling process is accustomed to the nature of OSN which can be represented as graphs, with nodes representing users and edges denoting connections. We perform a breadth-first search using Nutch to acquire Twitter public user-graphs and capture the tweets through their webpage portal for broader streams. As a result, we were able to collect over 5 billion tweets from all languages over the course of 3 months, providing a reliable database for evaluation purposes.

5 Topic-Based Event Analysis and Visualization

Topic-based analytics is crucial for event characterization in terms of revealing ***topics*** and ***time***. In this section, we introduce algorithms to extract topical and temporal information with regard to an event, as well as visual representations that communicate the topical and temporal aspects.

5.1 Extracting Topics from Text Data

We start by organizing text streams based on topics. Given a text corpus, there are multiple ways to extract meaningful topical themes. Among them, probabilistic topic models [10] are considered to be advantageous compared to traditional vector-based text processing techniques. In LeadLine, we first employ the most widely used topic model, LDA [11], to extract semantically meaningful topics from text corpora. LDA generates a set of latent topics, with each topic as a multinomial distribution over keywords, and assumes each document can be described as a probabilistic mixture of these topics [11].

5.1.1 Topic Streams

In addition to topics, another important attribute of an event is time. In order to highlight the temporal dimension, we organize the extracted topics along the temporal axis. Considering each topic as a stream that evolves over time, the calculation of each topic stream is performed by computing a spline based on the volume of textual information associated with the topic in each time frame. A time frame is a unit that the texts are temporally aggregated upon. The time frame unit varies by datasets and tasks, ranging from minutes for social media data to days for news stories.

5.1.2 Visualization of Topic Streams

To visually represent topical themes over time, we adopt a flow-like visual metaphor in which each stream (row) represents a topic and the height of each topic changes as the amount of textual information related to the topic varies (Figure 1 left) We represent each stream in a separate row for the later discovery of events within each topic stream.

However, with each topic stream represented separately, the visualization does not capture the overall trend of how all topics evolve over time. In order to provide the overall context, a ThemeRiver representation is placed in the background of the visualization. Therefore repetitive patterns exhibited by a text stream (such as a weekly pattern for news stories) are still portrayed.

5.2 Automatically Detecting Events in Topical Streams

A key technical contribution of this paper is automatically identifying temporal peaks in topics as indicators of events. To detect events from topic streams, we consider each stream as a time series. Each time series is computed by aggregating texts related to the topic within every timespan.

To automatically discover events from these topical time series, we apply an early event detection (EED) method to look for bursts in topic streams. More specifically, we adopt the cumulative sum control chart (CUSUM) widely used for change detection [31]. CUSUM is good at detecting shifts from the mean in a time series by keeping a running sum of "surprises". We implemented our version of CUSUM for detecting changes in topical volume. For each

```
Algorithm 1: CUSUM
Input: topic time series X collected at time $i = 1, \ldots k$
Steps:
    1. Calculate the mean $\mu$ and standard deviation $\sigma$ of the time series;
    2. Calculate running sum S from the starting time frame
    $S_1 = \max[0, x_1 - \mu]$
    $S_i = \max[0, S_{i-1} + x_i - \mu]$.
    3. When $S_k$ exceeds a threshold H (in units of $\sigma$), event alarm triggers. The beginning and end of the event are determined by the nearest positive $S_i$ to the triggering point.
```

topic stream, the algorithm keeps a cumulative sum of topic volume at each timespan that are higher than the mean topic volume. Once the cumulative sum exceeds a threshold, the event alarm triggers. Our implementation of CUSUM is shown in Algorithm 1.

The result is a set of automatically detected events within all topic streams, with each event labeled by a start and an end along the time dimension. Our approach of detecting events on time series is generalizable. It could be used as an extension on any time-series visualization to highlight changes as indicators of events.

5.2.1 Visualizing Detected Events

To visually represent the detected events, we use a combination of contours and highlights to indicate events within each topical stream. As shown in Figure 1 left, we provide an overview of events by drawing contours in topical streams to signal events. The time span of the contours is determined by the event detection results.

In addition to using contours to provide an overview of all events, LeadLine supports highlighting events of interest through user interaction. For example, a user could interactively highlight events related to a specific named entity such as a person or location (the extraction of entities is introduced in Section 6). To provide details on demand, LeadLine allows users to access documents (news or microblog messages) that discuss a highlighted event by clicking on the event.

5.2.2 Interactive Event Detection

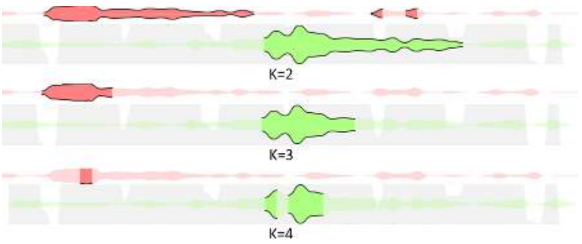


Figure 3: Comparison of different granularities of events. K indicates detecting changes K*standard deviation above the mean.

Although one advantage of our event detection algorithm is to automatically discover events that trigger topical bursts, we still want to provide end users the ability to specify how big a change warrants an event. By adjusting the scale of the "change", users will be able to view either finer or coarser-grained events.

In step 3 of Algorithm 1 , the event alarm triggers when the difference between the mean and running sum S is greater than a threshold. The threshold is usually measured by a fixed number of standard deviations (K) [33]. In LeadLine, we allow users to interactively adjust K between one and four standard deviations above the mean of one topic stream. When K is smaller, the algorithm is able to detect smaller shifts from the mean, leading to more events that cause less drastic change to be discovered. On the other end, when K is larger, the algorithm is only able to detect big shifts from the mean so only larger "bursts" are detected. Figure 3 shows three event detection results with varying thresholds. Once a user adjusts the K value using the slider in the interface, LeadLine re-runs the event detection algorithm to produce new event results. Through allowing users to interactively manipulate the fineness of events, LeadLine provides overviews with more or less details for given text corpora.

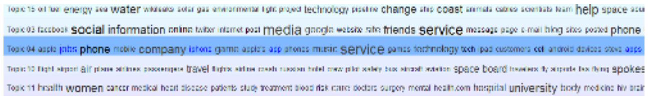
5.3 Topic Ranking

When placing the topics in our visualization, we want similar topics to be close so that the events later derived based on topics are also posited in proximity. Since LDA does not explicitly model the relationship between topics, we rank the topics based on their similarity measured by Hellinger distance. Specifically, Hellinger distance measures the distance between two probability distributions of topics over the entire vocabulary in the text streams:

$$topic - similarity_{i,j} = \sum_{v=1}^{N} (\sqrt{\beta_{i,v}} - \sqrt{\beta_{j,v}})^2 \qquad (1)$$

β is the probability of the ith topic over term v in the vocabulary, and N denotes the vocabulary size.

5.3.1 Visualization of Topical Content



Topic 15 oil fuel energy sea water wildlife solar gas environmental light project technology pipeline change ship coast animals cables scientists team help space

Topic 03 facebook social information online twitter internet post media google website site friends service message page e-mail blog sites posted phone

Topic 04 apple jobs phone mobile company iphone game apple's app phones music service games technology tech ipad customers cell android devices steve apps

Topic 10 flight airport air plane airlines passengers travel flights airline crash russian hotel crew pilot safety bus aircraft aviation space board travelers fly airports tsa flying spokes

Topic 11 health women cancer medical heart disease patients study treatment blood risk care doctors surgery mental health.com hospital university body medicine hiv

Figure 4: Topic Cloud with the topic related to mobile device/technology highlighted. Keywords in blue are the time-sensitive keywords in a certain time span. The topics are extracted from CNN news corpora (Aug 15 - Nov 5, 2011).

To represent topics in the form of a set of keywords, we develop the Topic Cloud view based on tagCloud representation (Figure 4). In the Topic Cloud, the size of each keyword reflects its number of occurrences in all topics. Since topics are ranked based on their similarities, the Topic Cloud view starts with a random topic at the top and places the topic with the highest similarity score next to the one above. As a result, similar topics are placed next to each other visually. To color the topics in Topic Cloud, we choose colors from the HSV (hue, saturation, value) space by dividing the space based on similarities between the list of topics. As a result, similar colors are assigned to topics with greater similarity. The same color scheme is also applied to the Topic Stream View (Figure 1 Left).

5.4 Time-sensitive Keyword extraction

To accurately summarize what an individual event is about, we need to re-rank the keywords within a topic by considering which keywords are more prominent given the timespan of the event. In order to do so, within each topic, we first extract the most prominent keywords for each time frame. We adopted the keyword-ranking algorithm that Wei et al. proposed [41] to re-rank terms in each topic based on time factor. More specifically, we followed Algorithm 2 to extract time-sensitive terms given a topic-term distribution matrix, and the number of desired terms for each time frame N. The input for this algorithm is a text corpus divided into sub-collections using time frame and topics. Each sub-collection contains only the documents focusing on the specific topic and also published during the time frame. The algorithm follows a TFIDF heuristic to determine time-sensitive terms: (a) if a term occurs frequently in the sub-collection, it is important; (b) if the term also occurs in many other sub-collections, the importance is discounted.

```
Algorithm 2: extract time-sensitive terms
Input: topic-term distribution matrix $\phi$; desired number of keywords per time frame N
Steps:
  1.
  for each topic i do
      for each time frame t do
          Identify a collection of documents $D_{i,t}$ focusing on topic i from entire text stream;
      end for
  end for
  2.
  for each term W in topic i from $D_{i,t}$ do
      calculate term frequencies TF
  end for
  3. Re-rank the TF scores with topic-term probabilities: for each term $W_m$, weight(W$_m$) = $\lambda_1 \frac{TF_{i,t,m}}{\sum_t TF_{i,t,m}} + \lambda_2 \phi_{i,m} \log \frac{\phi_{i,m}}{(\prod_{k=1}^{K} \phi_{k,m})^{1/k}}$
  4. Within each topic and time frame, select the top N terms as time-sensitive terms.
```

The extracted time-sensitive terms could not only aid the exploration of the text collections temporally, but also provide accurate labels for discovered events. The association of events with corresponding time-sensitive keywords is accomplished through highlighting time-sensitive keywords (in the topic view) when a topical burst is examined by users. But for the clarity of the figures, we manually annotated events with the time-sensitive keywords.

5.4.1 Interaction and View Coordination

The Topic Cloud view and Topic Streams are coordinated through user interaction. Hovering the mouse over one topic stream in the Topic Streams would highlight the corresponding topic in the Topic Cloud. Hovering the mouse over a specific timespan within a topic stream would highlight the time-sensitive keywords. Double-clicking the same area further allows users to access the content of the documents (as later briefly shown in figure 8). In addition, events from different topics can be highlighted by selecting named entities. Details of extracting named entities from events are introduced in section 6. In addition to view coordination, one important interaction LeadLine provides is choosing the "fineness" of the detected events. By simply moving the slider, a user can determine the scale of the examination.

In summary, section 5 presents algorithms that extract topical and time information that characterizes events. Visual representations of events are also designed based on the algorithmic results. The visualization not only allows interactive examination of the detected events, but also provides users the power to steer the event detection algorithm

6 Entity-based Event Analysis and Visualization

Complementing the topics and time attributes that are depicted in topic analysis, entity analysis is yet another crucial component. The named entities further reveal relationships between events by connecting them with information regarding people and location.

The primary goal for our name entity extraction is to extract people and locations from the textual streams and associate these entities with events. As illustrated in Figure 2, our named entity extraction identifies the named-entities once the data has been cleaned and prepared. The current entity extractor uses the LingPipe package [3] with Statistical Chunking and customized Dictionary-based Extraction. The three categories extracted are people, location, and organizations. This information is then inserted into the HBase storage platform, and attached to each corresponding news article. Note that location entities have also been enriched with geo-tags that are acquired alongside the contents.

While our current approach identified a sufficient amount of meaningful entities for news articles, the initial results from this approach on microblogs was fairly noisy. Similar findings have been identified in research in the AAAI community on the same issue, even with more state-of-the-art part of speech (POS) tagging [19]. We believe improvement on entity-based analytics for tweets is much needed, and do appreciate other representative research by MacEachren et al. on extracting entities from tweets [27]. Given the focus of this paper, we will focus on entities in news and demonstrate the utility of associating entities and their relationships in characterizing events and constructing narratives.

6.1 Entity Graph and Geo-Mapping

Entities are primarily used in LeadLine to connect multiple events into a meaningful narrative structure. Similar to the term used in Jigsaw [39], our distributional similarity between entities is computed based on the commonality of their contexts of occurrence in text. In particular, the association between entities is determined based on the co-occurrence of entities within the corresponding text. For example, two entities that co-occur in at least two news articles are considered connected. This gives us a contextual background and a baseline to depict the interplay between entities. Similar to the interactive event detection process, LeadLine enables users to interactively adjust the granularity of entity correlations based on their co-occurrence.

As shown in Figure 1 top right, we have created an interactive entity graph visualization to represent the connection and correlation between entities. A basic force-directed graph is utilized to allow users to dynamically explore the graph by showing and hiding links and entities. Instead of visualizing all the entities within a large graph, LeadLine constructs the graph content based on user's selections of topics and time factor as well as the types of entity. Different types of entities are color-coded. For example, people are shown in blue while organizations are shown in green.

Furthermore, location information is plotted on an interactive map to reveal the spatial distribution of entities contained in the text corpora (see Figure 1 bottom right). The geospatial view and entity graph are coordinated through user interactions. Selections within one view is used to filter the associated entity in all views. Through such view coordination, different aspects of the entities can be examined simultaneously.

In summary, LeadLine is designed to utilize results from named entity extraction to provide people and location information for exploring and inferring relationships between events. Interactive visualizations are further used to enable users to navigate through implicit connections between entities.

7 Uncovering Narratives By Associating Events

As a tightly coupled topic and entity analysis and interactive visual interface, LeadLine enables the user to perform investigative analysis of events inside a text corpus with a guided exploratory environment. While this environment is useful in interactively examining the four attributes < ***Topic, Time, People and Location*** > that comprise an event, summarizing the events into an intuitive narrative is more convenient for sharing and reporting.

Inspired by Segel and Heer's narrative genres [37], we have developed a Partitioned Poster style interface designed to assist in summarizing events into narratives (Figure 5). The interface is implemented using Bostock's D3.js library [12] to increase accessibility and to allow further exploration of the data by remote users.

The narrative poster can be considered as a simplified version of LeadLine. The hierarchy of information in the narrative poster is: 1) Entity, 2) Event, 3) Topic. The entity list can be very long for a large text corpus, so an autocompleted text entry field is used for

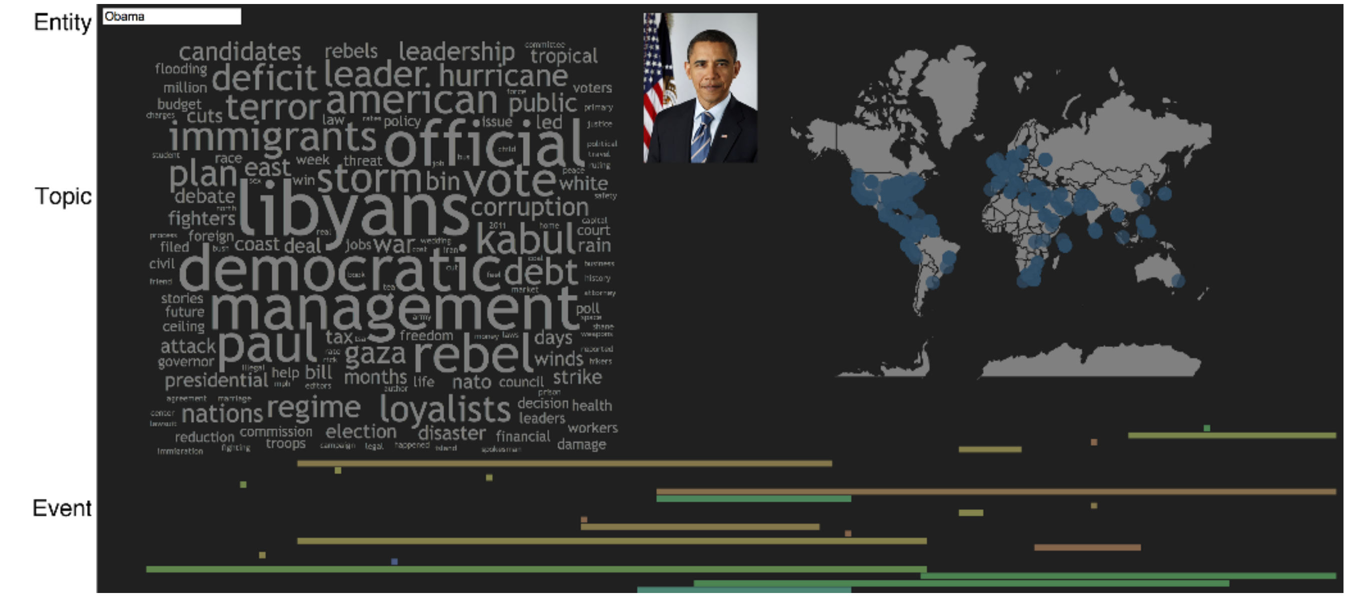


Figure 5: A narrative poster constructed from news events related to President Obama. Top left: a tag cloud summarizing all events related to the selected entity (president Obama); top right: locations mentioned in the events; bottom: event timeline.

entity selection. After selecting an entity, the user is presented with aggregate wordcloud and map views and a timeline view showing the events. The wordcloud view displays time-sensitive keywords of all shown events. The map view provides geolocations mentioned in the events. The timeline view gives a scope of events' durations and overlap (simplified version of the event bursts in the Topic Stream view).

The three views are coordinated with the timeline view being the main point of interaction. Hovering over an event in the timeline view causes the map and wordcloud views to display only keywords associated with that event rather than the aggregate for the entire entity. The Partitioned Poster format provides an intuitive way for telling stories. The simplified views and interactions provide users a succinct temporal summary of a given corpus, as well as details about events of interest.

8 Case Studies

To evaluate the efficacy of LeadLine in automatically detecting events from text data and facilitating exploration of events based on topics, time and named entities, we conducted case studies using real-world textual data from CNN news [14] and Twitter [6]. For the case studies, we recruited 8 users to explore the events from the CNN and Twitter datasets. We provided training to our participants regarding the LeadLine system before the exploration. During the exploration, we asked the participants to think-aloud and the experimenter took notes of the findings. In the end, we conducted a post-study interview to collect feedback about the visualization.

8.1 Case Study 1: Exploring the Occupy Wall Street Movement

In this section, we first describe the dataset we use for the study. We also characterize challenges in analyzing such a large-scale social movement. We then present how LeadLine assists users in discovering major events and understanding the OWS movement through exploring these events from multiple aspects.

8.1.1 Data set and Background

The Occupy movement is an on-going series of demonstrations and is known for using social media for publicity and organization. The Occupy movement is long-lasting and wide-spread without central leadership. This creates challenges in understanding the direction of the movement. In addition, a wide range of goals were reported [7], including more jobs, more equal distribution of income, and bank reform. Given the prominent use of social media in the Occupy movement, it makes sense to analyze the voices from protesters and citizens.

To provide an overview of the OWS movement, we filtered our tweet collection described in Section 4 using the hashtag #occupy. The resulting dataset contains more than 100,000 tweets from Aug 19 to Nov 01. Given the length of the dataset, the time unit in the visualization is set to 6 hours, allowing users to see how an event evolves within a day. 15 topics were extracted using LDA.

8.1.2 Exploring the OWS Movement Through Major Events

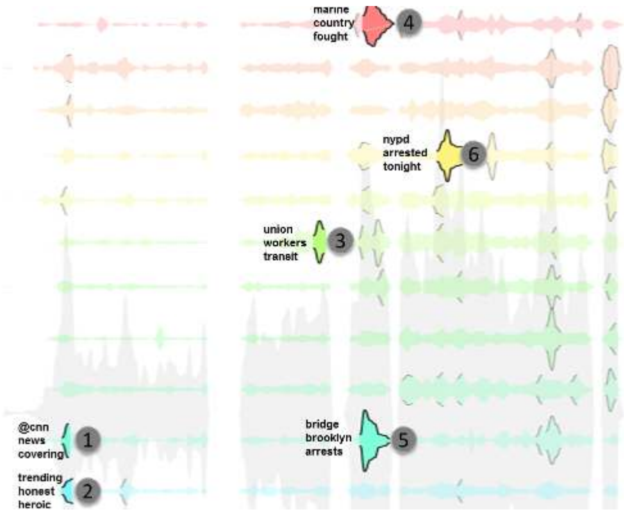


Figure 6: Major events discovered by the participants. Only events discussed in section 8.1.2 are highlighted.

Before the studies, we asked our participants if they were familiar with the OWS movement. All of them indicated that they had heard or read about the movement from news media, but none had a good understanding of the movement. In this task, we asked the participants to identify major events of the OWS movement and report what they have learned through exploring with LeadLine.

The participants usually started by quickly going through all topics in the Topic Cloud to gain an overview of the topics summarizing the tweets. Some participants then mainly focused on the Topic Stream View as they explored major events, while glancing at the Topic Cloud for information regarding a topic or a specific event. To make sense of what event triggered a certain topic burst, the participants clicked on the burst and accessed all tweets contributing to the burst. As they performed explorations in LeadLine, the experimenter took notes of their findings. For reporting purposes, we collected major findings from our participants, shown in Figure 6. Here we describe a few common discoveries from the participants:

Topic bursts #1 and #2 clearly mark the beginning of the Occupy movement on Sep 17, 2011, with tweets noting the coverage from news media on the first day of OWS gathering (burst #1), and the trend of tweets mentioning the OWS movement picking up steam on Twitter (burst #2).

Topic bursts #3 and #4 mark the events of two forces joining the OWS movement to support the protestors. On Sep 28th, the board of the NYC Transport Workers Union of America voted to support OWS, which splashed a lot of celebratory tweets (burst #3). Similarly, on Oct 1st, the burst of tweets (#4) was triggered by the marines joining the movement to protect protestors from the police.

Topic bursts #5 and #6 indicate marches by the occupy protesters. Topic burst #5 is triggered by protesters marching across the Brooklyn Bridge on Oct 1st. Some tweeted live updates of the march and of police arrests of the protestors, while others complained that there was not enough media coverage of the event. On Oct 5th, more than 10,000 protesters marched near Zuccotti Park. Tweets within initial time of the burst were about the march and the massive number of people involved. However, tweets in the later part of the burst were ranting about protesters being pepper sprayed, beaten and arrested by police. The event clearly triggered some outrage on Twitter (burst #6).

The events mentioned above are just a sample of all events explored and commented on by our participants. After the study, most of the participants mentioned that they now had a much better understanding of the OWS movement, including when the movement started, who was involved, and some major events within the large-scale movement.

8.1.3 Evaluating Detected Events

In addition to having participants identify and explore major events of the OWS movement using LeadLine, we compared our event results against timelines from Wikipedia [2] and the L.A. Times [1] regarding the same movement. Between Aug 19 and Nov 1, the Wikipedia timeline provided 24 events built by online communities, while The L.A. Times provided around 20 new articles describing the OWS movement. Depending on the level of details a user may want, LeadLine could provide events indictors ranging from 20 to more than 50.

We also examined the event indicators to match them with events from Wikipedia and The L.A. Times. Figure 7 shows our matching results. The results indicate that most of the automatically detected events by LeadLine match the timelines from Wikipedia and The L.A. Times, as annotated in Figure 7 with both logos after the majority detected bursts. Wikipedia also published events before September 17th that may have led to the OWS movement, which our automated detection did not identify due to insufficient tweets. However, LeadLine is able to show pre-September 17 activities that are worthy of further investigation. One interesting finding from our comparison is that The L.A. Times as a media organization reported more "photo-op worthy" events related to celebrities and politicians, while Wikipedia covered events centered around the OWS movement. As expected, LeadLine identified events that triggered more commentary on Twitter.

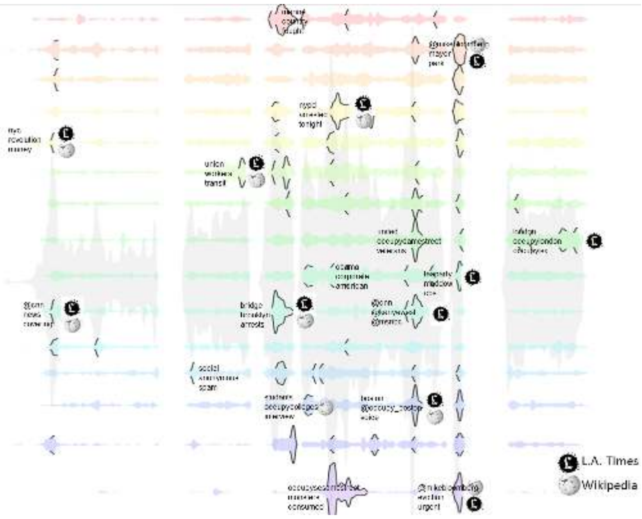


Figure 7: Comparing events identified by LeadLine against timelines from Wikipedia and L.A. Times.

In summary, through interactively exploring the topical bursts LeadLine, the participants were able to identify interesting events and gain insights regarding the OWS movement. Through further comparing automatically detected events in LeadLine against timelines provided by Wikipedia and L.A. Times, we validated that the topical bursts accurately reflect true events.

8.2 Case Study 2: Constructing Narratives based on Events from News Stories

In the second study, we asked participants to first explore topical bursts and named entities related to events with LeadLine, and then focus on creating narratives based on their exploration.

8.2.1 Data preparation

In this study, the dataset was CNN news stories from Aug 15, 2011 to Nov 5, 2011. The news articles were harvested and organized using methods introduced in Section 4. 30 topics were modeled from a total number of 3,130 news articles. Named entities including people and location were extracted from the news corpus.

8.2.2 Uncovering Narratives based on Events

The exploration process of news events was similar to the first case study in that participants started by browsing all news topics to gain a quick overview of the corpus. Some participants then identified a topic of interest and examined the bursts indicating events within that topic. Others turned to the entity graph and the map to see the locations and people mentioned in the news corpus.

Here we report findings identified by our participants. Starting by exploring the entity graph, one participant was interested in gathering events related to former Apple CEO Steve Jobs. By clicking on the node of Steve Jobs in the entity graph, the participant filtered all events that mentioned Jobs' name in the Topic Stream View (Figure 8). Jobs was consistently mentioned in the events in the mobile-technology related topic (in blue), which made sense to the participant. He further browsed some of the events, such as Jobs' resignation from Apple in late August (blue burst in the middle), and the unveiling of iPhone 4S on the date of Oct 4th, followed by

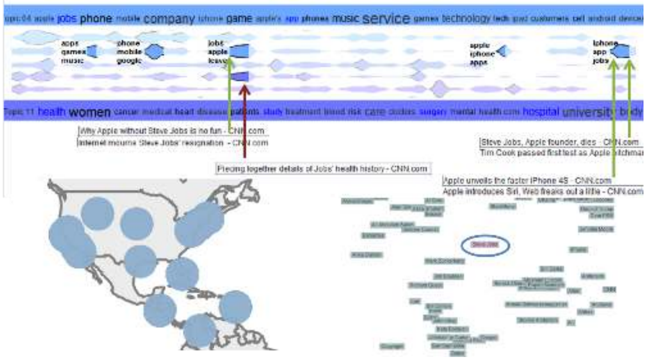

opic 04 apple jobs phone mobile company iphone game apple's app phones music service games technology tech ipad customers cell android device
apps games music
phone mobile google
jobs apple leave
apple iphone apps
iphone app jobs
Topic 11 health women cancer medical heart disease patients study treatment blood risk care doctors surgery mental health.com hospital university body
Why Apple without Steve Jobs is no fun - CNN.com
Internet mourns Steve Jobs' resignation - CNN.com
Piecing together details of Jobs' health history - CNN.com
Steve Jobs, Apple founder, dies - CNN.com
Tim Cook passed first test as Apple pitchman
Apple unveils the faster iPhone 4S - CNN.com
Apple introduces Siri, Web freaks out a little - CNN.com

Figure 8: Events related to former Apple CEO Steve Jobs. The green arrows link news articles with the events related to the blue topic, while the red arrow connects to the purple topic.

the news of Steve Jobs' death (last burst in blue). The participant investigated further when he saw a topic burst in the health related topic in purple also highlighted by LeadLine to indicate relevance to Jobs. Upon inspection of the news articles related to the event burst (shown as the call-outs in Figure 8), the participant discovered that one CNN program was trying to piece together a detailed history Job's of health in light of his resignation from Apple. "I didn't know CNN reported on the health history of Steve Jobs right after he stepped down, but I guess this kind of program might raise awareness for pancreatic cancer", the participant commented.

One of the most examined entities was President Obama. Participants discovered that he was a busy man during the three month period. Figure 1 shows topical bursts indicating events related to the president. He appeared in events such as immigration issues, political debates, natural disasters and international relations. After exploring the events in LeadLine and gaining insights about the news stories, multiple participants commented that they can now tell a story about the entities they focused on. They also expressed their desire to have a tool to gather and report their findings.

Inspired by the comments from the participants, we further developed a structure (described in Section 7) for building a narrative based on selected events. The narrative presents all events-related information regarding one person or location in a concise manner. The interface is web-based, so it is easily accessible from anywhere for reporting and revisiting.

9 Discussion

In this section, we discuss the limitations of LeadLine, and outline future directions to extend the current research.

9.1 Limitations arise from the Use of Topic Modeling, Early Event Detection, and Named Entity Recognition

Our approach relies on automated algorithms to discover information regarding who, what, when, and where in order to characterize an event. Inevitably, the final results are affected by the performance of each algorithm. For instance, the interpretability of each "burst" in the topical streams depends on the topic modeling results. The accuracy of detected entities relies on the performance of the named entity recognition (NER) algorithm. In addition, the same NER algorithm performs differently on different datasets. Solving the issues in the short term is challenging, but we think it is useful to make users aware of these issues. In order to do so, we plan to borrow methods from uncertainty visualization to annotate different layers of uncertainty so that users can make more informed decisions during investigation and analysis.

9.2 Identifying Inter-topic Events

In the scope of this paper, we assume that each event is associated with one major topic. However, certain events may have an impact on multiple topics. In order to identify inter-topic events, entities and timespan can contribute to grouping bursts from different topics into one triggering event. In other words, if bursts in several topics occur at the same time and share the same set of entities including people and location, we can assume that they are triggered by the same event. If one event encompasses multiple topics, the number of topics may further be used to evaluate the impact of the events.

9.3 Future Improvements

There are several improvements we would like to pursue in the future. First, when visualizing events identified by our event detection algorithm, it is clear that the algorithm favors upward trends in the topic stream. Although the current results include the starting time for each event, we want to improve the event detection algorithm to include downward trends in the topic stream so that the event cycle is complete.

Second, an interesting challenge we face when cleaning the text data is the duplication issue in both the news and microblog data. For news corpus, the issue arises because news websites keeps the article content up-to-date by reusing the same URL with a different timestamp. For Twitter, the issue lies in spam tweets and advertisements from different agencies. Our current implementation of addressing this duplication challenge is two-fold: first, we use the unique parameters for news and Twitter contents (e.g. URL, posted timestamp, author, etc) to construct a unique identification (UUID). This unique identification is used as checksum in the first stage to filter duplicates with exact match. The second stage, our implementation utilizes the standard Longest Common Substring (LCS) algorithm to compare the content length between two articles that share similar contents. We believe this is of great use in reducing the skewing effects that are introduced by the duplicated in the text streams.

While this performs comparatively well on the news data, such two stage process suffers from a performance penalty due to the largely fragmented nature of tweets. This issue needs to be addressed since in the event characterization process, both the named entity extraction algorithm as well as the topic modeling relies on word frequency calculations. Such duplication will undoubtedly skew the analysis results, producing inaccurate if not false event information.

10 Conclusion

In this paper, we present an interactive visual analytics system, LeadLine, that identifies meaningful events and allows users to examine the events that trigger changes in topical themes in news and social media. To discover events, LeadLine extracts information regarding who, what, when and where by integrating topic modeling, event detection, and named entity recognition methods. LeadLine also supports interactive exploration of events based on the 4Ws. Two case studies were conducted based on both news and social media data. The results indicate that LeadLine can not only accurately identify meaningful events given a text collection, but can also contribute to users' understanding of the events through interactive exploration.

Acknowledgements

This work was supported in part by a grant from the National Science Foundation under award number SBE-0915528. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] L.a. times' report on ows. http://timelines.latimes.com/occupy-wall-street-movement/, 2011.

[2] Wiki occupywallstreet. http://en.wikipedia.org/wiki/Timeline-of-Occupy-Wall-Street, 2011.

[3] Alias-i. lingpipe 4.1.0. http://alias-i.com/lingpipe, 2012.

[4] Apache hadoop. http://hadoop.apache.org, 2012.

[5] Apache nutch. http://nutch.apache.org, 2012.

[6] Twitter, inc. http://www.Twitter.com, 2012.

[7] Occupy wallstreet report [online], http://www.cnn.com/2011/10/05/opinion/rushkoff- occupy- wall- street/index. html.

[8] J. Allan, editor. *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[9] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 303–312, New York, NY, USA, 2010. ACM.

[10] D. Blei and J. Lafferty. *Text Mining: Theory and Applications*, chapter Topic Models. Taylor and Francis, 2009.

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

[12] M. Bostock, V. Ogievetsky, and J. Heer. D¡sup¿3¡/sup¿ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, Dec. 2011.

[13] S. A. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Crawling facebook for social network analysis purposes. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 52:1–52:8, New York, NY, USA, 2011. ACM.

[14] CNN. Cnn online. http://www.cnn.com/.

[15] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2412–2421, Dec. 2011.

[16] W. Dou, X. Wang, R. Chang, and W. Ribarsky. Paralleltopics: A probabilistic approach to exploring document collections. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 231 –240, oct. 2011.

[17] Event. Merriam Webster, http://en.wikipedia.org/wiki/Event.

[18] A. Feng and J. Allan. Finding and linking incidents in news. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 821–830, New York, NY, USA, 2007. ACM.

[19] J. Foster, Ö. Çetinoglu, J. Wagner, J. L. Roux, S. Hogan, J. Nivre, D. Hogan, and J. van Genabith. hardtoparse: Pos tagging and parsing the twitterverse. In *Analyzing Microtext*, volume WS-11-05 of *AAAI Workshops*. AAAI, 2011.

[20] A. Goldenberg, G. Shmueli, R. A. Caruana, and S. E. Fienberg. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8):pp. 5237–5240, 2002.

[21] V. Guralnik and J. Srivastava. Event detection from time series data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 33–42, New York, NY, USA, 1999. ACM.

[22] S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*, INFOVIS '00, pages 115–, Washington, DC, USA, 2000. IEEE Computer Society.

[23] S. A. Khan. Handbook of biosurveillance, m.m. wagner, a.w. moore, r.m. aryel (eds.). elsevier inc. isbn-13: 978-0-12-369378-5. *Journal of Biomedical Informatics*, 40(4):380–381, 2007.

[24] M. Krstajic, E. Bertini, and D. Keim. Cloudlines: Compact display of event episodes in multiple time-series. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2432 –2439, dec. 2011.

[25] C. A. Kurby and J. M. Zacks. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79, Feb. 2008.

[26] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim. Eventriver: Visually exploring text collections with temporal references. *Visualization and Computer Graphics, IEEE Transactions on*, 18(1):93 –105, jan. 2012.

[27] A. M. MacEachren, A. R. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *IEEE VAST*, pages 181–190, 2011.

[28] H. Mannila, H. Toivonen, and A. Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data Min. Knowl. Discov.*, 1(3):259–289, Jan. 1997.

[29] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 227–236, New York, NY, USA, 2011. ACM.

[30] R. Mckee. *Story - Substance, Structure, Style, and the Principles of Screenwriting*. Methuen, 1999.

[31] D. C. Montgomery. *Statistical quality control*. Wiley Hoboken, N.J., 2009.

[32] D. Neill and G. Cooper. A multivariate bayesian scan statistic for early event detection and characterization. *Machine Learning*.

[33] D. B. Neill and W.-K. Wong. Tutorial on event detection tutorial. http://www.cs.cmu.edu/ neill/papers/eventdetection.pdf, 2009.

[34] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[35] R. Sabhnani, D. Neill, and A. Moore. Detecting anomalous patterns in pharmacy retail data. *Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection*, Aug. 2005.

[36] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[37] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, Nov. 2010.

[38] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. Zhou. Understanding text corpora with multiple facets. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 99 –106, 2010.

[39] J. Stasko, C. Gorg, Z. Liu, and K. Singhal. Jigsaw: Supporting investigative analysis through interactive visualization. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 131 –138, 30 2007-nov. 1 2007.

[40] X. Wang, W. Dou, Z. Ma, J. Villalobos, Y. Chen, T. Kraft, and W. Ribarsky. I-SI: Scalable Architecture of Analyzing Latent Topical-Level Information From Social Media Data. *Computer Graphics Forum*, 31(3):1275–1284, 2012.

[41] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 153–162, New York, NY, USA, 2010. ACM.

[42] J. Weng, Y. Yao, E. Leonardi, and F. Lee. Event Detection in Twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[43] J. M. Zacks and B. Tversky. Event structure in perception and conception. *Psychological Bulletin*, 127:3, 2001.