International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 140 - 145, 2014, ISSN: 2319–8656

# An Interactive visual Textual Data Analysis by Event Detection and Extraction

Danu.R
Sri Manakula Vinayagar
Engineering College
Pudhucherry, India

Pradheep Narendran. P
Sri Manakula Vinayagar
Engineering College
Pudhucherry, India

Ranjith Kumar. C
Sri Manakula Vinayagar
Engineering College
Pudhucherry, India

Bharath. B
Sri Manakula Vinayagar
Engineering College
Pudhucherry, India

**Abstract:** Now a days, searching for the text data in a large ocean like location is quite challenging and more inaccurate task. Data that holds with the relation to its event can be evolved with certain changes with some intervals of time. Already existing techniques provides a trendy manner in order to extract a textual data with the visual analysis based on the event.But few data may have topic meaning that representing the kind of data to be extracted. In this paper, we propose a analytic system as an interactive manner called LeadLine, to recognize a data automatically by some semantic events in news blog as well as social media and deploys expansion or retrieval of the events. To organize such an events, LeadLine combines topic modeling, event detection, and named object or an entity recognition techniques to retrieve information automatically based on who, what, when, and where for each event. In order to make text data to be an effective one, LeadLine enables users to analyze interactively valid events by using 4 Ws to build an reviewing of mainly how, when and why. Bulky text data can be present normally as also the outdated one. These data can be concise with the help of LeadLine. LeadLine also provides the most simple process just by the exploration of events. To prove the effectiveness of LeadLine, These were implemented in the news blogs and social media data.

**Keywords**: Event detection, Topic modeling, LeadLine, Entity recognition.

## 1. INTRODUCTION

News blogs and online news like various text data present as a real-time dependent that is purely periodical based were located as worldwide. In the news blog, it has certain events that follows chain manner and in social media, the data can be simply like a user comments about something in the social aspect. Matching of certain patterns in terms of comprehensive can be either constant set of feed ot a changeable set of data. Some data in both the social media as well as news blog can be hidden in some case because of their privilege. So,a process to filter data that are in the form of text can be chosen based on their topics, and the relevant set of information can be triggered in order to get the assembling of complete appropriate information as a result. While examining the text data among the numerous amount of data, there will be more problem that can be faced from an event perspective .

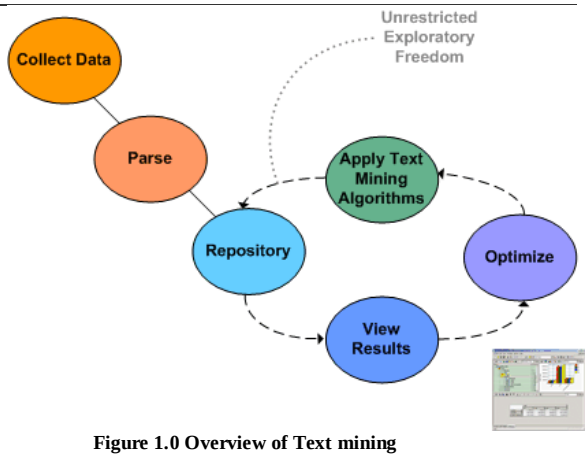


**Figure 1.0 Overview of Text mining**

There are many communities visualizing the working of the topic modeling with the time perspective. But here, we focusing on the topical trend based on the time, that doesn't meant for the complete event based technique but the major change in the temporal trends of the particular events.

### 1.1 LeadLine - An Introduction

A congested model that allows us to deploy computational methods to perform auto-extraction process for the events from text data. To explore such an events, we retrieve information based on what, who, where and when by simply integrating topic modeling, event detection, and entity recognition techniques. Initially, text data can be extracted in the social media and news blog sites on the conceptual themes using Latent Dirichlet Allocation (LDA) to provide topic in the formulation on events. To recognize the trending scale for each events, we have implemented an Early event detection algorithm to control the persistence of the events. This step of execution provides an attribute for representing the starting of any event that may also further expanded or depends upon other events as a ending event. To extract information about any people or location, related to the event, named entity recognition for the set of corpus of text and associate them with the events. With the above four processes are modeled in a system as an explicit one, our approach reinforces identification and extraction of events by topical, entity level and in trendy manner. To correlate and combine the events results as an effectively, we built a visual interface that suggests some related results for the event. Such an interface enables users to interactively traverse events and mainly to adjust or modify the event detection process based on the level of detailed set of data. Shaping the text data based on the event has additionally provides a base line for building such ideas as a creative. We have extended LeadLine that has a capacity to validate data, which allows its user to access and revisit the extracted findings easily. Especially, our approach provides three different benefits:

- Provides creative examining interface that makes users to get back their findings.
- A common process that integrates topic modeling, entity recognition, and already existing event detection mechanisms to identify semantic events from text data.
- An interactive visual system for analyzing user searchable textual data in the forms of 4W's set of questions.

### 1.2 Formulating Events

There are several questions that leads to critics to identify an event from the collection of text. How such meaningful events are carried out and extracted from the bulky collection of text ? Several properties that describes the characteristics of a specific event ? How to explore an event that in turn automatically discovers an appropriate event from the text corpora ? To reply these questions, we first make sure on what made up of an event:

Merrian - Webster defines a general definition that an event is a thing that happens or takes place, especially one of importance or any activity. In Topic Detection and Tracking (TDT) community and event detection [7,11], an event is defined based on its property as " a notation of something that represents the certain thing with corresponding time, topic and location on where it is associated ". Similarly the story telling concepts by McKee defines that an event refers to "creates semantic change in the temporal situation of a particular character" [13].

By integrating all these definitions about an event, is an " Occurrence reflecting any change in the larger amount of text data that utilizes the related topics at a specific time. This is defined in terms of topic and time, and related with the entities like an individual/ group of person and location ". We refers events with a four attributes like < Topic, Time, People, Location >. These refers to the 4W's questions : what, who, where and when.

## 2. RELATED WORKS

We mainly concentrate on the three areas such as named entity recognition, event detection, topic detection and analysis, and also text visualization techniques for a text with the background work of LeadLine.

### 2.1 Event Structure in Perception & Process

A different piece or the segment of time that denotes any person or location with the starting and ending stage is called as event. People can easily get them through the event just because of dividing and identify them with the different part of time continuously. People may use such an obeserved segments into an events or physical activities at mutiple set of timescales. Since, the same concept can be applied for even the abstract continuous streams, like topical streams, from the text corpora. Though, an event is treated the unit of making use ofactivities that serves more natural representation of any activities.

### 2.2 Event Detection

Over-the counter (OTC) medication sales, a type of as a source for detecting events indicating disease outbreaks describes a mutually growing system built for time detection of anthrax, a widespread occurrence of an infectious disease in a community at a particular time. This method comes into the category of common variation methods which concentrates on detecting events from time set of series [1]. As a more general approach, Guralnik et al. [2] presented steps to determine the change points in timely data dynamically without previous knowledge of the trending distributions. Other surveillance systems for a disease taken into consideration for both temporal and space related information. In addition, Neil at al. further developed a “multivariant Bayesian scanning statistic” (MBSS) [8] technique for fast and more relevant event detection. The already proposed event detection mechanisms are more efficient, but they lack the ability to handle text corpora,that may contain rich set of information that results with the symptoms and how they can be evlolved in the over time. In this paper, our approach allows to convert textual data into multiple semantic time information so that we can apply different ideas from the Biosurveillance community for early event detection on text data. The locality-sensitive hashing, enables first story detection on streaming data is chosen as a proposed system. However, the importance of the extracing events is not covered in the proposed technique.

### 2.3 Data Acquisition and Preparation

To explain the common techniques of our approach and their deployable domains, we have applied in two types of text data: CNN news and microblogs from Twitter. While both kind of origin that contain some collection of rich set of information resulting in a major real-world events, the main reason for choosing the two set of data sources is just because of its flexible editable module of style and the delay for responding to a particular event. In some specific, content from news media like CNN and others are customized by some journalists by specifying the topics with some set of background works. Not every but some posts in the social networks contains several information which are fixed range of commendatories [10]. These different type of text

information provide various levels of benchmarks that enables to validate the logical architecture.

## 3. SYSTEM ARCHITECTURE

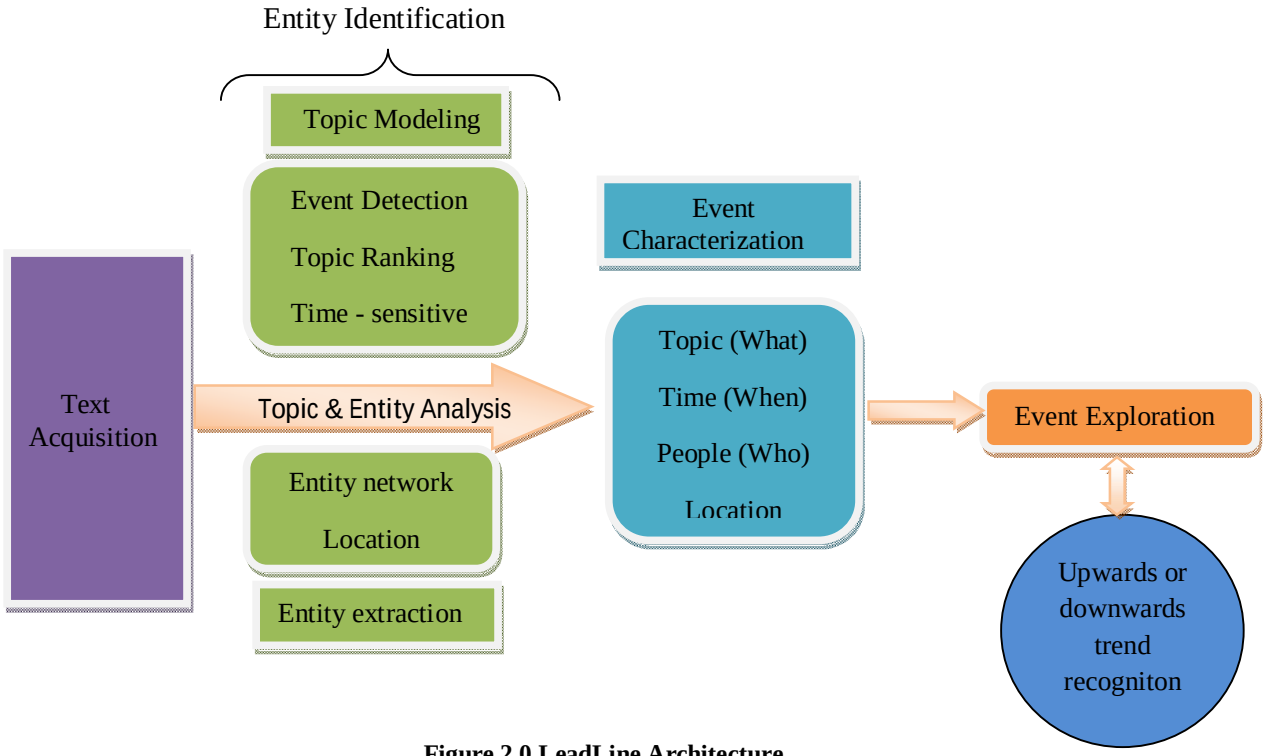


**Figure 2.0 LeadLine Architecture**

### 3.1 Data Acquisition and Preparation

To explain the common techniques of our approach and their deployable domains, we have applied in two types of text data: CNN news and microblogs from Twitter. While both kind of origin that contain some collection of rich set of information resulting in a major real-world events, the main reason for choosing the two set of data sources is just because of its flexible editable module of style and the delay for responding to a particular event. In some specific, content from news media like CNN and others are customized by some journalists by specifying the topics with some set of background works. Not every but some posts in the social networks contains several information which are fixed range of
commendatories [10]. These different type of text information provide various levels of benchmarks that enables to validate the logical architecture. These two sources belongs to public domain, there are no certain set of data that are not supposed available or visible as they are protected or under privacy enabled. Hence, we have extended our existing architecture in the news blogs and the social network with data crawling mechanism. The current approach extends the existing one is just by adding the news article crawling techniques. Both these news blog and the article data needs to be monitored and the extractor mechanism watches up to hour based set of tasks.

**News Blog Data Acquisition**: It needs to be customized with the page crawling and the RSS daemons, obviously. These methods generally implemented with universal techniques that tries to crawl the complete web domain information,copy all webpages, extract all the relevant textual articles, parse article time data. The data is stored into the HBase data structure that results in the faster access and MapReduce [9] based technique for the data cleaning and processing. Using these crawling techniques,data can be retrieved and filtered with the news articlesas the bottom up process.

**Twitter Data Crawling:** Some microblogs from Twitter, Facebook are also gathered in the form of dual crawling techniques. The primitive process that uses our MapReduce concurrently or a parallelized data crawler, which acts as between with the Internet through multiple

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

independent crawling techniques. Each crawlers may constantly gathers data from the social media by various public fields and moves it into HBase. As a result, we can able to collect over 5 billion posts or user tweets by providing a

reliable database from all languages over the course of 3 months for evaluation purposes. We implements a search technique called breadth-first search (BFS) using Nutch to obtain Twitter public user-graphs and capture them through their web portal for wider streams additionally.

**3.2 Analytics architecture for events detection and characterization**

To retrieve or extract an information from the text corpora, we can simply integrate the several kind of techniques to recongize <Time, People, Location, Topic>. To extract semantic topics with their timespan for any particular events, we holded topic models based on their themes and an Earlier Event Detection technique to identify a start and an end for each and every sort of event. To explore information about whom (individula person or a group) and where (associated location), we were performing named entity recognition and also analyzing relation between extracted data in the form of entities. We dividing the identification of topic themes and its span of time cycle as a topic-based way of analytics, in which we initially get through the topics from the input text corpora using Latent Dirichlet Allocation (LDA) as shown in the Fig 2.0. Then we applies, 1) topic - level event detection technique to automatically explore "events" as a triggers that are named by the timespan; 2) Time-tactful text or a phrase extraction that provides text information regarding an event with a set of brief keywords; 3) Topic ranking process to make easier of the discovery of event relation just by placing chunks of texts with similar topics nearby in a separate corpora; and finally Completing the topic-based analytics, our approach also focuses on named entity-based logic to identify people or/and a location relevant with each event. Especially, this process is interfaced as for extracting main/key entities from a textual data regarding whom and where. The visual interactive interface acts as a combining part of both logic processing to connect through the users and its complex analytic results. With this visual interactive interface, LeadLine mechanism supports interactive exploration of any events from various categories like whom, which, when, where as well as makes users to interact with the ongoing logical algorithms to partially makes adjusting the process of detecting and characterizing events from text corpora.

**3.3 Topic-Based Event analysis and visual perception**

Topic-based logic is a most crucial task in the event characterization in terms of exploring the topical theme and its time. Here, we just introduce an algorithm to extract topical and trendic information with based on an event, and some visual way of representations that can communicate with the topical as well as temporal ways.

**3.4 Extracting Topics from Text Data**

We begins by managing textual data streams depending on their topics. User simply gives a text as a word or a phrase, and there are different aspects to retrieve semantic topical themes. Among those aspects, Probabilistic topic models [12] are treated to be beneficious when comparing to traditional vector-based text analyzing techniques. In LeadLine, we first works with the most commonly used topic model called, LDA [14], to explore meaningful topics from text corpora.

**5.1.1 Visualization of Topic Streams**

To represent a data with the specialization of how it visually has to be presented, it merely concentrates on how the height of the topical themes that are changed in a searching domain. Each topic contains some relevant data information that can be carried out with the sort of holding some topical information about the searching data. Still, more effective algorithms are used, it wont results an exact crispy topical contents are retrieved in a system. In order to serve the complete context, a ThemeRiver representation is used in the backdrop of the visualization process. Thus redundant text patterns revealed by a text stream as a row (like weekly data pattern in the news stories) are still depicted.

**5.1.2 Topic Streams**

Time is more important attribute of an event than the topic. For making enabled of the clear process about the temporal change observing and exploring, we manage those topics along with the temporal central line. By considering each topic as a data information that exceeds over the time, the calculation of each topic information is done by processing a container based on the amount of text information related with the theme of the topic in each timescale. It is a unit that in which texts are integrated based on the temporal behavior. The time frame unit can be simply differs by collection of data and its tasks, that can be ranging also for minutes frequently in the social media data into days for newer stories.

**3.5 Topic Ranking**

During the exploration of any event, the results retrieved to be visually kept placed onto the similar set of visualization can be holded in a contiguous manner and also the events recently derived are topic-based ranked. LDA approach does not explicitly make the relationship between their corresponding topics, so that we need to rank the topics which is identical to be founded by Hellinger distance.

# 4. RESEARCH PROPOSAL

## STEP 1: Automatically Detecting Events in Topical Streams

A major task of this approach is to detect the temporal changes that are happening to the event. To detect such events, based on the topical theme, we need to consider it with the help of time series. Each and every time series is computed by relating or aggregating each topic with its assigned time scale. Most probably, we use the cumulative sum control chart (CUSUM) for the purpose of change detection [15]. It is effective for recognizing variations in the mean in a time series by maintaining a running sum of "revelation". We adopted CUSUM maily for detecting changes in topical data theme. For every topic theme, the program keeps itself a mutual integration maintains the topical theme and each stream has its own time span that are high when comparing to mean topic. To automatically retrieve data information in the topic streams, the mechanism called Earlier Event Detection (EED) can be used to identify bursts to a particular event. If the mutually integrated sum is more than a threshold, the event can be triggered out. 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. If the data can be expected for the future events are to be represented, then a file that contains relevant details between two dot operators are pulled out. Such a process of detecting timely events are more reliable task.

**4.1 Visualizing Detected Events**

To present any topical streams with the information as a visually interactive, we have an outline of its representation as well as highlighting the events of those particular topical stream which would have been chosen. The wider information of a time of the outlined data is choosen by the event detection resultant data. In addition to it, LeadLine mechanism supports starring of an events as a suggested or an interest via its user interaction process. To provide information in a crispy manner, LeadLine enables its users to access a documents like news feeds or microblog data can be defined as just by clicking on the event.

## ALGORITHM 1 : Cumulative Check sum

Algorithm : CUSUM
Input:Various topical time series $X$ collected $i = 1,...k$
Steps:
1. Calculate the mean $\mu$ and standard deviation $\sigma$ of the particular time series;
2. Calculate presently running sum $S$ from the starting time scale
S1 = max[0, $x_1$ - $\mu$]
Si = max[0, $S_{i-1}$ + $x_i$ - $\mu$].
3. When $S_k$ exceeds a value exists in $H$ (in units of $\sigma$), event triggers. The start and end of the event are determined by the closest positive Si to its triggering point.
4. If time is not mentioned, or any keywords like 'Upcoming', 'Future', 'List of any events' then
Date = Get (Today's date)
Explore all the topical data that consists of information within two dot operators that exceeds the Date.

**4.2 Detection of an Interactive Event**

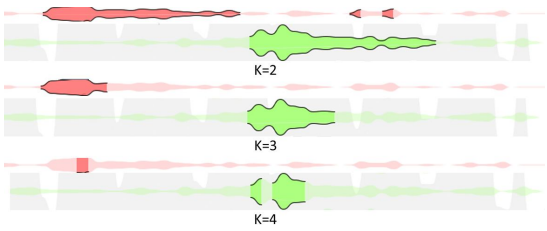


**Figure 3.0 Comparison of different variations of the events.**

One of the most striking advantage of this approach is just for providing automatically for an event detection that mainly triggers for the topical streams that are treated as a bursts which guarantees for the particular event. By simply clicking the button called as "Tune", the user can able to adjust the fine or coarse of the discovery . K refers to the standard deviation which are usually said to be a fixed mesasure of the threshold [16]. Users are allowed to adjust those K values . If the value of K is minimum, then there will be a situation of making sure that there are lesser number of mean of the variation on the particular event. If the value of K is greater as found, then there will a result of bigger range of shifts. If the user adjusts the tune level, then the LeadLine mechanism has to re-execute with the present values in the system.

## STEP 2: Time-sensitive Keyword extraction

To make an approach an efficient one, we need to refine the search and more recent information has to be provided to the user. In order to perform such an operation, we need to provide each event with its own time scale based retrieval process. The input for this algorithm is a text data that can be divided into sub-collections using their time scale and also in topics. Each sub - collection of data may have its own timespan and the topic recognized entity. The algorithm follows a TF-IDF (**Term Frequency**–Inverse Document Frequency) heuristic to choose time-sensitive terms: (a) if a term occurs many times in the sub-collection, it is marked; (b) if the term also occurs in many of other set of sub-collections, the importance is not marked.

## 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 Time Frequency
end for
3. Re-rank the Time Frequency scores with topic-term probabilities[17]
4. Within each topic and time frame, select the top N terms astime-sensitive terms.

# 5. CONCLUSION

In this paper, we were enhancing an visual interactive analytics system called as LeadLine, that identifies semantic events and enables users to validate the changes in the social media as well as news feeds topical streams from the triggering of events. To explore such an events, LeadLine mechanism uses who, what, when and where conditions to retrieve information based on the categories. It also provides a visually interactive process in a system. Finally, the results obtained by LeadLine doesn't only have semantic information, but also provides its user a complete information about his data.

# 6. REFERENCES

[1] 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.

[2] 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.

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

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

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

[6] 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.

[7] 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.

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

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

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

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

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

[13] LeadLine: Interactive Visual Analysis of Text Data through Event Identification and Exploration, IEEE Conference on Visual Analytics Science and Technology 2012.