

Hanbyul Yeon · Seokyeon Kim · Yun Jang

Predictive visual analytics of event evolution for user-created context

Received: 25 November 2015 / Revised: 1 February 2016 / Accepted: 29 February 2016 / Published online: 10 June 2016
© The Visualization Society of Japan 2016

Abstract Digital big data provide the vast potential of increasing effectiveness and efficiency for decision making. Since the volume of the data is enormous, the data analysis requires large amount of time and effort. It is more problematic when predictive analysis is necessary for futuristic decision making. For predictive analysis, there have been many studies to forecast future trends spatio-temporally. However, most of studies provide just future tendency per event using graphs or maps without contextual compositive analysis. In this paper, we present a predictive visual analytics system to provide predictive event patterns. We infer the future event evolution by combining contextually similar cases occurring in the past. We utilize social media data to detect interesting abnormal events and match the detected abnormal events within the past news media data to retrieve similar event patterns. Then, we extract future event patterns through compositing contextual relationship among topics included in the similar past patterns. To evaluate our VA system, we demonstrate three use cases in this paper and validate our system with possible predictive story lines. In addition, we present an informal user study and feedback to validate the effectiveness of our system and improve the system in the future.

Keywords Predictive analysis · Social media data analysis · Abnormal event · Future story lines

1 Introduction

Applications using digital media have been widely spread in our daily lives as hardware and software technologies advance. This leads to the era of digital big data since many applications produce lots of data every single second. Moreover, even mobile devices are used to overcome spatiotemporal limitations for generating and storing data. Recently, there has been lots of research on utilizing big data to analyze circumstances in our lives. The analysis is consequently connected to making decisions for proactive actions to mitigate possible risks, and to predict for future situations based on data. Although the potential for efficient decision making increases, it is not easy to analyze the data properly and effectively due to the

Categories and subject descriptors I.3.8COMPUTER GRAPHICSAplications.
General terms Visual Analytics System.

H. Yeon · S. Kim · Y. Jang (✉)
Sejong University, Seoul, South Korea
E-mail: jangy@sejong.edu

H. Yeon
E-mail: hbyeon109@sju.ac.kr

S. Kim
E-mail: ksy0586@sju.ac.kr

enormous volume of the data. It is even more problematic when predictive analysis is necessary to enable analysts to make decisions for future trends.

There have been also many studies on predictive analysis to forecast future trends spatiotemporally. However, most of studies provide future tendency per event or correlative future patterns on graphs and maps. These are based on the frequency of an event, such as numeric counts. If we analyze numerical data, these future tendency or patterns are what we want to predict but if we want to perform textual analysis of text data, then, it is not easy to perform predictive and contextual analysis. Since text data provide more semantic and contextual information compared to numeric data, many researchers focus on analysis of text data, such as web documents, blogs and social media data. These have been intensively investigated over past years to extract meaningful consequential information. Although it is relatively easy to obtain analysis results of text data, such as keywords and text patterns, predictive analysis of text data has been studied little due to the complexity of context within the data.

In this work, we present a predictive visual analytics system using topic composition for text data, especially social media data and news media data, to forecast how text data for certain event evolve over time in the future. We first extract abnormal topics from social media data, Tweets, to investigate interesting and unexpected events. Then we search for similar emergence patterns within the past news media data related to the extracted abnormal topics utilizing correlative analysis, assuming that certain events happen similar to the way of the past emergence. Once a user selects the most similar past pattern in the system, we provide relevant topics that appeared in the past news media data within the selected past pattern. The selective relevant topics by a user are combined to create new context for the contextual prediction. Our system extracts the new context within the past data and shows the evolution of the new context over time together with possible stories that might come out in the future. Since our system offers possible scenarios of an event evolution combining relevant topics, a user obtains an insight into future story trends of the event based on the historical data. This allows a user to make a decision for proactive preparation, prevention, or mitigation as early as possible. To evaluate our system, we demonstrate three use cases, *Germanwings crashed in alps on March 24, 2015*, *heavy snowstorm in east coast of US on January 26, 2015*, and *Paris attacks on November 12, 2015* and validate our approach with possible story lines. In addition, we present an informal user study and feedback to measure the effectiveness of our system and to improve our system in the future.

We first review previous work in Sect. 2 and present our system overview and its technical components in Sect. 3. We describe our predictive algorithm in Sect. 3 and visual interfaces in Sect. 4. In Sect. 5, we present case studies using our system. We show a user study and user feedback in Sect. 6 and the discussion about limitations of our system in Sect. 7. Finally, the conclusion and future directions are discussed in Sect. 8.

2 Related work

News, RSS feeds, and weblogs were the most popular data sources to detect and analyze events in the past and many research utilized temporal visualization focusing on automatic analysis with limited visual interaction to assist user to understand meaning of news and events. ThemeRiver (Havre et al. 2002) presents a temporal visualization of similar grouped topics from a large volume of news collection using the stack graph. BlogPulse (Glance et al. 2004) extracts key phrases in weblogs and visualizes their trends to investigate popular interests within the documents. OpinionSeer (Wu et al. 2010) visualizes sentiment trends in real time from RSS news feeds about the U.S. presidential election in 2008. Sayyadi et al. (2009) propose a event detection algorithm for social network data to analyze events by a network of keywords based on their co-occurrence. Recently many research for event detection use social media data from social network services, such as Twitter, Flickr, Facebook, since many users generate situational reports where they create and share contents, and give comments on contents from other people (Meyer et al. 2011; Best et al. 2012; Lee and Sumiya 2010; Chae et al. 2012). The social media data tend to consist of GPS information, time stamps, texts, images, and videos that can be analyzed to understand what is happening. The social media data report events faster than traditional news media due to mobile capabilities. Therefore, many researchers have focused on real-time monitoring and event detection of these social media streams (Wanner et al. 2014).

Although an event is defined variously in many domain areas, in this work, we use the definition by Zhao and Mitra (2007). They define an event as a set of relation between social actors on a specific topic over a

certain time period. Also Becker (2011) derives event properties, such as planned vs. unplanned, trending vs. non-trending, exogenous vs. endogenous. For event analysis, CloudLines Krstajic et al. (2011) and EventRiver Luo et al. (2012) present temporal news density visualization. StoryTracker Krstajic et al. (2013) shows visualizations of temporal flows for major topic groups as parallel color bars ranked by importance utilizing clustered news stream generated by the Europe Media Monitor (EMM). An interactive monitoring system, ScatterBlogs2 Bosch et al. (2013), is used to analyze geo-located microblogging data and Thom et al. (2015) analyze how visual analytics of social media can be used for proactive action in the crisis management. Sakaki et al. (2010) propose a disaster alert system using twitter messages where they calculate an epicenter using time and GPS interval between tweet mentions. Thom et al. (2014) present two approaches to identify probable locations using large-scale aggregation from tweet datasets. Takahashi et al. (2014) propose an approach to detect the emergence of topics in a social network stream, which is to use a probabilistic model that can capture normal mentioning behaviors of a user. Zhao et al. (2014) present a FluxFlow to reveal and analyze anomalous information spreading processes in twitter based on one-class conditional random fields model.

Watanabe et al. (2011) present a local event detection in the real word using location information from microblog. Zhao et al. (2011) compare twitter messages with traditional news media, such as New York Times, using unsupervised topic modeling and Becker et al. (2010) present an identification of real-world events by learning multiple feature similarity metrics in social media data. Kraft et al. (2013) show a visual analytics system that automates event detection for the interactive investigative visualization to analyze Twitter data. Chae et al. (2012) use the seasonal trend decomposition to detect abnormal events in microblogging data. This VA system alerts abnormal topics real-time and analyze the cause of the events. Chae et al. (2014) present a system designed to assist analysts with an visual spatiotemporal analysis and decision support in evacuation planning and disaster management. They analyze public behavior responses during natural disasters. Terpstra et al. (2012) investigate nearly 97,000 tweets to enable real-time and automated analysis during natural disasters. They argue that twitter can contain high-quality information for the decision making in disaster response. Vieweg et al. (2010) study the difference in reaction to different two crisis events about flooding and grass fire. Additionally, a technique is proposed to extract useful, relevant information from twitter messages during emergencies. SensePlace2 provides situational awareness for crisis management in the form of search and monitoring using visual components. This system extracts crisis-relevant information from geolocated Twitter data. SensePlace2 (MacEachren et al. 2011) provides situational awareness for crisis management in the form of search and monitoring using visual components. This system extracts crisis-relevant information from geolocated Twitter data. Thom et al. (2012) present a visual analytics system that extracts anomalies from location-based microblog messages in real-time and visualizes them on interactive map.

Some researchers apply cross-media analysis that provides occurrence pattern, correlation, or influence among multiple resources. Itoh et al. (2014) present inter-media comparison framework through images extracted from different types of media to understand societal behaviors. This framework visualizes image flows that provide us to comparing occurrence of topical images and tracking the origin of these images in multiple media resources, such as blogs and TV news. Yang and Leskovec (2011) analyze temporal variation in the popularity of online contents. They develop K-Spectral Centroid clustering algorithm which finds cluster centroids using similarity to find the common temporal patterns. They detect temporal patterns of the hashtags on twitter, and analyze propagation the news on twitter. Adar et al. (2007) analyze behavioral aspects using multiple resources including query logs, blogs, news, and identify a number of behavioral patterns based on correlative analysis.

Previous studies mentioned above are focused mostly on monitoring and analyzing data stream. There have been also predictive studies that foretell future events. Asur and Huberman (2010) demonstrate how social media content can be used to predict real-word outcomes, such as box office revenues. Bollen et al. (2011) study public mood from Tweets using correlative and predictive analysis of Dow Jones Industrial Average values. A similar research is proposed to predict political election using sentiment analysis (Tumasjan et al. 2010; Sang and Bos 2012). These systems provide basic visualizations for the prediction results including linechart, scatterplots without user interactions. Assady et al. (2014) show two approaches for predictive analysis to predict movie rating and box office gross of upcoming movies. These approaches are based on machine learning and visual interaction. The machine learning approach uses algorithms such as neural network, regression, classification. On the other hand, visual interactive approach mainly relies on the background knowledge. Hao et al. (2011) propose a visual analytics system for the long-term prediction for seasonal time series data using peak-preserving smoothing. Lu et al. (2014) propose a framework to

predict the opening weekend box-office gross of upcoming movies. This system combines features and machine learning algorithm, such as SVM and neural network. Malik et al. (2014) present a predictive visual analytics system that supports proactive decision making environments to help efficiency in resource allocation and deployment. Maciejewski et al. (2011) propose a predictive analytics approach for temporal prediction by geographically visualizing and aggregated distribution to prevention perceived threats utilizing the seasonal trend decomposition by loess smoothing for temporal prediction and kernel density estimation for event distribution. Bryan et al. (2014) introduce epidemic disease simulation system to predicting spatial spread of the epidemic using agent-based models. TiMoVA Bögl et al. (2014) is a predictive visual analytics system to guide whole process for the time series analysis. This system provides prediction of future values with respect to the selected time series model and verification of the selected model based on the prediction results.

3 Predictive visual analytics

We introduce our predictive visual analytics system to provide future trends of an event. We monitor input twitter messages and extract abnormal topics for unexpected events. Then, our system allows user to explore similar past events and to composite-related topics for predictive analysis. As we mention earlier, most previous studies provide a future trend per event based on the numeric frequency count; however, the numeric-based future trend does not tell what causes the trend evolution. In contrast, our system allows us to predict future trends utilizing various contextual information besides the frequency count. The main contribution of our work is the predictive analysis of events based on user-created context that implies detail of the event evolution. We describe our system briefly in the following sections. Note that this paper is an extended work published in VINCI2015 (Yeon and Jang 2015). We have added more case studies in Sect. 5.3, user study in Sect. 5.3, and discussion in Sect. 7.

3.1 System overview

Our predictive analysis, framework is designed with three main parts including topic extraction, predictive analysis and visual analytics as shown in Fig. 1. Topic analysis module consists of topic modeling, abnormal event detection, temporal correlation, and related topic extraction. We employ Latent Dirichlet Allocation (Mallet toolkit) McCallum (2002) to extract topic groups from the tweets. To detect abnormal events, we use Seasonal Trend Decomposition based on Loess smoothing (STL) Cleveland et al. (1990) to calculate topics abnormality in tweets. A user is provided with similar past patterns using temporal correlation for the selected abnormal topic through past news media data and related topics from the past news media data are presented for the predictive analysis. Predictive analysis module includes topic composition, prediction, and verification. The related topics from the topic analysis module are ready to combine to create a new context. The new context is investigated again in the past news media data and we extract possible candidates for predictive story lines. The candidates are presented with actual past stories for the verification. Visual analytics module displays abnormal topics and their contextual information with various visual components and interactions. Our system is built using Java, R, and Javascript. Java machine crawls tweets and news documents, and R calculates abnormality and temporal correlations. Javascript is used to create visual

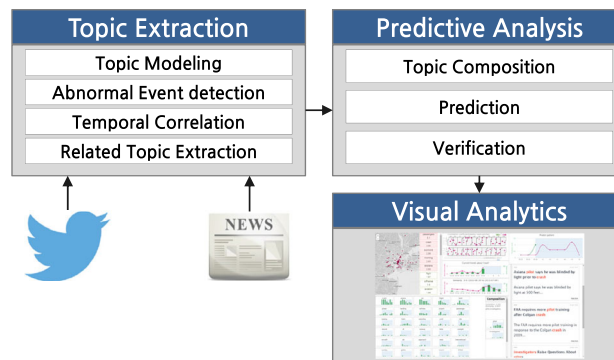


Fig. 1 System diagram for predictive visual analytics

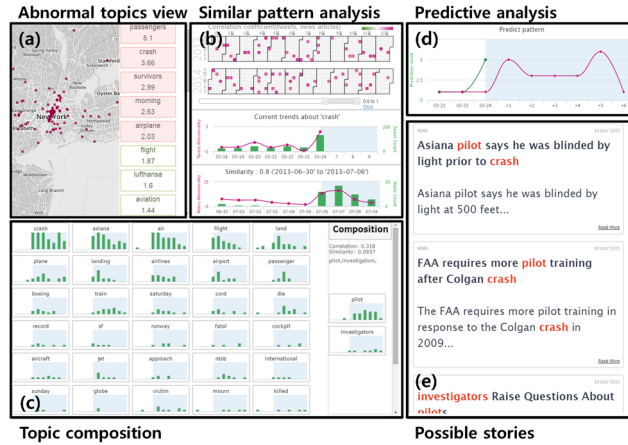


Fig. 2 Our predictive visual analytics system

analysis system with map, chart, and graphical components for the interactive visualization system as shown in Fig. 2. We utilize twitter data and news media data to validate our system including 52,945,433 twitter messages from January 1, 2014 to November 22, 2015, and 145,628 news documents from January 1, 2013 to November 22, 2015. Many people use twitter to publish and update their activities, experiences, thought, and feelings in everyday life. In addition, tweet messages generally contain texts, timestamps, and geo-locations. Therefore, tweet data can be a good source to analyze unexpected situations in real time. However, we obtain only abstract information from twitter since twitter only allows 140 characters; therefore, it is not easy to perform a deep analysis of the events only with twitter. In contrast, there is no limit in the news data unlike twitter. In addition, news data contain more reliable information because only facts are mentioned mostly. Due to these reasons, we apply cross-media analysis since it is not easy to trace events over time and to extract unbiased contextual patterns only within twitter data. Tweet data are used to detect current main issues including abnormal events, whereas news media data are used to extract similar precedents in the past to the abnormal events detected from the tweet data. We apply cross-media analysis since it is not easy to trace events over time and to extract unbiased contextual patterns only within twitter data. Our predictive visual analytics system as presented in the figure consists of abnormal topics view in (a), similar pattern analysis in (b), topic composition in (c), predictive analysis in (d), and possible story lines in (e).

3.2 Topic extraction

We collect twitter data and traditional news media data using their APIs. Tweet data include text messages with geo-location and temporal information. Tweet data are useful to observe circumstances of users and provides event information spatiotemporally. We extract topics for abnormal events from this user situational characteristics. Since the number of topics for abnormal events is much less than the number of topics for normal events in regular lives, such as *morning*, *coffee*, and *park*, it is not desirable to retrieve abnormal topics by ordering the frequencies. To extract potentially interesting topics from tweet messages, we employ Latent Dirichlet Allocation(LDA). LDA is a probabilistic topic model to extract interesting main subjects and topics analytically. Given a probabilistic distribution (θ) of main subjects within a document and a probabilistic distribution (z) of words within the main subjects, LDA selects main subjects in a document and words within the main subject repeatedly according to the probabilities. We aggregate tweet data by 10 min to increase the effectiveness of abnormal topic extraction. The tweets include ungrammatical but human-understandable expression. Therefore, we perform normalization utilizing porter stemming algorithm Porter (1997) before using LDA. LDA is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. Our system adopts MALLET toolkit McCallum (2002) for topic modeling. Table 1 shows an example of LDA topic extraction results. We extract topics from tweets posted at 07:10 am (EDT) on March 24, 2015 around New York. The most topics describe ordinary morning but the last topic group whose probability is 0.025 tells an unusual situation in the morning.

Table 1 Topic extraction examples

Probability	Topics
0.342	Morning coffee spring mood park cold
0.275	School start leave ready change manhattan
0.275	School morning tired train follow Tuesday
0.275	Morning cold change word street drive
0.025	Germanwings plane flight crash passengers Alps

3.3 Detection of abnormal topics using STL

Oxford dictionary defines *anomaly* as something that deviates from what is standard, normal, or expected. In this work, we define a *abnormal topic* as a topic that is far off from the average trends. Since our tweet data contain context and temporal information, we can investigate a topic trend over time. To stress abnormal topics, we utilize Seasonal-Trend Decomposition-based Loess smoothing (STL) Cleveland et al. (1990). STL is a filtering procedure for decomposing a time series data into three components including trend (T), seasonal(S), and remainder (R) as follows.

$$Y = S + T + R, \quad (1)$$

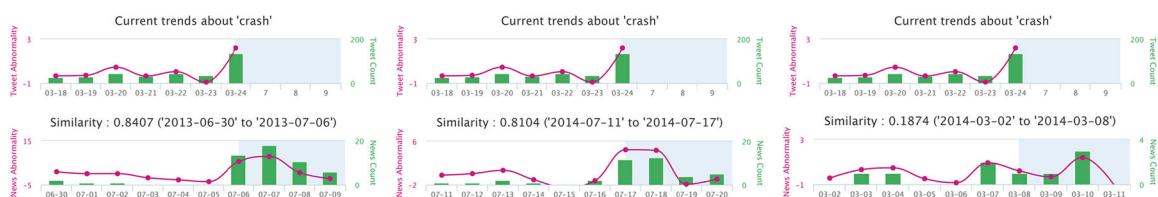
where Y is the original time series of data. We borrow an algorithm from Chae et al. (2012) to calculate topic abnormality. We calculate the remainder values and apply a 7-day moving average for the remainder values to compute z -score, $Z = (R(d) - \text{mean})/\text{std}$, where $R(d)$ is the remainder value of the day, mean is the last 7 days mean of the remainders, and std is the standard deviation of the remainders. The remainder component is the residuals from the seasonal plus trend fit. The Trend(T) component provides regular patterns in the data, whereas the Remainder(R) indicates abnormalities from the regular patterns. For example, a topic *morning* is always highly scored in the morning as T ; the topic *morning* would not be high in the evening except unusual circumstances; therefore, we use the Remainder(R) component to calculate topic abnormality. If z -score is greater than 2.0, our system alerts the topic as abnormal. Note that we set the abnormal z -score as 2.0 since the abnormality of z -score 2.0 is about 2.5 % out of all topics.

3.4 Temporal correlation between events

Most events have similar cases in the past assuming that certain events happen similar to the way of the past emergence. Since we focus on predicting future trends for an abnormal event, we search for similar emergence patterns within the past news media data related to the extracted abnormal topic utilizing correlative analysis. In our work, we use Pearson's correlation of the abnormal topic patterns between tweet and news data. Pearson's correlation is a measurement of correlations between two temporal datasets and its coefficient is defined as the following.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}, \quad (2)$$

where n is the length of the temporal data, which is 7 days toward the past from the current time in this work. x and y indicate abnormalities of a topic from tweet data and news media data at a certain time, respectively. \bar{x} and \bar{y} are the averages and S_x and S_y are the standard deviations. Note that we apply a 7-day moving window to correlate computation between two patterns. The reason why we use two different media data is that we want to extract objective meanings of the topics through comparing different datasets. Figure 3 shows three similar pattern candidates for the selected abnormal topic, *crash*, in the past news data. Note

**Fig. 3** Similarity pattern analysis for the abnormal topic, crash. Three similar past pattern candidates are shown

that current date is set to 2015-03-24 and three candidate dates are 2013-07-06, 2013-07-17, and 2014-03-08. Therefore, we can guess that the current trend of the selected topic might be one of the candidates.

3.5 Prediction using topic composition

Certain event has similar cases in the past; for example, abnormal event, *Germanwings crashed in alps on March 24, 2015* which has similar cases, such as *missing Malaysia airlines in Indian ocean on March 8, 2014*, *Missile hit Malaysia airlines plane in Ukraine on July 17, 2014*, and *Asiana airlines crashed in San Francisco airport on July 6, 2013*. These cases are different from each other; however, the original nature of these cases is similar, which denotes *airplane crash*. Although the evolutions of the events over time are not same perfectly, the common abstract evolution is similar, for example, event happening–response–investigation–prevention. The main idea of our work for the predictive analysis is to foretell possible future stories from the past similar cases. For the predictive analysis, we propose a topic composition to create a new interesting context. We first retrieve news documents related to the abnormal topic during days between -3 and $+6$ from the selected date by a user as presented in Sect. 3.4. In the section, a user can select one of three candidate days. We, then, extract topics from the retrieved news documents. Several extracted topics are provided and a user can select a few topics to create new interesting contexts. Since each topic contains temporal occurrence pattern, we also show the temporal topic evolutions over time. Besides, our system provides contextual information such as correlative analysis and mutual information among selected topics to indicate the relevance of the selected topics. Correlative analysis is performed with Pearson's correlation described in Sect. 3.4. Mutual information is computed with a probability ($p(x, y)$) that two different topics occur at the same time and individual probability ($p(x)$ and $p(y)$) that each topic occurs as defined in the following.

$$MI(x, y) = \frac{p(x, y)}{p(x)p(y)} \quad (3)$$

In this work, $p(x, y)$ is a probability that two selected topics appear together in news media data at the same time. Similarly, $p(x)$ and $p(y)$ are the probabilities that each topic appears individually in news media data. Then we provide possible story line candidates for the new context from the past news media data together with the future trend of the new context.

4 Visual interfaces

We have developed the real-time predictive system as shown in Fig. 2. This system consists of five visual interactive components including abnormal topics view (a), similar pattern analysis (b), topic composition analysis (c), prediction analysis (d), and news and tweet explorer (e).

4.1 Abnormal topics view

Abnormal topic view provides most interesting topics currently. As shown in Fig. 4, this abnormal topic view consists of topic list view on the right and map view on the left. Topic list view shows most abnormal topics extracted by the algorithm introduced in Sects. 3.2 and 3.3. Note that the topics are extracted only from tweet data within last 10 min that can represent current circumstances. The abnormal topics are listed in order by z -score abnormalities. The red-colored topics are categorized as abnormal, whereas the green-colored topics are as normal. We can recognize main events by combining the topics from the list. The map view shows the topics plotted on a map and provides the locations of the tweet messages. We can investigate where the abnormal topics are produced. We simply plot the message locations on a map; however, we plan to study the topic diffusion in the future.

Figure 4 shows an example of abnormal topic view on January 26, 2015, 9am (EST). There are a few general topics related to everyday life, but topics including *snowstorm*, *snow*, *snowday*, *winter*, and *snowfall* are detected as abnormal events.

4.2 Similar pattern analysis

The purpose of similar pattern analysis is to search for similar correlative patterns in the past to the current interesting topic. Figure 5a presents a calendar view of correlation coefficients calculated as described in

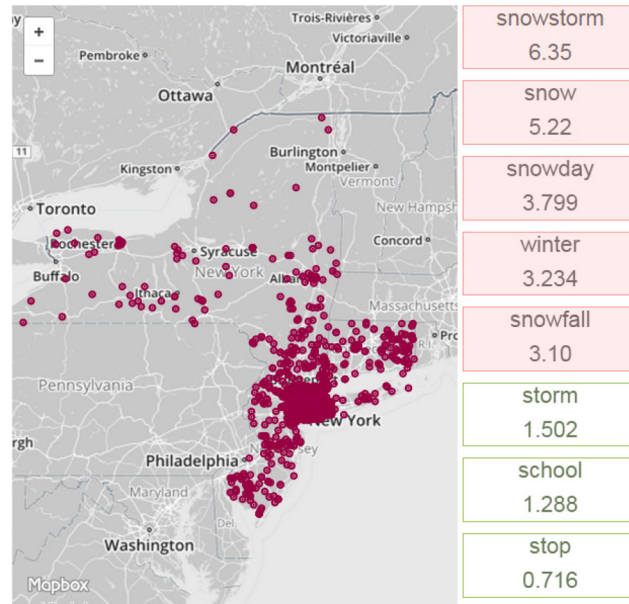


Fig. 4 Abnormal topic view. Tweet data are plotted on a map and abnormal topics are listed with their abnormalities



Fig. 5 Similar pattern view. Similar patterns to the current abnormal event trend are marked based on the similarities in the calendar view (a). **b** The current abnormal topic trend over time and **c** is the past similar trend selected from the calendar view (a)

Sect. 3.4. We can adjust the filtering values to color the cells in the view. (b) Shows the trend of the current selected topic from tweet data. The red trend line indicates the abnormalities, whereas the green bars represent the counts of tweet messages or news including the topic. (c) Presents the past similar pattern by selecting a day in the calendar view (a). As shown in the figure, the current interesting topic on January 26, 2015 in Fig. 5b has similar trend pattern to the past event around November 19, 2014 in Fig. 5c. We can expect that this topic trend might follow the past trend. Note that the blue area indicates the future time frame. Most related past events are found in winter as seen in (a), there are two cases in July and August, 2013. After investigating news articles, we find that a topic *snow* is found in the last name of a basketball player whose performance was great in July 2013. Another article including *snow* is related to news that a



Fig. 6 Our visual interfaces for topic composition in (a) and predictive story lines in (b) are presented. Note that *similarity* indicates the mutual information in Eq. 3

snowboarder died by a tragic accident in August 2013. Although correlation coefficients are marked as high in July and August, the trend patterns cannot be used in the similar pattern analysis.

4.3 Topic composition analysis

To investigate an event, a topic is not sufficient to represent the entire event since a topic does not contain any context information. Therefore, in this work, we provide a topic composition analysis to investigate an event by different contextual situations. For example, the combination of *snow* and *storm* leads to totally different meaning compared to the combination of *snow* and *basketball* as introduced in the previous section. Therefore, we offer a capability of combining different topics to create new contextual events. As seen in Fig. 6a, we extract related topics from news media data containing the abnormal topic chosen from tweet data in Fig. 4. Therefore, most topics shown in Fig. 6a are related to the event that we want to investigate. The topics are presented with frequency bar charts from -2 to +6 day about the selected date introduced in Fig. 5a. We can combine multiple topics by dragging interesting topics into the composition area on the right. Once topics are combined, we provide correlation coefficient and mutual information introduced in Eqs. 2 and 3 to indicate how the selected topics are related within the data. The goal of the topic composition analysis is to examine various possible future event trends by comparing the current abnormal event.

4.4 Predictive analysis

Predictive analysis provides possible story lines in the future for the abnormal event based on the topic composition result. Figure 6b shows the composite topic trend and possible stories. In the graph, the green line indicates the current news counts containing the abnormal topic, whereas, the red line represents the trend of the composite topic. Note that the *x*-axis indicates dates and the *y*-axis indicates the number of news. In the figure, we can find that the number of news counts is high at the reference date and slowly decreases, which implies that the composite topic event might fade away slowly as time goes. We can also pick any date within the graph and our system provides the possible news story lines in the below.

5 Case study

In this section, we evaluate our system using three abnormal event cases. One is the Germanwings crash case on March 24, 2015 and the other is the heavy snowstorm case on January 26, 2015. We present how our system is used to investigate the abnormal events and examine their predictive trends.

5.1 Germanwings crash

Our system alerts several abnormal topics at 7:10 am (EDT) on March 24, 2015. The abnormal topics include *crash*, *airplane*, *passengers*, and *morning* but the topic *morning* does not indicate much information since it is actually morning. However, we can conjecture that there is a plan crash from other abnormal topics, *crash*, *airplane* and *passenger*. Our system also provides more topics, *Lufthansa*, *Germanwings*, which infers that either Lufthansa or Germanwings airplane might have crashed. These abnormal topics are presented in Fig. 2a. We can further analyze the situation through our system and we select *crash* as an abnormal topic in this paper. Our system searches for similar past patterns in the news media data and presents the similarity scores in the calendar view in Fig. 2b. User can select any date to compare the topic evolution patterns. Note that the blue area in the graph indicates the future time frame. We find three similar cases as shown in Fig. 3, including *Asiana airlines crashed in San Francisco airport on July 6, 2013*, *Missile hit Malaysia airlines plane in Ukraine on July 17, 2014*, and *missing Malaysia airlines in Indian ocean on March 8, 2014*. Therefore, we can guess that the stories for the current abnormal topic in the future might be similar to one of these similar past patterns. Once the interesting past event is selected, our system shows most relevant topics from the news media data during 10 days (-6 and +4 days from the selected date) as presented in Fig. 2c.

Now we are ready to investigate the predictive analysis of the topics using the past news media data. We combine several topics to create new interesting context. Simply a user can drag any topic into the composition box in Fig. 2c. Then our system starts to search for news media data containing the new context. The evolution of the new context over time is displayed in Fig. 2d. Note that the evolution indicates the frequency of the new context in the news media data over time. User can select any date in (d) to see what might happen on the date, which is presented in Fig. 2e. We present three selected contexts with topic combinations in Fig. 7. First context in Fig. 7a is generated by combining two topics, *investigators* and *pilot*. The predictive trend pattern is presented in the upper graph and its corresponding stories are shown under the graph. The actual topic evolution as a validation is presented in the lower graph. Two graphs evolve similarly and we expect similar story lines for the current topic in the future. The second context in Fig. 7b is picked from Malaysia airlines case. Although the graphs are not perfectly same, the main story lines are similar as the volume of news media about *plane* and *victim* increases at the first stage, it suddenly decreases, and then slowly increases again. Since our system provides all story lines from the past news media data, a user is able to read through related stories in the past to obtain insight into possible future story lines.

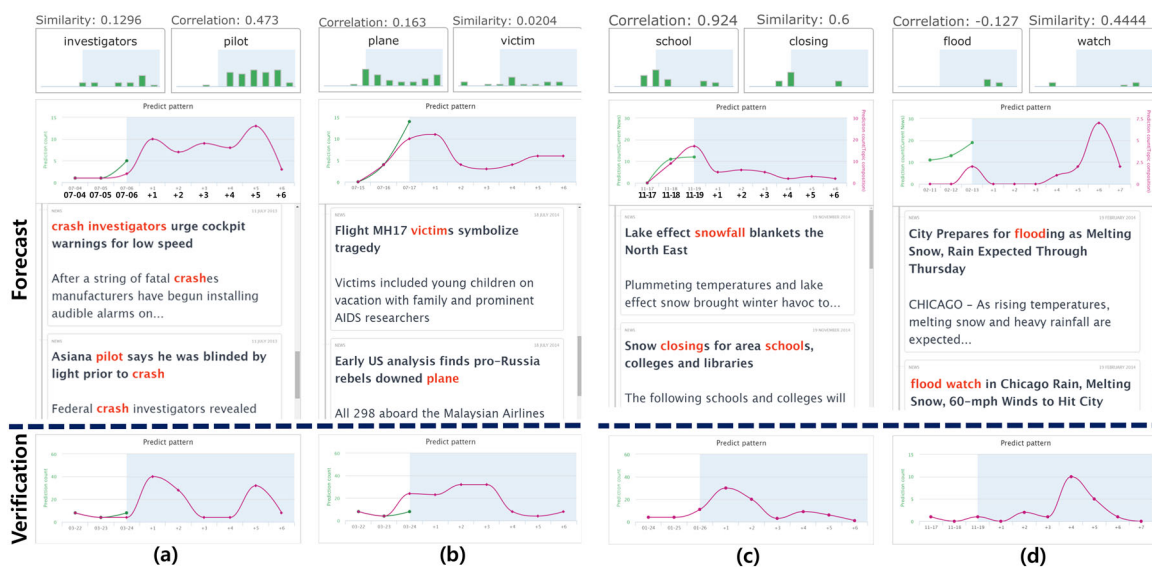


Fig. 7 Two new contexts created by combining topics for *Germanwings* case in (a) and (b). Another two contexts by topic composition for *snowstorm* case in (c) and (d). Note that *similarity* indicates the mutual information in Eq. 3

5.2 Heavy snowstorm in east coast of USA

According to US national weather service, blizzard is expected in east coast between January 26, 2015 and January 27, 2015. Our system detects abnormal topics from tweet data from 9am on January 26, 2015. As seen in Fig. 4, several abnormal topics are alerted, such as *snowstorm*, *snow*, *snowday*, *winter*, *snowfall*. Moreover, *school* and *stop* are also ranked in the high abnormality list. From this list, we can guess that there might be severe winter snow that affects school and something might stop during the winter storm. We start to search for related news and we find that there is severe blizzard warning in the east coast and all transportation except emergency vehicles may not be running during this blizzard. At this point, we are interested in what will be happening related to this blizzard event in the future. First we search for similar past patterns from news data by correlative analysis. We compute all correlation coefficients with 7-day window over last 2 years and visualize the coefficient values in Fig. 5a. Note that we adjust filtering values to emphasize highly correlative dates. The current tweet data pattern for this abnormal event in Fig. 5b is compared with the past similar pattern in Fig. 5c. Our system provides many similar patterns during winter and we select an event as a reference event on November 19, 2014 among those after investigating news data. Therefore, our predictive analysis is based on the event on November 19, 2014.

To predict the event trend related to *snow*, our system extracts frequently mentioned keywords from news data containing *snow* on the selected reference event date as shown in Fig. 6a. Although there are many keywords in the figure, one context with a few keywords might be different from another context with different keywords. In this work, we choose two different composite contexts with two sets of keywords. The first context is created with *school* and *closing* as shown in Fig. 7c. As seen in the figure, the forecast by our system is similar to the actual event trend in the bottom and we see that school closing is a big issue during snowstorm but eventually fades away out of the issue list. The second context is made with *flood* and *watch* as presented in Fig. 7d. The context is not concerned in the beginning of the event but eventually it becomes an important issue when huge amount of snow starts to melt after a few days. This is also verified by comparing our predictive trend with the actual news trend at the bottom.

5.3 Paris attacks at November 2015

In the evening around 9:20 pm (CET) on November 13, 2015, a series of coordinated terrorist attacks consisting of massive shootings, bombings, and hostage-taking occurred in Paris, the capital of France, and its northern suburb, Saint-Denis. The attackers killed 130 people, and 368 people were injured (Wikipedia 2015). Although some tweets were already captured at 9:30 am (CET) immediately, they were not sufficient enough to detect abnormal topics. Our system alerts a lot of abnormal topics at 9:37 am (CET) on November 13, 2015. The abnormal topics include *prayforparis*, *parisattacks*, *attack*, *terror*, *france*, and *peace*. Moreover, *dead* and *victims* are also ranked in the high abnormality list. We can guess that a horrible terror happened in Paris and a lot of people were dead. We, first, select *attack* as an abnormal topic to analyze the future event evolution by similar cases occurring in the past. We start to search for related news and we find that multiple similar cases from the daily correlative analysis, such as *Charlie Hebdo shooting in Paris* on January 7, 2015, *Copenhagen shootings* on February 24, 2015, and *Ankara bombings in Turkey* on October 10, 2015. We compare the current event (Fig. 8a) with the past similar patterns and we select *Charlie Hebdo shooting* case (Fig. 8b) that is the most similar to the current event pattern considering that *Charlie Hebdo shooting* occurred in Paris 10 months ago.

As seen in Fig. 9, our system provides the most relevant topics from news media data on the user-selected date in Fig. 8b. Each topic has its own trend over time and some of topics have similar trends. This indicates that it is possible to extract an event status from the topic list and corresponding trends. For example, topic *charlie*, *hebdo*, *attack*, *paris*, *pray*, *france*, *shooting*, *terror* have similar trends in the *Charlie Hebdo shooting* case as shown in Fig. 9a. The topic list provides the information that a gun-related terror occurred in Paris and it might be related to *charlie hebdo*. In addition, we can extract an information that some people were dead in this terror by looking at the topic *victim*, *pray*. Moreover, people had discussed a countermeasure as indicated in *law*, *vigils*, *demonstration*. In Fig. 9b, we can find similar information in the *Paris attacks*. However, *Paris attack* has more critical incidents than *Charlie Hebdo shooting*, such as *suicide*, *bombing and war*. In addition, the topic trends changed during the event as happening–response–investigation–prevention. Therefore, the topic list with the trends provides various information and allows us to create many new contexts using our topic composition.

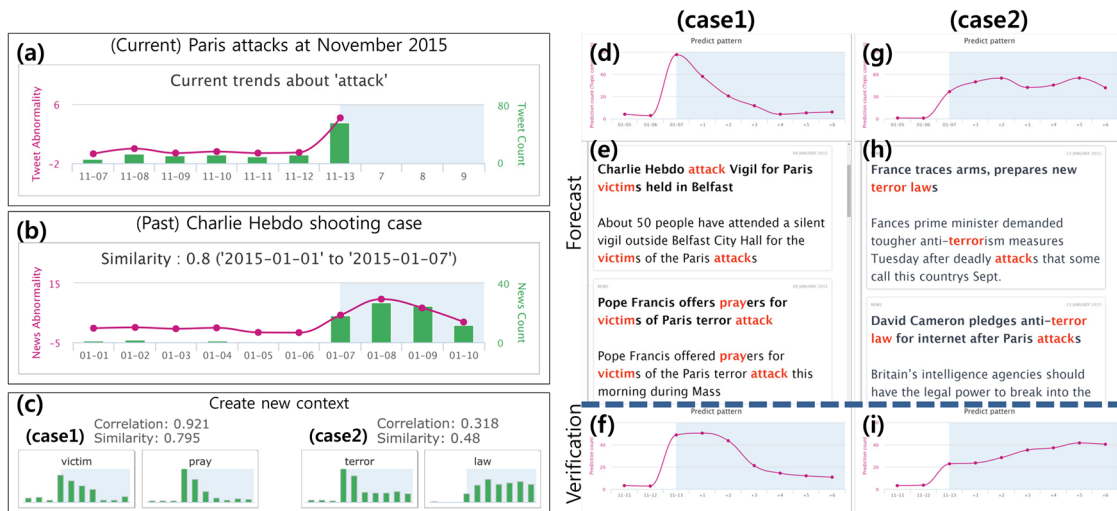


Fig. 8 Two new contexts created by combining topics for Paris attack case

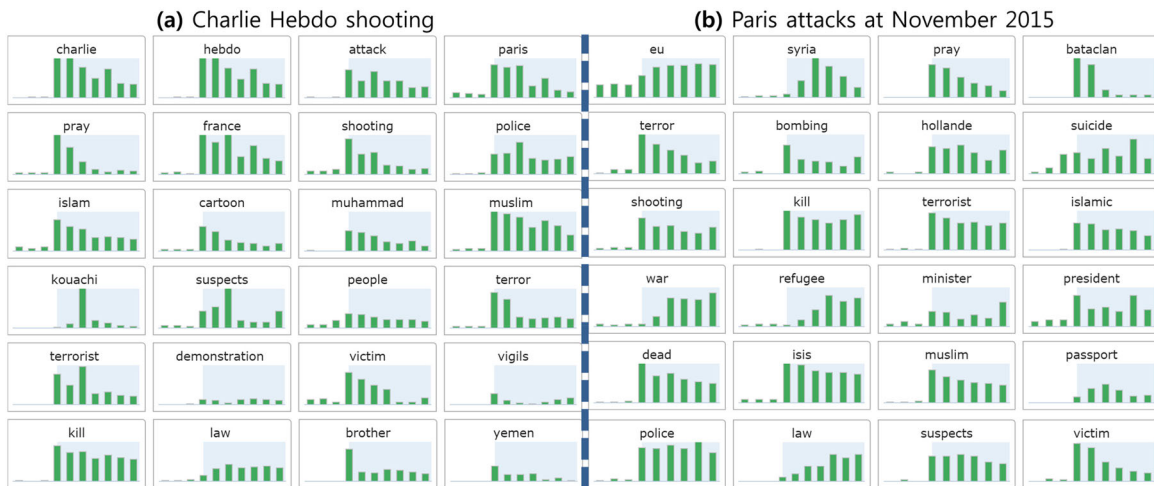


Fig. 9 Example of topic composition about Charlie Hebdo shooting in (a) and Paris attack in (b). Note that we discard highly ranked topics to explore various topics in (b), such as Paris, France, attack, response, theater, and explosion

We combine several topics to create new interesting context to predict the current event evolution over time. We, first, create a new context using *victim* and *pray* as shown in Fig. 8c. These topics are very relevant to *Charlie Hebdo shooting* case by looking at correlation and similarity scores. This context was a big issue during *Charlie Hebdo shooting* as shown in Fig. 8d. Many people prayed for the victims of the terrible incidents in Fig. 8e. However, it is seen that the attention by the public was changed elsewhere as time passed. We verify that this pattern is similar to the actual current event as presented in Fig. 8f.

We create another context with *terror* and *law* in Fig. 8c. These topics are not highly correlated compared to *victim* and *pray* but we are interested in how this new context would evolve in the near future. These two have the higher similarity score than the correlation score. Therefore, we guess that the topic *terror* contains a broad meaning for the event, whereas the topic *law* is seen as a part of the events. Figure 8g is obtained from the context evolution by combining two topics, *terror* and *law*. This context became more important as time passed as seen in Fig. 8g. It has been raised steadily to the legislation immediately after the attacks occurred through the press and the government had prepared new terror laws

after the event in *Charlie Hebdo shooting* case. In *Paris attacks in November 2015* case, the prime suspects arrived at France since EU refugee policy allowed them to stay in EU. Therefore, many countries in Europe have requested EU to change its decision on the refugees law. Moreover, some countries have already passed a bill that makes it more difficult for refugees from certain countries to enter their countries. We verify this by comparing our predictive trend in Fig. 8g with the actual news trend in Fig. 8i.

6 User study and feedback

We conducted an informal user study and obtained user feedback to evaluate the effectiveness of our predictive visual analytics system. We recruited eight graduate students who majored in computer engineering and were not coauthors of this paper. We divided them into two groups (G1 and G2) to evaluate the user interactions in our visual analytics system. We explained our system thoroughly to G1 before the user study. However, G2 utilized our system without any explanation. Each user study took about total 60 min including 10 min of system demonstration (only G1), 30 min of free exploration, and 20 min of discussion. G1 and G2 were interested in predicting future event evolution using past similar cases. Especially, they were impressed by creating meaningful contexts using topics extracted from past similar news documents. They commented that our predictive visual analytics system could be used to make decisions for proactive preparation, prevention or mitigation as early as possible when huge emergence events occur. In addition, most of interviewees agreed that the system would be very useful for natural disease situations.

- Abnormal topics view** Both groups (G1 and G2) were able to understand the Abnormal topics view at once. Two groups easily guessed an abnormal event by analyzing the topic abnormal scores from the list. However, both groups commented that it would be helpful to investigate event using word phrases instead of only individual topics since the word phrases could provide more accurate information about the events at the start of the event investigation.
- Similar pattern analysis** The goal of this analysis is to find similar past pattern candidates for the selected abnormal topic. They agreed that cross media analysis would be better than just one data source to trace events and to extract contextual pattern over time. However, most of users pointed out that it was difficult to find an interesting pattern using our calendar view. Especially, G2 spent long time on understanding this module. After the user study, we had an in-depth discussion about the problem. A color-coded cell in the calendar view indicates the correlation coefficient between the current tweet trend and the past news trends as shown in Fig. 5a. A user had to manually explore high correlative dates to search for similar cases; however, there were too many similar patterns in the calendar view. They noted that the visualization including text information would be clearer than that using only the color-coded cells.
- Topic composition analysis** Both groups commented that it was very simple and straightforward to create a meaningful context by dragging interesting topics into the composition area. In addition, the mutual information and correlation scores were very helpful to figure out how much the selected topics were related within the data. Many users commented that it was difficult to know where to start for the investigation since there were too many topics and combinations similarly to the calendar view. Theoretically, our system is able to create nCr contexts, where n is the number of topics and r is the number of topics in a context. Moreover, our system does not support semantic analysis in the topic composition. Therefore, these comments and limitations would lead the future work to improve the topic composition analysis in our system.
- Predictive analysis** All users agreed that the prediction pattern and possible story lines were very useful for predicting events flow. They especially liked the possible story lines. However, some in the group G2 spent long time in understanding the meaning of the prediction pattern in Fig. 2d. They commented that it would be helpful to understand possible stories if videos or images were provided together with the text results.

7 Discussion

Our web-based predictive visual analytics system consists of five visual interactive components. We have discussed how our system allows the user to search for the predictive information and insight into an abnormal event. There exist limitations in our system based on the feedback from users as presented in the following.

First, it is hard to explore similar patterns in the past using similar pattern view as shown in Fig. 5. We utilize a color-coded calendar visualization and the color indicates the pattern similarity to the current abnormal event trend as shown in Fig. 5a. A user can select high similarity patterns using the calendar view. However, it is not easy to find an interesting pattern among the similar patterns since there are too many patterns that a user have to compare. Second, our system provides possible story lines in the future based on the new contexts obtained as the topic composition results. However, the possible story lines are extracted by matching words without any semantic meaning of the words or context. For example, topics *law* and *terror* in Fig. 9a might indicate the prevention of the tragic accident but our system provides only story lines containing both topics without the prevention context. Therefore, there is a limitation to analyze the actual event evolution semantically.

Third, our predictive visual analytics system utilizes the LDA and STL algorithms to detect the current abnormal topics from the tweets. STL requires long computing time due to the large number of topics since all topics have to be calculated in STL to extract the abnormal topics among many topics. In our experiment, the computing time is 5–10 min in average for the abnormalities of all topics contained in about 750 tweets (about 11,000 topics) using Intel i5 3.50 GHz with 16 GB memory.

There are several directions for generalizing and extending our system. First, we will study automatic case classifications to enable users to search for similar patterns easily. In our system, we use Pearson's correlation of the topic patterns between tweet and news media data for 10 days. We will investigate best conditions for the correlation computation by changing parameters. Moreover, since Pearson's correlation does not reflect the meaning of the event, we will study other comparison methods, such as document clustering, to analyze the meaning of the event. Second, we are considering automatic topic composition candidates for better context investigation and semantic analysis among topics. Third, we will improve the event detection process for the real-time monitoring and analysis by adopting parallel computing and GPU computation. Last, we will study general evaluation for our prediction results.

8 Conclusion

In this paper, we have presented a predictive visual analytics system using topic composition for text data, especially social media data and news media data, to forecast how text data for certain event evolve over time in the future. We first detected an abnormal topic and correlated the temporal trend of the abnormal topic with ones of past similar topics to search for interesting time frame. New topics were provided for the time frame and combined by a user to create a new contextual event for the predictive analysis. A user was able to investigate the predictive results from the trend graph and possible story views. We demonstrated our system with three cases and proved that the new contextual topic combination provides similar results to the actual future evolutions. We conduct an informal user study to evaluate the effectiveness of our predictive visual analytics system. Most users agreed that our predictive visual analytics system could make a decision for proactive preparation, prevention or mitigation as early as possible when huge emergence event occurs.

As future work, we will apply machine learning approach for finding and classifying similar cases automatically and apply multiple cases to predict future stories. We will also apply semantic analysis to cluster meaningful information about relevant topics. We will study rule-based classification for more accurate predictive analysis. Moreover, we plan to extract contextually similar documents with whole texts instead of a few keywords. Finally, we plan to evaluate our predictive process using ground truth data from the past cases.

Acknowledgments This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the the Korea government(MIST) (R0190-15-2016), Development of Complex Fast Stream Big Data Processing based on In-memory Technology in Distributed Environment) and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A1011170). We would like to thank the reviewers for their valuable suggestions and comments, which helped to improve the presentation of this work.

References

- Adar E, Weld DS, Bershad BN, Gribble SS (2007) Why we search: visualizing and predicting user behavior. In: Proceedings of the 16th international conference on World Wide Web, ACM, pp 161–170
- Asur S, Huberman BA (2010) Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, vol 01, pp 492–499
- Becker H (2011) Identification and characterization of events in social media. PhD thesis, Columbia University
- Becker H, Naaman M, Gravano L (2010) Learning similarity metrics for event identification in social media. In: Proceedings of the 3rd ACM international conference on web search and data mining, pp 291–300
- Best D, Bruce J, Dowson S, Love O, McGrath L (2012) Web-based visual analytics for social media. In: Proceedings of the workshop on social media visualization (SocMedVis) at ICWSM
- Bögl M, Aigner W, Filzmoser P, Gschwandtner T, Lamarsch T, Miksch S, Rind A (2014) Visual analytics methods to guide diagnostics for time series model predictions. In: Proceedings of the IEEE VIS 2014 workshop visualization for predictive analytics
- Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *J Comput Sci* 2(1):1–8
- Bosch H, Thom D, Heimerl F, Puttmann E, Koch S, KrÄijger R, WÄurner M, Ertl T (2013) Scatterblogs2: real-time monitoring of microblog messages through user-guided filtering. *IEEE Trans Vis Comput Graph* 19(12):2022–2031
- Bryan C, Ma K-L, Mniszewski S (2014) Integrating predictive visualization with the epidemic disease simulation system (episims). In: Proceedings of the IEEE VIS 2014 workshop visualization for predictive analytics
- Chae J, Thom D, Bosch H, Jang Y, Maciejewski R, Ebert DS, Ertl T (2012) Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In: IEEE Conference on visual analytics science and technology, VAST, pp 143–152
- Chae J, Thom D, Jang Y, Kim SY, Ertl T, Ebert DS (2014) Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Comput Graph* 38:51–60
- Cleveland RB, Cleveland WS, McRae JE, Terpenning I (1990) STL: a seasonal-trend decomposition procedure based on loess (with discussion). *J Off Stat* 6:3–73
- El Assady M, Jentner W, Stein M, Fischer F, Schreck T, Keim DA (2014) Predictive visual analytics approaches for movie ratings and discussion of open research challenges. In: Proceedings of the IEEE VIS 2014 workshop visualization for predictive analytics
- Glance NS, Hurst M, Tomokiyo T (2004) Blogpulse: automated trend discovery for weblogs. In: WWW 2004 workshop on the weblogging ecosystem
- Hao MC, Janetzko H, Mittelstädt S, Hill W, Dayal U, Keim DA, Marwah M, Sharma RK (2011) A visual analytics approach for peak-preserving prediction of large seasonal time series. *Comput Graph Forum* 30(3):691–700
- Havre S, Whitney P, Nowell L (2002) Themeriver: visualizing thematic changes in large document collections. *IEEE Trans Vis Comput Graph* 8:9–20
- Itoh M, Toyoda M, Zhu CZ, Satoh S, Kitsuregawa M (2014) Image flows visualization for inter-media comparison. In: Pacific visualization symposium (PacificVis), 2014 IEEE, pp 129–136
- Kraft T, Wang DX, Delawder J, Dou W, Li Y, Ribarsky W (2013) Less after-the-fact: Investigative visual analysis of events from streaming twitter. In: IEEE symposium on large-scale data analysis and visualization, LDAV, pp 95–103
- Krstajic M, Bertini E, Keim DA (2011) Cloudlines: compact display of event episodes in multiple time-series. *IEEE Trans Vis Comput Graph* 17(12):2432–2439
- Krstajic M, Najm-Araghi M, Mansmann F, Keim DA (2013) Story tracker: incremental visual text analytics of news story development. *Inf Vis* 12(3–4):308–323
- Lee R, Sumiya K (2010) Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In: Proceedings of the 2Nd ACM SIGSPATIAL international workshop on location based social networks, LBSN '10, pp 1–10
- Lu Y, Kruger R, Thom D, Wang F, Koch S, Ertl T, Maciejewski R (2014) Integrating predictive analytics and social media. In: Proceedings of the visual analytics science and technology (VAST), 2014 IEEE conference on, pp 193–202
- Luo D, Yang J, Krstajic M, Ribarsky W, Keim DA (2012) EventRiver: visually exploring text collections with temporal references. *IEEE Trans Vis Comput Graph* 18(1):93–105
- MacEachren AM, Jaiswal AR, Robinson AC, Pezanowski S, Savelyev A, Mitra P, Zhang X, Blanford J (2011) Senseplace2: geotwitter analytics support for situational awareness. In: IEEE VAST, IEEE Computer Society, pp 181–190
- Maciejewski R, Hafen R, Rudolph S, Larew S, Mitchell M, Cleveland W, Ebert D (2011) Forecasting hotspots: a predictive analytics approach. *IEEE Trans Vis Comput Graph* 17(4):440–453
- Malik A, Maciejewski R, Towers S, McCullough S, Ebert D (2014) Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement. *IEEE Trans Vis Comput Graph* 20(12):1863–1872
- McCallum AK (2002) MALLETT: a machine learning for language toolkit. <http://mallet.cs.umass.edu>
- Meyer B, Bryan K, Santos Y, Kim B (2011) Breaking news detection and visualization through the geo-tagged twitter network. In: Proceedings of the ISCA 26th international conference on computers and their applications, pp 84–89
- Porter MF (1997) An algorithm for suffix stripping. In: Jones KS, Willett P (eds), Readings in information retrieval, Morgan Kaufmann Publishers Inc., pp 313–316
- Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World Wide Web, WWW '10, pp 851–860
- Sang ETK, Bos J (2012) Predicting the 2011 dutch senate election results with twitter. In: Proceedings of the workshop on semantic analysis in social media, pp 53–60
- Sayyadi H, Hurst M, Maykov A (2009) Event detection and tracking in social streams. In: Proceedings of the international conference on weblogs and social media
- Takahashi T, Tomioka R, Yamanishi K (2014) Discovering emerging topics in social streams via link-anomaly detection. *IEEE Trans Knowl Data Eng* 26(1):120–130

- Terpstra T, de Vries A, Paradies G, Stronkman R (2012) Towards a realtime twitter analysis during crises for operational crisis management. In: Proceeding of 9th international conference on information systems for crisis response and management (ISCRAM)
- Thom D, Bosch H, Koch S, Wörner M, Ertl T (2012) Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In: Pacific visualization symposium (PacificVis), 2012 IEEE, pp 41–48
- Thom D, Bosch H, Krüger R, Ertl T (2014) Using large scale aggregated knowledge for social media location discovery. In: System Sciences (HICSS), 2014 47th Hawaii international conference on, pp 1464–1473
- Thom D, Krüger R, Ertl T, Bechstedt U, Platz A, Zisgen J, Volland B (2015) Can twitter really save your life? A case study of visual social media analytics for situation awareness. In: Pacific visualization symposium (PacificVis), 2015 IEEE, pp 183–190
- Tumasjan A, Sprenger T, Sandner P, Welp I (2010) Predicting elections with twitter: what 140 characters reveal about political sentiment. In: Proceedings of the 4th international AAAI conference on weblogs and social media, pp 178–185
- Vieweg S, Hughes AL, Starbird K, Palen L (2010) Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '10, New York, NY, pp 1079–1088
- Wanner F, Stoffel A, Jäckle D, Kwon BC, Weiler A, Keim DA (2014) State-of-the-art report of visual analysis for event detection in text data streams. In: Borgo R, Maciejewski R, Viola I (eds) EuroVis - STARS. Swansea, London, pp 125–139
- Watanabe K, Ochi M, Okabe M, Onai R (2011) Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In: Proceedings of the 20th ACM international conference on information and knowledge management, pp 2541–2544
- Wikipedia. Wikipedia: November 2015 paris attacks. https://en.wikipedia.org/wiki/November_2015_Paris_attacks
- Wu Y, Wei F, Liu S, Au N, Cui W, Zhou H, Qu H (2010) Opinionseer: interactive visualization of hotel customer feedback. IEEE Trans Vis Comput Graph 16(6):1109–1118
- Yang J, Leskovec J (2011) Patterns of temporal variation in online media. In: Proceedings of the 4th ACM international conference on web search and data mining, pp 177–186
- Yeon H, Jang Y (2015) Predictive visual analytics using topic composition. In: Proceedings of the 8th international symposium on visual information communication and interaction, pp 1–8
- Zhao J, Cao N, Wen Z, Song Y, Lin Y-R, Collins C (2014) Fluxflow: visual analysis of anomalous information spreading on social media. Vis Comput Graph IEEE Trans 20(12):1773–1782
- Zhao Q, Mitra P (2007) Event detection and visualization for social text streams. In: International conference on weblogs and social media (ICWSM'2007), Boulder, Colorado
- Zhao WX, Jiang J, Weng J, He J, Lim E-P, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In Proceedings of the 33rd European conference on advances in information retrieval, pp 338–349