TopicOnTiles: Tile-Based Spatio-Temporal Event Analytics via Exclusive Topic Modeling on Social Media

Minsuk Choi
Korea University
Seoul, South Korea
mchoi@korea.ac.kr

Sungbok Shin
Korea University
Seoul, South Korea
sungbokshin@korea.ac.kr

Jinho Choi
Korea University
Seoul, South Korea
rigrome@korea.ac.kr

Scott Langevin
Uncharted Software Inc.
Toronto, ON, Canada
slangevin@uncharted.software

Christopher Bethune
Uncharted Software Inc.
Toronto, ON, Canada
cbethune@uncharted.software

Philippe Horne
Uncharted Software Inc.
Toronto, ON, Canada
phorne@uncharted.software

Nathan Kronenfeld
Uncharted Software Inc.
Toronto, ON, Canada
nkronenfeld@uncharted.software

Ramakrishnan Kannan
Oak Ridge National Laboratory
Oak Ridge, TN, USA
rkannan@gatech.edu

Barry Drake
Georgia Tech. Research Institute
Atlanta, GA, USA
barry.drake@gtri.gatech.edu

Haesun Park
Georgia Tech.
Atlanta, GA, USA
hpark@cc.gatech.edu

Jaegul Choo
Korea University
Seoul, South Korea
jchoo@korea.ac.kr

ABSTRACT

Detecting anomalous events of a particular area in a timely manner is an important task. Geo-tagged social media data are useful resource for this task, but the abundance of everyday language in them makes this task still challenging. To address such challenges, we present TopicOnTiles, a visual analytics system that can reveal the information relevant to anomalous events in a multi-level tile-based map interface by using social media data. To this end, we adopt and improve a recently proposed topic modeling method that can extract spatio-temporally exclusive topics corresponding to a particular region and a time point. Furthermore, we utilize a tile-based map interface to efficiently handle large-scale data in parallel. Our user interface effectively highlights anomalous tiles using our novel glyph visualization that encodes the degree of anomaly computed by our exclusive topic modeling processes. To show the effectiveness of our system, we present several usage scenarios using real-world datasets as well as comprehensive user study results.

CHI'18, April 21–26, 2018, Montréal, Canada

ISBN 123-4567-24-567/08/06...$15.00

DOI: http://dx.doi.org/10.475/123_4

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Spatio-temporal data analysis; visual analytics; social media; anomalous event detection

INTRODUCTION

Social networking services (SNS) are becoming part of our daily lives, enabling us to exchange various types of data such as images, videos, and geo-spatial data with others. Leveraging these ever-increasing social media data and extracting useful information from them has emerged as a critical task.

Target task. Among many of such tasks, this paper focuses on detecting and analyzing anomalous events from social media data. To clarify the keywords used in this paper, we define topic as a coherent theme derived through topic modeling technique, and event as something that happened in particular time and region. The definition of anomaly in the context of events can be subjective, but in this paper, we define it as spatio-temporal exclusiveness. That is, if a particular event happened without being on a regular basis over time, e.g., daily, weekly, monthly, and yearly, or over a wide area, we think of it as being exclusive to a particular time and an area, which we consider as anomalous events.

Detecting anomalous events in a timely manner is a problem of utmost importance in various applications, e.g., detecting a disease outbreak, a natural disaster, etc. It was shown that user postings from Twitter, a widely-used SNS service, are functioning as a real-time social sensor for detecting natural disasters [29]. In particular, location-based social media are playing a crucial role in analyzing important events occurring in a particular region and a particular time.

Front-end interface. A multi-zoom level tile-based map interface has been used in various map systems such as the Google Maps, Bing Maps, and so on. Accordingly, this type of interfaces has been actively adopted in various visual analytic systems that support particular analytic capabilities based on geo-tagged data on a map. For example, the tile-based map interface was shown to be useful in summarizing each region for density strips [12]. A zoom-in/out function allows users to analyze geo-tagged data seamlessly from an overview to details. Furthermore, the computation on a tile-based interface is easily parallelizable as long as it can be done independently per tile without having to update the entire map.

Back-end computational model. Topic modeling is a popular technique in text mining as it is effective in generating a meaningful summary from a large collection of documents. When analyzing topic modeling results, each topic is usually represented as a set of its representative keywords that are often semantically meaningful as a whole, as well as its associated documents.

In the backend, noise and everyday language often dominate the resulting topics generated by standard topic modeling, resulting in its limited usability in anomalous event analytics systems. Our system leverages a recently-proposed topic modeling technique called STExNMF [30] that, given multiple sets of documents, can extract exclusive topics of each compared to the others. By integrating this technique with visual analytics environment in a spatio-temporal setting, our system can reveal spatio-temporally exclusive topics with respect to adjacent spatial regions and time points.

In our work, we propose a tile-based visual analytics system based on geo-tagged social media data, such as Twitter data, which allows users to detect and analyze anomalous events in real time. Currently, our system contains 3,598,242 geo-tagged tweets that we collected from New York city in 2013. In general, only a fraction of tweets are known to have geo-tags due to a privacy concern, but we were able to collect about 35,000 tweets on average per day, which can potentially reveal meaningful information about anomalous events occurring in a particular region.

To the best of our knowledge, our system is one of the first such systems integrating the topic modeling techniques effectively with a tile-based map interface. In addition, our system offers a rich set of informative visualization techniques such as heatmaps [1] and glyphs [32, 18] to support the anomalous event detection tasks.

In summary, our visual analytics system called TopicOnTiles includes the following contributions:

- We propose a spatio-temporal visual analytics system that integrates novel topic modeling techniques with a tile-based map interface.
- Our work proposes novel visualization components, such as glyphs and heatmaps, that provides crucial insights in supporting various analytic needs for anomalous event detection.
- We present comprehensive usage scenarios demonstrating the capabilities of our system in detecting anomalous events.
- We conduct comprehensive user study to show the superiority of our system compared to baseline settings.

RELATED WORK

This section discusses the existing studies related to our work from visual analytic perspectives on (1) social media and (2) spatio-temporal data.

Visual analytics on Social Media Analysis

Traditionally, much research has been conducted on visual analytics systems that involve textual data. Diakopoulos et al. [14] provide a visual analytics system that finds and assesses useful information for journalists. ThemeCrowds [2] presents multi-scale tag clouds over time using Twitter data. UTOPIAN [13] utilizes a 2D embedding technique to visualize the extracted topics from user-generated review data.

Recently, the demand for visual analytics systems that take advantage of multimedia data (e.g, images, videos, and so on) is ever-growing. Qian et al. [28] suggest a multi-modal event topic modeling system that can effectively model social media documents, including lengthy text documents with related images. Cai et al. [6] propose a new topic modeling method that utilizes multiple types of data from Twitter such as images, text documents, and so on. Bian et al. [3] develop a social event summarization framework using microblogs of multiple media types including images and videos. The Vox Civitas [15] interface integrates videos from an event with the ability to visually assess the textual social media data. Niu et al. [27] introduce a multi-source-driven asynchronous diffusion model that characterizes the video spread-out behavior and predicts the activation time on social media. Chen et al. [10] present an interactive visual analytics system to extract mobility patterns from geo-tagged social media, which provides variety of visualization views. They proposed additional modules to generate exploratory maps using social media data to support the analysis of events from different perspectives. [11, 9]

Spatio-Temporal Visual Analytics on Anomaly Detection

Visual analytics on time-evolving social media data has been an active research area. Xu et al. [36] propose the streaming text visualization integrating a ThemeRiver-style of visualization with topic modeling approaches for time-evolving topics [33, 35, 36]. Using other types of visualization techniques for time-evolving text, a recent system called FluxFlow [37] visualizes retweeting activities as data points in a timeline to detect the anomalous information spreading patterns in social media. TwitInfo [26] displays a large collections of twitter data using a timeline-based plot and a map visualization. Visual Backchannel [16] presents a timeline visualization using the online discussions from Twitter data. TargetVue [7] presents a novel glyph visualization of anomalous users' temporal usage patterns in social media. More recently, a sedimentation-based metaphor has been proposed to visualize highly dynamic text streams and their corresponding topical evolution [23].

A geographical map-based visual analytics has been actively developed for supporting spatio-temporal document analysis tasks. Andrienko et al. [1] provide heatmap visualization showing the information of events and travels on a map using Flickr data. ScatterBlog [5] extracts topics from microblog messages and visualizes them on a map with spatio-temporally formulated clusters. LeadLine [17] provides meaningful events from social media data using text-processing techniques on a map view. SensePlace2 [25] identifies a particular event such as natural disasters or pandemics using geo-tagged twitter data with a map-based visualization. Thom et al. [34] have built a tag cloud visualization of frequent keywords found in geo-tagged tweets on top of a map interface, and another spatio-temporal visual analytics system has aimed at detecting anomalous events by utilizing topic modeling on Twitter data and their temporal pattern analysis called a seasonal trend analysis [8]. Another system implements a Magic Lens interface [4] that dynamically shows the representative keywords of topics in a user-selected area on a map. Lu et al. [24] present a spatio-temporal visual analytics system that integrates visual tools with a time-series intervention model that enables users to detect interesting events. Similar to our model, this system presents a geographical map that provides most-frequently appearing keywords of the region. However, in a visual analytical system using the social media such as Twitter, such method may not be appropriate since the social media data usually involves much noise.

In our work, we utilize not only the textual content information but also the geo-spatial and temporal information in detecting anomalous events in a particular region and a time point. Using multiple types of data, we leverage a state-of-the-art topic modeling technique that extracts spatio-temporally exclusive topics, enabling users to detect anomaly through topics visualized on the partitioned tile-based map interface. TopicTiles is one of the first visual analytic systems to tightly integrate the effective summarization through topic modeling with a tile-based map interface. Furthermore, we adopt additional visual analytics tools such as heatmaps [1] and glyphs [32, 18] on our system, to reveal anomalous events in a timely manner.

OVERALL DESIGN OF TOPICONTILES

As shown in Fig. 1, TopicOnTiles presents a multi-zoom-level tile map interface where the entire regions are partitioned by tiles, each of which contains spatio-temporally exclusive topics extracted from Twitter data corresponding to the tile. TopicOnTiles also provides various visual elements in the form of glyphs to visualize the analysis results of the topic and its keywords of each tile. Furthermore, the system supports easy access to those raw tweets containing keywords of interest to users.

This section discusses key design considerations of TopicOnTiles, which aims at facilitating real-time detection and in-depth analysis of anomalous events on a tile-map interface. To this end, our system provides exploratory analysis in terms of (1) tile-wise information, (2) user-selected keyword of interest, and (3) spatio-temporal frequency patterns of raw tweets. For the most part, we followed the basic principle based on Shneiderman's mantra [31], *'Overview First, Zoom and Filter, and Details on Demand.'* In the following, we describe the rationales of our system from the perspectives of these three steps.

R1. Providing tile-wise topical summary

Initially, our system shows the geographical overview of tweet contents by providing topic keywords per tile on a map interface. Users can explore topic keywords with flexible spatial granularity by zooming and panning tiles and their corresponding topic keywords on a map. The system provides topic keywords and tile-wise glyphs as a summary of the tile. Topic keywords are computed by our exclusive topic modeling algorithm, which works as a concise, but insightful summary of tile-wise tweet contents. To obtain topics describing the characteristics of events generated in the tile, we extract topics using the spatio-temporally exclusive topic modeling technique, which works by suppressing those topics also found in the neighboring tiles. The tile-wise glyph in the upper-left part of the tile shows the various statistics pertaining to the tile.

R2. Revealing anomalous tiles along with keyword-based topical information

As topics serve as a summary about a document corpus, the topic keywords can reveal what is going on in a particular tile. Hence, analyzing tiles at a keyword level can effective in detecting anomalous events. To this end, We compute the most dominant topics per tile and guide users to potentially anomalous tiles.

R3. Allowing access to raw data with their geospatial and temporal frequency patterns

Users can recognize spatio-temporal patterns of raw tweets containing the user-selected keyword by displaying them in our system. We present the geospatial patterns of tweets by highlighting the locations containing the tweets via the density heatmap. We enable the temporal pattern recognition through visual encodings such as glyphs and vertical grid. In addition, we shows users the raw tweets pertaining to the user-specified keyword.

HOW TOPICONTILES WORKS

This section presents the system design of TopicOnTiles, an interactive visual analytics system that analyzes social media data on a tile-based map for anomalous event detection. We describe backend computations and user interfaces along with their associated design rationales.

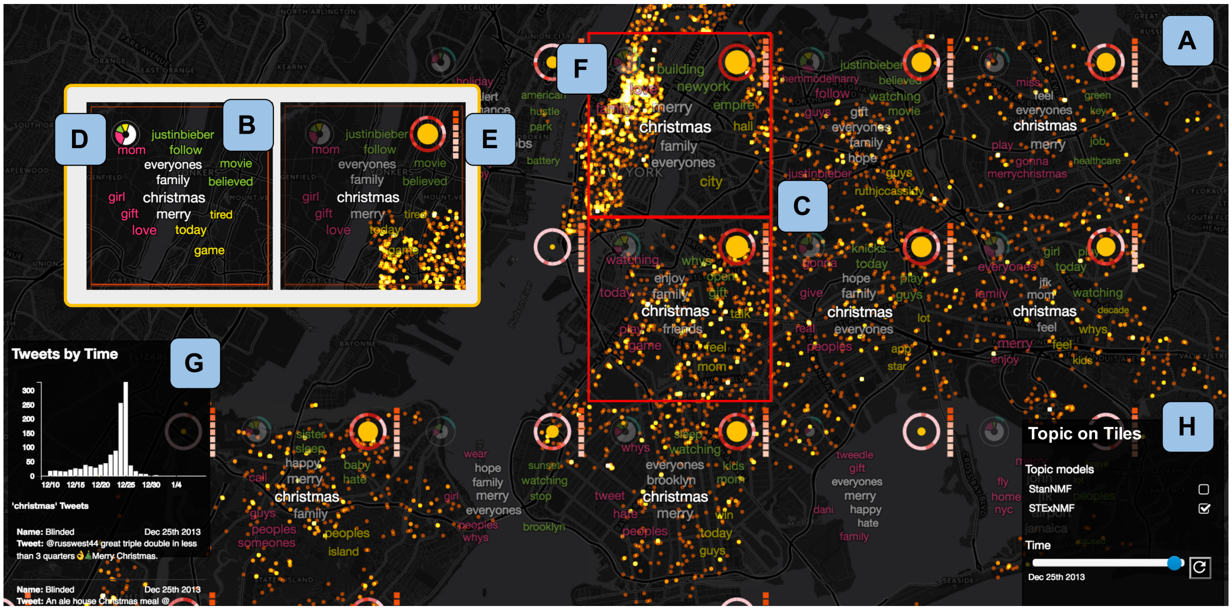

Figure 1. User interface of TopicOnTiles. (A) The background shows a dark-colored geo-spatial map to highlight our main visualization components. The map is divided into a grid of tiles. (B) In each geographical tile, topic keywords computed based on the exclusive topic modeling are visualized, where their colors encode topic indices and the font size indicates the word frequency corresponding to the tile. (C) A tile with a thick rectangular border represents a high anomaly score so that the user can easily pinpoint such a region. (D) A small pie chart per tile shows the relative frequency of topics in the tile, and its radius encodes the tweet count of the tile. The outer layer of the glyph describes the general exclusiveness score of the topics. The size of the inner layer represents the total frequency of the keyword within the tile. A ring shape shows the distribution of tweets containing the user-selected keyword over 24 hours. (E) A vertical grid describes how frequently the selected keyword occurred over the last seven days. (F) The geometric point heatmap represents the locations of each document containing the selected keyword. (G) Users can further explore the bar chart showing the tweet count over time and their raw tweet contents. (H) Users can compare the results according to the topic modeling methods and change the date in the control panel.

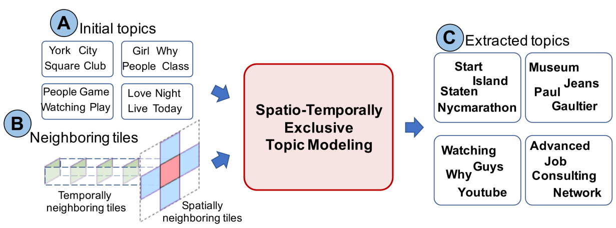

Figure 2. Exclusive topic modeling via STExNMF. (A) For each tile as a center tile, STExNMF first computes topics from its neighboring tiles using the standard topic modeling technique. (B) Afterwards, using the documents of the center tile and those topics extracted from the neighboring tiles, (C) STExNMF computes exclusive topics by iteratively removing the explainable part of documents using the topics from the neighboring tiles.

Spatio-Temporally Exclusive Topic Modeling

The core backend algorithm of our system is a novel exclusive topic modeling method, as illustrated in Fig. 2. The main objective of topic modeling is to provide a comprehensive summary of social media data of a particular tile as a small number of topics (R1). By checking the tile-wise topic keywords displayed on the map, users can pinpoint the anomalous events with their associated tiles.

As Twitter data have a significant amount of noise and everyday language that distracts users from detecting the topics relevant to important spatio-temporal events, the standard topic modeling methods have limited use. In response, we extend a state-of-the-art topic model called STExNMF [30], which extracts the topics of a particular tile exclusive to spatio-temporally neighboring tiles (R2). As illustrated in Fig. 2, we refer to the spatially neighboring tiles as those spatially adjacent to the given tile, and the temporally neighboring ones as those from k previous days (e.g., $k = 7$ in this paper) in the identical location.

Overall, STExNMF works in the following procedure, as shown in Fig. 2. First, given a tile T, we compute topics from the neighboring tiles of T using the standard topic modeling technique [22]. Afterwards, taking as input the documents of T and those topics extracted from the neighboring tiles of T, STExNMF extracts exclusive topics by exhaustively discarding the explainable part of documents in T using the topics from the neighboring tiles of T.

Identification of Anomalous Tiles

We guide the users to the anomalous tiles by using a measure called the anomaly score (R2). Using the topic keywords associated with a given tile, its anomaly score is defined as the total sum of frequency of the unusual topic keywords that had not appeared for the past p days in the corresponding tile, divided by the number of topics in the tile.

User Interfaces for Anomalous Event Detection

In order to facilitate the detection process of anomalous events, our interaction capabilities are designed for two purposes of (1) analyzing the characteristics of the tile using various measures (R1, R2), and (2) drilling down into the spatio-temporal

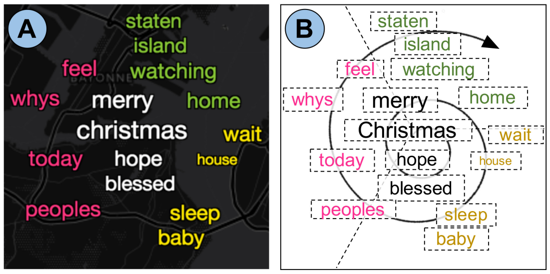

Figure 3. Clustered word cloud for spatio-temporally exclusive topic modeling. We used a spiral layout, placing the four representative topic keywords with the most frequently appearing topic at the center and others along the expanding circumference. The topic keywords are assigned colors in a way that the most dominant topic is shown to be white, the second dominant topic being red, the third being blue, the fourth being green, and the fifth being yellow. The font size of a keyword represents its frequency within the tile.

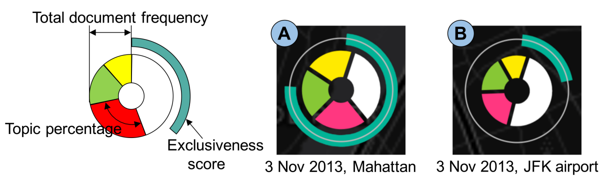

Figure 4. Glyph-structure for tile-based analysis. (A) shows that the exclusiveness score of the Manhattan tile on November 3, 2013 is very high and the distribution of topics is even. This means that the Manhattan area is highly exclusive with respect to its neighboring tiles, both locally and spatially. We can infer that there are many people gathering and talking about different subjects. On the other hand, as shown in (B), JFK airport shows a low exclusiveness score, and the score of the first Topic is very high. This means that when people are visiting airports, they talk about similar topics, and we can probably infer that they send a lot of messages to say hello or good bye at the airport. In reality, "jfk", "john", "kennedy" and "airport" topic are always high around JKF airport.

frequency pattern of tweets as well as their raw content information (R3), as will be described below.

Word cloud visualization of exclusive topics

As the document size of each tile varies, we extract a different number of topics depending on the number of documents contained in a single tile. That is, we obtain a single topic for the tile with less than 150 tweets, two topics for the tile with more than 150 and less than 400 tweets, three for that with more than 400 and less than 1,500 tweets, and four for that with more than 1,500 tweets. In this manner, our system naturally provides a detailed summary on the tiles containing a relatively large number of tweets (R1, R2). In addition, if multiple topics contains the same top keywords within a single tile, we display only one of them to avoid redundancy and visual clutter.

We visualize the topic keywords on top of geo-spatial tiles by using a spiral layout, as shown in Fig. 3.

Tile visualization modules showing keyword-based tile analysis

We provide two visualization modules related to tiles that derive from keyword-based analysis (R2). One is visual emphasis of anomalous tiles based on the anomaly score described

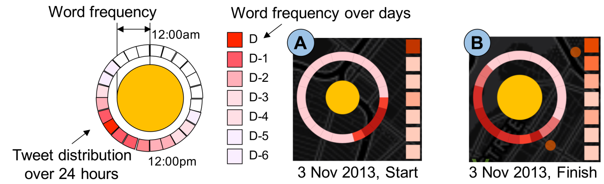

Figure 5. Glyph-structure for keyword-based analysis. (A) and (B) show the glyphs of the words 'Start' and 'Finish' on November 3, 2013, respectively. If today's vertical grid is the darkest, then we can infer that today is the day of the marathon. The tweet distribution containing the word 'start' is high in the morning, and the tweet distribution containing 'finish' is also high in the afternoon. The user can guess that it started in the morning and ended in the afternoon.

previously. The other is the glyph located at the upper-left of each tile.

Visual emphasis of anomalous tiles. The visual emphasis of anomalous tiles works by putting red-colored edges surrounding the anomalous tile, as shown in Fig. 1(A). The thickness and the brightness of the edge encodes the anomaly score, so the brighter and thicker the edge becomes, the more anomalous context the tile contains.

The upper left glyph for topical analysis. Our glyph design is composed of two parts: one on the inner layer and the other on the outer layer of the glyph, as shown in Fig. 4. The size of the circle is determined by the total tweet count within the tile. The inner part of the circle shows a pie chart, representing the relative tweet counts corresponding to different topics. This pie chart highlights the dominant topic among the extracted topics. Each portion of the pie chart is drawn as the percentage of the topic score of the topic with respect to the all of the topic scores combined. In here too, the topic score is measured by the frequency of the keywords that make up the topic.

The outer layer of the glyph describes the general exclusiveness score of the topics belonging to the tile $V_{C,t}$. We express the score as $G_{C,t}$. $G_{C,t}$ is defined as the Jaccard distance between the topics extracted from the standard topic modeling and the topics from exclusive topic modeling of the same tile. In computing the standard topic modeling, we utilize the standard NMF algorithm. Consequently, $G_{C,t}$ is defined as

$$G_{C,t} = 1 - \frac{|S_{ex} \cap S_{st}|}{|S_{ex} \cup S_{st}|}, \quad (1)$$

where S_{st} and S_{ex} refer to the top l keywords from the topics of each tile using standard topic modeling and exclusive topic modeling, respectively. $|\cdot|$ denotes the cardinality of a set. $G_{C,t}$ measures the amount of changes in keywords between the standard topic modeling and our exclusive topic modeling. This measure indicates how much the tile is affected by removing the topics from the neighboring tiles when extracting exclusive topics of the tile. In other words, the tiles with a high value of $G_{C,t}$ share a significant amount of topics with the neighboring tiles.

Interactive visualization of raw tweets

The analysis of raw tweets aims at providing keyword-wise pattern analysis of the raw tweets, both (1) spatially and (2)

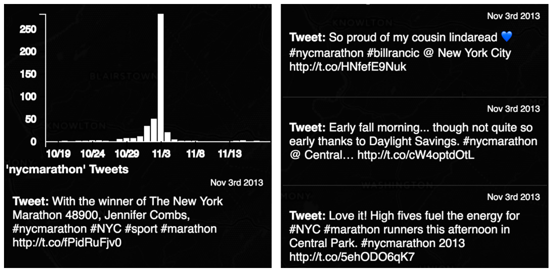

Figure 6. Temporal pattern visualization of raw tweets. The bar graph shows the frequency of the tweets containing the user-selected keyword and the panel below the bar graph shows the content of these raw tweets.

temporally (R2, R3), and (3) displaying the keyword-related raw tweets for further details.

Spatial pattern analysis. When the user clicks on a keyword, the system illustrates a heatmap showing the spatial distribution of the tweets containing the keyword. A brighter dots corresponds to a region with more tweets.

Temporal pattern analysis. Temporal patterns of raw tweets are visualized in two forms: a circular glyph and a vertical grid, as shown in Fig. 5. When clicking a keyword on a tile, the circular glyph is visualized at the upper right part of the tile.

This glyph has the inner and the outer layers. The size of the former represents the total frequency of the keyword within the tile. A ring shape at the outer layer of the circular glyph shows the distribution of tweets containing the user-selected keyword over 24 hours. That is, the ring is divided into 24 equal-sized slots corresponding to each hour in a day. A dark color in a particular slot represents a high frequency of tweets posted in the corresponding hour.

A vertical grid next to the circular glyph describes how frequently the selected keyword occurred over the last seven days, including the current day. A dark color represent a high frequency. Using this vertical grid visualization, one can identify useful patterns, such as a temporally sudden increase in the keyword frequency in recent days or a uniformly distributed frequency patterns over past days.

Giving access to the raw tweets. As shown in Fig. 6, we provide a pop-up visualization to enable users to check the details of raw tweets associated with a particular events. Once clicking a keyword, the window pops up, showing a bar graph and a window containing the raw tweets. The bar graph at the top of the dashboard shows the frequency of the tweets containing the user-selected keyword. The panel below the bar graph shows the content of these raw tweets.

Tile-based map interface

TopicOnTiles supports zooming and panning interactions on a tile-based map, as shown in Fig. 8(A) and (B). Each zoom-in enlarges the map size twice while maintaining the same number of tiles in a screen. That is, the size of each tile region gets smaller by half of the previous size of the tile region, providing a finer-grained topical information per tile.

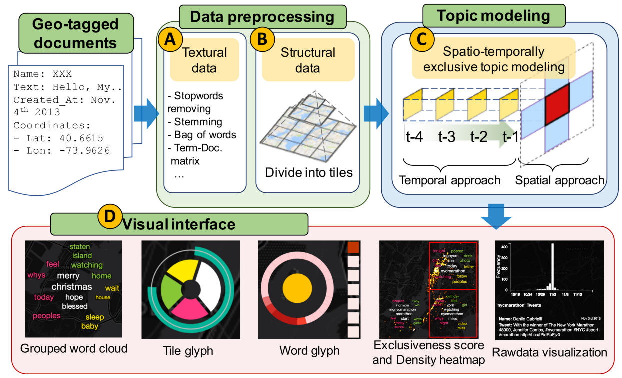

Figure 7. The overall architecture of TopicOnTiles. TopicOnTiles is composed of three parts: data preprocessing, topic modeling, and interactive visualization interfaces. (A) In the data preprocessing step, we collect geo-tags, time stamps, and raw text of each tweets. The raw text is preprocessed via bag-of-words vector encoding with stopword removal and stemming. Then, (B) we split the entire set of bag-of-words vectors with respect to different dates and geo-locations for tiling. (C) We precompute all the topic modeling results to ensure real-time interactive visualization in TopicOnTiles. Afterwards, (D) we calculate other analytic measures, such as anomaly scores, topic scores, the spatial and the temporal distribution of tweets, and so on, on the fly once a user interaction is issued.

System Architecture

As shown in Fig. 7, the overall architecture of TopicOnTiles is composed of three parts: (1) data preprocessing, (2) topic modeling, and (3) interactive visualization interfaces.

In the data preprocessing step, we collect geo-tags, time stamps, and raw text of each tweets. The raw text is preprocessed via bag-of-words vector encoding with stopword removal and stemming (Fig. 7(A)). Then, we split the entire set of bag-of-words vectors with respect to different dates and geo-locations for tiling (Fig. 7(B)). We pre-compute all the topic modeling results to ensure real-time interactive visualization in TopicOnTiles (Fig. 7(C)). Afterwards, we calculate other analytic measures, such as anomaly scores, topic scores, the spatial and the temporal distribution of tweets, and so on, on the fly once a user interaction is issued (Fig. 7(D)).

System Implementation

TopicOnTiles is composed of the front-end user interfaces and the back-end computational modules. These two communicate with each other using RESTful API over HTTP, where the web-server for the front-end is implemented using Golang[1]. The user interfaces for interactive visualization are written in Javascript while \the back-end computations are done in Python. The web-server of the back-end is implemented using the Flask library. To guarantee real-time responses, we precompute some processes which takes long time to process. A precomputing module is implemented in Matlab, and the data are stored in MongoDB[2] and file format. TopicOnTiles operates on the desktop server with Intel Zeon E5-2687W v3 and 384GB memory. When ignoring the network latency, the response time of all the supported user interactions is less than a second.

[1] https://golang.org/
[2] https://www.mongodb.com/

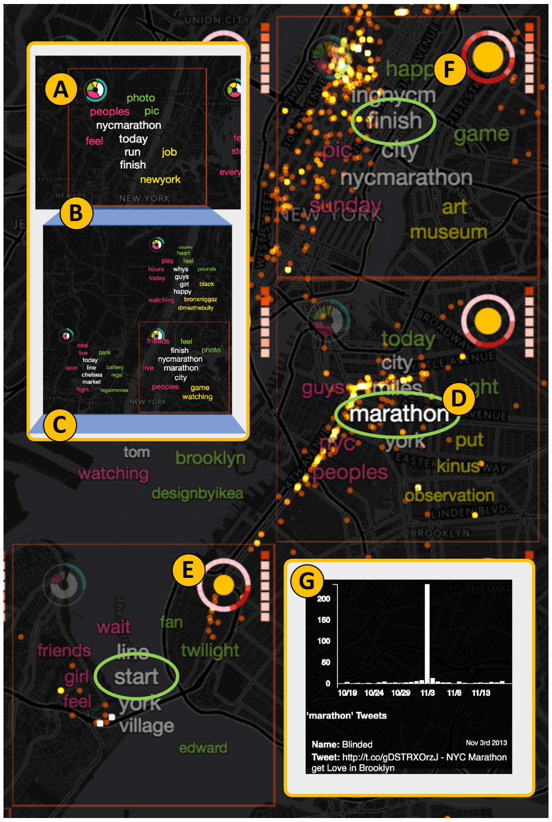

Figure 8. Snapshot of TopicOnTiles showing the tweets posted on Nov. 3, 2013. (A) Thick edges of the tile containing the topic keywords related to marathons indicates an anomalous event in the corresponding tile. (B) Zooming in the region shows more details. (C) The further zooming in reveals such keywords as 'start,' 'marathon,' and 'finish,' as highlighted in green ellipses. (D) Clicking the keyword 'marathon,' its density heatmap (bright orange or yellow dots) shows the actual geo-locations of the tweets containing 'marathon.' Additionally, (E)(F) the glyph visualization of 'marathon' shows the total frequency as an inner circle as well as the temporal frequency pattern over 24 hours as an outer circle and that over the last seven days as vertical grids. (G) Finally, the temporal frequency of tweets over a one-month period and the original tweet contents are provided.

USAGE SCENARIOS

This section presents two usage scenarios of TopicOnTiles based on the tweets collected from New York City in year 2013.

ING New York City Marathon

A user wants to detect an anomalous event that took place on November 3, 2013, and we assume ING NYC Marathon is one such event. Initially, using TopicOnTiles, the user analyzes topic keywords from the lowest zoom level and finds several tiles containing the keyword 'marathon.' After several minutes of hesitation, he decides to zoom in the tile having keywords 'finish,' 'city,' 'york,' and 'marathon.' After viewing topics of some tiles, he assumes that a marathon competition is taking place in New York, but is not sure whether the event is anomalous. For more information, the user zooms in the map to get more insights.

First, as an overview, on November 3, we see topics from various tiles, as seen in Fig. 8(A). The edge of the tile containing the topic with keywords 'finish,' 'run,' 'city,' and 'nycmarathon' is the thickest, meaning that the tile has the highest anomaly score.

Then, as shown in Fig. 8(B) the user zooms in the map to discover topics from the smaller tiles. He finds out that the edges covering the tiles with keywords 'nycmarathon' are the thickest.

The user zooms in again (Fig. 8(C)) and finds out that the edges near the tile from the previous zoom level are the thickest. Keywords that imply the starting and the finishing lines of the course, such as 'start' and 'marathon,' and 'finish' are revealed in some tiles. The user focuses on the tile having the keyword 'marathon,' as shown in Fig. 8(D). The user clicks on the keyword, 'marathon.' The density heatmap shows that there are many people in the area who tweeted using the keyword 'marathon.' By looking at the vertical grids, he finds out that the keyword started to appear only one or two days before. The yellow inner part of the glyph indicates that the keyword 'marathon' appeared most frequently near the Central Park. The outer layer of the glyph shows the time distribution of the keyword by hours. The user starts to feel that the marathon was not a common event.

In the tile near the Staten Island (Fig. 8(E)), the keyword 'marathon' mostly appears in the morning, whereas in the tile near the Central Park (Fig. 8(F)) the keyword appears throughout the day.

Judging from these facts, the user thinks that the NYC Marathon has been one of the dominant events happening in New York City, where the starting point of the course seems to be near Staten Island, and the course is mainly held in Central Park. Then, the user refers to the pop-up window displayed at the lower-left part of the system (Fig. 8(G)) to check the raw tweets for the event, and finds that the marathon has taken place on that day.

Likewise, our system is capable of detecting and guiding the user to anomalous events such as the *2013 ING NYC Marathon*. Through topics and various spatio-temporal visual components, users can analyze on anomalous events such as the marathon event, and can also find out about the interesting facts such as the starting and the ending lines of the course and a sketchy knowledge on the entire marathon course.

The Trayvon Martin Protest and the MLB All-Star Game

We present the usage scenario that takes place in July 14 of New York City, the day where the Trayvon Martin Protest and the 2013 Major League Baseball (MLB) All-Star Game took place. The user is trying to analyze the the anomalous events that took place in July 14, 2013.

Initially, our system puts a thick rectangular edge on the area near the Central Park and the Citifield, the homeground of the New York Mets, a New York City-based MLB team. As shown in Fig. 9, the edges are colored in the areas around the Central Park (Fig. 9(A)) and the Citifield (Fig. 9(B)). It seems that an anomalous event has taken place near those areas.

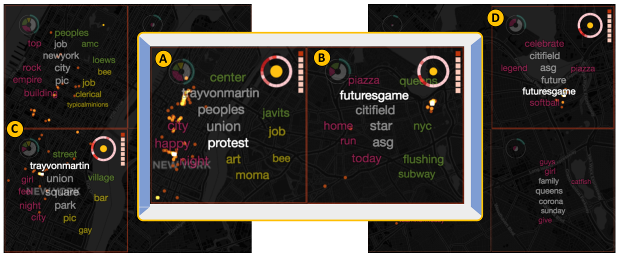

Figure 9. The Overview image of Jul. 14, 2013. (A), (B) The edges of the tile containing the topic with keywords related to Trayvon Martin Protest and MLB All-Star Futures Game were highlighted. Zoom in the tile of Trayvon Martin Protest to analyze on detail, (C) the keywords such as 'trayvonmartin,' 'union,' 'square,', and 'park' are revealed. The vertical grids of those keywords indicate that these keywords started to appear heavily starting from July 14. The temporal glyph shows that the event has mainly taken place during the night. Zoom in the tile of MLB All-Star Futures Game to analyze on detail, (D) the keywords such as 'celebrate,' 'piazza,' 'legend,', and 'softball' are revealed. It shows another event, the All-Star Legends & Celebrity Softball Game, was took place, and Piazza took outstanding performance in this game.

The user starts to look into other visual analytic tools. The vertical grids of those keywords indicate that these keywords started to appear heavily starting from July 14. The temporal glyph shows that these keywords were mainly used at night, implying that the event has mainly taken place during the night.

He clicks on the keyword 'trayvonmartin' of the tile having other keywords such as 'union' and 'square.' The density heatmap shows a specific point colored in white, a probable location of the anomaous event (Fig. 9(C)). The temporal grid shows that the keyword emerged extensively in July 14. Having seen all of those tools, the user senses that an anomalous event had taken place, having to do with 'protest' and 'trayvonmartin.' For further clarification, the user checks on the pop-up window. After looking at the tweets, the user is sure that a massive protest happened near the Central Park on July 14, 2013.

Indeed, on that day, the protest event against racial discrimination related to Trayvon Martin took place near the Central Park on July 14.

Then, once the user checks on the Trayvon Martin Protest, the user looks into another tile covered in thicker edge, the one near the Citifield stadium (Fig. 9(B)). Keywords such as 'citifield,' 'star,' 'futuresgame,' and 'asg(All Star Game)' seem to imply that the MLB All-Star Game was held at the Citifield Stadium. The vertical grid show that the keyword 'futuresgame' has started to on July 14. We can also infer that the keyword is mainly used during the afternoon. The user assumes that the Futures' All-Star Game took place on that day (Fig. 9(D)).

The news in July 14 tells us that the Futures' All-Star Game took place in Citifield Stadium on July 14, 2013.

EVALUATION: USER STUDY

We conducted a user study with 30 participants in total. The goal of our study is two fold. The first is to understand how much our system helps a user recognize anomalous events, and the second is to identify the effective components of our system in detecting anomalous events. In the following, we describe the study design and the results.

Study Design

For our study, we used geo-tagged tweets of New York City from September 20 to December 31 in 2013. Using these data, we prepared for three different settings. The third setting is TopicOnTiles with its full capabilities while the second one shows only the topic keywords on a tile-based map without our novel visualization components such as glyph visualization, tile edges representing the anomaly score, and the density heatmaps. Finally, the first setting is identical to the second one except that the topic keywords are obtained by the standard topic modeling rather than STExNMF.

We designed our study as a between-subject study where each of the three different settings were tested separately by ten different participants. That is, each participant was exposed to only one setting without being informed of the other two.

In each of the study session, which took about 40 minutes on average, we started with the brief description about the goal of the system and how to use it, and the participant was then asked to '*find and write down all of the events that happened in New York City between November 1 and 5, 2013'* in a free-text form, where each event was described as a full sentence, a phrase, or a single keyword. Each participant was given eight minutes to perform this job, but s/he was allowed to end the session earlier. Next, we repeated a similar task with the same participant for a different time period from November 29 to December 2 in 2013. Afterward, we asked each participant to fill out the the computer system usability questionnaire (CSUQ).[3]

Once collecting the event descriptions from all participants, we first removed obviously meaningless keywords, such as 'love,' 'people,' and 'hi.' Next, each of the three authors of this paper separately classified each event description as either being spatio-temporally ordinary or anomalous without knowing which setting it was collected from. We then collected those event descriptions unanimously labeled as being anomalous. Finally, using these anomalous events, we compared the performance of the three settings in term of the total number of anomalous events a user found divided by the amount of time s/he spent.

Study Results

Quantitative comparisons. The comparisons of the number of results of the first part of the questionnaire is shown in Fig. 10. We measured the number of anomalous events the participants found for 8 minutes, averaged the scores by their groups, and presented them in events per minute. The bar graph shows the mean values of the scores and the error bar refers to the standard deviation of them. The results in Fig. 10 show the averaged performances of the first and the second experiments.

A one-way between subjects analysis of variance (ANOVA) [19] was conducted between the three groups. to evaluate the effect of visualization components in

[3] http://garyperlman.com/quest/quest.cgi

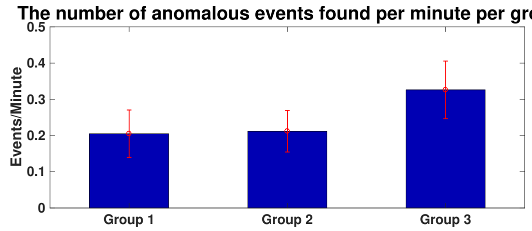

Figure 10. Comparison of the number of anomalous events detected per minute. Group 3, which corresponds to TopicOnTiles with its full capabilities, is shown to be significantly different from each of Group 1 and Group 2, while Group 1 and Group 2 does not show a statistically significant difference.

finding the anomalous events. The ANOVA test showed that at least one of the groups were significantly different from the others. We set the significance level α as 0.05. The p-value of the test was $p = 0.010$ ($F(2,57) = 4.972$).

To identify the difference in the means of the different groups, we conducted a post-hoc analysis using the Bonferroni test [20]. The results indicated that the mean score for the Group 3 (M=.3258, SD=.1590) was significantly different than those of the other groups, Group 1 (M=.2048, SD=.1309) and Group 2 (M=.2119, SD=.1148.), with the significance probabilities .020 and .032, respectively. This experiment reveals that the visual interfaces installed in the system play an important role in detecting anomalous events.

In addition, the graph shows that the score of the Group 2 is slightly higher than those of the Group 1, although the Bonferroni test results suggest that Group 2 did not differ significantly from Group 1 with the significance probability of almost 1. Be that as it may, some participants were able to distinguish the differences in the functionalities of the two methods. One of the participants said,*'the standard topic modeling is helpful in detecting major events, but as the keywords are more likely to appear again in other tiles, the exclusive topic modeling can be more useful in inferring various topics because it removes the prevalent keywords that are common in the neighboring tiles.'*

The CSUQ. The second part of the questionnaire is the CSUQ. With the CSUQ, we conducted the subjective surveys of the participants. The results are shown in Fig. 6.2. The results show that the system helps accomplishing the work efficiently (Q6, Q7) and that the interface is pleasant (Q11, Q12).

We presume that the high scores in the 'efficient work' is mainly owing to the novel topic modeling technique and the visual encodings that enable both the spatial and the temporal analysis of the topics in extracting anomalous events. In addition to that, the various kinds of visually analytical tools on the system contributed to arising the curiosity of some users. One participant said, *'allowing a facilitated spatial and temporal analysis of the events using visual components helped me to gain various information on events, and personally such visual display has even aroused my curiosity of wanting to know more.'*

However, though not as low in the objective sense, the questionnaire shows that the relatively lowest score on the question

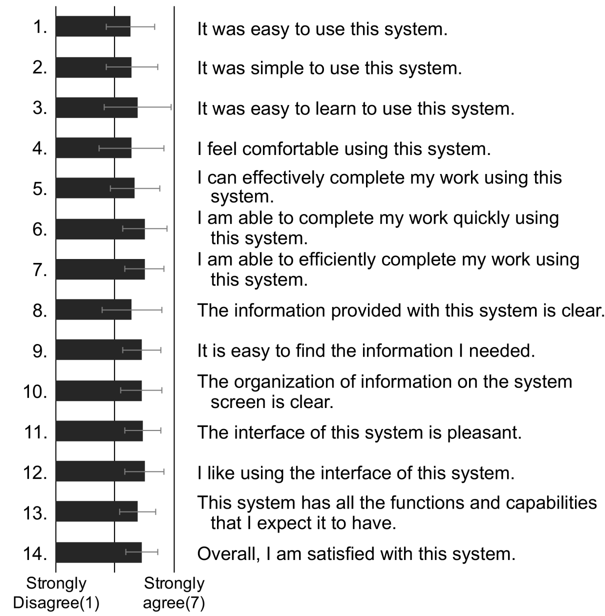

Figure 11. The summary of CSUQ questions and answers. TopicOnTiles has been highly rated in many areas, especially in terms of its efficiency and its interface.

was 'the information provided with this system is clear' (Q8). For some people, our visual components could have been perceptually difficult to use. One participant said, *'the visual ingredients gives us useful information such as the spatial and temporal patterns of a particular keyword. However, as there are too various analytical tools displayed, sometimes I get a little hesitant in deciding which tool to use.'*

DISCUSSIONS

This section discusses the limitations of TopicOnTiles based on the results and the comments from the user studies.

Determining the number of topics

There has been numerous research that attempted to automatically determine the optimal number of topics, or clusters in topic models. The hierarchical topic modeling by Kuang et al. [21] utilizes the information gain as a way to determine the number of topics of a document corpus. Although mathematically optimized, the topics extracted from the algorithm may not reveal all of the topics pertaining to the anomalous events. In some cases, it may deliver superfluous topics that do not contain significant meanings. Moreover, as our system intends to deliver information of the tile in a limited amount of space, in some cases, it may not be able to capacitate all of the topics that can be delivered. Another topic modeling system, UTOPIAN [13] merges or splits the topics with the help of user interaction. However, in systems such as ours that deal with large-scale data, providing interactions to all of the tiles can be can be a burden to our system, and could severely lower the performances of our system.

In coping with such vulnerability, we assumed that the number of anomalous events is proportional to the number of datasets, and set the number of topics of a tile region according to the number of tweets residing in each tile. However, such assumption has flaws in several aspects. It is not capable of distinguishing the everyday language topics from the meaningful topics. The topics displayed on tile may not detect all of the meaningful topics that may contain information on anomalous events. We aim to devise an algorithm that can determine the number of topics to be extracted in anomaly detection.

Events distributed at the boundary of the tiles

In a tile-based map, the regions are subdivided into equally sized grids, regardless of the district boundaries. This can cause problems when an anomalous event takes place near the tile boundaries. The tweets that talk about the event would be fragmented into multiple tiles, resulting in the diminishment of the event. In our system, we tried to address these issues by providing users summarized topics from various zoom levels. By enabling users to view topics of the area from differently fragmented tiles, not only it provides users the topical summary of the tile, but users can also find out the different topics, which could not be found in tiles from the smaller zoom levels. However, if the topic that appeared in the tile on lower zoom level disappears in those from higher zoom levels, it may not provide meaningful information for some users that want to know about the details of the event they found on the lower zoom level. We plan to develop a method that can provide a more sophisticated method that can subdivide the regions considering its districtual context, or the location the anomalous event took place.

Facilitating user interfaces for first-time users

When asked about the usability of our system, a majority of the participants gave positive responses. One of them commented, '*the visual components were helpful, because they guided us to the tiles where the anomalous event had taken place.*' Moreover, many of the participants said that the temporal pattern analysis of keywords using glyph visualizations helped them to notice the anomalous events.

However, there have been comments from some participants that the system may take some time to get accustomed to for those having not much knowledge on the related domains. One of participants said, '*there exists too much visual components, and sometimes it is confusing which one I should be looking at.*' This accords with the CSUQ score, where the relatively lowest score was Q8. In designing TopicOnTiles, the proper balance between simplicity and complexity in design has always been a big issue. As the target users of TopicOnTiles are not solely confined to experts, we aim to reorganize the analytical tools of our system under a purpose-driven objective.

CONCLUSION AND FUTURE WORK

In this paper, we have introduced a tile-based visual analytics system for anomalous event detection called TopicOnTiles. TopicOnTiles conducts a spatio-temporal analysis based on a tile-based map interface using the spatio-temporally exclusive topic modeling technique that extracts topics with respect to the neighboring tiles. In addition, the system provides various visual encodings such as the glyphs, vertical grid and heatmaps to facilitate anomalous event detection. We have shown usage scenarios by detecting and discovering the contents of the *2013 ING NYC Marathon* and the Trayvon Martin protest. Furthermore, we have provided user studies showing the efficiency and the usability of TopicOnTiles through the quantitative comparison and the CSUQ.

As our future work, we plan to address the issues mentioned in the previous section. Furthermore, we plan to extend our visual system to a real-time spatio-temporal anomaly detection system that can solve various problems arising in urban areas such as pollution, crime and so on by associating the geo-tagged textual social media data with other types of geo-tagged data, such as the air pollution data, GPS data and so on.

ACKNOWLEDGEMENTS

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. DOE and supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. NRF-2016R1C1B2015924). The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

REFERENCES

1. G. Andrienko, N. Andrienko, P. Bak, S. Kisilevich, and D. Keim. Analysis of community-contributed space- and time-referenced data (example of flickr and panoramio photos). In *In Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 213–214, 2009.
2. D. Archambault, D. Greene, P. Cunningham, and N. Hurley. Themecrowds: Multiresolution summaries of twitter usage. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents (SMUC)*, pages 77–84, 2011.
3. J. Bian, Y. Yang, H. Zhang, and T. S. Chua. Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia*, 17(2):216–228, Feb 2015.
4. E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose. Toolglass and magic lenses: The see-through interface. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 73–80, 1993.
5. H. Bosch, D. Thom, M. Wörner, S. Koch, E. Püttmann, D. Jäckle, and T. Ertl. Scatterblogs: Geo-spatial document analysis. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 309–310, 2011.
6. H. Cai, Y. Yang, X. Li, and Z. Huang. What are popular: Exploring twitter features for event detection, tracking and visualization. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM)*, pages 89–98, 2015.
7. N. Cao, C. Shi, S. Lin, J. Lu, Y. R. Lin, and C. Y. Lin. Targetvue: Visual analysis of anomalous user behaviors

in online communication systems. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(1):280–289, 2016.

8. J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 143–152, 2012.

9. S. Chen, S. Chen, L. Lin, X. Yuan, J. Liang, and X. Zhang. E-map: A visual analytics approach for exploring significant event evolutions in social media. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2017.

10. S. Chen, X. Yuan, Z. Wang, C. Guo, J. Liang, Z. Wang, X. L. Zhang, and J. Zhang. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(1):270–279, 2016.

11. S. Chen, W. Z., J. Liang, X. Yuan, N. Cao, and W. Y. D-map: Visual analysis of ego-centric information diffusion patterns in social media. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 41–50, 2016.

12. D. Cheng, P. Schretlen, N. Kronenfeld, N. Bozowsky, and W. Wright. Tile based visual analytics for twitter big data exploratory analysis. In *Proceedings of the IEEE International Conference on Big Data (Big Data)*, pages 2–4, 2013.

13. J. Choo, C. Lee, C. K. Reddy, and H. Park. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 19(12):1992–2001, 2013.

14. N. Diakopoulos, M. De Choudhury, and M. Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2451–2460, 2012.

15. N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 115–122, 2010.

16. M. Dörk, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 16(6):1129–1138, 2010.

17. W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 93–102, 2012.

18. C. Dunne and B. Shneiderman. Motif simplification: Improving network visualization readability with fan, connector, and clique glyphs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2013.

19. D. Howell. *Statistical Methods for Psychology*. Duxbury, 2002.

20. G. Keppel and T. Wickens. *Design and analysis: a researcher's handbook*. Pearson Education Inc., 4th edition, 2004.

21. D. Kuang and H. Park. Fast rank-2 nonnegative matrix factorization for hierarchical document clustering. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 739–747, 2013.

22. D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

23. S. Liu, J. Yin, X. Wang, W. Cui, K. Cao, and J. Pei. Online visual analytics of text streams. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(11):2451–2466, 2016.

24. Y. Lu, M. Steptoe, S. Burke, H. Wang, J. Y. Tsai, H. Davulcu, D. Montgomery, S. R. Corman, and R. Maciejewski. Exploring evolving media discourse through event cueing. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(1):220–229, 2016.

25. A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 181–190, 2011.

26. 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 SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 227–236, 2011.

27. G. Niu, X. Fan, V. O. K. Li, Y. Long, and K. Xu. Multi-source-driven asynchronous diffusion model for video-sharing in online social networks. *IEEE Transactions on Multimedia*, 16(7):2025–2037, 2014.

28. S. Qian, T. Zhang, C. Xu, and J. Shao. Multi-modal event topic model for social event analysis. *IEEE Transactions on Multimedia*, 18(2):233–246, 2016.

29. 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*, pages 851–860, 2010.

30. S. Shin, M. Choi, J. Choi, S. Langevin, C. Bethune, P. Horne, N. Kronenfeld, R. Kannan, B. Drake, H. Park, and J. Choo. Stexnmf: Spatio-temporally exclusive topic

discovery for anomalous event detection. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 435–444, 2017.

31. B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336–343, 1996.

32. A. Spoerri. Infocrystal: A visual tool for information retrieval. In *Proceedings of the 4th Conference on Visualization*, pages 150–157, 1993.

33. G. Sun, Y. Wu, S. Liu, T.-Q. Peng, J. J. Zhu, and R. Liang. Evoriver: Visual analysis of topic coopetition on social media. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 20(12):1753–1762, 2014.

34. D. Thom, H. Bosch, S. Koch, M. Worner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *Proceedings of the IEEE Pacific Visualization Symposium*, pages 41–48, 2012.

35. F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 575–582, 2004.

36. P. Xu, Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J. J. Zhu, and H. Qu. Visual analysis of topic competition on social media. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 19(12):2012–2021, 2013.

37. J. Zhao, N. Cao, Z. Wen, Y. Song, Y. R. Lin, and C. Collins. Fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 20(12):1773–1782, 2014.