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

Minsuk Choi[‡,*], Sungbok Shin[‡], Jinho Choi[‡],
Scott Langevin[§], Christopher Bethune[§], Philippe Horne[§], Nathan Kronenfeld[§],
Ramakrishnan Kannan[∨], Barry Drake[∧], Haesun Park[¶], Jaegul Choo[‡†]

[‡]Korea University
Seoul, South Korea

[§]Uncharted Software Inc.
Toronto, Canada

[¶]Georgia Tech.
Atlanta, USA

[∨]Oak Ridge
National Laboratory
Oak Ridge, USA

[∧]Georgia Tech.
Research Institute
Atlanta, USA

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.

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 an integral 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. This study 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; however, in this paper, we define it as spatio-temporal exclusiveness. That is, if a particular event occurred once, but not repeatedly 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. These are considered 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 or a natural disaster. It was shown that user postings from Twitter, a widely-used SNS service, can function 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 Google Maps and Bing Maps. 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 because of its effectiveness in

*E-mail: mchoi@korea.ac.kr.
†E-mail: jchoo@korea.ac.kr, the corresponding author.

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

DOI: https://doi.org/10.1145/3173574.3174157

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 the 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 geotagged tweets that we collected from New York city in 2013. In general, only a fraction of tweets are known to have geotags 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 involving textual data. Diakopoulos et al. [14] provide a visual analytics system that finds and assesses useful information for journalists. ThemeCrowds [2] presents multiscale 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-increasing. 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 and text documents. 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 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 a 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 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 system 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 spatiotemporally 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. Another spatio-temporal visual analytics system has aimed at detecting anomalous events by utilizing topic modeling on Twitter data; their temporal pattern analysis is referred to as 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 social media such as Twitter, such method may not be appropriate because social media data usually involves much noise.

In our work, we utilize not only the textual content information but also geo-spatial and temporal information in detecting anomalous events in a particular region and time. 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. TopicOnTiles is one of the first visual analytic systems to robustly integrate 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 a topic and the 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 rationale 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 commonly 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 be 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 a 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 show the temporal patterns through visual encodings such as glyphs and vertical grids. In addition, we show 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 supports anomalous event detection using social media data on a tile-based map interface. We describe the user interfaces and the backend computations along with their associated design rationales.

User Interfaces for Anomalous Event Detection

TopicOnTiles is built on a tile-based map system, where the entire spatial regions are divided into equal-sized grids. Like other map-based systems, our system supports zooming and panning interactions, as shown in Fig. 8(A) and (B). We also provide additional user interactions based on the spatio-temporal analysis of social media data that can work as effective social sensors [29]. To be specific, our interaction capabilities are designed for two purposes: (1) guiding users to anomalous events (R1, R2) and (2) providing the users with tile-wise spatio-temporal pattern analysis (R1), exploration of keywords associated with events (R2), as well as their raw content information (R3) using various visual tools, as will be described below.

Visualizing Anomalous Topics

This section describes the backend spatio-temporally exclusive topic modeling algorithm and the visualization of its results

Spatio-temporally exclusive topic modeling. The core backend algorithm of our system is a novel exclusive topic modeling method designed to extract spatio-temporally exclusive topics. The main objective of topic modeling in our system is to provide a comprehensive summary of social media data of a particular tile as a small number of topics (R1). By checking tile-wise topic keywords displayed on the map, users can pinpoint the anomalous events with their associated tiles. The algorithm utilized in our system, STExNMF [30], summarizes

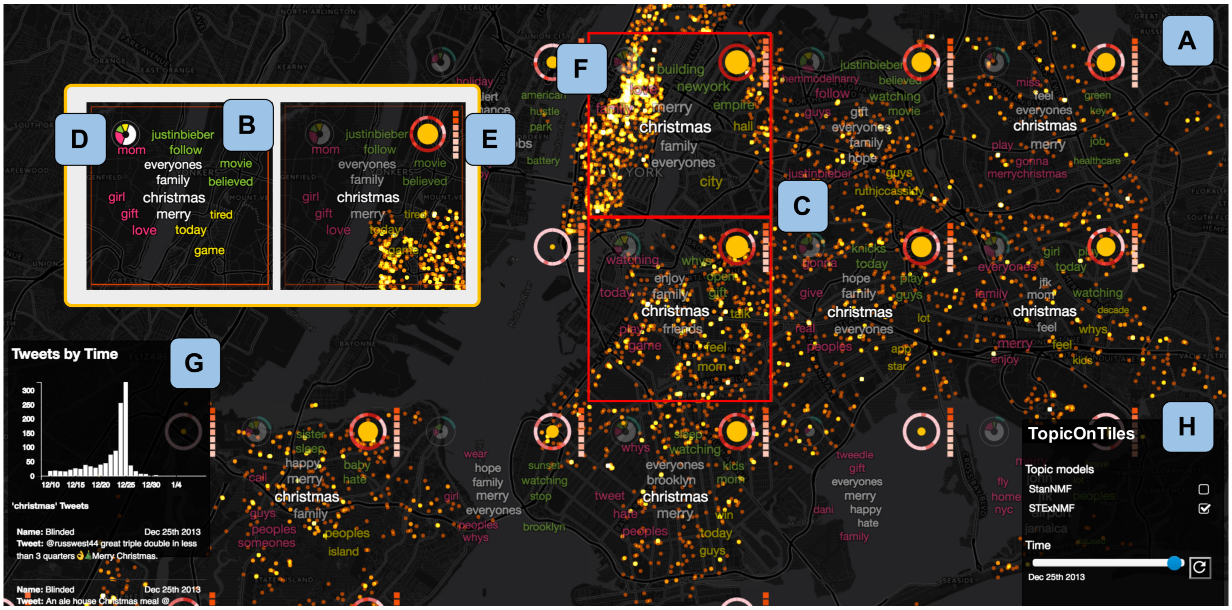

Figure 1. User interface of TopicOnTiles. (A) The background shows a dark-colored geo-spatial map to highlight our main visualization components. The map is divided into a grid of tiles. (B) In each geographical tile, topic keywords computed based on the exclusive topic modeling are visualized, where 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.

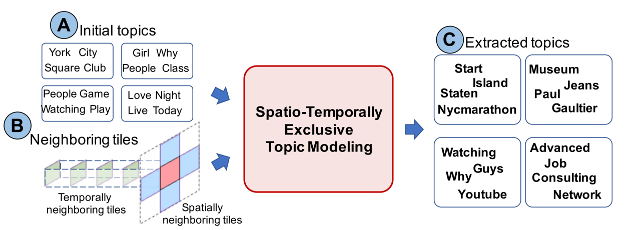

Figure 2. Exclusive topic modeling via STExNMF. (A) For each tile as a center tile, STExNMF first computes topics from its neighboring tiles using the standard topic modeling technique. (B) Afterwards, using the documents of the center tile and those topics extracted from the neighboring tiles, (C) STExNMF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

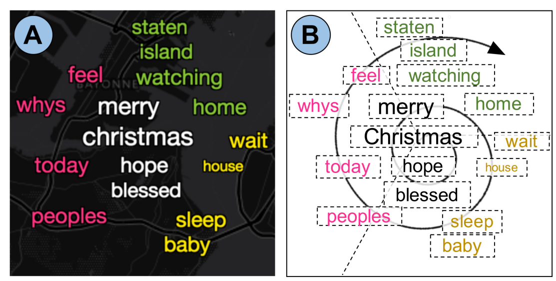

Figure 3. Topical word cloud for spatio-temporally exclusive topic modeling on a spiral layout. It places the four representative topic keywords having the most frequently appearing topic at the center and others along the expanding circumference. The most dominant topic is colored in white, the second dominant topic in red, the third in green, and the fourth in yellow. The font size of a keyword represents its frequency within the tile.

the tweets of a given region and time point via their unique topics that are not found in the spatio-temporally neighboring tiles (R2) while suppressing spatio-temporally common topics (R2). The topics computed in this manner do not only remove a significant amount of noise and everyday language that distracts users from detecting meaningful topics, but also give us important clues for anomalous events.

As shown in Fig. 2, STExNMF works as follows. We refer to the spatially neighboring tiles as those spatially adjacent to a given tile and the temporally neighboring ones as those from k previous days (e.g., $k = 7$ in this paper) in the identical location. 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 iteratively discarding the explainable part of documents in T using the topics computed from the neighboring tiles of T.

Topical word cloud. We display topics under the assumption that there are more anomalous events in tiles where there are more tweets. The number of topics computed and displayed are determined by the number of the documents contained in a tile, e.g., two topics for less than 150 tweets, three for more than 150 and less than 400, four for more than 400. In this manner, users can get an idea about the number of tweets

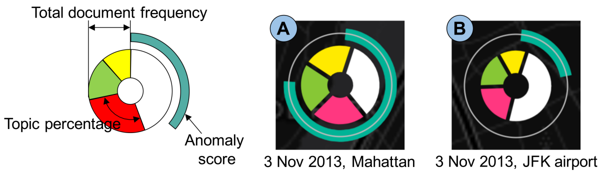

Figure 4. Glyph design for tile-wise anomaly and topic analysis. (A) The example in the center shows that the exclusiveness score of the Manhattan area on November 3, 2013 is high while the distribution of topics is balanced. This means that the topic contents of this area are highly exclusive with respect to its neighboring tiles, both spatially and temporally. It implies that many people gather and talk about diverse subjects. On the other hand, (B) JFK airport shows a low exclusiveness score, and the score of the first topic is relatively high. This indicates that when visiting airports, people talk about common topics, and we conjecture that they usually send similar messages to say hello and/or goodbye at the airport. Our further examination shows that the topic keywords, 'jfk,' 'john,' 'kennedy,' and 'airport,' are frequently found around the JFK airport area.

in a tile by examining the topics displayed. In addition, if multiple topics contain the same top keywords within a single tile, we display only one of them to avoid redundancy and visual clutter.

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

Discovering and Examining Anomalous Tiles

In addition to a topical word cloud, TopicOnTiles offers three novel visual encodings per tile for detecting anomalous tiles and their event contents: (1) highlighting anomalous tiles using colored thick edges, (2) encoding the topic proportions and the anomaly score of a tile as a pie chart-like glyph, and (3) encoding the overall, 24-hour, and 7-day frequency patterns of a user-selected keyword as donut chart-like glyph and vertical grids.

Highlighting anomalous tiles via thick edges. We guide users through the anomalous tiles by using a measure called the anomaly score of a tile (R2). Using the topic keywords of a tile computed by STExNMF, 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 nor in the spatially neighboring tiles, divided by the number of topics in the tile. The visual emphasis of anomalous tiles works by putting red-colored edges surrounding the anomalous tile, as shown in Fig. 1(C). The thickness and the brightness of the edge encode the anomaly score, so that as the brighter and thicker an edge becomes, the more anomalous contents the tile contains.

Visualizing topic proportions and anomaly score of tiles. Our design of this glyph, shown in the upper-left part of a tile, is composed of two parts: the inner and the outer layers, as shown in Fig. 4. The size of the circle in the inner layer represents the total tweet count within a given tile. The pie chart represents the distribution of different topic proportions in terms of their associated tweet counts (R3). Colors in the circle are in accordance with the topic colors displayed in the topical word cloud. Finally, the arc length of the outer layer of the glyph encodes the anomaly score of a tile, as described above. While the edge thickness of a tile plays a role of

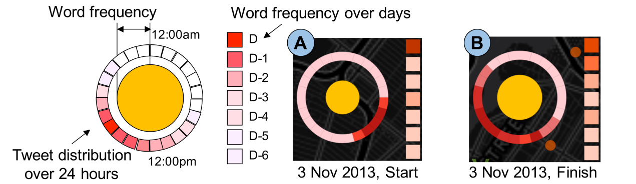

Figure 5. Glyph visualization of temporal frequency patterns of a user-selected keyword. (A) and (B) show the glyphs of the words 'Start' and 'Finish' on November 3, 2013, respectively. If today's vertical grid is the darkest, then we can infer that today is the day of the marathon. The tweet distribution containing the word 'start' is high in the morning, and the tweet distribution containing 'finish' is also high in the afternoon. The user can get an idea that it started in the morning and ended in the afternoon.

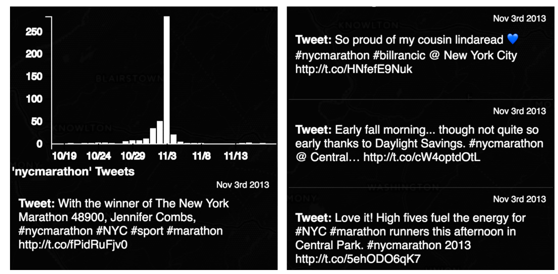

Figure 6. Temporal pattern visualization of raw tweets. The bar graph shows the frequency of those tweets containing the user-selected keyword and the panel below the bar graph shows the content of these raw tweets.

quickly leading users to highly anomalous regions, such a glyph allows users to compare the anomaly score among tiles more accurately.

Visualizing temporal frequency patterns of a user-selected keyword. Once a user clicks on a particular keyword, another glyph visualization gets shown in the upper-right part of a tile. This visualization is in two forms: a donut chart-like glyph and vertical grids, as shown in Fig. 5. The former has two sub-parts; the inner and the outer layers. The inner layer represents the total frequency of the keyword within the tile while the outer one shows the hourly distribution of tweets containing the user-selected keyword over 24 hours. As the keyword frequency gets higher, the color becomes darker. The outer ring can reveal the particular time when the event occurs.

The vertical grids next to the donut chart-like glyph represent the distribution of tweet counts containing the user-selected keyword over the last seven days, from a user-chosen day. The darker color represents a higher frequency of tweets. Using this vertical grid visualization, users can identify useful patterns, such as a sudden increase of a keyword frequency in recent days or a uniformly distributed frequency patterns over past days.

Visualizing actual tweet-posting geo-locations. When a user clicks on a keyword, the system highlights the geo-locations of individual tweets containing the keyword, as bright yellow dots (R3). Using this visualization, users can identify the region with dense bright dots as the location where the keyword has been heavily used. Fig. 1 shows the spatial distribution of tweets of the keyword 'christmas' in New York City on December 25.

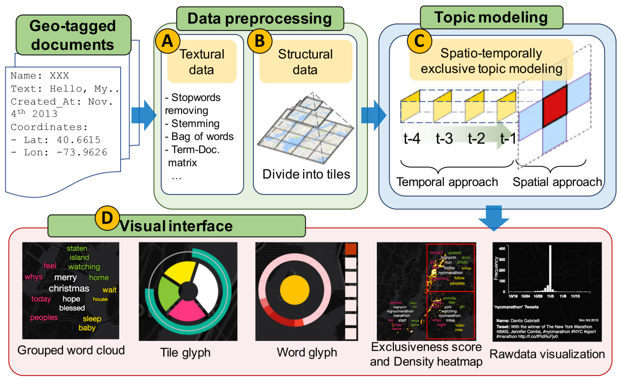

Figure 7. The overall architecture of TopicOnTiles. TopicOnTiles is composed of three parts: data preprocessing, topic modeling, and interactive visualization interfaces. (A) In the data preprocessing step, we collect geo-tags, time stamps, and raw text of each tweets. The raw text is preprocessed via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

Analyzing raw tweets of interest.

As shown in Fig. 6, TopicOnTiles provide an additional visualization that enables users to check the details of raw tweets associated with a particular keyword (R3). When clicking a keyword, a dashboard containing a bar graph and a panel containing the raw tweets pop up. The bar graph shows the daily frequency of the tweets containing the user-selected keyword For a month's period. The panel below the bar graph shows the content of these raw tweets so that the user can further understand the event details.

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 tweet text. The raw text is preprocessed via bag-of-words vector encoding with stopword removal and stemming (Fig. 7(A)). Next, we split the entire set of bag-of-words vectors with respect to different dates and geo-locations for splitting them into grids (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 pre-compute some processes which takes long time to process. A pre-computing 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/

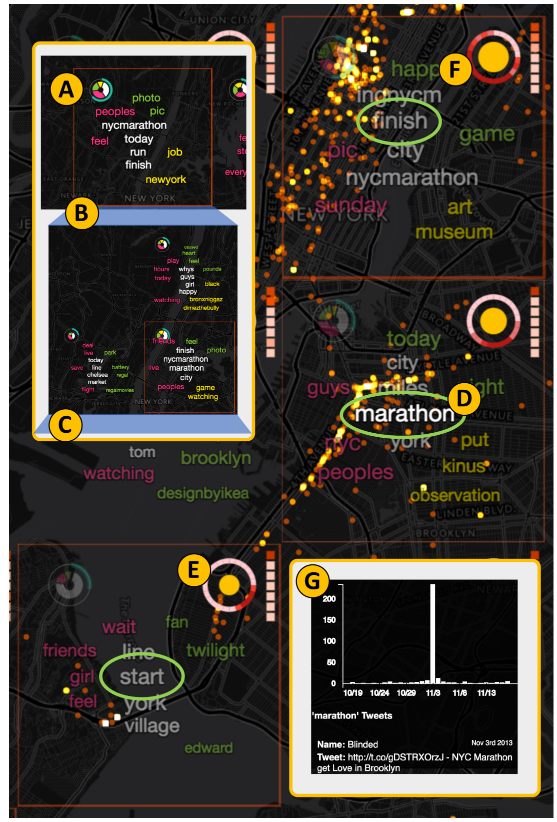

Figure 8. Snapshot of TopicOnTiles showing the tweets posted on November 3, 2013. (A) Thick edges of the tile containing the topic keywords related to marathons indicates an anomalous event in the corresponding tile. (B) Zooming in the region shows more details. (C) The further zooming in reveals such 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. First, he focuses on the tile with thick edges, where ‘nycmarathon,’ ‘today,’ ‘run,’ and ‘finish’ are the main topic keywords, as seen in Fig. 8(A).

Next, as shown in Fig. 8(B), she zooms in the map to discover topics from the smaller tiles. She finds the tiles with keywords ‘nycmarathon’ covered with thick edges.

She zooms in again (Fig. 8(C)). She focuses on the tile having the thick, red-colored edges (Fig. 8(D)). On that tile, keywords such as ‘marathon,’ ‘city,’ and ‘ingnycm’ are revealed as the main topic keywords. She is curious about this event, and it is related to marathon. Therefore, she clicks on the keyword ‘marathon’ for more information.

Shortly after, the bright yellow dots and the glyph at the upper-right part of the tile are shown on the map. The yellow dots show that many people tweeted using the keyword ‘marathon.’ The dots are bright in areas near the tile having ‘marathon’ as the main topic keyword, and the area near the Central Park. They make up a long line of bright dots along these tiles. She assumes that this bright line is the marathon course. By looking at the vertical grids of the tiles, she finds that ‘marathon’ started to heavily appear on the selected day.

She selects this keyword and looks at the outer ring of the glyph to analyze the hourly patterns of the event. She discovers that in the tile having ‘start’ as its main keyword, ‘marathon’ appears mostly in the morning (Fig. 8(E)). In the tile containing ‘marathon,’ it mostly appears around noon. Near the Central Park (Fig. 8(F)), it appears throughout the day.

The yellow circle of the glyph indicates that ‘marathon’ appears in many tiles, appearing more frequently near the Central Park.

Judging from all these findings, she is quite sure that NYC Marathon is taking place in New York City. She now wants to check the raw tweets for further investigation. Thus, she examines the pop-up window (Fig. 8(G)). After checking these tweets, she identifies that a marathon competition has taken place and finds out various reactions of the event from the Twitter users.

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 obtain the interesting information such as the starting and the ending lines of the course and the entire marathon course.

Trayvon Martin Protest and MLB All-Star Futures Game

Suppose a user wants to find the anomalous events of July 14 in New York City, the day when the Trayvon Martin Protest and the 2013 Major League Baseball (MLB) All-Star Futures Game took place.

[2]https://www.mongodb.com/

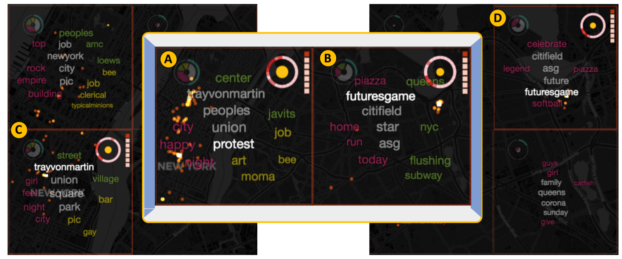

Figure 9. Snapshot of TopicOnTiles showing the tweets posted on July 14, 2013. (A), (B) The edges of the tile containing the topic with keywords related to Trayvon Martin Protest and MLB All-Star Futures Game were highlighted. Zooming in the tile of Trayvon Martin Protest to analyze detailed contents, (C) keywords such as ‘trayvonmartin,’ ‘union,’ ‘square,’ and ‘park’ are revealed. The vertical grids when selecting each of these keywords indicate that they started to appear heavily starting from Jul. 14. The temporal glyph shows that the event has mainly taken place during the night. Zooming in the tile of MLB All-Star Futures Game for further analysis, (D) keywords such as ‘citifield,’ ‘asg,’ ‘future,’ and ‘futuresgame’ are revealed, which clearly shows another event, All-Star Futures Game.

Trayvon Martin Protest. Initially, our system indicates the anomalous tiles by putting a thick rectangular edge on the tiles 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(A) and Fig. 9(B). In the tile covering the Central Park, we can find keywords such as ‘trayvonmartin,’ ‘union,’ and ‘protest,’ and near the Citifield, there are keywords such as ‘futuresgame,’ ‘citifield,’ and ‘asg.’ The user first decides to delve into the tile covered by thick edges, which contains a keyword ‘trayvonmartin.’

She zooms in the map and sees tiles as shown in Fig. 9(C). She focuses on the tile containing the keyword ‘trayvonmartin.’ She clicks on the keyword ‘trayvonmartin’ to check other visual interfaces. She looks at the upper-right part of the glyph. The vertical grids indicate that the keyword emerged extensively at the chosen day. The outer ring shows that the keyword is mainly used at night, signalling that the event took place during the night. The yellow dots disclose a specific point in white, a potential location of the anomalous event. As there are keywords ‘union,’ and ‘square’ in the same topic, she assumes that the location is associated with these keywords.

Using all these tools, she senses that an event presumably related to ‘trayvonmartin’ has taken place, and the location has to do with ‘union’ and ‘square.’ For verification, she checks the pop-up window. She identifies that a massive protest occurred near the Union Square and also obtains detailed information on the event.

MLB All-Star Futures Game. As she is also a huge baseball fan, she looks into another tile covered by a thicker edge, the one having keyword ‘futuresgame’ and ‘citifield.’ After zooming in the map, she clicks on the keyword ‘futuresgame,’ as shown in Fig. 9(D). A glyph visualization appears, and the vertical grids manifest that the keyword started to appear on the chosen day. The outer ring denotes that the keyword is mainly used during the afternoon. The yellow dots pinpoint the exact location of the place, presumably Citifield Stadium. Looking at keywords such as ‘futuresgame’ and ‘asg(All-Star Game),’ she assumes that All-Star Futures Game took place on this day. After looking at the raw tweets from the pop-up window, she recognizes that the game had taken place on that day in Citifield Stadium.

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-subjects 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 given a different time period of data from November 29 to December 2 in 2013.

Once this task is completed, we exposed all the participant to TopicOnTiles with its full capabilities, let them play with it for about five minutes, and asked them 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 being either 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, we computed the total number of anomalous events a user found and divided it by the amount of time s/he spent, which represents the number of anomalous events found per unit time. We then averaged this measure across participants within each of the three different settings.

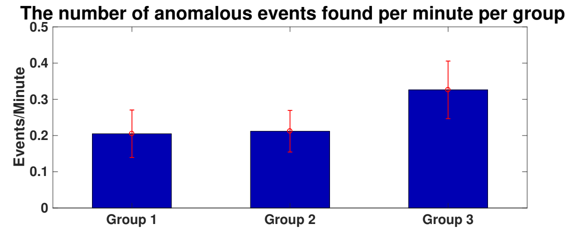

Figure 10. Comparison of the number of anomalous events detected per minute. Group 3, which corresponds to TopicOnTiles with its full capabilities, is shown to be significantly different from each of Group 1 and Group 2, while no statistically significant difference is found between Group 1 and Group 2.

Analysis Results

Comparative analysis. Fig. 10 shows the comparison of the performances among the three settings tested in our study. On average, Group 3, TopicOnTiles with its full capabilities, shows the best performance while Group 1, which has no additional visual components than topic keywords nor STExNMF, performs worst.

A between-subjects one-way analysis of variance (ANOVA) [19] was conducted between the three groups, in order to evaluate the effectiveness of visualization components in finding the anomalous events. At the significant level α as 0.05, the ANOVA test showed that at least one of the groups were significantly different from the others, with $p = 0.01$, $F(2,57) = 4.972$.

To identify the difference in the means of the different groups, we conducted a post-hoc analysis using Bonferroni test [20]. The results indicated that the mean score for Group 3 ($\mu = .3258$, $\sigma = .1590$) was significantly different from each of Group 1 ($\mu = .2048$, $\sigma = .1309$) and Group 2 ($\mu = .2119$, $\sigma = .1148$), with $p = 0.02$ and $p = 0.03$, respectively. This result indicates that our proposed visual interfaces in TopicOnTiles play a crucial role in detecting anomalous events.

Additionally, Fig. 10 shows that Group 2 performs slightly better than Group 1 although the result from Bonferroni test shows no statistically significant differences between them. Nevertheless, some participants were able to distinguish the differences of the two settings. For example, one participants said *that '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 diverse topics because it removes the prevalent keywords common in the neighboring tiles.'*

Analysis on CSUQ answers. Fig. 6.2 shows the survey results of CSUQ answers collected from the participants. The results show that the system helps to accomplish the task efficiently (Q6, Q7) and that the interface is useful and pleasant (Q11, Q12).

We conjecture that the high scores in the former part is mainly owing to the novel topic modeling technique of STExNMF and the novel visual interfaces that supports both the spatial and the temporal analyses of the topics in detecting anomalous

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

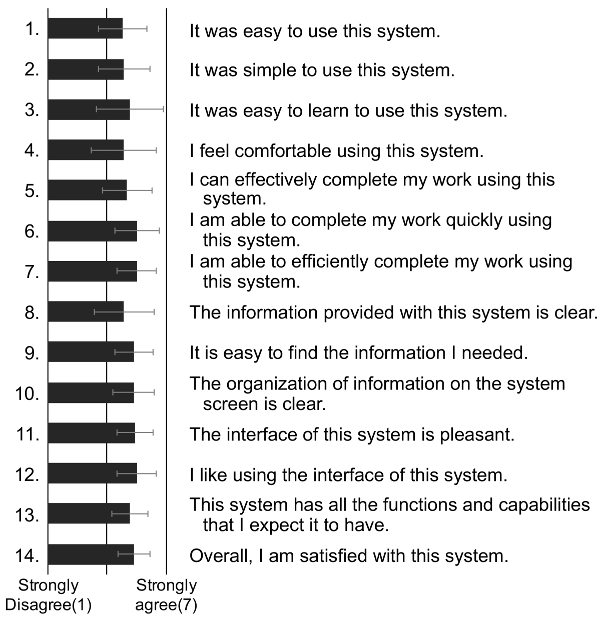

Figure 11. Summary of the collected CSUQ answers. TopicOnTiles is rated high in many questions, especially in terms of its efficiency and its interface.

events. One participant said that *'allowing a facilitated spatial and temporal analysis of the events using visual components helped me gain various information on events, and personally, such a visual display has even aroused my curiosity of wanting to know more.'*

At the same time, the relatively low score on the question 'the information provided with this system is clear' (Q8) implies that our visual components may have room to improve from the perspective of non-expert users. 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 many analytical tools displayed, sometimes I get a little hesitant in deciding which functionalities of the 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 anomalous topics

There has been much research that attempted to automatically determine the optimal number of topics or clusters in topic modeling. A hierarchical topic modeling [21] utilizes the information gain as a measure 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 a space, it may not be able to visualize all the computed topics based on an optimal topic number. Another interactive topic modeling system called UTOPIAN [13] merges or splits the topics with the help of user interactions. However, in systems such as ours that deal with large-scale data, if a user has to perform manual interactions to adjust the number of topics for individual tiles, it would be burdensome from the perspective of human efforts and dynamic topic modeling computation.

In coping with such limitations, we assume 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 an assumption has flaws in several aspects. It is not capable of distinguishing the everyday language topics from meaningful topics. The topics displayed on a tile may not detect all of the meaningful topics that contain information on anomalous events. Our system could benefit from an advanced algorithm that can determine the number of topics to be extracted in anomaly detection.

Events distributed around tile boundaries

In a tile-based map, the regions are subdivided into equal-sized regular grids without carefully determining tile boundaries. This can cause problems when an anomalous event takes place near the tile boundaries. That is, the tweets mentioning the same event may be split into multiple adjacent tiles, causing a possible blind spot in our system. We tried to address these issues by providing users with a topic summary from different zoom levels. However, this method is not without limitations. If a user zooms in the map to delve into an event found from a lower zoom level but does not find the relevant topic in a higher zoom level, then it may not provide the information the user would want. A sophisticated approach would be helpful so that the system can subdivide the regions considering the administrative or natural boundaries.

Facilitating user interfaces for novice users

Regarding the usability question 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 participants told that the temporal pattern analysis of a user-selected keyword using glyph visualizations helped them notice the anomalous events.

However, we received negative comments from some participants, saying that the system may take some time to learn how to use for those having little knowledge on the related domains. One participant said, *'there exist too many visual components, and sometimes it is confusing which one I should be looking at.'* This goes in accord with the relatively low score of Q8 in the CSUQ answers. In developing TopicOnTiles, how to add full capabilities into our system without losing simplicity to use our system has always been a big issue. Considering potential casual users for our system, developing a new visual design and brief tutorials would greatly improve the usability of TopicOnTiles.

CONCLUSIONS AND FUTURE WORK

This paper presented a tile-based visual analytics system TopicOnTiles for anomalous event detection using geo-tagged social media data. TopicOnTiles is mainly built upon a tile-based map interface using the novel topic modeling technique that extracts spatio-temporally exclusive topics with respect to the neighboring tiles. TopicOnTiles also provides various visual encodings such as the glyphs, vertical grids, and heatmaps to facilitate anomalous event detection tasks. We showed usage scenarios using Twitter data from New York City, where our system effectively reveals the event of the *2013 ING NYC Marathon* and the Trayvon Martin protest. Furthermore, we conducted user studies showing the efficiency and the usability of TopicOnTiles compared to other baseline settings.

As our future work, we plan to extend our visual system to a real-time monitoring system that can solve various problems arising in urban areas such as natural disasters, crimes, and so on, by associating the geo-tagged textual social media data with other types of geo-tagged data, such as mobile user data.

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 *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.