Eurographic Conference on Visualization (EuroVis) 2012
S. Bruckner, S. Miksch, and H. Pfister
(Guest Editors)

Volume 31 (2012), Number 3

I-SI: Scalable Architecture of Analyzing Latent Topical-Level Information From Social Media Data

Xiaoyu Wang, Wenwen Dou, Zhiqiang Ma, Jeremy Villalobos, Yang Chen, Thomas Kraft, and William Ribarsky

University of North Carolina at Charlotte

Abstract

We present a general visual analytics architecture that is constructed and implemented to effectively analyze unstructured social media data on a large scale. Pipelined based on a high-performance cluster configuration, MPI processing, and interactive visual analytics interfaces, our architecture, I-SI, closely integrates data-driven analytical methods and user-centered visual analytics. It creates a coherent analysis environment for identifying event structures, geographical distributions, and key indicators of emerging events. This environment can support monitoring, analyzing, and responding to latent information extracted from social media. We have applied the I-SI architecture to collect social media data, analyze the data on a large scale and uncover the latent social phenomena. To demonstrate the efficacy and applicability of I-SI, we describe several social media use cases in multiple domains that were evaluated by experts. The use cases demonstrate that I-SI can benefit a range of users by constructing meaningful event structures and identifying precursors to critical events within a rich, evolving set of topics.

Categories and Subject Descriptors (according to ACM CCS): D.2.11 [Software Architectures]: Domain-specific architectures

1. Motivation, Analysis Goal, and Challenges

We are moving toward a ubiquitous social era, in which mobile communications, social technologies and sensor-based services connect people, the Internet and the society into one immensely interconnected community. With the rapid growth of such ubiquitous communication infrastructures, we are living in a world where nearly everyone is connected in real time. Our society as a whole is being greatly influenced by such intimate connections, affecting every aspect of people's social behaviors. Moreover, the evolution towards such interconnected, real-time social discourse has changed the way people ***organize and respond to*** social events (e.g., happenings, protests or campaigns), enabling people to form, share, discuss, and react to social activities instantaneously.

As a result of all this personalized, digital communication, massive amount of data, including both textual and multi-media data, are collected in real-time regarding who we are, where we are, and what we are talking about. Particularly, the emergence of microblogging has yielded an overwhelming amount of such data, ranging from status updates on Twitter and Facebook, to extended comments on Google+, all often accompanied by images and more and more by video. As one example of the explosive growth, Twitter rose from about 6 million visitors per month in January 2009 to over 37 million per month as of November 2011[Twi]. Based on multiple estimates, on an average day, users globally submit 140 million "tweets" on Twitter; and for each month, users share about 30 billion pieces of content on Facebook.

These massive, large scale social media datasets bear extremely rich information that, if visualized, can lead to a profound impact on depicting patterns for emerging social events (and their underlying topics). This can contribute in new, important ways to the understanding of social phenomena. The need to assess the related social phenomena in a systematic way has increased for both ***citizens*** and ***government*** (e.g. emergency responders and law enforcement). Analyzing this rich social media data gives them the ability to understand and even predict people's interests, and to further depict the shifts and turns of social activity at the individual, group and global level. For example, analysis of social media could give government valuable information on how to effectively mitigate problematic situations (natural disasters or chaotic scenes). Citizens or citizen groups could learn about the development, history, and spread of ideas of social movements (e.g., Occupy Wall Street).

1.1 Analysis Goal: Depicting Social Phenomena and Their Event Structure

The key point of our research resides in the core analytical architecture that focuses on combining data-driven analytics methods with human-centered analytical approaches through the use of an interactive visual analytics interface.

Email to: xwang25 @uncc.edu

© 2012 The Author(s)
Computer Graphics Forum © 2012 The Eurographics Association and Blackwell Publishing Ltd. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

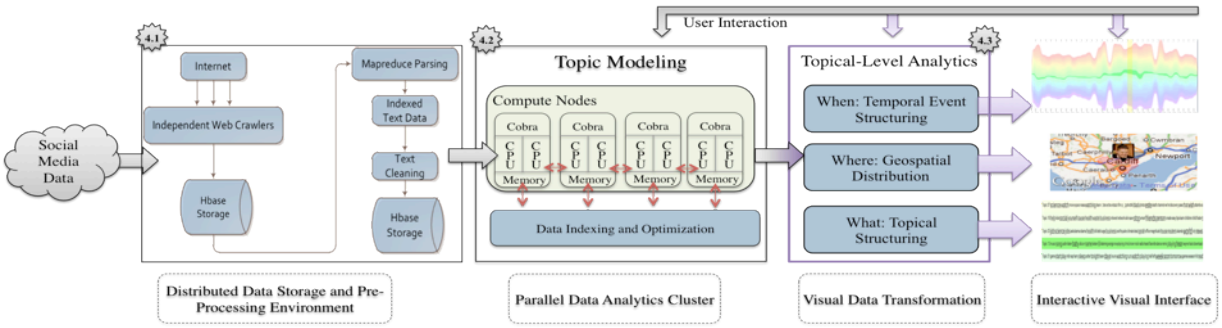

Figure 4: *An overview of I-SI architecture. There are 4 major components in the I-SI architecture: Distributed Data Storage and Pre-Processing (Section 4.1), Parallel Data Analytics Cluster (Section 4.2) ; Visual Data Transformation; Interactive Visual Interface (Sec. 4.3).*

The process of collecting such data (microblogs), analyzing it on a large scale with data-driven technologies, and creating human-centered visual analytical environments, can be generalized across different domains.

The analytics architecture promotes the interpretability of social phenomena, through presenting a platform for conducting larger scale topic-level analysis. Scalability is quite important because these ever-changing, often streaming media, require constant observation and interaction with very large, dynamic data. A detailed description of the components of the architecture and its related analysis processes, along with the usable software tools, are the focus of this paper.

1.2 Major Challenges and Opportunities

Despite continual efforts, analyzing such large-scale, loosely structured, and less-contextual social media data to support analytical reasoning remains extremely challenging. There is a scarcity of methods to extract the latent semantic information in the massive text corpora. Specifically, the challenges are two-fold:

- **Motivating Challenges I: Depict Latent Social Activities**
  At an individual level, as streams of diverse information constantly bombard users, it is difficult for them to keep up with, let alone harvest important and interesting messages. In addition, a user might want to identify useful content outside of her selected focus, or to discover trendy topics that other people have been discussing on social media. This task involves not only a meaningful summary of vast information streams, but also the support for interactive exploration of content according to individual interests.
  At an organizational level, analyzing social media data streams allows institutions to grasp up-to-date topical trends and identify critical events that may require appropriate action. For instance, a commercial organization might be interested in reviewing consumer responses to products or the company's general image. Analyzing relevant information from social media may be a better way to gather honest opinions from possible customers than conducting targeted surveys on a sample population. Similarly, campaign strategists might be interested in knowing people's general opinions towards different parties and politicians. Social media is a perfect place for collecting such data since large groups of users voice a rich set of attitudes over time and respond to events through Facebook or Twitter.
  In addition to the above examples, valuable information extracted from the noisy social media data could inform other entities such as emergency responders and police departments about future events that are being organized or current events as they unfold.
- **Motivating Challenges II: Establish Meaningful Social Event Structures**
  Nowadays people do not need to be powerful to launch a successful media campaign thanks to social media. A 27-year old art gallery owner started a national movement "Bank Transfer Day" against big banks through one Facebook post [FAC]. The movement led more than 1 million customers (estimated) to transfer their cash out of big banks to credit unions. An on-going national campaign–Occupy Wall Street–used social media abundantly to spread nationwide [CBS].
  The scale of such movements amazes people, but little is known regarding how they were initiated and organized. Through analyzing social media information related to the occupy movement, for example, one can construct a timeline or even an event structure to investigate the progression of the movement, and answer questions such as who were the initial organizers, who joined the campaign at what time, when exactly did the movement start, what ideas and issues developed, and which events might have led to this massive national campaign.

2. Introducing *I-SI* Architecture

In response to these challenges, we have developed a visual analytics architecture to support topical-level investigative analysis of social media data. Our architecture, *I-SI*, is centered on the combination of data-driven topic modeling approaches with human-centered visual analytics techniques; topic modeling is enhanced by interactive visual interfaces, providing results that can be explored, filtered, and managed by users. *I-SI* creates a coherent analysis environment for identifying event structures, geographical distributions, and key indicators of emerging events. In other words, *I-SI* can help analysts identify and follow social phenomena as they emerge, evolve, and mature.

On the high-level, as shown in Figure 1, the data analytics capability of *I-SI* comes from leveraging High-Performance clusters to apply automated topic modeling to social media data such as large collections of tweets. Extended on our previous work, ParallelTopics [DWCR11],

© 2012 The Author(s)
© 2012 The Eurographics Association and Blackwell Publishing Ltd.

the visualization component of *I-SI* creates an investigative visual analytics environment [WMS*08] that not only provide a summary of "what happened" in terms of meaningful topics, but also allow inferences of causal relationships between a critical event and the effect of the event. In particular, our architecture could represent the progression of social events through underlying latent common themes over time, which allows users to discover the overall trend as well as rise and fall of individual social activities.

We have currently applied the *I-SI* architecture to collections of unstructured social media data (e.g., Twitter data), analyzing the data on a large scale and uncovering the latent social phenomena (who-when-where-what-why). The results bring forth meaningful semantic information that is otherwise hidden in the large aggregation of noisy tweets. The results contain topics summarized based on the tweets; dynamic patterns of topics, and emerging events.

We have reached out to multiple user communities, as detailed in Section 6.1. During this outreach, *I-SI* has been demonstrated to and evaluated by multiple users, including political campaign strategists and law enforcement experts. Feedback from these experts suggested that our *I-SI* architecture contributes to the social media analysis in the follows aspects:

- Analyzing social media data on event/topical level instead of keyword level. The cohesive themes from the otherwise noisy social media data are nicely summarized and presented to users. The purpose is to analyze and ultimately predict social activities/behaviors.
- Ability to handle large amounts of data. Studies have seldom focused on analyzing social media data on a large scale. We have utilized parallel computing methods to handle billions of microblog messages at once.
- Straightforward identification of critical events and even precursors to the events via temporal and geospatial visualization. In addition, through providing interactive exploration capabilities, the *I-SI* architecture enables users to perform investigative analysis regarding certain events/topics and answer who-what-where-when-why questions.

3. Related Work

There is a wide range of research on social media analysis, especially on Twitter data because of the public nature of tweets.

3.1 Analysis of space and time in Social Media

A large portion of the published research about Twitter has focused on questions related to Twitter's spatial and temporal properties with little or no semantic analysis on the textual content of tweets. For example, Java et al. [JSFT07] studied the topological and geographical properties of Twitter through constructing a social network based on users and their "friendship" information without considering content of tweets. More recently, MacEachren et al. [MJR*11] has developed SensePlace2 - a geovisual analytics system that supports situational awareness for crisis events using Twitter data. SensePlace2 focused on extracting explicit and implicit geographic information for tweets, and combining geospatial with temporal information to promote understanding of situations evolving in space and time.

3.2 Topical Analysis of Textual Content in Social Media

Other work has presented analysis of the textual content of social media data.

3.2.1 Identifying Relevant Content

There has been numerous research on recommending, filtering and searching social media content [BSH*10, CNN*10, DCCC11]. Beinstein et al. [BSH*10] proposed a Twitter application called "Eddi" that organizes a user's *own* feed into coherent clustered topics for more direction exploration. Chen et al. [CNN*10] explored content sources, topic interest models, and social voting as three separate dimensions for designing a recommender of social media content. To identify the most relevant content in social media, Choudhury et al. leveraged information diversity and user cognition [DCCC11]. Our work differs from these research in that not only we allow users to identify tweets based on topics, we also provide an overview that doesn't limit a user to either search based on contents in her own Twitter stream or having to know what to look for in the first place.

3.2.2 The Use of Topic Models

Blei et al. introduce latent Dirichlet Allocation (LDA) in 2003 [BNJ03]. The aim of LDA is to discover the hidden thematic structure in large archives of documents. Lots of content analysis was performed using LDA or extended versions of LDA. For example, Ramage et al. maps the content of the Twitter feed into dimensions using Labeled LDA [RDL10], with the four dimensions corresponding roughly to substance, style, status, and social characteristics of posts (4S). One limitation of the work comes from the authors manually assigning the predetermined labels (4S) to learned topics without leaving room for users to explore and attach other meanings to the topics. Ritter et al. have applied LDA and other unsupervised approaches for the purpose of modeling conversations within Twitter streams, as the sequential dialogue reflects the shape of communication in the online platform [RCD10]. More recently, Sizov proposed a framework, GeoFolk [Siz10], which combines textual content with spatial knowledge (e.g. geotags) to construct better algorithms for content management, retrieval, and sharing.

3.3 Visual Analysis of Social Media Data

Most of the aforementioned [JSFT07, RDL10, Siz10] work only focuses on data-driven techniques with a limited scale. Therefore, their objective and approach differs from our approach of combining both topic-level analysis and human-centered visual analytics methods.–In the realm of interactive visualization, aside from SenseSpace2 [MJR*11], researchers have presented systems to track on-

© 2012 The Author(s)
© 2012 The Eurographics Association and Blackwell Publishing Ltd.

going social events and to support the use of social media as supplemental information sources for journalists [DGWC10,DNKS10].

Dörk et al. introduced Topic Streams, a web-based interactive visualization system to follow and explore conversations on Twitter about large-scale events [DGWC10]. The authors also presented several design goals such as summarizing the conversation, providing flexible time windows, etc, which are quite informative for future design. In addition, Diakopoulos et al. [DNKS10] presented a visual analytics tool, Vox Civitas, to help journalists extract news value from social media content around broadcast events. The visualization component of the *I-SI* architecture differs from Topic Streams and Vox Civitas in defining topics. As opposed to a keyword-based approach, we extract topics using LDA to pick out stronger and more cohesive themes from the entire social media corpus.

4. *I-SI*: Scalable Architecture for Topical Analysis of Social Media Data

In this section, we present our architecture and its implementation. Pipelined based on Hadoop servers, a high-performance cluster configuration, MPI processing, and visual analytics interfaces, our architecture closely integrates data-driven analytical methods and user-centered visual analytics. It creates a coherent analysis environment for identifying event structures, geographical distributions, and key indicators of emerging events. The core components of our architecture include Data Collection, Data Cleaning, Topic Modeling, and finally Interactive Visual Analytics Interfaces. As shown in the overview pipeline (Figure 1), the benefit of our componentized modules is that the structure can incorporate more efficient and advanced analysis components to enrich the analytic capability of the architecture.

4.1 Distributed Data Storage and Pre-Processing Environment

As shown in Figure 1, our architecture aims at incorporating multiple sources of social media data such as Twitter updates, editorial news, and blogs. The heterogeneous and streaming nature of these data sources poses a significant challenge in data management schema and data cleaning.

Given the scale of data that our architecture focuses on, standard SQL data management schema are not optimized to handle the I/O of different kinds of social media data. Specifically, considering the intrinsically fragmented and loosely structured nature of tweets, our architecture requires a powerful distributed database processing approach to achieve sufficient data access and effective data processing. After experimenting with several nuance NoSQL structures (e.g. Cassandra [CAS], MongoDB [MDB11]), we adopted the MapReduce framework [DG08], and configured Hadoop [APA] to provide an efficient and scalable distributed computing and data storage platform. HBase, an open-source realization built on the Hadoop Distributed File System (HDFS), is used in this platform to store the collected social media data.

Besides providing a stable data management platform, we also utilized Hadoop to create a robust parallel data crawling and cleaning process. As shown in Figure 1, such process is interfaced with the Internet through multiple independent crawlers. Each of the crawlers constantly collects social media data from various public domains and dumps it into HBase. Specifically, we have created crawlers to tap into Twitter's public API to collect tweets. It acquires such information on the "Garden-hose" level, which constantly delivers 10% of tweets with a statistically significant sample of all contents [Twi]. As a result, we were able to collect over 3 Billion tweets from all languages over the course of 11 weeks, providing us a reliable database for our evaluation and outreach purposes.

Concurrently, such textual data is being cleaned, parsed, and prepared for topic modeling through multiple Mapreduce jobs that perform these analytics tasks in the background. During these tasks, noise symbols and stopwords are removed. Basic statistical analysis, such as word count, is also performed over the data, preparing for the topic modeling procedure. The implementation of both the data cleaning and basic statistical tasks contain two stages (i.e. map and reduce stages), detailed in the Algorithm 1. In the map stage, data is distributed into working nodes for intermediate computation; the output of the map stage follows the <key, value> pair format. After merging all the values with identical keys into an array, the merged intermediate results are collected for further computation in reduce(s). The final output of the reduce(s) has the same <key, value> pair format as map's. Extracted data is then stored into HDFS data repository and is distributed across multiple nodes within our Hadoop cluster to guarantee reliability.

Input: Text data from HBase storage
Output: Cleaned data, frequencies of each distinct word in every document with its associated document ID
Map stage:
    Create stop words hash table;
    Read input from HBase;
    Tokenize each line;
    Remove stop words;
    Output <docID+word, 1>;
Reduce stage:
    Input <docID+word, [1, 1, …]>;
    Count the number of elements in the value of input value array [1,1,…] ;
    Output <docID, [word1+freq, word2+freq …]>;

Algorithm 1. Map-Reduce Process for Data Cleaning.

4.2 Parallel Topic Modeling using High-Performance Computing Cluster

In order to have a comprehensive understanding about the latent social media data, one needs to extract and correlate information from massive amounts of data. With new tweets reaching a billion every five days, performing such analysis is beyond the scope of computing power of any single-node configuration, either it would be impossible to process the data (i.e. memory issues) or it would take too long to obtain analysis results. This suggests yet another

© 2012 The Author(s)
© 2012 The Eurographics Association and Blackwell Publishing Ltd.

significant scalability challenge in social media analysis.

To alleviate such scalability issues, our architecture incorporates the use of a high-performance cluster to strengthen the analytical capability over the social media data. In particular, once the data has been cleaned and stored in HDFS, it is then ready to be processed by parallel computing clusters for topical-level analysis. Such a process has its most notable performance bottleneck at the learning and inference stages [NEW06]. In order to reduce the time to complete this stage, we extend on Google's PLDA MPI implementation [PLD]. The algorithm is a general implementation of LDA with parallelization implemented into key portions of the algorithm. Such a process utilizes Gibbs Sampling, a Monte Carlo approach, to compute the result towards a conversion point as the number of iterations increase.

During this process, we use Portable Batch System (PBS) to schedule the jobs and the Message Passing Interface (MPI) is used to make parallel use of the cluster nodes. As shown in Figure 1, each node of our cluster has 12 cores and a total of 36GB of memory, with a fast Gigabit Ethernet to communicate results. Our cluster converts the input data (e.g. tweets) into output data (topic-based probabilistic information), using LDA to create a probabilistic model that uses the documents, the words, and their utterances to build the topic model. The benefit of using such infrastructure is two-fold:

- We can now process the social media on a scale that a single node computer would not be able to handle. Such infrastructure and its parallelized algorithm granted us to capability to peek into the topics that are embedded in the large unstructured text corpus. For example, we have tested the *I-SI* architecture to investigate topics from 17,651,186 English tweets, roughly around 1.8Gb data over the course of 5 weeks (after data cleaning).
- This setup reduces the processing time for data in the range of 150mb to 300mb, providing our architecture the iterative capability to search most interpretable topic modeling results. This would be important for certain critical response situations such as presented in Scenario I and III (see section 5.1 and 5.3).

4.3 Visual Data Transformation and Interactive Visual Interfaces

To support the analysis needs from different user communities, we designed a coordinated multiple-view interface to create an interactive visual analytics environment. Each view is designed via transforming the output from topic models to showcase one distinct aspect of the underlying social media data.

As shown in Figure 1, the *I-SI* interface is designed to support understanding of spatial and temporal patterns of social activities, identification of event structures, and topical trends through analysis of the growing social media database. A key goal for these interfaces is to permit users to interactively explore, characterize, and compare the space-time aspects associated with topics in tweets. The default interface includes a topic cloud view, temporal and geospatial views, and detailed text view. Tight coupling between these views via interactive techniques permits this interface to effectively visualize highly dynamic and fragmented social media data. The four primary display views are dynamically coordinated. Each view is introduced below and their coordination is further discussed.

Topic Cloud: revealing major topics. We present the topics as a tagcloud for quick overview/summary of the social media corpus. In the topic cloud, each line displays a topic, which consists of multiple keywords. The order of the keywords within a topic indicates their importance to the topic. In addition, since one keyword may appear in multiple topics, the size of each keyword reflects its number of occurrences within all topics.

Temporal View: presenting topic evolution. The temporal view is created as an interactive ThemeRiver [HHN00], with each ribbon representing a topic. The length of the time frame in the ThemeRiver can be changed by users based on their investigative needs. Similar to previously developed topic-based text methods [WLS*10], tweets are divided into corresponding time units based on their time stamps after the time frame has been chosen; then the height of each ribbon is calculated by summing the number of tweets in each time unit.

Geospatial View: displaying geographical distributions. We utilize Google Map [GOO] to provide users with interactive geospatial analysis (see Figure 1). By placing the tweets with geo-tagging (i.e. GPS location associated with tweets) onto the scalable map, detailed geographic relationships and patterns immediately become apparent. In addition, we extended Google Map to effectively display the topical distributions of the social media data. The geospatial view incorporates a client-side clustering algorithm to overlay large amounts of geo-coordinated tweets over the map, creating a density Heatmap [SMKH05] to show the tweet clusters.

View coordination and interactions Since investigative analysis of social activities may involve the utilization of all views, coordination among the views is supported. On the *topic* level, hovering over a ribbon in the temporal view would highlight the corresponding topic in the topic cloud so that users could quickly synthesize information regarding topic content and temporal trend. On the *temporal* level, filtering tweets that were posted within a certain time period is supported. A user could further filter tweets by geospatial area and topics. For instance, clicking on an intersection of a topic ribbon and a time frame in the temporal view would lead to the selection of tweets that are highly related to the topic and posted during the time period. These selections support detailed examination of topical trends and events.

5. Case Study

To demonstrate the efficacy and applications of *I-SI*, we describe three scenarios based on analysis of social media data. In these scenarios, the *I-SI* architecture supported summarizing a large amount of social media information and interactive exploration of the generated topical trends. More specifically, the scenarios demonstrated that the

© 2012 The Author(s)
© 2012 The Eurographics Association and Blackwell Publishing Ltd.

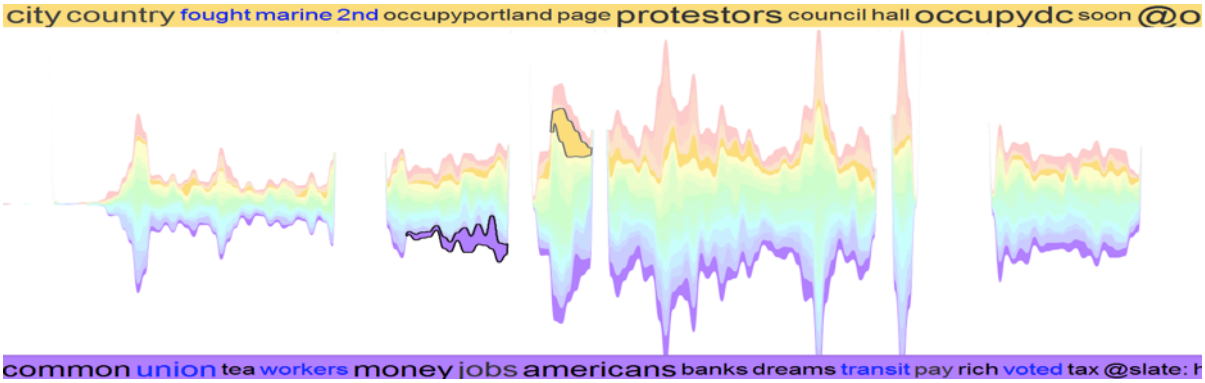

Figure 2: *Difference forces joined the Occupy Wall Street movement. Highlighted portion of the yellow topic -marine joined to movement to protect the protestors from the police. Highlighted portion of the purple topic – union workers voted to support the OWS movement.*

interactive analysis could distill meaningful and otherwise hidden information from noisy social media data, such as revealing critical events and pre-cursors of such events.

5.1 Scenario I: Depict Meaningful Event Structures

The Occupy movement is an ongoing series of demonstrations and is known for using social media to attract more protestors. The Occupy movement is long-lasting and widely spread; people in almost every major city within the U.S. and around the world have joined it and created other related protests such as Occupy Seattle, Occupy London, etc. The ***challenge*** in understanding such a movement lies in distilling the main topics and trends from a movement with massive participation and a wide range of goals such as more and better jobs, more equal distribution of income, bank reform, and a reduction of the influence of corporations on politics [RUS11]. Given the prominent use of social media in organizing the Occupy movement, it should be possible to summarize and analyze how the movement unfolded through analysis of these media.

5.1.1 Data collection and preparation

Since we want to focus on the Occupy movement in this scenario, we further filtered for all tweets with hashtag #occupy from our tweet collection. A hashtag is a Twitter convention used to simplify search and indexing. Users include specially designed terms starting with # into the body of each post. The resulting dataset includes more than 100,000 tweets starting from Aug 19 to Nov 01. Our architecture then automatically removed stopwords and performed topic modeling. Such automated process enables us to experiment with different numbers of topics, and results in the choice of 15 topics for interpretability.

5.1.2 Investigating the Occupy movement

Exploring the unfolding of Occupy movement. Our analysis environment enables users to explore and follow the evolution of the movement. As the user, a campaign strategist, inspected topics in the temporal view, she noticed that this movement had been evolving gradually over the course of two months. This trend had been exemplified by two significant forces joining the Occupy movement. Specifically, as shown in figure 2 (yellow topic), marines joined the Occupy Wall Street (OWS) movement to protect the protestors from the police on Oct 1. The user reached this conclusion by selecting tweets related to the topic of interests within the burst of topic volume for this event (see Figure 2). She noticed that people were shouting out on Twitter about this event: "…the marines coming to protect protestors" and "marine - 2nd time fought for my country time 1st time I've know my enemy".

A similar pattern was also seen for another topic, which suddenly gains momentum as the NYC transit union workers joining OWS (shown in figure 2, purple topic). People sounded excited on Twitter about the event: "200 000 transport workers union votes support!!!", "new york transit workers union voted unanimously support #occupwallstreet. 38000 active march oct. 5". More interestingly, based on reading the last tweet, the user suspected there might be an organized march on Oct 5. Indeed, another big increase in volume of the same topic occurred on Oct 5, and the tweets were related to the march though the Financial District of Wall Street, which was joined by thousands of union workers.

Identifying pre-cursor to the Occupy movement. In addition to identifying meaningful events based on the sudden topical volume change in the ThemeRiver, our analysis environment also enables users to construct a comprehensive story by looking at the overall movement. As shown in figure 3, the overall volume of tweets with "#occupy" became significant around Sep 17, 2011, which is the protestors' self-proclaimed start date of the movement. However, our temporal view clearly indicated relevant tweets were posted well before Sep 17, dating all the way back to Aug 19, 2011 (highlighted region in figure 3).

This unique pattern could suggest a pre-cursor to the Occupy movement, and motivated the user to look further into the details of the tweets. With our coordinated views, she was able to directly click on each time frame to inspect tweets one time step at a time. Upon reading the tweets, the user immediately realized that the OWS was a well-organized event. Specifically, organizers had been using Twitter to advertise the upcoming event and to raise media attention as early as Sep 11, with tweets stating: "trainings! medic! legal support! communication training. facilitator trainer #occupywallstreet #sept17". As the actual event (Sep 17th) drew near, the organizers were giving more specific instructions, as they posted on the 14th :"bringing tent sleeping bag food water to new york this weekend!"

© 2012 The Author(s)
© 2012 The Eurographics Association and Blackwell Publishing Ltd.

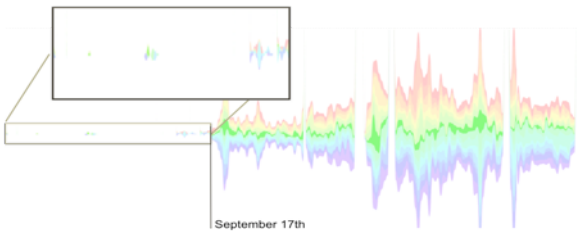

Figure 3: *Precursor to the Occupy movement. People started to organize and advertise the event way before the official beginning date of Sep 17.*

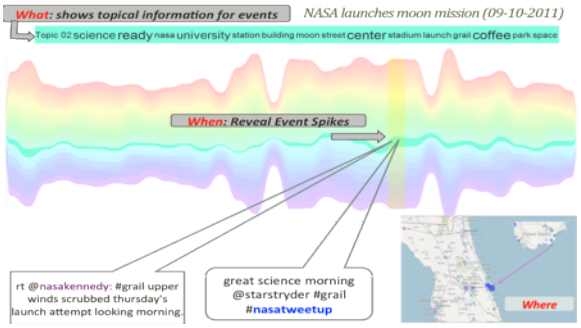

Figure 4: *Topical burst indicates NASA's moon mission.*

And even on the early morning of the 17th, protestors were provided maps of how to get to Zuccotti Park: "hashtags user-friendly time-table map uploaded". At this point, it became obvious that the initial protest was orchestrated by a group of organizers.

To seek the possible origin of the movement, the user kept retracing the tweets published earlier than the 14th and noticed that there were other hashtags that frequently co-occur with #occupywallstreet during the first few days of the movement. These hashtags include: #usdayofrage, #yeswecamp, #nyccamp, etc. At this point, the user could carry the investigation further by looking into tweets with these hashtags around that time. Such observation was validated by recent Wikipedia's updates on Occupy movement where the U.S. Day of Rage (#usdayofrage) was considered the governing body of the OWS group [IBT].

In summary, the *I-SI* framework supports the analysis of social media data regarding the Occupy movement. The backend topic modeling plus frontend interactive visualization supports investigative analysis of the otherwise unorganized and noisy information, and enables the answering of questions such as how did the movement evolve, which forces joined the movement at which time, are there any precursors to the Sep 17th protest, etc. As detailed in Section 6.2, the implication of this finding can be significant to public safety personnel in that, if they are able to acquire the key indicators hours/days before the protest, they can develop a better oversight and management strategy.

5.2 Scenario II: Establishing Investigative Analysis

We use this scenario to demonstrate the scalability of our architecture, demonstrating that the *I-SI* interactive analysis environment allows users the capability to tap into relevant data on a large scale. Over 12 million tweets were examined over the course of three weeks, with 30 topics extracted for interpretability. Unlike the previous scenario, these tweets were not filtered by hashtag. Thus exploration of the dataset will permit *it to tell the user* what it is about.

In this scenario, a summer research intern began by examining this tweet collection to discover interesting events he might have missed during the past few weeks. Upon highlighting different ribbons in the ThemeRiver (When) view, he notes that the cyan ribbon (see Figure 4) exhibits a unique temporal pattern. A closer examination of the timeline reveals a volume burst around Sep 10, suggesting more tweets were related to this topic within that time period.

The student then associates this timeline view with the topic cloud view (What), and finds that the topic refers to science and NASA. This becomes very interesting to the student, who happens to be an enthusiast in astrophysics. Quickly he references the tweets with their geospatial location on the map (Where) view. He observes that the mention of such event is mostly centered in Orlando, FL, where one of NASA's launching sites is located. Further browsing through the actual tweets suggests that people (Who) across the country are excited about this event. At this point, the student has linked all these investigative hints together and checked into the news database to find media coverage. He then correctly concludes that the event is the NASA GRAIL launch on Sep 10 to study the moon from crust to core.

The investigation in this scenario included all four of the W's and ended in the correct hypothesis of how a single space event could stir discussion and may inspire people toward scientific activities. It uncovers the cause (an event) of a certain volume burst in large-scale social media data and clarifies the trend and pattern of social phenomena.

5.3 Scenario III: Identifying Epidemic Spread

In this scenario, we demonstrate how *I-SI* can support investigation of the spread of an epidemic and pinpoint when the epidemic happened by analyzing microblog messages.

5.3.1 Data Source and Data Preparation

In contrast to the other two scenarios, the dataset in this case is provided by the VAST Challenge committee [VCO]. The dataset contains more than a million microblog messages collected within a major metropolitan area, over the course of a month. While synthetic, this dataset can be a great benchmark for its careful integration of the epidemic theme, the investigation of which requires robust analysis capability since the epidemic is buried in a mass of irrelevant microblog data. This VAST challenge is also a great evaluation of our *I-SI* environment, since it permits comparing the patterns observed from the visualization interface with the ground truth that comes with the dataset.

Relying on the robust text analytics capability in *I-SI*, we were able to effectively perform topic modeling over this textual corpus, with a vocabulary of 13,284 unique terms. 10 topics were extracted and visualized from the corpus after several experiments of topical interpretability.

© 2012 The Author(s)
© 2012 The Eurographics Association and Blackwell Publishing Ltd.

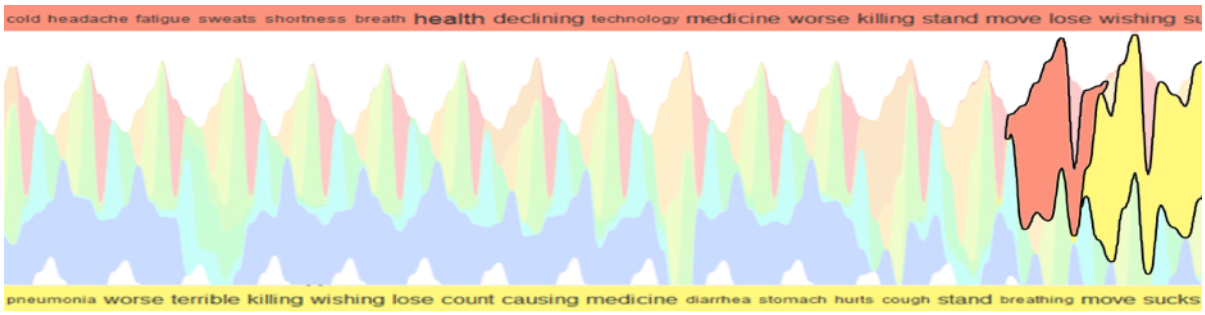

Figure 5: *Identifying the start of an epidemic spread. The orange topic captures flu-like symptoms, which bursts on day 1. The yellow topic shows on the next day, the symptoms have evolved to more severe ones such as pneumonia and diarrhea.*

5.3.2 Characterizing Epidemic Spread

As shown in Figure 5, the temporal patterns of the ten extracted topics are presented in our ThemeRiver view over the course of a month. Each time unit in the figure denotes 4 hours, which is adjustable to support inspection of different temporal granularities. Upon exploration of the temporal view, one can easily discover that multiple topics share a repetitive characteristic, such as the repeating mentions of TV shows every night (un-highlighted topic in red).

What really attracted the users' attention, however, is the sudden disappearance of the repetitive patterns during the last 3 days. Instead emerging topics in that time frame were about flu-like symptoms, such as "cold, headache, fatigue, sweats, etc." (see orange topic in Figure 5) and "pneumonia, diarrhea, cough, etc." (see yellow topic in Figure 5). These two topics signify exactly when the outbreak has begun. Moreover, our temporal view clearly suggested a progression of the illness from cold and headache to more serious symptoms, such as pneumonia, diarrhea and difficult breathing, since the orange topic stream appeared before the yellow topic stream. The results conform to the ground truth provided by the challenge committee. Finally, with the ability to pinpoint when the epidemic has begun, one can further conclude that the disease did not seem to be contained based on the volume of the yellow topic in the last day. Therefore, if the microblogs were collected and analyzed as the epidemic unfolded, the results could inform emergency responders to take actions to prevent the disease from spreading.

In summary, our *I-SI* architecture supports processing and analysis of the microblog messages. Though interactive exploration of the visualization results, a user could successfully identify latent information regarding the outbreak and depict temporal patterns of the epidemic spread.

6. Preliminary User Feedback

Compared to typical visual analytics evaluations, we recognize the challenges in conducting thorough evaluation of the *I-SI* architecture. This evaluation process, to bring it to completeness, may require experts from multiple research domains to collaboratively examine the efficiency and effectiveness of the scalable architecture. We believe that any findings from such evaluation would be of tremendous value; yet, conducting it can be a longitudinal process that needs more strategic consideration and explanation that beyond the scope of this paper.

Instead of focusing on evaluating the architecture as a whole, we seek users' understanding about the analysis environments that our architecture enables. In this section, we report user feedback based on several preliminary user evaluations with our investigative visual analysis interfaces. These evaluations were conducted to assess the effectiveness and efficiency of such an interface in supporting understanding latent social phenomena such as the three case studies shown above.

In particular, we report our interactions with three groups of experts in political campaign planning (CP) (5 experts), finance (4 analysts) and emergency response (3 lead experts). While the evaluations were conducted informally, these outreach activities granted us a sufficient amount of time to introduce our architecture and its visual interfaces as well to gather their feedback. First we presented our system by demonstrating the investigative scenarios described in the previous section. Then the experts were given some time to ask questions regarding the system and the interface. Finally, we concluded the evaluation by asking them to give feedback and comments. Given privacy concerns, we are removing all these experts' affiliations. However, they all agreed to have their comments published in this paper.

6.1 Monitor and Analyze Social Phenomena

One of the benefits that all these experts see in the *I-SI* architecture is its capability in helping to depict latent social phenomena that are otherwise hidden in the data. Especially to CP strategists, who are responsible for analyzing hundreds of political blogs and news on a daily basis, the capability to identify and summarize the latent topics from their data is of great value. One of the experts mentioned that, "this tool is very exciting in that it could give me a way to effectively assess what people are talking about with regard to political events." He further commented that this would provide a great baseline analysis for their strategic planning work, "where [their] line of business is about finding the right people and talking about the right things".

While the analysis environment was well received, one of the CP experts pointed out that trust issues and uncertainty might affect analysis outcomes. Given the accuracy needed in the CP's work, they were interested in learning

© 2012 The Author(s)
© 2012 The Eurographics Association and Blackwell Publishing Ltd.

how they can effectively validate the outcomes of the overall analysis architecture. The validations and quantification of analysis outcomes is certainly a crucial future direction as we continue enriching our architecture. Of course, it would be possible to validate a particular topical outcome by merely reading enough of the blog entries organized around that topic, but it would be good to have the visualizations of the automated results show this at once or with limited probing of details. Otherwise the approach is not scalable.

6.2 Potentially be Proactive to Key Event Indicators

As mentioned in scenario I (section 5.1), the *I-SI* architecture helped depict key event indicators for the Occupy movement. This capability is highly appreciated by emergency responders; and our results demonstrated great potential in facilitating their duties. The ER experts we interviewed were very excited to see the system in action. One of them indicated that the potential of having *I-SI* in their working environment could not only help them "follow up with what they knew", but also raise their awareness on "what they didn't expect". One usage case they pictured to use our system is for proactive measures for political events. They would like to utilize our tool to deploy their manpower in more targeted directions.

Due to the "limited resources (financially and personal-wise)", an ER manager mentioned that, "we can't respond to every small indicator that the system provided us." This requires our architecture to be able to perform more comprehensive event structuring, producing more a hierarchy with key indicators. His comment is well received, and we are working extensively on researching a quantifiable event structuring metric for ascertaining where attention is needed and resources should be deployed.

6.3 Follow the Influence of Social Events

Based on our interaction with marketing experts, one of their strongly emerging interests is utilizing the social media data to depict marketing impacts that are generated by certain social events. They see our architecture could potentially help them to follow their customer base, and understand their interests. As summarized by one of the experts, "this system ties the marketing loop back to us…It could help us to find a targeted audience and pursue that market with customized approaches".

While we demonstrated our data connection between structured data (e.g. GPS locations) and unstructured text, the further fusion of heterogeneous data is another important aspect for which these experts would like to have further evaluation. In particular, they are interested in learning how we could effectively associate information from different text corpora. This should certainly be doable at the topic level.

7. Limitations

There are limitations to this research that need to be addressed. The current analytics capability of our architecture is limited because this research was conducted within the specific Natural Language Processing area of topic modeling. We attempted to mitigate this limitation by componentizing our architecture, which opens up opportunities to incorporate other text analytics methods such as sentiment analysis and named entity recognition. Nevertheless, different characteristics, other natural language processing algorithms, and their scalability constraints could engender different analytical environments.

In addition, we undertook this architectural research to depict information of social media data from an analysis perspective. Our data management schema, which resides in Hadoop clusters, is still preliminary. We are in the process of determining more comprehensive data collecting and integration schema to handle the ever-growing complexity and scale of social media data. An important benefit of integrating an optimization process into our architecture is the potential for improved efficiency of the infrastructure, allowing informed resource management, avoiding replicated work.

The presented architecture illuminates the strong role that a combined approach of data-driven modeling algorithms and user-center visual analytics plays in revealing the latent phenomena within complex social media. It is our hope that by identifying these system limitations, the research fields of visual analytics, parallel computing and databases might be brought together, providing scalable solutions for social media analysts and new techniques for revolutionizing the analysis environments.

8. Future Work and Conclusion

In the future, we would like to enrich each component within the *I-SI* architecture. For the cluster computing, we would like to optimize LDA parallel processing algorithm and further improve its scalability and efficiency. As for the data analytics stage, we would like to add in techniques such as sentiment analysis and named entity recognition to automatically extract more information other than semantic topics from social media data.

In this paper, we presented a visual analytics architecture, *I-SI*, to effectively analyze unstructured social media data on a large scale. *I-SI* integrates data driven analytics methods such as topic modeling with human-centered visual analytics, via an interactive visual interface. We demonstrate multiple investigative visual analysis environments that I-SI is able to provide for monitoring, analyzing, and potentially enabling response to latent topical information extracted from social media.

Acknowledgement

We express our sincere appreciations to the domain experts for their valuable feedback. We would also like to thank supports by the National Science Foundation under award number SBE-0915528 and IIS-1019160. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

© 2012 The Author(s)
© 2012 The Eurographics Association and Blackwell Publishing Ltd.

References

[APA] APACHE HADOOP [online]:. http://hadoop.apache.org. 4

[BNJ03] BLEI D. M., NG A. Y., JORDAN M. I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3 (Mar. 2003), 993– 1022. 3

[BSH*10] BERNSTEIN M. S., SUH B., HONG L., CHEN J., KAIRAM S., CHI E. H.: Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology* (New York, NY, USA, 2010), UIST '10, ACM, pp. 303– 312. 2

[CAS] Apache Cassandra Project. http://cassandra.apache.org/. 4

[CBS] CBS News Online. URL:http://www.cbsnews.com/8301-501465_162-20117291-501465.html, Oct 7, 2011. 2

[CNN*10] CHEN J., NAIRN R., NELSON L., BERNSTEIN M., CHI E.: Short and tweet: experiments on recommending content from information streams. In *Proceedings of the 28th international conference on Human factors in computing systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 1185–1194. 2

[DCCC11] DE CHOUDHURY M., COUNTS S., CZERWINSKI M.: Identifying relevant social media content: leveraging informa- tion diversity and user cognition. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia* (New York, NY, USA, 2011), HT '11, ACM, pp. 161–170. 2

[DG08] DEAN J., GHEMAWAT S.: Mapreduce: simplified data processing on large clusters. *Commun. ACM 51*, 1 (Jan. 2008), 107–113. 4

[DGWC10] DORK M., GRUEN D., WILLIAMSON C., CARPENDALE S.: A visual backchannel for large-scale events. IEEE Transactions on Visualization and Computer Graphics 16 (2010), 1129–1138. 4.

[DNKS10] DIAKOPOULOS N., NAAMAN M., KIVRAN-SWAINE F.: Diamonds in the rough: Social media visual analytics for journalistic inquiry. In Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on (oct. 2010), pp. 115 –122. doi:10.1109/VAST.2010.5652922. 4

[DWCR11] DOU W., WANG X., CHANG R., RIBARSKY W.: Paralleltopics: A probabilistic approach to exploring document collections. In Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on (oct. 2011), pp. 231 –240. doi: 10.1109/VAST.2011.6102461. 2

[FAC] Christian K.' facebook page [online]: http://www.facebook.com/Nov.Fifth, Nov 5 2011. 2

[GOO] Google Map 2012 [online]: URL: http://map.google.com. 5

[HHN00] HAVRE S., HETZLER B., NOWELL L.: Themeriver: visualizing theme changes over time. In Information Visualization, 2000. InfoVis 2000. IEEE Symposium on (2000), pp. 115 –123. 5

[IBT] Ibtimes report [online]. URL: http://newyork.ibtimes.com/articles/215511/20110917/occupy-wall-street-new-york-saturday-protest. htm. 7

[JSFT07] JAVA A., SONG X., FININ T., TSENG B.: Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (New York, NY, USA, 2007), WebKDD/SNA-KDD '07, ACM, pp. 56– 65. 3

[MJR*11] MACEACHREN A., JAISWAL A., ROBINSON A., PEZANOWSKI S., SAVELYEV A., MITRA P., ZHANG X., BLANFORD J.: Senseplace2: Geotwitter analytics support for situational awareness. In Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on (oct. 2011), pp. 181-190. 3,4

[MON] MongoDB [online]: URL: http://www.mongodb.org/. 4

[NEW06] NEWMAN D., SMYTH P., STEYVERS M.: *Scalable parallel topic models* (2006). Journal of Intelligence Community Research and Development (2006). 5

[PLD] Google pld [online]: http://code.google.com/p/plda/. 4

[RCD10] RITTER A., CHERRY C., DOLAN, B.: Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (HLT'10), Stroudsburg, PA, USA (2010). 3

[RDL10] RAMAGE D., DUMAIS S., LIEBLING D.: Characterizing microblogs with topic models. In *Proceedings of the Fourth-International AAAI Conference on Weblogs and Social Media*, AAAI (2010). 3

[RUS11] RUSKOFF D. : Occupy wallstreet report [online]. October 2011. URL: http://www.cnn.com/2011/10/05/ opinion/rushkoff- occupy- wall- street/index. html. 6

[Siz10] SIZOV S.: Geofolk: latent spatial semantics in web 2.0 social media. In *Proceedings of the third ACM international conference on Web search and data mining* (New York, NY, USA, 2010), WSDM '10, ACM, pp. 281–290. 4

[SMKH05] SLOCUCM T., MCMASTER R., KESSLER F., and HOWARD H. *Thematic Cartography and Visualization*. Pearson Prentice 2005. 6

[Twi] TWITTER:. http://www.Twitter.com. 4

[VCO] VAST Challenge Committee 2011[online]: http://hcil.cs.umd.edu/localphp/hcil/vast11/. 8

[WLS*10] WEI F., LIU S., SONG Y., PAN S., ZHOU M. X., QIAN W., SHI L., TAN L., ZHANG Q.: Tiara: a vi- sual exploratory text analytic system. In Proceedings of the 16th ACM SIGKDD international conference on Knowl- edge discovery and data mining (New York, NY, USA, 2010), KDD '10, ACM, pp. 153–162. 2

[WMS*08] WANG X., MILLER E., SMARICK K., RIBARSKY W., CHANG R.: Investigative visual analysis of global terrorism. In *Computer Graphics Forum (2008), Computer Graphics Forum,* pp. 919–926. 3

© 2012 The Author(s)
© 2012 The Eurographics Association and Blackwell Publishing Ltd.