A Survey on Visual Analytics of Social Media Data

Yingcai Wu, Nan Cao, David Gotz, Yap-Peng Tan, and Daniel A. Keim

***Abstract*—The unprecedented availability of social media data offers substantial opportunities for data owners, system operators, solution providers and end users to explore and understand social dynamics. However, the exponential growth in the volume, velocity, and variability of social media data prevents people from fully utilizing such data. Visual analytics, which is an emerging research direction, has received considerable attention in recent years. Many visual analytics methods have been proposed across disciplines to understand large-scale structured and unstructured social media data. This objective, however, also poses significant challenges for researchers to obtain a comprehensive picture of the area, understand research challenges, and develop new techniques. In this paper, we present a comprehensive survey to characterize this fast-growing area and summarize the state-of-the-art techniques for analyzing social media data. In particular, we classify existing techniques into two categories: gathering information and understanding user behaviors. We aim to provide a clear overview of the research area through the established taxonomy. We then explore the design space and identify the research trends. Finally, we discuss challenges and open questions for future studies.**

***Index Terms*—Visual analytics, visualization, social media data.**

I. Introduction

Social media, such as Twitter and Facebook, have become prevalent in recent years. They can serve as powerful online communication platforms that allow millions of users to produce, spread, share, or exchange information at any time and any place. Such information typically includes multimedia content, such as text, image, and video. The huge amount of multimedia data spreading on social media imply rich knowledge and cover a wide spectrum of social dynamics occurring across the globe on an unprecedented scale and in real time. This phenomenon provides great opportunities to address important issues, which seem impossible to solve in the past. For example, the use of social media data has exhibited a huge potential in various applications, such as detecting breaking news, spreading news, coordinating rescue efforts, participating in local events, tracking a sports event, and gaining situational awareness during a crisis. However, the effective use of social media data is challenging because of its high heterogeneity, huge volume, and fast changing rate.

Yingcai Wu is with State Key Lab of CAD&CG, Zhejiang University. E-mail: ycwu@zju.edu.cn

Nan Cao is with NYU Tandon School of Engineering & NYU ShangHai. E-mail: nan.cao@nyu.edu

David Gotz is with University of North Carolina at Chapel Hill. E-mail: gotz@unc.edu

Yap-Peng Tan is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. E-mail: eyptan@ntu.edu.sg.

Daniel A. Keim is with University of Konstanz, Germany. E-mail: Daniel.Keim@uni-konstanz.de

Visual analytics is an emerging research direction for data exploration and analysis. It has been successfully used to understand multimedia data on social media. Compared with computer graphics, which aims to create photorealistic or non-photorealistic pictures or movies, visual analytics focuses on data analytics facilitated by interactive visual interfaces. Visual analytics enables human-centric computational intelligence by effectively integrating human knowledge and expertise into powerful computational algorithms through a high-bandwidth visual processing channel and user interactions. Visualization presents data and analysis results in context, and thus, it can provide rich evidence that supports or contradicts the analysis results, and consequently, help with data interpretation and result validation. Analysts can annotate on (e.g., place labels on) or adjust results via interactive visualizations to supervise the underlying analysis procedure based on advanced active learning algorithms, for example, and thus, gradually produce increasingly precise analysis and correct results.

Recent progress on visual analytics creates new opportunities for understanding large-scale, dynamic, and heterogeneous social media data. Compared with traditional data, social media data have a unique characteristic of having both network structure and heterogeneous information content, such as text and images. The two aspects often influence each other. For instance, information propagating in a social network can have great impact on dynamic changes in the network structure. The network structure, on the other hand, could also play roles in spreading information. Thus, visual analysis of social media data, for example, understanding user behaviors, should take into account both aspects in most cases. Moreover, social media data frequently come in huge volume with a high level of heterogeneity and substantial noise.

The characteristics of social media data present great challenges to traditional visual analysis methods, which only deal with structured, homogeneous, and small-scale data. Interest in research on visual analytics of social media data has been increasing. Researchers have conducted visual analytics studies published in conferences and journals across different fields, such as visualization, data mining, and social computing, to understand social media data. These studies on visual analytics have generated insightful results and findings, which demonstrate the success and effectiveness of using visual analytics in dealing with complex social media data. Along with these studies, many different visual analysis techniques have been invented to handle social media data.

Numerous prior studies scatter across disciplines, which poses a significant challenge for researchers to generate a comprehensive picture of the area and to develop new techniques, particularly for researchers who are investigating the area for the first time. In this paper, therefore, we survey the state-of-the-art research on the visual analytics of social media data

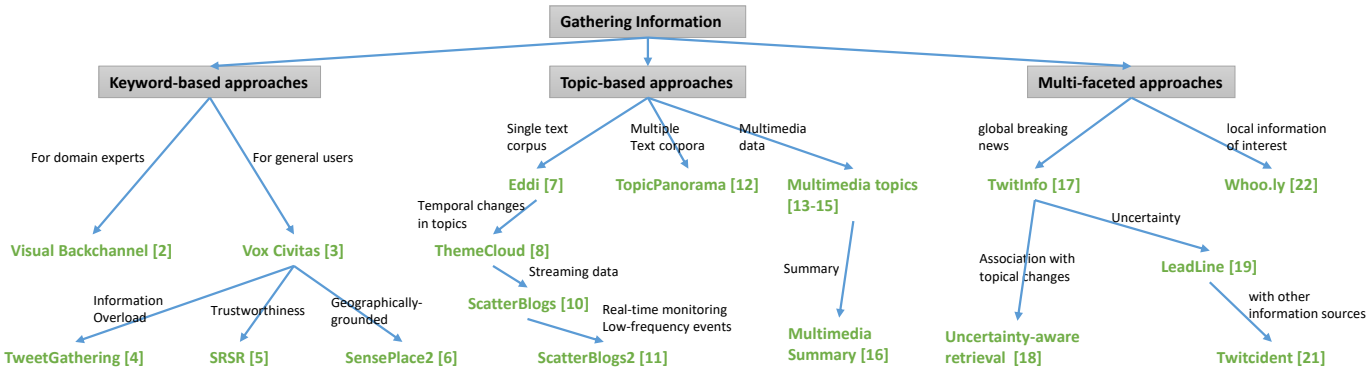

Fig. 1. The classification of visual analysis approaches for gathering information on social media.

and place prior studies in a unified and coherent framework. Through the survey, we aim to identify common approaches across different domains and cover the entire spectrum of related research. We also offer a new perspective on the challenges in the area and suggest promising future research directions. Furthermore, we establish a common research agenda by putting previous studies under a unified umbrella. This common agenda allows researchers from different fields to learn from one another and to identify common lessons, findings, and challenges across different fields. We contribute a new taxonomy of existing techniques and discuss the strengths and weaknesses of each category in the taxonomy. In particular, we classify existing methods into two major categories, namely, understanding user behaviors and gathering information.

In Section II, we introduce the background and methodology used in the survey. Then, we present a taxonomy of existing visual analytics methods in Sections III and IV. Visual analysis tools and engineering are introduced in Section V. We then summarize research trends in Section VI and discuss research challenges and agenda in Section VII. Finally, we conclude the survey in Section VIII.

II. Background and Methodology

Social media are web-based applications, such as Twitter and Facebook, which allow people to produce, spread, and exchange information, as well as to form online social networks. Compared with other media, social media are characterized by user-generated content and virality, which is defined as the tendency of information to be circulated quickly and widely among social media users via online social networks. Among different social media platforms, Twitter has been the most widely used in academic research because of its open policies that are friendly to academia. Twitter restricts message length to less than 140 characters, which enables crisp and targeted communication. Moreover, 500 million tweets are posted every day on Twitter, which capture thoughts from millions of users on practically all subjects. This survey mainly focuses on Twitter data. Apart from short text messages, information spreading on Twitter can have other forms, such as images, videos, and URLs. Therefore, social media data are inherently multimedia data, which may contain images, videos, text, as well as the underlying social networks. The rapid development of social media has been generating a huge amount of data, which enable data-driven applications, such as crisis management and viral marketing. Although a vast amount of data are being generated daily, these data often come in huge volume with high heterogeneity and substantial noisy information. Moreover, social media are dynamic and continuously updated, which creates social streams. Consequently, maximizing the advantage of data is a challenge.

Visual analytics is an emerging research direction that focuses on "analytical reasoning facilitated by interactive visual interfaces" [1]. It is a multidisciplinary field that combines visualization, human factors, and data analysis. Visual analytics aims to empower analysts to obtain insights from massive, dynamic, and ambiguous data via a seamless integration of computational intelligence with human knowledge, intuition, and expertise. Visual analytics has been adopted as an effective means to handle social media data. Researchers have proposed various new methods across different disciplines, which pose a challenge to derive an overall picture of related research. In this paper, we use an iterative method to review the literatures. We subsequently develop a taxonomy of existing research, which can adequately characterize and distinguish among the existing methods in the literature. The taxonomy categorizes prior studies into two classes, namely, gathering information and understanding user behaviors.

III. Gathering Information

Social media have been generating a large amount of highly diverse information, such as celebrity news, personal updates, and breaking news, across the globe on a daily basis. Thus, social media offer valuable sources for people to acquire their desired information. However, the unprecedented scale of social media data with considerable noisy information can easily overwhelm people and prevent users from acquiring meaningful information. Assisting users in seeking valuable information on social media has received significant attention in recent years. Accordingly, various approaches have been proposed to assist general users in searching for their desired information, journalists in gathering news information, and decision makers in maintaining situation awareness.

We summarize existing techniques into three categories, namely, *keyword-based*, *topic-based*, and *multi-faceted ap-*

Fig. 2. *Visual Backchannel* [2] introduced to visualize online conversations regarding a large-scale event on Twitter. The system integrates Topic Streams to represent topical development, People Spiral to indicate the activity of participants, Post List to show the recent posts, and Image Cloud to display shared photos.

proaches. Keyword-based approaches mainly visualize information which is retrieved by using a set of manually provided keywords or terms (see Fig. 1). Topic-based approaches empower users to retrieve information from social media using a combination of topic-based visualization and topic modeling and clustering methods. These approaches assist analysts in obtaining comprehensive insights from large-scale textual data. Notably, if a topic is simply regarded as one word by a method, then the method is still considered a keyword-based method. Multi-faceted approaches allow users to gather information from multiple perspectives.

A. Keyword-based approaches

Social media contain a huge volume of information concerning nearly every aspect of human society. The information abstracted as stream data is generally overwhelming, with a significant proportion of noise and useless information. Such flooding of information makes it difficult for users to acquire information of interest. Researchers in the visualization community depict this issue as information overload and propose various methods to address this issue.

Visual Backchannel (Fig. 2) was developed to follow and explore online conversations regarding a large-scale event on social media for general users [2] . Notably, the event to follow is defined by a manually-specified keyword or term. Tweets regarding an event can be continuously retrieved via Twitters open API. The collected tweets are subsequently processed to remove stop words and merge similar words, which results in a number of word stems that are regarded as topics. Apart from the traditional post listing, Visual Backchannel includes three novel interactive visualizations, namely, topic streams demonstrating topic evolution, people spiral indicating participants and their activity, and image cloud displaying shared photos. Topic streams are a primary view of Visual Backchannel that uses stacked graphs to visualize the dynamic changes in the frequencies of word stems over time from a live-changing social stream, such that both current and previous changes in the backchannel conversations can be clearly displayed. With such coordinated views, Visual Backchannel provides a visual summary of the backchannel conversations from temporal, topical, social, and pictorial facets.

Apart from helping general users seek information, social media have become valuable sources of newsworthy information for domain experts. Several interactive visualization systems have been developed to assist journalists in their search for information from social streams. Vox Civitas was designed and developed to help journalists and media professionals extract valuable news based on large-scale visual aggregations of social media contents [3]. Apart from the simple keywords specified by users, four types of automatic content analysis methods, namely, relevance, uniqueness, sentiment, and keyword extraction, are utilized to assist users in searching and filtering tweets related a large-scale event. Vox Civitas offers an easy-to-use user interface, which aligns a video of an event to several simple views, such as keyword and message volume graph views. However, Vox Civitas still has several limitations, such as information overload, lack of trustworthiness, and no support for situational awareness.

To address the issues of information overload and lack of context, Zubiaga et al. [4] introduced a user interface, *TweetGathering*, which Twitter users could easily adapt. It integrates flexible filters, ranks trending keywords according to their newsworthiness, displays representative tweets to hasten information access, and add necessary context to short tweets.

Although social media offer abundant and valuable information to journalists, finding reliable and trustworthy information is difficult. Diakopoulos et al. [5] presented a visual analysis system called “Seriously Rapid Source Review” (*SRSR*) that would enable journalists to find and assess information sources on social media. SRSR uses an eyewitness detector obtained from a dictionary-based technique to extract first-hand, on-the-ground tweets regarding breaking news. These authors also used a k-nearest neighbor method to classify information sources (users) into three categories, namely, organizations, journalists/bloggers, and ordinary users. SRSR offers an intuitive visual interface to help journalists find and assess information sources.

Maintaining geographically-grounded situational awareness, which is critical for crisis management, has received much attention. *SensePlace2* [6] was developed to support situational awareness by displaying spatiotemporal information of social streams. It adapts a heatmap visualization technique to show the frequencies of the retrieved tweets with respect to a particular topic. Thus, professionals can flexibly use temporal and spatial filters to obtain their desired information from huge volumes of information on social media.

B. Topic-based approaches

Keyword-based approaches enable users to efficiently retrieve information of interest using a set of provided keywords. However, the volume of gathered information can easily exceed the analysis capabilities of users. Prior studies have used keywords or hashtags to organize messages into topics, but the keywords or hashtags may fail to adequately characterize underlying topics and can easily lead to several “topics”

that are difficult to distinguish and examine. To cope with these issues, topic-based approaches that adopt advanced text mining, information retrieval, and natural language processing techniques to extract semantic topics from social media messages have received considerable attention in recent years.

Eddi [7] presented a novel topic clustering method, namely, *TweeTopic*, which transforms a tweet into a search query and sends the query to a search engine. A set of topic descriptors can be obtained in the search results. TweeTopic utilizes a search engine as an external knowledge source to overcome the length limitation of tweets and improve retrieval accuracy. Through TweeTopic, Eddi enables users to browse the content of a particular topic using the tag cloud, timeline view, topic dashboard, and navigation list.

ThemeClouds [8] aims to understand the views of users regarding a specific topic over time. Thus, compared with TweeTopic, ThemeClouds can help reveal temporal changes in topics of a text corpus. It creates a profile document of a user by combining all of his/her tweets for a time step. Subsequently, it adopts a scalable min-max linkage agglomerative clustering method to cluster the profile documents of all the users for that time step. The clustering algorithm is applied to each time step to generate a series of cluster trees, which is used to produce multilevel tag clouds. Once a cluster is selected, ThemeClouds automatically determines a suitable resolution and summarizes the collective discussion over time using multiple views of multilevel tag clouds.

Although TweeTopic and ThemeClouds can provide an overall picture of a text corpus, these tools cannot handle streaming text data. In recent years, several topic-based visualization tools [9]–[11] have been developed to support event detection and monitoring using streaming data from social media. *ScatterBlogs* is a scalable and interactive visualization approach that supports the detection and exploration of abnormal events and topics from various social media streams, such as Twitter, Flickr, and YoutTube [10]. ScatterBlogs extracts inherent topics from social media messages using *Latent Dirichlet Allocation* (LDA), a widely accepted topic modeling method in text mining. The extracted topics are subsequently examined to identify unusual and unexpected topics using a seasonal-trend decomposition algorithm. Abnormal events marked by peaks and outliers are then detected via z-score evaluation. ScatterBlogs uses an interactive visual interface with map visualization, tag clouds, and a histogram to support the interactive exploration and visualization of spatio-temporal social media data. However, the automatic method that applies LDA slows down system performance. Furthermore, real-time monitoring is unsupported by ScatterBlogs. In addition, low-frequency events and topics may be ignored by ScatterBlogs.

Accordingly, *ScatterBlogs2* was developed to overcome the aforementioned issues using a two-stage strategy [11]. The first stage allows users to create, modify, and test classifiers and filters using recorded microblog messages. A graph-based filter orchestration view is adopted to visually and interactively create a filter graph with a sequence of filters or classifiers. Users can define filters to cope with unusual and low-frequency topics and events. The second stage enables users to monitor a real-time social stream using the classifiers

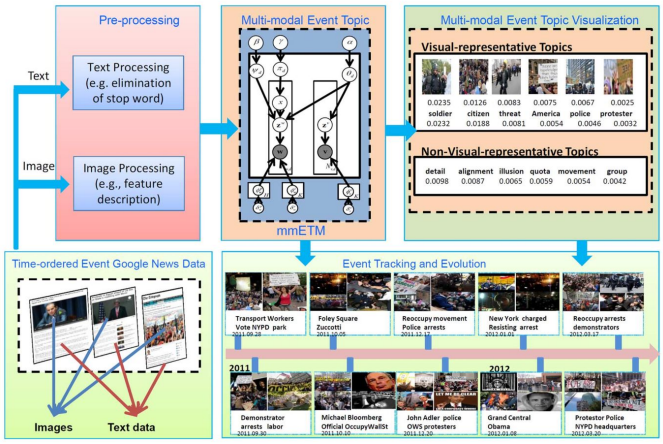

Fig. 3. A multi-modal framework designed to detect topics from multi-modal information, track the evolution of the topics, and visualize the topics with text and images over time [13].

and filters created during the first stage. The visual interface of the second stages consists of a set of coordinated views, such as a map view and an LDA topic view, to allow for real-time monitoring and exploration of a social stream.

Most existing tools generally deal with a single text corpus, such as Twitter, or individually handle different corpora. These tools may not be give a full picture of ongoing events on social media. *TopicPanorama* was recently proposed to allow researchers to simultaneously analyze and correlate the topics of different corpora simultaneously [12]. TopicPanorama models each textual corpus as a topic graph; then, the topic graphs of different corpora are matched using a tailored consistent graph-matching algorithm. The matching algorithm enables joint optimization among different topic graphs via real-time modifications of the matching result. A complete picture of relevant topics across different corpora can be revealed via a combination of radially stacked tree visualization and density-based graph visualization. TopicPanorama is highly interactive and assists users in interacting with matched topic graphs at different granularity levels. However, TopicPanorama only handles small-scale graphs and simultaneously visualizes several corpora.

Although significant progress has been made, the aforementioned approaches only exploit textual user-generated information, which is often noisy, short, and sparse. Thus, it can be difficult to detect meaningful topics. To this end, recent work takes advantage of multimedia information, such as image and video, to detect semantically meaningful topics on social media [13]–[15] and create a visual summary of multimedia information [16]. A popular method is the similarity graph method that transforms multimedia data into a graph [14]. Graph-based clustering algorithms are then applied to identify the topics. Qian et al. [13] introduced a multi-modal event topic model (mmETM) that can identify the correlations between textual and visual modalities, such that the semantic topics, including both visual-representative and non-visual-representative topics, and their evolutionary trends can be obtained. A multi-modal event topic visualization is used to vi-

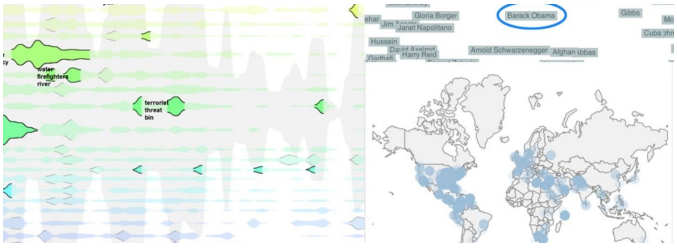

Fig. 4. *LeadLine* [19] designed to show dynamic changes in topics using a flow metaphor. An entity graph, a map visualization, and a tag cloud are used to examine and explore events from the perspectives of the 4 Ws.

sualize those topics and their trends with texts and images over time (Fig. 3). Cai et al. [15] proposed a generative probabilistic topic model named STM-TwitterLDA that jointly models five Twitter features (text, image, timestamp, location, and hashtags) to detect topics on Twitter. A maximum-weighted bipartite graph matching method is applied in tracking the detected topics. The topics are visualized using representative images determined based on three criteria, namely, visual relevance, visual coherence, and distinctiveness.

C. Multi-faceted approaches

Multi-faceted approaches assist users in acquiring information regarding social media events from multiple perspectives. Compared with other methods that mainly focus on understanding the textual content of social media messages, multi-faceted approaches provide a comprehensive overview by using a combination of advanced data mining methods to maintain situation awareness. *TwitInfo* is a typical multi-faced approach that aggregates and visualizes a large collection of microblog messages to explore an event [17]. It uses a novel streaming algorithm based on signal processing techniques to detect tweet volume peaks in real time. The detected peaks are highlighted and labeled with meaningful text extracted from tweets in a timeline-based display. Users can examine the peaks (subevents) from the timeline display. A set of linked views is used to visualize and explore multiple facets of an event on social media, namely, message frequency, relevant tweets, overall sentiment, and the geospatial distribution of tweets. A new method was also proposed to appropriately normalize overall sentiment aggregation to address the issue of different recall rates in positive and negative sentiment classifiers. Through these techniques, TwitInfo enables users to visually track user-specified events on social media by allowing users to collect, aggregate, and visualize tweets regarding these events.

However, retrieval uncertainty was not addressed in the work. A recent method was proposed to deal with uncertainty in the retrieval process, such that data can be retrieved more accurately [18]. The method uses a mutual reinforcement graph model to retrieve a multifaceted uncertainty-aware result. A composite visualization which combines a graph visualization, an uncertainty glyph, and a flow map was employed to visually analyze and explore the retrieval results.

Existing methods can help visualize temporal changes in topics using stacked graphs [2], [20]. However, only a few methods can visually relate topical changes to the associated events. To address this issue, *LeadLine*, an interactive visualization system was introduced to extract topics automatically and detect prominent events on social media [19]. A combination of topic modeling, event detection, and named entity recognition techniques is used to extract multi-faceted information, namely, the investigative 4 Ws (who, what, when, and where), with respect to each detected event. LeadLine adopts a flow metaphor to represent a topic visually in a row. Multiple flows visually encode the evolution of multiple topics over time. Furthermore, an entity graph, a map visualization, and a tag cloud are utilized to allow users to examine and explore events from the perspectives of the 4 Ws (Fig. 4).

The above multi-faceted approaches rely on one information source, which may have the risk of losing information or showing biased information. To ensure the reliability of the obtained information, information from other channels can be adopted. A framework called *Twitcident* was proposed to support filtering, searching, and analyzing information regarding incidents on social media [21]. The framework can monitor emergency broadcasting services to detect incidents in an accurate and reliable manner. These incidents are further enriched by using relevant information obtained from social media. The enrichment module contains a four-stage process, namely, the named entity recognition, message classification, linkage preservation, and metadata extraction. Twitcident also adopts two core filtering strategies for efficient filtering, namely, keyword-based filtering and semantic filtering. A visual interface that integrates a message list, line charts, map visualization, and pie charts is provided to support multi-faceted search and real-time analytics.

Apart from acquiring information regarding global significant events or breaking news, researchers also study visualization applications in which users can gather local information of interest. For example, *Whoo.ly* enables users to search for their desired local information using five coordinated views for displaying recent posts, active events, top topics, active people, and popular places [22]. Active events are detected using a new scalable statistical event detector, which first extracts trending features from posts and then groups topically-related features into event clusters via a nearest neighbor clustering algorithm. Two types of extractors, namely, template-based information extractor and learning-based information extractor, were developed to detect popular places from posts. Whoo.ly also ranks users to enable viewing of active members based on their mentioning and posting activities via an adopted algorithm similar to PageRank. The most frequently mentioned terms and phrases are detected as topics using a fast TF-IDF approach.

IV. Understanding User Behaviors

The increasing availability of various social media data, such as those from Twitter and Facebook, provides opportunities to gain a deeper understanding of various types of user behaviors on social media. Therefore, this topic has received increasing attention in the fields of computational social science and computer science. Among several related works in this topic [23], we focus on reviewing visualization techniques

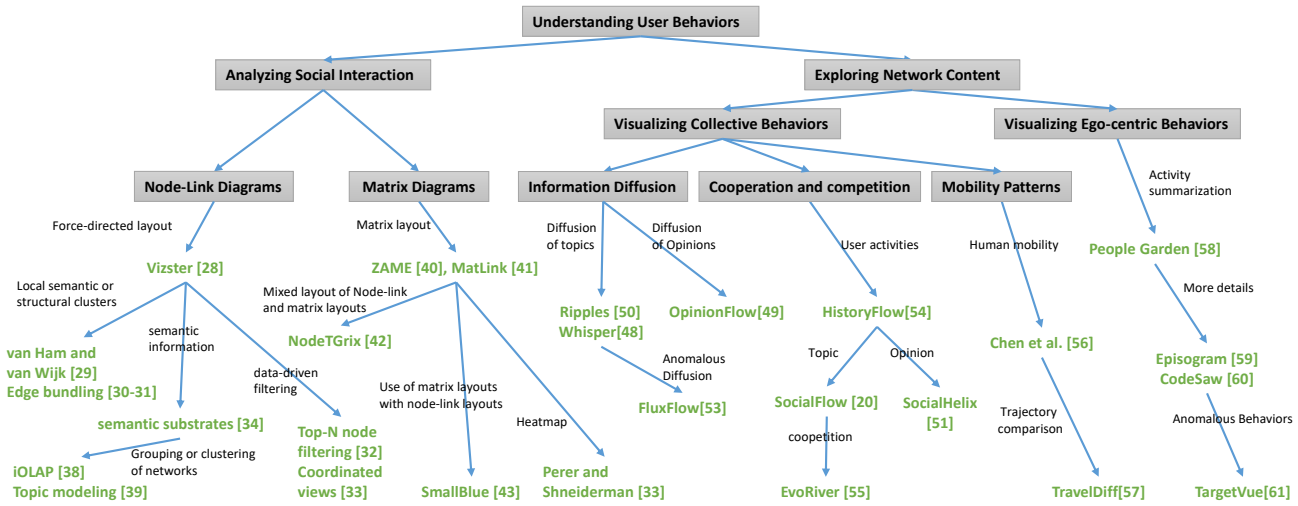

Fig. 5. The classification of visual analysis approaches for understanding user behaviors on social media.

that have been developed to improve the understanding of user behaviors on social media in this section.

In general, as shown in Fig 5, existing techniques can be largely classified into two broad categories, namely, analysis of social interaction and exploration of network content based on the explicit and implicit uses of social network, respectively. There are two general classes of approaches for visually analyzing social interactions among users: node-link diagrams and matrix diagrams. Exploration of network content can also be further divided into two categories based on different types of user behaviors (i.e., collective behaviors and ego-centric behaviors) to be visualized. Most existing research focused on visualizing collective behaviors, and the research topics in this category includes the visualization of (1) information diffusion, (2) social competition and cooperation, as well as (3) human mobility. In terms of visualizing ego-centric behaviors, existing research focused either on summarizing the behaviors or detecting anomalous behaviors.

The remaining section reviews the details of the techniques in each category based on their design objectives and application domains.

A. Exploring Social Interaction

Social media networks can be found in a wide array of contexts. While popular general audience websites (such as Facebook or Twitter) often garner much of the attention, similarly structured media networks exist in many other settings. Such networks often form around domain-specific topics for like-minded users to share information, as with the medical website PatientsLikeMe [24]. Alternatively, networks can form to support specific professional purposes, such as expert networks within large corporations.

Analyzing social interactions among users in social media networks can play important roles in a wide range of applications, such as friend recommendation on social media [25] and detection of leading roles and the corresponding communities in movie [26]. Prior methods and findings aim to derive insights from the interactions between entities in a social network. That is how entities are related, including both local structure and higher level constructs within the network. Research in this area typically combines interactive visual analytics methods with techniques developed in related research fields such as network analysis, graph theory, and domain sciences such as epidemiology or biology.

The structure of a social network is typically defined as a set of nodes (e.g., the people with accounts on a social media website) and links (e.g., the interactions or connections established between people using a social medial website). Both nodes and links can have one or more attributes (e.g, the number of "friends" a user has on a social media website; or the number of messages communicated between each pair of people connected by an edge). These networks can exist at vastly different scales (from single digits to billions of nodes) and can vary over time.

Considered in combination, these factors can make the visual analysis of network structure a major challenge. A sustained research effort over more than a decade has led to a range of technologies designed to support varies types of analysis activities. Adopting a classification that is similar to one proposed by Correa et al. [27], we identity two specific sub-specialties of research within the network structure class: (a) Node-Link Diagrams; (b) Adjacency Representations.

1) Node-Link Diagrams: Node-link diagrams are perhaps the most common way to visually represent a social network. The challenge is that most traditional graph drawing techniques do not scale effectively when social networks grow beyond tens or hundreds of nodes. As a result, research in this area often addresses the challenge of interactive exploration of a focus area of the network while maintaining a user's context of the large graph structure. For example, the *Vizster* system from Heer and Boyd uses a force-directed layout of nodes and links to visualize a social network, and allows users to interactively select focus areas to highlight specific subsets of the network [28] (see Figure 6).

In related work, van Ham and van Wijk [29] used similar force-based layout algorithms along with "circle and line"

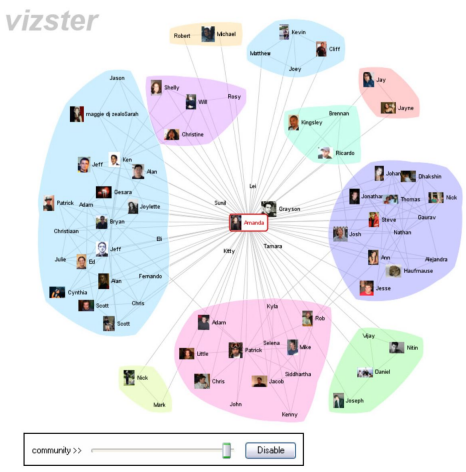

Fig. 6. The *Vizster* interface [28] introduced to visualize network structures based on a force-directed layout algorithm. The colored overlays represent communities identified within the network.

visual representations. However, they proposed alternative layout algorithms designed to emphasize local semantic or structural clusters rather than global layout metrics. Edge-bundling techniques, such as Hierarchical Edge Bundles [30] and geometry-based edge clustering [31], can also help highlight structure within the network by adjusting the routes of related edges into aligned bundles to reduce visual clutter.

Other attempts to identify local structure or otherwise focus analytic attention have utilized various forms of data-driven filtering. For example, one can use "Top-N" node filtering [32] to focus attention on important nodes as measured by various metrics. Alternatively, multiple coordinated views can be provided to visualize statistical measures about the structure of the graph. Users can then interact with the coordinated views to limit the number of nodes and edges on display at any one time in a semantically meaningful way [33].

When higher-level semantic information is available, such as when nodes have been classified into a set of categories, more structured representations can be explored. This includes, for example, the use of semantic substrates [34] to organize nodes within a higher-level groups before determining a final layout. This approach allows users to analyze patterns in edges of the graph as they connect nodes across or within groups. Following a similar approach, grouping or clustering of networks prior to layout can be seen in more recent social network visualization work [35]–[38]. For instance, Chi et al. [38] presented a unified framework called iOLAP based on a polyadic factorization method to directly model four important dimensions, namely, people, relation, content, and time in social network data. The network clusters detected by iOLAP can be used to create clear and structured node-link diagrams. More sophisticated algorithmic methods, such as probabilistic topic modeling [39], can also be used to create structured representations.

2) Adjacency Representations: While node-and-link diagrams provide an intuitive visual representation for the graph-based structure of a social network, the visual design adopted in these systems can also be critically limited in its ability to scale to large networks. Large numbers of nodes can produce a very large number of edge crossings, which in turn can produce difficult-to-interpret visualizations. Recognizing this limitation, alternative representations have also been explored.

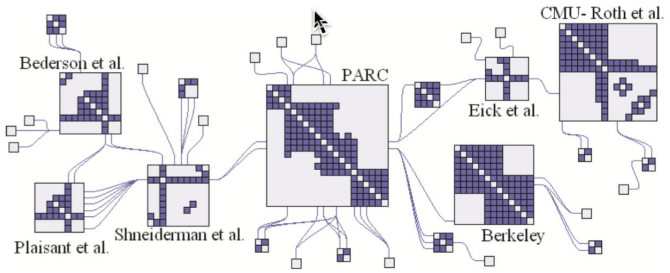

Fig. 7. The *NodeTrix* interface [42] to show a hybrid visualization combining adjacency tables with node-link diagrams to provide a multi-level view of the underlying network.

Adjacency diagrams are perhaps the most widely used alternative to node-and-link representations. These visual representations encode connectivity between nodes within some form of adjacency matrix rather than individually drawn edges, eliminating the need to embed the actual network structure within a two-dimensional plane [40], [41]. This idea has been used widely, often in clever combination with traditional node-link structures to provide views of structure at multiple levels of granularity. For example, Figure 7 shows the *NodeTrix* system from Henry et al. [42].

Some systems make only minor use of adjacency matrices, using them to complement other views. For example, Perer and Shneiderman used an adjacency matrix to show information about the "top 30 nodes" in a network by degree, resulting in a heatmap-like view [33]. Similarly, Lin et al's *SmallBlue* system uses adjacency representations together with traditional node-link diagrams [43].

B. Exploring Network Content

1) Visualizing Collective Behaviors: Similar behaviors from different people are generally aggregated first before being visualized because of the complexity and diversity of user behaviors on social media. Most existing techniques that follow this principle were designed to represent behaviors formulated by groups of social media users and identify behavioral patterns formulated spontaneously by the public. A wide range of collective behaviors, such as collective listening behaviors [44] and user attributes [45], have been explored and studied. In the following sections, we discuss and review the behaviors of spreading information, collaborating or competing with each other, and geospatial mobility.

a) Visualizing the process of information propagation: Understanding what, when, where, and how information is spread across space and time in social media platforms, such as Twitter, is a popular research topic in multimedia and visualization [46]–[49]. Studies have been conducted from different aspects, such as visually monitoring the information diffusion process [48], characterizing video diffusion [46], showing the diffusion history of messages/topics [50] or opinions [49],

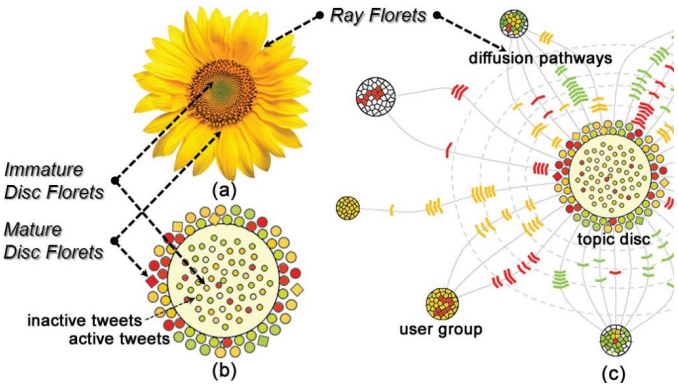

Fig. 8. The design of *Whisper* [48] follows a flower-metaphor, which uses different components of a sunflower to represent the key components of a diffusion process, including topics, communities, and the diffusion paths.

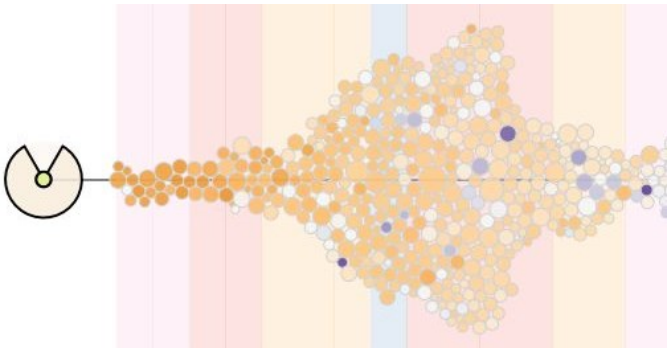

Fig. 9. *FluxFlow* [53] designed to illustrate the diffusion process of anomalous information on social media, such as the spreading of rumors in Twitter.

[51], understanding diffusion of users' rating behavior [52], and analyzing anomalous diffusion patterns [53].

Researchers have proposed various models to characterize the diffusion of information on social media. Niu et al. [46] analyzed substantial video propagation traces and revealed several interesting diffusion patterns. They found that a user's neighbors often have different activation times and the activation latency follows an exponential mixture model. Based on the findings, a comprehensive multi-source-driven asynchronous diffusion model was proposed to characterize the video diffusion behavior and predict the activation time on social media. The model parameters can be learned by an expectation maximization algorithm. Zhao et al. [52] studied the diffusion of user rating behavior and proposed a unified matrix-factorized framework. The framework typically models four important factors related to user and the item's topics, user interest, rating behavior habits, and behavior diffusion to gain insight into the diffusion process. Lei et al. [47] indicated that image diffusion is highly related to the relevance between image annotations and user preferences. Based on the finding, they presented a social diffusion model called common-interest model for image annotation. The diffusion of information captured or characterized by the aforementioned models can be visually analyzed using various visualization methods, such as timeline visualization [20] and radial visualization [48].

Google Ripples [50] and *Whisper* [48] are the earliest examples of techniques that were designed to trace the topic diffusion process on social media. Google Ripples follows a simple but effective design, which provides the diffusion history by showing a diffusion tree based on a circular tree layout. However, the approach does not support visual analysis of more complex spatio-temporal diffusion processes.

The design of Whisper is more sophisticated and can address the problem effectively. It tends to visually show three major properties of spreading processes in social media, namely, the temporal trend, the social-spatial extent, and the community response on a topic of interests. The entire information diffusion process centers around a focal topic that is visualized based on a sunflower metaphor (Fig. 8). The seeds of a sunflower are generally spread far and wide, thereby mimicking the spreading of tweets in Twitter. In general, Whisper visually summarizes the collective responses from communities on a given topic by representing how tweets are retweeted by groups of users. In particular, the visualization consists of three major components. A topic disc is placed in the middle of the view, which shows the focal topic, from which the corresponding tweets pop up in real time. A set of user groups are circularly arranged at the periphery of the display and connected to the topic disc by a set of diffusion pathways, along which, tweets are spread over time. Once a diffusion pathway is selected, it is visualized as a timeline, in which the key roles in the diffusion process, such as those of opinion leaders will be highlighted. This design facilitates the understanding of when and where a piece of information is dispersed and the identification of the social responses of the crowd to large-scale events.

Despite the diffusion of topics, research has focused on visualizing the diffusion of opinions among people on social media. One of the most important works in this category is *OpinionFlow* [49], which illustrates the diffusion history of opinions based on Twitter data. The visual design of OpinionFlow uses a Sankey diagram to represent user flow across multiple topics because of its simplicity and the intuitiveness. On top of the Sankey diagram, a density map produced based on a directional Gaussian kernel is introduced to help visualize the diffusion trends and the directions of opinions among users for the given topics over time. A node-link diagram is also used to specifically highlight the individual diffusion paths on top of the density map, thereby providing additional details regarding the diffusion.

Recently, research attention has also been focused on understanding anomalous diffusion processes. For example, *FluxFlow* [53] was designed to help understand the diffusion process of rumors detected by an underlying analysis model. FluxFlow uses multiple coordinated views to help put the analysis results in a rich context, which helps interpret and understand the results. In particular, it introduces an aggregated temporal circle packing design (Fig. 9), which demonstrates how an original message is visualized and propagated among people over time. In this design, each circle denotes a user who has retweeted the original tweets. The size of a circle denotes the importance of a user, which is defined based on the number of his/her followers. The color of each circle indicates its anomaly score that is computed in the analysis model. Each

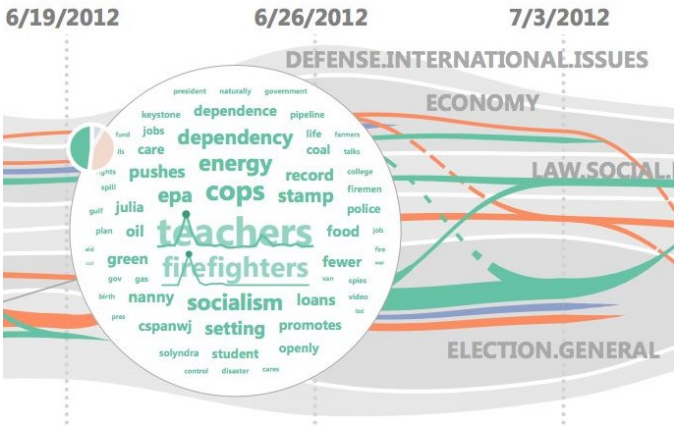

Fig. 10. SocialFlow [20] illustrates the competition of topics for attracting public attention in the process of information diffusion.

user is initially placed at the time point when he/she retweets the raw tweet along a timeline. The circle packing is used to remove overlaps and shows data patterns, such as the change in user volume. This design clearly represents the temporal diffusion process in a flow, which illustrates overview patterns and individual details. Thus, the design enables an elaborate comparison between normal and abnormal diffusion processes.

b) Showing cooperation and competition: As two of the most commonly occurring collective social behaviors, cooperation and competition attract considerable research interests. Several flow-based visualization techniques have been proposed to demonstrate cooperation and competition behaviors among groups of social media users from different aspects [20], [54], [55]. Viegas et al. [54] first investigated these types of behaviors based on social media data. In particular, they introduced *HistoryFlow* that was designed to illustrate how different users collaboratively or competitively edited on the same page in Wikipedia. In their work, a flow-based visualization was introduced to demonstrate the editing behaviors of users on different parts of a wiki page over time. Many interesting patterns were found in this study.

Recently, increasing interest has been focused on understanding "topic conflict", i.e., the process in which various topics compete for public attention while spreading on social media. In particular, Xu et al. [20] introduced a visual analysis system to illustrate topic competition. This system captures the competition process of multiple topics promoted by various opinion leaders based on a competition model. To facilitate visual analysis, a flow-style visualization was introduced (Fig. 10). The visualization can illustrate the change in competitiveness of each topic over time. Opinion leaders who demonstrate the behavior of promoting or demoting the spread of topics are visualized as threads, which indicates that their roles change over time. Sun et al. [55] expanded this work and introduced *EvoRiver* to visually analyze competition and cooperation (jointly called "coopetition") behaviors simultaneously. Compared with the work of Xu et al. [20], EvoRiver can capture and illustrate the complex relationships among topics based on the effects of carry-over, coopetition recruitment, and coopetition distraction.

In addition, Cao et.al [51] introduced *SocialHelix*, which illustrates another type of competition, i.e, sentiment divergence, on social media. In particular, sentiment divergence indicates disagreements in public opinions, which are generally triggered by events such as political campaign or promotions of competing products. To visually analyze such type of behavior, SocialHelix detects divergence through sentiment analysis and presents the results via a helix metaphor. In particular, two user communities with the most dramatic opinion divergence over time are visualized as two bends that form a helix. The representative tweets for the events that triggered the divergence are detected over time, and are grouped and visualized in the middle vertical bars that connect the two community bends in context. Thus, the divergence between two groups of people is thus visualized as two twisting belts that illustrate the development of the divergence and the change in group opinions.

c) Illustrating mobility patterns: Geo-tagged social media data have been used to uncover user mobility patterns. These data are generally extremely sparse (e.g., less than 1% of Twitter data have geo-tags), which makes analysis and visualization difficult. Chen et al. [56] introduced a visual analysis system that helps detect mobility patterns from sparsely sampled social media data. The system employs a wide range of visualization views to illustrate context from different perspectives to help detect interesting mobility patterns from sparse social media data. A heuristic model is further used to reduce data uncertainty, thereby guiding the appropriate selection of reliable data for further analysis and visualization. Users can explore the semantics of movements, such as the transportation methods, frequent visiting sequences, and keyword descriptions, based on this system.

Krüeger et al. [57] followed this research direction and introduced *TravelDiff*, a system designed to investigate and compare the travel trajectories of users using microblog data. The system introduces an interactive interface to facilitate visual comparison of mobility patterns extracted from the data set collected from different sources or events. In particular, the proposed visual comparison method highlights trajectory difference by normalizing and contrasting trajectories. The corresponding density maps are displayed within the same view. The system also hierarchically aggregates the resulting trajectories to produce a summary with a high-level structure.

2) Visualizing Ego-centric Behaviors: Compared with the aforementioned visualization techniques designed to analyze and represent various collective behaviors, only a small number of visualization techniques have been proposed to represent the behaviors of a single user on social media. Most existing methods draw a behavioral portrait of a user via a glyph-based design, which visually summarizes user behavior records in a single glyph, thereby enabling an effective visual comparison.

PeopleGarden [58] is one of the earliest works in this category. A flower-style glyph is used to visually summarize user activity histories in an online discussion group. The glyphs of different users are randomly placed in a "garden" area. Although it summarizes the interactions of users, it does not visualize other details, such as "when did an interaction occurred and who were involved in it".

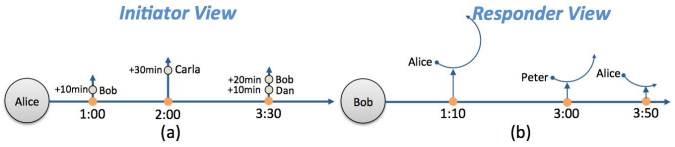

Fig. 11. *Episogram* [59] illustrates the following behaviors : (a) social interaction initiated by a centric user and (b) social interaction responded to by a centric user.

This issue has been addressed in follow-up designs. For example, to motivate developers in open source projects *CodeSaw* [60] used small multiples of line charts to visualize the activities of code contribution and social communication between developers over time. Cao et al. [59] introduced *Episogram*, (Fig. 11), which uses a glyph design to summarize the social interactions (e.g., posting or retweeting tweets). Episogram visualizes each interaction thread using a vertical line on a timeline. Two glyph designs were proposed to represent the social interactions which are initiated (e.g., posting tweets) by initiators or responded to (e.g., retweeting tweets) by responders, respectively. The first type can be simply visualized as a vertical thread line with its length indicating the duration of communication recorded in the data (Fig. 11(a)). A crescent-shaped glyph was introduced to represent how a centric user responds to interactions on top of the vertical thread line (Fig. 11(b)). In this design, the length of a crescent arc indicates the life circle of the associated communication.

Motivated by the preceding works, Cao et al. [61] further developed *TargetVue* to detect users with anomalous behaviors in Twitter. The system initially extracts a set of behavior features for each user account and uses an anomaly detection algorithm to identify a group of suspicious users in feature space. The most suspicious users are then visualized using two types of glyphs that show their communication behavior (i.e., posting or retweeting) and the corresponding behavior features, respectively. These glyphs follow a similar design scheme, in which a user is visualized as a circle in the middle. The color indicates the anomaly score computed using the analysis algorithm, whereas the size indicates the importance of the user as determined by the number of his/her followers.

V. Tools and Engineering

In recent years, significant open-source activity has resulted in a powerful ecosystem of general libraries and software, which can be used to support visual analysis of social media data. D3.js [62] is a representative JavaScript library, which is used widely to not only create simple diagrams and charts but also develop sophisticated visualization systems with complex data sets. Commercial software, such as Tableau and Microsoft Power BI, which allow non-developers to create data visualizations have also received much attention. However, most existing tools only support interactive visualization of structured data like data tables used in traditional databases.

To our knowledge, few libraries and software are specifically designed for visual analysis of social media data, which may contain different types of unstructured data, such as images, text, and network data. To visually analyze unstructured multimedia data, visualization practitioners circumvent the issue by first transforming the unstructured data to meaningful structured data by data mining or other analysis approaches. For instance, topic-based methods for gathering information described in Section III extract topics from unstructured text or images on social media. The structured topics are then interactively visualized and explored using various visual analysis approaches.

Many graph-based visualization libraries and tools have been created to help analyze social interactions and social structures in social networks. Given the difficulty of working with graph-based data in many traditional tools (which often assume tabular matrix-based data representations), several graph visualization libraries have been developed. Commonly used software packages include: *UCINET* [63], *Gephi* [64], *NodeXL* [65], and Pajek [66]. While not specific to graphs and networks, the widely used D3.js [62] visualization toolkit also includes basic implementations for several network-focused visualization algorithms. As these techniques mature, the need to engineer systems that work with ever larger sets of data is of growing importance. Many of these tools can be in combination with emerging large-scale graph analysis capabilities. In particular, large-scale and high-performance graph-based analysis platforms have been developed including the graph database platform *Neo4j* [67] and the *GraphX* analytics API from the *Apache Spark* project [68]. Visualization methods themselves must also be developed with scalability as a core design requirement. If done properly, such systems can scale to very large datasets without sacrificing interactive performance (e.g., [40] and [69]).

VI. Research Trends

This section discusses the research topics which have been explored so far using our taxonomy, and then discusses the strengths and weaknesses of the major techniques. Based on discussion, we identify the recent research trends and summarize our findings.

A. Gathering Information

Seeking meaningful information from social media has been extensively studied in recent years. Advanced text mining, natural language processing (NLP), and information retrieval methods have been commonly adopted to extract semantic information, such as name entities, topics, and events from social media. For example, LeadLine uses both topic modeling and name entity extraction methods to detect and characterize events, and thus, facilitate multi-faceted exploration and visualization [19]. Eddi [7] utilizes a search engine to overcome the character limitation of a tweet by transforming a tweet into a query to retrieve relevant long articles, which helps enrich the content of a tweet.

Furthermore, the trustworthiness or uncertainty of the retrieved information has received considerable attention [18]. For example, Vox Civitas [3], which aims to improve news gathering, defines the newsworthiness of a message based on both relevance and uniqueness to improve the trustworthiness of the retrieved result. TopicPanorama [12] utilizes a glyph

design to intuitively visualize the degree of uncertainty. The SRSR interface discovers news sources from social media using a distance-dependent method to generate a list of potential targets, through which prediction accuracy is improved [5].

Moreover, researchers have placed increased emphasis on multi-media contents, such as text, images, and videos in recent years. Visual Backchannel visualizes pictures intuitively as an image cloud, the layout of which is controlled by two parameters, namely, spacing constant and rotation range [2]. Vox Civitas [3] integrates microblog messages from social media and videos from a broadcast media event to enable journalists to determine public response to that event.

B. Understanding User Behaviors

With the rapid development of online social media platforms, understanding user behaviors become more and more important. This is a relatively new but active research area. Visual analysis of social interaction has been an active research area for many years and significant progress has been made. Early systems tended to focus on raw network structure. Advances in this area have resulted in more sophisticated methods, which allow for users to visually focus on important local structures or semantically organize a visualization for specific types of analyses. At the same time, the visual analytics community has developed novel visual approaches for representing social network data including adjacency representations and visual designs that focus on temporal dynamics. The maturity of many of these approaches has resulted in a wide range of application-focused work in this area. This includes visual analysis applications designed to understand network dynamics for applications where social media data is seen as clearly relevant: marketing, politics, or entertainment (e.g., [70]). Increasingly, however, these same capabilities are impacting a broader range of disciplines including epidemiology [71] and biology [71]. Others, however, adopt the method and apply them to new but structurally similar linked network data. In this way, the advances developed for social media applications are becoming enabling technologies for new discovery in fields like computational biology.

Most existing visualizations were specifically designed to reveal certain types of user behaviors in a specific application domain. Early systems, such as *HistoryFlow* [54] and *PeopleGarden* [58], visually summarize the behaviors and activities of a small group of people. Recently, considerable research has been conducted to explore and understand collective behaviors using big social media data corpus. Especially, there is growing interest in visualizing the process of information diffusion, which has been studied from different aspects, such as the topic and sentiment perspectives. Researchers have achieved notable progress in recent years on affective image understanding [72], [73], which helps shed light on sentiment propagation on social media using multimedia data in a more comprehensive manner. Understanding user cooperative and competitive behaviors is another hot research topic in this direction. Although some preliminary research (e.g., [20], [51]) has been conducted, this topic is still at an early stage. More research is needed to obtain a deeper understanding of how and why the cooperation or competition among topics takes place. Finally, in recent years, understanding and detecting anomalous or even malicious user behaviors via visual analysis approaches is also a promising direction due to the innate limitation of the existing automatic approaches and techniques. Many visual analysis systems have been developed to supervise anomaly detection with users' domain knowledge and experiences.

VII. Research Challenges

In this section, we share our perspectives on the research challenges and suggest an agenda for future research on visual analytics of social media data.

Recent progress shows that visual analytics has great potential in assisting users in seeking meaningful information. However, several challenges remain. First, the issue of information overload can be addressed by combining interactive visualization, NLP, and multimedia techniques [74], [75], but ensuring the robustness and efficiency of existing techniques remains a huge challenge. As expected, more efficient methods using parallel computing technologies will likely be proposed in the future to handle large-scale streaming social media data. Second, prior studies [2], [3] only display images and videos without dealing with embedded features or semantic information. Advanced imaging and computer vision techniques must be investigated to assist in gathering multimedia information from social media. Third, analyzing and understanding the trustworthiness of information gathered from social media streams remain difficult. The trustworthiness of information can be affected by various factors, such as the credibility of users and the reliability of the applied techniques. New methods should be explored to analyze and visualize the trustworthiness and uncertainty of information. Forth, the explosive growth of data on social media results in difficulty in visualizing data in a display screen with a relatively limited size and resolution. One possible solution is to extend the limited screen size to an unlimited virtual or augmented world using state-of-the-art virtual/augmented reality technologies. Nevertheless, addressing the occlusion problem remains difficult while providing natural user interaction to handle large-scale streaming social streams. Collaborative interactions in a virtual/augmented environment to explore and monitor information streams also warrant further study.

Research challenges for understanding user behaviors are diverse. First, neither a standard design principle in visualization nor a theory in social science exists to support visualization designs that illustrate the behaviors of users. Therefore, theoretical research in both social science and visualization is necessary to ensure a comprehensive understanding and correct representation of complex user behaviors. Second, in big data era, understanding large-scale behavior patterns and finding interesting user behaviors generally require both visualization and data analysis methods to achieve high performance, so that data can be processed in real time or near real time. Third, most existing techniques only focus on several aspects of various behaviors. A more comprehensive analysis of behaviors from multiple perspectives is needed given the complexity of the problem.

Despite the research community's record of progress in exploring social networks, many outstanding challenges remain. Chief among these is the challenge of scale. Widely adopted social networks have billions of users with seemingly limitless content being generated each day. Developing visual analytics methods that work effectively and at interactive rates when working real-world scale remains an enormous challenge. Similarly, real-world data is noisy and uncertain, due both to quality issues with real-world data and limitations of the analysis algorithms applied to that data. Visual analytics methods which can help convey uncertainty and other quality issues to users, and allow them to make judgments about the data despite those quality limitations, will be of critical importance. Finally, the multi-modal aspect of social media data is becoming ever more challenging. Data is being generated in more forms, from more sources, and being communicated to more people than ever before. This means that social network visualization techniques—which utilize primarily structured content—must be combined with analytics that make it easier to incorporate unstructured data such as videos or text. In addition, the visual analytics methods must be designed to work with complex datasets with very large numbers of variables.

VIII. Conclusion

Visual analytics is a promising research direction. It aims to empower people to analyze and explore complex data by integrating visualization, human computer interaction, and data analysis techniques. In recent years, much interest in developing visual analytics tools has been particularly intense on exploring and understanding social media data. Such data are a typical form of multimedia data that comprise text, images, videos, and networks. In this paper, we present a comprehensive overview of the visual analytics of social media data. A taxonomy of prior studies is introduced to classify the state-of-the-art techniques into two categories, namely, visual analytics for gathering information and understanding user behaviors. We trust that the proposed taxonomy can provide a coherent vocabulary for researchers to share knowledge and simplify analysis tasks.

The visual analytics of social media data is rapidly developing with numerous new methods emerging every year. However, the area is still in its infancy with many challenges and open questions. Many of the challenges cannot be addressed using techniques from only one discipline. We believe that multi-disciplinary research that combines visualization, multimedia, NLP, and human computer interaction will lead to more powerful and enabling approaches and technologies to handle and understand social media data. Thus, we encourage various communities to focus on this promising research area.

Acknowledgment

The work is supported by National 973 Program of China (2015CB352503), the Fundamental Research Funds for Central Universities (2016QNA5014), NSFC (61502416, 61602306), the research fund of the Ministry of Education of China (188170-170160502), 100 Talents Program of Zhejiang University, a grant from Microsoft Research Asia, and is based in part upon work supported by the National Science Foundation under grant no. DMS-1557593.

References

[1] J. Thomas and K. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Press, 2005.

[2] M. Dörk, D. M. Gruen, C. Williamson, and M. S. T. Carpendale, "A visual backchannel for large-scale events," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1129–1138, 2010.

[3] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in *Proceedings of IEEE Conference on Visual Analytics Science and Technology*, 2010, pp. 115–122.

[4] A. Zubiaga, H. Ji, and K. Knight, "Curating and contextualizing twitter stories to assist with social newsgathering," in *Proceedings of the International Conference on Intelligent User Interfaces*, 2013, pp. 213–224.

[5] N. Diakopoulos, M. De Choudhury, and M. Naaman, "Finding and assessing social media information sources in the context of journalism," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 2451–2460.

[6] A. M. MacEachren, A. R. 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 IEEE Conference on Visual Analytics Science and Technology*, 2011, pp. 181–190.

[7] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi, "Eddi: interactive topic-based browsing of social status streams," in *Proceedings of Annual ACM Symposium on User Interface Software and Technology*, 2010, pp. 303–312.

[8] D. Archambault, D. Greene, P. Cunningham, and N. Hurley, "ThemeCrowds: Multiresolution summaries of twitter usage," in *Proceedings of the Workshop on Search and Mining User-generated Contents*, 2011, pp. 1–20.

[9] 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*, To appear.

[10] 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 IEEE Conference on Visual Analytics Science and Technology*, 2012, pp. 143–152.

[11] H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl, "ScatterBlogs2: Real-time monitoring of microblog messages through user-guided filtering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 1077–2626, 2013.

[12] X. Wang, S. Liu, J. Liu, J. Chen, J. Zhu, and B. Guo, "TopicPanorama: A full picture of relevant topics," *IEEE Transactions on Visualization and Computer Graphics*, To appear.

[13] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 233–246, 2016.

[14] J. Pang, F. Jia, C. Zhang, W. Zhang, Q. Huang, and B. Yin, "Unsupervised web topic detection using a ranked clustering-like pattern across similarity cascades," *IEEE Transactions on Multimedia*, vol. 17, no. 6, pp. 843–853, 2015.

[15] 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 ACM International Conference on Multimedia*, 2015, pp. 89–98.

[16] J. Bian, Y. Yang, H. Zhang, and T.-S. Chua, "Multimedia summarization for social events in microblog stream," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 216–228, 2015.

[17] 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 ACM SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 227–236.

[18] M. Liu, S. Liu, X. Zhu, Q. Liao, F. Wei, and S. Pan, "An uncertainty-aware approach for exploratory microblog retrieval," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 250–259, 2016.

[19] 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 IEEE Conference on Visual Analytics Science and Technology*, 2012, pp. 93–102.

[20] P. Xu, Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J. J. H. Zhu, and H. Qu, "Visual analysis of topic competition on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2012–2021, 2013.

[21] F. Abel, C. Hauff, G. Houben, R. Stronkman, and K. Tao, "Twitcident: fighting fire with information from social web streams," in *Proceedings of the international conference on World Wide Web*, 2012, pp. 305–308.

[22] Y. Hu, S. D. Farnham, and A. Monroy-Hernández, "Whoo.ly: facilitating information seeking for hyperlocal communities using social media," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 3481–3490.

[23] L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos, "Understanding user behavior in online social networks: A survey," *IEEE Communications Magazine*, vol. 51, no. 9, pp. 144–150, 2013.

[24] "PatientsLikeMe." [Online]. Available: http://patientslikeme.com/

[25] S. Huang, J. Zhang, L. Wang, and X.-S. Hua, "Social friend recommendation based on multiple network correlation," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 287–299, 2016.

[26] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "RoleNet: Movie analysis from the perspective of social networks," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 256–271, 2009.

[27] C. D. Correa and K.-L. Ma, "Visualizing Social Networks," in *Social Network Data Analytics*, C. C. Aggarwal, Ed. Springer US, 2011, pp. 307–326.

[28] J. Heer and D. Boyd, "Vizster: visualizing online social networks," in *Proceedings of IEEE Symposium on Information Visualization*, 2005, pp. 32–39.

[29] F. v. Ham and J. J. v. Wijk, "Interactive visualization of small world graphs," in *Proceedings of IEEE Symposium on Information Visualization*, 2004, pp. 199–206.

[30] D. Holten, "Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 741–748, 2006.

[31] W. Cui, H. Zhou, H. Qu, P. C. Wong, and X. Li, "Geometry-based edge clustering for graph visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1277–1284, 2008.

[32] A. Perer, I. Guy, E. Uziel, I. Ronen, and M. Jacovi, "Visual social network analytics for relationship discovery in the enterprise," in *Proceedings of IEEE Conference on Visual Analytics Science and Technology*, 2011, pp. 71–79.

[33] A. Perer and B. Shneiderman, "Balancing systematic and flexible exploration of social networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 693–700, 2006.

[34] B. Shneiderman and A. Aris, "Network visualization by semantic substrates," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 733–740, 2006.

[35] C. Dunne and B. Shneiderman, "Motif simplification: Improving network visualization readability with fan, connector, and clique glyphs," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 3247–3256.

[36] S. Ghani, B. C. Kwon, S. Lee, J. S. Yi, and N. Elmqvist, "Visual analytics for multimodal social network analysis: A design study with social scientists," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2032–2041, 2013.

[37] L. Shi, N. Cao, S. Liu, W. Qian, L. Tan, G. Wang, J. Sun, and C.-Y. Lin, "HiMap: Adaptive visualization of large-scale online social networks," in *Proceedings of IEEE Pacific Visualization Symposium*, Apr. 2009, pp. 41–48.

[38] Y. Chi, S. Zhu, K. Hino, Y. Gong, and Y. Zhang, "iOLAP: A framework for analyzing the internet, social networks, and other networked data," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 372–382, 2009.

[39] R.-A. Negoescu and D. Gatica-Perez, "Modeling flickr communities through probabilistic topic-based analysis," *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 399–416, 2010.

[40] N. Elmqvist, T. N. Do, H. Goodell, N. Henry, and J. D. Fekete, "ZAME: Interactive large-scale graph visualization," in *Proceedings of IEEE Pacific Visualization Symposium*, 2008, pp. 215–222.

[41] N. Henry and J.-D. Fekete, "MatLink: Enhanced matrix visualization for analyzing social networks," in *Proceedings of International Conference on Human-Computer Interaction - INTERACT*, 2007, pp. 288–302.

[42] N. Henry, J. D. Fekete, and M. J. McGuffin, "NodeTrix: A hybrid visualization of social networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1302–1309, 2007.

[43] C. Y. Lin, N. Cao, S. X. Liu, S. Papadimitriou, J. Sun, and X. Yan, "SmallBlue: Social network analysis for expertise search and collective intelligence," in *Proceedings of IEEE International Conference on Data Engineering*, 2009, pp. 1483–1486.

[44] Y.-H. Yang and J.-Y. Liu, "Quantitative study of music listening behavior in a social and affective context," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1304–1315, 2013.

[45] Q. Fang, J. Sang, C. Xu, and M. S. Hossain, "Relational user attribute inference in social media," *IEEE Transactions on Multimedia*, vol. 17, no. 7, pp. 1031–1044, 2015.

[46] G. Niu, X. Fan, V. O. Li, Y. Long, and K. Xu, "Multi-source-driven asynchronous diffusion model for video-sharing in online social networks," *IEEE Transactions on Multimedia*, vol. 16, no. 7, pp. 2025–2037, 2014.

[47] C. Lei, D. Liu, and W. Li, "Social diffusion analysis with common-interest model for image annotation," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 687–701, 2016.

[48] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu, "Whisper: Tracing the spatiotemporal process of information diffusion in real time," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2649–2658, 2012.

[49] Y. Wu, S. Liu, k. Yan, M. Liu, and F. Wu, "OpinionFlow: Visual analysis of opinion diffusion on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1763–1772, 2014.

[50] F. Viégas, M. Wattenberg, J. Hebert, G. Borggaard, A. Cichowlas, J. Feinberg, J. Orwant, and C. Wren, "Google+ Ripples: A native visualization of information flow," in *Proceedings of the international conference on World Wide Web*, 2013, pp. 1389–1398.

[51] N. Cao, L. Lu, Y.-R. Lin, F. Wang, and Z. Wen, "SocialHelix: visual analysis of sentiment divergence in social media," *Journal of Visualization*, vol. 18, no. 2, pp. 221–235, 2015.

[52] G. Zhao, X. Qian, and X. Xie, "User-service rating prediction by exploring social users? rating behaviors," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 496–506, 2016.

[53] J. Zhao, N. Cao, Z. Wen, Y. Song, Y. Lin, and C. Collins, "#FluxFlow: Visual analysis of anomalous information spreading on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1773–1782, 2014.

[54] 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*. ACM, 2004, pp. 575–582.

[55] G. Sun, Y. Wu, S. Liu, T.-Q. Peng, J. J. H. Zhu, and R. Liang, "EvoRiver: Visual analysis of topic coopetition on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1753–1762, 2014.

[56] 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*, vol. 22, no. 1, pp. 270–279, 2016.

[57] R. Krüeger, G. Sun, F. Beck, R. Liang, and T. Ertl, "TravelDiff: Visual comparison analytics for massive movement patterns derived from twitter," in *Proceedings of IEEE Pacific Visualization Symposium*, 2016.

[58] R. Xiong and J. Donath, "Peoplegarden: creating data portraits for users," in *Proceedings of the ACM symposium on User interface software and technology*, 1999, pp. 37–44.

[59] N. Cao, Y.-R. Lin, F. Du, and D. Wang, "Episogram: Visual summarization of egocentric social interactions," *Computer Graphics and Application*, vol. To appear, 2015.

[60] E. Gilbert and K. Karahalios, "Using social visualization to motivate social production," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 413–421, 2009.

[61] N. Cao, C. Shi, S. Lin, J. Lu, Y.-R. Lin, and C.-Y. Lin, "TargetVue: Visual analysis of anomaly user behaviors in online communication systems," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 280–289, 2015.

[62] M. Bostock, V. Ogievetsky, and J. Heer, "D3 Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[63] S. P. Borgatti, M. G. Everett, and L. C. Freeman, "UCINET," in *Encyclopedia of Social Network Analysis and Mining*. Springer, 2014, pp. 2261–2267.

[64] M. Bastian, S. Heymann, M. Jacomy, and others, "Gephi: An open source software for exploring and manipulating networks." *ICWSM*, vol. 8, pp. 361–362, 2009.

[65] D. Hansen, B. Shneiderman, and M. A. Smith, *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Morgan Kaufmann, 2010.
[66] W. d. Nooy, A. Mrvar, and V. Batagelj, *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2011.
[67] "Neo4j: The World's Leading Graph Database." [Online]. Available: http://neo4j.com/
[68] "GraphX | Apache Spark." [Online]. Available: http://spark.apache.org/graphx/
[69] J. Abello, F. V. Ham, and N. Krishnan, "ASK-GraphView: A large scale graph visualization system," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 669–676, 2006.
[70] P. A. Gloor, J. Krauss, S. Nann, K. Fischbach, and D. Schoder, "Web Science 2.0: Identifying Trends through Semantic Social Network Analysis," in *Proceedings of International Conference on Computational Science and Engineering*, vol. 4, 2009, pp. 215–222.
[71] N. A. Christakis and J. H. Fowler, "Social network visualization in epidemiology," *Norsk epidemiologi= Norwegian journal of epidemiology*, vol. 19, no. 1, p. 5, 2009.
[72] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang, "Visual affect around the world: A large-scale multilingual visual sentiment ontology," in *Proceedings of the ACM International Conference on Multimedia*, 2015, pp. 159–168.
[73] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang, "Object-based visual sentiment concept analysis and application," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 367–376.
[74] S. D. Roy, T. Mei, W. Zeng, and S. Li, "Towards cross-domain learning for social video popularity prediction," *IEEE Transactions on multimedia*, vol. 15, no. 6, pp. 1255–1267, 2013.
[75] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Computing Surveys*, vol. 46, no. 3, pp. 38:1–38:38, 2014.

Yingcai Wu is a ZJU100 Young Professor at the State Key Lab of CAD&CG, Zhejiang University in China. His main research interests are in visual analytics and information visualization, with focuses on user behavior analysis, urban informatics, social network analysis, and text visualization. He received his Ph.D. degree in Computer Science from the Hong Kong University of Science and Technology. Prior to his current position, Dr. Wu was a researcher at Microsoft Research Asia, Beijing, China. For more information, please visit http://www.ycwu.org.

Nan Cao is a professor at TongJi University in China and the director of the HCI and Big Data Visualization Lab in TongJi College of Design and Innovation. Before joining TongJi, he was a research staff member at IBM T. J. Watson Research Center. His research interests include data visualization, visual analysis, and data mining. He creates novel visualizations for representing complex (i.e., big, dynamic, multivariate, heterogeneous) data in the domains of social science and medical informatics.

David Gotz is Associate Professor of Information Science in the School of Information and Library Science at the University of North Carolina at Chapel Hill (UNC). Dr. Gotz directs the Visual Analytics and Communication Lab and conducts research on a range of topics at the intersection of data visualization, HCI, machine learning, and statistical analysis. Dr. Gotz is also Assistant Director for the Carolina Health Informatics Program and an Associate Member of the UNC Lineberger Comprehensive Cancer Center. Dr. Gotz earned his PhD in Computer Science from UNC in 2005. He spent nearly a decade as a Research Scientist at the IBM T.J. Watson Research Center in New York before returning to join the UNC faculty in 2014.

Yap-Peng Tan received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1993, and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, in 1995 and 1997, respectively, all in electrical engineering. From 1997 to 1999, he was with Intel Corporation, Chandler, AZ, and Sharp Laboratories of America, Camas, WA. In November 1999, he joined the Nanyang Technological University of Singapore, where he is currently Associate Professor and Associate Chair (Academic) of the School of Electrical and Electronic Engineering. His current research interests include image and video processing, content-based multimedia analysis, computer vision, pattern recognition, and data analytics.

Daniel A. Keim is professor and head of the Information Visualization and Data Analysis Research Group in the Computer Science Department of the University of Konstanz, Germany. He has been actively involved in data base, data analysis, and information visualization research for more than 20 years and developed a number of novel visual analysis techniques for large data sets. Dr. Keim got his Ph.D. and habilitation degrees in computer science from the University of Munich, Germany. Before joining the University of Konstanz, Dr. Keim was associate professor at the University of Halle, Germany and Technology Consultant at AT&T Shannon Research Labs, NJ, USA.