Chapter 16
Analysis of Social Media Data: An Introduction to the Characteristics and Chronological Process

Pai-Lin Chen, Yu-Chung Cheng, and Kung Chen

16.1 Introduction

Mass communication scholars often refer to the impact social media has on the mass communication ecosystem. Not only would scholars like to have a greater understanding of social media content, but also the government and enterprises, who would like to see the developing trends of social consensus and the ongoing movement of consumers through social media content analysis. However, social media data analysis has characteristics of big data, which differs from the data analysis methods traditionally used by social science and previously applied by mass communication scholars. Consequently, social media data analysis is a current area of academic interest.

Even though many scholars believe the application of big data methods is a necessary trend (boyd and Ellison 2007; Mahrt and Scharkow 2013; Parks 2014), results put forth by the discipline of mass communication indicate that research articles concerning data-mining and the methods in which social media is analyzed are still rudimentary (boyd and Crawford 2012). The field of social media data analysis demands additional academic attention and contribution. This preliminary chapter seeks to describe the characteristics, elements, and the chronological process of social media data analysis from a mass communication scholar's perspective. Through case study, this chapter seeks to present ways a researcher analyzes social media data, and how that researcher poses questions and deals with the data during the process.

P.-L. Chen (✉)
College of Communication, National Chengchi University, Taipei, Taiwan

Y.-C. Cheng
Hsuan Chuang University, Hsinchu, Taiwan

K. Chen
Department of Computer Science, National Chengchi University, Taipei, Taiwan

S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_16

16.2 The Data Analysis of Social Media

Social media appearing at the start of the twenty-first century provides "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system" (boyd and Ellison 2007, p. 211). Facebook, Twitter, and PTT can be viewed as different types of social media. We view photographs and watch videos uploaded by our social media friends; we leave comments on social media posts; and we paste and/or share useful or interesting messages. In light of this, social media's information producing mechanism is very different from previous mediums of mass communication. It refers to platform mechanism for "a conversational, distributed mode of content generation, dissemination, and communication among communities" (Zeng et al. 2010, p. 13). The birth of social media not only revolutionizes the way people share information, but also hugely impacts mass communication research.

When a hundred thousand audiences use social media to produce, cooperate, and share every different type of message, the messages form a huge unprecedented dataset through the automatic categorization, record, and preservation by the media platform. The data generated by social media not only includes the content produced by human beings, but also comprises of metadata produced by machines. These data sets are huge in quantity and variety, which Lev Manovich (2012, p. 2) called "Social big data." The content of the data carried by social media results from the huge amount of facts, opinions, imagination, and feelings people have produced (Yang and Hhao 2016, p. 2). It provides a huge database that can be used as the target for collecting and analyzing (Stieglit and Dang-Xuan 2012). Many scholars, entrepreneurs, politicians and media workers seek to discover a social, political, cultural and/or industrial niche within the enormous data set. Tufekci (2014, p. 1) provides a vivid analogy: "the emergence of big data from social media has had impacts in the study of human behavior similar to the introduction of the microscope or the telescope in the fields of biology and astronomy." This metaphor points out how the birth of big data has brought a qualitative change to the research of social science: the thing the birth of big data changes is not just the scale of analysis, but also its vision and depth.

The birth of social media analytics is a field of knowledge that corresponds with the birth of social media. For scholars of mass communication, the birth of big data brings a profound change toward research method paradigm. The greatest significance of big data is not only the sudden multiplication of the quantity and scale of research data, but also the way mass communication scholars (boyd and Crawford 2012, p. 663).

Social media analytics is an emerging field to which scholars have provided different definitions. For example, Zeng et al. (2010, p. 14) maintain that social media analytics is a set skills for "developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application." Whereas

Stieglitz et al. 2014, p. 90 argue that social media analytics is "an approach toward research that involves multiple disciplines of knowledge." The mentioned scholars have not only provided the methodological foundation to other scholars from the perspective of different disciplines, such as business management, economics, and sociology, but also collected, mined, and analyzed the data, ultimately using large-scale social media to construct a data model intended to solve problems posed by the given academic or practical field.

Social media and data analysis can be viewed through two models when concerning contemporary society's ability to process information. A Nobel Prize winner in economics Daniel Kahneman points out that there are two thinking or decision-making process models: first, there is the model of thinking or decision-making process which generates responses immediately after the event (system I), second, there is the model of thinking or decision-making process that makes decisions through deliberation (system II). Respectively known as "fast thinking" and "slow thinking" (Kahneman 2012). On the one hand, social media of contemporary society has generated large-scale, diverse, and highly dense data which is conducive to the "reactive system" enabled by contemporary society. On the other hand, the data analysis of social media, conducted by researchers through applied methods in data collecting, filtering, and analyzing suggest a social reality rather similar to the "reflective system" Social media and its data analysis parallels the push and pull which maintains the equilibrium of social information embodied by "fast thinking" and "slow thinking."

16.3 The Challenge of Analyzing Social Big Data

Upon reviewing research articles from different disciplines, we discover that social big data analysis has faced various challenges including: the enormity and complexity of the data, the difficulty applying automated technologies when processing a dataset derived through human behavior or human expression, not to mention the questionable completeness and transparency of a data set still in its initial stages.

16.3.1 The Enormity of Amount and Scale

The quantity and scale of social data is large, the data types are diverse and in continual output. The preferred network platform in the age of web 2.0 is social media, allowing users to both produce and share data (Brügger and Finnemann 2013, p. 78). Data aggregated on social platforms include not only digital content, but also meta-data that characterizes the data (Cheng 2014, p. 80). Therefore, the amount and scale of social media accumulated data is bigger when compared to traditional media platforms. This is why researchers often fail to process this data when applying existing and traditional research methods (Stieglitz et al. 2014, p. 91; Boumans and Trilling 2016).

16.3.2 Human Generation

Social data is generated by a mechanism of human-generated computing. This type of data largely comes from the action of human languages and the human-machine interaction. This is different from other forms of instrument measured big data such as: outdoor temperature, accumulated rainfall, or atmospheric particulates observed.

Most social data includes two kinds of data, meta-data and content data. Meta-data is the "data that describes data," for example, the user account, the time of post, and the serial number of articles. They are usually in list form, and created by the computer system. Analysis of the content data is much more difficult. First, a lot of human and material resources are required to clean the data, because of the way human's use language and the diversity of data formats (a great amount of titles, keywords, tags, emotional symbols, plus the content of streamlined audio/video files can be included along with the textual content), the content's attributes are highly disparate; the connotations of the text are very complex. Second, there are several different ways of using human language: opinions, evaluation, irony, etc. One word may contain multiple meanings. Although researchers can look for the patterns of text through data science techniques (such as text-mining or machine learning), the intended meaning of a given text is still difficult to grasp. Third, although many people believe that patterns of online interaction are reflected in social media data, online interaction are interconnected with social situation, and the meanings of human interactions are complex. It is an oversimplification to interpret the meanings of complex human interactions from limiting, textually derived data accumulated from social media platforms. For example, clicking the "like" or "share" buttons on Facebook, statistics on the number of online messages, or the centrality of a social network connection. In this light, the degree to which automated techniques help people understand social media interaction is still very limited.

16.3.3 The Integrity and Transparency of Data

The other challenge facing social media analysis is data integrity. Social media data is one of the most important forms of revenue for social media companies. Social media makes social media data profitable by selling it. Therefore, social media companies restrict access to all their data. The data released through the API is a very small amount of the total data. For example, Twitter only releases 1% of its total data through the API. Payment is required to access the rest. Social media owns and safeguards the big data; the only organizations who can use social media data in its entirety are the enterprises and government organizations who can pay the high fees for big data access, as noted by Lev Manovich (2012).

Data analysts from academic institutions and average size enterprises must accumulate their own materials from data released by social media platforms, or hire data mining companies to provide the materials for them. When data is collected by

a third party, the integrity of the data is often linked to the technical capabilities of that party. For instance, social media is composed of many platforms; therefore, data sources vary and weighting among the data is difficult. An ongoing challenge is whether a "cross-the-board" analysis of Facebook, Twitter, and YouTube, is credible.

Besides data integrity, the other dimension worth noticing is data transparency. One criterion for scientific research quality is whether the data can be reproduced, whether the same result can be achieved by the same procedure. Thus, the procedures of data mining or analysis must be transparent. In all the fields collecting social media data, transparency is exists as debates regarding whether the algorithmic mechanism should be open source. The owners of contemporary social media platforms use algorithmic mechanisms to mine and release data. Due to this, data suppliers (the external data collectors) are incapable of knowing the algorithmic content of social media platforms through their own data crawling, or from purchasing data from social media platforms. Social media companies often consider their algorithms a market competitive business secret. Consequently, such companies will not allow open sourcing. Therefore, the degree of data transparency is insufficient, researchers cannot measure the rationality backing the data collecting procedures nor the integrity of the data accumulated through the algorithmic mechanism.

Due to the above reasons, the analytical requirements of collecting social media data are relatively high. Therefore, there is currently relatively little research based on empirical data. According to the meta-analysis made by Felt (2016, p. 4–5), in the 294 social media articles collected in the communication studies database *Communication and Mass Media Complete*, 83% still applied traditional data collection methods, only 17% of them collected data using information science techniques. The articles using traditional data collection methods largely used content analysis methods (21%) and survey questionnaires (20%; especially the online questionnaire). Not to mention, the same research found that the social media data collected was primarily from Facebook (69%) and Twitter (46%). There was infrequent transplatform research (<10%). According to the bibliographical review of Chiang and Lin (2015), of the 39 articles they sorted from the SSCI database, only 12 articles used numerical data for empirical research (the rest were mainly articles on theoretical or conceptual analysis). There were only two articles in the domestic TSSCI database in which only one article used data analysis methods. The research mentioned above also analyzed the number of authors. It showed that big data research is often coauthored by many contributors reflecting the nascent state of this field, and the necessity of teamwork due to its transdisciplinary nature.

16.3.4 Summary

As we mentioned above, traditional research approaches have difficulty handling large-scale, multi-material, multiform data filled with various kinds of noise. However, through the assistance of data science, social media related questions

posed by communication scholars can be solved. This means social media analysis possesses two processing qualities.

16.4 The Dual Processes of Social Media Analysis

Social media analysis bears the characteristics of a dual process. On the one hand, communication scholars need to develop problems worth probing into. Followed by discovering an answer through viewing what comes forth after data processing. On the other hand, there is a definite procedure for data processing. Researchers will collect, clean, and present data through visualization, in accordance with the posed problems. We will start by delineating the contents of these dual processes. Then we will demonstrate the relationship between them.

Data analysis is essentially the process "from problem formulation to problem solving." Academic researchers and practical workers both use social media data analysis to solve problems. While communication scholars may focus on questions like, how do messages of tremendous catastrophes communicate, disseminate, and/or converge through a social media platform (e.g., Chen and Cheng 2014)? Public relations employees in corporations or organizations use social media data to understand trends of consensus toward particular events, and take that data as a reference for maintaining the image of a brand or strategically handling a crisis (Li 2011). Although academic scholars and practical workers have different problems, question/answer strategies, meticulousness, and expectations toward analysis, both require consideration of data related questions and problems, as well as the process of finding an answer through social media data.

From the perspective of posing and positioning problems, social media data analysis is not much different from other academic or practical research processes. The real difference between other research methods and social media data analysis, which includes the subject of big data, are the details in handling the numerical data. As we mentioned above, social media data is generated by human beings. Social media data is not only in large amount, multiple in form, and filled with noise, but also incomplete. Therefore, not only does the problem formulation to problem-solving process require data science knowledge but the data processing process also requires further explanation.

16.4.1 Data Processing

The second factor in is data processing. Data processing includes five elements: (1) Data/Metadata: metadata is the data about data, or the structural message of data. For instance, cookies or user ID, data generated geographical, temporal information. (2) Algorithm: algorithm means that platforms will use a formula and variables to calculate social interaction. The algorithm decides the competitiveness of the

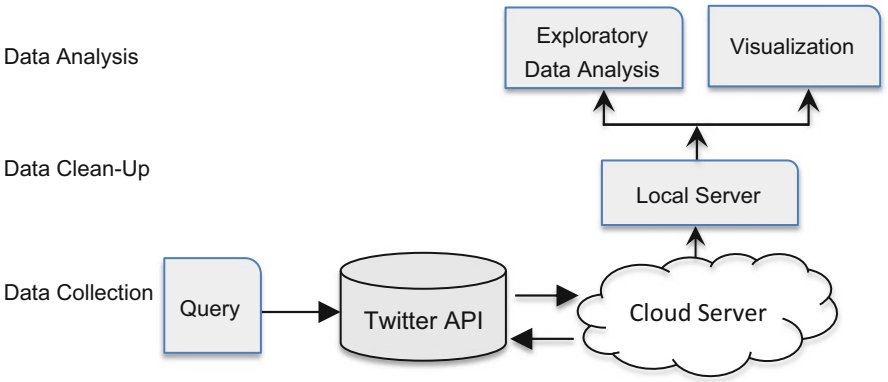

Fig. 16.1 A basic procedure to process data from social media platform

social platforms such as business secrets. (3) Protocol: protocol unites the data formats of different systems and implicitly directs user behavior in a manager-favored direction; (4) Interface: visible interface means the end user interface which is iconized and easy to use, the invisible interface is the one which is used to connect the hardware and software, and the API is the one between the visible interface and invisible interface; and (5) Default: the software has the function of directing the user (van Dijck 2013). We can induct social media data processing as the stage of data collection, the stage of data cleanup, and the stage of data analysis as per the following (Fig. 16.1).

16.4.2 Data Collection

Data collection is the initial stage of processing social media data. This stage is mainly about the process of tracking, monitoring, mining, cleaning, and arranging the digital footprints left by specific platforms or audiences, then making them into targets for analysis.

Because the quantity of social media data is big, the data-mining process often relies on scientific techniques which mine data through the automated software. The composite of specific data obtained through the information technology mining process is called the dataset. Generally speaking, there are two ways of obtaining a social media dataset: that is, you can access the dataset by API, or you can access the dataset by parsing the RSS/HTMAL.

The so-called "access by API" is the process in which researchers write the program, and log into the API created by the owners of the social media platform in accordance to specifications of different columns and limits provided by social media. One can set several vocabularies in the program, mine the data from social media servers, download and save data in a database, and wait for the subsequent data cleaning up and analysis (Fig. 16.2).

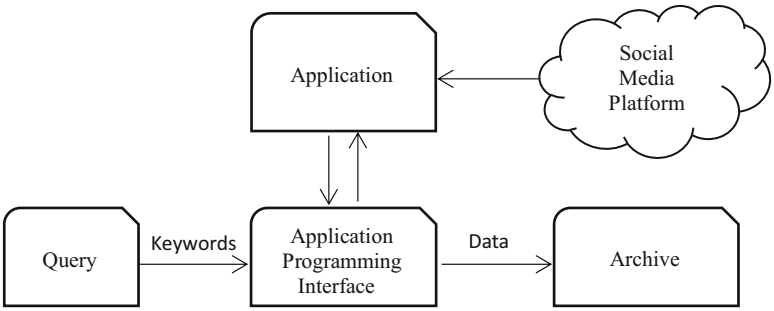

Fig. 16.2 Assessment by API of the social media platform's application interface

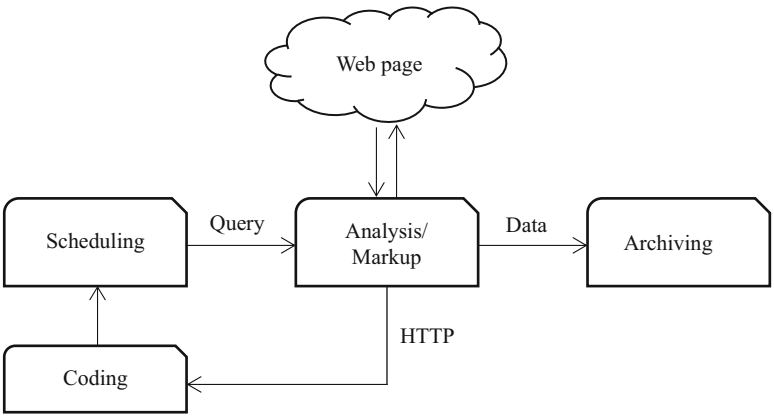

Fig. 16.3 The process of data crawling from the social media platform

On the other hand, the so-called "RSS/HTMAL parsing" is a program that attempts to simulate the human act of browsing a webpage, or the process that attempts to mine data from webpages of social media websites. Concerning the webpage of social media platforms, users can write web crawler programs to simulate human webpage browsing behavior. The schedule of these web crawling programs can often be set, manipulated to send messages to specific websites, and used to mine data from a specific section of the website through one-pass programming. Once the web crawler program has crawled the data (text and metadata), it downloads it to the database, analyzes them, and does parse/markup before saving them in a database, awaiting the cleanup and analysis of researchers. Because web crawling involves arranging the format in accordance with the webpage, the web crawler program must be updated whenever the social media platform owners change the webpage format (Fig. 16.3).

For researchers, whether to use API or Web crawler is a difficult decision to make. APIs usually set ceilings of the data volumes; and the versions of access protocol vary from time to time. On the other hands, web crawling imitates human

actions in browsing web pages, the right to access may be taken over when its behavior is considered "hostile" by the Platform. The researchers have to choose either method and live with it.

What data to collect? Bruns et al. (2013) pointed four major types of data used in Twitter analysis, namely keywords, Hashtags, Mentions, and URL. Particularly, researchers dealing with data analytics on significant public events usually use the above data sets. Among other ways, keywords are the most common way in accessing social media data. Researchers can identify or mine the targeted data using one or several social media keywords. Keywords are usually used for data crawling specific events (e.g., the presidential election, social movements, or catastrophes). The keyword's setting either comes from researcher experience, or from judgments made after reviewing the data. For example, collecting data on airline accidents using the airline company name, flight number, or accident location. Using words like "Malaysia Airlines" or "MH370" targets the March 2014 Malaysia Airlines aircraft disappearance.

The keywords used for data crawling are not only the keywords researchers set in relation the research goal, but also the hashtags provided by social media platforms. Hashtags are phrases set by social media users in accordance with a communication goal. They are often used to arch, group, and connect social issues. For instance, users of Twitter use #Orlando to mark the Orlando, Florida gunshot incident in June, 2016. By comparing the keyword/hashtag amounts during a specific period, researchers can understand the Orlando accident's social media trend and area of discussion.

An ever-lasting challenge to the social science researchers is: how to solve their specific research questions in this limited metrics of data?.

16.4.3 Data Cleanup

Every dataset can include characters, numbers, icons, or other formats. Because social media data contains a very large amount of non-structural data, researchers must clean the data.

The goal of data cleaning is to unify the format, reduce noise and convert data into an easy-to-process scale. It is the most time-consuming stage of data processing. As we mentioned above, social media data is generated by human beings. The content of social media data is mainly non-structural. Social media data must be converted into a machine-readable format before being processed by automatic information technology.

We can induct the data cleaning procedures as follows: format processing, selection/filtering, and integration or separation.

16.4.3.1 Data Format

Researchers often transform or arrange the data format into a desirable state. For example, the time of a social media post is usually recorded as a 15-number string. The string is converted into year/month/day by processing different sections. Another example is that much textual information—which is saved in Big-5 format—must be converted into UTF-8 format for convenient computing.

For analysis usage, data usually requires a cut after processing the data format. The cut is the handling process of data selection, filtering, integration, separation, etc. Due to this process, selection and filtering of data are vital.

16.4.3.2 Data Selection

After processing the format, the dataset is arranged like one to many matrixes. It can be saved as several tables. Each table stores its social media data in column and row. Social media data has many dimensions, for instance, the user account, the post time, and the message content. These dimensions are each stored in a vertical column. All columns are arranged from left to right. The so-called data selection is when researchers identify and preserve corresponding columns keeping with analysis requirements.

16.4.3.3 Data Filtering/Sifting

Every dataset contains several forms of data. Each is presented horizontally. Each presents the variables consistent with the sequence of individual columns. All data has a horizontally fixed, corresponding to the columns' sequence. The filtering/sifting of data is when researchers give instructions to save or delete data, according to the analysis requirement and the individual column's range of variables. For example, researchers only need to delete those appearing twice in the user number columns if they wish to delete those which repeat, so each user number appears only once.

16.4.3.4 Data Integration/Separation

From the start, datasets researchers usually analyze contain no corresponding or analyzable columns. Due to this, researchers must combine or separate different columns, as well as create new columns for analysis, according to the researches problematic. The so-called data integration is when researchers combine several columns into one. For example, the "like," "drop your message here," and total number of times one picture or article is shared exists in three different columns. The researchers wish to merge the three columns into one variable by the weighting of the three, and saving them as one new column. Conversely, data separation is

when researchers separate a single column into several columns. For example, the post's time column contains the year, month, and day of the original post, which are separated into three individual columns, before analyzing only the month.

16.4.3.5 Data Dimensions

The above actions such as data processing, selection, filtering, integration, or separation are used by researchers manipulating data, to make the data analyzable. During the process of updating social media data, the "data dimensions" are the datasets having the same properties. In social media data, the most common types of data dimensions are temporal types, spatial types, numeric types, categorical types, and relational types. For example, the time of the post is presented in the dataset in year/month/day columns. These columns co-present temporal dimensions. The post's location co-presents spatial dimensions in longitude and latitude columns. These dimensions are cleaned up and categorized as the desired goal the researchers wish to pursue.

However, not all the data dimensions correspond to the research questions. The dimensions are not immediately suitable for data analysis. To construct the research metrics, researchers usually organize the dimensions of the data according to particular research goals.

16.4.3.6 Data Metrics

The so-called "data metrics" is a function formula constructed by the dimensions of the data in accordance with the goals of research. The metrics are usually used as research variables. For example, the Facebook monitor integrates the quantity of the clicked "like" button, the number of dropped messages, and the number of "shares" generated by users into a set of functions:

The engagement rate of the fans page per day = (the quantity of the clicked "like" button + the number of dropped messages + the number of "shares")/the number of fans that day × 100%.

The researchers require the "quantity of the clicked 'like' button," the "number of dropped messages," and the "number of 'shares.'" From the data dimensions, they crop, combine, and weigh the above information forming the "engagement rate of the fan page per day," this obtains the engagement rate between the fan page and its audience.

The above mentioned "engagement rate" is constructed by researchers. Although all the data have the same amounts, scale, or dimensions. The result is different dependent on the different ways of constructing the metrics (e.g., the data will be weighted according to different extents of the clicked "like" buttons, dropped messages, shares, and participation). Thus, the effectiveness of the metrics must be further evaluated and proved.

Regarding "engagement rate" use, when researchers consider them as equal value, the amounts of the "like," "share," and critics can be aggregated and present the above mentioned functions. However, the weighting or the other ways of data processing must be considered while integrating the three values mentioned above, if the researchers are to deem the difference between the three above mentioned activities' extent of participation (e.g., Mayfield 2006; Forte and Lampe 2013) and decide that the difference should be made when the quantitative value is put forth (e.g., the researchers decide that the quantity of those clicking "share" > quantity for those of the critic>the quantity of those clicking "like"). In other words, researchers construct a set of function formulas based on his or her theories to calculate the data. Thus, the result of the analysis might be different according to the different metrics of calculation the researchers apply based on his or her theories, even though the used data set is the same.

16.4.4 Data Analysis

According to the goal of the data, there are two types of data analysis: exploratory analysis and argumentative analysis.

16.4.4.1 Exploratory vs. Argumentative Analysis

Exploratory Data Analysis (EDA) is when researchers discover the state or trend of data distribution through simple statistic indexes and visualization tools, and consider the state or trend as the basis of further data analysis (Tukey 1977; Seltman 2015). The explorative data analysis method is often used in the initial stage of data processing, and when the data's connotation is unclear. Since the characteristics of big data are that the sample is very similar to the pool (Schönberger and Cukier 2013), and every dataset to possess its singularity, the state of the data requires exploration, and explorative data analysis is vital during the initial stage of analysis. The process is both explorative and hypothetical, just as its surface meaning suggests. It has an explorative meaning in which the researchers can make the state of the data surface through simple statistic methods and visualization tools. It also has the hypothetical meaning in which researchers try to observe the state of data distribution and the possible irregularities, concerning highly uncertain data, to find out key points for questioning. For example, the convergent or divergent trend of the quantity of posted articles can be observed by constructing a simple temporal model in reference to the time and the number of posts, when the critical public event breaks out. The temporal relationship of the trend can be visualized, allowing researchers to understand the opinions or trends of social media, making analysis possible.

On the other hand, the "argumentative data analysis" is the process when researchers hypothesize or prove the result of the initial exploration by ways of classifying, grouping, correlating, or predicting data, including the construction of statistical models through the application of various kinds of statistical tools. Besides, researchers can reduce the data's scale and undergo qualitative analysis according to results from the initial exploration of quantitative analysis, in order to discover the data's connotation.

16.4.4.2 Types of Data Analysis

The analytical tools of social media data are very different from traditional social science analytical methods. This chapter aims to discuss the following types of data analysis: temporal/trend analysis, relational analysis, numerical/type analysis, text analysis, and sentiment analysis. The future types of data analysis will be increasingly diverse with the evolution of information science technology and research methods. The connotation of the different data analysis types are:

16.4.4.3 Temporal/Trend Analysis

This type of data analysis aims to observe the transformation of social media data's variables in a particular period according to sequence of time by using time and other dimensions as the materials. For instance, researchers observe the transformation of the extent of concern of social media users by referring to the number of social media posts or the migration of frequency of particular texts posted during a catastrophic event (Jungherr 2014). The transformation of the quantity's scale under the sequences of time can be illustrated by considering sequences of time as the X variable, and frequency of the post as the Y variable. For instance, concerning the Tweets of the 2012 Australian presidential election, Burgess and Bruns (2012) discuss how people aggregate and the election on Twitter. They collect 41,500 posted articles containing the hashtag #ausvote for 38 days before and after the election, and compare the frequency of the posted articles on different days. The research found that 22% of the articles were posted on the day of the vote. However, only 1900 out of 3700 people (51%) posted articles during the past 38 days, also posted articles on the day of the vote. In other words, over half of the users gave election opinions on the day of the vote.

Temporal analysis can be used to predict the trend or pattern over an issues lifetime. To predict the developing trend of social media discussion (Stieglitz et al. 2014, p. 92), this kind of data analysis survey records the life pattern of social media discussion, and stores every kind of time sequence models in the database, according to the algorithmic models constructed from information science/statistics (e.g., hidden Markov model). The cross-strait data analysis for Twitter on the 2012 presidential election is presented on the graph below. The graph shows the distribution frequency of posted articles by three different languages groups

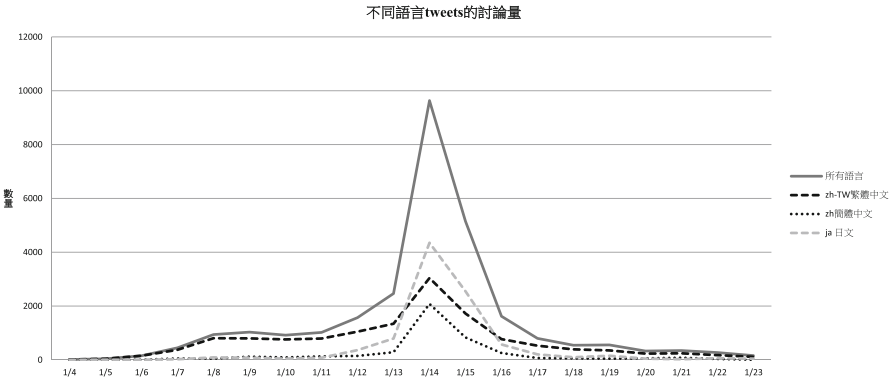

Fig. 16.4 The time sequence analysis of tweets during the 2012 presidential election (Cheng and Chen 2014)

(traditional Chinese, simplified Chinese, and Japanese) 2 weeks before and after the election. The increase in public opinions of those who apply traditional Chinese is earlier than the other two language groups (Cheng and Chen 2014) (Fig. 16.4).

16.4.4.4 Relational Analysis

Relational analysis uses relational data produced by the social media platform. The main goal is to observe the relational context among platform users. For example, when Twitter users retweet another user's article, that user's account is posted on the platform. This kind of record relates two users, and it can be considered as a node and link between them. The above type of data can be analyzed through mining and coding, in order to represent the social relation between those who post and those who receive the posts. For example, Kogan et al. (2015) collects tweets/retweets during Hurricane Sandy. They differentiate four kinds of social networks. First, by classifying users in affected areas or unaffected areas according to the tweet locations; second, by referring to tweets before, during, and after the Hurricane; third, by referring to the related data constructed by authors of the tweets/retweets, and by considering the authors of the tweets/retweets as nodes. Their research found that Twitter users in affected areas sent much more messages while under the hurricane's affect. These tweets form the tight and mutual related social networks during the Hurricane through tweeting. The data below comes from the cross-strait Twitter data during the 2012 presidential election. The analysis goal is to relate the data between Tweeters/Retweeters. Using graphic software from the network Gephi, the connotation this network graph shows is: although the three language groups (traditional Chinese, simplified Chinese, Japanese) focus on discussing the 2012 presidential election, users of the three language groups refer to different subjects. Showing that

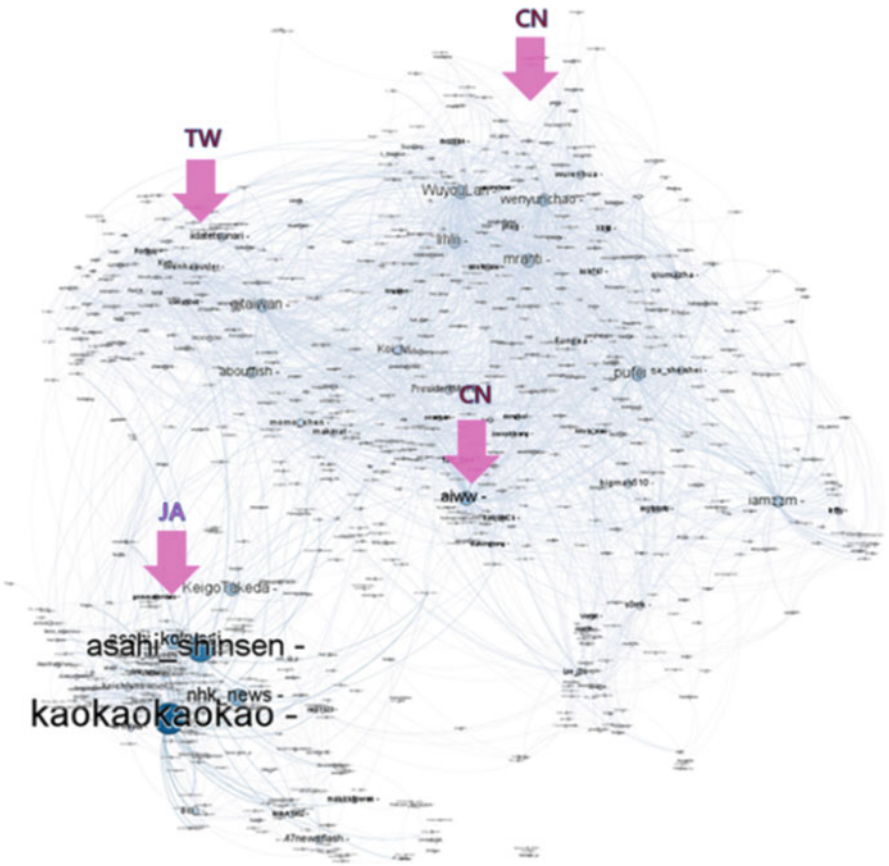

Fig. 16.5 The network analysis of language groups during the 2012 presidential election (Cheng and Chen 2014)

during the 2012 presidential election, the phenomenon is that Twitter users of the three language groups gave opinions concerning the same issue (Fig. 16.5).

16.4.4.5 Numerical/Type Analysis

Numerical/type analysis concerns the relationship between two or multiple data dimensions. This kind of analysis usually considers two groups of data dimensions as the variables, and cross analyzes them through statistical tools (such as SPSS or SAS). The data dimension can be either numerical data or type data. The researchers illustrate the relationship between data dimensions through the statistical results. For example, Bruns (2014) analyzes the relationship between the hashtags of buzzwords and reposted articles. For instance, to discuss the different states of related tweets,

in particular the interaction between different language groups, Burns and Burgess (2013) arrange user groups who speak Latin-based and non-Latin-based languages differently, and according to the quantity of their tweets, differentiate between the "active users," the "highly participant users," the "relatively inactive users" of the different systems of languages, by using millions of tweets which contain the hashtags #Egypt and #Libya, and by applying language recognition tools, concerning the protests people held in Egypt and Libya during "Arabic Spring." From observing the number of tweets with #egypt and #libya, this research finds that there are several differences concerning the number and transfer of the tweets. The most active 1% of Egyptian users usually use #egypt, particularly after the most common buzzword hashtag is transferred from #Jan25 to #egypt, and the Arabic users increased hence leading the whole discussion, even surpassing English users. This research compares the number of tweets between different language communities, and then discovers the differences in the event's transfer of messages between different language communities, according to the characteristics of different tweeted language systems.

16.4.4.6 Text Analysis

Text analysis of traditional media is when researchers construct categories by considering human being as the coder and the page or single article as the sample for analysis (Chew and Eysenbach 2010). This kind of analysis is often called content analysis. The scale of social media data is huge to the degree that artificial analysis cannot handle it. Recently, data scientists analyze text through Natural Language Processing (NLP). Often referred to as "text analysis" or "computer-assisted text analysis," it first considers each word as the basic unit, and then undergoes the statistical analysis of word frequency or the relationship analysis of the vocabulary (Brooker et al. 2016). Chinese words are a little bit more complex than English words, because there is no interval among Chinese words. In order to do text analysis, the different types of words in one sentence must be cut, and the obsolete words must be reduced, by applying word segmentation technology (Chen and Cheng 2014; Cheng and Shih 2016). The below graphic example describes the algorithm of Chinese word segmentation which applies the surveillance pedagogy: the initial text needs to be cut and marked in the dictionary. First by noting the prefixes, mid-section suffixes, or single words, according to the position the single word has in the vocabulary; second, put into the language model for learning through the marked texts; third, provide for further analysis by outputting the text after the processing of word segmentation (Fig. 16.6).

Researchers can process vocabulary (such as word frequency, or the co-occurrence of vocabulary in certain paragraphs) through automatic software, after huge amounts of words in social media texts are segmented (Fig. 16.7).

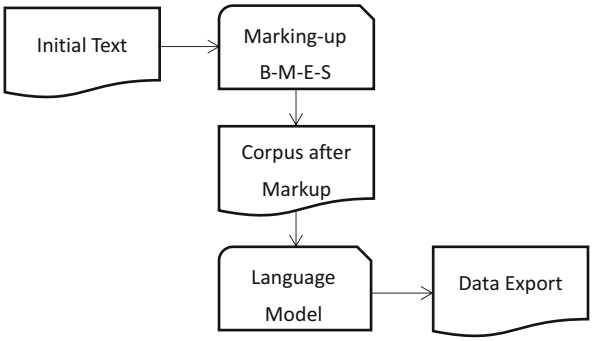

Fig. 16.6 The process of Chinese word segmentation (the word database group, academic sinica)

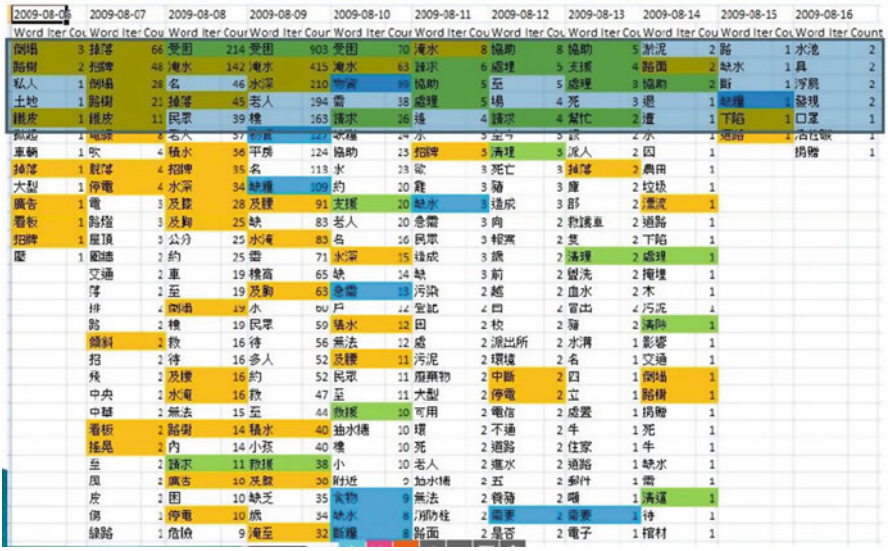

Fig. 16.7 The distribution of the time sequence of word frequency on web sites during Typhoon Morakot in 2009 (Chen and Cheng 2014)

16.4.4.7 Sentiment Analysis

Sentiment analysis can be viewed as a particular kind of text analysis. In sentiment analysis, researchers will select the target vocabulary from the text, and compare the shared sentimental traits, to judge whether the text has a positive or negative sentiment (Stieglitz et al. 2014, p. 92). In artificial or fully automatic surveillance, the sentiment analysis must undergo word segmentation first, and then classify the positive/negative categories of the word according to the pre-constructed sentiment dictionary (i.e., the categorized vocabulary groups according to the sentimental properties), or by applying methods of machine learning. Through statistics and integration, sentiment analysis can predict the text's sentimental disposition. For example, American scholar O'Connor et al. (2010) analyzed the mutual referencing

between Tweets and keywords of consumer confidence by applying Twitter data analysis methods, and comparing that with the results of surveys done by survey companies. The author analyzed the sentimental dimension of the text by using software. The research found that there is a positive correlation between the word frequency of sentimental words on Twitter and the traditionally surveyed consumer confidence index. The correlation coefficient is up to 80% on specific issues. The research shows that the traditional public survey can be potentially substituted or supplemented by the sentiment analysis of the text. Thus, the author maintains that the social media data analysis can be used to measure public opinions and predict consumer confidence.

16.4.5 Data Visualization

Data visualization is when researchers transform the data from numbers to graphic materials through software assistance, in order to display the distribution types or data trends. In social media data analysis, data visualization can be used in two situations: the first situation is when the visualization process is considered a tool for discovery. The second situation is when the visualization process is considered tool for narration, which usually takes place in the late stages of data analysis. The researchers will communicate with the audience through graphic display, and focus on the external manifestation through data visualization of the analysis. When the visualization process is used for narration, the researchers decide how the image is presented according to the data dimensions, analysis types, and graphic genre.

16.4.6 Using and Combining Types of Analysis

The relationship between the research questions (goals) and the types of data analysis (means) is not a one-to-one correlation when discovering or solving a single question. Several different analytical tools may be applied for one research. The researchers will do the discovering analysis, for example, the relatively simple sequential or numerical analysis and describe the complete figure of event time and the discussion hotspots first, converting the key strings the users form into different times in accordance with the hotspots (in order to mine the text's contents or participants), and finally apply semantic web analysis or social network analysis of the users. For instance, in the research of Tweets in the 2012 Australian presidential election, Burgess and Bruns (2012) show the complete data first, followed by further numeric analysis of the Tweets' language groups second. Cheng and Chen (2014) use the same strategy to present the complete trend using numeric analysis first, then differentiate the groups through comparative numeric analysis according to the surfacing core issues second.

16.5 Data Evolution

Data evolution refers to the process relating problem to data processing. As we mentioned above, social media data analysis is both the process from questioning to answering and the process of data analysis and data cleanup. The two processes share data in common. The mined and produced data evolve between these two processes. We can illustrate how the two processes evolve data by referring to the graph below (Fig. 16.8).

16.5.1 Data Collection

The social media data analysis usually originates from questions. In the initial stage of data analysis, researchers imagine how the data would look like, according to the posed research questions. For example, what kinds of platforms and what types of data do we need to find? What methods should we apply to mine the data? The research team gets into the collection stage of data processing while imagining it. It will set the key strings per the imagination, include the key strings, then mine the data from the social media platform through API-mining software (or set the crawling software), to obtain the raw materials.

16.5.2 Data Cleanup

The raw materials of social media platforms may have many mojibakes, data lags, or misplaced formats. The researchers need to transform these raw materials into an intuitively usable data format, for instance, transforming the codes of the character (Big5 → UTF8) or transforming time zones of Greenwich to that of Taiwan. Researchers use many ways to structuralize the datasets and convert them

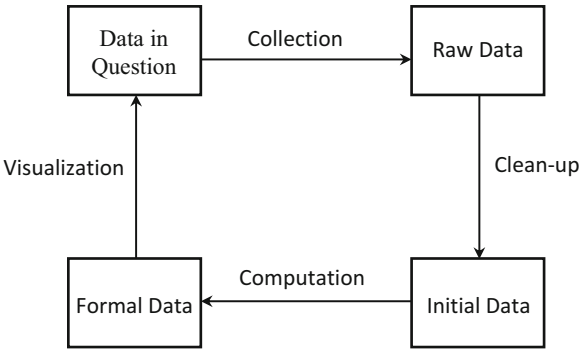

Fig. 16.8 The process of data evolution

into initial data usable for discovery. At the same time, the researchers will use simple visualized graphs to undergo data discovery analysis (e.g., the researchers understand the life pattern of a product through collating time, posted articles, and quantity of users), or use descriptive statistics to quickly find the state of data currently presented, and to predict questions such if the data is complete? Is the data clean enough? Is there any noise? Was there a problem in the data collection process? Even the question "Is the dataset useful?" is worthy of being raised.

16.5.3 The Computation/Verification of Data

The columns of the initial data still require researchers to select, integrate, or separate according to the data's dimensions. They number of variables still require filtering from researchers to be the execution target of computational calls. In the meantime, researchers will select and integrate desirable analysis types, and actually them according to the posed research questions. For example, researchers will transform the data of the reposted articles into a correlated dataset, and process the correlated data by applying the social network analysis method. Or the researchers will do one type of analysis, then do the other types of analysis from the results of the previous analysis. In other words, the analysis, as a means of discovery, is constantly occurring during the research process.

Researchers should judge if the analysis results answer the research questions after every kind of analysis has been applied. If the data can't answer the question, the researchers should evaluate and reflect upon this phenomenon in accordance with different standards, such as the research question, the data, and the results of the analysis. Researchers should also evaluate the relationship between the dataset and the posed questions, such as whether the construction of data dimensions can become its metrics, whether the construction of data dimension can be used for analysis, and whether these metrics can sufficiently answer the research questions? In other words, researchers compare the data's representation and the data's dimensions/metrics, then see if these two aspects mutually correspond.

Finally, researchers will present the data by means of visualization. Researchers select the tools and present the data comprehensively when he or she deems the data is sufficient for analysis, argumentation, and interpretation.

In the data analysis process, the discovery questions and the problem-solving process, as well as the process of data processing appear parallel. The processes are mutually related, not existing independently. Thus, researchers need to constantly care about the correlation between his or her own problematics and the data processing. This process of care may be difficult to describe in words. Therefore, we will use one data analysis case here to illustrate it.

16.6 Discussion

The contemporary social media data analysis is when researchers develop and evaluate every information tool and data structure, in order to collect, observe, analyze, note, or present social media data, in accordance with the requirements of the institution. The essence of social media data analysis is knowledge production/reproduction according to the data constructed by social media. This article aims to analyze the characteristics, elements, and process toward approaching social media data analysis. Based on bibliographical analysis and case studies, we have obtained several viewpoints for further research.

We have a few points to make before the end of the chapter.

16.6.1 Data Analysis as the Field of Activity

This field of data analysis, which contains human behavioral data on the social media platform, is not only large in scale but also diverse in form and full of noises. Thus, social media data analysis is an activity where researchers come from different knowledge backgrounds (in particular the social and data sciences) and go through data analysis in mutual cooperation and communication.

Social media data analysis is a field of transdisciplinary knowledge in which researchers come from different intellectual backgrounds, and posess different domains of knowledge. Social scientists are good at transforming social reality into problems and interpreting messages and meanings. On the other hand, data scientists are good at processing big diverse data, and converting it into forms of visual communication.

Social media data analysis is located at an intersection between social science and information science. For communication scholars, the greatest challenge in social media data analysis is converting posed problems into data processing (Brooker et al. 2016, p.1). Thus, the key is transdisciplinary teams and communication.

16.6.2 Considering the Characteristics of the Data Thinking Process

The core of social media data analysis is thinking using data. This process is a dynamic one. Data analysis is the process of problem-posing, problem-solving, and answer-seeking (Smith et al. 2014). The problems and data are continually evolving, and researchers will develop problem-solving strategies according to changes in situation, as the "Thinking in Action" Scribner (1986) described. Data analysis is not only related to the mental structure of the researchers, but also embodied

by connection between mind and body. Thus, data thinking can be viewed as the connected and collaborative process between mental state, body, and artificial products.

16.6.3 The Connection of Research Questions and Materiality of Social Media Data

Social media researchers must understand the questions communication scholars pose, and frame those questions in the context of social media data. Researchers look for the connection between questions and data, and discover a problem-solving approach. As Gibson (1979) said, what social media data analysis looks for is the affordance between the posed questions (the subjective desire of the researchers) and the situation (the materiality in the situation). Researchers not only require research questions in mind but also need to understand the affordance between software tools and data. Data processing of social media data is usually handled by software tools. Every software tool has its own materiality (e.g., researchers will present the dimensions, sequence, and amounts of data by using rows and columns), in order to form the data's characterization. Consequently, researchers must use pros and cons of the material characteristics wisely, in order to find the best problem-solving strategy.

As for educators in social media data analysis, teaching/learning in a moving context is very important. The traditional pedagogy of research methods puts emphasis on method, rule, and universality. However, the essence of research activities focuses on tacit knowledge. It does not necessarily involve oral or textual explanation. The datasets are disparate. The important thing is in all cases to find the connection between the research questions and the data's materiality, and explore the best solution within the diverse problem-solving toolbox. Therefore the application of teaching strategies such as using cases as teaching materials, learning by doing, and using real-world cases to lead student thinking is necessary.

16.6.4 The Importance of Emphasizing Explorative Data Analysis

Data analysis contains explorative and hypothetical processes. It contains the explorative connotation. Researchers use simple statistical methods and visualization tools reveal the different types. It contains the hypothetical connotation. Researchers tend to observe the data distribution types and the possible data singularities to find key questions in highly uncertain data areas.

The datasets must be explored, because every one of them is both stand-alone and similar to the other datasets occurring at the same time. In order to give researchers

an understanding of the data, and hence find the best solution to the problems, the functional scaffold allows researchers to represent different types of data before other researchers through simple descriptive statistics and visualized graphs.

However, traditional statistics (no matter if it is used for pedagogical purposes or research purposes) usually put more focus on verification process and model construction, rather than explorative analysis. Researchers need to observe the different data types of social media big data. Thus, the exploratory process is very important in social media data analysis. The question regarding the proportion that explorative data analysis has on future education is worthy of attention.

16.7 Concluding Remarks

Contemporary social media is like the "responsive system" of immediate response our contemporary society provides in the sense that it can generate very large scale, diverse, and highly dense data. On the other hand, data analysis is like the "reflective system" in the sense that researchers collect, filter, and analyze the data through meticulous and slow processes, hence representing social reality.

Social media developed at a very fast pace, while its data analysis is still in its initial stage. There is a huge gap between "fast thinking" and "slow thinking." The community of social media big data analysis is continually growing, while the reflective system of social media still lags.

Social media data analysis is an emerging field of knowledge. This field is still awaiting the inputs of different academic communities. This chapter is a preliminary analysis which attempts to draw a rough outline for social media data analysis. Owing to the format limits of this research chapter, we have selected the parts we wish to express. Complete topic coverage is impossible. We believe that the chapter would be a worthy reference for novices of social media data analysis. The goal of this chapter would be achieved if, in the future, the conversation and critique as regards social media data analysis can be advanced further.

References

Boumans, J., & Trilling, D. (2016). Taking stock of toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism, 4*(1), 8–23.

boyd, D. M., & Crawford, K. (2012). Critical questions for big data. *Information Communications Society, 15*(5), 662–679.

boyd, D. M., & Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*, 210–230.

Brooker, P., Barnett, J., & Cribbin, T. (2016). Doing social media analytics. *Big Data & Society, 3*(2), 1–12.

Brügger, N., & Finnemann, N. (2013). The web and digital humanities: Theoretical and methodological concerns. *Journal of Broadcasting & Electronic Media, 57*(1), 66–80.
Bruns, A., Highfield, T., & Burgess, J. (2013). The Arab spring and social media audiences: English and Arabic Twitter users and their networks. *American Behavioral Scientist, 57*(7), 871–898.
Burgess, J., & Bruns, A. (2012). (Not) the Twitter election: The dynamics of the #ausvotes conversation in relation to the Australian media ecology. *Journalism Practice, 6*(3), 384–402.
Burgess, J., Bruns, A., & Hjorth, L. (2013). Emerging methods for digital media research: An introduction. *Journal of Broadcasting & Electronic Media, 57*(1), 1–3.
Chen, P. L., & Cheng, Y. C. (2014). From information to social convergence: Discovering emerging channels in major disasters. *Mass Communication Research, 21*, 89–125; (in Chinese).
Cheng, Y. C. (2014). The computational turn for new media studies: Opportunities and challenge. *Communication Research and Practice, 7*, 45–61; (in Chinese).
Cheng, Y. C., & Chen, P. L. (2014). Emerging communities in social media during the 2012 Taiwanese presidential election: A big-data analysis approach. *Mass Communication Research, 120*, 121–165; (in Chinese).
Cheng, Y. C., & Shih, S. F. (2016). News sources in social media during the 2012 presidential election in Taiwan. *Chinese Journal of Communication Research, 29*, 107–133; (in Chinese).
Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS One, 5*(11). online14118.
Chiang, Y. & Lin, T. T. C. (2015). Big data in communication studies: A systematic review. In Peng, Y. (Ed.), The Proceedings of "2015 Big data, New Media & Users" Conference, Taoyun, Taiwan: Yuan-Tze University (pp. 355–368) (in Chinese).
Felt, M. (2016). Social media and the social sciences: How researchers employ big data analytics. *Big Data & Society, 3*(1), 1–15.
Forte, A., & Lampe, C. (2013). Defining, understanding, and supporting open collaboration: Lessons from the literature. *American Behavioral Scientist, 57*(5), 535–547.
Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.
Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of Communication, 64*, 239–259.
Kahneman, D. (2012). *Thinking: Fast and slow*. New York: Penguin Books.
Kogan, M., Palen, L., & Anderson, K. M. (2015). Think local, retweet global: Retweeting by the geographically-vulnerable during Hurricane Sandy. Paper presented at the *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work*, Social Computing, Vancouver, BC, Canada.
Li, B. (2011). *The rising wind of opinion mining: On spatial and temporal structures in the diffusion of online hotspot events*. Beijing: People's Daily Press; (in Chinese).
Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media, 57*(1), 20–33.
Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 460–475). Minneapolis, MN: The University of Minnesota Press.
Mayfield, R. (2006, April 27). *Power law of participation*. Ross Mayfield's Blog. Retrieved from http://ross.typepad.com/blog/2006/04/power_law_of_pa.html
O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM, 11*(122–129), 1–2.
Parks, M. R. (2014). Big data in communication research: Its contents and discontents. *Journal of Communication, 64*, 355–360.
Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.
Scribner, S. (1986). Thinking in action: some characteristics of practical thought. In R. Sternberg (Ed.), *Practical intelligence*. New York: Cambridge University Press.
Seltman, H. (2015). Exploratory data analysis. In *Experimental design and analysis* (pp. 61–98). Pittsburgh, PA: Carnegie Mellon University.

Smith, A., Molinaro, M., Lee, A., & Alberto, G. (2014). Thinking with data. *The Science Teacher, 81*(8), 58–63.

Stieglitz, S., & Dang-Xuan, L. (2012). Social media and political communication: A social media analytics framework. *Social Network Analysis and Mining, 3*(4), 1277–1291.

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social media analytics: An interdisciplinary approach and its implications for information systems. *Business & Information Systems Engineering, 6*(2), 89–96.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, Ann Arbor, Michigan, USA, June 1–4, 2014.

Tukey, J. W. (1977). Exploratory data analysis. Reading, PA: Addison-Wesley.

van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. New York: Oxford University Press.

Yang, L. W., & Shao, K. H. (2016). *Social big data: Listening & Analysis*. Taipei: Future Career Publishing; (in Chinese).

Zeng, D., Chen, H., Lusch, R., & Li, S. (2010). Social media and intelligence. *IEEE Intelligent Systems, 25*(6), 13–16.