Technische Universität München

Fakultät für Luftfahrt, Raumfahrt und Geodäsie

Lehrstuhl für Kartographie

Perception of social-event-induced human behavior from geotagged social media data

Ruoxin Zhu

Vollständiger Abdruck der von der Fakultät für Luftfahrt, Raumfahrt und Geodäsie der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr.-Ing. Uwe Stilla
Prüfer der Dissertation:	1. Prof. Dr.-Ing. Liqiu Meng
	2. Prof. Dr.-Ing. habil. Dirk Burghardt Technische Universität Dresden

Die Dissertation wurde am 09.07.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Luftfahrt, Raumfahrt und Geodäsie am 09.09.2020 angenommen.

Abstract

Understanding human behavior has always been one of the key issues in social sciences. The occurrence of various social events may have various effects. A timely perception of human behavior induced by social events is crucial for governmental agencies, enterprises, and citizens to trace the consequences of social events as well as take necessary measures to amplify/reduce the positive/negative impacts. Researchers have mainly used traditional sociological methods (e.g., interviews and sample surveys) to study the impact of social events, leading to a large number of insightful empirical findings which, however, are difficult to validate. In recent years, the rapid technological development, especially in communication technology and mobile positioning technology, has radically boosted social sensing, which in turn has induced a data-driven channel for understanding various human behaviors induced by social events.

As an effective means of social sensing, social media services are well accepted by users wanting to share their thoughts about nearby events. One of the main driving forces is the instinctive human urge to share knowledge about real-world events. The spatial-temporal and semantic information embedded in geotagged social media data provides valuable indicators of human behavior induced by social events. While current research dealing with geotagged social media data mainly focuses on event detection and tracking, this thesis sets its focus on the sensing and comprehension of human behavior induced by known social events.

Unlike the monitoring of natural environment where the observation could be continuously conducted in real-time by means of automated sensors, in social sensing systems, people acting as social sensors are characterized by their spontaneity and discontinuity. Therefore, adaptations of existing approaches and new approaches are needed to extract valuable information and mine the hidden knowledge from social media data. Combining machine learning, natural language processing, and visualization methods, this thesis put forward a generic analytical framework dedicated to geotagged social media data for the understanding of the impacts of known social events on human behavior. In addition to the interpretation of human inner behavior, based on spatial, temporal, semantic and sentimental constraints, a density-based clustering algorithm extended from DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is deployed to explore the crowd behavior. Furthermore, an approach of geospatial network analysis is presented to explore the regional crowd mobility patterns based on the changing spatiotemporal states of social media users.

The introduced ideas and methods are implemented and verified in selected case studies with two social media datasets from Sina Weibo - the largest microblogging platform in China. The impacts of selected known events, including tragic and festive events, on human behavior are addressed to demonstrate the feasibility and performance of the generic framework. Though this thesis has succeeded in its aim of perceiving human behavior induced by social events from geotagged social media data, there is still room for improvement. The follow-up research will make use of more advanced data scientific approaches, more diversified multi-source social media data and higher dimensional training datasets of longer time series to optimize the methods and expand the scope of application.

Zusammenfassung

Das Verständnis menschlichen Verhaltens war schon immer einer der Schwerpunkte in den Sozialwissenschaften. Das Auftreten verschiedener sozialer Ereignisse kann verschiedene Auswirkungen haben. Eine frühzeitige Erfassung des menschlichen Verhaltens, das durch soziale Ereignisse ausgelöst wird, ist für Regierungsbehörden, Unternehmen und Bürger von entscheidender Bedeutung, um die Folgen sozialer Ereignisse zu verstehen und die notwendigen Maßnahmen zum Verstärken/Verringern der positiven/negativen Auswirkungen zu ergreifen. Bisher haben die Forscher hauptsächlich herkömmliche soziologische Methoden (z.B. Interviews und Stichprobenerhebungen) angewandt, um die Auswirkungen sozialer Ereignisse zu untersuchen, was zu einer großen Zahl aufschlussreicher empirischer Ergebnisse geführt hat, die jedoch schwer zu validieren sind. In den letzten Jahren hat die rasante technologische Entwicklung, insbesondere in der Kommunikationstechnologie und der mobilen Positionierungstechnik, *social sensing* stark gefördert, was wiederum einen datengetriebenen Kanal für das Verständnis verschiedener menschlicher Verhaltensweisen, die durch soziale Ereignisse verursacht werden, hervorgerufen hat.

Als wirksames Mittel von *social sensing* werden Social-Media-Dienste von Nutzern, die ihre Gedanken über Ereignisse in der Nähe mitteilen wollen, gut angenommen. Eine der Hauptantriebskräfte ist der instinktive menschliche Drang, Wissen über Ereignisse in der echten Welt zu teilen. Die raum-zeitlichen und semantischen Informationen, die in geotaggten Social-Media-Daten eingebettet sind, liefern wichtige Indikatoren für menschliches Verhalten, das durch soziale Ereignisse ausgelöst wird. Während sich die aktuelle Forschung, die sich mit geotaggten sozialen Mediendaten befasst, hauptsächlich auf die Erkennung und Verfolgung von Ereignissen konzentriert, liegt der Schwerpunkt dieser Arbeit auf der Erfassung und dem Verständnis menschlichen Verhaltens, das durch bekannte soziale Ereignisse hervorgerufen wird.

Im Gegensatz zur Überwachung der natürlichen Umwelt, wo die Beobachtung mittels automatisierter Sensoren kontinuierlich und in Echtzeit durchgeführt werden könnte, zeichnen sich Menschen, die als soziale Sensoren fungieren, bei sozialen Erfassungssystemen durch ihre Spontaneität und Diskontinuität aus. Daher sind Anpassungen bestehender Ansätze und neue Herangehensweisen erforderlich, um wertvolle Informationen zu extrahieren und das verborgene Wissen aus Social-Media-Daten zu gewinnen. Durch die Kombination von maschinellem Lernen, Natural-Language-Processing und Visualisierungsmethoden wird in dieser Arbeit ein generisches analytisches Rahmenwerk vorgeschlagen, das sich dem Verständnis der Auswirkungen bekannter sozialer Ereignisse auf das menschliche Verhalten durch geogetaggte Social-Media-Daten widmet. Zusätzlich zur Interpretation des menschlichen inneren Verhaltens, basierend auf räumlichen, zeitlichen, semantischen und emotionalen Einschränkungen, wird ein dichtebasierter Clustering-Algorithmus, erweitert von DBSCAN (Density-Based Spatial Clustering of Applications with Noise), eingesetzt, um das Verhalten von Menschenmengen zu untersuchen. Darüber hinaus wird ein Ansatz der raumbezogenen Netzwerkanalyse vorgestellt, um die regionalen Mobilitätsmuster der Crowd auf der Grundlage der sich ändernden raum-zeitlichen Zustände der Social-Media-Nutzer zu untersuchen.

Die vorgestellten Ideen und Methoden werden in Fallstudien mit ausgewählten Social-Media-Datensätzen von Sina Weibo - der größten Microblogging-Plattform in China - umgesetzt und verifiziert. Die Auswirkungen ausgewählter bekannter Ereignisse, einschließlich tragischer und festlicher Ereignisse, auf das menschliche Verhalten werden angesprochen, um die Machbarkeit und Leistungsfähigkeit des generischen Rahmens zu demonstrieren. Es ist der Dissertation gelungen, das durch soziale Ereignisse induzierte menschliche Verhalten anhand von geotaggten Social-Media-Daten wahrzunehmen. Aber es gibt immer noch Raum für Verbesserungen. In der weiterführenden Forschung werden fortgeschrittene datenwissenschaftliche Ansätze, diversifiziertere Multi-Source-Social-Media-Daten und höherdimensionale Trainingsdatensätze längerer Zeitreihen genutzt, um die Methoden zu optimieren und den Anwendungsbereich zu erweitern.

Contents

Abbreviations

VSM	Vector Space Model
LSA	Latent Semantic Analysis
LDA	Latent Dirichlet Allocation
TF	Term Frequency
TF-IDF	Term Frequency–Inverse Document Frequency
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
ST-DBSCAN	Spatial Temporal-Density Based Spatial Clustering of Applications with Noise
SCSTSC	Sentiment-Constrained Spatiotemporal Semantic Clustering
NB	Naive Bayes
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
PLSA	Probabilistic Latent Semantic Analysis
BOW	Bag of Words
PMN	Population Mobility Network
RFDR	Relative Flow Difference Ratio
SFT	Spring Festival Travel
GDP	Gross Domestic Product
LBS	Location-Based Service

List of Figures

List of Tables

1 Introduction

Understanding human behavior has always been one of the key issues in social sciences. The occurrence of various social events may have various effects that cannot be ignored. A timely perception of human behavior induced by social events is crucial for governmental agencies, enterprises, and citizens to trace the consequences of social events as well as take necessary measures to amplify/reduce the positive/negative impacts. The emergence of social sensing makes it possible to perceive human behaviors induced by social events in near real-time. This thesis focuses on related methods of using geotagged social media data to perceive various human behaviors induced by social events.

1.1 Motivation

The world has witnessed a tremendous upsurge of documented and communicated social events since the late nineteenth century. A variety of social events, especially negative ones (e.g., terrorist attacks, violent incidents), have undeniably impacts on a variety of aspects regarding individuals and society at large, which are worthy of recognition as a distinct research field.

Researchers have mainly used traditional sociological methods (e.g., interviews and sample surveys) to study the impact of social events, leading to a large number of insightful empirical findings which, however, are difficult to validate. In recent years, the rapid technological development, especially in communication technology and mobile positioning technology, has radically boosted social sensing, which in turn has induced a data-driven channel for understanding social events (Goodchild & Glennon, 2010). Social sensing uses various spatial big data sources to study the characteristics of human behavior in space and time, revealing spatial and temporal distributions, connections, and processes of social-economic phenomena. Different types of big data (e.g., social media data, trajectory data, and check-in data) generated by social sensors have brought about new opportunities for the understanding of the natural and social environment we live in (Y. Liu et al., 2015). The related research includes traffic management, urban planning, environment, public health, and many further domains. The decisions supported by social sensing data are becoming more people-oriented.

As an effective means of social sensing, social media services (e.g., Facebook and Chinese Sina Weibo) are particularly popular for people wanting to share their thoughts about nearby events. One of the main driving forces is the instinctive human urge to share their knowledge of real-world events, such as a government election or a traffic jam (Valkanas & Gunopulos, 2013). These data may be extracted from social media and processed for various purposes. In particular, the spatial-temporal and semantic information embedded in geotagged social media data provides valuable indicators of human behavior induced by social events.

While current research dealing with geotagged social media data mainly focuses on event detection and tracking, this thesis shifts the attention from event-oriented monitoring to the sensing and comprehension of human behavior induced by social events.

1.2 Research tasks

Human behavior can be classified as outer and inner behavior (Duncan, 1971). The outer behavior is everything we can externalize so that an ordinary observer could perceive, be it scratching one’s head, typing on a keyboard, running across a field, or talking to a friend. The inner behavior involves the internal domain of activity within us such as thoughts and feelings that an ordinary observer cannot perceive. The access to inner behavior is not straightforward and has to rely on the externalized behavior.

The temporal-spatial human behavior caused by social events can also be inner and outer nature. Perceptions of social events, attitudes and emotional changes caused by social events, for example, are typical inner behavior, while geospatial movements and activities of humans caused by social events are typical outer behavior. Geotagged social media data contain rich spatial-temporal and semantic information. They allow us to explore the behavioral characteristics of the crowd and the related social phenomena. The framework in Figure 1.1 gives an overview of the research contents in this thesis concerning the perception of human behavior induced by social events from geotagged social media data.

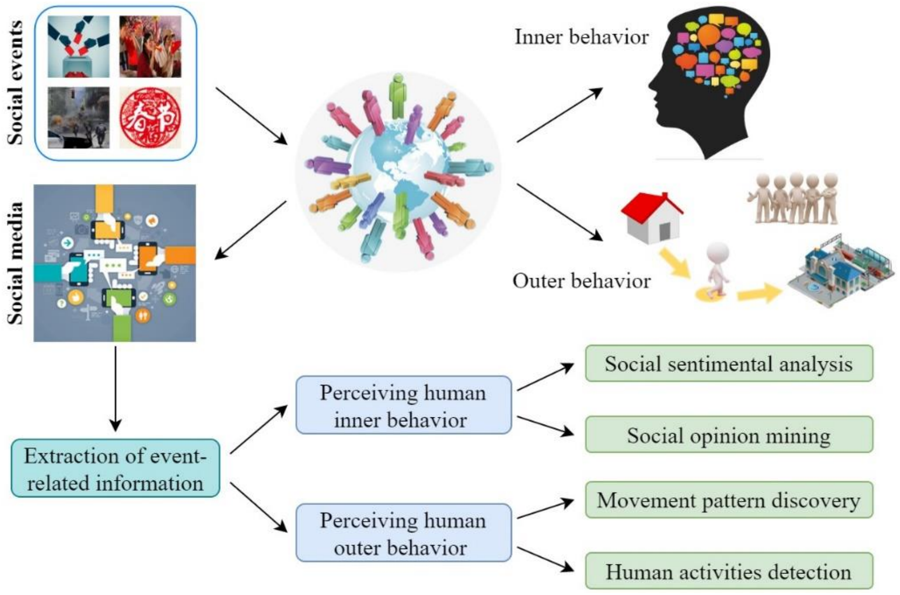

Social events
Social media
Inner behavior
Outer behavior
Extraction of event-related information
Perceiving human inner behavior
Perceiving human outer behavior
Social sentimental analysis
Social opinion mining
Movement pattern discovery
Human activities detection

Figure 1.1 Research contents of perceiving human behavior induced by social events from geotagged social media data.

Using crowd-sourcing, users have contributed to social sensing, which simultaneously comes with an inevitable challenge (Ali et al., 2011). Unlike natural environment monitoring where the expected data can be delivered in real-time through automated sensors, in social sensing systems, people acting as social sensors are less controlled. Therefore, different approaches are needed to extract valuable information from social sensing data and mine the hidden knowledge. More concretely, we have the following research tasks:

- **Development of approaches for the extraction of event-related information from geotagged social media.** Extracting unlabeled event-related information from geotagged social media datasets could be formulated as a binary classification problem and solved with machine learning methods. The construction of training data, feature selection, and the construction and evaluation of classifiers require a targeted design.
- **Detection of human inner behavior caused by a known social event from geotagged social media data**. Social events may affect people's inner behavior in terms of emotions and cognitive processes. Elaborated methods of social sentiment analysis and social opinion mining are required to explore the impacts of social events on human inner behavior.
- **Detection of human crowding behavior induced by known/unknown social event from geotagged social media data**. Social events may provoke people's crowding behavior, which is reflected in geotagged social media data as a high concentration of similar messages within a certain period in a specific area. Spatiotemporal semantic clustering algorithms combined with sentimental constraints are required to explore clues and related crowding behaviors in geotagged social media data.
- **Detection of movement patterns of population induced by a known social event from geotagged social media data**. Due to their powerful social impact, some social events can even drive people to move across regions. It is necessary to develop a feasible method to recognize people's cross-regional movement from social media data, and then to recognize changes in human space-time behavior patterns induced by social events.

1.3 Thesis structure

The thesis is structured in six chapters.

Chapter 2 is dedicated to the theoretical and practical foundations of research on human behavior and social events and summarizes the research progress in understanding human behavior caused by social events. **Chapter 3** addresses the application potential of social sensing in analyzing human behaviors induced by social events. In particular, the author introduced in detail the perception capabilities based on social media, especially the characteristics of geotagged social media data and the related scientific values. The author summarizes the research progresses with regard to the perception of human behavior induced by social events from geotagged social media data.

A methodological framework for sensing human behavior induced by social events from geotagged social media data is put forward in **Chapter 4**. It is composed of three analytical modules of geotagged social media data. These modules are used to perceive from spatial, temporal and semantical perspectives the inner behavior, crowded behavior, and changes in population mobility induced by social events. The detailed implementations with a number of experiments on selected case studies are elicited and evaluated in **Chapter 5**. Finally, the major findings of the thesis and envisions of further developments are summarized in **Chapter 6**.

2 Analysis of human behavior caused by social events

This chapter introduces research works on human behavior and social events, with the aim to analyze the possible impacts of social events on human behavior. Section 2.1 introduces necessary definitions, influencing factors, observation methods, and research values related to human behavior. Section 2.2 sets a focus on event research, starting with social events along with the definitions and research progresses. Section 2.3 discusses the importance of studying human behavior caused by social events and summarizes current research progresses.

2.1 Human behavior

Behavioral science as a relatively modern branch of science deals primarily with human action and often seeks to generalize human behavior in society (Merriam-Webster dictionary). Its research object involves a wide range of topics including thinking process, communication, consumer behavior, business behavior, and social change. The application scope of behavioral science involves almost all fields of human activities, forming many branch disciplines, such as organizational management behavior, medical behavior, criminal behavior, political behavior, organizational behavior, and so on.

2.1.1 Basic concepts of human behavior

Human behavior is a response of individuals or groups of humans to internal and external stimuli. In scientific research, human behavior can be classified as outer and inner behavior (Duncan, 1971). Outer behavior is everything we do that an ordinary observer (e.g., bare eyes, physiological sensors) could perceive, such as running across a field, swimming in the pool, typing on a keyboard, fishing by the river, or talking to a friend. Behavioral actions can take place on various time scales, ranging from sweat gland activity to sleep or food consumption. On the other hand, inner behavior involves the domain of activity within us that an ordinary observer cannot perceive and includes the person’s cognition, and emotions.

Cognitions describe thoughts and mental images you carry with you, and they can be both verbal and nonverbal. “I need to submit my conference paper next week before the deadline” or “I would like to know what the reviewers think of my paper” can be considered as verbal cognitions. In contrast, imagining the scene of your oral presentation in the upcoming meeting is a nonverbal cognition. Cognitions comprise knowledge and skills, such as knowing how to swim, the ability to program in Python, and mastering English spelling. And emotion refers to any relatively short-lived conscious experience characterized by a mental activity and is not characterized as a result of either reasoning or knowledge. Usually, emotions exist on a scale ranging from positive to negative.

Outer behavior and inner behavior do not operate independently of each other. The interaction between outer and inner behaviors can alternately respond to external and internal stimuli. When you see a friend you haven't seen for a long time, it may lead to a sudden increase in joy, accompanied by a corresponding recognition. When you receive an email that your paper is rejected, it may cause sad emotions and a series of reflections, and then you may go to sports to release your emotions or revise the content of the paper according to the review comments.

2.1.2 Classification of human behavior

Research on human behavior addresses how and why people behave in their own way. However, human behavior is very complicated because it is affected by many factors. Human behavior can be classified and analyzed from multiple perspectives.

- Instinct behavior and social behavior. Instinct behavior refers to the typical, stereotyped behavior pattern that is inherited innately and can appear without learning, and it is purposeful directional behavior, such as eating, drinking, defense, sex, sleep, and maternal behavior. According to the nature of social behavior, it can also be divided into economic behavior, political behavior, legal behavior, cultural behavior, and ideological behavior, etc.

- Conscious behavior, subconscious behavior, and unconscious behavior. Consciousness is the sentience or awareness of internal or external existence. Conscious behavior in psychology derives from conscious thinking processes of which we are actively aware. It is any behavior that is completed willingly with intent derived from your own volition. Subconscious behavior is driven by the subconscious mind, while unconscious behavior is affected by the unconscious mind.

- Overt behavior and covert behavior. Overt behavior refers to behaviors that can be observed, such as running movements, changes in facial expressions. Covert behaviors are unobservable actions which can only be deduced by oneself, such as thoughts and feelings.

- Rational behavior and irrational behavior. Rational behavior is goal-oriented. It refers to the decision-making process that results in the optimal level of benefit or utility for an individual. It is predictable, sensible, and logical. In contrast, irrational behavior is neither objectively logical nor in accordance with human preferences and priorities.

- Voluntary behavior and involuntary behavior. Voluntary behavior is the behavior that happens as a result of conscious choice. Involuntary behavior is the behavior that loses control or subconsciously without realizing what you are doing.

Many human behaviors are rational, overt, and conscious, but there are still many human behaviors that are covert, involuntary, irrational, or driven by the subconscious mind.

2.1.3 Main factors affecting human behavior

Human behavior is mainly affected by the interaction among four factors of heredity, maturity, environment, and learning to develop into a unique behavior pattern for individuals.

- **Heredity.** Heredity refers to the biological process in which the morphological physiological, psychological, and behavioral characteristics of the parents can be genetically passed on to children. The theory of hereditary determination emphasizes the role of genetics in the development of human behavior (Galton, 1869), according to which the cognitive development is determined by innate genetic genes. While the intrinsic genetic factors influence the development of human behavior, the role of the environment is to initiate, promote, or delay the realization of this process.
- **Environment**. The environment refers to everything that an individual grows in time and space and affects his or her physical and mental development. The intrauterine environment of the fetus, the family environment after birth, and the social and natural environment can all affect the development of human behavior. The theory of environmental determination holds that the development of human behavior is mainly a passive result of the external environment (Horowitz, 1992). Similar to changes in the growth of animals and plants, the human behavior pattern is also greatly affected by the environment. Through environmental training and professional learning, everyone may become a professional.
- **Maturation**. Maturity refers to the individual’s physiological tissue structure and function and various instinctive behaviors developed step by step according to the time sequence of genetic expression. Gesell’s Maturational-developmental theory holds that genetics plays a major role in the development of human behavior and emphasizes the importance of maturity for the development of human behavior (Herman, 2001).
- **Learning**. Learning is any relatively permanent change in behavior that results from past experiences (Busemeyer & Myung, 1992). For example, learning knowledge, skills, laws, and moral qualities at school all serves the purpose of adaptation to society. People also need to constantly innovate and develop, which again has an important relationship with learning. Therefore, learning plays an important and decisive role in the formation of many human social behaviors.

The development of human behavior is mainly the result of the interaction between inheritance and the environment. Heredity brings individual potential, and the environment assists the development of potential. Heredity has a greater impact on body structure and special talents, while the environment has a greater impact on the development of language, interests, emotions, and other psychological characteristics. Maturation can be seen as an expression of an individual's physical and psychological genetic talent. It is the foundation of learning which in turn is the manifestation of maturity. At an early age, the development of individual behavior is greatly affected by maturation factors; after maturation, it is greatly affected by learning.

2.1.4 Measuring human behavior

In order to describe and explain human behavior, researchers have designed many methods and various devices to capture human behavior characteristics (Bernstein & Livingston, 1982; Liégeois et al., 2019). Common observation methods include:

- **Behavioral observation**. Behavior observation is one of the most traditional methods of psychology research on human behavior. Researchers can either go to the field, observing the behavior of people in the natural environment, or invite individuals or groups to conduct laboratory observations.

- **Survey**. The survey is a suitable method to collect and analyze the self-reported behaviors, emotional states, and personality characteristics of respondents.

- **Eye-tracking**. It refers to tracking eye movement by measuring the position of the eye's fixation point or the movement of the eyeball relative to the head. This is a more objective way of studying an individual's visual attention and can be used to infer mental activities from eye movements.

- **Electroencephalography**. Electroencephalography is a neuroimaging technology that uses sensors and amplifier systems to measure the electrical activity generated by the brain from the surface of the scalp. It helps researchers to infer perceptual, cognitive, and emotional processes from brain activities.

- **Magnetic resonance imaging**. Magnetic resonance imaging can achieve brain imaging with higher spatial resolutions and therefore can be used to monitor the size, shape, and activity of specific parts of the brain to explore human cognitive processes.

- **Electrodermal activity**. Electrodermal activity is a property of the human body that causes constant changes in the electrical characteristics of the skin. Skin conductance is an indicator of psychological or physical stimulation. When emotionally stimulated, it can cause emotional sweating, especially on the forehead, hands, and feet. In this way, the skin conductance can be used as a measure of emotional and sympathetic responses.

- **Electrocardiogram**. The electrocardiogram can be used to monitor people's heart rate and pulse, so as to gain a deeper understanding of the respondent's physical state (e.g., anxiety and stress levels) and to explore the relationship between human physiological changes and human behavior.

The data processing techniques can be either qualitative or quantitative:

- **The qualitative study** aims to process non-numerical data for meaning-making. For example, open questionnaires and unstructured interviews may allow us to explore people's views on specific social events. The qualitative field research provides another possibility to observe people's responses to changes in the surrounding environment and explore the underlying reasons.
- **The quantitative study** describes and analyzes human behavior through mathematical, statistical, or computational techniques based on quantifiable data. Typical quantitative techniques include structured surveys and quantitative output obtained through various physical sensor observations (e.g., eye tracker, smartwatch).

Traditional survey methods (such as questionnaires, face-to-face communication) can help understand people's consciousness and emotions. With the advancement of science and technology, various types of biosensors and measurement equipment have emerged and made it possible to further explore the deep consciousness of human beings and understand how the mind, emotion, brain, and body interact.

2.1.5 Application fields

Human behavior research is relevant for all applications where people are involved as subjects or objects of study. Here are some examples:

- Criminal behavior research. Criminology is an area that necessitates the study of physical or psychological characteristics of people with the purpose to identify offenders. Studies in this area may be focused on criminal anthropology, criminal physiology, criminal psychology, etc. It is also possible to study a large number of crime phenomena and analyze what kind of social environmental conditions may cause crime to occur. In addition, the cause of crime can also be derived from the interactions of various factors in the personal and social environment.
- Understanding consumer behavior. Consumer behavior is an area concerning the characteristics of psychological activities and behavioral rules of consumers in the process of acquiring, using, consuming, and disposing of products and services. Understanding consumer behavior is essential from the perspective of marketing in order to design products that can optimally match consumer behavior.
- Exploring behavioral medicine. Behavioral medicine is a combination of behavioral science and medicine. It studies health-related knowledge and technologies in behavioral sciences and applies related knowledge and technologies to the subject areas of disease prevention, diagnosis, treatment, and rehabilitation. The focus is set on behaviors that are closely related to human health, so as to guide people to establish healthy behaviors, correct abnormal behaviors, and change unreasonable lifestyles and bad habits.

- Management optimization. Management psychology addresses the laws of human behavior and psychological activity in organizations. The main task is to explore the psychological basis for improving management and seek various ways and methods to inspire people, enhance their enthusiasm and creativity, and improve their productivity.

Human behavior is obviously multifaceted and dynamically changing phenomenon. Understanding human behavior is a difficult task, requiring multidisciplinary scientific observation and collaboration.

2.2 Social events

2.2.1 Event and event study

Event is perhaps the most extensive information container for dynamic geo-historical phenomena. Societal structures are both bearers and products of events through which they are consolidated or reformed. Events are pervasive in various fields, such as the political field, military field, economic field, etc. Events can occur in people's social life, or they can come from sudden changes in nature, such as natural disasters and climate change. Various events (planned event/ unplanned event, social event/ natural event) enrich our lives and influence our lives. Understanding and explaining an event is therefore a cognitive process to specify what structural change an event brings about, and to determine how the change took place.

Due to the universality of its existence, the study of events has attracted the attention of worldwide researchers. However, different scientists may have different interpretations of events because of different backgrounds or different perspectives of thinking. It may not be possible to reach a common core notion at the current stage, but they share an invariant common core of characteristic features (Nina, 2015). Three core points for events are worth pointing out: an occurrence at a given place and time; a special set of circumstances; a noteworthy occurrence. Every event has some objects (e.g. participants) and involves some relationships (e.g. causal relationships) with other events. Therefore, in order to explain any event well enough, we should take into account its objective and results, its individual participants, its position in space and time, and its relationships to various other events.

The event-oriented study is somewhat all-encompassing and interdisciplinary. It draws from a large number of foundation disciplines (e.g., anthropology, sociology, philosophy, psychology, management, economics, political science, human geography) and closely related professional fields (e.g. leisure studies, tourism studies, education, cultural management, sport management, theatre studies, health studies, urban and community studies, rural studies, multicultural studies). The interconnections between these areas of study can help us better understand why they exist/occur, and how we can manage them better to derive positive outcomes as well as minimizing undesirable and unforeseen consequences.

As a research area of growth, event studies encompass the planning and management, outcomes, the experience of events and meanings attached to them, and all the dynamic processes shaping events and the reasons why people attend them. Event studies as a field of study are defined by its holistic approach towards events as a phenomenon, including all those issues surrounding events. In other words, event studies can exist without event management, and in fact it already does. When an economics researcher or sociologist examines the impacts of an event, regardless of any interest in its planning or production, that is an approach to event studies.

2.2.2 Social events

Social events happen every day with a local, regional or global reach. The Brexit, the Fire at Notre-Dame, the #MeToo Movement, the Murder of Jamal Khashoggi, Donald Trump's Election of U.S. Presidency, the Malaysia Airlines Flight 370 Disappears, the Boston Marathon bombings, are some prominent events in the recent years. These social events have a wide impact on various fields and aroused widespread concern, urging the research world to conduct theoretical and empirical studies.

Events can be divided into natural and social events. While events that occur in the natural environment are not directly induced by people, events occur in the human environment contain people as the core element of society. Human participation is a necessary condition for social events. In addition, if what happened only involves the individual but does not affect others, such as a person meditating, it cannot be called a social event. That means social events must involve human behavior. It occurs under certain spatiotemporal conditions and has an impact on others. In summary, a social event is an event that occurs in a specific social environment and involves human participation and has a significant social impact. The characteristics of social events are as follows:

- **Human participation**. The main body of social events is human. Among the various elements of social events, there must be elements that involve related groups of people. These elements show the interconnectedness between events and people. In addition, it also involved human subjectivity. The understanding of the same event may be similar or different among different individuals. Likewise, human behavior caused by the same event may also be different.

- **Causality**. The occurrence of social events may be affected and induced accidentally or predetermined by a variety of factors. At the same time, the impact of social events may become the cause of other events.

- **Complex**. The complexity of social events is correlated with their causality. Each effect may have multiple causes which may be both natural and artificial. The existence of multiple causes makes it difficult for the government and the public to recognize social events. In addition, the evolution of social events is complicated.

After the occurrence of social events, there may be more factors involved, forming a “domino effect” or “fission effect”. The integration of more factors not only promotes the evolution of social events but also increases the complexity of social events. The rapid spread of social events can easily lead to other problems, making the originally complicated events more complicated under the condition of continuous fission or chain reaction.

Social events have various forms and types. In order to improve the understanding of social events, we can classify social events from the aspects of nature, causes, and time of occurrence.

- Nature - Social events can have a positive or negative nature. Positive social events can have a beneficial social impact and contribute to the sustainable development of society. For example, Ethiopia and Eritrea signed a peace deal in 2018, which may facilitate regional peace and regional economic development. A tree-planting day is another statutory festival which some countries set in order to awaken people’s passions for planting and protecting trees, increasing the forest area and protecting the natural living environment. On the contrary, negative social events will have unfavorable social impacts and are not conducive to social development. A typical case is the 9/11 attacks - a series of suicide terrorist attacks in the United States on September 11, 2001. The 9/11 attacks caused serious health and psychological trauma to the victims, rescuers and their relatives, and a serious damage to the world economy.
- Causes - Social events can be induced by natural factors, such as the Indian Ocean earthquake and tsunami 2004 and the Sichuan earthquake 2008. Although they are induced by natural factors, these disasters have a non-negligible impact on people, which in turn induced a series of social responses. Social events can be also directly caused by human factors, which may then lead to considerable consequences, the aforementioned 9/11 attacks and the 2020 United States presidential election are just two examples.
- Time of occurrence - Social events can occur at different time points along a temporal scale of a certain granularity. Events that occurred in pre-modern, early modern and later modern times can be called historical events, such as the Protestant Reformation, the American Revolution, the Battle of Waterloo and the World War I. Events occurring in modern society can be defined as contemporary events, such as the Korean War, the Vietnam War and the Persian Gulf War.

2.2.3 Research on social events

Research on social events usually focuses on four aspects: the cause, the tracking methods, the impact on human mental health, economy, policy, etc., and the lessons learned from social events in management and prevention aspects. Here are some examples:

- Research on the causes of social events plays an important role in emergency response, social management and even international relations. At present, Coronavirus disease is spreading around the world, which poses a huge threat to the health of people and seriously affects the daily lives of residents. Various countries have formulated different response measures, which have also induced a series of different reactions among residents (such as protest march, gathering). Coronavirus disease has also challenged the capacity of medical care and evolved into a serious social event that has caused global attention with a huge impact on human society. Various hypotheses and speculations involving conspiracy theories, leakage of biological and chemical weapons, and technological warfare etc. are virally spread. However, as of now, there is no conclusion about the origin of the COVID-19 virus, and related research work is still in progress. For example, Forster, Forster, Renfrew, & Forster (2020) distinguished the most common type of coronavirus in different areas based on amino acid changes. They found out that three central variants (A, B, C), the proportions of Type A and C in Europe and America are relatively significant, and Type B is popular in East Asia. Type A has the closest relationship with viruses found in bats and pangolins, and it may be the reason of the COVID-19 outbreak, Type A undergoes two mutations to produce B, and C mutated from B by one variation.

- Research on the tracking method of event evolution is helpful for timely predictive analysis and emergency response, especially in terms of taking measures to avoid potential negative impacts. Hall, Tinati, & Jennings (2018) analyzed the role of social media in political events such as the Brexit and US presidential election. It shows that social media may better predict the outcome of two political events than traditional polling and political forecasting. Qian, Zhang, Xu, & Shao (2016) proposed a multi-modal event topic model to explore the interesting events from massive social media data and then employ an incremental updating strategy to track and get detailed temporal evolution of social events. This framework was conducted on two social events (i.e. Occupy Wall Street, United States Presidential Election), showing a better performance than the traditional models.

- The influence of social events has always been the focus of scholars in various fields. For example, Winter et al. (2016) explored the methods to assess the economic loss caused by four selected Scottish landslide events and a flood event during 2004-2007. This method can be used to analyze the direct economic loss and direct consequential economic loss caused by these disasters. However, the indirect consequential economic costs of both natural disasters are hard to estimate and require further research. Akresh, Verwimp, & Bundervoet (2011) examined the impacts of crop failure in 1989 and the armed conflict during 1990-1991 in Rwanda on children's health. Their study shows that the body heights of both girls and boys born during the conflict are lower than the standard. On the other hand, only girls are adversely affected by poor harvests, and this condition is more serious in poor families.

- Learning based on social events is a very critical part in the study of social events. We need to learn from the causes and consequences of social events and optimize management plans. For example, a fire broke out in the Notre-Dame de Paris in the afternoon of April 15, 2019, the roof was destroyed, and a firefighter was seriously injured while fighting fire. This event caused the attention of people all over the world. Political figures from various countries expressed their condolences to French people, and many entrepreneurs also donated money for the reconstruction of the Notre-Dame de Paris. Investigators speculated that the cause of the fire may be traced back to electrical failure or cigarettes. The painful lesson of the Notre-Dame de Paris fire led to enhanced awareness of the scientific maintenance of historical buildings, such as adding fire sprinkler system (Tannous, 2019).

Social events may have various effects on economy, social stability, human behavior, management, etc. In recent decades, a widespread attention has been drawn to human behavior influenced by social events, for example, population migration induced by wars, and the physical and mental trauma caused by terrorist attacks. Therefore, understanding the dynamic impact of social events on human behavior can help governments, communities, and individuals to take timely and effective measures to enhance beneficial effects and reduce harmful effects.

2.3 Analysis of human behavior caused by social events

A variety of social events, especially the unfavorable ones (e.g., terrorist attacks, violent incidents), have long-lasting impacts on individuals and society at large in many aspects. It is therefore necessary to take a closer look at them. This section introduces the research progress made by sociologists in applying various sociological methods to study the impacts of different social events.

- Sports events - Ohmann, Jones, & Wilkes (2006) studied the impact of the 2006 Football World Cup on the local residents through face-to-face interviews with 132 people. Most of the interviewees believed that the 2006 Football World Cup had a positive impact, especially in terms of urban renewal, increased security, positive fan behavior, and the general atmosphere surrounding the event. Based on a questionnaire approach, Fredline & Faulkner (2001, 2017) used logistic regression analysis to interpret the influence indicators on residents' different attitudes towards local motor-racing events. The results show that most local residents support the two activities to continue in their current locations. However, the noise and traffic problems caused were the main reasons why some residents opposed the incident. The relocation strategy and compensation strategy provide a reference for targeted reduction of the induced negative effects. Using a database of social conflicts, Moreno (2017) analyzed the short-term causal impact of sports events on social unrest events in Africa and found that victories have led to reduced social unrest events whereas defeats not. Moreover, the influence of victory is related to the degree of ethnical division of the country and the degree of political autocracy.

For countries with more ethnical groups and less autocracy, the role of victory is more obvious. The analysis proves that victory can promote national unity, and the role of emotional change needs further observation.

- Sexual abuse - The #MeToo movement is a typical spontaneous movement aimed at combating sexual harassment and assault against women, and it was spread widely after the sexual abuse cases of Harvey Weinstein. O'Neil, Sojo, Fileborn, Scovelle, & Milner (2018) commented on the possible effect of movement on public health based on some social statistics. This research points out that revealing inequalities in social and economic aspects that lead to sexual harassment is essential for the improvement of women's health conditions. And sexual harassment and the accompanying stress that leads to chronic disease may likely decrease as generations become more accustomed to gender equality. Gender equality helps more women participate in political activities and bring down the rates of depression and post-traumatic stress disorder, even population mortality. But limited evidence has yet proven that the campaign of preventing sexual harassment in the workplace has effectively affected health outcomes.

- Regional conflicts – Conflicts may destroy the local ecological environment and economy, and seriously damage human physical and mental health. Considering political terror ratings and intensity of traumatic events, Charlson et al. (2012) established a model based on previous review and meta-regression, and this model was used to estimate post-conflict prevalence in population affected by the 2011 Libya Conflict in six "medium intensity conflict" areas: Misrata, Benghazi, Zintan, Tripoli/Zlitan, Misrata and Ras Jdir camps. The population suffering from post-traumatic stress disorder and depression prevalence was evaluated by employing a benchmark model usually used in low- and middle-income countries. The cases of severe post-traumatic stress disorder and depression prevalence caused by the conflict were about 123,400 and 228,100, respectively, while the total population in the whole research area is just 1,236,600. And the rate of the post-traumatic stress disorder cases comorbidity with severe depression was 50%. To stabilize these prevalence, 154 full-time staff would be needed and the security environment should be restored. Charlson et al. (2019) also estimated the prevalence of mental disorders in the emergency settings by the systematic review and applying Bayesian meta-regression techniques. 22.1% people in post-conflict settings suffer from mental disorders, 9% has moderate to severe mental disorder. These new numbers reveal that previous estimates have underestimated the impact of conflicts on human mental health.

- Political aspects - Healy, Malhotra, & Hyunjung (2010) collected original survey data during basketball championships to study the impact of irrelevant events on voters' evaluations of government performance. And the experiments have shown that personal well-being may influence voting decisions at the subconscious level. Dhingra, Ottaviano, Sampson, & Reenen (2016) evaluated foreign investment in the UK affected by the Brexit by making use of a statistical model. It shows there may be an unfavorable impact: the foreign investment to the UK after the Brexit will reduce by appropriate 22%, which will destroy UK productivity and contributes to a fall in real income of 3.4%. Case studies in the car and financial industries also indicate the Brexit would reduce EU-related output of goods and services, and it seems unlikely for the UK to strike large trades with non-EU countries. Portes (2016) analyzed the short and long-term migration flow and migration policy affected by the UK referendum. After Bretix, there will be a sharp fall in the net migration from other European Union countries due to the continuous slow employment growth in the UK, negative impact of referendum on economic growth and output and some legal and psychological reasons. For example, migrants who are sensitive to the exchange rates may prefer to stay at hometown where they can earn more, and EEA citizens may not have much confidence in their future in the UK and contemplate to move to other EU countries or return to their hometowns. These negative effects can likely lead to a reduction of skilled workers and an increase in the burden of enterprise management, which are all problems that policymakers cannot easily solve. An analysis based on the annual population survey of private households in the UK shows that immigration from the other EU countries to the UK increased rapidly before the 2016 referendum, however, the number fell sharply from mid-2016 (Vargas-Silva & Fernández-Reino, 2019).

- Public safety - Smith, Rasinski, & Toce (2001) analyzed public responses to the 9/11 attacks through a random telephone survey of 2126 U.S. residents. The results showed that the 9/11 attacks had a profound impact on Americans. This terrorist attack did not undermine the confidence of the American people in the country. On the contrary, anger was the most profound response. In addition, some interviewees were concerned about their future and safety, as well as national development. Many people's physical and mental health has also been affected. Arvanitidis, Economou, & Kollias (2016) studied the impact of terrorist attacks on citizens' risk perceptions based on European social surveys and showed that terrorist incidents can affect people's trust in the government, but such impacts are short-lived and will soon disappear. Ebola virus disease was first discovered in 1976 and outbroke widely in West Africa during 2014–2016 with high fatality rate. According to the report of WHO, this outbreak resulted in more than 11000 deaths among infected 28652 cases. O'Leary, Jalloh & Neria (2018) investigated the Ebola epidemic impact on human mental health driven under fear and culture by reviewing previous studies. They indicated that fear-related behaviors and stigmatization can negatively affect the mental health of infectors, and cultural dynamics may benefit the people in the study area, but those related to the infected person may need a lot of psychological support, and emotions caused by traumatic experiences can be fragmented and "cooled down". The mental health caused by Ebola is not only an issue related to survivors, health workers, but also hits the general population in the affected area. Jalloh et al. (2018) assessed the Ebola impact on the mental health of the general population in Sierra Leone for the time period between 2014 and 2016 by administering a cross-sectional survey based on multi-staged cluster sampling. As a result, people with depression and post-traumatic stress disorder were very common one year after Ebola outbreak, undermining the importance of handling the mental health during the epidemic period of an infectious disease.

- Natural disasters - As one of the most severe disasters, the Tōhoku earthquake with a magnitude nine happened in the Pacific coast of Tōhoku 2011, followed by a tsunami and the Fukushima nuclear accident. The sequence of disasters killed more than 15,000 people, 120,000 people became homeless and 80,000 residents were evacuated. Worse still, people also suffered from unsafe water and unpredictable radiation threat in subsequent years. To access the impact of this earthquake on marriages, births and the secondary sex ratio in the whole Japan, Hamamatsu, Inoue, Watanabe, & Umezaki (2014) used a quadratic regression equation to fit the number of marriages and births before the disaster period and then compared the observed figures with the predicted number of two categories under 95% confidence limits during the post-disaster. In this research recorded male birth ratio is contrasted with empirically estimated 95% confidence interval of a binomial distribution. The result shows all the three indices decreased after the destructive Tōhoku earthquake. In 2019, a fire happened to Notre-Dame de Paris and brought devastating damage to art and built material which was estimated to be billions of dollars. Tannous (2019) surveyed people from different fields of expertise like structural engineers and heritage specialists to estimate the immediate, short- and long-term consequences of the fire. The economic loss is significant, and it will long last for residents around and near the church, and even for the whole France.

These examples of social events and their impacts on human behaviors are typically studied using traditional sociological methods (e.g., interviews, questionnaires and surveys, and documents and records). In recent years, there is an increasing development towards the analysis of human behavior driven by multiple social data sources. For example, cell phone mobility data was used to explore the relationship between various types of social events and participants' origins (Calabrese, Pereira, Lorenzo, Liu, & Ratti, 2010). Various types of network communication structures were used to detect political abuse for special political events in social media (Ratkiewicz et al., 2011). With regard to political opinion mining, an opinion diffusion model was constructed to analyze and predict public reactions based on crowd social media content (Sobkowicz, Kaschesky, & Bouchard, 2012). Overall, current sociological studies on the human behaviors caused by social events mainly rely on sample surveys and in-depth interviews. It is irrefutable that the sample survey has been the core methodology of sociology. However, the intensified cooperation among sociology, computer science, and linguistics will empower us with new data and methods for the recognition of human behavior induced by social events (Savage & Burrows, 2007). On the basis of widely used social media platforms, we can gather larger amounts of geotagged social media data and detect human behaviors embedded in various social events.

3 Social sensing as a data-driven approach to perceive human behavior induced by social events

As can be seen from the introduction in section 2.3, human behaviors caused by social events have been mainly studied using traditional sociological methods. However, the development of science and technology has enabled the continuous emergence of social sensing data. We have now more options to understand human behavior. Social sensing can provide data-driven field study for the perception of human behavior induced by social events, which is addressed in this chapter. Section 3.1 introduces the origin, concept and main research contents of social sensing. Section 3.2 provides more details of perception capabilities based on social media along with the characteristics and application potential of geotagged social media data. Section 3.3 summarizes the research progresses on the perception of human behavior induced by social events in geotagged social media data.

3.1 Social sensing

The advent of the information age has considerably enriched the human capacity of acquiring, sharing and processing the data. How to make use of the data for a better understanding of human behavior and for the improvement of social services is an important issue currently involving interdisciplinary fields such as sociology, computer science, and geography. Social sensing is a good starting point. It aims to capture relevant clues about human behavior in real-time based on large-scale and multiple types of sensing devices deployed in human living spaces.

3.1.1 The origin of social sensing

Human beings have never stopped investigating themselves and the society in which they live. The recognition of human behavior has always been a fundamental research issue of many scientific communities. Initially, researchers collected data on human activities and social phenomena through traditional social survey methods such as written questionnaires and interviews, and used statistical induction methods to understand the laws and characteristics of human activities (Lameck, 2013). However, the survey data is hardly representative due to the limited sample size. In addition, personal preferences are embedded in the survey data, making it difficult to gain objective results.

The emergence of scientific computers has enabled efficient computing and simulation of human behavior and social phenomena. The agent-based social simulation is one of the popular methods, according to which an agent-based model is established to simulate social phenomena (X. Li, Mao, Zeng, & Wang, 2008). Different elements (persons) of a social system are represented by autonomous agents, and placed in a simulated society within which the actions of the agents are monitored. Specifically, the agent-based social simulation method does not require survey data. Instead, it sets initial conditions and interaction rules for each agent, and simulates real-world social interactions through the interactions between agents. The design of initial conditions and interaction rules relies on the support of classical theories and assumptions.

The development of Internet technology and the popularization of network applications have provided researchers with massive data resources for analyzing online behavior and social networks. Meanwhile, social computing is booming, and various social applications, such as mining of behavior patterns of Internet users, analysis of online public opinion, and online counterterrorism, are constantly emerging. Compared to earlier social survey methods, Internet-based research methods address large-scale data. However, online behavior is a virtual behavior, which is fundamentally different from human behavior in the real world.

The steadily progressing information technology has considerably changed the human lifestyle. In the past ten years, the amount of information accumulated by human activities has exceeded the total amount of information that was ever recorded by humans ten years ago. The term “big data” is increasingly being used to describe and define the massive data generated in the information age, and to name the related technological developments and innovations (Gandomi & Haider, 2015). In recent years, big data-related technologies of storage, processing and analysis are rapidly evolving and widely used in different fields such as sociology, economics, finance, tourism, and management. Being driven by the demands of social computing and the development of big data technologies, social sensing emerges. It adopts large-scale multiple types of sensing devices, such as universal sensors (e.g., motion sensors, audio and video sensors), smartphones (e.g., GPS, call records, text messaging), email and web app (e.g., forums, social networking sites, blogs, wikis), to obtain large-scale real-time field data on human behavior, which is then analyzed and interpreted. In addition, social sensing also emphasizes intelligent assistance and decision support for individuals, groups, and society.

3.1.2 The concept of social sensing

Humans live in a mixed network environment formed by the fusion of communication networks, the Internet, and sensor networks. The digital footprint generated by humans in a mixed network environment converges into a complex picture or data source of individual and group behavior. A scientific understanding of human behavior can improve the quality of our daily lives, such as reducing traffic congestion, limiting the spread of disease and optimizing public resource scheduling. Social sensing can provide new research methods, tools and scientific data for social sciences, while social sciences provide theoretical support and research questions for social sensing. The traditional research of social science usually uses questionnaires or observations to obtain real-world data and acts on the real world through indirect feedback such as interpretation and prediction. Social sensing-based approaches can better connect social science with the real world. However, there is not yet a uniform description of social sensing across multiple disciplines, although the current research on social sensing is in full swing. Here are three popular perceptions of social sensing:

- **Social awareness-oriented computing**. Some of the preliminary ideas of social sensing are reflected in computer supported cooperative work (CSCW). The group perception emphasized by CSCW is one kind of social sensing. It is about the understanding of all aspects of the group activity and is an important factor to ensure the efficiency and quality of collaboration. In 2005, Pentland (2005) first mentioned ”Socially aware computation”. This paper quantifies and visualizes social situations (such as speaking tones, facial movements, and postures) in interpersonal communication to promote people’s social communication. In the subsequent few years, the definition, research scope, theoretical basis, and typical applications of socially aware computation have gradually become clear and enriched. In 2009, Lazer et al. (2009) elaborated the understanding of individuals, organizations, and societies by collecting and analyzing massive real-life data streams. The socially aware computation uses large-scale multiple types of sensing devices to capture and identify individual social behaviors in real-time, analyzes and mines the social interaction characteristics and rules of groups, thereby assisting individual social behaviors and supporting community interaction and collaboration. Socially aware computation emphasizes the use of computer science and technology to perceive real-world individual behaviors and group interactions, providing intelligent assistance and support for individual and group interactions (Lukowicz, Pentland, & Ferscha, 2012).

- **Social media-based sensing**. The popularity of online social media such as Twitter and Instagram allows users to share information about themselves and their surroundings. The vast amount of information posted on social media makes it possible to understand the real world through social networks. Monitoring all aspects of the state of the world in this way is called social sensing (D. Wang, Szymanski, Abdelzaher, Ji, & Kaplan, 2019). Humans are treated as sensors and social media as measurement tools. As a matter of fact, social sensing precedes the use of physical sensors and social media. Thousands of years ago, words were spoken or written to convey observed experiences to others. However, today’s technologies have made it easier for people to formulate and share ideas immediately with networked friends worldwide. The widespread use of social media also enables new possibilities to be involved in collaborative activities for broader groups of people without limitations of time or geographical zones (Lévy, 2010). The main challenge for social media-based sensing is how to find and understand valuable information from the vast amount of social media content.

- **Social sensing from a geographic perspective**. The majority of the real-world data has a temporal and spatial reference and multi-source spatiotemporal data is growing exponentially. Currently, the widely used spatiotemporal big data include global positioning trajectories of mobile devices, public transportation information, location-based social network data, mobile phone signaling data, and pictures with geographic coordinates, etc. The research paradigm of space-related disciplines (such as geography and urban science) has become increasingly data-driven. Spatiotemporal big data has created an opportunity to better understand the spatiotemporal associations and patterns of human behavior, thereby promoting the formation and development of social sensing theories. Liu et al. (2015) proposed a conceptual framework of social sensing in the field of geography, which refers to theories and methods of studying the characteristics of human spatiotemporal behavior with the help of various types of spatiotemporal data, and then revealing the spatiotemporal distribution, connection and process of socio-economic phenomena. The spatiotemporal behavior of individuals is random and it is difficult to extract valuable features. However, as the sample becomes larger, the regularity of group behavior becomes more obvious. Generally speaking, this kind of regularity, especially the socioeconomic characteristics, are related to the geographical environment. Therefore, social sensing based on spatiotemporal big data provides a new observation method of the socioeconomic environment.

Although the understanding of social sensing is different in various fields, and the definition of social sensing lacks a uniform interdisciplinary description, the research methods of social sensing share three common characteristics:

- It is based on sensing. Social sensing acquires continuous, real-time, and dynamic data from the real world through a large number of different sensing devices. Compared with manually collected data, social sensing data is more objective and accurate. In addition, it can better reflect the spatiotemporal characteristics of human behavior and social interactions in the real world.

- It is driven by data. The data-driven social sensing provides a new perspective for social science research. The agent-based social simulation method is difficult to reflect the dynamic development of real social systems. The emergence of social sensing data can quantitatively depict human collective behavior. The knowledge gained through the analysis of large-scale social sensing data is more convincing.

- It is a field study. Social sensing focuses on the user’s real experience. Unlike laboratory research, where users are invited to the laboratory for investigation, field study requires researchers to enter the actual living environment in which users are located for data collection and experimental analysis. Therefore, the biggest advantage of social sensing is its authenticity.

In this thesis, social sensing is defined as the perception and response to human behavior and the surrounding environment through large-scale multi-sensors in social living space, so as to assist individual behavior, group interaction and social development.

3.1.3 Research contents related to social sensing

This part mainly introduces the research content of social sensing from three aspects of data, analysis and service, as shown in Figure 3.1. Real-time sensing in the real world provides data support for analysis of human behavior, which in turn improves human life and provides various intelligent services to individuals, groups and society.

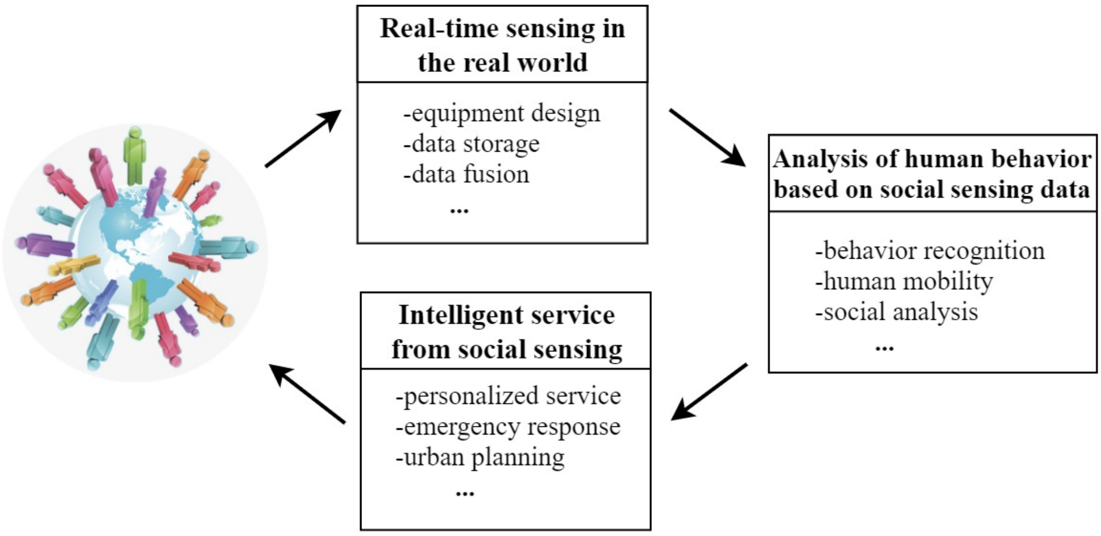


Figure 3.1 Research contents related to social sensing.

Real-time sensing in the real world

Social sensing is essentially the use of a variety of sensing devices to capture the data of human behavior in the real world. The main research contents include the design of sensing equipment, the collection of multi-source heterogeneous data, and the storage and fusion of large-scale sensing data. Currently reported work focuses on the design and development of new sensing devices, as well as the extraction and processing of multiple data sources (e.g., wearable sensors, email, social media).

In order to capture individual and collective patterns of behavior, Olguín et al. (2009) designed wearable devices to measure human behavior data within a group (such as face-to-face interaction, conversational time, physical proximity to other people, and physical activity level). The obtained data can be used to predict human satisfaction with the environment and the relationship between groups. Gips & Pentland (2006) designed a smart badge that can collect data on the behavior of conference attendees. This smart badge can also be used to measure user's research interests. Research on communication characteristics and patterns of e-mail exchange proves that electronic communication data can be used for characterizing individual behavior and identifying latent structure in human populations (Malmgren, Hofman, Amara, & Watts, 2009). Ruflin, Burkhart, & Rizzotti (2011) introduced the data model differences of eight commonly used storage systems (five types) for social sensing data and evaluated their scalability and their ability to query and process data. Although MySQL database has been used in various social media platforms (e.g., Twitter, Facebook) and column store type is also becoming more and more popular, in fact each storage system has its own limitations due to its storage characteristics. Further research is needed to create more scalable systems. In particular, the combination of storage systems and processing should be studied in more detail to provide support for efficient data analysis. In response to travel recommendations, Yerva, Jeung, & Aberer (2012) proposed a conceptual framework that can integrate and analyze heterogeneous social sensing data and physical sensor data. By mapping weather-related social media data and meteorological sensing data into a two-dimensional mood space, it is possible to predict tourist emotions at tourist locations based on historical data and weather forecasts.

Many of the above-mentioned research works mainly use a single device to capture a single data source. With the increasing availability of multiple social sensing methods, efforts on the integration and comprehensive utilization of various data sources with better results are expected in the years to come.

Analysis of human behavior based on social sensing data

Based on social sensing data, individual behavior characteristics can be identified, and various types of analytical methods (e.g., social network analysis, machine learning) can be used to explore group social interactions (e.g., organizational structure, group activities, communication patterns and their dynamic evolution).

Human behavior recognition is mainly achieved by applying various learning models to social sensing data. Feasible learning models can be divided into supervised models (e.g., support vector machines, Bayesian networks, decision trees) and unsupervised models (e.g., clustering, pattern mining) (Chen, Nugent, Cook, & Yu, 2011). Considering various characteristics (i.e., user movement, user location, voice, and environmental information), Wang, Gu, Tao, Chen, & Lu (2011) studied two temporal probability models to model and identify interaction processes involving multi-user activities. These two models were dedicated to identifying multi-user activities based on wearable sensors in a smart home to capture each user's behavior and user interactions. Experiments show that the performance of the two methods is not stable for various types of activities. More features from other sensors (such as gyroscopes, 3D compasses, etc.) are worth further consideration. Su, Tong, & Ji (2014) introduced the research progress of using smartphone sensors to identify user activities. The smartphone carries a variety of sensors (e.g., accelerometer, GPS, light sensor, temperature sensor, gyroscope, barometer). Using various classifiers based on time-domain features and frequency-domain features, users' various daily activities (e.g., jogging, walking, sitting down) can be effectively monitored. However, some challenges related to location sensitivity and activity complexity remain.

On the other hand, various analysis methods (e.g., principal component analysis, linear regression analysis and correlation analysis) have been used to reveal the mobility and activity laws of human beings. The various roles people play in social networks can be reflected in social sensing data. Considering the spatiotemporal factors, Yu, Si, Song, Li, & Yen (2014) designed three metrics to monitor the co-location behavior of mobile phone users. Based on the co-location metrics, the supervised method was used to identify the diverse social relationships between mobile phone users (e.g., family, colleagues). Social relationship recognition based on social networks and communication relies heavily on the connectivity between users, while this method based on co-location features can better perform the relationship recognition task, especially for finding users pairs with weak relationships. Social sensing data (such as mobile phone data, bus data) can be used to understand human mobility patterns. Taking Shenzhen as an example, D. Zhang et al. (2014) analyzed the correlation and difference between multi-source data (i.e., cellphone data, taxicab GPS data, smart card data for subway & bus and bus GPS data) and proposed an inference algorithm based on the mobility graph to predict human mobility. Experiments show that there are certain differences in the mobility patterns of people based on different data sources, and comprehensive reasoning based on multi-source data can address such biases. Sun, Fan, Helbich, & Zipf (2013) explored the potential of perceiving human space-time activities based on Flickr photos. Taking Vienna as an example, they used kernel density estimation and spatial scan statistics to analyze geotagged photos related to tourist accommodation obtained by keyword retrieval. The results can reflect the seasonal trend of tourist accommodation distribution. However, representativeness still needs to be improved.

While the social sensing data can support us to recognize human behavior and detect human social activity patterns, which in turn may trigger further analysis and thus provide intelligent services for individuals, groups, and society.

Intelligent service from social sensing

Under the premise of understanding individual behavior and group interaction rules, social sensing can provide intelligent support for human activities (e.g., personal recommendation, group collaboration support) and serve domain applications (e.g., infectious disease prevention, emergency warning, road traffic collaborative monitoring, cities development plan).

In order to provide personalized location recommendations, Baral, Wang, Li, & Chen (2016) used the matrix factorization method to model human behavior by fusing the social, categorical, geographical, and temporal features from the user's historical check-in data. Experiments based on two real-world datasets proved that the performance of this fused model is better than other recommended models that only consider features from a single channel.

In serving social interaction, Konomi, Inoue, Kobayashi, Tsuchida, & Kitsuregawa (2006) used an identification technology of radio-frequency to obtain and visualize social clusters of conference participants from the publication database, thereby effectively promoting the exchange of conference participants. Arb, Bader, Kuhn, & Wattenhofer (2008) have developed a platform for mobile social networks to promote user communication. By matching the mobile phone contact list between users, the common contacts can be confirmed so as to promote users to develop potential friends.

Campbell & Lane (2008) designed the people-centric sensing application, which can extract and analyze user status from sensory information in personal mobile devices, sports equipment (e.g., running shoes or bicycles), and civil infrastructure. Users can share this information through online social networks (such as Facebook and Twitter) to enhance social interaction with other users.

In the field of intelligent transportation, Calabrese, Colonna, Lovisolo, Parata, & Ratti (2011) used a large amount of sensing data from mobile phones, buses, and taxis to analyze urban mobility in real-time, providing decision support for intelligent traffic management in Rome. In urban planning, various social sensing data can be used to identify land-use types. Based on social media data and points of interest data, Y. Wang et al. (2016) analyzed human activity patterns, social media discussion topics, and the distribution of urban facilities in different regions to infer urban land use patterns. Experiments show that multi-source data analysis based on these three types of features can effectively identify seven types of land use clusters and two types of mixed land use areas. For emergency response, Sakaki, Okazaki, & Matsuo (2013) developed an earthquake reporting system used in Japan based on Twitter data streams. Support vector machine is used to extract earthquake-related information, and particle filter algorithm is used to estimate earthquake location. Experiments show that this system can quickly and effectively detect earthquakes and can issue warning announcements in a shorter time than Japan Meteorological Agency.

Social sensing is not only a precondition for the recognition of real-world human behavior and the understanding of human social activity patterns, but more importantly, it can provide intelligent assistance for individuals, groups, and society in a large number of application fields, such as social networks, public health, public safety, large-scale system engineering, intelligent traffic management, urban planning. No doubt, when the research continues, more intelligent support for the sustainable development of society will emerge.

3.2 Geotagged social media data as a new channel to perceive human behavior

Information about human behavior is a key for the understanding of people's living habits and social dynamics. It is richly embedded in various social sensing data, especially in geotagged social media data which covers many aspects, is widely distributed and rapidly updated. In this section, we introduce the origin of social media and the potential of geotagged social media data in perceiving human behavior.

3.2.1 Social media

Development of information dissemination

Information dissemination is the transportation of information to the intended recipients. People's daily activities are constrained by space, time, and energy, and it is impossible to maintain physical contact with a large society in which they live. Therefore, they have to rely on various types of "information providers" (such as traditional media and social media) to understand the complex society.

The information medium is the content carrier serving information dissemination. Since the appearance of language and text, the continuous development of technology has promoted the evolution of information mediums from print media, electronic media, digital media, Internet media and social media. A comparison of these evolving media is given in Figure 3.2. The first two industrial revolutions greatly improved the efficiency of printing production and led to the emergence of various electronic media. The third industrial revolution promoted the emergence and development of digital media. During this period, along with the popularity of the Internet, the Internet media made it possible to spread diverse information to users around the world in real-time. Subsequently, the Internet application revolution triggered by the Web2.0 ushered in the social media era of user-generated content (UGC).

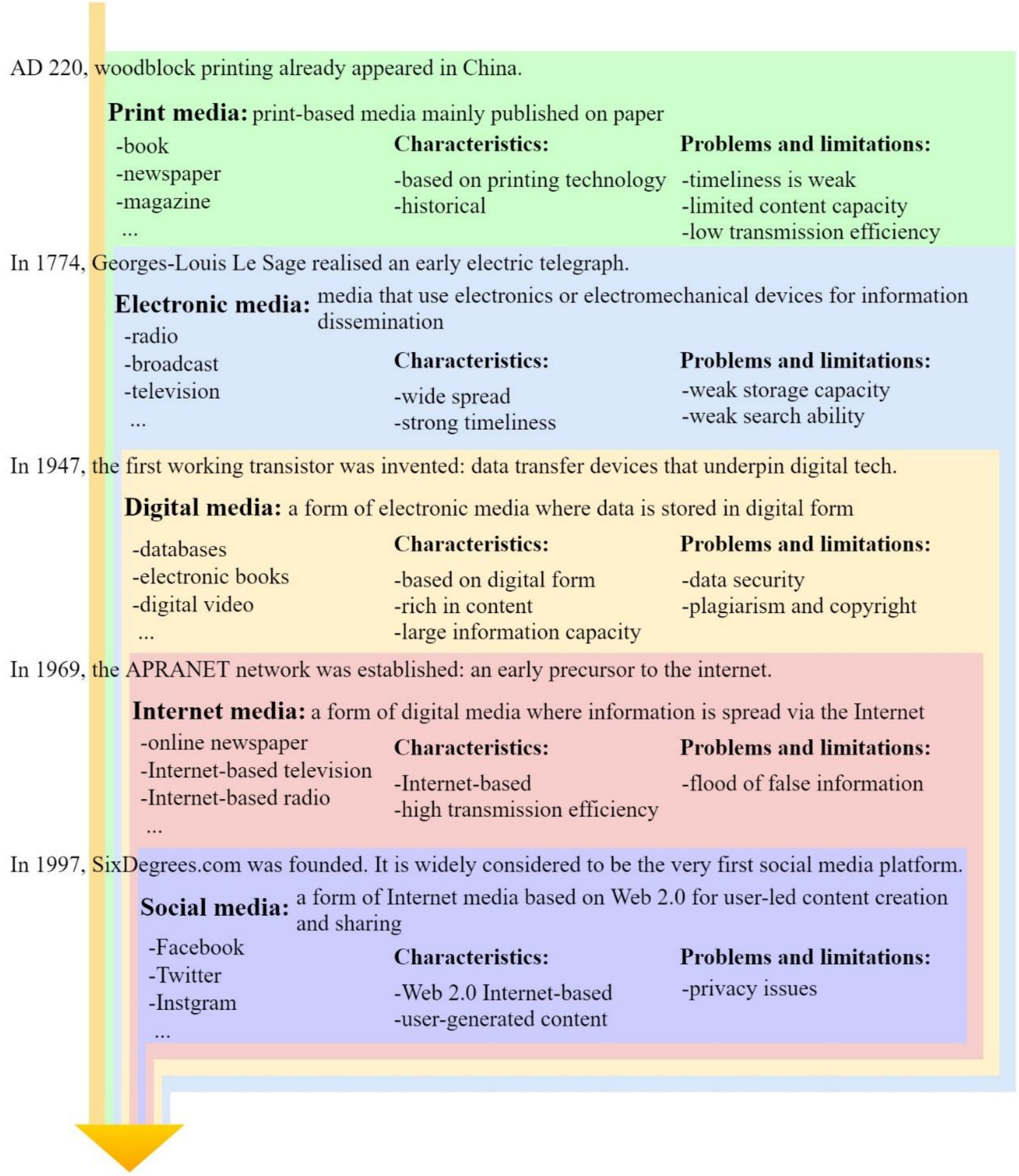

AD 220, woodblock printing already appeared in China.
Print media: print-based media mainly published on paper
-book
-newspaper
-magazine
...
Characteristics:
-based on printing technology
-historical
Problems and limitations:
-timeliness is weak
-limited content capacity
-low transmission efficiency
In 1774, Georges-Louis Le Sage realised an early electric telegraph.
Electronic media: media that use electronics or electromechanical devices for information dissemination
-radio
-broadcast
-television
...
Characteristics:
-wide spread
-strong timeliness
Problems and limitations:
-weak storage capacity
-weak search ability
In 1947, the first working transistor was invented: data transfer devices that underpin digital tech.
Digital media: a form of electronic media where data is stored in digital form
-databases
-electronic books
-digital video
...
Characteristics:
-based on digital form
-rich in content
-large information capacity
Problems and limitations:
-data security
-plagiarism and copyright
In 1969, the APRANET network was established: an early precursor to the internet.
Internet media: a form of digital media where information is spread via the Internet
-online newspaper
-Internet-based television
-Internet-based radio
...
Characteristics:
-Internet-based
-high transmission efficiency
Problems and limitations:
-flood of false information
In 1997, SixDegrees.com was founded. It is widely considered to be the very first social media platform.
Social media: a form of Internet media based on Web 2.0 for user-led content creation and sharing
-Facebook
-Twitter
-Instgram
...
Characteristics:
-Web 2.0 Internet-based
-user-generated content
Problems and limitations:
-privacy issues

Figure 3.2 The evolution of information medium.

The diversity and consistency of social media

The emergence of the Web 2.0 mode has promoted the popularity of social media, which provides instant and interactive platforms for users to create and share content. It greatly increases the frequency of audience participation in the production and dissemination of information. And the diversified needs of users have driven the formation of various social media. Some mainstream social media are shown in Table 3.1.

Table 3.1 Top 20 most used social media in the world (in 2019).

Name	Type	Monthly Active Users (in millions)	User distribution	Main content
Facebook	Social network	2,230	Worldwide	Text, photos and multimedia can be shared with other users with various privacy settings.
YouTube	Video sharing	1,900	Worldwide	Videos can be uploaded, viewed, shared, commented, and YouTubers can be subscribed.
WhatsApp	Instant messaging	1,500	Worldwide	Text, images, and multimedia can be shared to specified users or groups. In addition, users or groups can make voice and video calls.
Facebook Messenger	Instant messaging	1,300	Worldwide	Users can exchange messages, photos and multimedia and make voice and video calls.
WeChat	Instant messaging	1,060	Regional, mainly in China	Text messaging, voice messaging, video calls, photograph and video sharing, as well as location sharing are supported.
Instagram	Photo and video sharing	1,000	Worldwide	Photographs and short videos can be uploaded, edited and organized with tags and location information. Posts can be shared with various privacy settings.
Tencent QQ	Instant messaging	861	Regional, mainly in China	Online group and voice chat, social games, music, shopping, microblogging, etc.
Tumblr	Microblog	642	Worldwide, nearly half in the United States	Users can post multimedia and other content to a short-form blog with various privacy settings.
Qzone	Social network	632	Regional, mainly in China	Users can write blogs, keep diaries, upload photos and determine the visibility of the content.
Tiktok	Video sharing	500	Worldwide, nearly half in China	It is used to create short dance, lip-sync, comedy and talent videos.
Sina Weibo	Microblog	392	Regional, mainly in China	Texts, photos and videos can be shared and commented while instant messaging service is also available.
Twitter	Microblog	335	Worldwide	Users post and interact with messages known as “tweets”. Registered users can post, like, and retweet tweets, while unregistered users can only read them.
Reddit	Rating and discussion	330	Worldwide, nearly half in the United States	Various content can be submitted, discussed and voted by other users. Posts are organized by subject into user-created boards cover a variety of topics.
Baidu Tieba	Forum	300	Regional, mainly in China	Baidu Tieba uses forums as places for users to socially interact. There were more than eight million forums, which covered a variety of topics.
LinkedIn	Business network	294	Worldwide	Professional networking, including job search and recruitment. LinkedIn supports members to create profiles and connect to each other.
Viber	Instant messaging	260	Worldwide	Instant messaging, media exchange, and paid international landline and mobile calling service.

Snapchat	Photo and video sharing	255	Worldwide	Multimedia messages sharing. The most commendable is its “burn after reading” function, which is conducive to privacy protection.
Pinterest	Photo and video sharing	250	Worldwide	Users can create, discover and save information such as images and videos in the form of pinboards.
Line	Instant messaging	203	Regional, mainly in Japan	Line supports users to exchange texts, images, audio, video and conduct free VoIP conversations and video conferences.
Telegram	Instant messaging	200	Worldwide	Telegram supports instant messaging, data exchange, and provides data encryption for part services.

Source: Buffer Marketing Library and Wikipedia

Although a variety of available social media exhibit diverse features, they have some common characteristics (Kaplan & Haenlein, 2010; Obar & Wildman, 2015):

- Social media is an Internet-based interactive Web 2.0 application.
- User-generated content is the lifeblood of social media.

Based on these common characteristics, this thesis treats social media as a platform of Web 2.0 for user-led content creation and sharing. According to the Digital 2020 Global Overview Report, more than 4.5 billion people have internet access, and about 3.8 billion people enjoy social media services in 2020. Users are generating a tremendous amount of data. For instance, 317,000 status is updated every minute on Facebook. A comparative analysis with sociological surveys proves the advantages of social sensing based on social media in terms of coverage, content, data volume, and timeliness, as shown in Table 3.2.

Table 3.2 Comparative analysis of social sensing.

Method **Characteristic**	**Social sensing based on social media**	**Sociological surveys**
Data acquisition	application programming interface	questionnaire survey field research participant observation …
Resource costs	short time mainly equipment cost	time-consuming mainly labor cost
Data quality	uneven and usually requires data preprocessing before analysis	targeted and valuable
Analytical method	mainly data mining	mainly statistical analysis
Content	rich in content (e.g., text, photo, video) and metadata information (e.g., geo-reference, timestamp)	scarce, usually text and numbers
Data size	mass information	small data
Coverage	social media users	small sampling
Timeliness	time-sensitive	weak timeliness

3.2.2 Geotagged social media data

Geotagging in social media

The growing easiness of receiving positioning signals from GNSS (Global Navigation Satellite Systems) and mobile telecommunication systems has spawned a plethora of location-aware devices. A concept called "geotagging" was born and popularized in social media, allowing users to post location data as metadata along with shared contents. Location information can provide an anchor to the shared content and can be used to "check-in" to venues. Geotagging helps to connect the social network of the virtual world with the physical world and is therefore supported by many social media service platforms. The following are some commonly used ones:

- Flickr – It is a photo-sharing platform that allows users to upload images and share them with the public or designated groups. Starting from August 28, 2006, Flickr supports geotagging of photos. Flickr service supports the Exif format in uploaded images to provide location information. The user may set the availability of image location data.
- Twitter – It is a microblogging service that allows users to post text-based short posts. In November 2009, Twitter enabled the geotagging function. A twitter post may include a "location ID" (referring to a human-recognizable location) or the user's latitude and longitude at the time of posting. Users may make privacy settings to decide whether they are willing to enable geotagging and to publish publicly.
- Facebook – It is a social networking platform, allowing users to share text and multimedia content with others. In 2010, Facebook enabled a feature called "location". When users use Facebook from location-aware devices such as smartphones, this feature allows users to check-in. According to Facebook's default privacy settings, this information is only shared with the listed members of the user's friends.
- Foursquare – It is a location-based service available on mobile devices created in March 2009. Similar to Facebook's "location" feature, it allows users to check-in at their current location to announce their existence and share short text messages. Foursquare provides the function of linking with the user's Facebook and Twitter accounts to achieve synchronized updates with Twitter and Facebook so that users can share updated content with friends.

With the popularity and convenience of geotagging in social media, more and more users are accustomed to deploying this service feature, leading to exponentially increased amount of geotagged social media data. The geotagged data records users' check-in location, time, ID, and shared information (e.g., text, picture) which can reflect users' spatiotemporal behavior to a certain extent. The geotagged social media data of individual records the individual spatiotemporal trajectory and a large number of trajectories of many individuals may reflect group trends and social dynamics.

Application potential of geotagged social media data

Geotagged social media data, as an important part of geographic information in the era of big data, has the characteristics of the quick update, large amount, rich information, etc. The geotagged texts, pictures, videos, and other multimedia contents are quickly spread in user-centric social networks. What is recorded are not only the spatiotemporal locations but also users' knowledge and feelings at those locations. For this reason, the geotagged social media data is like a mirror, holistically reflecting users' living conditions in the corresponding socio-economic environments. The analysis of geotagged social media data has a high scientific and practical value for individuals, service providers, and governmental and non-governmental organizations.

For individual users, intelligent services can be provided through mining geotagged social media data. For example, analyzing photo-sharing communities can summarize popular tourist attractions and routes to provide travel recommendation services. By analyzing the traffic information shared by users in real-time, road conditions can be updated in time to provide users with intelligent navigation services.

For service providers, the mining of geotagged social media data may help them optimize their services to match users' lifestyles and preferences. For example, the review website can analyze users' check-in data, thus issue related electronic coupons to users based on their consumption habits and preferences, thereby promoting consumption.

For government and non-governmental organizations, geotagged social media data can be seen as a large-scale sample of human activities and can provide diverse decision support. For example, flu-related microblogs recording the spread of certain flu can be used to derive statistical estimations and predictions, thereby providing decision support for disease control and prevention. Likewise, based on users' check-in data, the distribution and development status of a business district can be analyzed, thus provide a reference for economic decision or industrial planning.

Data uncertainty

In spite of the aforementioned potential values of geotagged social media data in various fields, the shadow side should not be ignored. The analysis is highly dependent on geotagged posts published by users. It is worthwhile to take a close look at the context of data generation and retrieval so as to understand why the study based on geotagged social media data has some inherent limitations.

- **Representativeness**. The analysis based on geotagged social media data relies heavily on publicly posted posts with geotags. Although more and more people publish geotagged social media data, not everyone uses social media applications, and not all posts are geotagged. Therefore, such data does not adequately represent the behavioral characteristics of the entire population. For example, young people tend to be more interested in accessing the Internet, while some older people are not used to socializing on the Internet. Besides, due to the limitation of population distribution and technological development differences, most of the activities in online social networks are concentrated in urban areas, mainly in technologically developed North America, Europe, and Asia. In some underdeveloped areas, digital devices and Internet applications are only sporadically available. Furthermore, the popularity of different social media services varies in different countries and regions. Therefore, when conducting regional research, we need to consider the prevalence of different social media services in specific regions.

- **Credibility**. Social media has the characteristics of being fast, reciprocal and open, which brings great convenience to people to share information. However, these characteristics have also made social media a breeding ground for false information. Since users can publish content on social media almost without any restrictions, the wrong opinions caused by users' cognitive limitations, prejudice, or intended falsification can be arbitrarily posted on social media platforms. Moreover, fake information tends to raise public curiosity more quickly and therefore spread more virally than fact-based news, which may mislead an unforeseen large number of ignorant users. When the number of fraudulent users continues to increase, it may cause social problems and affect social stability. It has been observed that in 2010 U.S. midterm elections, social bots were used to guide the political inclination of the public (Ratkiewicz et al., 2011). Therefore, the credibility of information on social media has become an urgent research issue which is being explored by means of advanced computer technologies such as machine learning, complex network analysis, and natural language processing.

- **Access restrictions**. Some social media services provide APIs for information retrieval, thereby supporting academic research and commercial applications. For example, Twitter data is most commonly used in numerous studies from disaster perception to social network analysis (Stephens & Poorthuis, 2015). Twitter's Streaming API supports free real-time access to up to 1% of tweets for academic research. Corresponding to enterprise-level applications, the Firehouse API supports more open access permissions (Morstatter, Pfeffer, Liu, & Carley, 2013). However, due to strict restrictions in terms of quantity and geographic scope, not all open APIs are suitable for analyzing human spatiotemporal behavior. (Martí, Serrano-estrada, & Nolasco-cirugeda, 2019). For example, Instagram has imposed strict restrictions on data access. Each new application is launched in a sandbox mode, allowing only the latest 20 information records of up to ten authorized users. A strict review is required, if more rights of data access are required.

Privacy issues

The geotagging function connects the user's social interaction in the virtual space with the real-life in the physical space, which helps enhance the attractiveness and experience of using social media services. However, the widespread application of the geotagging function may lead to the leakage of user location information, which implies many potential threats. The attacker can infer personal privacy information such as the user's hobbies, sports patterns, and health status based on the location information. In addition, such information may be used for evil purposes, such as tracking, harassment, or burglary. In the past few years, there have been some third-party applications based on social media services that have violated the law of privacy protection. For example, Stalqer is a mobile application that tracks users' friends based on Facebook's location data. Website "pleaserobme.com" shows the location of the empty house based on the geotagged Foursquare and Twitter posts (Freni, Vicente, Mascetti, Bettini, & Jensen, 2010). The app called "Girls Around Me" integrates Facebook and Foursquare data to show users the geographic location of nearby women.

This alarming situation urges us to reinforce the privacy protection mechanism. Such a mechanism usually incorporates complex algorithms into the original system, which helps increase the uncertainty of user's identity and location, and eliminate the correlation between the identity and location. In this way, user's private information will not be explicitly exposed to potential attackers.

Although the privacy protection mechanism can enhance the protection of user privacy, it may reduce the availability of social media services and the efficiency of the system to a certain extent. Therefore, when designing a privacy protection mechanism, it is necessary to find a balance between privacy protection, the availability of social media service, and the system efficiency. There are still some difficulties in fully implementing location privacy protection in social media services:

- Location privacy protection and location-based services are somewhat contradictory. The higher the quality of location-based services, the easier it is to disclose users' location privacy.
- There are many ways to disclose users' location information. Users may willingly disclose their current location information so that friends can know their whereabouts. In addition, based on the published content information, it is possible to infer the exact location of the user. Some social media services that use real-name authentication bind personal identities with location information, which is more likely to cause leakage of user privacy information.
- Different users have different requirements for location privacy protection, and the same user may have different requirements for location privacy protection on different occasions. Currently, it is hard to design a fine-grained personalized privacy protection scheme that supports different levels of privacy protection.

Despite these difficulties, it is necessary to raise public awareness of privacy protection so that individual persons may take their own responsibility as far as possible to mitigate the possibility of misusing geotagged social media content. In addition, researchers have their part of the responsibility to develop novel strategies that can protect user privacy without affecting the normal operation of social media services.

The continuous innovation of the social media service model will take location privacy protection as a default constraint.

3.3 Social events, human behavior in geotagged social media data

The occurrence of social events will inevitably affect people's inner or outer behaviors. On the one hand, social events as external stimuli can trigger people's cognitive processes of social events. On the other hand, they may also affect people's emotions, opinions, and even trigger various outer behaviors. For example, terrorist attacks can cause severe emotional trauma to the families of the victims. A war may trigger the migration behavior of residents seeking survival chances.

Using social media services such as Twitter and Facebook, everyone can publish his or her own insights and feelings and record the spatiotemporal status anytime, anywhere. And the social event in the real world is one of the main driving forces for people to share information (Valkanas & Gunopulos, 2013). The spatial-temporal and semantic information embedded in geotagged social media data provides valuable indicators to investigate the human behavior induced by social events.

Social events can be perceived in the event-related information posted by social media users. The occurrence of social events induces spontaneous and unsupervised discussions by social media users. Further on, the geotags, if available, can be used to explore the time and place of the event and the evolution process. On the other hand, people will express their opinions, feelings, and changes in their spatiotemporal state through social media services. The analysis of geotagged social media data promotes the timely perception of people's inner and outer behaviors driven by social events. Knowing the impact of social events on people in a timely manner, especially the negative impact, is helpful for emergency response, post-disaster reconstruction, and assisting government and relevant agencies in decision-making.

Geotagged social media data has been used to perceive human behavior induced by social events from multiple aspects. Reported research works so far are mainly focused on answering when where what happened and what kind of impact can be induced. In the following sections, we provide a review from the perspectives of event detection, event tracking, and analysis of event impact.

3.3.1 Event detection

Events can be detected using techniques such as natural language processing, computer vision, machine learning, spatiotemporal analysis, to name but a few. The reported works on event detection fall into two categories. The one is based on Twitter-like textual streams, focusing on the analysis of textual features. The other is based on the Flickr-like multimedia streams, considering visual, textual, and spatiotemporal features for multimodal fusion (R. Zhu, Zuo, & Lin, 2019).

Event detection from Twitter-like textual streams

The approaches of event detection from textual streams can be either document-pivot or feature-pivot. The former focuses on the integration of documents, while the latter focuses on the clustering of bursty features from the documents.

The document-pivot detection

The document-pivot detection follows the principle of assigning documents that address the same event to the same cluster. The organization of documents is based on a bag-of-words model. Depending on the degree of semantic considerations, three common models are possible: Vector Space Model (VSM), Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

A VSM-based method can be used to describe a document as a vector of terms weighted by parameters such as term frequency (TF) and term frequency-inverse document frequency (TF-IDF). The vectors belong to the same cluster if they reveal a similarity degree beyond a given threshold. Each cluster refers to a potential event. The VSM-based method is intuitive but suffers from a low computational efficiency when dealing with a high-dimensional sparse matrix.

This drawback can be overcome in an LSA-based method whose basic idea is to map a sparse high-dimensional lexical space of texts to a low-dimensional latent semantic space and then compute similarity measures in the latent semantic space. Nevertheless, the LSA-based method cannot distinguish between events related to the same topic if a large number of words are common between these events. A remedy would be achieved if the textual information and the spatial-temporal information are jointly considered in document clustering.

The LDA is a topic model based on Bayesian probability with a three-layer structure of words, topics, and documents. It is usually adapted to handle the characteristics of social media data (e.g., short message) or extended to integrate spatial and temporal features for event detection. A case study is provided by C. Zhang et al. (2017) who modeled the text, time, location of tweets as a multimodal embedding, and utilized various features to explore the possible geo-topics via a Bayesian mixture model. A regression classifier was further applied to identify real local events from candidate geo-topics. In addition, some research works are dedicated to improving the efficiency of event clustering. Valkanas & Gunopulos (2013) used sentiment classification to assist event detection from Twitter streams. When a surge in any emotional state is observed at a certain location, a targeted event detection will be implemented for this area with various term-weighting techniques. In order to avoid considering the entire stream of messages when a specific event type is investigated, Sakaki, Okazaki, & Matsuo (2010) built a supervised classifier to filter out the tweets that are irrelevant to a target event.

The feature-pivot detection

The real-time detection tasks can be suitably handled by means of feature-pivot methods. The feature-pivot detection follows the working principle of finding bursty features from social media streams, and grouping these features based on some common characteristics. The process of feature-pivot detection can be divided into three main steps: feature extraction, feature selection, and feature clustering. At first, the content of each document is preprocessed in order to extract features along with their respective count statistics. Then, the bursty features are identified from a large number of extracted features based on various selection techniques such as discrepancy principle, wavelet analysis, and discrete Fourier transform (DFT). Finally, the bursty features are clustered as potential events using various algorithms that can be model-based, graph-based, or density-based.

As the number of posts related to local events is relatively small, frequency-based approaches alone may not be sufficient. Therefore, the location information of social media data can be integrated to enhance the detection of local events. Sugitani, Shirakawa, Hara, & Nishio (2013) applied a hierarchical method to cluster the tweets based on the geo-distance and identify the spatiotemporal burstiness of terms in each cluster to identify real local events. Krumm & Horvitz (2015) split the space hierarchically into relatively small regions and utilized the regression errors of geotagged tweet volumes to approach the regions with local events. Based on the analysis of the spatial concentration of temporal keywords, local event information could be detected in real-time from the Twitter stream.

Event detection from Flickr-like multimedia streams

Supervised methods and unsupervised methods are all widely adopted for event detection from Flickr-like social media streams. Early machine learning methods focused on the event detection in imagery, ranging from monomodal to multi-modal imagery created by fusion. For example, Petkos, Papadopoulos, & Kompatsiaris (2012) took space, time and visual features into account and developed a typical multimodal spectral clustering approach. Their method trains a classifier as an indicator matrix to supervise the multimodal fusion and clustering. Ahsan, Sun, Hays, & Essa (2017) proposed a web-based supervised learning approach to detect visual concepts relevant to specific event categories. On one hand, Wikipedia and Flickr tags were used to mine event concept with Google News dataset as supplementary resources. On the other hand, Microsoft Bing was used to search and build event-related training datasets. Classifiers were trained with deep convolutional neural network features extracted from a pre-trained network on all images. This Internet-based learning method requires only a few labeled examples and the results were surprisingly good.

Parallel to progresses in supervised learning, research works on event detection based on unsupervised learning continue. Papadopoulos, Zigkolis, Kompatsiaris, & Vakali (2011) built hybrid image-similarity graphs based on image and tag similarity. Considering visual features, text (i.e., title, description and tags) and time, Schinas, Papadopoulos, Petkos, Kompatsiaris, & Mitkas (2015) used a sliding window over the multimedia stream to build and maintain a multimodal same-event image graph and applies a graph clustering algorithm to detect events. This method enables the query of specific events and can classify events based on the duration and influence range of the detection cluster. Yang, Li, Liu, Ma, & Cheng (2017) addressed multiple heterogeneous features (i.e., time, location, tags, user identity and visual feature). They proposed a three-stage framework to discover events from Flickr-like data. First, a soft-voting strategy and a graph random walk model were combined to obtain fused features. Then, a multimodal feature coding model constrained by the dual structure was proposed to learn multimodal data representations. Finally, hybrid clustering models were applied to discover potential events. Their experimental results show that four clustering methods based on the proposed framework (DBSCAN, K-Means, and its two semi-supervised methods) can achieve improved performance compared with baseline approaches.

3.3.2 Event tracking

Event tracking refers to a process of recording the development of an already detected or known event. Geotagged social media data can be used to retrieve event-related information and understand the evolution of social events.

Extraction of event-related information

Once an interesting event has been detected, further observation and analysis may follow to keep track what else would happen to the event. An information extraction algorithm is needed to obtain more event-related information from the original data. In addition to the traditional keyword retrieval methods, other strategies are possible. Abel, Hauff, Houben, Stronkman, & Tao (2012) proposed a user-driven semantic filtering strategy to track real-world incidents or crises. Given an incident, collected messages are processed by means of a semantic enrichment module including named entity recognition (NER), classification of messages, linkage of messages to external Web resources and further metadata extraction. Their experiments show that this kind of semantic filtering strategy outperforms keyword-based filtering. Murzintcev & Cheng (2017) proposed an automated process to collect hashtags related to an interested event, thereby obtaining event-related information. As a result, highly relevant messages for every event separately are retrieved, which is an advantage compared with methods relying on a disaster lexicon. Yu Feng & Sester (2018) addressed flood-related information by conducting text-based and photo-based classification respectively and testing on various classification algorithms (e.g., random forest, ConvNets). This allows them to identify the best classifiers. This approach is also applicable to track other categories of events.

Semantic evolution based on event tracking

An event can be specifically tracked by observing its semantic evolution. Osborne et al.

(2014) designed a real-time event tracking and summarization method. For a given event, the search query is formed on the most informative terms (i.e., nouns, adjectives, verbs and cardinal numbers). When new related tweets come in, the redundant tweets would be eliminated and remaining ranked tweets are used to update the summary of the event. Cai, Yang, Li, & Huang (2015) applied a generative probabilistic model, which essentially relies on a maximum-weighted bipartite graph matching to trace the evolution of events along the temporal dimension. The effectiveness of this method is proved in a case study on the event "Snowden". Schinas et al. (2015) proposed to apply the graph-based clustering algorithm periodically on features of time, text, and image. This allows to detect and trace dense sub-graphs that correspond to events. Once new events are detected, the system tries to link them with detected events from earlier timeslots based on structural similarity between the underlying sub-graphs. For event summarization, representative and diverse sub-set of images will be selected based on a graph-based ranking algorithm.

Spatiotemporal evolution based on event tracking

The evolution of an event may also be tracked by its spatial changes in partitioned time units. X. Zhou & Xu (2017) used the time spectrum to find the starts, ends, or prime time of an event. Based on the density of tweets, contour lines were designed to simulate the spatial pattern of events. An empirical study was conducted to delineate the spatiotemporal evolution of a natural event (heavy precipitation) and a social event (Pope Francis' visit to the US). Based on spatial densities of both event-related posts and all posts, Gao, Wang, Padmanabhan, Yin, & Cao (2018) adopted the Epanechnikov kernel function to track and visualize the spatial and temporal trends of an event in a map of social media event rate (SMER). Their study used a sequence of historical SMER maps to estimate local event baseline and reveal potential spatiotemporal patterns of the underlying event. In addition, geotagged photos can be used to monitor the scene changes. Yan, Eckle, Kuo, Herfort, & Fan (2017) proposed a workflow to monitor and assess post-disaster tourism recovery from geotagged Flickr photos. A space-time bin method in both spatial and temporal dimensions was used to assess the recovery of scenic spots by comparing the similarities of tourist photos in the affected areas at different time periods. Timely tracking disaster recovery is beneficial to disaster management and visitor awareness.

3.3.3 Analysis of event impacts

A timely analysis of event impact helps to improve situational awareness and understanding public opinions, which is desirable for governments or enterprises to take necessary measures to amply/reduce the positive/negative impact. Currently reported explorations on the impact of events are based on the mining of situational awareness and public opinion from geotagged social media data.

Mining of situational awareness

One of the main service capabilities of geotagged social media data is situational awareness about the impact of events. In order to understand how social media is used by emergency management professionals, MacEachren et al. (2011) conducted surveys based on questionnaires. Their results show that mainstream social media software (e.g. Facebook, LinkedIn, Twitter) are all used commonly for personal or professional purposes. And maps (94.7%), photos/video collections (71.1%), time graphs (60.5%) are the top three data formats of the web-based application for emergency management. Crooks, Croitoru, Stefanidis, & Radzikowski (2013) studied the perception of Tweet data on earthquakes. Although this kind of social media data is not equivalent to a seismograph that monitors the intensity of an earthquake, the tweet-like social sensing is useful for the rapid identification and localization of the impact area of an earthquake event. Similar researches include flood, fire, and storm perception. Related photos attached to tweets can provide a clear idea of disaster situations (Dashti et al., 2014). In addition to natural disasters, geotagged social media data can also play an important role in the monitoring of influenza. Signorini et al. (2011) applied the support vector regression to perceive the disease activity in collected influenza-related tweets. Gao et al. (2018) analyzed the spread of influenza in real-time and at multiple geographical scales based on twitter. Their case study of flu seasons in the United States in 2013 and 2014 shows that the proposed procedure yields results that correlate strongly with national and local influenza-like illness (ILI) reports.

Mining of public opinions

Besides situational awareness, geotagged social media data can also be used to explore the impacts of the event on humans. Caragea, Squicciarini, Stehle, Neppalli, & Tapia (2014) performed sentiment classification (i.e., positive, negative and neutral) of posted messages during the Hurricane Sandy. Their map-based visual representation shows how users' sentiments change in relation not only to the locations of users but also the relative distance from the disaster. X. Zhou & Xu (2017) analyzed the potential of social media data for the perception of emotions induced by social events. Empirical research shows that event-related social media data can effectively reflect user's emotional changes towards the event. In addition, the time series display of the emotional values of the high quartile and low quartile can effectively reflect the different emotional changes of different groups on the same event. Tumasjan, Sprenger, Sandner, & Welpe (2010) investigated whether Twitter could be used as a valid indicator to mirror the offline political sentiment. A case study on the German federal election shows that Twitter is indeed used extensively for the political deliberation. The ranking by number of messages mentioning a party is basically consistent with the ranking by share of vote in the election results. Moreover, joint mentions of two parties are in line with political ties and coalitions in the real world. Jahanbakhsh & Moon (2014) studied the 2012 US presidential election combined with machine learning and the LDA model. The opinion mining from geotagged social media data is consistent with the actual public opinion. This also shows that it is an effective way to understand social attitudes toward social events through social media.

In general, current research on geotagged social media data is more concentrated on the detection and tracking of social events than the study of the impacts of social events on human behaviors. This is a natural development because the latter is built upon the results of the former. This thesis is dedicated to developing methods for the perception of human behavior induced by social events based on geotagged social media data.

4 Methodological foundations of human-behavior sensing from geotagged social media data

The review of the state of the art in Chapter 3 shows that the current research on the perception of social events based on geotagged social media data focuses on the detection and tracking of social events. The study of the impacts of social events on human behaviors has just begun and is the emphasis in this chapter. We present the methodological fundamentals of perceiving human behavior with regard to inner behavior, crowd behavior, and crowd mobility induced by social events. Section 4.1 introduces a conceptual framework of how social events trigger users' responses in geotagged social media data which is then used to derive human behavior. How the human inner behavior is induced by a specific social event is explained in Section 4.2. Section 4.3 and 4.4 address the crowd sentiments and crowd mobility induced by social events. The corresponding methods based on sentiment-constrained spatiotemporal semantic clustering and the geospatial network analysis are introduced.

4.1 A generic framework

Social media services have enabled users' contributions to social sensing, which simultaneously comes with an inevitable challenge (Ali et al., 2011). Unlike the monitoring of natural environment where the expected data can be delivered in real-time through automated sensors. In social sensing systems, users as social sensors are characterized by their spontaneity and discontinuity. Therefore, adaptations of existing approaches and new approaches are needed to extract valuable information and mine the hidden knowledge from social media data.

A considerable amount of social media data reflects social events near and far. In other words, social events will inevitably have various potential impacts on human behavior. They induce spontaneous and unsupervised communications among social media users. Users' opinions, feelings, and changes in spatiotemporal states circulating in social media are therefore rich data sources for us to perceive human behavior induced by social events. Figure 4.1 shows a generic framework of perceiving human behavior from geotagged social media data.

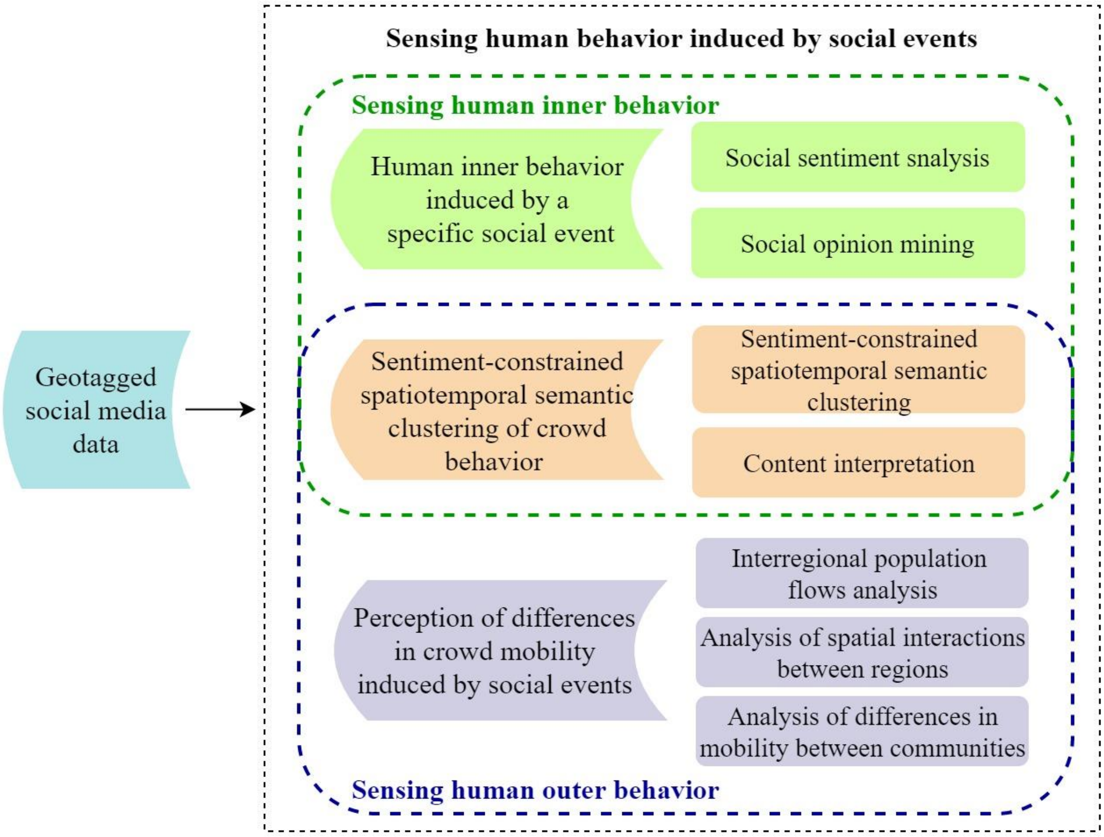

Sensing human behavior induced by social events
Sensing human inner behavior
Human inner behavior induced by a specific social event
Social sentiment snalysis
Social opinion mining
Geotagged social media data
Sentiment-constrained spatiotemporal semantic clustering of crowd behavior
Sentiment-constrained spatiotemporal semantic clustering
Content interpretation
Perception of differences in crowd mobility induced by social events
Interregional population flows analysis
Analysis of spatial interactions between regions
Analysis of differences in mobility between communities
Sensing human outer behavior

Figure 4.1 Perception of human behavior from geotagged social media data.

The framework contains three case studies selected to demonstrate how human behavior is perceived from different perspectives:

- **Analysis of inner behavior induced by a specific social event**. On the one hand, perceiving human inner behavior from geographic social media data is helpful to explore the impact of the specific social event on human emotions; on the other hand, people’s opinions as cognitive reflections on a specific social event can be obtained. With the easy accessibility of social media platforms, a large amount of information that mirrors a specific social event can be quickly collected, geotagged and used as a preliminary reference for emergency response or as a supplement to survey statistics.

- **Sentiment-constrained spatiotemporal semantic clustering for crowd behavior**. Crowd phenomena are very common and can be observed in many public places such as concerts and sport events. They are physical or virtual gatherings of people who share a purpose at the same time. Accordingly, the behavior of these people forms a crowd behavior. One of the manifestations of crowd behavior in social media data is the descriptions of similar or related emotions at specific times and places, which may be explored through a suitable clustering algorithm. Density-based spatial clustering of applications with noise (DBSCAN) is a popular clustering algorithm proposed by Ester, Kriegel, Sander, & Xu (1996). In order to discover knowledge from spatiotemporal data, Birant &

Kut (2007) extended the DBSCAN algorithm and designed the ST-DBSCAN algorithm considering the spatiotemporal neighborhood of objects as shown in Figure 4.2. ST-DBSCAN algorithm has been widely used in the field of spatiotemporal analysis (Huang, Li, & Shan, 2018; S. Xu, Li, & Huang, 2019). The ST-DBSCAN is further extended in this thesis to a sentiment-constrained spatiotemporal semantic clustering algorithm (SCSTSC), which is used to infer human crowd behavior from geotagged social media data.

- **Perceive differences in population mobility induced by social events**. The occurrence of major social events (such as regional conflicts, festivals, and political events) may trigger large-scale population movements. The current analysis mainly relies on various survey data. The widespread social media services have made it possible to find the laws of population movement induced by social events in the changing spatiotemporal states of social media users. A geospatial network analysis framework is proposed to explore the population mobility patterns induced by social events from geotagged social media data.

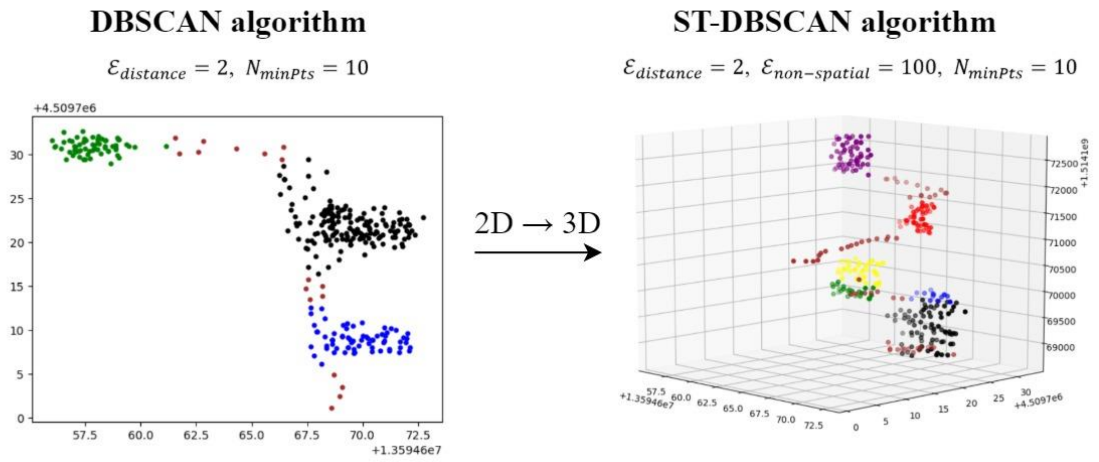


Figure 4.2 From DBSCAN to ST-DBSCAN: extension from two-dimensional to three-dimensional space.

4.2 Human inner behavior induced by a specific social event

In this study, we start from known or detected events, focusing on how they affect people's inner behavior, and for how long. By combining machine learning, natural language processing, and visualization methods in a generic analytical framework, we attempt to interpret the impact of known social events on human inner behavior based on geotagged social media data. The impact has multiple perspectives including time, space, and semantics. The analytical process consists of four parts: (1) preprocessing; (2) extraction of event-related information; (3) analysis of event impact; and (4) visualization, as shown in Figure 4.3. The data preprocessing aims at improving the quality of data and adapting the data to the subsequent treatment. The extraction of event-related information from geotagged social media data is described in Section

4.2.2. The analysis of social event impact on human inner behavior is divided into two parts: social sentiment analysis and social opinion mining. The former examines the public sentiment affected by social events (Section 4.2.3), and the latter emphasizes people's rational knowledge of events (Section 4.2.4). Finally, visualization methods for data and analysis results are discussed in Section 4.2.5.

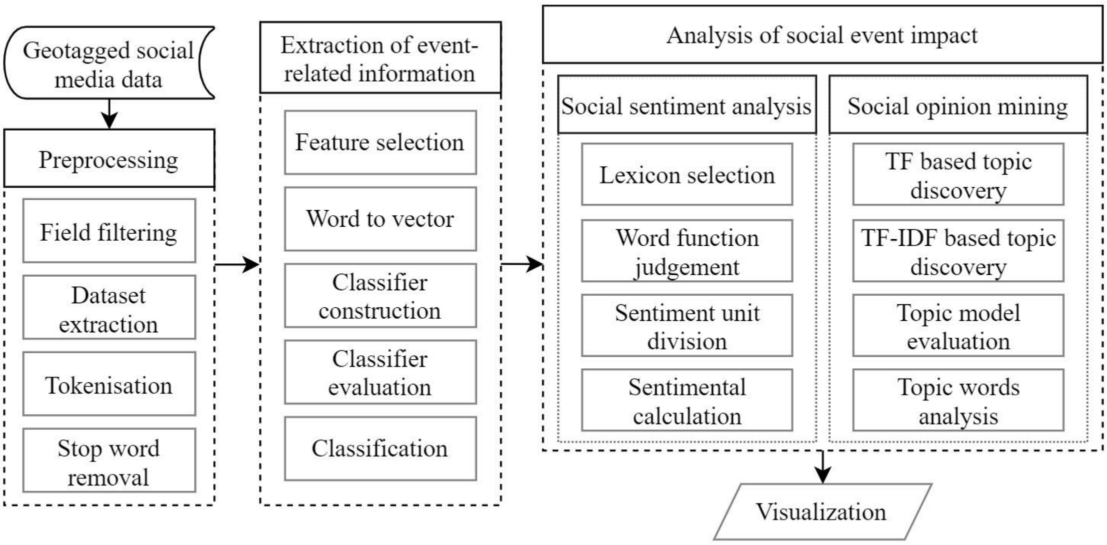


Figure 4.3 The analytical flow diagram of perceiving human inner behavior.

4.2.1 Preprocessing

The quality of user-generated content is rather heterogenous, varying from excellence to abuse and spam. Therefore, data cleaning is required. The preprocessing process includes field filtering, dataset extraction, text tokenization, and stop-word filtering. Bearing in mind that this study deals with the spatial, temporal, and semantic dimensions of geotagged social media data, related fields such as release time, location, and text messages are extracted from the raw data, and other unnecessary information is filtered away.

The hashtag is a good reference to label event-related information by the public. First, event-related hashtags are manually selected to construct an event-related information dataset. The individual posts in the dataset are then classified as being either event-related or event-irrelevant, leading to the construction of two corresponding training datasets.

In order to facilitate the subsequent classification and mining of texts, we undertake a tokenization process as well. Cohesive strings from social media posts are split up into single words using word segmentation technology. Texts in different languages can be segmented using corresponding strategies or specialized software, but the working principle remains the same. For example, words in the English language can be separated by spaces. However, in Chinese text, a semantic unit is often composed of one or more Chinese characters, and there is no obvious separator between words.

Therefore, a Python package specialized for Chinese text segmentation called “jieba” can be adopted. By building a directed acyclic graph based on prefix dictionary structure, potential word combinations can be efficiently identified. The most common and frequently occurring words lacking valuable information, which are called “stop words”, are then excluded to reduce noise among the remaining tokens. In this study, we used the standard stop word list released by the Information Retrieval Laboratory of the Harbin Institute of Technology to remove stop words.

4.2.2 Extraction of event-related information

Extracting unlabeled event-related information from geotagged social media datasets could be formulated as a binary classification problem and solved with machine learning methods. As shown in Figure 4.4, the whole process starts with the construction of training data, continues with feature selection, the construction and evaluation of classifiers, and terminates with extracted event-related information.

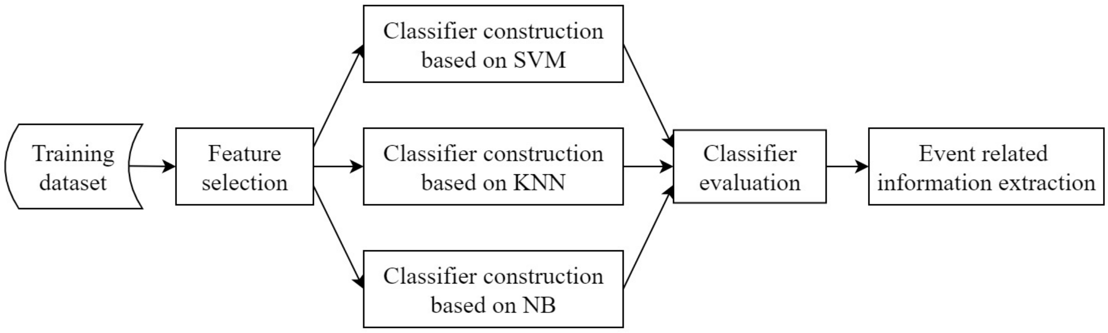

Training dataset
Feature selection
Classifier construction based on SVM
Classifier construction based on KNN
Classifier construction based on NB
Classifier evaluation
Event related information extraction

Figure 4.4 The workflow for the extraction of event-related information.

At first, two training datasets are constructed: an event-related dataset, and an event-irrelevant one. Hashtags serve as markers for an event-related training dataset, whereas a dataset posted under similar conditions but before the event happened is selected as an event-irrelevant training dataset. Subsequently, we apply the Vector Space Model (VSM), which represent text documents as vectors of identifiers that can be easily processed by computers. In addition, the term frequency - inverse document frequency (TF-IDF) is used as the semantic weighting factor, which is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

A text-based binary classification is required in order to extract more event-related information from the original data. With respect to a specific event, it is usually difficult to obtain large amount of labeled data for deep learning in a short period of time, so this paper considers three robust machine learning algorithms (Naive Bayes, the k-nearest-neighbors algorithms, and Support Vector Machine) as alternatives (Altman, 1992; Cortes & Vapnik, 1995; Feng & Sester, 2018; L. Yang, MacEachren, Mitra, & Onorati, 2018). The Naive Bayes (NB) method is a probabilistic classifier based on Bayes’ theorem and assumption of conditional independence. The k-nearest-neighbors (KNN) algorithm is a non-parametric method with the k closest training examples in the feature space as input. Support Vector Machine (SVM) constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space to do classification, regression, or other tasks. These three methods are perhaps insufficient for complex classification tasks, but they perform well on the binary classification of texts. Hyperparameters tuning is indispensable to build classifiers, and cross-validation is required to avoid overfitting. In the end, we use three common measures to compare the relative performance of the classifiers: precision, recall, and *F1_score*, as shown in Equations (4.1) - (4.3):

$$\text{Precision} = \frac{t_p}{t_p + f_p} \tag{4.1}$$

$$\text{Recall} = \frac{t_p}{t_p + f_n} \tag{4.2}$$

$$\text{F1\_score} = 2 \times \frac{precision \cdot recall}{precision + recall} \tag{4.3}$$

where t_p is the number of true positives, f_p is the number of false positives, and f_n is the number of false negatives. *Precision* is the fraction of positive items among the retrieved items. *Recall* measures the proportion of actual positives that are correctly identified as such, and *F1_score* is the harmonic average of the *Precision* and *Recall.* Among the three aforementioned classifiers, the one with the best performance is adopted to extract event-related data for further analysis.

4.2.3 Social sentiment analysis

This section describes how to analyze changes in public sentiment in social media messages. The common sentiment analysis task is to extract sentiment polarity or intensity from text, facial expression, body movement, or music (Cambria, Schuller, Xia, & Havasi, 2013). Regarding short text-based sentiment analysis, common methods can be categorized as supervised methods and lexicon-based methods. Since supervised approaches are usually domain-specific, while lexicon-based methods are more general, the lexicon-based methods are a better fit for our context.

We consider sentimental words, stop words, degree adverbs, and antonym words in analyzing the level of sentiment in the social media message. Three tasks are involved: judgment of the word function, construction of a sentimental unit, and calculation of the sentimental value. With regard to the first task, each word is determined to be a sentimental word, a degree word, or an antonym word. The sentimental units are then constructed according to the location of the sentimental word. Each sentimental unit contains a sentimental word as well as possible degree words and antonym words before it, which is typical for Chinese as well as English. Finally, we calculate the sentimental score for each sentimental unit and summarize all scores as the final sentimental score for the whole text, as shown in Equation (4.4):

$$P(T) = \sum_{i=1}^{n} P(U_i) \tag{4.4}$$

where $P(U_i)$ means the sentimental value of the ith sentimental unit, and $P(T)$ is the sentimental value of the text.

The degree wordlist is provided by HowNet knowledge base (Dong, Dong, & Hao, 2010). It has 219 words and is divided into six levels. Referring to (Yin Wang & Zhang, 2017), we assign these degree levels different weights in descending order (2, 1.5, 1.25, 1.2, 0.8, 0.5) and attach the weights to the corresponding sentimental words. The degree factor γ for the ith sentimental word w_i is defined in Equation (4.5):

$$\gamma(w_i) = \prod_{k=1}^{m} d_{ki} \tag{4.5}$$

where d_{ki} is the weight of the kth degree word for the ith sentimental word.

Antonym words are important for the judgment of the sentimental polarity in a sentence. If an antonym word precedes a sentimental word, the semantics for the word will be reversed, which will affect the sentimental polarity of the whole sentence. Since there is not a standard antonym wordlist, we extracted 44 antonym words to construct such a wordlist based on several related research works (Dang & Zhang, 2010; Wen, 2003). The impact factor of antonym words τ on the sentimental words w_i is defined as:

$$\tau(w_i) = (-1)^n \tag{4.6}$$

where n is the number of antonym words in the ith sentimental unit. Taking all the above characteristics into consideration, we obtain the sentimental value of the ith sentimental unit:

$$P(U_i) = P(w_i) \times \gamma(w_i) \times \tau(w_i) \tag{4.7}$$

where $P(w_i)$ is the sentimental value of the ith sentimental word. The sentimental wordlist provided by the Dalian University of Technology has a more detailed intensity division for sentimental words (L. Xu, Lin, Pan, Ren, & Chen, 2008). It contains 27,466 sentimental words and each word has a polarity and a sentimental intensity. Sentimental intensity is divided into five levels (1,3,5,7,9) in ascending order.

4.2.4 Social opinion mining

Social events as external stimuli can arouse cognitive reactions of individuals. The immediate perceptions tend to prevail, whereas the reflections on the concept, judgment, and inference of events remain less affected. Besides, social media users acting as smart social sensors are able to articulate their personal perception to social events to reveal the hidden relations between a core event and other events. In this section, we use the topic discovery model and the topic evaluation model to mine social opinions from social media messages. The analysis of the topic words obtained from topic discovery can assist us to discover the attitude of the public and the implicit relationships among events.

Topic Model: LDA

Topic model is a type of statistical model used to discover the latent semantic structures of a text body. It helps us to organize and gain insights into large collections of unstructured text bodies (Blei, 2012). Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) are the prevailing methods used for selecting representative terms in text collections (Blei, Ng, & Jordan, 2003; Hofmann, 1999). In our experiments, we chose LDA since it is the generalization of PLSA and can create results to better explain the text semantics. LDA is a Bayesian graphical model and has three layers of "document-topic-word" (Griffiths & Steyvers, 2004). The graphical model of LDA is shown in Figure 4.5 (R. Zhu et al., 2019).

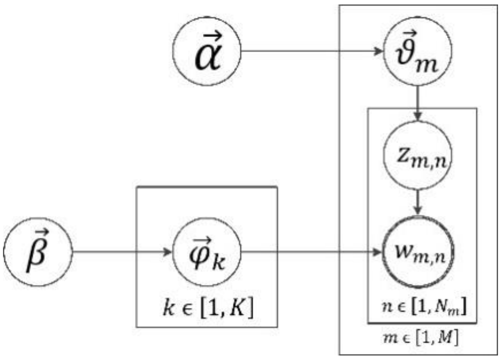


Figure 4.5 Graphical model of Latent Dirichlet Allocation (LDA).

Its generative process is presented as follows:

- For each document m, pick a multinomial distribution ϑ_m from a Dirichlet distribution with parameter α;
- For each topic k, pick a multinomial distribution φ_k from a Dirichlet distribution with parameter β;
- For the nth word in the mth document where $m \in \{1, \cdots, M\}$, and $n \in \{1, \cdots, N_M\}$
  - Sample the word $w_{m,n} \sim Multinomial(\varphi_{z_{m,n}})$
  - Sample the topic $z_{m,n} \sim Multinomial(\vartheta_m)$

The $w_{m,n}$ is the only observable variable, and other variables are latent variables. This study uses Gibbs sampling (Heinrich, 2008), and the word-topic matrix ϑ and φ could be calculated as follows:

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V}\left(n_k^{(t)} + \beta_t\right)} \tag{4.8}$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K}\left(n_m^{(k)} + \alpha_k\right)} \tag{4.9}$$

$$p(z_i = k|\vec{z}_{-i}, \vec{w}) \propto \frac{n_{m,-i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}\left(n_{m,-i}^{(k)} + \alpha_k\right)} \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{t=1}^{V}\left(n_{k,-i}^{(t)} + \beta_t\right)} \tag{4.10}$$

where i indicates a two-dimensional subscript of (m, n). z_i is the assignment of the nth word in document m to topic k. w_i indicates the nth word in document m and $-i$ means not including word w_i. $\vec{n}_m$ describes the vector $\left(n_m^{(1)}, \cdots, n_m^{(k)}\right)$, in which $n_m^{(K)}$ demotes the number of words belonging to the kth topic in document m. $\vec{n}_k$ indicates the vector $\left(n_k^{(1)}, \cdots, n_k^{(V)}\right)$, in which $n_k^{(t)}$ demotes the number of word t generated by topic k.

Topic Evaluation Based on Topic Coherence

The topic discovery based on LDA or PLSA requires that the number of topics is predetermined. How to find the right number of latent topics in a given corpus, however, remains an open question. Among the proposed evaluation methods, those that are based on topic coherence have revealed a desirable performance (Röder, Both, & Hinneburg, 2015).

Moreover, a topic evaluation model can be either intrinsic or extrinsic. Intrinsic methods do not use any external sources or tasks from the dataset, but for extrinsic tasks, a reference corpus is needed for creating a distributional semantic model. Since no good external domain datasets for sudden hot events currently exist, we chose the intrinsic measure UMass introduced by D. Newman, Lau, Grieser, & Baldwin (2010) to evaluate the topic coherence and define an optimal number of topics. For each topic, computing the sum of pairwise scores on the topic words $w_1, w_2, \cdots, w_n$, usually the top *n* words by frequency $\mathrm{p}(w|k)$:

$$Coherence = \sum_{i<j} score(w_i, w_j) \tag{4.11}$$

The UMass measure is expressed as a pairwise score function:

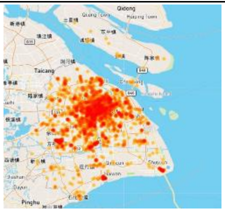
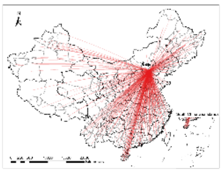
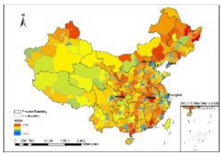
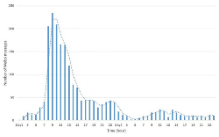
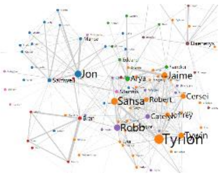
$$score_{UMass}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)} \tag{4.12}$$

$D(w_i)$ describes the count of documents containing the word w_i, $D(w_i, w_j)$ indicates the count of documents containing both words w_i and w_j. It is the empirical conditional log-probability $log\, p(w_j|w_i) = log\, \frac{p(w_i,w_j)}{p(w_j)}$ smoothed by adding one to $D(w_i, w_j)$.

4.2.5 Visualization

Visualization provides an effective means of data exploration and knowledge representation (Zheng, Wu, Chen, Qu, & Ni, 2016). This section introduces some visualization methods for spatial, temporal, and textual patterns, which are suitable for demonstrating our analytical results, as shown in Table 4.1.

Table 4.1 Some commonly used visualization methods for spatial, temporal, and textual patterns.

Visualization	Description	Application	Example
Heat map	A data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, displaying how the phenomenon is clustered or varies over space.	Spatial patterns	
Flow map	A map that shows the movement of objects from one location to another, such as the number of people in migration or the number of packets in a network.	Spatial patterns	
Choropleth map	A map that uses color to show variations in quantity, density, percent, etc. within a defined geographic area. Each color usually depicts a range of values.	Spatial patterns	
Bar chart	A chart that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.	Temporal patterns	
Spatiotemporal dot map	A map that shows the three-dimensional object and time-evolving physical quantity into three-dimensional space with spatial and temporal scale.	Spatiotemporal patterns	
Tag cloud	A visualization technique of text data to show word collection. The importance of each tag is usually shown with font size or color.	Textual patterns	
Word network	A text network visualization technique that encodes all the words as the nodes in the graph and their co-occurrences as the connections between them.	Textual patterns	

With regard to locations, point-based visualization helps to endow individual points within the spatial context. Each point represents an object, and its visual variables (e.g., color, size) carry related information. Heatmap is a good practice in expressing spatial hot spots as clustering results. Line-based visualization turns the discrete points into a fitted curve and can also be used to depict locations along trajectories. Region-based visualization is a good way to depict the aggregated information over regions of a predefined granularity.

Time is a necessary property of geotagged social media data. Charts, such as stacked graphs and bar charts, are a good conventional method of visualizing linear time. For spatiotemporal visualization, spatiotemporal dot map or time-series snapshots could be used to express the dynamic spatial phenomenon following specific temporal sequences.

Regarding text visualization, the tag cloud is a good choice to represent the word frequency in the text. Word networks could be used to reflect the internal structure and semantic relationship in the texts. For example, the contextual relationships of words can be illustrated in a suffix tree, while the hierarchical relationship of the topics can be packed in circles.

4.3 Sentiment-constrained spatiotemporal semantic clustering of crowd behavior

Clustering is one of the commonly used data mining methods for knowledge discovery in large-scale datasets. It is a process of grouping data based on their similarity. Density-based clustering algorithms, especially DBSCAN and its variations, are very effective for detecting clusters with arbitrary shapes from noisy dataset, without prior knowledge of the number of clusters. The main idea of applying clustering algorithm to spatiotemporal data is to consider the spatiotemporal neighbors of objects and search for high-density areas in the feature space.

Crowd behavior in social media services is characterized by descriptions of similar or related emotions at specific times and places. The geotagged social media data contains timestamps, geographic locations, and text content, which makes it a valuable data source for exploring human crowd behavior. However, compared to traditional survey research, it is not easy to cluster geotagged social media data. In a target area, a large number of active users contribute social media data, and the content may vary from the birthday celebration at home to sharing sadness at school because of failing to pass the exam. Therefore, it is necessary to combine multi-dimensional features in an appropriate method to cluster this complex data.

For the exploration of human crowd behavior from geotagged social media data, the DBSCAN is extended in this study to the sentiment-constrained spatiotemporal semantic clustering algorithm (SCSTSC). The extension has made it possible to consider the spatiotemporal proximity, text similarity, and sentiment constraints.

4.3.1 Computational structure

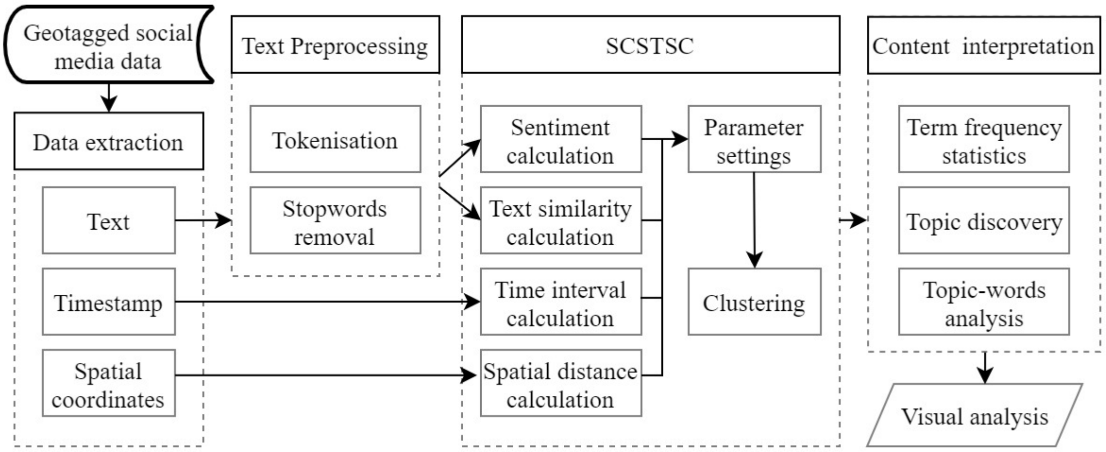


Figure 4.6 The workflow for the perception of human crowd behavior.

As shown in Figure 4.6 the overall workflow of perceiving human crowd behavior induced by social events from geotagged social media data is divided into four stages: data extraction, text preprocessing, clustering and content interpretation. The obtained results can be visualized.

- **Data extraction**. Data obtained from different social media services may contain various contents and metadata. But most geotagged social media data contains location information, timestamps, and text content. In the stage of data extraction, the geotagged social media data is filtered to retain the relevant location information, timestamp and text information, thereby reducing the data size and increasing the speed of subsequent data analysis.
- **Text preprocessing**. The time interval and spatial distance between geotagged social media posts can be calculated directly, while the text content needs to be converted into a word list for clustering and content interpretation. As mentioned in Section 4.2.1, the text preprocessing involves the tokenization and stop words need to be removed. The processed word list will be used for subsequent text similarity analysis and topic modeling analysis.
- **SCSTSC**. The spatial distance, time interval, text similarity, and sentiment value of posts are calculated separately and used to guide the clustering of geotagged social media data. The clustering algorithm is described in detail in Section 4.3.2.
- **Content interpretation.** Once the clusters are obtained, the word frequency is used as a statistical factor to apply LDA topic modeling introduced in Section 4.2.4 to analyze the content of the clustered data.

4.3.2 Sentiment-constrained spatiotemporal semantic clustering

The methodological core of our clustering task is DBSCAN which is an unsupervised clustering algorithm. It has two parameters, $\mathcal{E}_{distance}$ and N_{minPts}, as input, where $\mathcal{E}_{distance}$ shows the neighborhood search radius and N_{minPts} indicates the minimum number of points that need to be included in the search radius to construct density-based clustering. If the number of the points included in the search radius of a point reaches N_{minPts}, this point is regarded as the core point. The core point and its neighbor points within the search radius will be classified as a cluster. Other core points in the neighborhood will also be added to this cluster and the same procedure continues until no further core points can be found. This leads to a density-based clustering of spatial datasets.

In order to support clustering for spatiotemporal data, ST-DBSCAN uses two proximity metrics $\mathcal{E}_{distance}$ and $\mathcal{E}_{non-spatial}$ to measure the similarity of data points with one for the spatial distance between data points, and the other for non-spatial distances, such as time intervals. Therefore, ST-DBSCAN needs to define three parameters $\mathcal{E}_{distance}$, $\mathcal{E}_{non-spatial}$ and N_{minPts} to perform two-dimensional clustering. A point can be regarded as a core point if there are no less than N_{minPts} points in its two-dimensional neighborhood based on $\mathcal{E}_{distance}$ and $\mathcal{E}_{non-spatial}$.

The clustering of geotagged social media data for the perception of crowd behavior goes a step further. In addition to spatial proximity, time proximity, semantic proximity, sentimental constraints are added. This leads to the extended clustering algorithm SCSTSC (sentiment-constrained spatiotemporal semantic clustering), as shown in Figure 4.7. The algorithm requires spatial distance neighborhood ($\mathcal{E}_{distance}$), time interval neighborhood ($\mathcal{E}_{time}$), text similarity neighborhood ($\mathcal{E}_{textsimilarity}$), sentiment constraint ($\mathcal{E}_{sentiment}$), and a minimum number of points in the neighborhood (N_{minPts}) as input for geotagged social media data (D).

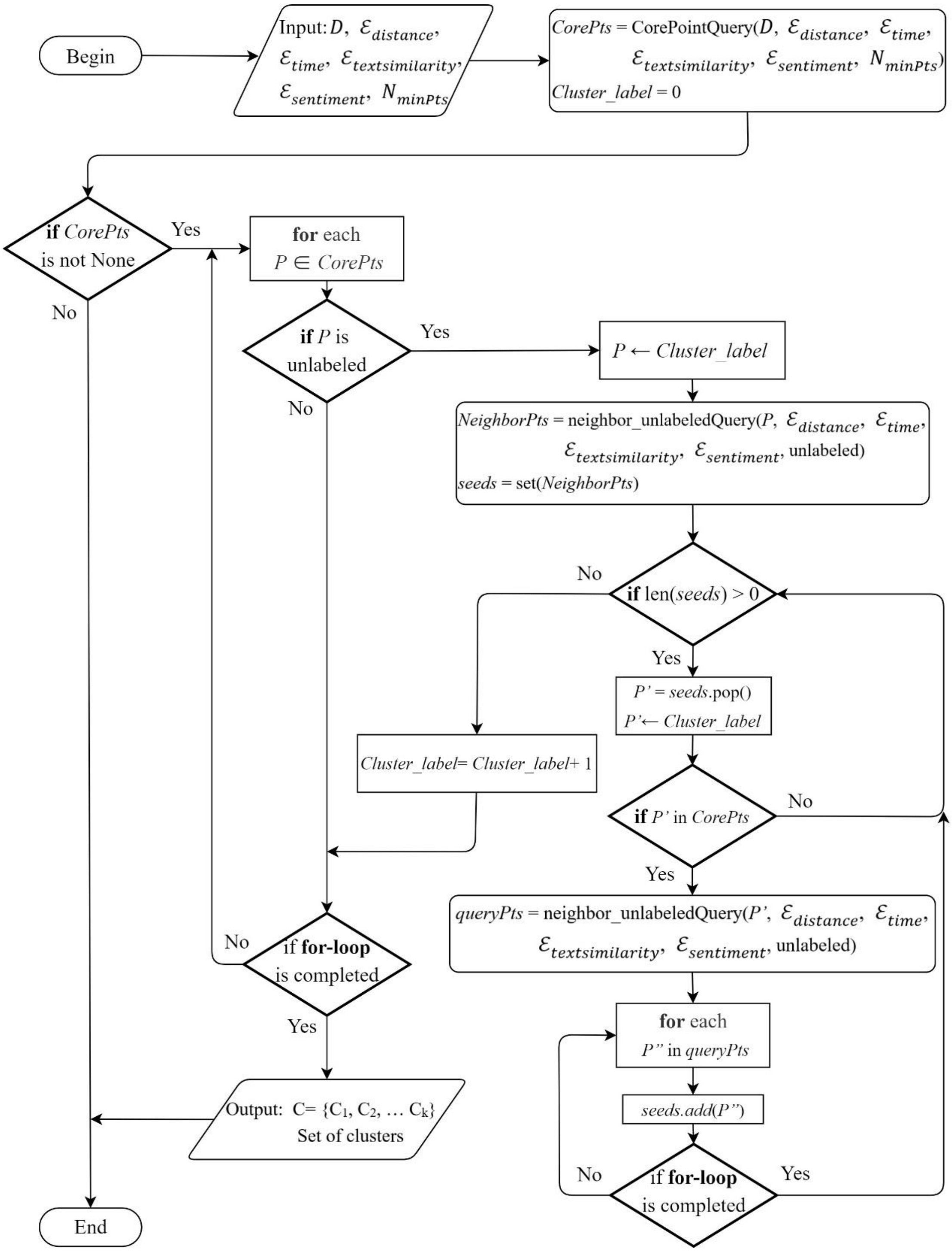

Begin
Input: D, $\mathcal{E}_{distance}$, $\mathcal{E}_{time}$, $\mathcal{E}_{textsimilarity}$, $\mathcal{E}_{sentiment}$, N_{minPts}
CorePts = CorePointQuery(D, $\mathcal{E}_{distance}$, $\mathcal{E}_{time}$, $\mathcal{E}_{textsimilarity}$, $\mathcal{E}_{sentiment}$, N_{minPts})
Cluster_label = 0
if CorePts is not None
Yes
No
for each $P \in$ CorePts
if P is unlabeled
Yes
No
P ← Cluster_label
NeighborPts = neighbor_unlabeledQuery(P, $\mathcal{E}_{distance}$, $\mathcal{E}_{time}$, $\mathcal{E}_{textsimilarity}$, $\mathcal{E}_{sentiment}$, unlabeled)
seeds = set(NeighborPts)
No
if len(seeds) > 0
Yes
P' = seeds.pop()
P'← Cluster_label
Cluster_label= Cluster_label+ 1
if P' in CorePts
No
Yes
queryPts = neighbor_unlabeledQuery(P', $\mathcal{E}_{distance}$, $\mathcal{E}_{time}$, $\mathcal{E}_{textsimilarity}$, $\mathcal{E}_{sentiment}$, unlabeled)
for each P" in queryPts
seeds.add(P")
No
if for-loop is completed
Yes
No
if for-loop is completed
Yes
Output: C= {C1, C2, … Ck} Set of clusters
End

Figure 4.7 The working diagram of sentiment-constrained spatiotemporal semantic clustering algorithm (SCSTSC).

As shown in Figure 4.7. D is the dataset after data extraction and text preprocessing. Each record $d \in D$ can be represented by tuples $[x, y, t, c, l]$, where x and y are geographic coordinates and t is timestamp, c is the word list after text preprocessing, and l is the clustering label that is predefined as noise at the beginning. The SCSTSC algorithm receives D as input and returns C. Each $d_l \in C$ in the result set has a label, which may be a cluster label or a noise.

The geodesic distance is used as the spatial distance and calculated with python-GeoPy package based on the method given by Karney (2013). The time interval is calculated based on the time stamp. The sentiment analysis method is introduced in Section 4.2.3. For the calculation of similarity between short texts in social media services, we adopt the cosine measure (Al-Anzi & AbuZeina, 2017). The short texts are converted to word vectors to represent word lists based on the bag-of-words model (BOW). Thus the similarity of two texts can be expressed as the cosine value of the angle $cos(\theta)$ between the corresponding vectors according to Equation (4.13).

$$sim(c_1, c_2) = cos(\theta) = \frac{v_{c1} \cdot v_{c2}}{\|v_{c1} v_{c2}\|} = \frac{\sum_{i=1}^{n} v_{c1}^{i} v_{c2}^{i}}{\sqrt{\sum_{i=1}^{n} {v_{c1}^{i}}^{2}} \sqrt{\sum_{i=1}^{n} {v_{c2}^{i}}^{2}}} \tag{4.13}$$

If the two word vectors are orthogonal, their cosine similarity is 0, and if they are identical, their cosine similarity reaches the maximum of 1. If the word list is empty after text preprocessing, it's semantic similarity to other texts is set to zero by default. In the SCSTSC algorithm, two geotagged social media posts are considered to be semantically similar if their text cosine similarity is greater than $\mathcal{E}_{textsimilarity}$.

The CorePointQuery() function retrieves each point to find those core points that contain no less than N_{minPts} points in the temporal, spatial, semantic and emotional neighborhood. The Neighbor_unlabeledQuery() function retrieves neighborhood points that are not yet grouped around a specific core point. The SCSTSC algorithm starts with calculating the core point set *CorePts*, and then traverses each core point to perform clustering judgment. If the visited core point is not labeled, it is assigned a new cluster with the unlabeled neighbor points and be labeled together. If there are other core points in this cluster, the unlabeled neighbor points of these core points will be added to this cluster and be labeled to realize the expansion of the cluster. By traversing all the core points, the clustering of geotagged social media data will be realized.

4.4 Perception of differences in population mobility induced by social events

Human behavior induced by a social event can be traced not only from the crowd sentimental messages, it can also be reflected from the spatial and temporal changes of the positions of the users' posts. The occurrence of major social events (such as regional conflicts, festivals, and political events) may trigger large-scale population mobility. The widespread social media services have made it possible to explore the changing spatiotemporal states of social media users and find the patterns and laws of population mobility induced by social events. The following sections are dedicated to the methodology of geospatial network analysis for the perception of population mobility patterns. Following the flow diagram given in Section 4.4.1, Section 4.4.2 introduces the method of using geotagged social media data to build the weighted directed population mobility network (PMN). Sections 4.4.3-4.4.8 introduce the involved analytical methods.

4.4.1 The flow diagram for the analysis of crowd mobility network

This section presents a geospatial network analytical framework for the exploration of regional population mobility patterns from social media data. Three major steps are involved as shown in Figure 4.8. Firstly, the regional population mobility difference is explored based on the population mobility difference ratio, PageRank algorithm, and attractiveness index. Secondly, the community detection method and rich-club coefficient are applied to further observe the spatial interactions between regions. Finally, the community activity index and attractiveness index are derived from population mobility patterns, which can be used to reveal the imbalance of population flows between communities.

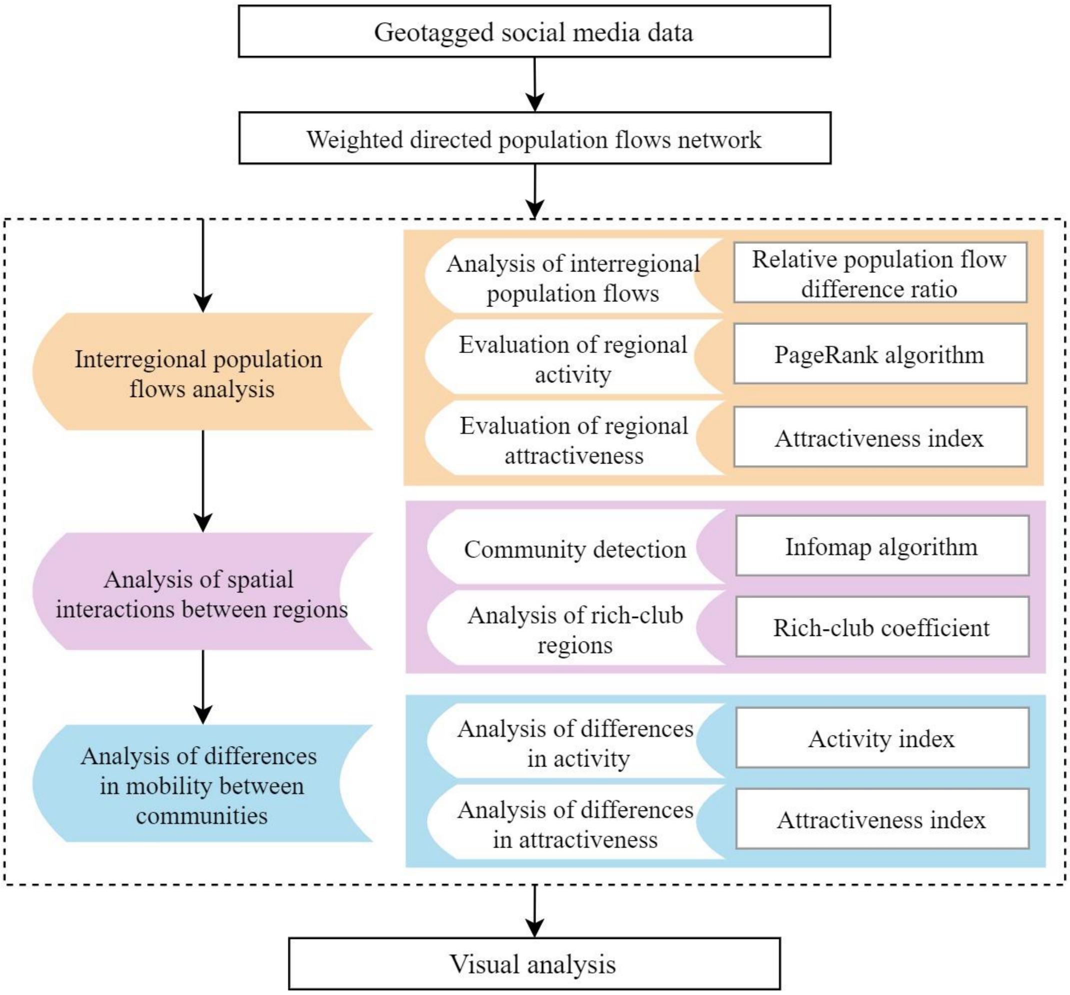

Geotagged social media data
Weighted directed population flows network
Interregional population flows analysis
Analysis of interregional population flows
Relative population flow difference ratio
Evaluation of regional activity
PageRank algorithm
Evaluation of regional attractiveness
Attractiveness index
Analysis of spatial interactions between regions
Community detection
Infomap algorithm
Analysis of rich-club regions
Rich-club coefficient
Analysis of differences in mobility between communities
Analysis of differences in activity
Activity index
Analysis of differences in attractiveness
Attractiveness index
Visual analysis

Figure 4.8 A flow diagram of crowd mobility exploration based on spatial network analysis.

4.4.2 Construction of Population Mobility Network (PMN) from geotagged social media data

Population mobility is a spatiotemporal phenomenon of the collection of human movement behavior, which can be represented by a weighted directed network consisting of the origin, destination, and other related attributes (e.g., mode of transportation, path, distance). Depending on the research purposes and research scales, the origin and destination can be either large administrative units (e.g., countries, cities) or small spatial entities (e.g., parking lots, bus stops, subway stations). According to the selected research scale and research unit, the weighted directed Population Mobility Network (PMN) can be constructed based on the changes in the spatiotemporal state of social media users. The process takes the following steps:

- Extraction of the geotagged social media posts of each user in chronological order from the raw data based on a unique user identity label and the timestamps of posts.
- Spatial overlay of the location information of each post with the selected regional divisions (e.g., city administrative divisions, national borders) to determine the region where the user posted information.
- Determination of whether the location records of the two sequential posts are made in the same region. If there is a region change, an interregional movement is recorded.
- Aggregation of all the interregional movement records of the same period to construct a weighted directed PMN, with the regions as network nodes and the direction and volume of population mobility as the direction and weight of the network edges.

By constructing a PMN in the event occurrence period and in the general period, the impact of major social event on population mobility can be explored from changes of population mobility patterns during the event occurrence period against the general period and the differing population mobility patterns in different regions during the event occurrence period.

4.4.3 Relative Flow Difference Ratio (**$RFDR$**)

The population mobility of a region may reveal a significant difference within a certain period of time. For example, during the World Cup game, there are far more incoming tourists to the host country of the World Cup than outgoing ones. This can be measured by the Relative Flow Difference Ratio ($RFDR$), which is a ratio between the difference of the inflow volume and outflow volume and the total flow volume:

$$RFDR = \frac{num_{inflow} - num_{outflow}}{num_{inflow} + num_{outflow}} \tag{4.14}$$

A positive $RFDR$ indicates that the inflow volume exceeds the outflow volume. When the population mobility in a region is balanced during a period, the $RDFR$ is approximately zero.

4.4.4 Attractiveness of a region

The attractiveness is another measure based on the expectation that people usually move to a more attractive region from a less attractive region. During the Oktoberfest, for example, Munich is more popular than other cities, attracting tourists from all over the world. J. Xu et al. (2017) defined attractiveness as the difference in the $RFDR$ between the event occurrence period and the general period. Whether the $RFDR$ is positive or negative in both periods indicates that the population continues to flow in or out, thus shows whether a region gets more attractive or not. Therefore, we consider two cases. If the $RFDR$ is positive or negative in both periods, we sum to express the attractiveness of the region; otherwise, we use the difference of $RFDR$ between the two periods:

$$Attractiverness = \begin{cases} RFDR_{event} - RFDR_{general}, if\ RFDR_{event} \times RFDR_{general} \leq 0 \\ RFDR_{event} + RFDR_{general}, if\ RFDR_{event} \times RFDR_{general} > 0 \end{cases} \quad (4.15)$$

If there are more people entering than leaving a region in both periods, then the attractiveness of the region is positive, and vice versa. Moreover, if there are more people leaving in the general period and more people entering in the event period, then the attractiveness is positive, and vice versa.

4.4.5 PageRank for regional activities

The PageRank, originally developed to rank the relevance of webpages for search engines has been widely used in social network analysis (Peng, Yang, Cao, Yu, & Xie, 2017), transportation planning (Q. Sun et al., 2019), and network security (Ramos, Lazar, Filho, & Rodrigues, 2017). As expressed in Equation (16), the PageRank takes the number and quality of hyperlinks between web pages as the main factors to analyze the importance of a web page (Page, Brin, Motwani, & Winograd, 1999). The basic assumption is that more important pages are often referenced by other web pages. The link from page A to page B is regarded as “page A votes for page B”, and the importance of the page is determined based on the source and the importance of the voters. The PMN is analogous to the internet in that more active regions in the PMN receive or transfer more populations through more routes from other regions. Hence, the PageRank can suitably measure the regional activities in the PMN during the event occurrence period.

$$PR(x) = \frac{(1-\sigma)}{n} + \sigma \sum_{i \in P_x} \frac{PR(Y_i)}{C_{out}(Y_i)} \quad (4.16)$$

where $PR(x)$ is the PageRank value of page x, P_x depicts the set of pages that link to page x, $PR(Y_i)$ is the PageRank value of page Y_i linked to page x, $C_{out}(Y_i)$ describes the number of links coming from page Y_i, n depicts the number of pages, and σ is the damping factor used to deal with pages that have no explicit outgoing nor incoming links. These pages are considered to be linked to all pages in the network, and the PageRank values of such pages are divided equally among all pages. The values of PageRank can be approximated with high accuracy through several iterations. The more links and the higher the PageRank value of the source page to page x, the larger the PageRank value of page x will be. The PageRank algorithm is implemented by the python-NetworkX package.

4.4.6 Community detection

As a typical feature of a complex network, community structure describes the characteristic that the connections within communities are relatively dense, while connections among the communities are relatively sparse. The recognition of network communities helps to disclose the latent relations between network nodes. Community detection methods, such as the label propagation algorithm (Raghavan, Albert, & Kumara, 2007; Z. Wu, Lin, Gregory, Wan, & Tian, 2012) and modularity-based algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008; M. E. J. Newman, 2006), are designed to explore the characteristics of different network types (e.g., topology network, binary network, and directed network) and widely used in various fields (e.g., social network analysis, biological network analysis, and internet network analysis).

The interregional PMN is a typical weighted directed network, and we can apply appropriate community detection methods to explore the interactions between regions. The Infomap algorithm (Rosvall & Bergstrom, 2007, 2008) performs well in weighted directed networks in a variety of comparative tests (Lancichinetti & Fortunato, 2009; Zhao Yang, Algesheimer, & Tessone, 2016) and provides a promising perspective to observe spatial interactions between regions from the flow of population between regions. For example, Thanksgiving Day in the United States is a day for family gatherings. In such an important social event, many young people who migrate to work and study in big cities will return to their hometowns to reunite with their families. Through community detection, the interregional migration pattern during Thanksgiving can be recognized, thereby revealing the potential influencing factors combined with regional development differences.

Based on information theory (Shannon, 1948), the Infomap algorithm is dedicated to identifying network communities by selecting the fewest bits to express the route generated by random walks in the network. Each node is encoded as the Huffman code (Huffman, 1952), and a node with a higher access frequency is assigned a shorter code. To divide nodes into different clusters, a two-level description strategy is adopted. Each cluster is given a unique name, and the nodes within each cluster are named with different Huffman codes, which can be reused in different clusters. Unlike a single-layer structure that does not consider community structure, the two-level description strategy allows the nodes within one cluster to share the same code, so that the nodes themselves can be depicted with fewer bit codes, and shorter average bits can be achieved. Regarding a community partition M of n nodes into m clusters, the average description length of a single step $L(M)$ is:

$$L(M) = q_{\curvearrowright}H(Q) + \sum_{i=1}^{m} p_{\circlearrowright}^{i} H(P^{i}) \tag{4.17}$$

where $H(Q)$ describes the information entropy of the clusters names; $q_{\curvearrowright}$ describes the probability that the random walk switches clusters (Shannon, 1948); $H(P^{i})$ depicts the information entropy of the movements within cluster i, including the exit code (a virtual node within each cluster expressing that the random walk is leaving the current cluster) for cluster i; and $p_{\circlearrowright}^{i}$ depicts the probability of movements within cluster i (including the probability of exiting cluster i). The access probability of each node is calculated by the "random surfer" method, which is similar to PageRank, and the possible partitions are explored based on a simulated annealing approach (Guimera & Amaral, 2005) and deterministic greedy search algorithm (Clauset, Newman, & Moore, 2004). The Infomap algorithm is applied to explore the community structure of the PMN during the general period and the event period separately.

4.4.7 Rich-club coefficient

The rich-club coefficient is a measure reflecting the spatial interactions between prominent regions. Facing the social and economic disparity among people in different societies and countries as early as the end of the 19th century, the 80/20 rule was proposed to describe the phenomenon that a selected minority of elements are responsible for the vast majority of the observed outcomes in many real-world settings. Accordingly, the rich-club phenomenon is defined to describe the tendency that prominent elements establish stronger interactions among themselves than would be expected by random chance. This phenomenon has proven to be applicable in many areas, including transportation networks (Colizza, Flammini, Serrano, & Vespignani, 2006), scientific collaboration networks, and interbank networks (De Masi, Iori, & Caldarelli, 2006). The rich-club coefficient was first defined based on the topology network (S. Zhou & Mondragón, 2004). Subsequently, Opsahl, Colizza, Panzarasa, and Ramasco (2008) improved the rich-club coefficient $\varphi_w(r)$ to make it suitable for a weighted directed network (Opsahl et al., 2008). All nodes in the network are ranked in terms of the rich parameter r. For each value of r, the nodes whose richness is larger than r construct a rich club. $E_{>r}$ denotes the number of edges among the members of rich club, $W_{>r}$ is the sum of weights of these edges, and $W_{l,rank}$ is the l^{th} ranked weight on all edges of the network. Then, we have $\varphi_w(r)$:

$$\varphi_w(r) = W_{>r} / \sum_{l=1}^{E_{>r}} W_{l,rank} \tag{4.18}$$

In addition, the rich-club coefficient $\varphi_{w,null}(r)$, which is obtained from the corresponding null model that is random but still comparable to the real network, is introduced as a benchmark for comparison (Colizza et al., 2006). Therefore, the rich-club effect is determined as follows:

$$\rho_w(r) = \frac{\varphi_w(r)}{\varphi_{w,null}(r)} \tag{4.19}$$

Only when the rich-club coefficient of the actual network $\varphi_w(r)$ is greater than that of its corresponding randomized network $\varphi_{w,null}(r)$, e.g. with a randomized distribution of all nodes, that is, only when the ratio $\rho_w(r)$ is greater than 1, can it be proved that there is a rich-club phenomenon in the network. For the weighted directed network, the richness parameter can be the out-degree (the number of links going out from a node) or out-strength (the sum of the weights attached to these links). High out-degree and out-strength prove the activity of a node because of its high external links and external participation in the network. In a PMN, the rich-club ratio with these two richness parameters is employed to access the spatial interaction pattern between the active regions during the event occurrence period.

4.4.8 Differences in mobility between communities

To further explore the differences in mobility between communities, the evaluation indexes are built based on the attractiveness and activity of the regions that make up the communities. The mean PageRank values of the regions included in the community is defined to describe the activity of the community:

$$PR(C_n) = \sum_{i\in C_n} PR(region_i)/Num_{C_n} \tag{4.20}$$

where $PR(region_i)$ depicts the PageRank value of region i, C_n is the region subgroup included in community n, Num_{C_n} is the number of regions included in community n, and $PR(C_n)$ is the PageRank value of community n. If the regions in the community have relatively high PageRank values, then the community’s PageRank value will be relatively high. Concerning the attractiveness of the community, the regional attractiveness is calculated based on the relative flow difference ratio in the inflow and outflow of the regional population, so the average regional attractiveness is a relative indicator and not suitable to quantify the attractiveness of the entire community. Concerning a more active region, its attractiveness should have a more significant impact on the overall attractiveness of the community. Therefore, the ratio of the regional PageRank value to the maximum PageRank value is used as the weight of regional attraction on the attraction of the community. The average weighted regional attractiveness within the community is used to define the attractiveness of the community:

$$Attractiveness_{C_n} = \sum_{i\in C_n} \frac{PR(region_i)}{Max_{PR(region)}} \times Attractiveness_{region_i}/Num_{C_n} \tag{4.21}$$

where $Max_{PR(region)}$ is the maximum PageRank value of the regions and $Attractiveness_{region_i}$ is the $Attractiveness$ of region i. The more active and attractive the regions in the community are, the more attractive the community is.

5 Case study implementations

In response to the conceptual framework of perceiving event-induced human behavior from geotagged social media data shown in Chapter 4, this chapter aims to implement the introduced ideas and methods in selected case studies. Three experiments are conducted to explore the inner behavior, the crowd behavior, and the population mobility pattern. The first experiment in Section 5.1 deals with the impact of specific social events on people's inner behaviors based on social sensing of a tragic event "Shanghai Stampede 2014" (R. Zhu et al., 2019). The second one in Section 5.2 is dedicated to extracting and clustering crowd behavior from geotagged social media data related to a positive event and a negative event. The final experiment in Section 5.3 explores how a social event such as the Chinese Spring Festival may induce the world's largest cyclic migration phenomenon (R. Zhu et al., 2020).

5.1 Inner behavior induced by a specific social event

In this experiment, we illustrate the feasibility of using social sensing to perceive human inner behavior induced by a specific social event. The proposed method in Section 4.2 has been implemented and tested using geotagged Sina Weibo data in Shanghai to gain insight into a particular event - the Shanghai Stampede Tragedy 2014 that happened on New Year's Eve.

5.1.1 Data

The Shanghai stampede tragedy 2014 began shortly before midnight of December 31 at about 23:35 local time and lasted almost 15 min. In total, 36 people were killed and 49 were injured. The research data used herein was collected via the Chinese social media platform - Sina Weibo. The official application programming interface was used to access and download public content from Sina Weibo. The detailed process of data acquisition was described by Jendryke, Balz, & Liao (2017). The whole dataset includes Weibo messages recorded from 29 December 2013 to 8 April 2015 - 11,784,344 records over Shanghai. Each record includes 33 attribute fields and a part of them is used for analysis in the current study.

During the preprocessing of Weibo data, four related fields (latitude, longitude, text, time of creation) were preserved, and other fields were excluded. After data preprocessing, two training datasets were constructed. Since the stampede event occurred on the New Year's Eve, we extracted data released from Sina Weibo in the month following the event on the basis of hashtags (#Shanghai The Bund Stampede Event (上海外滩踩踏事件)#, #The Bund Stampede (外滩踩踏)#, #Shanghai The Bund New Year's Eve Stampede (上海外滩跨年踩踏)#, #The Bund New Year's Eve Stampede Event (外滩跨年踩踏事件)#) as the event-related dataset. Correspondingly, Weibo data posted on New Year's Day of 2014, a year before the stampede, were selected to construct the event-irrelevant dataset. Table 1 summarizes two datasets (R. Zhu et al., 2019).

Table 5.1 Test datasets.

	The Number of Records	**Date**	**Description**
Event-related dataset	2093	*January* 2015	Posted with event-related hashtags
Event-irrelevant dataset	61,081	1 *January* 2014	Posted in the New Year 2014

5.1.2 Sensing inner behavior induced by the Shanghai Stampede Tragedy 2014

Extraction of event-related information

For the text-based binary classification, we used Term Frequency-Inverse Document Frequency (TF-IDF) features and five commonly methods - namely MultinomialNB, BernoulliNB, KNN, SVM with linear kernel, and SVM with radial basis function (RBF) kernel. In order to avoid over-fitting, we used grid-search with 5-fold cross-validation on the whole dataset to obtain the optimal hyperparameters for each classifier. In order to select the optimal classifier from the five candidate classifiers, we selected 90% of the dataset for training and 10% of the dataset for testing with a fixed random state (random state = 100). The optimal hyperparameters and classifiers evaluation are summarized in Table 5.2. Where α is the additive smoothing parameter of MultinomialNB and BernoulliNB. For KNN, K means the number of neighbors and *weights* describes the weight function used in prediction. About SVM, C is the regularization parameter for both SVM (Linear Kernel) and SVM (RBF Kernel) while γ is the kernel coefficient for SVM (RBF Kernel). This whole process was implemented by using the scikit-learn (a Python module for machine learning) (Pedregosa et al., 2011).

Table 5.2 The optimal hyperparameters and evaluation of classification methods.

Method	**Parameters**	**Precision**	**Recall**	**F1_score**
MultinomialNB	α = 0.5	1.000	0.685	0.813
BernoulliNB	α = 0.001	0.995	0.972	0.984
KNN	K = 3, weights = Euclidean distance	0.972	0.319	0.481
SVM (Linear Kernel)	C = 1	0.991	0.995	0.993
SVM (RBF Kernel)	C = 100, γ= 0.01	0.991	0.995	0.993

As shown in Table 5.2, the KNN and the MultinomialNB achieve a high precision, but their recall scores and *F1_scores* are relatively low. The BernoulliNB performs better, while both SVM methods achieve the best overall performance. This result fits the characteristics of the SVM explained by Joachims (2002). We chose SVM with linear kernel for its advantages in terms of fewer parameters and faster computing speed to classify Weibo data for the period of three months after the event occurred. 2425 event-related data records were obtained, as seen in Figure 5.1 (R. Zhu et al., 2019). We found that these data were mainly posted during the first week after the event. On the first day, up to 1729 Weibo messages were posted, and the attention to the event gradually decreased from the second day on. After one week, little attention was paid to this event, but on the 13th day and the 21st day, the concern increased significantly. On the 13th day, a related news item was released, according to which leaders of the Huangpu district abused public funds for extravagant meal and beverage in a nearby restaurant when the stampede happened. On the 21st day, the Shanghai municipality announced the investigation results of the stampede. Obviously, these two related events were the main reasons for the increased attention to the tragic event.

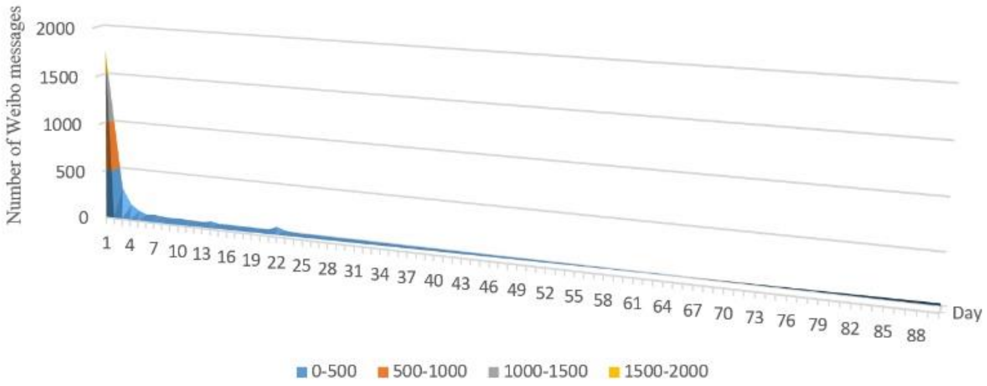


Figure 5.1 The number of relevant Weibo messages three months after the stampede event.

Figure 5.2 illustrates the frequency distribution of Weibo posts in the first 48 h after the event (R. Zhu et al., 2019). We can see that although this event happened at midnight, it stirred up some attention in the first few hours. A number of Weibo messages were posted from seven o’clock in the morning, peaking at nine o’clock. Early morning is a period during which all kinds of news travel fast. After 9 o’clock, the attention to the event slowly declined and reached the minimum at four o’clock the next morning before another wave of attention ensued.

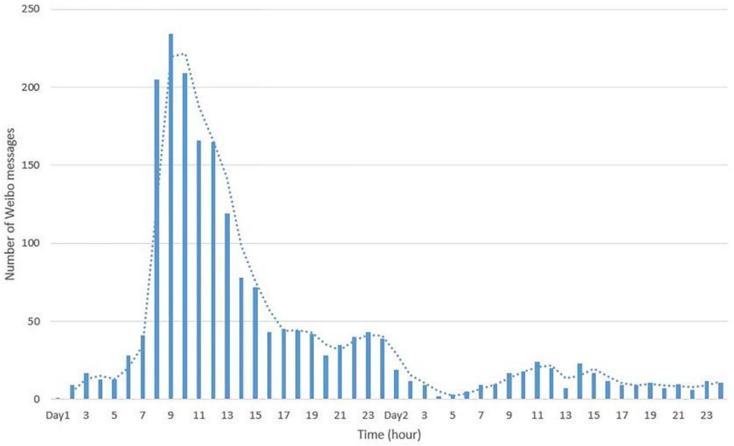


Figure 5.2 The number of related Weibo messages in the first 48 h after the stampede event.

In Figure 5.3, a time series of snapshots shows the evolution of Weibo posts (R. Zhu et al., 2019). A 3D spatiotemporal dot map in Figure 5.4 provides an alternative view of the distribution of event-related Weibo records in the first 24 h, where the vertical axis describes the time corresponding to 0-24 hours from down to top (R. Zhu et al., 2019). The black spot indicates the location of the stampede event. As can be seen from these two figures, a few Weibo data were posted near the event location in the first few hours, and the distribution became slowly dispersed. Seven hours later, people started to wake up and Weibo data grew rapidly and spread out. Twelve hours later, the news spread to almost the entire Shanghai city, causing the continued spread of the information over the whole area. Comparing the distributions from one week later to three months later, we find almost no further change.

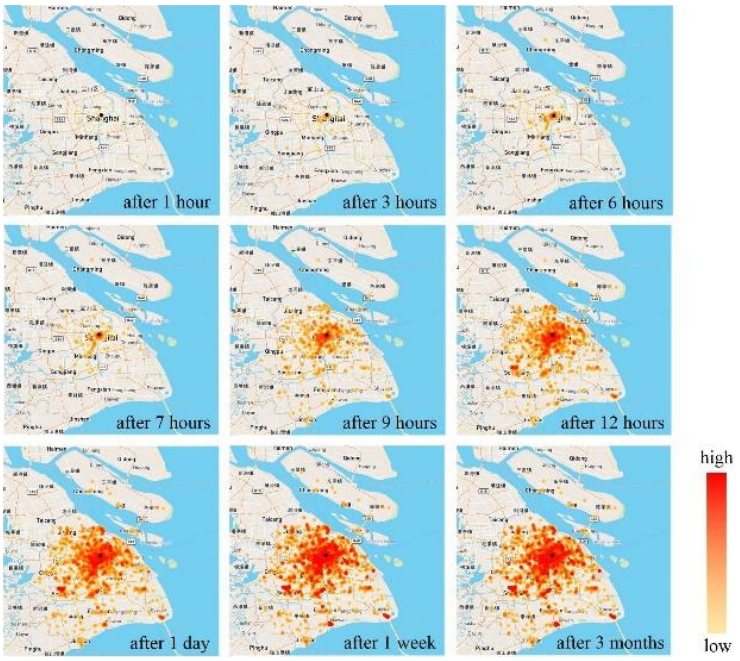
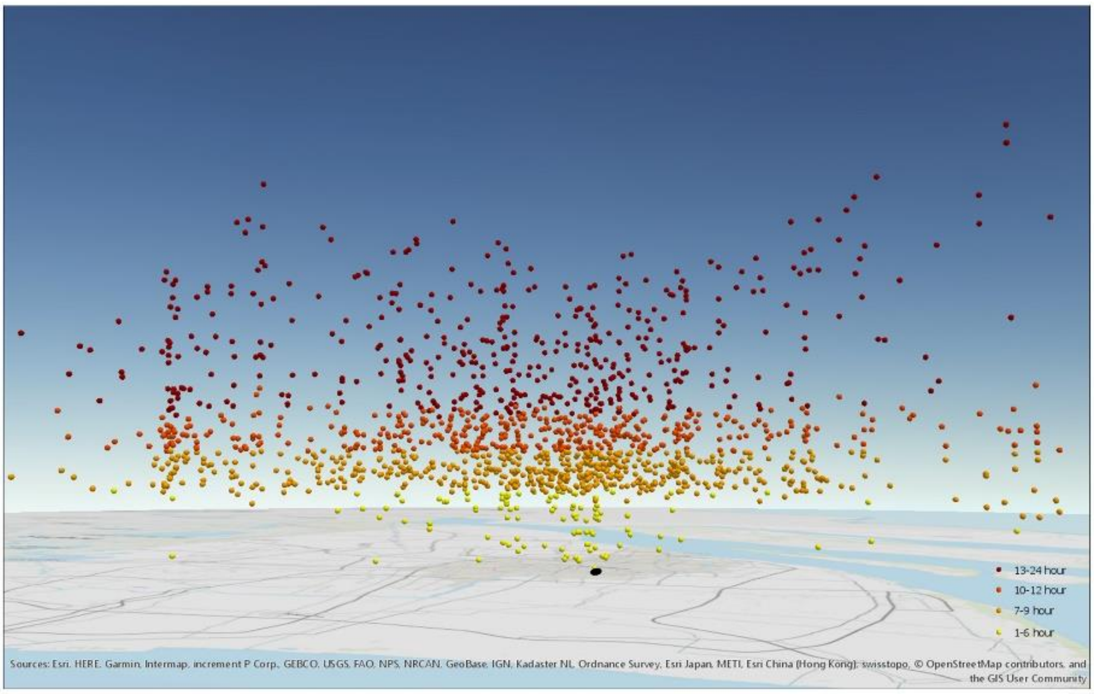


Figure 5.3 A time series showing the distribution of event-related Weibo records.



Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, © OpenStreetMap contributors, and the GIS User Community

Figure 5.4 A spatiotemporal dot map of event-related Weibo records in the first 24 h.

We used the torque-aggregation-function (torque-resolution = 2) provided by CartoDB to aggregate the geo-location of the event-related Weibo data. This function fetch points in clusters of 2x2 pixels based on the current zoom level and the numbers of points inside the clusters are used for data rendering. The aggregation result is visualized as a heatmap in Figure 5.5 (R. Zhu et al., 2019). We can see clearly some hotspots (i.e., yellow areas) distributed in different areas of the city. Then, we used the density-based spatial clustering of applications with noise (DBSCAN) algorithm to select areas with highest concentrations of Weibo messages. Based on the knowledge about the size of the Shanghai area and the number of event-related data, we tried with several different parameters and used one percent of the total number of event-related data (i.e., 24) as the minimum features per cluster and 1 km as the search distance. The subareas of the seven largest clusters are obtained and enlarged in Figure 5.5. Subarea 1 has the highest concentration of data and contains a black spot indicating the location of the stampede event. Subareas 2 and 7 are residential areas, and subarea 5 contains the Shanghai Hongqiao International Airport. The remaining three subareas are campus regions of universities and student dormitories. Students and intellectual groups tend to be more concerned about the event than other citizens.

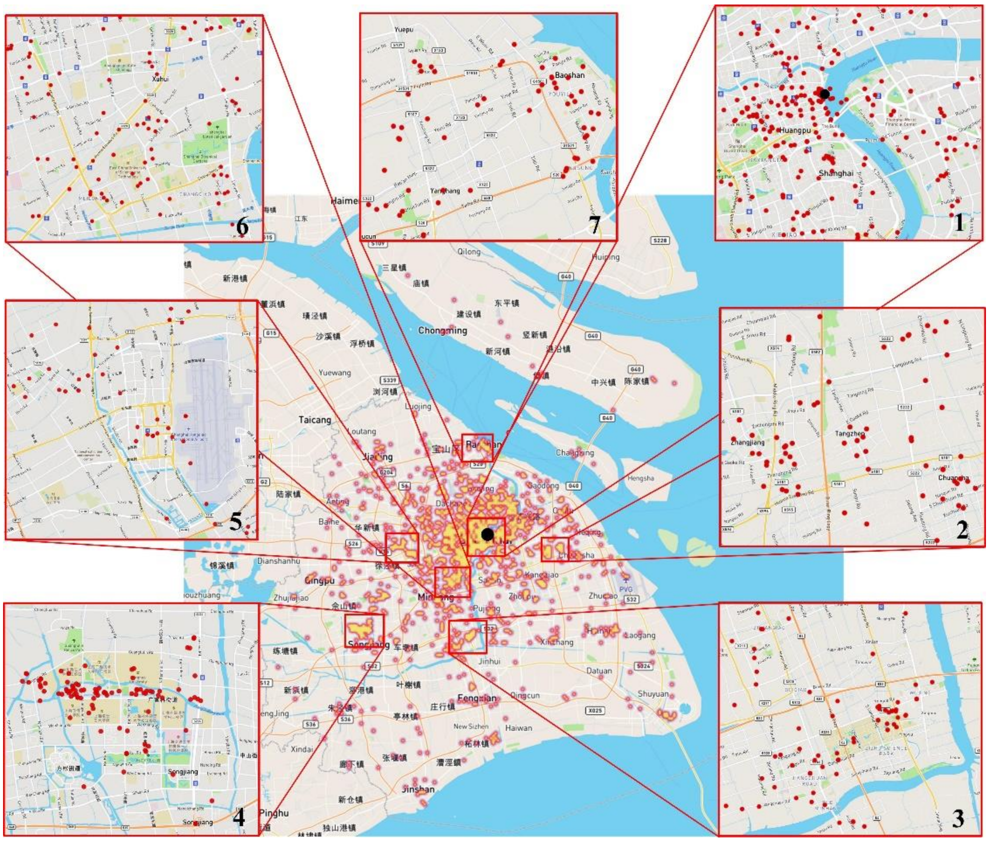


Figure 5.5 Spatial cluster analysis of event-related information.

Social sentiment analysis

Figure 5.6 illustrates the results of our statistical analysis on the daily average sentimental changes in different datasets around the week before and after the incident (R. Zhu et al., 2019). Blue bars represent the daily average sentimental value of the event-irrelevant information, red ones indicate the daily average sentimental value of all Weibo messages, and gray ones indicate the daily average sentimental value of the event-related information. Some fluctuation of the daily average sentiment in all Weibo data can be perceived. The daily average sentimental value declined gradually from 25 December and went up after 28 December. It peaked on 31 December but dropped back in the following days. A plausible reason for this change is that 25 December is Christmas day, so the public mood is slightly lighter. By the last day of the year, the public's mood peaks due to the coming new year. As can be seen, the Shanghai public's sentiment is very positive in daily life.

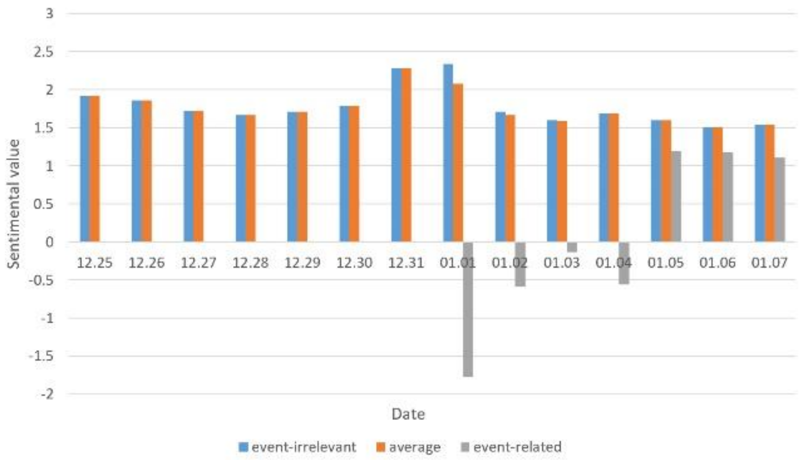


Figure 5.6 Daily average sentimental values in three datasets.

Since the 2014 Shanghai stampede event took place on New Year’s Eve, the average daily sentimental value in the event-related dataset was very low on 1 January, reaching around −1.8; the average daily sentimental value of event-irrelevant data and comprehensive data on the same day is above 2. Later, as time went on, it gradually eased up and became slightly positive on the fifth day after the event, but remained below the average daily sentimental value of comprehensive data.

The daily average sentimental value of event-irrelevant data and that of comprehensive data were equal before the event. However, on the first day after the event, the daily average sentimental value of comprehensive data was 11% lower than that of event-irrelevant data, 2% lower on the second day, and 0.6% lower on the third day. Then, public sentiment was slowly recovered to the level prior to the event. This indicates, to some extent, that people’s understanding of social events had undergone a process from perceptual to rational knowledge. The comparative analysis of the sentiment changes can support us in understanding the impact of events on the public mood, which may be used as the basis for judging the happiness of the public.

Based on the sentimental intensity and the locations where Weibo data were posted, a sentimental map was created, as shown in Figure 5.7, where black, red, and blue, represent negative, positive, and neutral sentiment, respectively; the color intensity corresponds to the intensity of the sentimental feeling (R. Zhu et al., 2019). There are obviously more black spots than red spots, which indicates a very negative impact of the stampede event on the public. As for the positive text content, we can see from the following social opinion mining that this is mainly comprised of the public expressing blessings for the victims.

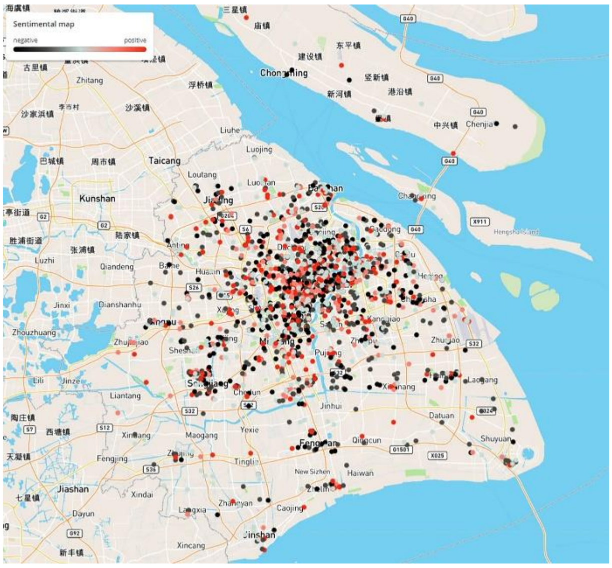
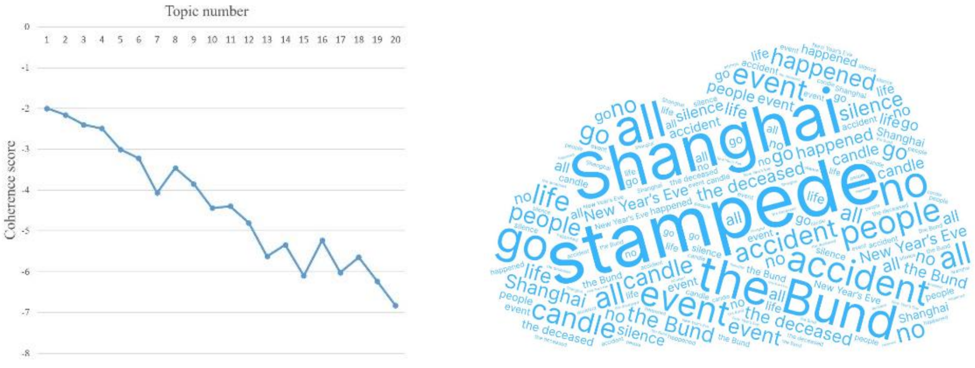


Figure 5.7 Sentimental map of the Shanghai stampede 2014.

Social opinion mining

At first, the term frequency (TF) was used and we performed the LDA calculation of event-related information with topic numbers from one to twenty. The 15 most frequently occurring words for each topic were selected and the UMass measure was adopted for the topic evaluation. The result is shown in Figure 5.8(a). LDA topic model with one topic is outstanding in comparison to the LDA topic model with other topic numbers. Then, the LDA topic model was run with 50 iterations for one topic. The results for the 15 most frequently occurring words on this topic are summarized in Figure 5.8(b), where the font size represents the probability of the word appearing on the topic (R. Zhu et al., 2019). Topic words such as “stampede”, “event”, and “accident” described the type of event. “Shanghai” and “the Bund” indicate the location of the event. “New Year’s Eve” reveals the time of the event. “Candle”, “life”, “silence”, and “the deceased” express people’s wishes for the dead. All these keywords reflect the characteristics of the event. They belong to the perceptual knowledge about the event including human feelings and representation of the event.



(a) (b)

Figure 5.8 (a) Topic evaluation diagram with topic numbers from one to twenty based on the term frequency (TF); (b) The weight chart of topic words for one topic based on TF.

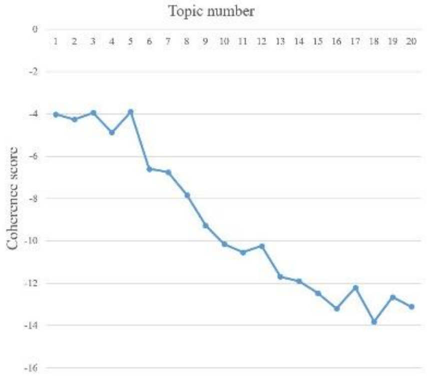


Figure 5.9 Topic evaluation diagram with topic numbers from one to twenty based on TF-IDF.

In a second step, TF-IDF was used to explore additional hidden knowledge according to the principle of reducing the weights of words such as “stampede” and “Shanghai” with high frequencies in all documents and increasing the weights of words which appear frequently only in parts of the documents. In this way, some local knowledge could be discovered. The result of topic evaluation for different topic numbers is shown in Figure 5.9, where five outstanding topics for event-related information are perceivable (R. Zhu et al., 2019). We used the LDA model with 50 iterations for these five topics. Each Weibo text has a corresponding probability for each topic and is assigned to the topic with the maximum probability. We calculated the average sentimental value of the Weibo messages included in each topic as shown in Table 5.3 (R. Zhu et al., 2019).

Table 5.3 Sentimental values of five outstanding topics.

Topic	Topic Word Content	Interpretation	Relationship	Sentimental Value
1	candle, event, report, event investigation, happened, Shanghai, people, leader, accident, that night, all, officers, went, eat, restaurant	Local officers ate dinner at nearby restaurants during the night of the incident.	Related event	-0.85
2	report, event investigation, survey result, event, candle, strong, cancel, light show, crowd, happened, that night, people, Shanghai, all, the deceased	Because of the stampede event, the light show was canceled in case of another security incident.	Causal relationship	-1.87
3	survey result, leader, people, restaurant, event, die, Shanghai, district mayor, dismissed, go, Huangpu district, courage,	The government handled the incident, dismissed the mayor, this may be related to a previous restaurant event.	Causal relationship	-3.20

	happened, accident, the Bund			
4	accident, Shanghai, the Bund, people, event, center, crowd, happened, mad, share, quick, come to see, candle, view, all	1. Overcrowding caused a stampede. 2. The public spreads event-related information through social media.	1.Reason 2.Public concentration	-0.55
5	candle, event, people, accident, Shanghai, all, go, the deceased, rest in peace, life, happened, the victim, strong, the Bund, silence	The reaction of the people to the incident, wishing that the dead rest in peace, hoping people can be strong.	Public reaction to death people	-1.14

As shown in Table 5.3, after excluding the description of the event itself, we can obtain some interesting events or knowledge from these topic words separately. These results include both the public attitude and people's judgments and inferences about the stampede event.

- In topic one, "event investigation", "that night", "officers", "have dinner", and "restaurant" show that the public was aware of a restaurant event which officers attended that night. Official reports show that some local officers used public funds to dine at a nearby restaurant that night, which caused public dissatisfaction. The stampede event also connected the public's attention to this restaurant event.

- In topic two, the "crowd" shows the reason for this stampede event. "Cancel", "light show", and "crowd" illustrate that, possibly due to concerns related to crowding, the lantern show was canceled. This is a causal event.

- In topic three, "district mayor", "dismissed", "Huangpu district", "restaurant", and "event" show that the local mayor was dismissed, which may be due to this event and the restaurant event. This should be a causal event.

- In topic four, "people" and "crowd" describe the cause of the stampede. Additionally, "share", "quick", "come to see", and "view" are some words used by the public to spread news of this stampede event to others. This topic shows the public dissemination of this stampede event.

- In topic five, "the deceased" and "victim" refer to death in this stampede event. "Rest in peace", "life", "strong", and "silence" are the reactions of the public to these victims. This topic shows the public's wish of peace for the victims of the incident.

It can be seen that the average sentimental value for each topic is negative. This result further illustrates the negative impact of this stampede tragedy on the public. Meanwhile, we can see that the average sentimental values of different topics are significantly different. This implies that the public's emotional responses to different topics on the same event are different. On topic 4, the public did not show strong negative emotions when discussing the cause of the incident and disseminating information about the incident. However, on topic 3, when the discussion involved the governmental penalty of the guilty officials for the stampede tragedy and the related restaurant event, the public expressed strong negative emotions.

5.1.3 Discussions

It is easy to extract event-related information according to related hashtags for events that receive a high level of attention or have a great impact. Together with event-irrelevant information, we tuned the hyperparameters and performed cross-validation to select the optimal classifier from several commonly used classification methods to obtain other related but unlabeled data. Both methods based on SVM achieved the best performance at the same time, while KNN and MultinomialNB did not perform well for the supervised learning of small-scale sparse matrices. However, for an event without a hashtag, we can refer to Gao, Wang, Padmanabhan, Yin, & Cao (2018) to filter the data through keywords, and an event-related dataset could then be constructed through manual annotation.

As seen in Figures 5.3 and 5.4, event information can spread rapidly through the internet. As the Shanghai stampede 2014 occurred at midnight and most people were sleeping at the time, the event-related information spread slowly in the first few hours until most people gradually woke up and communication of the event proliferated rapidly. It took just two hours for the event-related information to cover the entire Shanghai area, indicating both the wide reach of the internet and the great impact of this event. The impact on the everyday life of the public caused by the event is centered on the site of the incident and continues to spread outwards in space. To a certain extent, this reflects the law of network transmission of emergencies, which is spatially dispersed from the incident site to the outside.

A comparative analysis of the average daily sentimental value of the event-related information, event-irrelevant information, and comprehensive information could be used to assess the impact of the event on the public. As shown in Figures 5.6, on the first day after the stampede event, the average daily public sentimental value dropped by 11% due to this event. As time went on, the public sentiment gradually regressed and concern for this event declined. After about one week, the public sentiment resumed its normal state. The comprehensive statistical analysis of different events can help the government and related agencies to assess and predict the public sentimental changes and the recovery cycle so that the government can timely issue targeted response policies if a similar situation reappears.

As shown in Figure 5.8(b) and Table 5.3, using the LDA model based on TF, we can extract a few keywords to summarize the social event. Besides, using the LDA model based on TF-IDF, we can determine the public reaction and related hidden events. However, implicit knowledge needs to be obtained by parsing the topic words of each topic. The topic evaluation model based on the coherence score can help us determine the appropriate number of topics. In addition, the combination of topic modeling and sentiment analysis may facilitate the understanding of public sentimental attitudes toward different topics on the same event. In response to the Shanghai stampede, the public sentimental attitude towards each topic was negative. With respect to more controversial events (e.g., policy development, marketing activities), this approach can help deepen the understanding of diverse views held by the public.

5.2 Sentiment-constrained spatiotemporal semantic clustering for crowd behavior

In this experiment, we illustrate the feasibility of using social sensing to perceive human crowd behavior induced by a social event. The proposed method in Section 4.3 is implemented and tested using geotagged Sina Weibo data in Shanghai to gain insight into a negative event (Shanghai Stampede Tragedy 2014) and a positive event (New Year's Eve 2015).

The general introduction about the Shanghai Stampede Tragedy 2014 is described in 5.1.1. With this event as the background, this study uses negative-sentiment-constrained spatiotemporal semantic clustering algorithms to explore the human crowd behavior during and after the event.

When the Shanghai stampede 2014 happened, it happened to be the New Year's Eve 2015. In many countries and regions, people will gather at midnight to celebrate New Year's Eve and the arrival of the new year. This is a positive cultural event. This study uses positive-sentiment-constrained spatiotemporal semantic clustering algorithms to explore the manifestation of human crowd behavior in the geotagged social media data during the New Year's Eve 2015.

5.2.1 Data

The Shanghai stampede 2014 happened on the same night as New Year's Eve 2015. Sina Weibo data on Shanghai for December 31, 2014 and January 1, 2015 were extracted for subsequent processing and analysis. During the preprocessing of Weibo data, four related fields (latitude, longitude, text, and time of creation) were preserved and other fields were excluded.

Figure 5.10 shows the results of sentiment analysis of Sina Weibo posts on December 31, 2014, and January 1, 2015. The blue line shows the change in the number of Sina Weibo posts per hour. The red, yellow, and gray lines indicate the changes in the number of positive, neutral, and negative Sina Weibo posts per hour, respectively. The green line presents the change of the mean value of the sentiment of Sina Weibo posts every hour. The hourly sentiment mean in Weibo indicates that Weibo users have been showing a generally positive sentimental tendency over these two days. The number of Sina Weibo posts per hour peaks on New Year's Eve, as well as the number of positive Weibo posts and neutral Weibo posts per hour. However, the number of negative Weibo posts did not increase significantly during the Shanghai stampede event. In addition, the number of Weibo posts per hour experienced a process of sag and then swell from 0:00-7:00. Then, Weibo users are active from 7: 00-24: 00. This is in line with human physiological habits. Usually, people will go to sleep at midnight and wake up gradually at about 7 am to start a new day of life. 0-7 o'clock can be regarded as the dormant period of social intelligent sensors (people), and 7 o'clock-24 o'clock is the active period of social media users.

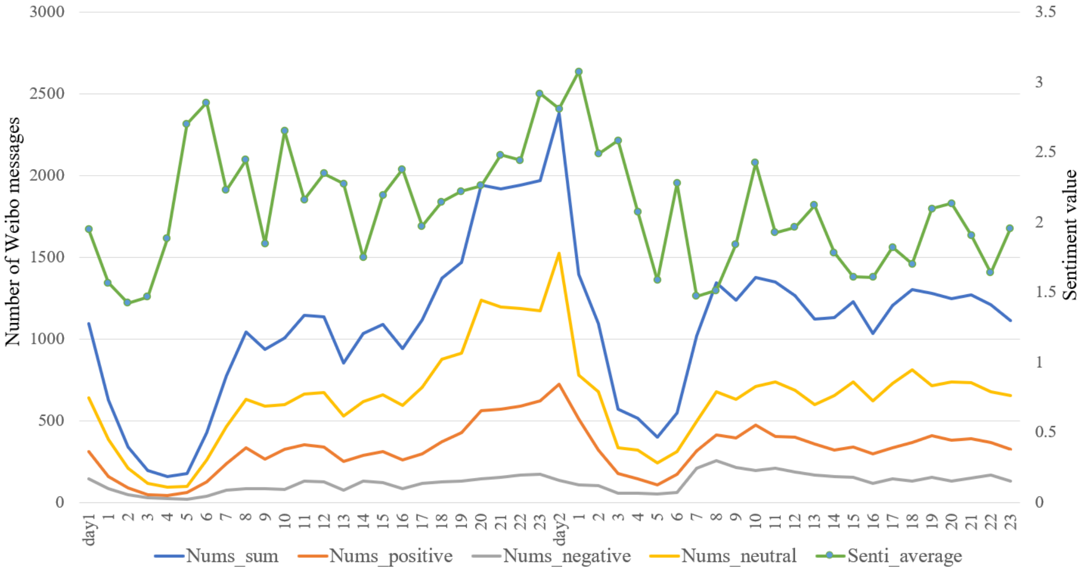


Figure 5.10 The results of sentiment analysis of Sina Weibo posts on December 31, 2014, and January 1, 2015.

Since the Shanghai stampede 2014 was a sudden incident that occurred at midnight, and the activity of Weibo users at this time gradually decreased, it may lack an immediate response. In the morning, when people wake up gradually, the posts related to the Shanghai stampede event on social media may increase accordingly. Therefore, we conducted exploratory cluster analysis on geotagged Sina Weibo data from three time periods (i.e., 23:30-00:30, 00:00-7:00, 7:00-8:00). For the New Year's Eve 2015, we performed a spatiotemporal semantic clustering analysis with positive sentimental constraints on geotagged Sina Weibo data at one hour midnight (23:30-00:30).

5.2.2 Sensing crowd behavior induced by Shanghai Stampede Tragedy 2014 and New Year's Eve 2015

Spatiotemporal semantic clustering analysis with negative sentimental constraints

First of all, we performed the spatiotemporal semantic clustering with negative sentimental constraints on the geotagged Sina Weibo data of one hour on New Year's Eve 2015 (23:30-00:30). Multiple sets of parameters are shown in Table 5.4. However, tests relying on multiple sets of parameters failed to extract any clusters from this hour of data. There is no concentrated discussion of certain events with negative emotions in a certain place during this time period. Therefore, we believe that the occurrence of this stampede incident did not immediately induce a strong reaction from social media users. Because this event occurred at midnight, related information about this sudden event was constrained by people's work and rest patterns and therefore difficult to spread effectively.

Table 5.4 Multiple sets of parameters for the clustering test.

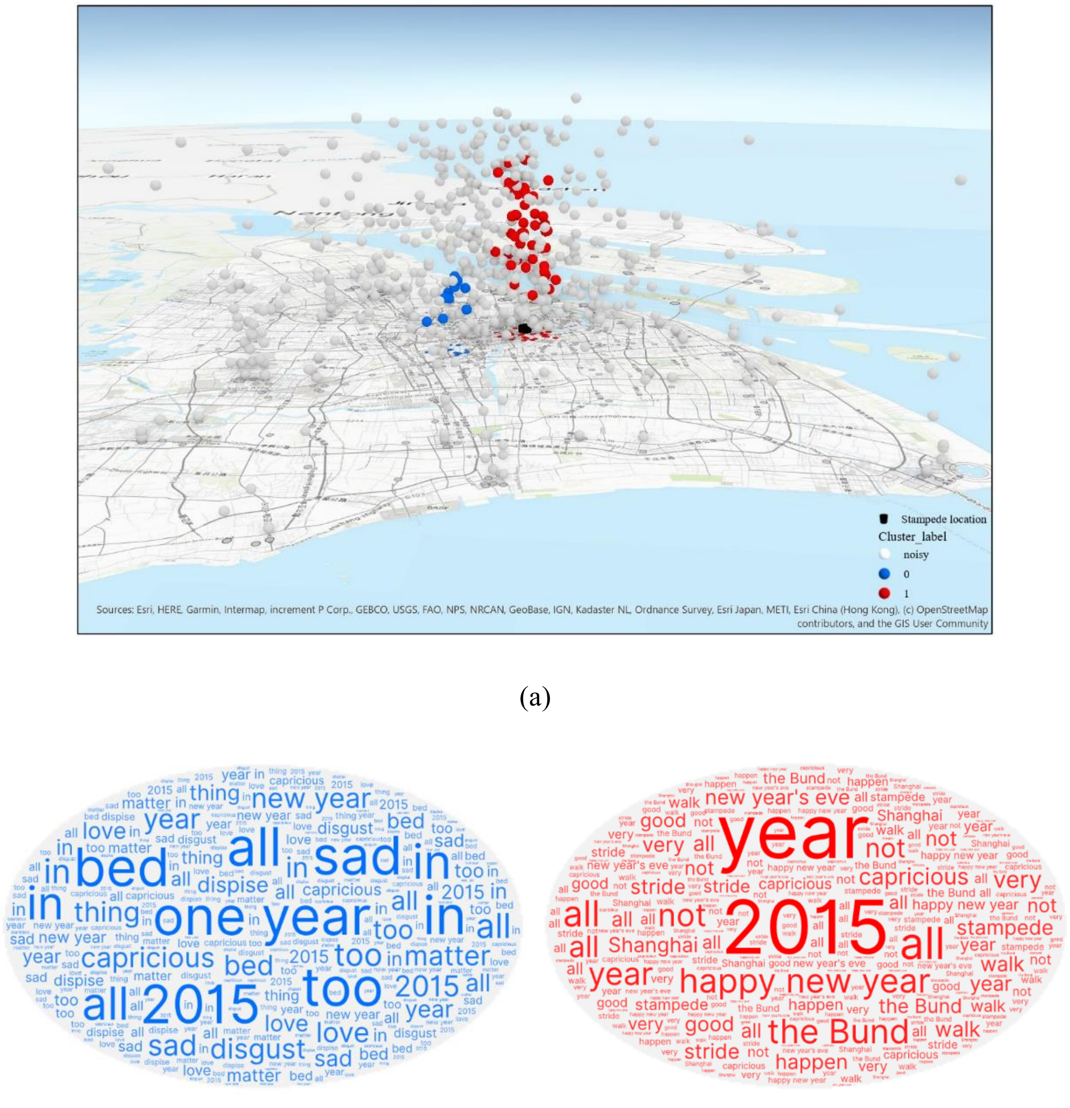
Test	$\mathcal{E}_{time}$	$\mathcal{E}_{distance}$	$\mathcal{E}_{textsimilarity}$	$\mathcal{E}_{sentiment}$	N_{minPts}	Clusters
1	10mins	3 km	0.50 (cos 60°)	negative	3	0
2	10mins	3 km	0.34 (cos 70°)	negative	3	0
3	10mins	3 km	0.34	negative	5	0
4	15mins	3 km	0.34	negative	5	0
5	15mins	5 km	0.34	negative	5	0
6	15mins	5 km	0.34	negative	3	0

In the next step, we carried out the spatiotemporal semantic clustering with negative sentiments from the geotagged Sina Weibo data within the first 7 hours after the Shanghai stampede event, that is, January 1, 2015, 0:00-7:00. From the change in the number of Weibo posts per hour in Figure 5.10, it can be seen that the number of Weibo posts in these 7 hours at night is less. Therefore, this test lengthened the time interval and appropriately increases the minimum number required for clustering. The clustering parameters were set to $\mathcal{E}_{time}$: 1 hour, $\mathcal{E}_{distance}$: 3 km, $\mathcal{E}_{textsimilarity}$: 0.34, $\mathcal{E}_{sentiment}$: negative, N_{minPts}: 10. Two data clusters were obtained as shown in Figure 5.11.

Figure 5.11(a) shows the distribution of geotagged Sina Weibo posts in these seven hours as dots in a space-time cube with the height corresponds to the time span of 0:00-7:00. The black spot indicates the location of the stampede event. The blue and red ones represent the spatiotemporal distributions of the two clusters. Figures 5.11(b) and 5.11(c) respectively show the 15 most frequently occurring topic words obtained from the topic analysis of the two clusters, where the font size represents the probability of the word appearing on the topic.

- The cluster data with label 0 has a time span of 00:51-02:15, and the location is the residential area (e.g., Xinqiao Village, Guilin Village). Topic words such as "one year", "2015", "new year" and "year" described the time while "bed" reveals where the user posted. "sad", "capricious", "disgust", and "despise" express people's strong negative emotions. "all", "in", "too" emphasize these negative emotions. On the contrary, "love" exudes hope for life. These keywords give such a scene, on New Year's Eve, some people are lying in bed to summarize the past year, vent their unpleasant experiences, and look forward to the new year.
- The cluster data with label 1 has a time span of 01:21-05:40, and the locations of these posts are closely around the occurrence of the stampede event. Topic words such as "New Year's Eve", "2015", "year", and "happy new year" reveal the time

of the event. “Shanghai” and “the Bund” indicate the location of the event. “stampede” describes the type of event. “not”, “good” express people’s feelings for this event. All these keywords reflect the characteristics of the event. These topic words belong to the perceptual knowledge about the event including human feelings and representation of the event. It can be seen that although the stampede event occurred at midnight, most people were sleeping for the next 7 hours. But this incident still induced discussions among people nearby on social media at this time.


Stampede location
Cluster_label
noisy
0
1
Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), (c) OpenStreetMap contributors, and the GIS User Community
one year
all
sad
in
bed
too
2015
capricious
disgust
year
2015
all
not
the Bund
happy new year
stampede
Shanghai

(a)

(b)

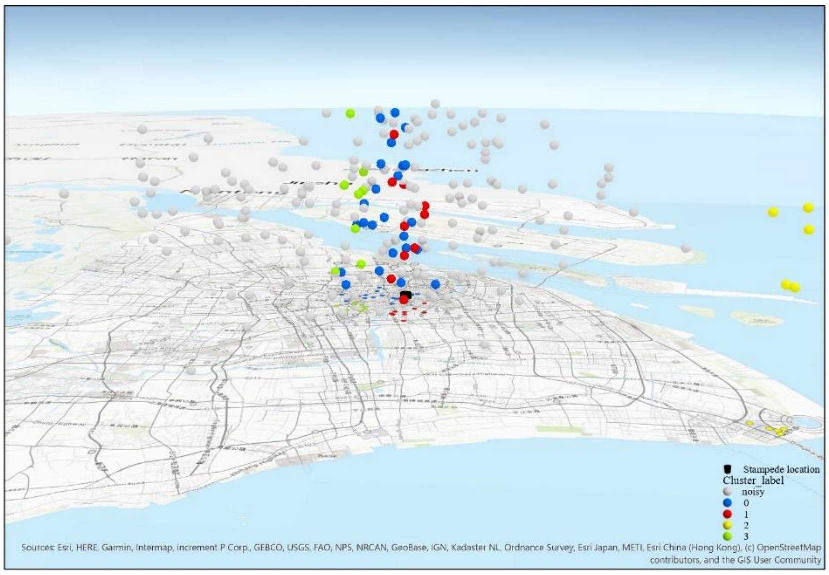
(c)

Figure 5.11 Clustering results of the Weibo posts within 7 hours after the Shanghai stampede event. (a) A space-time dot map showing spatiotemporal semantic clusters with negative sentiments; (b) The weight chart of topic words for cluster 0; (c) The weight chart of topic words for cluster 1.

Analog to the previous step, we carried out the spatiotemporal semantic clustering with negative sentiments from the geotagged Sina Weibo data in the eighth hour between

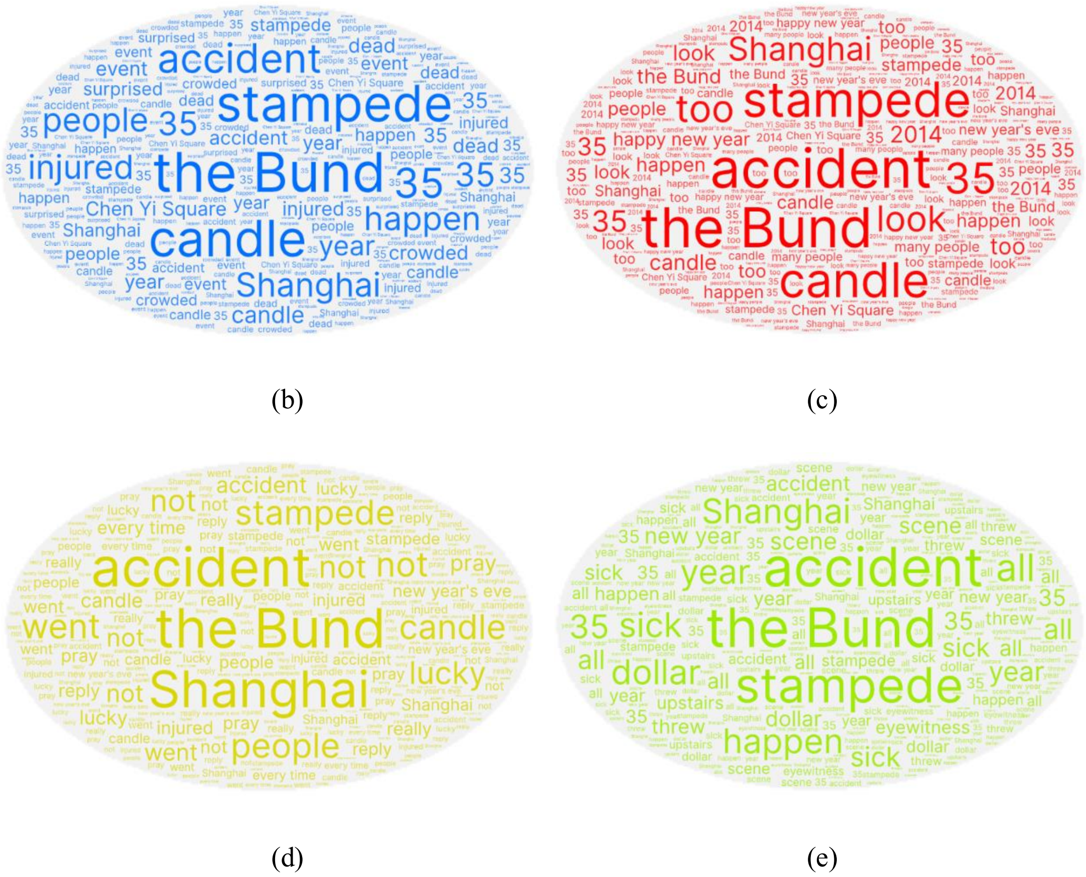
7:00 and 8:00 January 1, 2015 after the Shanghai stampede event. The clustering parameters were set to $\mathcal{E}_{time}$: 30 minutes, $\mathcal{E}_{distance}$: 3 km, $\mathcal{E}_{textsimilarity}$: 0.34, $\mathcal{E}_{sentiment}$: negative, N_{minPts}: 5. As shown in Figure 5.12, four clusters were obtained.

Figure 5.12(a) shows the distribution of geotagged Sina Weibo posts in this hour in a space-time cube. The black spot indicates the location of the stampede event. The blue, red, yellow, and green dots represent the spatiotemporal distribution of the four clusters. Figures 5.12 (b-e) respectively show the 15 most frequently occurring topic words obtained from the topic analysis of the four clusters, where the font size represents the probability of the word appearing on the topic. The topic words of the four clusters clearly represent the crowd reaction to the stampede event on social media (i.e., description of the event, public emotions, and attitudes). Among them, Cluster 0 (blue) containing the most posts is located around the place where the event occurred. Cluster 1 (red) and 3 (green) are near the incident. The possible reason is that similar scenes are more likely to induce event-related associations and discussions, making the Shanghai stampede event have stronger impacts on nearby people. This is a driving force to induce nearby people to share their own knowledge. In particular, people in the place where the event occurred are more susceptible to this type of effect, resulting in the largest clustering of posts. Cluster 2 (yellow) contains posts from the campus and dormitory area of two schools (i.e., Shanghai Ocean University and Shanghai Maritime University), reflecting the strong attention of the student community to a social event. One thing worth noting is that “eye witness”, “scene”, “threw”, “dollar”, and “upstairs” appeared in the topic words of cluster 3, describing the phenomenon that witnesses saw someone throwing money upstairs. This reflects some reasoning of eyewitnesses about the potential causes of the stampede incident. The Shanghai police deliberately investigated the matter and announced the results of the investigation on January 2. In fact, a local bar threw dozens of vouchers similar to US dollars onto a building near the place where the trample occurred. But this behavior occurred after the stampede, therefore, is not a cause of the stampede event.



Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), (c) OpenStreetMap contributors, and the GIS User Community

(a)



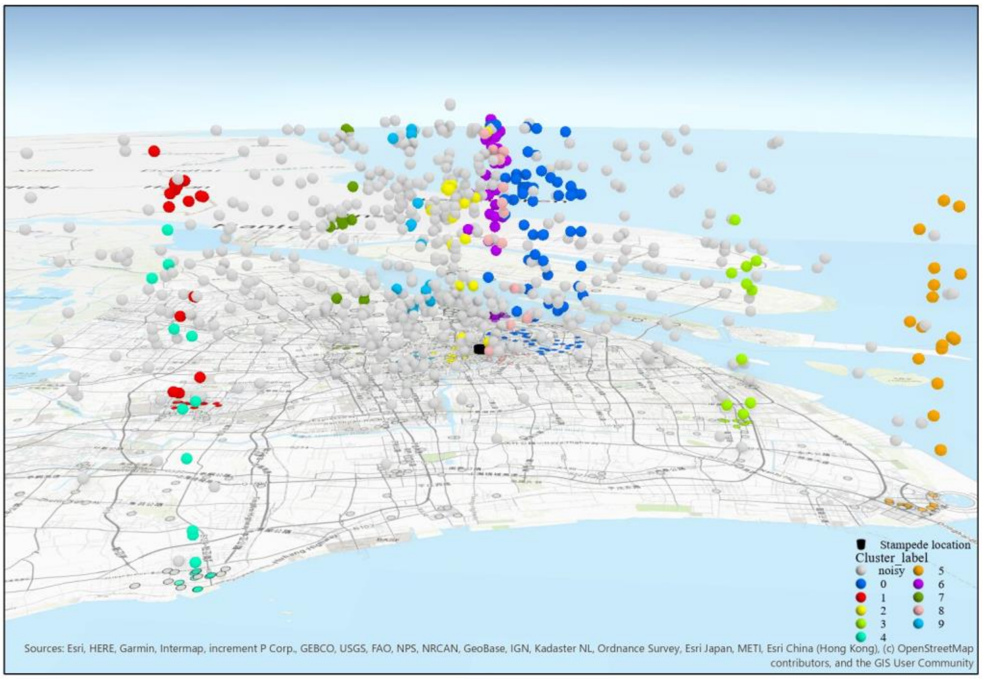
(b) (c)

(d) (e)

Figure 5.12 Clustering results of the Weibo posts between 7:00 and 8:00 in the following morning after the Shanghai stampede event. (a) A space-time cube dot map showing four spatiotemporal semantic clusters with negative sentiments; (b)-(e) The weight charts of topic words for cluster 0-3.

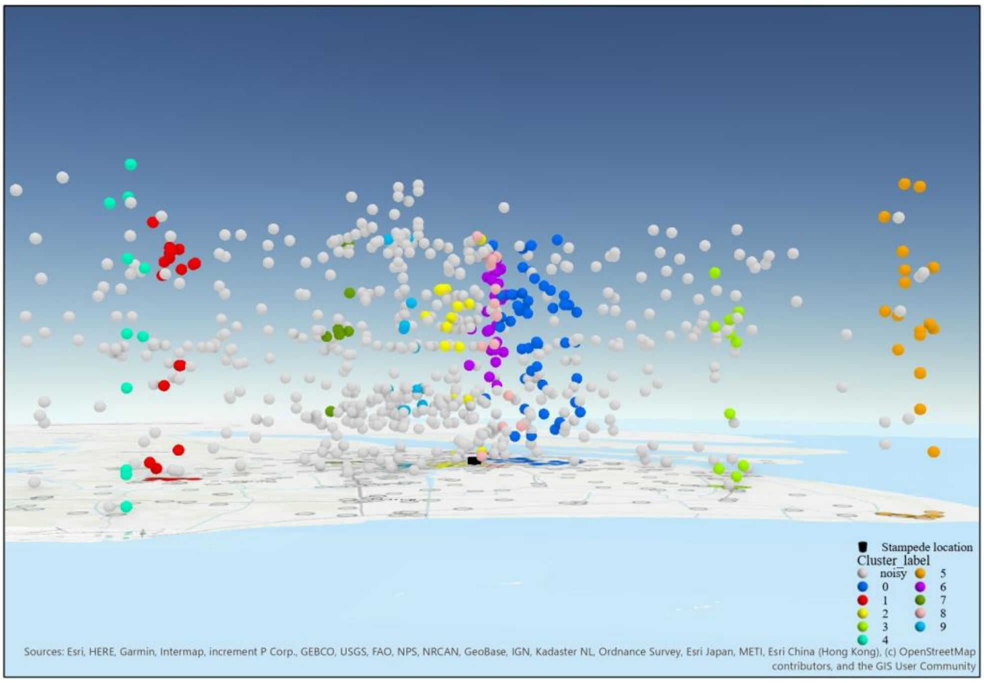
Spatiotemporal semantic clustering of posts with positive sentiments

Here, we performed the spatiotemporal semantic clustering of posts with positive sentiments from the geotagged Sina Weibo data for the hour around the New Year's Eve (23:30-00:30). The clustering parameters were set to $\mathcal{E}_{time}$: 30 minutes, $\mathcal{E}_{distance}$: 3 km, $\mathcal{E}_{textsimilarity}$: 0.34, $\mathcal{E}_{sentiment}$: positive, N_{minPts}: 10. A total of ten clusters were obtained as shown in Figure 5.13. Table 5.5 lists the 15 most frequently occurring topic words in each of the ten clusters and their locations. Five clusters are concentrated in the center of Shanghai, and the remaining five clusters are scattered in the east, west, south, and north of Shanghai. Although the locations of these datasets contain multiple types (scenic areas, schools, residential areas, commercial streets), the corresponding topic words all describe New Year's Eve activities and people's good wishes for the New Year. The clustering results of positive sentiments within this hour reflect the general mood of New Year's celebration among the local residents in the city.



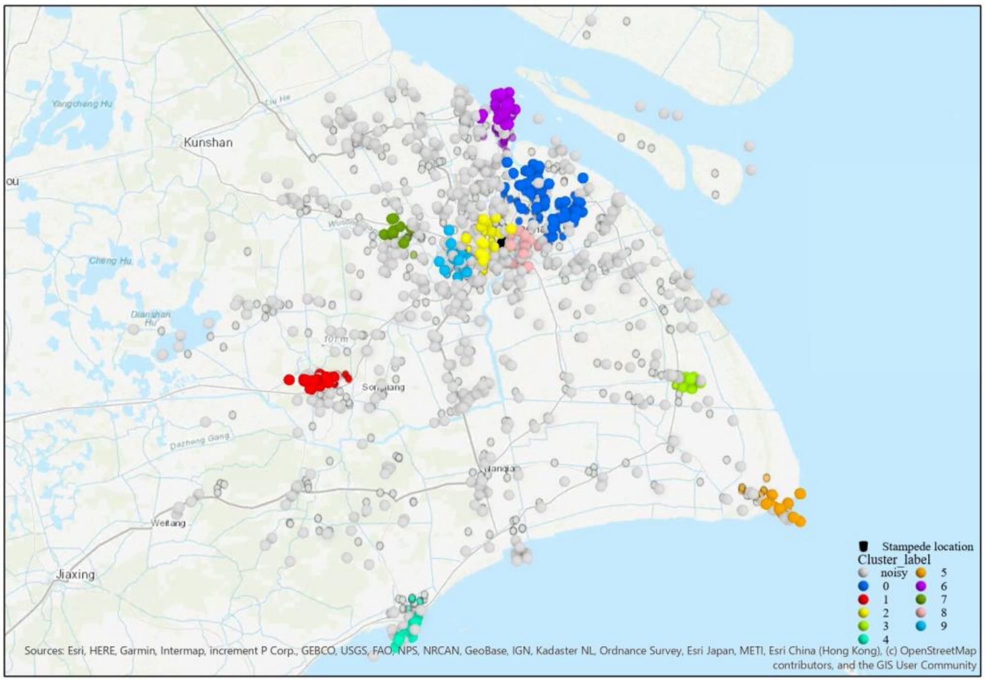
Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), (c) OpenStreetMap contributors, and the GIS User Community

(a)



Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), (c) OpenStreetMap contributors, and the GIS User Community

(b)



Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), (c) OpenStreetMap contributors, and the GIS User Community

(c)

Figure 5.13 Clustering results of the geotagged Sina Weibo posts within 1 hour around the New Year's Eve. (a) A space-time cube dot map showing the spatiotemporal semantic clusters with positive sentiments; (b) The side view of the dot map; (c) The top view of the dot map.

Table 5.5 Location and topic words of 10 posts clusters.

Cluster label	Location	Topic Word Content
0	Scenic area, commercial streets, residential areas	2015, love, happy new year, 2014, one year, good, happy, goodbye, new, people, all, warm, wish, very, new year's day
1	Campus and dormitory area	good, very, love, new, one year, happy, 2015, all, happy new year, yay, accompany, oh, people, happiness, wish
2	Scenic area, commercial streets, residential areas	love, all, new, 2015, warm, happy new year, new year's eve, love, one year, flowers, best, aircraft, wish, clothes, shy
3	Campus and dormitory area	2014, happy new year, 2015, new, play, click here, beauty shot, beat, love, thank, wish, look, record, everyone, my home
4	Scenic area, residential area	dream, 2015, try, just, snowing, wish, in my heart, white snow, dream, realize, year, new year, snowy, write down, for sure
5	Campus and dormitory area	2015, happy new year, year, good, new year's eve, love, everyone, lovely, happy, wish, one year, too, new, together, happy
6	Scenic area, residential area	love, 2015, happy new year, 2014, year, lovely, hello, everyone, new, happy, wish, one year, all, cheer, goodbye
7	Residential area	happy new year, hee hee, lovely, 2015, complacent, new, wish, happy, rose, like, really, fully, family, parents, health

8	Scenic area, residential area	happy new year, 2015, one year, work hard, wish, everyone, year, flowers, cheer, happy, new year, new, new year's day, accompany, continue
9	Scenic area, residential area	love, 2015, happy, year, best, shy, good, everyone, happy, into, new year's day, 2014, expect, happiness, forever

5.2.3 Discussions

The creation of social media data is somewhat influenced by human's physiological habits. Unlike physical sensors, human beings as social intelligence sensors do not work day and night. Therefore, what happens at night cannot get immediate and strong reactions from social media users. This has little to do with the significance of the event but is constrained by the patterns of people's work and rest. Since the Shanghai stampede incident happened at midnight, most people were unable to learn the news and respond to it. Therefore, the information related to this event spread extremely slowly in the first few hours until most people gradually wake up. Therefore, we cannot obtain effective clustering results during the time when the event occurs, but instead we can discover valuable information based on different time intervals such as 0:00-7:00 (human sleep time) and 7:00-8:00 (human active time). In particular, cluster analysis during human sleep periods (0:00-7:00) reflects that social sensing is capable of responding to sudden social events during periods of low human activity.

The setting of parameters has an important influence on the clustering effect, but how to find a suitable setting is dependent on the nature of social events in question. From the case study of the Shanghai Stampede event, we may provide some empirical suggestions.

- In the application of DBSCAN, a simple heuristic method is given to estimate $\mathcal{E}_{distance}$ and N_{minPts}, i.e., $N_{minPts} = \ln(n)$, where n is the size of the data. Then, *k-nearest neighbors* algorithm is applied to each point, where k is equal to N_{minPts}. In the *k-distance* value sorting graph, the distance corresponding to the first "valley" is the recommended $\mathcal{E}_{distance}$ (Ester et al., 1996). However, the determination of parameters in this method depends on the size and spatial distribution characteristics of the dataset, but the characteristics of event-induced human crowd behavior may not be related to the overall spatiotemporal characteristics of social media data. Therefore, this method is not suitable for mining human crowd behavior induced by social event in geotagged social media data. Considering the actual scene, using 3km as the initial spatial distance constraint ($\mathcal{E}_{distance}$) should cover most squares or venues. When the time span is restricted to smaller time intervals, there is fewer data in each time interval. Therefore, in order to ensure sufficient user participation, when performing small-scale spatiotemporal semantic clustering on social media data, 5 or 10 should be a reasonable initial number for N_{minPts}.
- For the time interval ($\mathcal{E}_{time}$), choosing half an hour as the initial parameter is advised by Huang, Li, & Shan (2018). Through our experiments, 30 minutes is a

suitable choice under normal circumstances. But at midnight, due to the decline in the activity of Weibo users, it is necessary to relax the time constraints in order to detect valuable content from the sparse data stream.

- In terms of semantic constraints, there is no valuable reference since this research is the first to combine sentimental constraints in density-based clustering while considering spatiotemporal semantic constraints. Our experiment reveals that the mean value of the cosine similarity of the Weibo posts in the entire Shanghai area fluctuated between 0.17 ($\cos 80°$) and 0.26 ($\cos 75°$) within the same time interval. There are two further facts. One the one hand, different users describe the same event differently due to their subjective cognition, the information related to the same event will be similar but not identical. On the other hand, the clustering needs to consider conditional constraints in multiple dimensions at the same time. Therefore, we recommend to slightly raise the standard on the basis of the mean value of text cosine similarity, with a cosine value of 70°, which is 0.34, as the initial constraint for text similarity ($\mathcal{E}_{textsimilarity}$).

From the sentiment analysis of Weibo posts, as shown in Figure 5.10, we can see that most of the Weibo posts are neutral. Therefore, spatiotemporal semantic clustering considering sentimental constraints is a feasible strategy to explore human crowd behaviors from social media data induced by specific event types, especially negative events. Spatiotemporal semantic clustering that considers negative sentiments can reduce the amount of computation on the one hand, and can keep us alerted to potential dangers on the other hand, which is helpful for emergency response and social management. In addition, experiments show the LDA topic modeling of the data clusters is helpful to describe the human crowd behavior induced by social events. Interpretation of topic words can reveal when where what happened or when where what people are doing together.

5.3 Perceiving differences in population mobility induced by social events

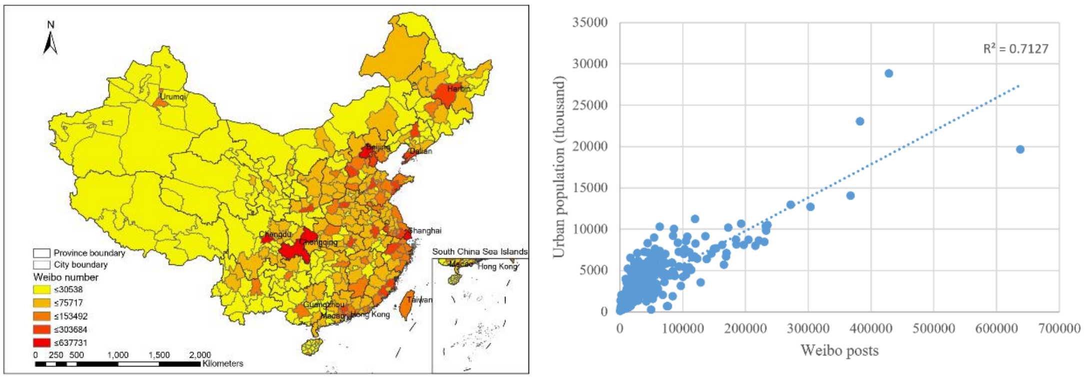
In this experiment, we illustrate the feasibility of using social sensing to perceive differences in population mobility induced by social events. The proposed method in Section 4.4 has been implemented and tested using Sina Weibo data in China to gain insight into the world's largest cyclical migration phenomenon - the Spring Festival travel (SFT) that happened around the Chinese Spring Festival. Specifically, the Spring Festival travel 2015 was used to analyze the impact of the Spring Festival on China's population movement.

The Spring Festival is a traditional Chinese festival and is regarded as an occasion for a family reunion. Therefore, many migrants leave the city where they work and go back to their hometowns for the family reunion before the Spring Festival, and after that, they return to the cities where they live. The phenomenon of SFT is rooted in the uneven development of the regional economy as well as the cultural tradition. Therefore, the migration pattern uncovered by the population flow of the SFT rush season can be used to indicate the state of uneven regional development in China.

As information sharing (e.g., updates of personal life) and social connection are two primary motivations for using social media services (L. Zhang & Pentina, 2012), the spatial-temporal information embedded in social media data provides valuable clues of human movements. Social sensing based on crowd-sourced geographical data provides a practical approach to explore the spatial behavior of the public and reveal the geographical features of the socioeconomy (Y. Liu et al., 2015). Sina Weibo is the most popular social media platform in China, with 57% of China's total Weibo users and 87% of China's Weibo activities. In 2018, Sina Weibo's monthly active users reached 462 million (2018 Sina Weibo User Development Report). Weibo data have been used in many areas, including public health (Tian, Batterham, Song, Yao, & Yu, 2018), environmental issues (S. Wang, Paul, & Dredze, 2015), natural disasters (Yandong Wang, Ruan, Wang, & Qiao, 2018), and urban land use (Y. Zhang et al., 2017). Based on the timestamp and location record of each Weibo user's posts, the user's movement trajectory can be constructed. The resulting intercity migration record can support this study.

5.3.1 Data

The Chinese Lunar New Year in 2015 falls on February 19. Traditionally, the preparations and celebrations of Spring Festival start on Xiao Nian Day (approximately one week before Spring Festival) and end at the Spring Lantern Festival (two weeks after the Chinese Lunar New Year day). As observed from the Tencent and Baidu LBS data, population flows usually reach a trough on the third day after the Spring Festival, and the population intercity flow before and after this day is reversed in direction and highly symmetrical in magnitude (J. Li, Ye, Deng, Liu, & Liu, 2016; Wei, Song, Xiu, & Zhao, 2018). Therefore, the typical SFT period in 2015 includes two periods: February 7-February 21 as the leaving period (two weeks), when the majority of migrants leave the city where they work and go back to their hometown for family reunion, and February 22-March 7 as the returning period (two weeks), when people return to the cities where they work after celebrating Spring Festival. In addition, March 8-March 21 (two weeks) is used as the ordinary period (two weeks) for comparison. For these three periods, we obtain a total of 18,152,016 Sina Weibo posts for 360 cities, including four municipalities, 293 prefecture-level cities, some county-level cities, Hong Kong, Macao, and Taiwan, which comprise the majority of China, with a fixed download frequency and uniform geographical distribution (Jendryke et al., 2017). Figure 5.14(a) shows the geographical distribution of Sina Weibo posts. Due to data acquisition restrictions, only a small part of Sina Weibo's records can be obtained for academic research. Even so, the number of Sina Weibo posts in each city has a high positive correlation with China's urban population (China Statistic Yearbooks 2015), as shown in Figure 5.14(b), proving that the obtained data represent a reasonable spatial sampling (R. Zhu et al., 2020).



(a) (b)

Figure 5.14 Chinese Sina Weibo data. (a) Geographical distribution of Sina Weibo posts; (b) The correlation between the number of Sina Weibo post and urban population per region.

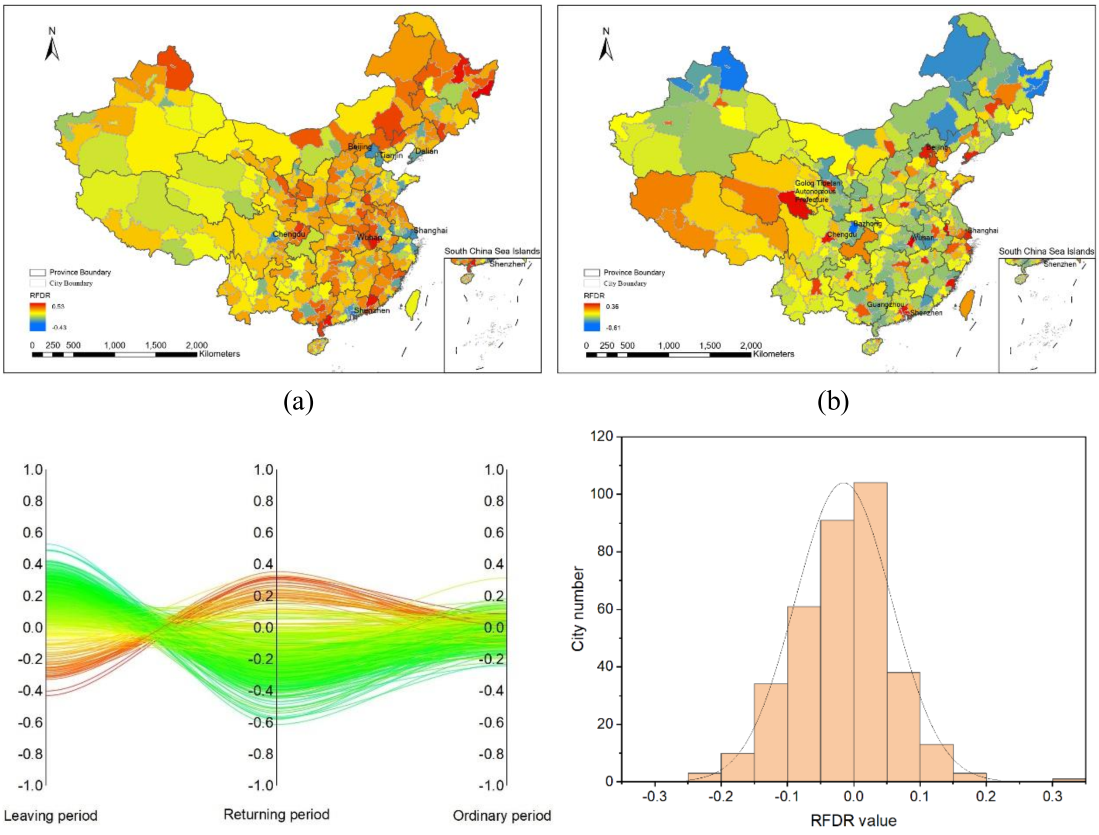
The posts of each Sina Weibo user were extracted from the raw data based on the unique user identity labels. Then, the location information of each Weibo post was spatially joined with the administrative division data of China (from the resource and environment data cloud platform of the Institute of Geographical Sciences and Natural Resource Research of the Chinese Academy of Sciences) through overlay analysis provided by Arcmap10.5.1, to determine the city where the user posted information. According to the timestamp, the location record of each Sina Weibo user can be obtained. Then, it was determined whether the position records of the two sequential posts are made in the same city, and if there is a city change, an intercity movement was recorded. Finally, all of the intercity movement records of the same period were summarized to construct a weighted directed population flow network, with the cities as network nodes and the direction and volume of population flow as the directions and weights of the network edges.

Compared with LBS migration data, Baidu migration data count population migration by hourly granularity, which can easily cut off unfinished journeys and induce an increase in the number of short trips, while Tencent migration data count population migration data on a daily basis, ignoring night traffic (J. Li et al., 2016). As personal life updates are one of the major motivations for users to post information, the spatiotemporal behaviors induced by Spring Festival become an incentive for Weibo users to record their lives. Therefore, the method of measuring intercity migration based on the location changes of users' posts in Sina Weibo is flexible and feasible. Based on the movement trajectory of the Sina Weibo users, this study constructed the weighted directed networks of intercity population migration flows considering the posts made during the entire SFT period, the leaving period, the returning period, and the ordinary period and obtained 996,901, 306,795, 429,907, and 175,911 intercity population flow records, respectively.

5.3.2 Sensing human mobility induced by Chinese Spring Festival

The difference in $RFDR$

Figure 5.15 displays the $RFDR$ of Chinese cities in the three periods, and an opposite trend is found between the leaving period and returning period (R. Zhu et al., 2020). In the leaving period, as shown in Figure 5.15(a), high positive values appear mainly in central, northeast, and northwest China, with Jixi in Heilongjiang Province having the highest value of 0.53. Beijing, Dalian, Tianjin, the Yangtze River Delta, and the Pearl River Delta show significant negative values, with Shenzhen of Guangdong Province showing the lowest value of −0.43. In the returning period (Figure 5.15(b)), significant positive values appear in Beijing, the Pearl River Delta, Dalian, the Yangtze River Delta, Guoluo, Chongqing, and some cities in western China. Wuhan in Hubei Province has the highest value of 0.36. Negative values are mainly observed in northeast, northwest, and central China, with Bazhong in Sichuan Province having the lowest value of −0.61. A strong negative correlation of $RFDR$ between the leaving period and the returning periods is presented in Figure 5.15(c). A city that has more relative population inflow in the leaving period has more relative population outflow in the returning period, and vice versa. During the ordinary period, the $RFDR$ distribution roughly shows a Gaussian distribution $N(-0.015,0.005)$, with 98% of the values distributed between -0.2 and 0.2 (Figure 5.15(d)).

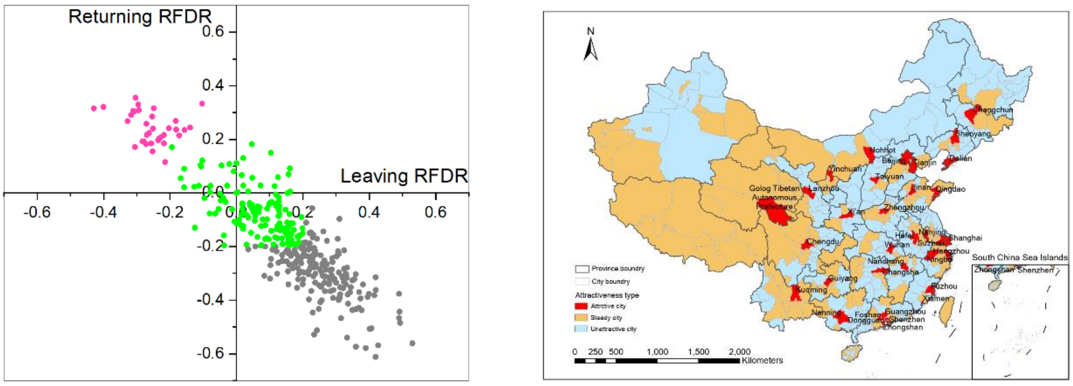


(a)

(b)

(c)

(d)



(e) (f)

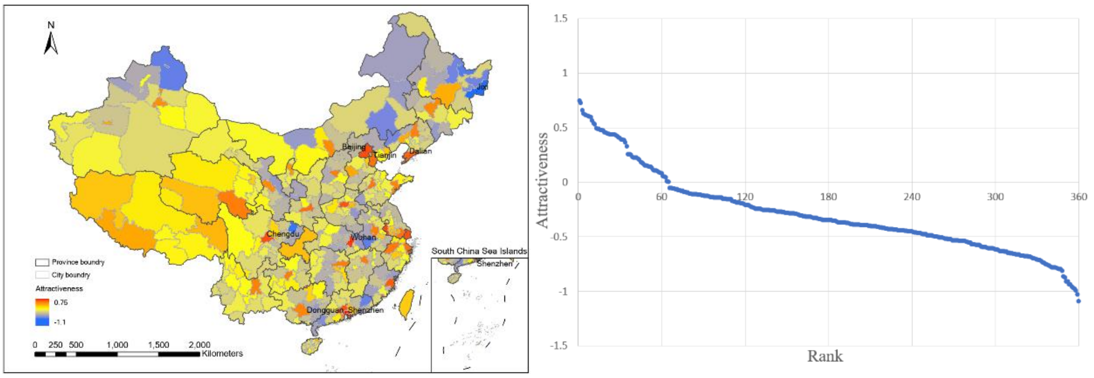
Figure 5.15 Relative flow difference ratio ($RFDR$) of cities in different periods. (a) Leaving period; (b) Returning period; (c) $RFDR$ change of cities in the three periods; (d) $RFDR$ distribution in the ordinary period; (e) City classification based on $RFDR$ change; (f) Geographical distribution of different types of cities.

It can be seen from the variation in $RFDR$ in three periods that the tradition of family reunion at Spring Festival caused a strong opposite population flow trend before and after the Spring Festival, and the trend of population flow becomes stable during the ordinary period. Therefore, according to the $RFDR$ distribution during the ordinary period, we use the interval $(-0.2,0.2)$ as the stability interval and judge a city's attractiveness based on the change in the $RFDR$ of the city in the leaving and returning periods (Figure 5.15(e)).

Thirty-four cities are classified as attractive cities because of the apparent negative value and positive value of the $RFDR$ in the leaving and returning period, respectively, which indicates that 10% of Chinese cities have significant appeal to migrant workers. In total, 145 cities are regarded as stable cities because the $RFDR$ values in both periods remain within the stable interval. Finally, 181 cities are rated as unattractive cities because of the prominent positive value and negative value of the $RFDR$ in the leaving and returning periods, respectively. Similarly, Long and Wu (2016) identified 180 shrinking cities in China, with a decline in population density, based on China's censuses in 2000 and 2010. As shown in Figure 5.15(f), three distinct attractive urban clusters are located in the Yangtze River Delta, Pearl River Delta, and Beijing-Tianjin-Hebei region. In addition, most provinces, except Xinjiang, Tibet, Taiwan, Hainan, and Heilongjiang, contain one or two attractive cities, surrounded by several stable cities and unattractive cities. In comparison with these five regions, Xinjiang, Hainan, and Heilongjiang contain several unattractive cities, reflecting the phenomenon of population loss in some subregions, while the population changes in Tibet and Taiwan are stable.

The difference in urban attractiveness

The spatial distribution of urban attractiveness based on the PMN during the leaving period and returning period appears to be similar to that of the $RFDR$ in the returning period (Figure 16(a)). Negative values appear mostly in northeast, northwest, and central China, with Jixi of Heilongjiang Province having the lowest value of -1.1. Higher values appear in Beijing, Dalian, Chengdu, Tianjin, the Pearl River Delta, and the Yangtze River Delta. Shenzhen yields the highest value of 0.75, followed by Dongguan and Wuhan. Figure 5.16(b) describes the distribution of the attractiveness value. Only 64 cities (17.8%) show positive attractiveness, while most cities have negative attractiveness, reflecting the severe imbalance in urban development in China (R. Zhu et al., 2020). A small number of influential cities are relatively developed and absorb labor from underdeveloped regions, while most cities lag behind in development and export labor to these developed cities.



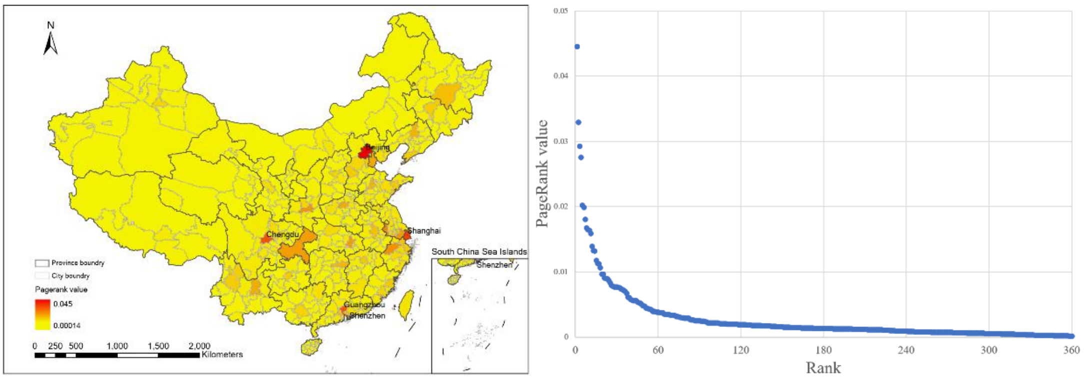
(a) (b)

Figure 5.16 The attractiveness of the Chinese cities. (a) Spatial distribution of urban attractiveness; (b) Attractiveness distribution.

The difference in urban activity based on PageRank

The activity of cities based on the PMN during the SFT period is similar to their classification and attractiveness, while the attractiveness focuses on the comparison of the relative population flow ratio and the activity is more prominent in the role of population flow volume (Figure 5.17(a)). The attractive cities classified based on the change in $RFDR$ have relatively higher PageRank values, with Beijing yielding the highest value of 0.045, followed by Shanghai, Chengdu, Guangzhou, and Shenzhen. Moreover, lower values appear mostly in the western and northern regions of China.

Figure 5.17(b) shows a heavy-tailed distribution of the PageRank values of cities, reflecting significant differences between Chinese cities (R. Zhu et al., 2020). For a few core cities, the higher the ranking is, the higher the PageRank value, and there are significant differences between cities. However, for most of the remaining cities, their PageRank values are low, and there is little difference between cities. This is somewhat similar to the 80/20 rule: a few core Chinese cities have greater influence, while most other cities lack the core competitiveness.



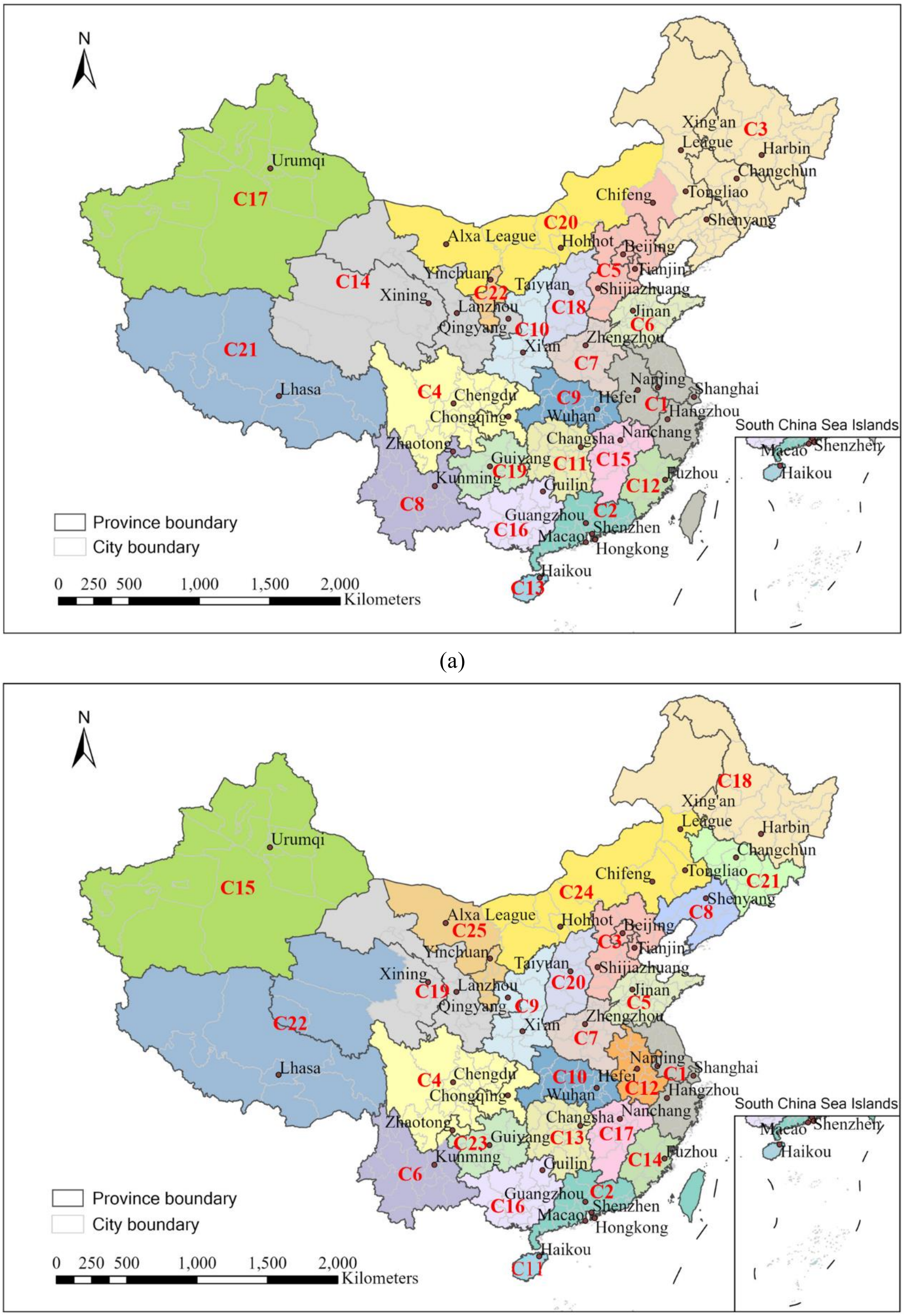
(a) (b)

Figure 5.17 Activity of population mobility in Chinese cities. (a) Spatial distribution of urban activity based on PageRank; (b) PageRank value distribution.

City communities based on the intercity migration network

Figure 5.18 describes the community structures of cities during the SFT period and the ordinary period and their mapping relationship. For the ordinary period, 25 communities of cities are detected, as shown in Figure 5.18(b). Most of the community divisions are consistent with the Chinese administrative division of the province, while some communities demonstrate two phenomena of interprovincial aggregation and fragmentation. Regarding interprovincial aggregation, community C1 (Yangtze River Delta) consists of Shanghai, Province Jiangsu, and Province Zhejiang; community C2 (Pearl River Delta) consists of Hong Kong, Macao, and Province Guangdong; community C3 (Beijing-Tianjin-Hebei region) consists of Beijing, Tianjin, and Province Hebei; and community C4 consists of Chongqing, Province Sichuan, and a city from an adjacent south province. These four regions have many cities with high PageRank and attractiveness, forming tight spatial organizations based on developed infrastructure, such as transportation and communication.

By contrast, the northeastern and western parts of Inner Mongolia join Province Heilongjiang and Province Ningxia to form communities C18 and C25, respectively, and the eastern and western parts of Province Qinghai are clustered with Province Gansu and Tibet to form communities C19 and C22. This segmentation phenomenon is affected by many factors, such as the economy, culture, and geography. For example, both western Qinghai and Tibet are mainly Tibetan living areas (Guo, 1996), while eastern Inner Mongolia and Province Heilongjiang share similar climates and living customs (Miao, 2016). The community structure of cities during ordinary time reflects that the majority of the trips are intraprovincial trips and interprovincial trips between neighboring provinces.



(a)

(b)

Figure 5.18 City communities. (a) Spring Festival travel (SFT) period; (b) Ordinary period.

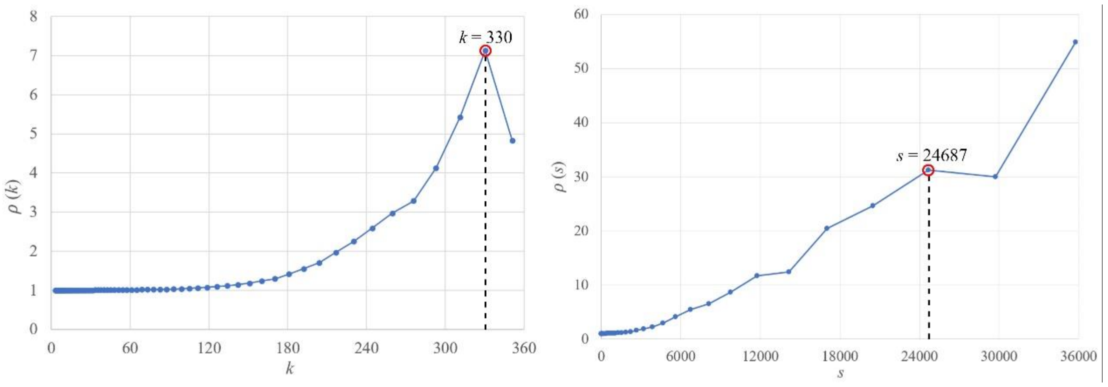
The community structure of cities during the SFT period is similar to that during the ordinary period, except that some communities merge and several cities move between neighboring communities, with 22 communities formed (Figure 5.18(a)). The northeastern part of China containing Province Heilongjiang, Province Liaoning, Province Jilin, and eastern cities of Inner Mongolia becomes Community 3; Province Anhui, Province Zhejiang, Province Jiangsu, and Shanghai in eastern China constitute Community C1. The cities of Alxa League, Zhaotong, Qingyang, and western Qinghai, which are separated by neighboring provinces during the ordinary period, re-establish communities C20, C8, and C14 with the cities belonging to the same province during the SFT period. However, the mobility behavior of city Chifeng is different. It detaches itself from its province and joins the neighboring community of Beijing, Tianjin, and Province Hebei during the SFT period. This phenomenon is mainly affected by the family reunion behavior induced by the traditional culture of the Spring Festival. The increase in the size of some communities also reflects the increase in long-distance travel during the SFT period.

Rich-club effect

Figure 5.19 shows that the vast majority of $\rho_w(k)$ and $\rho_w(s)$ are greater than 1 and present an upward trend, reflecting the remarkable rich-club phenomenon in the SFT network (R. Zhu et al., 2020). This means that the prominent cities in China tend to engage in stronger interactions among themselves. In addition, a significant "demarcation point" feature can be found from the change in the curves of both $\rho_w(k)$ and $\rho_w(s)$. Regarding $\rho_w(k)$, k = 330 is a demarcation point: for the number of points smaller than 330, the curve grows steadily, and beyond that point, the curve decreases. For $\rho_w(s)$, s = 24,687 is a demarcation point because the curve basically continues to rise, but there is a downward wave after this point. The prominent rich-club cities are extracted based on these demarcation points.

Using k > 330 as the selection criterion, nine cities are obtained - namely, Beijing, Shenzhen, Chengdu, Shanghai, Chongqing, Sanya, Guangzhou, Hangzhou, and Wuhan. Each of these cities has a migration connection with more than 90% of China's cities. With s > 24,687 as the selection threshold, five cities are selected, including Beijing, Shanghai, Chengdu, Guangzhou, and Shenzhen.

The analysis based on $\rho_w(k)$ emphasizes the topological connection between cities, while the analysis based on $\rho_w(s)$ focuses more on the strength of association between cities. Based on different selection criteria for the rich parameter, we can obtain rich-club cities at different rich levels. However, the core rich-club cities should have the most connections and the greatest connection strength. Therefore, we select the intersection of the two city lists as the core rich-club members, namely, Beijing, Shanghai, Chengdu, Guangzhou, and Shenzhen. The population migration flow involving these five rich-club members accounts for 23.2% of the total migration flow during the SFT period. Located in different city communities, these core rich-club cities form the backbone network of the SFT network and can promote cross-regional population flow through close interconnections.



(a) (b)

Figure 5.19 Rich-club coefficients in the SFT network. (a) r = k (out-degree); (b) r = s (out-strength).

Imbalance of regional development

To further understand the regional development differences in China, we use the urban community structure of the intercity migration network in the ordinary period as an embodiment of the urban agglomeration to analyze the importance and attractiveness of the regions, as shown in Figure 5.20-5.22. Figure 5.23 shows the PageRank and attractiveness of communities (R. Zhu et al., 2020). C1 (Yangtze River Delta), C2 (Pearl River Delta), and C3 (Beijing-Tianjin-Hebei region) are the three most active and most attractive city communities. Moreover, the southeastern coastal areas of China, including C8, C3, C5, C1, C14, and C2, have high PageRank values and positive attractiveness values. In contrast, inland communities adjacent to these communities, including C18, C21, C24, C20, C7, C12, C17, C13, and C16, show relatively significant negative attractiveness values, regardless of their PageRank values. The four regions distributed in western and northwestern China (C25, C19, C15, C22) yield the least influence, but the population flow in these regions is relatively stable and does not show significant negative attractiveness. The remaining five communities (C4, C6, C9, C10, C23), clustered in central and southwestern China, have relatively high PageRank values and positive attractiveness values.

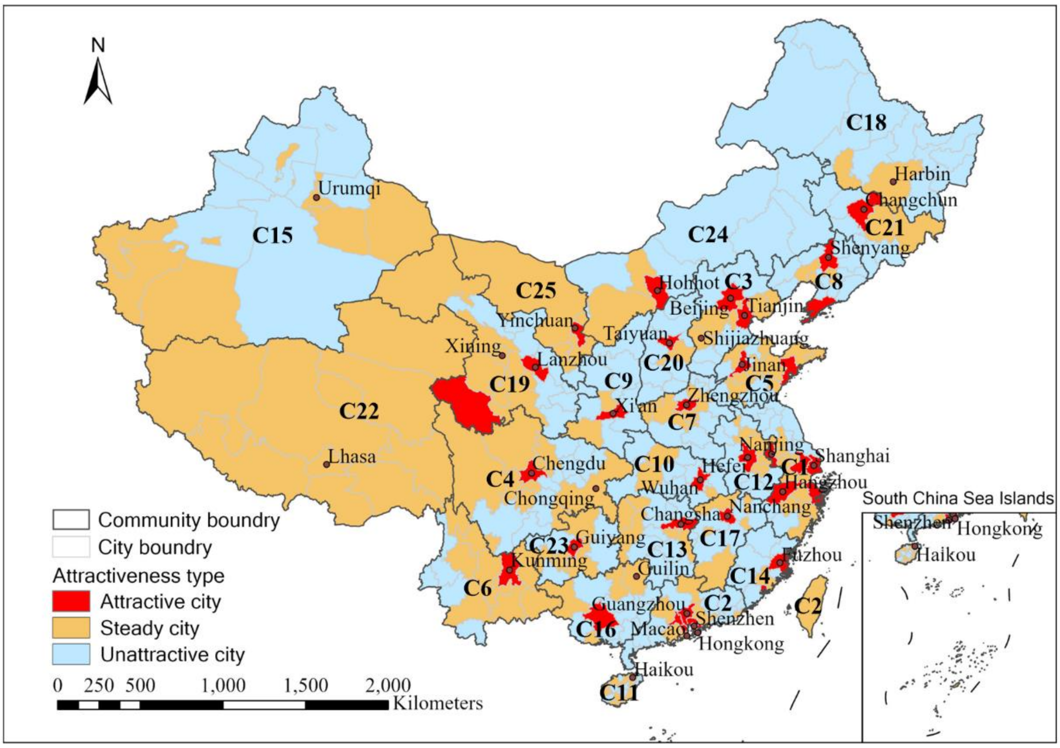


Figure 5.20 Overlay of city types and city communities.

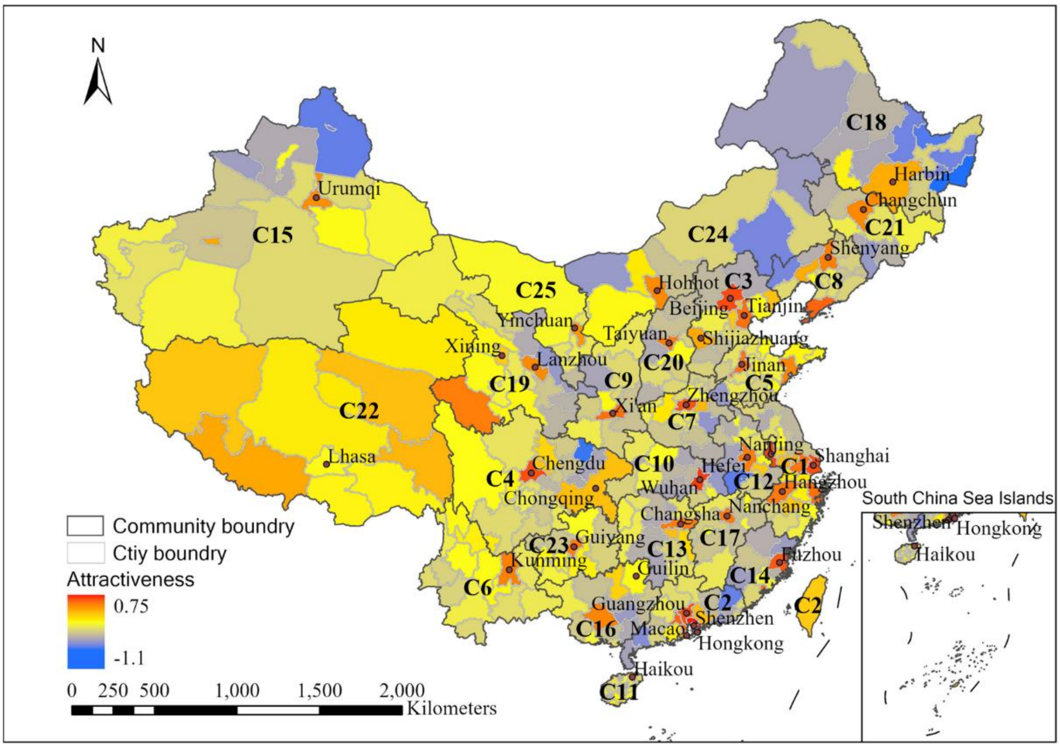


Figure 5.21 Overlay of the attractiveness values of cities and city communities.

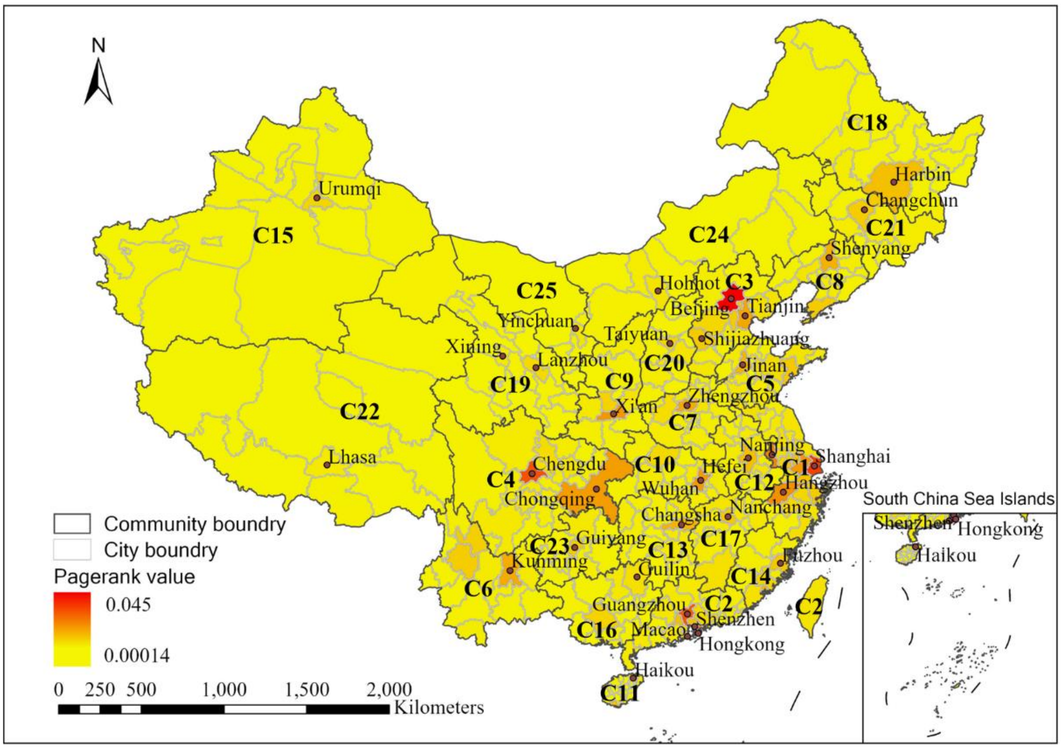


Figure 5.22 Overlay of PageRank values of cities and city communities.

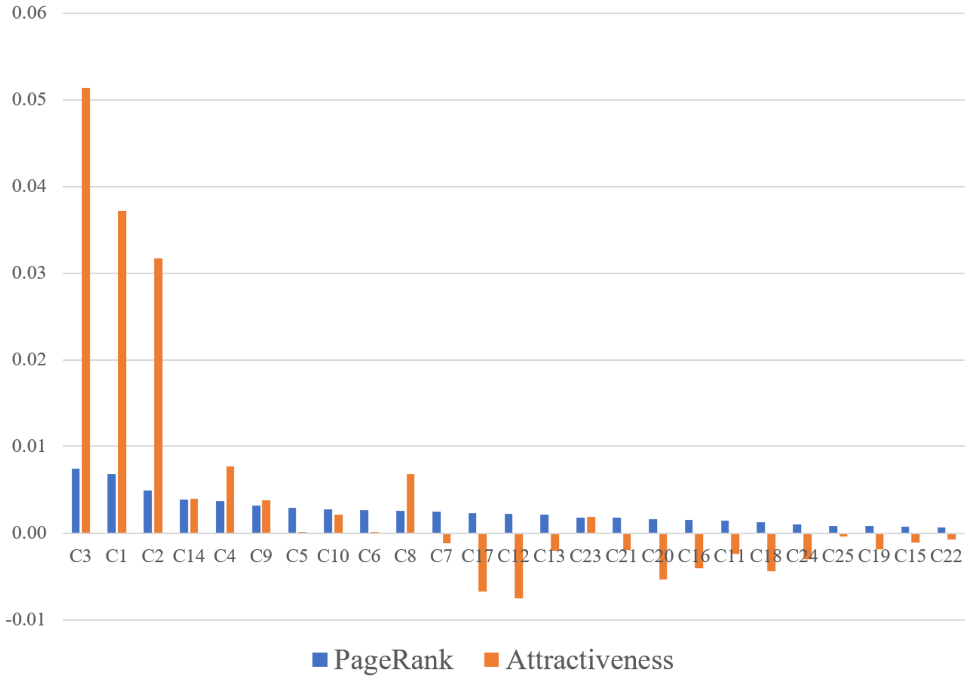


Figure 5.23 The PageRank value and attractiveness of communities, sorted by PageRank value.

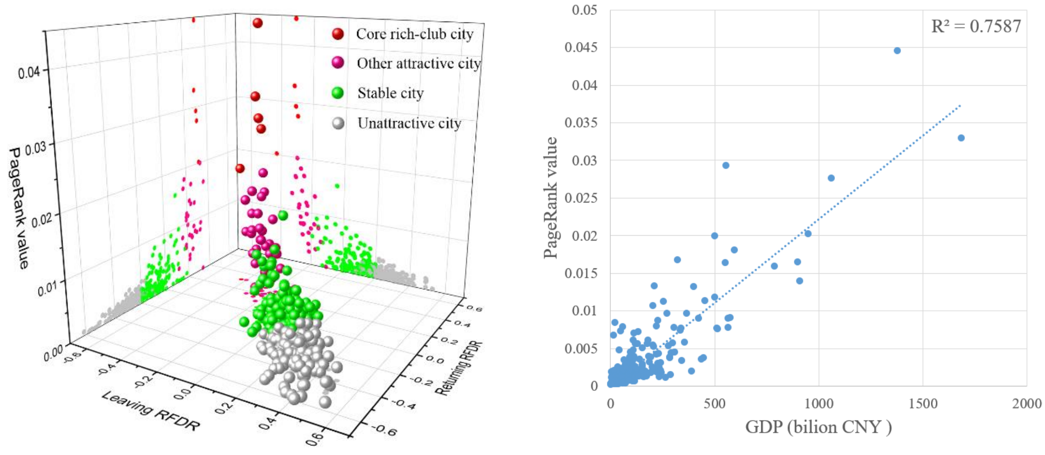
5.3.3 Discussions

The indices of population mobility based on PMN with urban development

There is a clear correlation between the PMN-based urban development indices, as shown in Figure 5.24(a) (R. Zhu et al., 2020). The relationship between PageRank and $RFDR$ presents a trend in which the PageRank value decreases while the city's $RFDR$ value goes from low to high during the leaving period and from high to low during the returning period. Additionally, the attractive cities have higher PageRank values than the less attractive cities. It can be inferred that cities with higher PageRank values are more competitive, thus showing significant and opposite $RFDR$ values during the two periods of Spring Festival. In addition, the core rich-club cities show the top five PageRank values and have significantly negative $RFDR$ values in the leaving period and significantly positive $RFDR$ values in the returning period, reflecting the prominent influence of these cities in the PMN.

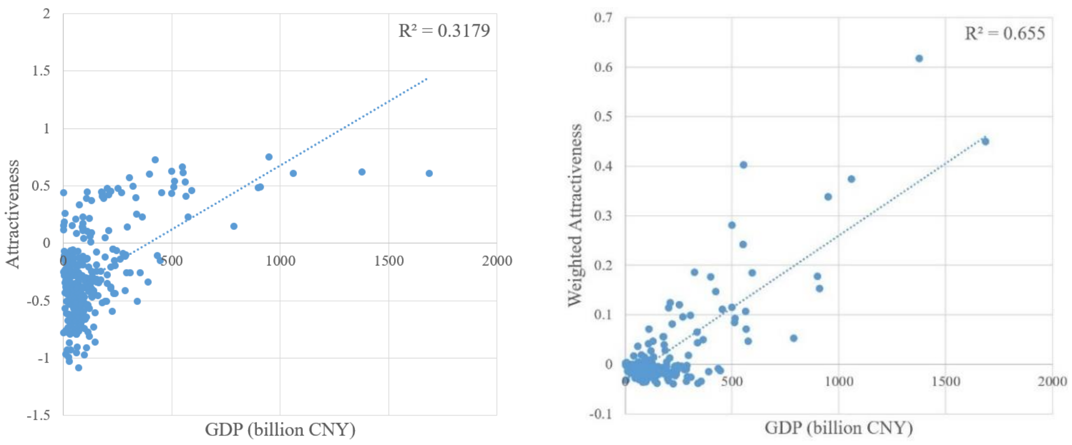
A study based on Tencent location big data found that indices based on migrant population can represent the level of urban development to a certain degree (J. Xu et al., 2017). To validate this finding and explore the relation between the indices used in this thesis and urban development, we calculated the values of three variables - the PageRank, attractiveness, and weighted attractiveness (the weight could be the ratio between the PageRank value and the maximum PageRank value among all cities) from the migration network and compared them with the urban GDP values from the China Statistical Yearbook 2015.

The correlation coefficient between the PageRank and GDP yields a very high value (Figure 5.24(b)), indicating that the PageRank value based on the migration network can reflect the development level of a city. Although attractiveness and weighted attractiveness are both positively correlated with GDP, the correlation coefficient between weighted attractiveness and urban GDP is higher than that between attractiveness and GDP. This is reasonable because the attractiveness value of a city is only based on its relative population flow ratio without comparing with other cities, whereas the weighted attractiveness considers the difference in the level of development between cities and can thus reflect the difference in the attraction between the cities. For the same reason, the weighted attractiveness can be transferred to analyze communities.



(a) PageRank - $RFDR$

(b) PageRank - GDP



(c) Attractiveness - GDP

(d) Weighted attractiveness - GDP

Figure 5.24 The relationships between indices based on SFT data and urban economic indices. (a) PageRank versus $RFDR$; (b) PageRank versus GDP; (c) Attractiveness versus GDP; (d) Weighted attractiveness versus GDP.

Although the PageRank has a higher correlation with urban GDP than attractiveness and weighted attractiveness, a high PageRank value does not necessarily mean that a city also has lower and higher $RFDR$ values in the leaving period and returning period, respectively (Figure 5.24). This is because the attractiveness of a city depends on not only its level of development but also various natural and social factors, especially the development levels of surrounding cities. Taking two cities with similar development situations as an example, one is adjacent to more developed cities and the other to underdeveloped cities. The influence from the surrounding cities makes one city less attractive than the other. This rule also applies to city communities. The PageRank value of C14 is slightly higher than that of C4, but C4 is almost twice as attractive as C14, largely because C14's attractiveness is affected by the adjacent C2 and C1 with higher PageRank values, and the PageRank values of the communities around C4 are relatively low.

Rich-club cities with interregional transportation

The rich-club coefficients based on two rich parameters - connections and weights prove that there is a rich-club phenomenon in the SFT network, which means that influential cities are not isolated from each other but have a close interaction. Although this finding is similar to the finding of a study based on the 2015 Baidu LBS data (Wei et al., 2018), there are two subtle differences. In the PMN based on the Baidu LBS data, only one city has an out-degree greater than 300, and the breakpoint of k = 200 is selected to determine 11 rich-club cities based on their connections. However, in the PMN based on the Sina Weibo data, this study finds better connectivity between cities, with the maximum out-degree of the urban nodes being 353, and nine cities connecting with more than 330 cities. Additionally, the six discovered core rich-club cities based on the Tencent LBS data are all located in the coastal areas (four of which are consistent with our results) but ignore the central hub of Chengdu. Moreover, the connections involving these six core rich-club cities account for 49.57% of the population flow, more than twice the 23.3% of movements involved in the five core rich-club cities in this study.

This inconsistency is mostly due to data limitations. Li et al. (2016) pointed out that the 2015 Baidu LBS data provide only the top 10 inflows and top 10 outflows per city per hour and extra top 4000 flows per collection. Therefore, the resulting PMN cannot adequately describe the intercity Spring Festival migration. On the other hand, the flow data provided by the Baidu LBS data lack passenger information and cannot integrate the segments of a traveler's trip into the entire travel route and thus are unable to describe long-distance travel.

The community structure of the cities demonstrates that the migration between cities is influenced by geographical proximity, and the cities within the community are spatially adjacent. Moreover, five core rich-club cities are scattered in the north, central, eastern, and southern parts of China and serve as transportation hubs, connecting various scattered areas through close interconnections. Additionally, these core rich-club cities are attractive cities and have the highest PageRank values, meaning that they have higher competitiveness.

It is inferred that a small number of rich-club cities dominated by Beijing, Shanghai, Chengdu, Guangzhou, and Shenzhen serve as critical regional nodes for local economic development, integrating spatially dispersed areas and promoting effective interaction across the country with close interconnection. China is currently systematically building multilevel transportation hubs to promote interconnectivity across the country, and the five core rich-club cities identified in this paper - namely, Beijing, Shanghai, Guangzhou, Shenzhen, and Chengdu - are listed as the first comprehensive international transportation hubs (among a total of seven). It implies that the development of Chinese transportation network system will further enhance the rich-club characteristics in the population flow. A few rich-club cities form a dense and interconnected backbone network to attract, absorb, and disseminate large-scale population flows.

City communities with urban agglomeration planning

Figure 5.25 shows the distribution of urban agglomerations outlined in China's 13th Five-Year Plan (2016 - 2020) (R. Zhu et al., 2020). As shown in Figures 5.15 - 5.17 and Figure 5.20 - 5.23, these planned urban agglomerations include almost all attractive cities and high PageRank cities and cover all city communities except C22. However, among these urban agglomerations, only the development targets of the Yangtze River Delta (C1), the Pearl River Delta (C2), and Beijing-Tianjin-Hebei (C3) are world-class urban agglomerations, while other urban agglomerations are still in the early stages of development, which is consistent with our results. C1, C2, and C3, which contain these three urban agglomerations, have the highest PageRank values and significant attractiveness. In 2015, the economic aggregate of these three urban agglomerations accounted for more than 40% of the national economy (China Urban Agglomeration Integration Report 2019).

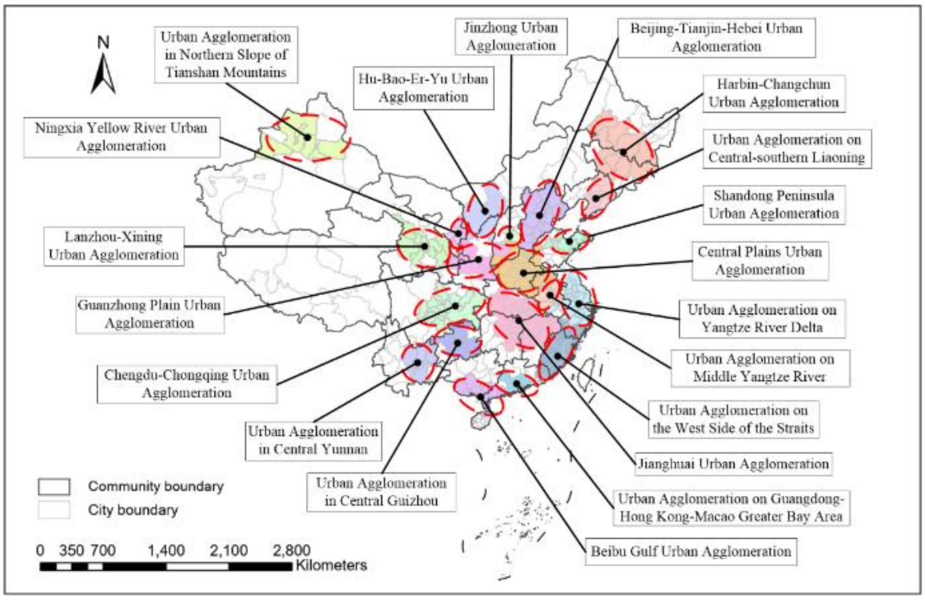

N
Urban Agglomeration in Northern Slope of Tianshan Mountains
Jinzhong Urban Agglomeration
Beijing-Tianjin-Hebei Urban Agglomeration
Hu-Bao-Er-Yu Urban Agglomeration
Harbin-Changchun Urban Agglomeration
Ningxia Yellow River Urban Agglomeration
Urban Agglomeration on Central-southern Liaoning
Shandong Peninsula Urban Agglomeration
Lanzhou-Xining Urban Agglomeration
Central Plains Urban Agglomeration
Guanzhong Plain Urban Agglomeration
Urban Agglomeration on Yangtze River Delta
Urban Agglomeration on Middle Yangtze River
Chengdu-Chongqing Urban Agglomeration
Urban Agglomeration on the West Side of the Straits
Urban Agglomeration in Central Yunnan
Jianghuai Urban Agglomeration
Community boundary
City boundary
Urban Agglomeration in Central Guizhou
Urban Agglomeration on Guangdong-Hong Kong-Macao Greater Bay Area
Beibu Gulf Urban Agglomeration
0 350 700 1,400 2,100 2,800
Kilometers

Figure 5.25 Anticipated urban agglomerations in China’s 13th Five-Year Plan.

However, due to the influence of the eastern coastal areas, the attractiveness of the adjacent northeastern and central communities is significantly negative, indicating a large number of unattractive cities. The cultivation of urban agglomerations in the northeast and central regions will help to promote the coordinated and sustainable development of the local economy. The four city communities in the western region yield the lowest PageRank value, and the corresponding urban agglomeration planning could revitalize local development. It is observed that China’s regional development level shows a gradient pattern of high to low from east to west, which is consistent with the results of studies based on DMSP/OLS nighttime light data 1992–2013 (H. Liu & Du, 2017), the Development and Life Index 2000–2012 (H. Liu & He, 2015), and the China City Statistical Yearbook 2010 (Guan, Fang, & Luo, 2012).

It is worth noting that C22 yields the lowest PageRank value, with no attractive city or high PageRank city, which indicates that the overall development level of this region lags behind. However, the attractiveness value of C22 is near zero, indicating that there is no significant population loss in this region. Despite the low economic level in the Tibet area, the local population has a strong tie to cultural tradition and lifestyle in the area, therefore, these outflows are rather limited (Chang, 2009). In addition, in the community structures of the two periods, Taiwan has established communities with the Pearl River Delta and the Yangtze River Delta, indicating population flow between Taiwan and the Pearl River Delta and Yangtze River Delta. However, during these two periods, Taiwan’s $RFDR$ values were stable and did not show significant population inflows or outflows. This is mainly attributed to the current cross-strait policy (Y. S. Wu, 2005). On the whole, the national urban agglomeration plan in China is reasonable.

6 Conclusion and outlook

6.1 Self-Evaluation

This thesis is among the first studies of social sensing along with the methods and their implementation for the perception of event-induced human behavior in geotagged social media data. An analytical framework is proposed to deploy the potential of social sensing for the discovery of human behavior based on spatial, temporal, and semantic clues in social media data. Three kinds of human behavior manifestations - inner behavior, crowd behavior, and human mobility are explored and experimentally verified. There are following highlights in the thesis:

- The analytical framework for perceiving event-induced human inner behavior has a generic nature of combining machine learning, natural language processing, and visualization methods, although its implementation is targeted to a special case. The combination of labeling function from hashtag and classification function from machine learning helps to obtain as much information as possible about a specific social event from raw social media data. The results of social sentiment analysis have revealed the impact of a specific social event on human emotions while social opinion mining can assist us in discovering the attitude of the public as well as the implicit relationships among events. This holistic analysis helps to gain insight into why an event happened, how and for how long it affected local citizens' inner behavior.

- An algorithm of sentiment-constrained spatiotemporal semantic clustering is proposed as the extension of DBSCAN and ST-DBSCAN for perceiving crowd behavior triggered by social events. It comprehensively considers spatiotemporal proximity, text similarity, and sentiment constraints reflected in social media data, and allows to obtain clusters containing similar descriptions with analogous emotions at specific times and places. Further, the content interpretation of clusters based on LDA topic modeling can reveal when where what happened or when where people do what together.

- A geospatial network analysis method combining the perspectives of geospatial science and complex network science is proposed and experimented to explore the population mobility patterns triggered by social events. The changing spatiotemporal states of social media users are used to trace population movement. By constructing PMN for the event occurrence period against a general period, the impact of major social event on population mobility can be explored from the changes in the population mobility pattern during the event occurrence period against the general period and the differing population mobility pattern in different regions during the event occurrence period. Further analysis based on the induced population mobility patterns may reveal hidden political, economic, social, and cultural characteristics.

- The social sensing is conducted from spatial, temporal, and semantical perspectives. The result can be used to characterize human behavior, including but not limited to inner behavior, crowd behavior, and human mobility induced by social events. On the other hand, the creation of social media data is confined by human behavior such as the regularity of work and rest, and by the characteristics of population distribution. Consequently, the acquired data has periodic differences in time series or regional differences in spatial distribution. The experiments in the thesis have demonstrated the feasibility of constrained social sensing to serve individuals, groups and society.

6.2 Outlook

This thesis has succeeded in its aim of perceiving human behavior from geotagged social media data, but naturally, there is still room for improvement. In the future, the following tasks will be further studied and investigated:

- This study has verified the social sensing approach for the observation and interpretation of human behavior induced by social events in geotagged social media data. However, social sensing based on crowdsourced geographic information has emerged only recently, and a much longer series and higher spatial-temporal resolution are needed to fully demonstrate the profound impact of major social events on human behavior.
- Existing methods can be refined and extended. The demographic information of social media users, e.g., age, gender, and education, and social media contents in various forms, e.g., text, image, and video, can be combined to further explore the impact of different types of social events on diverse human behaviors of different groups. Also, it would be beneficial to go beyond a single data source and carry out comprehensive analysis of different types of crowdsourced geospatial data, making existing methods more adaptive and innovative.
- With the continuous development of science, it is necessary to keep pace with the most advanced technologies in a timely manner and expand the service capabilities of social sensing. For example, a variety of deep learning frameworks (e.g., deep neural networks, convolutional neural networks, and recurrent neural networks) have been applied in various fields (such as computer vision, speech recognition, natural language processing, and bioinformatics) and acquired excellent effects. Integrating the newest progress of artificial intelligence into our proposed methods may further unlock the potential of social sensing. Likewise, the advanced virtual reality technologies are highly desirable for the improvement of visualization of human behavior in multiple dimensions of space, time, and semantics.

Bibliography

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., & Tao, K. (2012). Semantics + filtering + search = twitcident. exploring information in social web streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media - HT '12* (pp. 285–294). https://doi.org/10.1145/2309996.2310043

Ahsan, U., Sun, C., Hays, J., & Essa, I. (2017). Complex event recognition from images with few training examples. In *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017* (pp. 669–678). https://doi.org/10.1109/WACV.2017.80

Akresh, R., Verwimp, P., & Bundervoet, T. (2011). Civil war, crop failure, and child stunting in Rwanda. *Economic Development and Cultural Change*, *59*(4), 777–810. https://doi.org/10.1086/660003

Al-Anzi, F. S., & AbuZeina, D. (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. *Journal of King Saud University - Computer and Information Sciences*, *29*(2), 189–195. https://doi.org/10.1016/j.jksuci.2016.04.001

Ali, R., Solis, C., Salehie, M., Omoronyia, I., Nuseibeh, B., & Maalej, W. (2011). Social Sensing : When Users Become Monitors. In *Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering* (pp. 476–479). Association for Computing Machinery. https://doi.org/10.1145/2025113.2025196

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, *46*(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879

Arb, M. Von, Bader, M., Kuhn, M., & Wattenhofer, R. (2008). VENETA : Serverless Friend-of-Friend Detection in Mobile Social Networking. *2008 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*, 184–189. https://doi.org/10.1109/WiMob.2008.52

Arvanitidis, P., Economou, A., & Kollias, C. (2016). Terrorism ' s effects on social capital in European countries. *Public Choice*, *169*(3), 231–250. https://doi.org/10.1007/s11127-016-0370-3

Baral, R., Wang, D., Li, T., & Chen, S. C. (2016). GeoTeCS: Exploiting geographical, temporal, categorical and social aspects for personalized poi recommendation. In *Proceedings - 2016 IEEE 17th International Conference on Information Reuse and Integration, IRI 2016* (pp. 94–101). IEEE. https://doi.org/10.1109/IRI.2016.20

Bernstein, D., & Livingston, C. (1982). An interactive program for observation and analysis of human behavior in a long-term continuous laboratory. *Behavior Research Methods & Instrumentation*, *14*(2), 231–235.

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, *60*(1), 208–221. https://doi.org/10.1016/j.datak.2006.01.013

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Blondel, V. D., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P1008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General*, *121*(2), 177–194. https://doi.org/10.1037//0096-3445.121.2.177

Cai, H., Yang, Y., Li, X., & Huang, Z. (2015). What are Popular : Exploring Twitter Features for Event Detection , Tracking and Visualization. *MM '15 Proceedings of the 23rd ACM International Conference on Multimedia*, 89–98. https://doi.org/10.1145/2733373.2806236

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-Time Urban Monitoring Using Cell Phones : A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, *12*(1), 141–151. https://doi.org/10.1109/tits.2010.2074196

Calabrese, F., Pereira, F. C., Lorenzo, G. Di, Liu, L., & Ratti, C. (2010). The Geography of Taste : Analyzing Cell-Phone Mobility and Social Events. In *International conference on pervasive computing* (pp. 22–37). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12654-3_2

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, *28*(2), 15–21. https://doi.org/10.1109/MIS.2013.30

Campbell, A. T., & Lane, N. D. (2008). The Rise of People-Centric Sensing. *IEEE Internet Computing*, *12*(4), 12–21. https://doi.org/10.1109/MIC.2008.90

Caragea, C. ., Squicciarini, A. ., Stehle, S. ., Neppalli, K. ., & Tapia, A. . (2014). Mapping moods: Geo-mapped sentiment analysis during hurricane sandy. *ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management*, (May), 642–651. Retrieved from http://www.iscram.org/legacy/ISCRAM2014/papers/p29.pdf

Chang, L. (2009). The Current Status and the Development of Tibetan Culture Industry under the Global Horizon. *JOURNAL OF TIBET UNIVERSITY*, *24*(4), 32–37. https://doi.org/10.16249/j.cnki.1005-5738.2009.04.022

Charlson, F. J., Steel, Z., de genhardt, L., Chey, T., Silove, de rrick, Marnane, C., & Whiteford, H. A. (2012). Predicting the impact of the 2011 conflict in libya on population mental health: PTSD and depression prevalence and mental health service requirements. *PLoS ONE*, *7*(7), e40593. https://doi.org/10.1371/journal.pone.0040593

Charlson, F., van Ommeren, M., Flaxman, A., Cornett, J., Whiteford, H., & Saxena, S. (2019). New WHO prevalence estimates of mental disorders in conflict settings: a systematic review and meta-analysis. *The Lancet*, *394*(10194), 240–248. https://doi.org/10.1016/S0140-6736(19)30934-1

Chen, L., Nugent, C., Cook, D., & Yu, Z. (2011). Knowledge-Driven Activity Recognition in Intelligent Environments. *Pervasive and Mobile Computing*, *7*(3), 285–286. https://doi.org/10.1016/j.pmcj.2011.05.001

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6), 066111. https://doi.org/10.1103/PhysRevE.70.066111

Colizza, V., Flammini, A., Serrano, M. A., & Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nature Physics*, *2*(2), 110. https://doi.org/10.1038/nphys209

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, *17*(1), 124–147. https://doi.org/10.1111/j.1467-9671.2012.01359.x

Dang, L., & Zhang, L. (2010). Method of discriminant for Chinese sentence sentiment orientation based on HowNet. *Application Research of Computers*, *27*(4), 1370–1372. https://doi.org/10.3969/j.issn.1001-3695.2010.04.044

Dashti, S., Palen, L., Heris, M. P., Anderson, K. M., Anderson, T. J., & Anderson, S. (2014). Supporting Disaster Reconnaissance with Social Media Data : A Design-Oriented Case Study of the 2013 Colorado Floods. In *11th International ISCRAM Conference* (pp. 632–641).

De Masi, G., Iori, G., & Caldarelli, G. (2006). Fitness model for the Italian interbank money market. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *74*(6), 066112. https://doi.org/10.1103/PhysRevE.74.066112

Dhingra, S., Ottaviano, G., Sampson, T., & Reenen, J. Van. (2016). The impact of Brexit on foreign investment in the UK. In *BREXIT 2016 Policy analysis from the Centre for Economic Performance* (pp. 24–33). Retrieved from http://cep.lse.ac.uk/BREXIT/

Dong, Z., Dong, Q., & Hao, C. (2010). HowNet and Its Computation of Meaning. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 53–56). https://doi.org/10.5555/1944284.1944298

Duncan, A. D. (1971). The View from the Inner Eye: Personal Management of Inner and Outer Behaviors. *TEACHING Exceptional Children*, *3*(3), 152–156. https://doi.org/10.1177/004005997100300313

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). AAAI Press. https://doi.org/10.5555/3001460.3001507

Feng, Y., & Sester, M. (2018). Extraction of Pluvial Flood Relevant Volunteered Geographic Information (VGI) by Deep Learning from User Generated Texts and Photos. *ISPRS International Journal of Geo-Information*, *7*(2), 39. https://doi.org/10.3390/ijgi7020039

Forster, P., Forster, L., Renfrew, C., & Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences*, 202004999. https://doi.org/10.1073/pnas.2004999117

Fredline, E., & Faulkner, B. (2001). Residents' reactions to the staging of major motorsport events within their communities: a cluster analysis. *Event Management*, *7*(2), 103–114. https://doi.org/10.3727/152599501108751515

Fredline, E., & Faulkner, B. (2017). Variations in residents' reactions to major motorsport events: Why residents perceive the impacts of events differently. *Event Management*, *7*, 115–125. https://doi.org/10.3727/152599501108751524

Freni, D., Vicente, C. R., Mascetti, S., Bettini, C., & Jensen, C. S. (2010). Preserving location and absence privacy in geo-social networks. In *International Conference on Information and Knowledge Management, Proceedings* (pp. 309–318). https://doi.org/10.1145/1871437.1871480

Galton, F. (1869). *Hereditary Genius: An Inquiry Into Its Laws and Consequences*. London, Great Britain: Macmillan Publishers. https://doi.org/10.1037/13474-000

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gao, Y., Wang, S., Padmanabhan, A., Yin, J., & Cao, G. (2018). Mapping spatiotemporal patterns of events using social media: a case study of influenza trends. *International Journal of Geographical Information Science*, *32*(3), 425–449. https://doi.org/10.1080/13658816.2017.1406943

Gips, J., & Pentland, A. (2006). Mapping human networks. *Proceedings - Fourth Annual IEEE International Conference on Pervasive Computing and Communications, PerCom 2006*, *2006*, 159–168. https://doi.org/10.1109/PERCOM.2006.35

Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: A research frontier. *International Journal of Digital Earth*, *3*(3), 231–241. https://doi.org/10.1080/17538941003759255

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Guan, X., Fang, C., & Luo, K. (2012). Regional economic development disparity of China: an application of spatial field. *SCIENTIA GEOGRAPHICA SINICA*, *32*(9), 1055–1065. https://doi.org/10.13249/j.cnki.sgs.2012.09.004

Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature, 433*(7028), 895–900. https://doi.org/10.1038/nature03286.1.

Guo, D. (1996). China's Tibetan population and population in Tibet. *China Population Today, 13*(2), 7–8.

Hall, W., Tinati, R., & Jennings, W. (2018). From brexit to trump: Social media's role in democracy. *Computer*, *51*(1), 18–27. https://doi.org/10.1109/MC.2018.1151005

Hamamatsu, Y., Inoue, Y., Watanabe, C., & Umezaki, M. (2014). Impact of the 2011 earthquake on marriages, births and the secondary sex ratio in Japan. *Journal of Biosocial Science*, *46*(6), 830–841. https://doi.org/10.1017/S0021932014000017

Healy, A. J., Malhotra, N., & Hyunjung, C. (2010). Irrelevant events affect voters ' evaluations of government performance. *Proceedings of the National Academy of Sciences, 107*(29), 12804–12809. https://doi.org/10.1073/pnas.1007420107

Heinrich, G. (2008). *Parameter Estimation for Text Analysis*. Retrieved from http://www.arbylon.net/publications/text-est2.pdf

Herman, E. (2001). Families made by science. Arnold Gesell and the technologies of modern child adoption. *Isis, 92*(4), 684–715. https://doi.org/10.1086/385355

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289–296). https://doi.org/10.5555/2073796.2073829

Horowitz, F. D. (1992). John B. Watson's Legacy: Learning and Environment. *Developmental Psychology, 28*(3), 360–367. https://doi.org/10.1037/0012-1649.28.3.360

Huang, Y., Li, Y., & Shan, J. (2018). Spatial-temporal event detection from geo-tagged tweets. *ISPRS International Journal of Geo-Information*, *7*(4), 150. https://doi.org/10.3390/ijgi7040150

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE, 40*(9), 1098–1101. https://doi.org/10.1109/JRPROC.1952.273898

Jahanbakhsh, K., & Moon, Y. (2014). The Predictive Power of Social Media: On the Predictability of U.S. Presidential Elections using Twitter. *ArXiv Preprint*. Retrieved from http://arxiv.org/abs/1407.0622

Jalloh, M. F., Li, W., Bunnell, R. E., Ethier, K. A., O'Leary, A., Hageman, K. M., … Redd, J. T. (2018). Impact of Ebola experiences and risk perceptions on mental health in Sierra Leone, July 2015. *BMJ Global Health, 3*(2), e000471. https://doi.org/10.1136/bmjgh-2017-000471

Jendryke, M., Balz, T., & Liao, M. (2017). Big location-based social media messages from China ' s Sina Weibo network : Collection , storage , visualization , and potential ways of analysis. *Transactions in GIS, 21*(4), 825–834. https://doi.org/10.1111/tgis.12266

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines*. Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-0907-3

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Karney, C. F. F. (2013). Algorithms for geodesics. *Journal of Geodesy, 87,* 43–55. https://doi.org/10.1007/s00190-012-0578-z

Konomi, S., Inoue, S., Kobayashi, T., Tsuchida, M., & Kitsuregawa, M. (2006). Supporting Colocated Interactions Using RFID and Social Network Displays. *IEEE Pervasive Computing*, *5*(3), 48–56. https://doi.org/10.1109/mprv.2006.60

Krumm, J., & Horvitz, E. (2015). Eyewitness: Identifying local events via space-time signals in twitter feeds. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '15* (pp. 1–10). https://doi.org/10.1145/2820783.2820801

Lameck, W. U. (2013). Sampling Design , Validity and Reliability in General Social Survey, *3*(7), 212–218. https://doi.org/10.6007/IJARBSS/v3-i7/27

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms : A comparative analysis. *Physical Review E*, *80*(5), 056117. https://doi.org/10.1103/PhysRevE.80.056117

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., … Alstyne, M. Van. (2009). Computational Social Science. *Science*, *323*(February), 721–723. https://doi.org/10.1126/science.1167742

Lévy, P. (2010). From social computing to reflexive collective intelligence: The IEML research program. *Information Sciences*, *180*(1), 71–94. https://doi.org/10.1016/j.ins.2009.08.001

Li, J., Ye, Q., Deng, X., Liu, Y., & Liu, Y. (2016). Spatial-Temporal Analysis on Spring Festival Travel Rush in China Based on Multisource Big Data. *Sustainability*, *8*(11), 1184. https://doi.org/10.3390/su8111184

Li, X., Mao, W., Zeng, D., & Wang, F. (2008). Agent-Based Social Simulation and Modeling in Social Computing. In *Intelligence and Security Informatics* (pp. 401–412). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-69304-8_41

Liégeois, R., Li, J., Kong, R., Orban, C., Ville, D. Van De, Ge, T., … Yeo, B. T. T. (2019). Resting brain dynamics at different timescales capture distinct aspects of human behavior. *Nature Communications*, *10*(1), 1–9. https://doi.org/10.1038/s41467-019-10317-7

Liu, H., & Du, G. (2017). Spatial-temporal pattern of China's economic development and its dynamic evolution: based on city level DMSP/OLS night-time lights data. *Chinese Journal of Population Science*, *37*(3), 17–29.

Liu, H., & He, L. (2015). Regional disparity in China and its evolution (2000 - 2012) ——re-examination based on DLI. *Review of Economy and Management*, (1), 141–146. https://doi.org/10.13962/j.cnki.37-1486/f.2015.01.020

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., … Shi, L. (2015). Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers*, *105*(3), 512–530. https://doi.org/10.1080/00045608.2015.1018773

Long, Y., & Wu, K. (2016). Shrinking cities in a rapidly urbanizing China. *Environment and Planning A*, *48*(2), 220–222. https://doi.org/10.1177/0308518X15621631

Lukowicz, P., Pentland, S., & Ferscha, A. (2012). From context awareness to socially aware computing. *IEEE Pervasive Computing*, *11*(1), 32–40. https://doi.org/10.1109/MPRV.2011.82

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., … Blanford, J. (2011). SensePlace2: GeoTwitter analytics support for situational awareness. In *VAST 2011 - IEEE Conference on Visual Analytics Science and Technology 2011, Proceedings* (pp. 181–190). Providence, RI, USA: IEEE. https://doi.org/10.1109/VAST.2011.6102456

Malmgren, R. D., Hofman, J. M., Amara, L. A. N., & Watts, D. J. (2009). Characterizing individual communication patterns. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 607–615. https://doi.org/10.1145/1557019.1557088

Martí, P., Serrano-estrada, L., & Nolasco-cirugeda, A. (2019). Social Media data: Challenges , opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, *74*(September 2018), 161–174. https://doi.org/10.1016/j.compenvurbsys.2018.11.001

Miao, Y. (2016). Analysis of population change in Hulunbeier City. *INNER MONGOLIA STATISTICS*, (6), 59–61. https://doi.org/10.19454/j.cnki.cn15-1170/c.2016.06.025

Moreno, J. M. P. (2017). *Do football victories affect social unrest? evidence from Africa*. Pontificia Universidad Católica de Chile. Retrieved from https://repositorio.uc.cl/handle/11534/21444

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from twitter's streaming API with Twitter's firehose. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013* (pp. 400–408).

Murzintcev, N., & Cheng, C. (2017). Disaster Hashtags in Social Media. *ISPRS International Journal of Geo-Information*, *6*(7), 204. https://doi.org/10.3390/ijgi6070204

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 100–108). Association for Computational Linguistics. https://doi.org/10.5555/1857999.1858011

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Nina. (2015). *Event Cartography: A New Perspective in Mapping*. Technische Universität München.

O'Leary, A., Jalloh, M. F., & Neria, Y. (2018). Fear and culture: Contextualising mental health impact of the 2014-2016 Ebola epidemic in West Africa. *BMJ Global Health*, *3*(3), 1–5. https://doi.org/10.1136/bmjgh-2018-000924

O'Neil, A., Sojo, V., Fileborn, B., Scovelle, A. J., & Milner, A. (2018). The #MeToo movement: an opportunity in public health? *The Lancet*, *391*(10140), 2587–2589. https://doi.org/10.1016/S0140-6736(18)30991-7

Obar, J. A., & Wildman, S. S. (2015). Social Media Definition and the Governance Challenge: An Introduction to the Special Issue. *Telecommunications Policy*, *39*(9), 745–750. https://doi.org/10.2139/ssrn.2637879

Ohmann, S., Jones, I., & Wilkes, K. (2006). The Perceived Social Impacts of the 2006 Football World Cup on Munich Residents. *Journal of Sport & Tourism*, *11*(2), 129–152. https://doi.org/10.1080/14775080601155167

Olguín, D. O., Waber, B. N., Kim, T., Mohan, A., Ara, K., & Pentland, A. (2009). Sensible Organizations : Technology and Methodology for Automatically Measuring Organizational Behavior. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS*, *39*(1), 43–55. https://doi.org/10.1109/TSMCB.2008.2006638

Opsahl, T., Colizza, V., Panzarasa, P., & Ramasco, J. J. (2008). Prominence and Control : The Weighted Rich-Club Effect. *Physical Review Letters*, *101*(16), 168702. https://doi.org/10.1103/PhysRevLett.101.168702

Osborne, M., Moran, S., Mccreadie, R., Lunen, A. Von, Sykora, M., Cano, E., … O'Brien, A. (2014). Real-Time Detection , Tracking , and Monitoring of Automatically Discovered Events in Social Media. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 37–42). Baltimore, Maryland: Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-5007

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. *Stanford InfoLab*. Retrieved from http://ilpubs.stanford.edu:8090/422/

Papadopoulos, S., Zigkolis, C., Kompatsiaris, Y., & Vakali, A. (2011). Cluster-based landmark and event detection for tagged photo collections. *IEEE Multimedia*, *18*(1), 52–62. https://doi.org/10.1109/MMUL.2010.68

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.5555/1953048.2078195

Peng, S., Yang, A., Cao, L., Yu, S., & Xie, D. (2017). Social influence modeling using information theory in mobile social networks. *Information Sciences*, *379*, 146–159. https://doi.org/10.1016/j.ins.2016.08.023

Pentland, A. (2005). Socially aware computation and communication. *Proceedings of the Seventh International Conference on Multimodal Interfaces, ICMI'05*, (March), 199. https://doi.org/10.1145/1088463.1088466

Petkos, G., Papadopoulos, S., & Kompatsiaris, Y. (2012). Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval* (pp. 1–8). https://doi.org/10.1145/2324796.2324825

Portes, J. (2016). Immigration after Brexit. *National Institute Economic Review*, *238*(1), R13–R21. https://doi.org/10.1177/002795011623800111

Qian, S., Zhang, T., Xu, C., & Shao, J. (2016). Multi-Modal Event Topic Model for Social Event Analysis. *IEEE Transactions on Multimedia*, *18*(2), 233–246. https://doi.org/10.1109/TMM.2015.2510329

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, *76*(3), 036106. https://doi.org/10.1103/PhysRevE.76.036106

Ramos, A., Lazar, M., Filho, R. H., & Rodrigues, J. J. P. C. (2017). Model-Based Quantitative Network Security Metrics: A Survey. *IEEE Communications Surveys & Tutorials*, *19*(4), 2704–2734. https://doi.org/10.1109/COMST.2017.2745505

Ratkiewicz, J., Conover, M. D., Meiss, M., Goncalves, B., Flammini, A., & Menczer, F. (2011). Detecting and Tracking Political Abuse in Social Media. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media Detecting* (pp. 297–304). Barcelona: AAAI Press. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2850

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15* (pp. 399–408). Shanghai, China: Association for Computing Machinery. https://doi.org/10.1145/2684822.2685324

Rosvall, M., & Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, *104*(18), 7327–7331. https://doi.org/10.1073/pnas.0611034104

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123. https://doi.org/10.1073/pnas.0706851105

Ruflin, N., Burkhart, H., & Rizzotti, S. (2011). Social-data storage-systems. In *Workshop on Databases and Social Networks, DBSocial'11* (pp. 7–12). https://doi.org/10.1145/1996413.1996415

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *Proceedings of the 19th International Conference on World Wide Web*, 851–860. https://doi.org/10.1145/1772690.1772777

Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, *25*(4), 919–931. https://doi.org/10.1109/TKDE.2012.29

Savage, M., & Burrows, R. (2007). The Coming Crisis of Empirical Sociology. *Sociology*, *41*(5), 885–899. https://doi.org/10.1177/0038038507080443

Schinas, M., Papadopoulos, S., Petkos, G., Kompatsiaris, Y., & Mitkas, P. A. (2015). Multimodal Graph-based Event Detection and Summarization in Social Media Streams. In *Proceedings of the 23rd ACM international conference on Multimedia - MM '15* (pp. 189–192). Brisbane Australia: Association for Computing Machinery. https://doi.org/10.1145/2733373.2809933

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE*, *6*(5). https://doi.org/10.1371/journal.pone.0019467

Smith, T. W., Rasinski, K. A., & Toce, M. (2001). *America Rebounds : A National Study of Public Response to the September 11 th Terrorist Attacks*. Chicago.

Sobkowicz, P., Kaschesky, M., & Bouchard, G. (2012). Opinion mining in social media : Modeling , simulating , and forecasting political opinions in the web. *Government Information Quarterly*, *29*(4), 470–479. https://doi.org/10.1016/j.giq.2012.06.005

Stephens, M., & Poorthuis, A. (2015). Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems*, *53*, 87–95. https://doi.org/10.1016/j.compenvurbsys.2014.07.002

Su, X., Tong, H., & Ji, P. (2014). Activity recognition with smartphone sensors. *Tsinghua Science and Technology*, *19*(3), 235–249. https://doi.org/10.1109/TST.2014.6838194

Sugitani, T., Shirakawa, M., Hara, T., & Nishio, S. (2013). Detecting local events by analyzing spatiotemporal locality of tweets. *Proceedings - 27th International Conference on Advanced Information Networking and Applications Workshops, WAINA 2013*. https://doi.org/10.1109/WAINA.2013.246

Sun, Q., Guo, X., Jiang, W., DIng, H., Li, T., & Xu, X. (2019). Exploring the Node Importance and Its Influencing Factors in the Railway Freight Transportation Network in China. *Journal of Advanced Transportation*, *2019*, 1493206. https://doi.org/10.1155/2019/1493206

Sun, Y., Fan, H., Helbich, M., & Zipf, A. (2013). Analyzing Human Activities Through Volunteered Geographic Information: Using Flickr to Analyze Spatial and Temporal Pattern of Tourist Accommodation. In J. Krisp (Ed.), *Progress in Location-Based Services* (pp. 57–69). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34203-5_4

Tannous, W. K. (2019). the Fire of Notre Dame: Economic Lessons Learned. *WIT Transactions on The Built Environment*, *190*(13), 51–63. https://doi.org/10.2495/dman190051

Tian, X., Batterham, P., Song, S., Yao, X., & Yu, G. (2018). Characterizing Depression Issues on Sina Weibo. *International Journal of Environmental Research and Public Health*, *15*(4), 764. https://doi.org/10.3390/ijerph15040764

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 178–185). https://doi.org/10.1074/jbc.M501708200

Valkanas, G., & Gunopulos, D. (2013). How the Live Web Feels About Events. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 639–648). San Francisco: Association for Computing Machinery. https://doi.org/10.1145/2505515.2505572

Vargas-Silva, C., & Fernández-Reino, M. (2019). *EU Migration to and from the UK*. Oxford. Retrieved from https://migrationobservatory.ox.ac.uk/resources/briefings/eu-migration-to-and-from-the-uk/

Wang, D., Szymanski, B. K., Abdelzaher, T., Ji, H., & Kaplan, L. (2019). The age of social sensing. *Computer*, *52*(1), 36–45. https://doi.org/10.1109/MC.2018.2890173

Wang, L., Gu, T., Tao, X., Chen, H., & Lu, J. (2011). Recognizing multi-user activities using wearable sensors in a smart home. *Pervasive and Mobile Computing*, *7*(3), 287–298. https://doi.org/10.1016/j.pmcj.2010.11.008

Wang, S., Paul, M. J., & Dredze, M. (2015). Social media as a sensor of air quality and public response in China. *Journal of Medical Internet Research*, *17*(3), e22. https://doi.org/10.2196/jmir.3875

Wang, Yandong, Ruan, S., Wang, T., & Qiao, M. (2018). Rapid estimation of an earthquake impact area using a spatial logistic growth model based on social media data. *International Journal of Digital Earth*, 1–20. https://doi.org/10.1080/17538947.2018.1497100

Wang, Yandong, Wang, T., Tsou, M., Li, H., Jiang, W., & Guo, F. (2016). Mapping dynamic urban land use patterns with crowdsourced geo-tagged social media (Sina-Weibo) and commercial points of interest collections in Beijing, China. *Sustainability*, *8*(11), 1202. https://doi.org/10.3390/su8111202

Wang, Yin, & Zhang, S. (2017). Research of sentiment analysis for Chinese micro-blog topic. *Journal of Fuyang Normal University (Natural Science)*, *34*(2), 50–56. https://doi.org/10.14096/j.cnki.cn34-1069/n/1004-4329(2017)02-050-07

Wei, Y., Song, W., Xiu, C., & Zhao, Z. (2018). The rich-club phenomenon of China ’ s population flow network during the country ’ s spring festival. *Applied Geography*, *96*, 77–85. https://doi.org/10.1016/j.apgeog.2018.05.009

Wen, Z. (2003). *A Study on Negation in Modern Chinese.pdf*. Fudan University. Retrieved from http://cdmd.cnki.com.cn/Article/CDMD-10246-2003125349.htm

Winter, M. G., Shearer, B., Palmer, D., Peeling, D., Harmer, C., & Sharpe, J. (2016). The Economic Impact of Landslides and Floods on the Road Network. *Procedia Engineering*, *143*, 1425–1434. https://doi.org/10.1016/j.proeng.2016.06.168

Wu, Y. S. (2005). Taiwan’s domestic politics and cross-strait relations. *China Journal*, (53), 35–60. https://doi.org/10.2307/20065991

Wu, Z., Lin, Y., Gregory, S., Wan, H., & Tian, S. (2012). Balanced multi-label propagation for overlapping community detection in social networks. *Journal of Computer Science and Technology*, *27*(3), 468–479. https://doi.org/10.1007/s11390-012-1236-x

Xu, J., Li, A., Li, D., Liu, Y., Du, Y., Pei, T., … Zhou, C. (2017). Difference of urban development in China from the perspective of passenger transport around Spring Festival. *Applied Geography*, *87*, 85–96. https://doi.org/10.1016/j.apgeog.2017.07.014

Xu, L., Lin, H., Pan, Y., Ren, H., & Chen, J. (2008). Constructing the affective lexicon ontology. *Journal of The China Society for Scientific and Technical Information*, *27*(2), 180–185.

Xu, S., Li, S., & Huang, W. (2019). A spatial-temporal-semantic approach for detecting local events using geo-social media data. *Transactions in GIS*, *24*(1), 142–173. https://doi.org/10.1111/tgis.12589

Yan, Y., Eckle, M., Kuo, C., Herfort, B., & Fan, H. (2017). Monitoring and Assessing Post-Disaster Tourism Recovery Using Geotagged Social Media Data. *ISPRS International Journal of Geo-Information*, *6*(5), 144. https://doi.org/10.3390/ijgi6050144

Yang, L., MacEachren, A. M., Mitra, P., & Onorati, T. (2018). Visually-Enabled Active Deep Learning for (Geo) Text and Image Classification: A Review. *ISPRS International Journal of Geo-Information*, *7*(2), 65. https://doi.org/10.3390/ijgi7020065

Yang, Zhao, Algesheimer, R., & Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, *6*, 30750. https://doi.org/10.1038/srep30750

Yang, Zhenguo, Li, Q., Liu, W., Ma, Y., & Cheng, M. (2017). Dual graph regularized NMF model for social event detection from Flickr data. *World Wide Web*, *20*(5), 995–1015. https://doi.org/10.1007/s11280-016-0405-1

Yerva, S. R., Jeung, H., & Aberer, K. (2012). Cloud based social and sensor data fusion. In *15th International Conference on Information Fusion, FUSION 2012* (pp. 2494–2501). Singapore, Singapore: IEEE.

Yu, M., Si, W., Song, G., Li, Z., & Yen, J. (2014). Who were you talking to - Mining interpersonal relationships from cellphone network data. In *ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 485–490). https://doi.org/10.1109/ASONAM.2014.6921630

Zhang, C., Liu, L., Lei, D., Yuan, Q., Zhuang, H., Hanratty, T., & Han, J. (2017). TrioVecEvent: Embedding-based online local event detection in geo-tagged tweet streams. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Vol. Part F1296, pp. 595–604). https://doi.org/10.1145/3097983.3098027

Zhang, D., Huang, J., Li, Y., Zhang, F., Xu, C., & He, T. (2014). Exploring human mobility with multi-source data at extremely large metropolitan scales. In *Proceedings of the Annual International Conference on Mobile Computing and Networking, MOBICOM* (pp. 201–212). https://doi.org/10.1145/2639108.2639116

Zhang, L., & Pentina, I. (2012). Motivations and usage patterns of weibo. *Cyberpsychology, Behavior, and Social Networking*, *15*(6), 312–317. https://doi.org/10.1089/cyber.2011.0615

Zhang, Y., Li, Q., Huang, H., Wu, W., Du, X., & Wang, H. (2017). The Combined Use of Remote Sensing and Social Sensing Data in Fine-Grained Urban Land Use Mapping: A Case Study in Beijing, China. *Remote Sensing*, *9*(9), 865. https://doi.org/10.3390/rs9090865

Zheng, Y., Wu, W., Chen, Y., Qu, H., & Ni, L. M. (2016). Visual Analytics in Urban Computing: An Overview. *IEEE Transactions on Big Data*, *2*(3), 276–296. https://doi.org/10.1109/TBDATA.2016.2586447

Zhou, S., & Mondragón, R. J. (2004). The Rich-Club Phenomenon in the Internet Topology. *IEEE Communications Letters*, *8*(3), 180–182. https://doi.org/10.1109/LCOMM.2004.823426

Zhou, X., & Xu, C. (2017). Tracing the Spatial-Temporal Evolution of Events Based on Social Media Data. *ISPRS International Journal of Geo-Information*, *6*(3), 88. https://doi.org/10.3390/ijgi6030088

Zhu, R., Lin, D., Jendryke, M., Zuo, C., Ding, L., & Meng, L. (2019). Geo-tagged social media data-based analytical approach for perceiving impacts of social events. *ISPRS International Journal of Geo-Information*, *8*(1). https://doi.org/10.3390/ijgi8010015

Zhu, Ruoxin, Lin, D., Wang, Y., Jendryke, M., Xin, R., Yang, J., … Meng, L. (2020). Social Sensing of the Imbalance of Urban and Regional Development in China Through the Population Migration Network around Spring Festival. *Sustainability*, *12*(8), 3457. https://doi.org/10.3390/su12083457

Zhu, Ruoxin, Zuo, C., & Lin, D. (2019). Research on event perception based on geo-tagged social media data. In *Proceedings of the ICA* (p. 157). https://doi.org/10.5194/ica-proc-2-157-2019

Relevant publications

- Zhu, R., Lin, D., Wang, Y., Jendryke, M., Xin, R., Yang, J., ... & Meng, L. (2020). Social Sensing of the Imbalance of Urban and Regional Development in China Through the Population Migration Network around Spring Festival. *Sustainability*, *12*(8), 3457.

- Wang, Y., Deng, Y., Ren, F., Zhu, R., Wang, P., Du, T., & Du, Q. (2020). Analysing the spatial configuration of urban bus networks based on the geospatial network analysis method. *Cities*, *96*, 102406.

- Zhu, R., Zuo, C., & Lin, D. (2019, July). Research on event perception based on geo-tagged social media data. In *Proceedings of the ICA* (Vol. 2, 157).

- Cui, X., Wang, J., Wu, F., Li, J., Gong, X., Zhao, Y., & Zhu, R. (2019). Extracting Main Center Pattern from Road Networks Using Density-Based Clustering with Fuzzy Neighborhood. *ISPRS International Journal of Geo-Information*, *8*(5), 238.

- Lin, D., Zhang, Y., Zhu, R., & Meng, L. (2019). The analysis of catchment areas of metro stations using trajectory data generated by dockless shared bikes. *Sustainable cities and society*, *49*, 101598.

- Zhu, R., Lin, D., Jendryke, M., Zuo, C., Ding, L., & Meng, L. (2019). Geo-Tagged Social Media Data-Based Analytical Approach for Perceiving Impacts of Social Events. *ISPRS International Journal of Geo-Information*, *8*(1), 15.

Acknowledgements

Writing this final part of the doctoral thesis reminds me that the days of studying at the Technical University of Munich as a doctoral candidate are coming to an end. During these four years of PhD study, I have experienced various challenges and setbacks, accompanied by gaining knowledge and growth. I would like to express my heartfelt thanks to those who have given me great support and help during my PhD study.

First of all, I would like to express my sincere gratitude to my supervisor Prof. Dr.-Ing. Liqiu Meng, for her continuous support of my doctoral work from topic selection, research design to thesis writing. In addition, she always encouraged me to develop my soft skills by involving me in assisting course exercise, organizing meetings, and assisting the master's program. All kinds of interdisciplinary knowledge and ideas she shared with us during coffee breaks also opened up my horizons.

My gratitude also goes to Prof. Dr.-Ing. habil. Dirk Burghardt for reviewing my thesis and acting as the co-supervisor. His valuable comments and insights improved my understanding of the VGI and inspired further optimization of the thesis framework.

I am grateful to Prof. Jiayao Wang and Prof. Jianzhong Guo for supervising me during my study in Zhengzhou, China. They were the ones who have encouraged me to take the challenges of doing PhD in TUM and offered helpful advises at the critical moments of my PhD study.

I would like to offer my gratitude to my colleagues, who were always nice and helpful in numerous ways. I enjoy the pleasant time with them, including but not limited to the assisting exercise of course ‘Geoinformation’, discussion in the group meeting, communication during the coffee break, Cartorun team for TUM Campuslauf, and annual Christmas party.

I would like to express my thankfulness to all my friends for making my life interesting and beautiful. I cherish every moment with them, including but not limited to eating together in mensa, outdoor sports like hiking and boating, cooking together, cutting hair for each other, and helping each other move.

I would like to acknowledge the authors of the references for their rich and detailed works, which provided valuable reference and inspiration for my research. I would like also to thank Dr. Michael Jendryke for sharing with us the Sina Weibo Data.

I owe my deepest gratitude to my family for their selfless love and endless understanding and encouragement.

And financial support for my PhD study from China Scholarship Council (CSC) is gratefully acknowledged.