GEOGRAPHIES OF ONLINE SOCIAL INTERACTION: A BIG DATA ANALYTICS APPROACH TO SOCIAL MEDIA PLATFORM SINA WEIBO

CHAOWEN WANG
Bachelor of Science, Xiamen University, 2006

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfilment of the
Requirements for the Degree of

MASTER OF SCIENCE

Department of Geography
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

GEOGRAPHIES OF ONLINE SOCIAL INTERACTION: A BIG DATA ANALYTICS APPROACH TO SOCIAL MEDIA PLATFORM SINA WEIBO

CHAOWEN WANG

Date of Defence: September 4, 2018

Dr. W. Xu Supervisor	Professor	Ph. D.
Dr. J. Zhang Co-Supervisor	Associate Professor	Ph. D.
Dr. Y. Chali Thesis Examination Committee Member	Professor	Ph. D.
Dr. K. Klein Thesis Examination Committee Member	Professor	Ph. D.
Dr. A. Akbary Chair, Thesis Examination Committee	Professor	Ph.D.

Abstract

Social media has revolutionized many aspects of people's social life. However, few studies have utilized massive individual-level data from social media to examine the effects of geography.

In this study a program was developed to collect and analyze data from Sina Weibo in ten selected Chinese cities. Four geographic concepts, i.e., borders, distance, places, and urban system hierarchy were chosen to measure the geographic effects by investigating geographical distribution of people's connections and comparing tweets similarity between different cities. The results show that these geographic concepts are playing an important role in the formation of new online connections and shaping people's interests. Social media users still tend to establish connections and share more common interests with people who live in the same city or close to them. People who live in the first-tier cities have more opportunities to establish connections across the country and their interests cover a broader range.

Acknowledgement

I would like first to express my deepest appreciation to my supervisor, Dr. Wei Xu, who has provided constant support and help throughout the process. It is impossible to finish this program without your encouragement and guidance. I am not only extremely grateful for your irreplaceable academic expertise when I was struggling with my study, but also thank wholeheartedly for your patience and understanding when my research is progressing slowly. I have been truly fortunate to have the opportunity to study with a supervisor who cared so much and taught me a lot both in academic research and personal life. It is a great honor for me to be your student.

I also would like to take this opportunity to express my gratitude to my co-supervisor Dr. John Zhang. Your door is always open to me and the constructive advice and generous help are always available whenever I needed it. Without your instruction and guidance, I would have wasted much more time on finding proper research methods to collect and analyze the data.

Also, profound gratitude goes to Dr. Yllias Chali and Dr. Kurt Klein, my dedicated committee members. Your insightful and detailed suggestions largely improve the quality of this study. I am also truly indebted for your understanding of my application to extend my program.

A special thanks to my fellow graduate students, Li Yu and Rongxu Qiu, for your help, inspiration and of course friendship. I cannot remember how many times Mr. Yu and I discussed each other's researches on the balcony of 6th floor in Alberta's ruthless winter which definitely gave me countless inspirations and great courage to persevere. With you around, completing this program is no longer a lonely journey.

Finally, but by no means the least, thanks go to my wife, mom and dad for their unbelievable and continued support along the way. The decision to give up my previous career and pursue my passion for geography in Canada is still baffling you. However, you did your best to make my life easier even if you did not understand. I love you and I dedicate this thesis to you.

Table of Contents

List of Tables

List of Figures

Prologue

In one of Giuseppe Tornatore's epic trilogy films, <The Legend of 1900>, the protagonist pianist explained to his friend why he would never leave the ship to show the world his musical talent, "In all that sprawling city, there was everything except an end. Take a piano. The keys begin, the keys end. You know there are 88 of them and no-one can tell you differently. They are not infinite, you are infinite. And on those 88 keys the music that you can make is infinite. But you get me up on that gangway and roll out a keyboard with millions of keys, that keyboard is infinite. There's no music you can play. That's God's piano." Millions of people live in every corner of the cities, working in different occupations, developing different interests and holding different views about the world and society. Every moment, countless connections form and disappear between individuals, which make up and transform the urban social fabric whose complexity and grandness is beyond human imagination. I hope this study can be helpful in better understanding the cities in the era of social media, giving the pianist some courage and confidence to leave the ship and have the opportunity to see and appreciate the beauty and the magnificence of the cities.

CHAPTER ONE

1. Introduction

1.1 Background

Social media is an online interactive platform where a large number of users form an online virtual community, allowing them to communicate with each other, publish, share and disseminate information. As social media drastically increases its popularity, a growing number of people start to spend a great amount of time maintaining and expanding their social relationships, sharing information, and acquiring knowledge through various online social media platforms. The human social relationships and activities have extended from physical space to virtual cyberspace. According to the recent data, by the September of 2014, 52% of American online adults use two or more social media platforms (Duggan, Ellison, Lampe, Lenhart, & Madden, 2014). Social media has become an embedded part of people's lives.

The scope of influence of social media has continuously expanded, and it has revolutionized many aspects of people's social life. The traditional communication barriers caused by the friction of distance, the difference in social classes, and the variations in languages and cultures, are argued to be broken down to some extent. Social media also breaks the monopoly of traditional media and forms a self-media platform where everyone has the

ability to produce and disseminate information easily and timely. Moreover, compared with traditional media channels, social media is more effective and superior in acquiring information and interacting with multiple parties simultaneously (Huang, Yang, Baek, & Lee, 2016). Network events often ferment and spread rapidly through online social networks, propagated by public opinions in online discussions. Because of its growing reach and influence, social media promotes social changes in the real world.

From early forums, blogs, podcasts to comprehensive social media platforms, the development of online social network shows a trend of diversification in response to the need of online communities. Some social media platforms, e.g., Facebook, focus on maintaining and strengthening pre-existing social connections, while microblogging services, e.g., Twitter and Sina Weibo, are good at regrouping people and establishing new connections based on their common interests, beliefs or geographic locations (Boyd & Ellison, 2007). Microblogging platforms usually have a limitation on the lengths of a single post. In addition, relations on microblogging platforms are not mutual. These unique features enable microblogging services to become online channels where information can be quickly generated and diffused. As a type of social media, microblogging service mimics the social fabric of human beings, organizes a large number of users into social networks, and meets the needs of personalized information publishing, sharing and acquiring.

Traditionally, people have relied on social and geographic proximity to establish social connections, such as with classmates, neighbors, and co-workers. Since the advent of the internet, people have been given a new way to build relations. In virtual cyberspace, individuals have a lower cost to form and maintain social relationships across time and space boundaries than in offline physical societies. For example, a Toronto Raptors fan living in Calgary can easily find friends through online forums with whom he shares hobbies that he would otherwise struggle to reach in his daily life. An entrepreneur who runs a shoe factory in China can find potential customers from as far away as Africa via the Internet. Does this mean that the formation of social connections is no longer affected by geography?

Similarly, people's interests, opinions and attentions are deeply influenced and shaped by their friends and relatives around them. If this influence extends to the scale of a city, it will gradually form a city-specific culture and tradition, creating a sense of place at a city scale. But the internet allows people to have instant access to a wider variety of information and opinions, which leads to unprecedented cultural diversity. Then, are the unique culture and traditions belonging to one city gradually eroded?

Few studies have attempted to answer these questions, mainly because previously, it was expensive and time-consuming to obtain individual-level

data through surveys or interviews. Sometimes, these data are too sensitive to share with other scholars. Luckily, social media offers a way to resolve these problems by providing a massive source of self-reporting and publicly available data. In addition, people likes to express their "true selves" through online channels (Marriott & Buchanan, 2014) which means the data collected from social media might be more reflective of what their users really think and do. Meanwhile, the development of related research methods and algorithms as well as ever-growing computational capacity provide an opportunity to investigate these problems at an unprecedented scale and detail, which is not an option for the previous generations of geographers.

1.2 Research Objective

In the past few decades, the rapid development of high-speed transportation networks, the explosive rise of information and communication technology (ICT), and the spread of globalization have led many researchers to question the influence of geography on human behavior patterns. Some argue that traditional geographic barriers have been diminished and even eliminated. Others believe that the role of geography as a core force driving economic and social change is constantly decreasing. Some commentators and researchers even declare the death of distance and the end of geography (Friedman T. L., 2005). This view was certainly refuted by many researchers (Stiglitz, 2007) (Blij,

2007). However, these arguments and empirical studies were mostly based on the data of location of corporation headquarters, internet infrastructure distribution, and telecommunication call logs (Zook, 2001) (Townsend, 2001) (Agnes, 2009) (Moka, Wellman, & Basu, 2007). Given that social media has drastically reshaped the way in which people interact with each other and the information is disseminated, findings and conclusions from previous studies might no longer be applicable in the era of social media. It is therefore particularly necessary and important to revisit this debate by leveraging the large amount of high-quality individual-level behavioral data provided by social media.

Big data are large and complex data sets that usually come from new data sources. Big data from social media will undoubtedly enable researchers to reveal unprecedented details of human behavioral patterns and thus provide more insights into the modern social fabric. However, how to collect this data effectively and turn this relational and textual data into relevant data still remains a challenge. On the one hand, many existing studies on social media conducted by computer scientists focus more on the network structure than on the effects of geography on human behaviors. On the other hand, geographers are constrained by the limitation of analytical tools and methods and hence are largely unable to make better use of this vast amount of data to study this new

type of online social network. Therefore, an interdisciplinary study is urgently needed to bridge this gap.

To fill the above mentioned research gap, this study aims to combine with the knowledge of computer science and geography to deepen our understanding of online human behaviors and the effects of geography in online social media. Specifically, two research objectives are proposed:

1) To develop an analytical framework for collecting, cleaning and converting social media data.

Social media not only provides platforms for people to interact, but also provides massive data for geographers to study people's social behaviors. However, this new data source has rarely been utilized by geographers largely because of a lack of analytical and computational tools to collect and analyze this data. This study aims to develop a computer program that can be used to efficiently collect and clean social media data without being constrained by social media platforms' application programming interfaces (APIs). Also, this study intends to explore a way to convert relational and textual data into vector data that can be measured and compared. The developed research methods and analytical tools can be applied to other related social media research.

2) To answer the question of whether geography still matters by systematically analyzing these data in a quantitative manner.

This study plans to address this issue from two perspectives. The first is to investigate the effects of geography on the formation of new social relationships on social media. Now that social media has broken the spatial barrier and given people maximum freedom to establish new online connections, is the distribution of these connections still highly concentrated in a certain area? In other words, is the distribution still heavily influenced by the geographic attributes of social media users?

The second perspective is to study the contents published on social media and compare their similarity between cities. To tackle this problem, this study plans to calculate the contents' similarity between every two social media users. Pairwise contents similarity then will be aggregated to the city level. Like the first perspective, this study aims to find patterns to reveal the underlying relationships between contents similarity and geography.

In this study, in order to present the results in a quantitative manner, the effects of geography are conceptualized into four geographic concepts, i.e., borders, geographic distance, places and urban system hierarchy, each of which is employed to measure the friction of distance.

1.3 Outline

To accomplish the proposed research goal, the remaining chapters of this dissertation is organized as follows.

Chapter 2 first reviews social media related studies and points out the importance and effectiveness of using social media either as a study object or data source. Then, the definitions and development of social network analysis are outlined. The long-standing debate about whether information technology has declared the end of geography as well as the definitions and related research about the four key geographic concepts are also elucidated. This chapter introduces various methods to solve the problem of finding similar objects. The reason to choose Sina Weibo as the study object is presented in Chapter 3. This chapter also introduces the characteristics of the study area and details the process of data acquisition, cleaning, and storage. Finally, why and how to use TF-IDF and cosine similarity algorithms to convert textual data to numeral data and calculate similarity values between cities are introduced. Chapter 4 scrutinizes the findings of this study by classifying them into four categories according to geographic concepts, each of which is investigated from two perspectives including formation of new online connections and contents similarity between cities. In the end, Chapter 5 discusses the results by comparing them with previous findings reported in the literature. Also, the contributions and limitations of this study are presented.

CHAPTER TWO

2. Literature Review

This chapter starts with a review of previous works on social media and points out the importance and effectiveness of using social media either as a study object or data source. Next, the essential definitions and development of social network analysis are outlined. The historical background of the debate about whether geography still matters as well as the definitions and related research about four key geographic concepts are also introduced. This chapter continues to evaluate various methods to solve the problem of finding similar objects on social media.

2.1 Importance and Usefulness of Social Media

Numerous studies have analyzed and stated the importance of social media from different academic fields and perspectives, such as sociology, economics, environmental science and political science. This section aims to demonstrate the importance and usefulness of related research in this area by reviewing some of them.

Social media platforms play a vital role in organizing public demonstrations, especially in the authoritarian regimes. Usually, people do not have the right to establish associations and freedom of speech in these countries. However,

through online social media, people who hold opposing political views can communicate and organize together. Social media platforms also provide channels for people to express their dissatisfaction and anger. As Huang and Yip (2012) argue, online social media has become more and more important as an information-disclosure and discussion platform. The information and news disperse so quickly online that they have been already transferred to a multitude of different channels before an authoritarian government finds them out and deletes the original sources. In addition, for some local social movements, even though the local government can control the local media, it is difficult to censor national or global online channels in time. Numerous empirical studies of Arab Spring also note that social media has had crucial impacts on social reform and revolution (Eltantawy & Wiest, 2011), especially under circumstances where open media and freedom of speech are absent (Khondker, 2011).

Social media also promotes government's openness and transparency (Bertot, Jaeger, & Grimes, 2010). Various government departments start to use social media to connect with the public. Online platforms help to break down the communication barrier, establish trust, collect diverse opinions and enhance the quality of decision/policy making process. User-generated contents in social media have become an important channel to collect corruption information.

Empirical studies indicate that social media can reduce corruption both internally and externally (Shim & Eom, 2008).

With the timely and cost-effective user-generated contents, which include the georeferenced information, online social media platforms evolve as reliable data sources for analyzing social issues. A growing body of literature focuses on utilizing this data to improve disaster management and emergency planning (Yin, Lampert, & Cameron, 2012). Longueville et al. (2009) analyze Twitter's data created during a forest fire in Marseille in 2009 from four aspects: the temporal dimension, geographic dimension, user classification and information dissemination. The study demonstrates the online social networks' positive impacts on improving emergency planning. Another research about wildfires confirms the importance of social media during disasters in not only disseminating relevant and accurate information in time, but also providing platforms for people to support each other (Sutton, Palen, & Shklovski, 2008). By comparing tweets and re-tweets density, researchers have proved that social media can provide timely and accurate supplemental information about earthquakes that may be crucial to better allocate rescue resources (Liang, Caverlee, & Mander, 2013). Similarly, Stollberg and Groeve (2012) state that, as a valuable data source, social media can improve the awareness of disaster situations and provide additional information for emergency responders to make better decisions.

Online education is not an innovative concept, but with the evolution of technology and advancement in the recent decades, new rising online platforms can make students and instructors communicate and interact with each other efficiently and timely. Also, these platforms grant students more flexibility and opportunities to explore themselves, which was usually impossible previously because of geographic distance, time conflicts or economic limits. Many prestigious universities and world-famous professors have joined online platforms to offer online classes (Friedman & Hershey, 2013). An analysis of empirical studies of online learning finds that students who take online education deliver better performance compared with those receiving face-to-face instruction, because learners have more control on time and can take advantage of multimedia of online learning platforms (Means, Toyama, Murphy, Bakia, & Jones, 2009).

A growing body of literature focuses on the relations between social media and migrants. By interviewing 65 migrants in Ireland, Komito (2011) found that the social media helped migrants keep connections with their community in their home country. On the other hand, this kind of connection inevitably slows down the process of integration in the host country. On the contrary, another study argues social media plays a vital role to facilitate the process of integration and migration (Dekker & Engbersen, 2014).

The popularity of social media goes hand in hand with the popularity of smartphones. More and more users are sharing their real-time location information generated by the smartphone's GPS module on social media platforms. Using this massive amount of geotagged data to study human mobility patterns has become a research hotspot in the field of Geographic Information Science (GIS). Unlike the taxi trajectory data (Jiang, Yin, & Zhao, 2009) (Liu, Kang, Gao, Xiao, & Tian, 2012) or cellular network data (Becker, et al., 2011) (González, Hidalgo, & Barabási, 2008), which were mainly used by previous studies, geotagged data are easier to acquire and have a finer granularity. Liu et al. (2014) utilize the check-in data to confirm the distance decay effect in spatial interactions. Hasan et al. (2013) discover a relationship between the popularity of a place and the likelihood of choosing it as a destination. By using Twitter data, Hawelka et al. (2014) prove that geotagged data can be considered as a validated proxy for human mobility.

Using textual data posted by users on social media to estimate geographic locations is another hot topic in GIS and computer science. Based only on textual contents on Twitter, researchers have designed various probabilistic frameworks to effectively predict users' city-level locations (Cheng, Caverlee, & Lee, 2010) (Chandra, Khan, & Muhaya, 2011). Han et al. (2014) tackle this problem by focusing on location indicative words and argue that textual content posted by social media users can reflect their geospatial information

because some words are used disproportionately in different regions. However, previous studies focus only on geotagged data or textual data, with few studies combining both of them. In addition, geotagged data need to be published voluntarily by social media users. It certainly can reveal specific details, but these findings can only represent a subset of social media users. There are still a large number of users who are reluctant or rarely publish their real-time geotagged information for privacy or other reasons. On the contrary, the location data that are presented in every user profile are rarely utilized by scholars. This needs more attention.

The existing literature has demonstrated the great potential and importance of social media in promoting social reform, improving quality of online education and disaster management, as well as studying human mobility patterns. However, the relevant research on the relationship between geography and human behavior patterns is scarce and calls for more research in this field.

2.2 Social Network Analysis

Social media, even though it has many new features, is still a type of social network. Thus, to review the development of social network analysis and give related definitions is important and essential.

2.2.1 Definition of Social Network Analysis

Social media is also called social network sites (SNS). There are several different ways in which social network is defined. According to Ellision (2007),

"Social network sites as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system" (p. 211). Marin and Wellman (2011) define "a social network is a set of socially relevant nodes connected by one or more relations" (p. 11). Social network is built upon node and relation. Node (actor, vertex, agent, or player) is the unit that is connected by the relations (Marin & Wellman, 2011). Bellotti (2015) defines the node as "the first basic element of networks and it represents the unit of analysis or actors in the network" (p. 6). Nodes are persons and organizations, and any other units that can be connected. Relation (link, edge, tie or arrow) indicates the tie between individuals and may indicate the strength of the tie or the frequency of interaction between individuals (Daly, 2010). Bellotti (2015) defines the relation as "the second basic element of networks and it represents the tie that connects two nodes" (p. 6). A directed relation starts with one node and ends with another, while an undirected relation connects two nodes with no particular direction. Relations can be friendships, citations, collaborations, web links or any other connections

between these particular units (Wasserman & Faust, 1994). In addition, between two nodes, there might be multiple different relations. For example, A is B's manager in a company. At the same time, A is also B's team player on a basketball team.

There are several other concepts that are useful in characterizing social network. Size is the number of nodes in a network. Distance is the length of the shortest path between two nodes. The average distance is the average length of all the shortest paths in a network. Density is the number of all actual ties divided by the number of all possible ties in a network.

Social network analysis is the study of social interactions among social actors. As Leinhardt (2013) states, social network analysis is a dominant paradigm that "guides the selection of the social behavior data that are studied, influences the way these data are organized for analysis, and specifies the kinds of questions addressed" (p. xiii). Scott and Carrington (2011) suggest that "social network analysis is a structuralist paradigm: it conceptualizes social life in terms of structures of relationships among actors, rather than in terms of categories of actors" (p. 6).

Three types of networks that have been studied: ego-centric, socio-centric, and open-system networks (Kadushin, 2012). Ego-centric networks are networks where nodes are all connected with a single node. Socio-centric networks are networks where all nodes are confined in a closed system, such as all employees

in a company. Open-system networks are networks with no clear defined boundary.

2.2.2 Development of Social Network Analysis

It is widely accepted that the development of social network analysis began from the 1930s when Jacob Moreno introduced a quantitative method for measuring social relationships - sociometry. Moreno also invented a way to visualize social networks with points and links - sociogram (Scott & Carrington, 2011). However, by the 1970s, the social network approach was still not recognized as a paradigm to social science research (Freeman, 2004). Sociologist Harrison White and his students changed the situation, and their works developed a formal methodology for social network analysis. White not only created many mathematical models and theories to understand a society but also changed the traditional individual-based approach to network-based perspective to study it (Wellman, 1997).

In the 1990s, Watts and Strogatz's revolutionary research about Milgram's previous work on the "small world" phenomena (Milgram, 1967) opened a door for social network analysis (Watts & Strogatz, 1998). "Small world" phenomena or "six degrees of separation" is one of the fundamental theories of social network analysis. In the 1960s, Milgram chose 296 participants who lived in the Midwest states, such as Nebraska, Kansas, and asked them to forward one letter to a person who lived in Boston. If the participants personally knew

the person, they could mail the letter directly to the target person; if not, they could only mail the letter to one of their acquaintances who might know the target person. The average length of forwarding chain was six which was surprisingly short. This experiment revealed that we all live in a "small world" and are closely connected with each other.

From then on, many studies focus on small world phenomena and degree distribution. A recent study of Facebook reveals that more than 99% of active Facebook users belong to a single network, and the average distance between users is 4.7 (Ugander, Karrer, Backstrom, & Marlow, 2011). Another study on Twitter found the similar results that the average distance between Twitter users was 4.12 (Kwak, Lee, Park, & Moon, 2010). These studies prove that the "small world" phenomenon also exists on social media platforms.

According to density, social networks can be characterized as highly clustered networks or networks with structural holes. Burt (1992) argues that a densely connected network is inefficient, which only returns less diverse or even redundant information. A structural hole is a relationship between non-redundant contacts and provides additive network benefits. In the same way, "weak ties" theory also focuses on structural holes in the social network (Granovetter M. , 1977). Granovetter claims that there are two kinds of ties connecting people in a social network. The strength of ties is a general sense of closeness between people. Strong ties refer to stable and closer interpersonal

relationship, such as relatives and friends. On the contrary, weak ties refer to more tenuous relationships, such as acquaintances. Granovetter states that weak ties are much more valuable in the job market because they can facilitate the flow of information across different groups. As Granovetter noted, the connections between microblogging service users are typical weak ties. Other studies reveal the same effect of the strength of weak ties on online social media platforms. As a critical factor to bridge distant clusters within a social network, weak ties are not only ubiquitous on social media, but also are easier to maintain and more likely to be converted to strong ties because of the power of information technology (Grabowicz, Ramasco, Moro, Pujol, & Eguiluz, 2012) (Ellison, Steinfield, & Lampe, 2007).

Social network analysis, as a paradigm for the systematic study of society and human beings, has developed rapidly in recent years. In the past, it was limited by the difficulty of obtaining large-scale network data. Now with the popularity of social media, researchers have more opportunities and challenges to advance our understanding of society and humanity to a deeper level. Existing literature has shown the great potential of utilizing social network analysis in social media research, such as community detection, event, link and interest prediction, information dissemination and so on. However, most studies focus on the relationships while ignoring the contents created by social media users.

In addition, many research methods can only be applied on a specific social media platform which lacks flexibility.

2.2.3 The End of Geography or Geography still matters

By introducing the first law of geography: "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p. 236) , Tobler laid the foundation of the meaning of geographic study. However, the friction of distance is not static but dynamic, subject to changing technology. As Castells suggests, a shift is occurring from a "space of places" to "space of flows" (Castells, 1989).

There has been a long-standing debate about whether technological progress, especially the ICT sector development, will erode the effect of geographic distance and eventually declare the death of geography. The declining role of geography has been long recognized because of the advance of transportation and communication technology. For example, Harvey (1989) argues that space and time have been compressed by this new technology which enables instant communication, and traditional spatial-temporal boundaries are collapsed. Robins and Hepworth (1988) claim that the ideas of space and time have been fundamentally reconstructed by modern technologies, which have become decisive factors to shape people's social life. There are also researchers questioning the relevance of distance in cyberspace social media. Mitchell (1996) suggests that the cyberspace is "fundamentally and profoundly antispatial"

and is “ambient - nowhere in particular but everywhere at once” (p. 8). Cairncross (1997) even contends a borderless world will be shaped by communication revolutions and proclaims “the death of distance”.

However, geographers in general insist on the role of distance in constructing our social society even though such a role has declined due to technological development. Kitchin (1997) summarizes three reasons why geography still matters and remains significant: (1) globalization is not uniformly distributed among different countries; (2) information is often more useful in the local area; (3) cyberspace exists in real world that has a “spatial fixity”. Scott (2010) further suggests that, contrary to the predication of the end of geography (O'Brien, 1992) (Graham, 1998), geography is becoming more important because “globalization enhances the possibilities of heightened geographic differentiation and locational specialization” (p. 813). Empirically, Takhteyev et al. (2012) examine the effect of distance on online social media’ connections and argues that distance still has its influences on online social networks. These researchers believe that the traditional friction of distance still plays a vital role in cyberspace.

Most of these studies were conducted in the 1980s and 1990s, mainly from the perspective of telecommunication technologies and globalization. However, they rarely provide quantitative evidences to demonstrate their claims. With the rapid rise of social media, large data sets that contain geographic data

enable us to revisit this debate from the perspective of social media in a quantitative manner.

2.3 Conceptualize Geographic Effects

Because of the friction of distance, costs of social interaction vary and geographies of social interaction merge. The friction of distance can be conceptualized in different ways. To investigate the importance of geography in the age of social media, four geographic concepts are selected to explore whether geography still matters in social media, including borders, geographic distance, places, and urban system hierarchy, each of which is employed to measure the friction of distance in the literature. The very reason to select these four measures is that data about these four geographic concepts can be extracted from social media platforms and used for comparison, making it feasible to solve the research problems of this study.

2.3.1 Borders

Borders are lines that separate political, social and economic spaces, which usually correlate with boundaries of culture and language, as well as topographical features (Newman, 2006). People live in compartments which are delimited by borders. These lines may be invisible, but people's daily lives, economic activities, and social interactions are influenced by them, which in turn form the culture, habits, and traditions of a particular region and shape the

identity of the individual who lives within them. Borders create a certain social order (Albert, Jacobson, & Lapid, 2001) by distributing administrative responsibilities and allocating resources. People living in the same territorial subdivisions often speak the same language or dialect, share similar religious beliefs, and have more opportunities to interact with each other. So, borders provide an important reference for the study of human activities and behaviors (Thiemann, Theis, Grady, Brune, & Brockmann, 2010).

With the development of information technology and globalization, more and more border areas have been transformed from barriers to economic centers. People who live on both sides of borders have begun to understand, cooperate and coexist with each other (Gallusser, 1994). Therefore, some researchers assert the world has become “borderless” (Ohmae, 1991) (O'Brien, 1992), arguing that traditional boundaries are blurred and no longer important.

However, the majority of such studies focus on national boundaries, and very little has been done on the municipal and provincial ones. At the same time, there have been fewer studies using borders to group social media users and compare their online social connections and behaviors, which requires the attention of geographers.

2.3.2 Distance

China has a proverb called "close neighbors are better than distant relatives", which reflects the distance to the impact of interpersonal relationships. Distance

means the amount of space between two places or things (The Merriam-Webster Dictionary, 2016). In human geography, only considering the absolute physical distance sometimes cannot explain the geographic pattern of phenomena under study. It is necessary to integrate other factors, such as transportation cost, ethnicity, religion, language, and so forth, which constitute the relative distance. For two places or individuals, it is possible the physical distance between them is quite close, but their relative distance or cultural distance is large.

As a critical factor in determining spatial interactions, the importance of distance is always a central theme in geography (Know & Marston, 2002). Distance introduces different cultures, languages and time-zones which all require effort and time to overcome. Therefore, spatial interactions tend to take place more within closer regions. Distance's deterrent effect on human activity has been investigated by many geographers. As Butts (2002) claims, the relationship between distance and social structure is so strong there are few other factors that can claim such strength and generality. However, with the advent of telecommunication technology and rapid development of transportation networks, maintaining interactions with people over long distances becomes easier (Axhausen & Gärling, 1994). Since the traditional friction of distance is diminished in some respects, many researchers believe that physical distance is no longer a barrier for information and knowledge

diffusion. Some of them even proclaim "the death of distance" (Cairncross, 1997) (Hepworth, 1991). On the contrary, some researchers oppose this view and claim that the traditional sense of distance is only reconstructed and distance still matters in modern society (Olson & Olson, 2000) (Bradner & Mark, 2002) (Miller, 2004) (Takhteyev, Gruzd, & Wellman, 2012).

2.3.3 Places

Place is one of the fundamental concepts in geography, as Tuan (1979) claims "space and place together define the nature of geography". Place does not only mean a point or area on the surface of the earth. It is often associated with people's feelings and perceptions (Agnew & Duncan, 1989). This "sense of place" has a huge impact on human activities because it involves people's subjective and emotional attachment. As Cresswell (2014) claims "place is much more than a thing in the world but a way of seeing, knowing and understanding the world" (p. 18).

Place-related research topics, including housing, crime, gentrification, and employment have dominated the geographic literature over the past few decades. However, some researchers argue that modern technology and mass culture have destroyed the uniqueness of places and led to global homogenization (Seamon, 1979). Castells (1989) declares that "space of places" is shifting to "space of flows". Meyrowitz (1986) argues that electronic media is producing an increasingly "placeless" world which has no traditional behavior

patterns. The rise of social media along with the huge amounts of fine-grained volunteered geotagged data has given geographers an unprecedented opportunity to revisit the role of places in a cyberspace context (Haklay, 2010).

2.3.4 Urban System Hierarchy

As an important component of the urban system, urban system hierarchy is a traditional subject of urban geography and urban planning. Urban system hierarchy is a system to rank cities based on population, size or other social-economic factors.

Walter Christaller (1966) laid down the foundation of the idea of urban system hierarchy in his renowned Central Place Theory. He explained the reason behind hierarchical arrangements of different categories of central places by analyzing retail and service networks in the south of Germany. Larger cities, which are high in the urban hierarchy, provide various goods and services, whereas settlements lower in the urban hierarchy can only provide limited and basic ones. As Fekete (2014) states, urban hierarchy offers "a means to evaluate and compare metropolitan environments" (p. 251). Some researchers try to classify cities according to population (Borchert, 1967) (Partridge, Rickman, Ali, & Olfert, 2008), airline services (Taaffe, 1962), foreign headquarters (Centonze, 1989), employment (Winsborough, 1960) and other factors. Others tackle this problem from a perspective of networks (Beaverstock, Smith, & Taylor, 1999) (Neal, 2011). Either approach attempts to tease out the varying roles of cities in

organizing human activities in space, assuming higher order cities, such as global cities, have a greater spatial reach and implication.

Since the 1990s, the traditional urban system has been complicated by the popularity of the internet. The existing urban system hierarchy is widely questioned and challenged. Some researchers utilize domain names, internet infrastructure and web search activity to analyze and establish new urban hierarchies (Zook, 2001) (Townsend, 2001) (Lu & Huang, 2012).

However, there is a lack of research on the ever-changing urban networks in developing countries since most literature emphasize the global cities (Derudder, et al., 2010), especially from the perspective of social media. In addition, most studies aim to develop new ranking schemes rather than investigate how urban system hierarchy influences the spatial origination of human activities.

By breaking the abstract concepts of geography down to four key geographic concepts, this study aims to analyze and evaluate the importance of geography on social media and contribute to a better understanding of the long-standing debate about whether geography still matters.

2.4 Similarity Measures

In order to evaluate the effect of geography in terms of the four identified friction of distance attributes, it is essential to identify and employ a metric to

quantify social interactions on social media. Social media platforms host a great amount of user-generated contents, including videos, photographs, and textual contents. Numerous studies focus on utilizing them to create similarity measures to identify events, detect online communities and recommend friends and products. How to find similar objects is a significant issue in many fields, such as clustering, recommender systems and search engines (Ganesan, Garcia-Molina, & Widom, 2003). A similarity measure is usually an inverse distance metric to quantify the closeness between a pair of objects. Many similarity measures have been introduced and evaluated (Guy, Jacovi, Perer, Ronen, & Uziel, 2010), such as Euclidean, Mahalanobis, Hamming, Cosine, Jaccard, and so forth. However, there is still no firm theoretical foundation for any similarity measure (Findler & Leeuwen, 1979). Almost all the related studies were based on solid empirical evidences. There are generally two ways to calculate similarity measures according to the different data types used (Menczer, 2004): The first approach is link-based**.** Links or relations can be friendship, co-authorship, or any other relationship between two objects. In general, two objects are deemed to be similar if they are connected to similar objects in a network (Cai, et al., 2009). Popular link-based similarity measures include SimRank (Jeh & Widom, 2002), P-Rank (Zhao, Han, & Sun, 2009), Co-Citation (Small, 1973) and Bibliographic coupling (Kessler, 1963).

The second approach is content-based. Contents can be texts, videos, or any other multimedia contents related to a certain object. Among them, textual data are mainly used (Yoon, et al., 2014). For textual data, in general, two objects are deemed to be more similar when they share more common words in their contents.

Some researchers aim to find out the characteristics of similar users. Prantik et al. (2011) use social media user profile entries (tags) as keywords to calculate their semantic similarity and found that users who are directly connected are more similar, while similarity measure between users who are not connected shows no clear difference despite their topological distance. Anderson et al. (2012) investigate the online evaluation between similar users and find that users are more positive toward users who share more common interests or social ties.

Using similarity measures to detect community has been approached by many researchers (Pan, Li, Liu, & Liang, 2010) (Zhou, Lü, & Zhang, 2009). A community is a densely connected subset of a network. People who belong to the same community usually have similar behavior patterns due to social influences. Meanwhile, people who are similar to each other have higher possibility to establish connections in the future (Crandall, Cosley, Huttenlocher, Kleinberg, & Sur, 2008) (Guy, Jacovi, Perer, Ronen, & Uziel, 2010).

Accurate similarity measures are also crucial to the success of recommender system. Effective marketing requires recommendation of personalized services or products to potential clients who are similar to their existing clients rather than blindly targeting many advertisements to people who might not be interested. Social media platforms also need to recommend similar accounts or groups to users to keep them active (Ricci, Rokach, & Shapira, 2011). Chen et al. (2009) compare link-based, content-based and hybrid similarity measures and found that all of these algorithms are effective in recommending new relations. The link-based algorithm performs better at finding known contacts, while the content-based algorithm is good at discovering new friends.

Due to Sina Weibo's restrictions, a user's full online relationships cannot be acquired. Therefore, link-based similarity measures are not realistic for this study. In addition, link-based similarity measures are usually tailored to a specific social media platform, which lacks adaptivity. Meanwhile, they are severely affected by outliers who have no connection with other nodes in a social network (Mizzaro, Pavan, & Scagnetto, 2015). Besides, social connections sometimes are hard to be identified on some social media platforms. Although content-based similarity research has been widely studied in computer science for detecting communities and recommender systems, combining the contents similarity with people's geographic attributes to analyze the influence of geographic concepts on the formation of social relationships and interests'

development is still a new subject that deserves more attention from geographers.

2.5 Summary

Social media has proved its importance and usefulness in many disciplines. With its massive self-reporting and public available individual-level data, which previously were expensive and time-consuming to acquire, social media grants researchers a new opportunity to investigate the interplay of geography and social behavior. By conceptualizing the geographic effects into four key concepts, combined with the methods of social network analysis and similarity calculation, a comprehensive and quantitative analysis can be conducted to measure the complex social fabric and shed new light on the understanding of the role of geography in this digital age by providing empirical evidence to the existing literature.

CHAPTER THREE

3. Methodology

This chapter begins with detailing the reason to choose Sina Weibo as study object and why ten Chinese cities are selected as study area. As stated in Section 1.2, one of the research objectives of this study is to develop an analytic framework. After data collection, it is critical to extract data and transform them into a data format that can be analyzed in conjunction with the four key geographic concepts. Therefore, this chapter goes on describing the research methods and procedures of data collection, data cleaning and data conversion.

3.1 Selection of Study Object

There are different types of social media platforms. According to the contents, they can be categorized as music-focused social media, book-focused social media, game-focused social media, etc. If based on the format of posts, they can also be categorized as image-based social media, video-based social media, and audio-based (podcast) social media. In a broad sense, the most popular social media platforms can be divided into three types, including comprehensive or general social media platforms, vertical social media platforms and micro-blogging social media platforms. Comprehensive social media platforms like Facebook and WeChat provide all kinds of services for users to create and share

contents. The news and information spread on these platforms are not usually confined to a specific area. Users are connected based on their personal relationships rather than common interests or shared values. In contrast, vertical social media platforms, such as LinkedIn on which business and professional opportunities are shared, target their users to a specific group. Besides allowing users to connect and interact online like other social media platforms, the most remarkable feature of microblogging social media platforms like Twitter is the restriction of the length of each published post which fulfills the need to communicate faster (Java, Song, Finin, & Tseng, 2007).

3.1.1 Microblogging

Since Twitter was launched in 2006, microblogging service has gained worldwide popularity. Twitter co-founder Jack Dorsey initially intended to combine the cell phone carriers' Short Message Service (SMS) with social networking. At that time, one SMS could have only up to 160 characters. Twitter left 20 characters for profile information and restricted each post known as tweets to 140 characters. Even though this inconvenient length restriction is a compromise rather than a purposeful decision, it became one of the most successful product designs. On the one hand, brief text tweets force users to write concisely and straightforwardly, which saves users time and enables them to publish tweets frequently. On the other hand, short text can be spread quickly and widely. By restricting the length of posts, Twitter saves the time of

creating and processing information, which eventually speeds up the flow of information.

Another fascinating feature of microblogging service is its non-mutual relationship. Unlike some other social media platforms that require reciprocal connections (Kwak, Lee, Park, & Moon, 2010), a user can follow anyone without their permission on microblogging social media platforms. In other words, all the contents on microblogging platforms are public. It is easy for individuals to expand their online connections, obtain and share the information from a variety of sources in a short time, and at the same time maximize the speed of information dissemination.

Compared with traditional blog service (e.g., MySpace), which also establishes a unidirectional online network, connections on microblogging platforms are easier to identify (Takhteyev, Gruzd, & Wellman, 2012), establish and analyze. In addition, on reciprocal social media platforms, users may worry that their tweets are not interesting or even offend their connected friends. This concern sometimes is expressed in a decreased willingness to publish something about which they really care about or are interested. Obviously, if one does not want to offend anyone, the topics one can choose to publish online are limited. As a result, these social media platforms are not ideal study objects to investigate the similarity of people's interests. In contrast, for those using microblogging service, where users don't know each other in most cases, presenting and

expressing the true self not only fulfills the need of speaking out but also helps users build their online social networks and gain popularity. Taking into account the characteristics of microblogging social media platforms and the fact that they can provide appropriate data for this research, this study selects one of the microblogging social media platforms, Sina Weibo, as the research object.

3.1.2 Sina Weibo

Twitter would be an ideal microblogging social media platform for this study if it had enough users located in Mainland China. However, it was blocked by the Chinese government since 2009 which gave its Chinese competitor - Sina Weibo (means 'microblog' in Chinese) an opportunity to thrive. According to the financial report of the third quarter of 2017, after eight consecutive years of increase, Sina Weibo has accumulated 376 million monthly active users (MAU), surpassing 326 million MAU on Twitter (Figure 3.1).

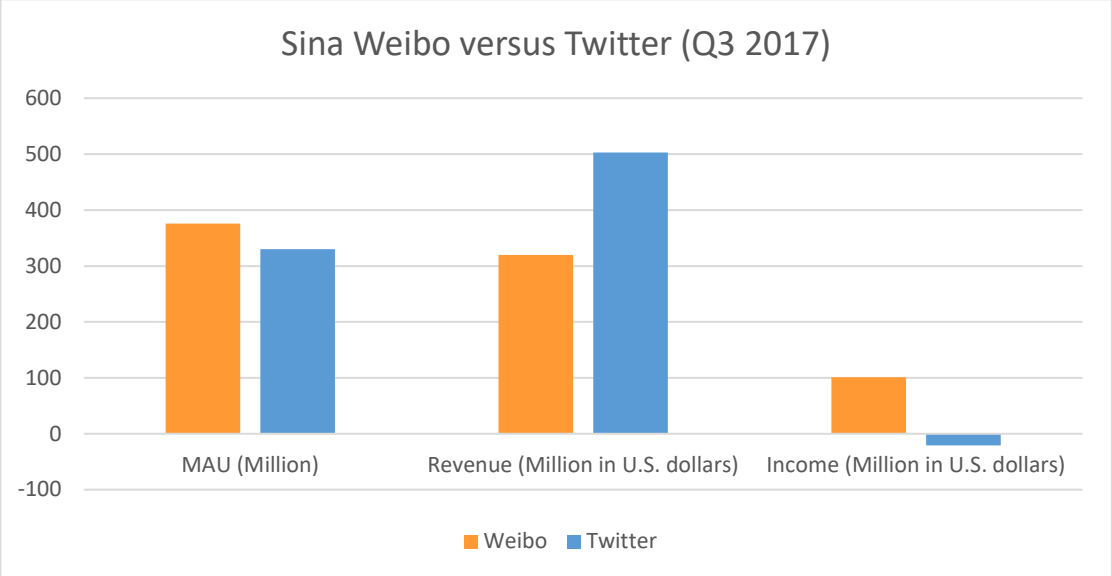

Figure 3.1 Sina Weibo versus Twitter

Weibo is so successful and popular that it has become the only survivor left in Chinese market after fierce competition against other Chinese microblogging service providers. Sina Weibo's popularity can be demonstrated by another interesting number. By the end of 2016, there is a total of 164,522 government Weibo accounts. Among them are 125,098 government institution accounts, and 39,424 public officer accounts (China Internet Network Information Center (CNNIC), 2017). It has become a common phenomenon that people use Weibo as a powerful tool to force the government to take action against the problems or injustice they cannot solve offline. On the other side, the government is willing to use Weibo to propagandize its ideology, respond to public concerns and show its governing capacity. It is not only a platform where people keep in touch with their friends and relatives but also a force to push societal changes.

Sina Weibo is not merely a replica of Twitter. Except for the length restrictions, Weibo introduced many innovative features to encourage online interactions and information dissemination, including threaded comments, live video, trends categorization, etc. A recent survey compares the top reasons of using social media between Sina Weibo and the other two popular social media platforms: WeChat Friends Circle and Momo (Figure 3.2) (China Internet Network Information Center (CNNIC), 2016). Contrasting with the other two social media platforms, which are mostly used to maintain pre-existing relationships, Weibo users are more inclined to use it to get and discuss the

latest news, acquire and share knowledge, and find interesting contents. One recent survey (China Internet Network Information Center (CNNIC), 2017)

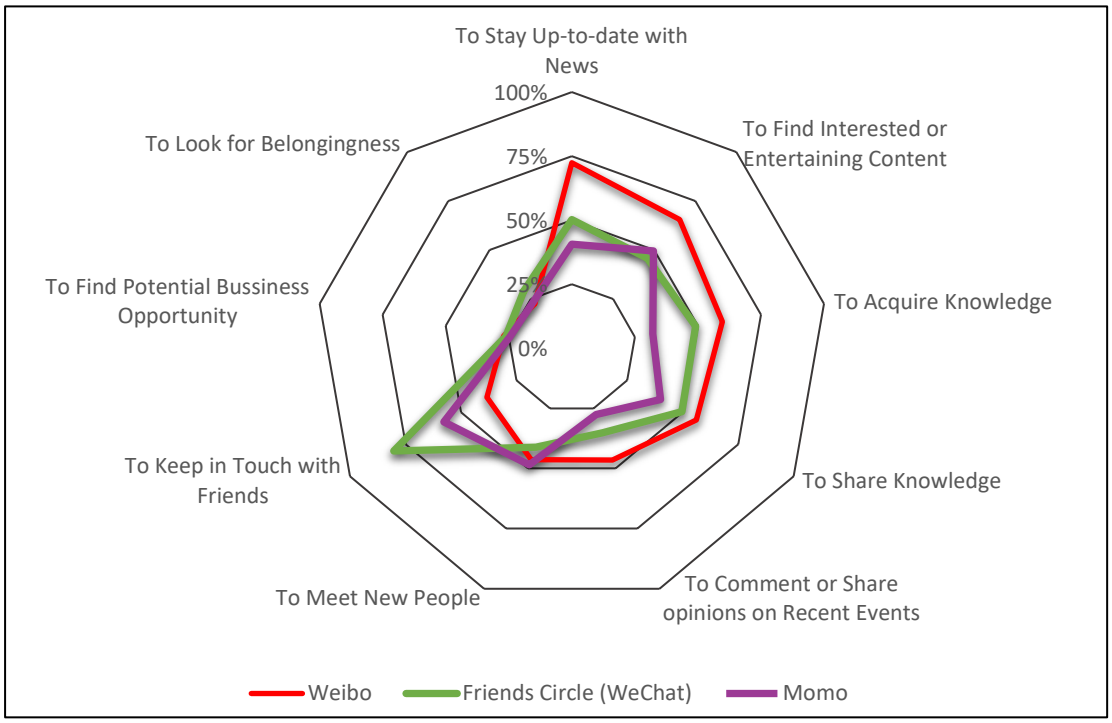

Figure 3.2 Reasons of Using Social Media

shows that Sina Weibo focuses more on establishing new connections with strangers, accounting for 43.3% of all connections. Moreover, Sina Weibo users follow more celebrities, opinion leaders, enterprise or institution accounts, which are not common for other social media platforms. The contents published on Weibo cover a broad range of topics, which makes it easier to analyze the geographical differences in social media users' interests. However, research related to Sina Weibo is far from sufficient in spite of its incredible influence on Mainland China today. Because of different culture background, online behavior (Chen, Liu, Wang, & Gu, 2012), platform design and tight control of online speech and severe censorship, a totally different

microblogging environment has been created in Mainland China, which means the related research results using Twitter or other microblogging platforms in western countries cannot be simply assumed to be the same as those based on Sina Weibo. Due to above reasons, Sina Weibo was chosen as the study object.

3.2 Study Area

Ten cities in mainland China, including Beijing, Shanghai, Guangzhou, Shenzhen, Nanjing, Suzhou, Chongqing, Hangzhou, Ningbo, Xiamen, were selected for this study (Figure 3.3).

Figure 3.3 Study Area

There are several reasons why these ten cities were selected. First of all, they have enough qualified data sample in a limited data collection time. Among a

total of 145,366 collected Weibo users, 16,458 users did not provide location information and need to be excluded. The others are dispersed in 569 cities and counties. The Weibo users' location is a long-tailed distribution. About 40% users are located in 20 cities. However, not all these users are qualified for this study. By applying a filtering mechanism, which will be elaborated in the following section, only 16,440 users are left and only 28 cities have more than 100 qualified users.

Secondly, in order to examine the influence of distance, some of these study areas need to be situated in different geographical regions, while others need to be located closely. As a result, Shanghai, Nanjing, Suzhou, Hangzhou and Ningbo are selected, since they are located in the Yangtze River Delta Region which is the most developed metropolitan area in China. These five cities are located in the east coastal area, and the longest distance between any of these two cities is less than 200 kilometers. In the southernmost part of mainland China, the Pearl River Delta, which is another metropolitan region, Guangzhou and Shenzhen were selected. The distance between them is less than 100 kilometers and to travel between them will only take half an hour by high-speed train. For Beijing and Chongqing, one is located in the northern part of China, and the other is in the southwest region. They are more than 1,000 kilometers away from the other eight cities. Xiamen is located in the southeast coastal area

between Yangtze River Delta and Pearl River Delta. The distance between Xiamen and these two metropolitan regions is less than 1,000 kilometers.

Similarly, in order to investigate the influence of urban system hierarchy on content similarity, this study also needs to include cities from at least two different tiers. The process of urbanization in mainland China has been extremely accelerated since 1978 when reforms and opening-up policies were implemented. The unprecedented growth of Chinese cities reshapes the traditional urban system hierarchy (Chen X. , 1991). Numerous studies have investigated this topic from different perspectives, including urban centrality (Zhou, Zhang, & Wu, 2001), transportation network (Zhong & Lu, 2011) (Yu, Gu, & Li, 2008), internet infrastructure (Wang & Ning, 2005), etc. There is no universally accepted classification scheme of China's urban system hierarchy. However, considering economic development level, population, city size and political administration, most researchers agree that Beijing, Shanghai, Guangzhou, and Shenzhen are the most developed cities in Mainland China. Therefore, they are categorized as the first-tier cities in this study. While the rest six cities which are categorized as second-tier cities, they are either provincial capitals or regional economic centers.

To make sure the analytical results are not biased by some outlying data, each city needs to have a sufficient number of users. Suppose ten cities are selected as study areas, and each of them has 100 qualified users, then a total of 499,500

($C_2^{1000} = \frac{1000!}{2!*(1000-2)!} = \frac{1000*999}{2} = 499{,}500$) pairwise similarity measures will be calculated. Meanwhile, if this sample size increases, not only does the number of pairwise similarity calculation increase but also the computation time for each calculation increases, because the TF-IDF matrix becomes bigger. Considering the computational capacity and the time taken for calculation, the sample size is limited to ten cities.

Admittedly, more cities as well as users might reveal greater details about social interactions. But these ten representative cities are considered to be sufficient to examine the patterns from the perspective of distance and urban system hierarchy to answer the research questions.

3.3 Database and Data Tables

MySQL is a database management system (DBMS) that is extremely popular and widely used by many programmers because of its high performance and on-demand scalability. Its free version cannot only fulfill all the requirements of data storage and queries of this study but also can facilitate access and manipulation of data in a reliable and fast manner. Therefore, MySQL was employed as the DBMS for this study.

Four relational data tables were created for this study (Table 3.1), including Account Table (Table 3.2 stores attribute data for Weibo users), Content Table (Table 3.3 stores the contents of each tweet and related attribute data),

Segmentation Table (Table 3.4 stores the results of segmentation and similarity values for each user) and Similarity Table (Table 3.5 stores pairwise similarity values between each two users).

Table 3.1 Data Tables

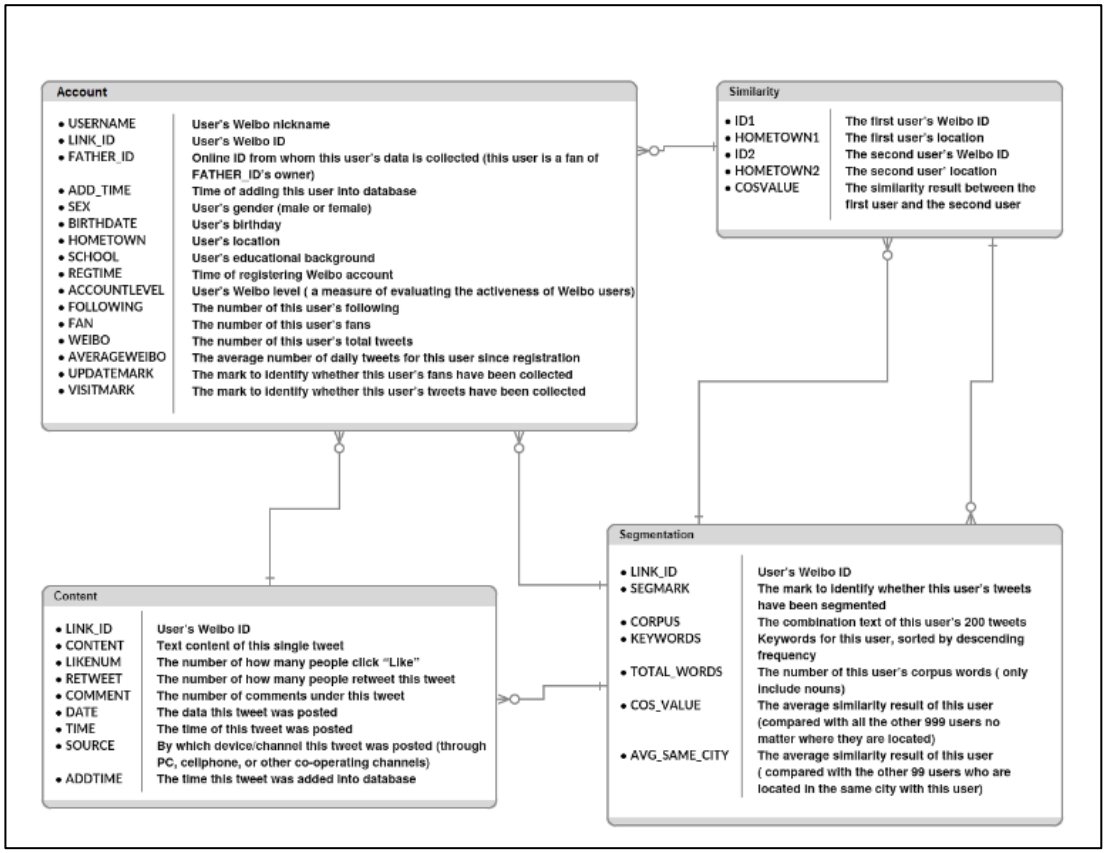

Table 3.2 Data Structure of Account Table

Attribute	Description
USERNAME	User's Weibo nickname
LINK_ID	User's Weibo ID
FATHER_ID	Online ID from whom this user's data are collected (this user is a fan of FATHER_ID's owner)
ADD_TIME	Time of adding this user into database
SEX	User's gender (male or female)
BIRTHDATE	User's birthday
HOMETOWN	User's location
SCHOOL	User's educational background
REGTIME	Time of registering Weibo account

ACCOUNTLEVEL	User's Weibo level (a measure of evaluating the activeness of Weibo users)
FOLLOWING	The number of this user's following
FAN	The number of this user's fans
WEIBO	The number of this user's total tweets
AVERAGEWEIBO	The average number of daily tweets for this user since registration
UPDATEMARK	The mark to identify whether this user's fans have been collected
VISITMARK	The mark to identify whether this user's tweets have been collected

Table 3.3 Data Structure of Content Table

Attribute	Description
USERID	User's Weibo ID
CONTENT	Content of this single tweet
LIKENUM	The number of how many people click "Like"
RETWEET	The number of how many people retweet this tweet
COMMENT	The number of comments under this tweet
DATE	The date this tweet was posted
TIME	The time of this tweet was posted
SOURCE	By which device/channel this tweet was posted (through PC, cellphone, cellphone's brand and model, other co-operating channels such as music website)
ADDTIME	The time this tweet was added into database

Table 3.4 Data Structure of Segmentation Table

Attribute	Description
LINK_ID	User's Weibo ID
SEGMARK	The mark to identify whether this user's tweets have been segmented
CORPUS	The combination of this user's 200 tweets
KEYWORDS	Keywords for this user, sorted descending
TOTAL_WORDS	The number of this user's words (only nouns)
COS_VALUE	The average similarity value of this user, compared with all the other 999 users
avg_cos_value_same_city	The average similarity result of this user, compared with the other 99 users who are located in the same city with this user

Table 3.5 Data Structure of Similarity Table

Attribute	Description
ID1	The first user's Weibo ID
HOMETOWN1	The first user's location
ID2	The second user's Weibo ID
HOMETOWN2	The second user' location
COSVALUE	The similarity result between the first user and the second user

3.4 Procedures

For this study, four program modules were developed with the Python programming language to collect and analyze data, including User Collection Module, Tweet Collection Module, Words Segmentation Module and Similarity Calculation Module (Figure 3.4).

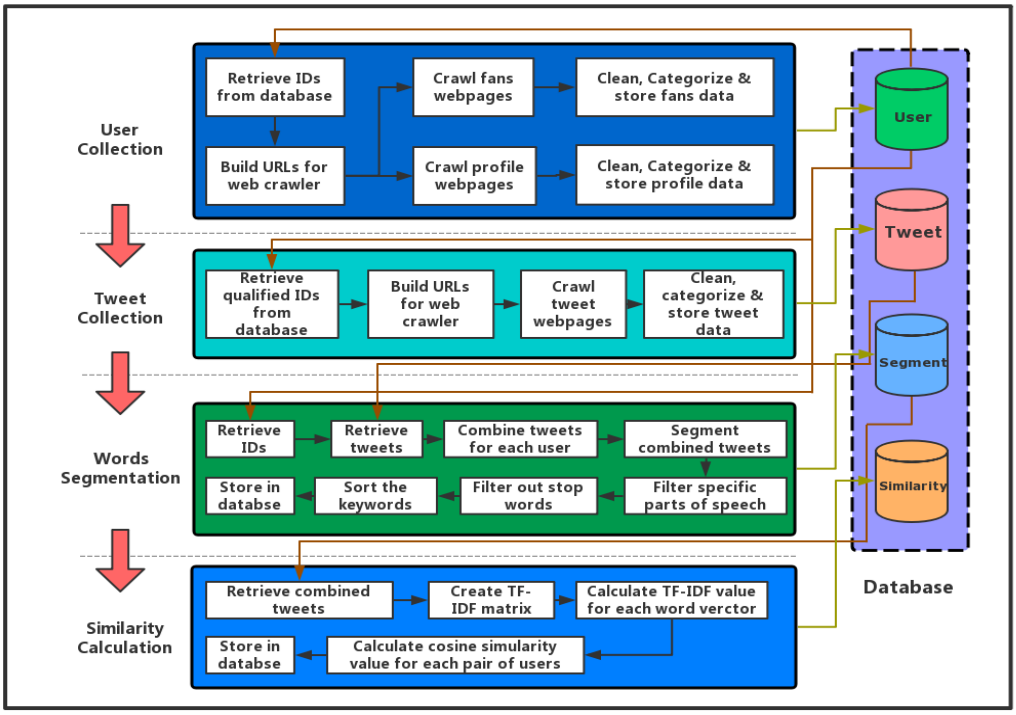

Figure 3.4 Program Modules and Flowchart

3.4.1 Users collection

3.4.1.1 API Versus Web Crawler

Almost all social networking sites provide APIs for researchers or programmers to use. APIs are a set of communication protocols that allow developers to connect their programs to service provider easily to either make use of its functions or share data. The most important advantage of APIs is that it provides a convenient and efficient way to connect. By using APIs, researchers can still successfully connect their programs with social media platforms without having to master the programming language and data structures used by specific service providers. In addition, APIs provide well-defined data structures which can significantly reduce the workload taken for subsequent data processing.

However, there are many limitations about the data that you can collect through APIs. Take Sina Weibo as an example; researchers need to go through a lengthy and complicated application process to obtain connection authorization. Besides that, if researchers aim to collect Weibo users' profiles or relationship data, authorization from each Weibo user is also required, which is impossible and unrealistic for this study. Even though all data are available through APIs, researchers still need to stick to their rate-limitations, which require waiting a certain time between two consecutive requests. In other words, the amount of data that can be collected per unit time is restricted by the service

provider, which makes the data collection process a tedious time-consuming process. Due to these drawbacks, other data collection options have to be considered.

Web crawlers or spiders refer to programs that automatically harvest web page content in large scale. They usually start from a Uniform Resource Locator (URL) address or a queue of URL addresses, then download the webpage, extract data and links which will be put into the queue. By repeating the process, web crawlers can be run automatically. For data collection, web crawlers act differently by providing much more control over time and data. The main benefit of web crawlers is that researchers can collect almost whatever they can see from a webpage. Unlike APIs, web crawlers are not bound to any specific websites. Sometimes people need data from various websites, and by some minor changes, web crawlers are flexible enough to collect data from different sources. However, web crawlers will cause a detrimental impact on websites if they send too many requests in a short time. Therefore, most websites have various anti-crawling mechanisms that need to be bypassed carefully, otherwise web crawlers will be detected and banned. How to collect a large volume of data in an acceptable time in a responsible and nondestructive manner still challenges many researchers. To make random sleep calls between requests to restrict the crawling speed is a good way to make the web crawler

look like a real human and avoid being detected by websites. This study adopted this strategy.

3.4.1.2 Regular Expression

Online data are categorized and stored in a service provider's DBMS and displayed on web pages in a readable format, which enables ordinary web users to view, understand and navigate the content without programming knowledge. However, when given a URL, the web crawler will return almost everything in an unstructured format with all Hypertext Markup Language (HTML) symbols and other unwanted data. Therefore, when data were captured, regular expressions were used to store data in order to extract and categorize data for subsequent analysis.

A regular expression is a computer science terminology, which is a search pattern to find a specific string in a block of text. It consists of operators, constructs or one or more characters to describe a string. Regular expressions are a productive and efficient tool to manage and process data. When given a text (a sequence of characters), a program can use a pattern to parse and find any matched strings. For example, a Weibo user's profile webpage consists of many different attributes about this user, such as User Name, Birthday, Gender, etc. When this page is crawled, the program receives unstructured text data where all profile attributes are mixed and cannot be automatically distinguished by the program. After analyzing this text, the pattern that

describes the gender attribute of the user can be identified. Let's say the string which includes gender attribute always starts with "Gender:". The length of this attribute is always two, and the string is either "Male" (男) or "Female" (女). By describing this pattern with regular expression syntax, all matched strings can be found and inserted into the database. Combination of multiple regular expressions is a common practice to search complex string. Most popular programming languages provide a built-in regular expression engine, even though the syntax might be slightly different. In this study, a third-party Python library called "Regex" was used due to the original built-in regular expression module lacks functionality and flexibility.

3.4.1.3 User Collection

In this study, the Snowball sampling method (Lee, Kim, & Jeong, 2006) was employed. The data collection starts from an initial user whose outgoing links are then followed to reach new users. By observation, movies are a unanimously favored topic by Weibo users. No matter whether people are from south-east metropolis or north-west inland villages, most are willing to talk about recent movies online. Therefore, an online opinion leader was selected as the initial user. This user posts tweets about his original and incisive review of recent movies, which helps him attract about 50,000 fans. Another reason to select this user is that his latest fans are distributed evenly in mainland China, which will minimize the sampling bias.

As the first module in Figure 3.4 shows, after being fed the initial user's ID, the web crawler can build an URL and start to crawl both the initial user's profile page and his fans' list pages.

User's profile data include information such as user's ID, gender, location, registration time, the number of fans, the number of followings and the number of all tweets. Educational background and birthday information can also be collected (Figure 3.5); however, these data are not mandatory when registering, thus many users have not provided them.

Figure 3.5 Example of Profile Webpage

A social media user usually follows hundreds of other users. Since most of his followings are celebrities whose online behaviors are not typical representations of ordinary social media users, this study only crawled users'

fans' data. However, due to privacy concerns, Sina Weibo only allows people to view a user's latest 100 fans' information which means a user's full network structure cannot be accessed. This restriction forces this study to focus on users' newly established online connections. When a social media user creates an online account, he will try to connect people who he already knows offline in the beginning. To a large extent, these initial connections are located in the same geographical region with the user. Gradually, they can establish some new online relationships with people who they are not acquainted offline. These new connections are built upon their common interests and similar perspectives rather than their pre-existing relationships and geographical proximity. Recursively collecting the latest 100 fans for each user not only provides a continuously growing database but also enables this study to analyze the geographical distribution of social media users' new online connections.

These profiles and fans' data were stored in the Account data table. Meanwhile, for each user, two status indicators were created in the database to indicate whether this user's fans and tweets content data have been collected or not. Then the program will check the database to find out and feed the web crawler the next ID whose profile and fans' information has not been collected. By repeating this process, the Account Collection module can collect all Weibo users' information if time allows.

3.4.1.4 User Filtration

Like all social media platforms, Sina Weibo also suffers from rampant bots. Bots (botnets, cyborgs) on social media refer to fake accounts that pretend to be a real person and are created for a specific purpose other than expressing the real opinions and feelings of a human being. Including these bots will affect the accuracy and objectivity of this study. Luckily, close observation can reveal some behavior patterns of most bots, which further helps remove them from the database. Furthermore, even though some collected Weibo accounts are used by real persons, their online behaviors cannot represent ordinary people. A typical case is celebrity accounts which usually have thousands, even millions of fans. In most cases, these accounts will only post tweets about a particular topic that is closely related to their career. An opposite example is that some users just use social media as a news platform to acquire the information in which they are interested. These people might not be willing to or good at expressing their feelings and opinions online. As a result, they barely post anything on Weibo. Even though they do represent a subset of social media users, there is no way to collect enough tweets data from them for subsequent analysis.

In this study, some filtering thresholds are set to remove bots and other noisy data. Besides locating in one of ten chosen cities, to be a qualified Weibo user for this study, the first criterion is to have more than 200 tweets and at least 50

fans. This criterion aims to remove bots that are created to be sold. Their existences are only to serve as some people's fans to appear popular and help them gain false influence (Confessor, Dance, Harris, & Hansen, 2018). This kind of bots will usually follow thousands of users while posting nothing online. Because they are just a piece of code, they do not have any existing offline friends to connect. Posting nothing or just some gibberish created automatically by a program will not help them attract fans either. As a result, posting very little content online and having a small number of fans are important characteristics for most social media bots. In addition, this criterion can also help remove accounts that barely post anything online even though they are used by real persons.

The second criterion for a qualified user is to have more than 600 followings. Just like the first criterion, this filtering mainly focuses on removing bots that are batch-created and sold to numerous Weibo users as fans. According to the theory of Dunbar number (Dunbar, 1992), an individual can only maintain stable relationships with about 150 persons. In fact, if a user follows too many users on social media, the flood of information will require hours to read and digest every day. So, a mature social media user keeps his following list from being too long in case they miss information in which they are really interested. This situation obviously does not apply to bots since they don't read anything posted by their followings. People can choose to reduce social interactions with

an estranged acquaintance in their daily life. While on social media, unfollowing somebody is often considered impolite. Therefore, the threshold of the number of followings should be bigger than 150. Kwak et al. (2010) find that the online behavior of Twitter users who have more than 1,000 followings or 1,000 fans is dramatically different.

The third criterion is to have less than 600 fans, which mainly aims to remove celebrity users, such as movie stars, famous sport players, opinion leaders, etc. These people attract huge number of fans because of their celebrity expertise, unique insights or deep understanding in a certain field. To maintain their online image and expand their influence, their tweet topics tend to focus mainly on areas they already excel at. It is fair to say that the topics they post online are not as broad as those by ordinary Weibo users. Including these celebrity users in this study will result in biased similarity calculation.

The last criterion is to remove users who post tweets with almost the same content. The number of these users' tweets, followings and fans all meet the filtration standard mentioned above, but their online behavior proves to be not a typical social media user. For example, some users are not actually active Weibo users, but they spend lots time on streaming music websites where their accounts are bound to their Sina Weibo accounts. Whenever they click "favorite" or "share" buttons on these websites, an automatically generated tweet will be posted on Sina Weibo. To remove these users, in the Words Segmentation

module, the program calculates and stores the number of total segmented words for each user and his high-frequency keywords list. Users who have less than 500 total segmented nouns and a short high-frequency keyword list, which usually only includes nouns, such as ‘Music’, ‘Video’, ‘Article’, are removed.

3.4.2 Tweets Collection

Figure 3.6 shows a tweet webpage. Like most websites, Sina Weibo constructs their URLs with certain patterns, which can be identified after observation and experiment. For example, the URL of a user’s profile webpage always starts with 'https://weibo.cn/', followed by the user’s ID, and ends with ‘/info’. So, a typical profile webpage URL looks like this:

https://weibo.cn/1764452651/info

Similarly, the URL of a user’s tweet webpage starts with ‘'https://weibo.cn/u/'', followed by the user’s ID, and ends with ‘'?page='’. By retrieving users’ IDs from the database, the program can construct the URLs of each users’ tweets webpage and feed them to the web crawler.

Two hundred tweets were crawled and stored into the database (Figure 3.7) for each qualified user using the same method mentioned in Section 4.4.1. In the end, 200,000 tweets were gathered.

Figure 3.6 Example of Tweets Webpage

PID	USERID	CONTENT	LIKENUM	RETWEET	COMMENT	DATE	TIME
632030	2255330640	转发了质馆咖啡的微博:#质馆-世界咖啡馆#【北美最小的咖啡店】这个"仅容一人居"的咖啡屋名为La Distributrice，	1	0	0	04月22日	10:34
632031	2255330640	转发了王燃phailik的微博:"不拍照，岁月会偷走我们的时间"愿有一间小小的咖啡馆，咖啡，甜点，傻傻的萨摩，	0	0	2	04月22日	10:06
632032	2255330640	转发了微博：抱歉，此微博已被作者删除。查看帮助：http://t.cn/zWSudZc赞[0]原文转发[0]原文评论[0]转发理由:	0	0	0	04月21日	10:03
632033	2255330640	转发了就爱英伦范的微博:别人的肚子别人的腿 你却还在躺着刷微博[组图共9张]原图赞[9]原文转发[62]原文评论[1]	1	0	0	04月19日	07:32
632034	2255330640	转发了重庆苏大姐火锅南京店的微博:#苏大姐火锅4月特惠，羊肉继续送 #1、关注@重庆苏大姐火锅南京店并转发	1	0	1	04月16日	18:57
632035	2255330640	又来次啦~@重庆苏大姐火锅南京店南京.升州路显示地图原图赞[1]转发[0]评论[0]收藏04月16日 18:57 来自三星 GAL	1	0	0	04月16日	18:57
632036	2255330640	转发了大爱猫咪控的微博:我就想知道，宋仲基和猫都想要的有多少人~原图赞[721]原文转发[431]原文评论[202]转	1	0	0	04月16日	11:29
632037	2255330640	转发了GIF博士的微博:其实我也不懂英语。。。太坏了！[哈哈]秒拍视频赞[39]原文转发[120]原文评论[21]转发理由	1	0	0	04月16日	11:28
632038	2255330640	转发了365道美食DIY的微博:【焗烤番茄蛋】西红柿炒鸡蛋、西红柿鸡蛋汤、西红柿鸡蛋面，我们已经吃得很多了	2	0	0	04月16日	11:26
632039	2255330640	转发了StyleLog的微博:大口大口吃冰淇淋的季节到了 [甜筒][甜筒][甜筒] [馋嘴][组图共9张]原图赞[620]原文转发[26	1	0	0	04月11日	12:07
632040	2255330640	转发了StyleLog的微博:终于要到这个季节了不是吗 [馋嘴][组图共9张]原图赞[1065]原文转发[529]原文评论[370]转发	0	0	2	04月11日	12:06
632041	2255330640	转发了无水印手机壁纸精选的微博:这才是完整版！[微笑][组图共9张]原图赞[357]原文转发[433]原文评论[65]转发理	0	0	0	04月11日	11:53
632042	2255330640	转发了全球奇闻趣事的微博:换毛季到了，养萨摩耶的同学一起来感受下！！！[笑cry]秒拍视频.赞[1062]原文转发[	1	0	0	04月07日	08:13
632043	2255330640	转发了AGM旅行手机的微博:妈妈问我为什么跪着刷微博？？因为这脑洞真不是正常人能想得出来的[doge][doge]转	1	0	2	04月06日	19:06
632044	2255330640	转发了微博搞笑排行榜的微博:据说每个人身边都有这样一个逗比朋友。哈哈哈萌贱萌贱的小黄人。[笑cry]秒拍视	0	0	0	04月06日	18:42
632045	2255330640	转发了微博：抱歉，此微博已被作者删除。查看帮助：http://t.cn/zWSudZc赞[0]原文转发[0]原文评论[0]转发理由:	1	0	0	04月03日	11:41
632046	2255330640	转发了澳帝焙咖啡Ultimate的微博:盛会也好，自嗨也罢，予人幸运，于己收获。病倒的伙伴，鸡血的再战，千"锤	1	0	0	04月02日	09:34
632047	2255330640	转发了慕素咖啡2015的微博:撒馆了，澳帝焙馆主robin开着咖啡车回家路上海.上海新国际展览中心[组图共3张]原图	0	0	0	04月02日	09:32
632048	2255330640	转发了高夫烘焙的微博:钢铁侠和蜘蛛侠也来展会了，全铜的咖啡机，机械手臂的冲煮，还有一位大哥全手工烘焙	0	0	0	04月02日	09:31
632049	2255330640	转发了快乐的咖啡伙伴的微博:[太阳][咖啡][好]『上海咖啡之旅——Base周末市集』上海咖啡之旅——Base周末市集	0	0	0	04月01日	14:52
632050	2255330640	上海展会游~蒙田咖啡车~[组图共9张]原图赞[1]转发[0]评论[0]收藏04月01日 14:52 来自三星 GALAXY S5	1	0	0	04月01日	14:52

Figure 3.7 Example of Tweets Content Data Table

3.4.3 Words Segmentation

Words segmentation is a technique that separates a sentence or a paragraph into a sequence of individual words. Words segmentation is the foundation of natural language processing (NLP). It has been used for search engines, machine translation, text mining, text similarity, keyword extraction and so on. In written English, words are separated by blank spaces and punctuations

while in Chinese, the absence of natural word delimiters (such as space) in a sentence makes the situation complicated. Although modern written Chinese has adopted punctuation symbols since the early 20th century, which simplifies the problem of segmenting a whole document into the problem of segmenting multiple single sentences, a long Chinese sentence might still contain dozens of words without delimiters. In addition, most Chinese characters can be seen as an individual word with semantic meaning. However, when combined with another character or more characters to form a new word, it can have a totally different meaning. Without delimiters, readers need to consider the context to resolve this kind of ambiguity and decide when and where to segment words. For example, '中' (Zhong) means 'middle', '学' (Xue) means 'learn'. A sentence like "他从中学的社团中学会了合作" (Ta Cong Zhongxue De Shetuan Zhong Xuehuile Hezuo) can be translated as 'He learned how to cooperate in a middle school community'. In this sentence, the first '中学' (Zhongxue) need to be segmented as a single word which means 'middle school', while the second '中学' (Zhong Xue) need to be segmented into two words as 'from' and 'learn'.

This study employed Jieba ("结巴", means stutter in English) toolkit to decompose Chinese sentences. A recent study compares most popular Chinese segmentation tools developed in the last ten years and finds that Jieba has an acceptable accuracy rate (Yang, Chen, & Zhao, 2017). Another study concludes Jieba has faster processing speed than other Chinese segmentation tools (Peng,

Cambria, & Hussain, 2017). Meanwhile, Jieba can be used across platforms by providing multi-versions for many popular programming languages including Java, C++, Python, etc. One disadvantage of Jieba is that it does not provide a human-machine interface and can only be called from programs; therefore, people who have no programming skills will find it difficult to use. Because this study has already used Python as the programming language to collect and analyze data, this disadvantage is not a problem.

When given an input text, Jieba builds a directed acyclic graph (DAG) for all possible words based on Trie Tree structure. Then it finds the most possible words segmentation based on word frequency. For words that are not listed in the dictionary, Hidden Markov Model (HMM) is used (Sun, 2012).

After segmentation, another question needs to be considered before data analysis. Take an adjective word 'happy' as an example. One person can feel happy when he gets married, while the other person can feel happy when his favorite hockey team wins the championship. Even though both of them posted some sentences on social media that all contain the adjective word 'happy', they are actually talking about different topics. Similarly, one person 'makes' a table in his garage, while the other person 'makes' a decision to move to Canada. Both of them used the verb 'make', but they were not talking about the same thing. On the contrary, if one person mentions a noun 'Donald Trump' in his post, it does not matter whether he likes, hates, agrees or makes fun of him, he

shows a certain interest in 'Donald Trump'. Therefore, this study only keeps nouns for content similarity analysis. The other parts of speech, including adjectives, adverbs, verbs, pronouns, prepositions, conjunctions, interjections, numerals, articles and determiners, are removed from the segmentation result. Fortunately, Jieba is a fabulous tool that cannot only segment a text quickly and accurately but also can identify the part of speech for each separated word.

Finally, before all these segmented nouns are stored into the database, some words need to be filtered out. In natural language processing, stop words mean the most common words that occur very frequently but have little meaning, such as 'the', 'a', 'how', 'on', etc. By removing these words, search engines or other NLP programs can focus on more important and meaningful words. There is no universally accepted stop words list. Most search engines compile their own list based on their experience and understanding. Jieba provides a simple stop words list that contains only dozens of most frequent words. After investigating the result of segmentation, other hundreds of nouns were added to this list. Some of these newly added stop words are embedded in users' tweets by Weibo platform. For example, if a user posts some pictures on Weibo, a word '照片' (Zhaopian, means photo) will be automatically added to the content of tweets. Some of them are representations of emojis which are a set of symbols or ideograms to express one's emotions, such as '笑脸' (Xiaolian, means smiley face), '香烟' (Xiangyan, means cigarette).

Filtering out stop words and non-nouns is essential as it saves huge storage space. More importantly, keeping these words in the database would make the TF-IDF matrix extremely sparse which requires much more computation time to calculate the similarity values. In addition, their presence would seriously affect other meaningful nouns' TF-IDF value, resulting in an inaccurate similarity calculation.

3.4.4 Similarity Calculation

To calculate content similarity among Weibo users, their tweets need to be converted from characters and words to numbers. The Vector Space Model (VSM) (Salton, Wong, & Yang, 1975) is a statistical model to represent text by means of vectors. VSM can simplify the processing of the text content to a vector operation in a high-dimensional vector space. A document can be abstracted as a vector, which consists of many terms. Each term has an associated weight, which indicates the importance of the term in the document. The similarity between different documents can be measured by the spatial distance of vectors. The effectiveness of a VSM relies significantly on the vector's term weighting. Among many different weighting schemes, TF-IDF is the most popular one.

3.4.4.1 TF-IDF

TF-IDF refers to Term Frequency – Inverse Document Frequency, which is a widely used information retrieval model in practical applications such as search engines and text mining. It is a statistical method used to assess the importance

of a word to one document in a set of documents or a corpus. If one word appears in a document with a high frequency and is rarely present in other documents, this word is considered to have a good ability to distinguish this document from the other documents.

Given a corpus D that contains m documents: d_1, d_2, … d_m, $n_{i,j}$ denotes the number of times of term t_i appears in document d_j. The mathematical equation to calculate TF value of term t_i in document d_j can be given as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

where $\sum_k n_{k,j}$ is the sum of number of occurrences of all terms in document d_j. The mathematical equation to calculate IDF value of term t_i in corpus D can be given as:

$$idf_{i,D} = \log \frac{|D|}{|\{d \in D: t_i \in d\}|} \quad (2)$$

where $|D|$ is the total number of documents in corpus D, and $|\{d \in D: t_i \in d\}|$ is the number of documents where term t_i appears. No matter how many times t_i occurs in a document, it is still counted as 1.

Combining above two equations, the TF-IDF value for term t_i in document d_j can be calculated as below:

$$tfidf_{i,j} = tf_{i,j} \times idf_{i,D} \quad (3)$$

As an example, suppose we have three sentences as below:

Sentence 1: I love Lethbridge.

Sentence 2: I love playing hockey.

Sentence 3: I move to Calgary from Lethbridge.

For all 9 indistinct words in these sentences (corpus), TF-IDF value for each word can be calculated (Table 3.6, Table 3.7, Table 3.8).

Table 3.6 TF Value for Each Word

Words / Sentences	I	love	Lethbridge	playing	hockey	move	to	Calgary	from
Sentence 1	0.33	0.33	0.33	0	0	0	0	0	0
Sentence 2	0.25	0.25	0	0.25	0.25	0	0	0	0
Sentence 3	0.17	0	0.17	0	0	0.17	0.17	0.17	0.17

Table 3.7 IDF Value for Each Word

Words	I	love	Lethbridge	playing	hockey	move	to	Calgary	from
IDF Value	1.00	1.18	1.18	1.48	1.48	1.48	1.48	1.48	1.48

Table 3.8 TF-IDF Value for Each Word

Words / Sentences	I	love	Lethbridge	playing	hockey	move	to	Calgary	from
Sentence 1	0.33	0.39	0.39	0.00	0.00	0.00	0.00	0.00	0.00
Sentence 2	0.25	0.29	0.00	0.37	0.37	0.00	0.00	0.00	0.00
Sentence 3	0.17	0.00	0.20	0.00	0.00	0.25	0.25	0.25	0.25

In this simple example, even though the frequency of word 'I' is the same as other words in sentence 3, since it occurs in all three sentences, its TF-IDF value is lower than the other words in sentence 3. For the word 'hockey' in sentence 2, because it appears only once in all three sentences, its TF-IDF value is higher. In summary, the TF-IDF value is directly proportional to the number of occurrences of a word in one document, and it is inversely proportional to the number of occurrences of the word in the corpus.

As Figure 3.4 shows, all 200 tweets for each qualified user are combined into a single document for words segmentation. Then the Similarity Calculation module builds a TF-IDF matrix, which has 1,000 rows and each row represents a qualified Weibo user from the study sample. This matrix also has hundreds of thousands of columns, each representing a word that appears in the corpus. It contains all TF-IDF values for each distinctive word that was used at least once by one Weibo user. With this matrix, users' similarity value can be calculated.

3.4.4.2 Cosine Similarity

The TF-IDF algorithm offers a means to present a user's tweets as a vector in a high-dimensional space whose direction is determined by its TF-IDF values. This study uses a cosine similarity algorithm to score the similarity of users.

Cosine similarity is a widely used similarity metric in information retrieval, text classification, text mining and so on (Han, Kamber, & Pei, 2011). It measures the cosine of the angle between two vectors. The problem to calculate two documents' similarity is converted to comparing the distance between the two corresponding vectors in high dimensional vector space.

There are many different measures to calculate similarity, such as Euclidean distance, City-block distance, Mahalanobis distance, Hamming distance, etc. (Xu & Wunsch, 2005). One advantage of cosine similarity is that it can ignore the magnitude of different vectors, which means the difference of documents'

length does not matter. It only focuses on how closely two vectors are oriented in space. Besides that, cosine similarity also delivers more accurate results for high-dimensional sparse data (Shirkhorshidi, Aghabozorgi, & Wah, 2015).

Given two vectors: A = [a_1, a_2, …, a_m], and B = [b_1, b_2, …, b_m] where a_i and b_i are term t_i's TF-IDF value in A and B, m is the number of dimensions of A and B. Then the cosine similarity value between A and B is:

$$\mathrm{sim}(A, B) = \mathrm{cosine}(\theta) = \frac{\sum_{i=1}^{m} a_i b_i}{\sqrt{\sum_{i=1}^{m} a_i^2}\sqrt{\sum_{i=1}^{m} b_i^2}} \quad (4)$$

From this equation it can be deduced that, in the extreme case, if there is not an identical word in documents A and B, the similarity value between them is zero. In contrast, if A and B are completely duplicated which means a_i and b_i are always the same, the maximal similarity value is one.

Each user's 200 tweets were combined into one document, so the 200,000 tweets were converted to one corpus that contains 1,000 documents. In the Similarity Calculation module, total 499,500 pairwise similarity values are computed for all 1,000 users from 10 study cities. Combining these values with users' location attributes, the average similarity values between cities can be delivered by using Standard Query Language (SQL) statements in database. However, an additional filtering process needs to be done to remove noisy data that cannot be filtered out in Section 3.4.1.4.

After running the program, each user's average similarity value with the other 999 users was calculated. Then all these users were sorted by their average similarity value. For users with the lowest average similarity value, their high-frequency keyword lists and tweets' contents were manually checked in order to remove any advertising accounts. These users only posted tweets to promote particular products or services. It is hard to say if these users are bots or real persons, but their online behaviors are not typical representations of ordinary social media users. Among these users, some mostly posted tweets about their favorite singers or beloved sports teams. Even though they also have lower average similarity value since their tweets are concentrated in a fairly narrow range of topics, they are kept as qualified users because their tweets did reflect their interests. In the end, after reviewing 300 users with lowest average similarity value, around 140 advertising accounts were removed and replaced with qualified accounts for better data quality. For the remaining 700 users, their tweets were randomly checked to ensure that these was no omission.

3.5 Summary

Unlike traditional methods of obtaining data through interviews and surveys, using computer science technology to collect online data, though it is efficient and cost-effective, is still a novelty in geography. This chapter systematically introduces the entire process from data collection, data cleaning, data

conversion to data analysis. Other researchers can make use of the methods presented in this chapter in their related studies without starting from scratch or reinventing the wheel. Meanwhile, using TF-IDF and Cosine Similarity algorithms to transform the textual data into mathematical model, without being constrained by semantics, is also promising and realistic. This innovative analytical framework enables people to deepen their understanding of contemporary society and human being themselves.

CHAPTER FOUR

4. Results and Discussions

In this study, the collected data were analyzed to answer the research questions from two perspectives: geographical distribution of new online connections and geographical difference of the content similarity.

4.1 Borders still matter

Thomas Friedman (2005) claims in his bestselling book, The World Is Flat, that ten forces, such as the information and communication technology (ICT), have "flattened" the earth, which makes traditional national borders less important. However, due to many reasons including trade barriers, cultural identity, and transportation cost, borders still impede the flow of information, goods, services, even capital. This study finds that the borders of cities and provinces still matter on Sina Weibo. The world is not flat, not even on the internet.

4.1.1 Borders still matter in establishing new online connections

In the Data Collection module, a total of 3,583 Weibo users' fans data were collected. Among them, 790 users are from one of ten selected study cities. The geographical distribution of their latest 100 fans is shown in Table 4.1.

Table 4.1 Fans Distribution

Fans' Location / Users' Location	Bei-jing	Guang-zhou	Shen-zhen	Shang-hai	Nan-jing	Su-zhou	Hang-zhou	Ning-bo	Chong-qing	Xia-men	Total
Beijing	1214	278	235	402	106	97	134	47	161	55	2729
Guangzhou	334	358	163	183	38	49	66	27	99	40	1357
Shenzhen	172	102	127	92	35	26	40	14	50	30	688
Shanghai	457	199	146	827	73	64	115	52	116	40	2089
Nanjing	78	26	25	70	189	16	13	2	21	13	453
Suzhou	95	46	31	116	15	168	13	5	23	5	517
Hangzhou	156	55	50	116	31	26	333	28	39	19	853
Ningbo	35	25	18	41	10	4	17	32	15	4	201
Chongqing	123	57	43	75	21	17	23	13	295	11	678
Xiamen	56	28	24	46	6	7	19	2	24	67	279

Except for Shenzhen and Ningbo, the users from the other eight cities have the most fans coming from their own city. Shenzhen is a special case since the city was just a small remote fishing village 30 years ago and was completely built up by immigrants. Take Shanghai as an example; in the database, Shanghai users have 7,907 fans in total. Among them, 827 users are from the same city. Although the number of fans from Beijing is the second highest, the figure is just about half that from Shanghai.

When the scale of comparison is enlarged to provincial level (Figure 4.1), a similar pattern is revealed. From all ten selected cities, the largest fans base is always in the provinces where the cities are located. Except for Xiamen, all of these percentages are bigger than 10%. Because of the special political system of mainland China, cities within the same province tend to be more closely connected, as reflected in the intense intra-province social and economic interactions (Liu, Sui, Kang, & Gao, 2014). In other words, the provincial

administrative borders have a negative impact on connections between cities in disparate provinces. The results show that borders of cities and provinces, as one geographic concept, still play a vital role in establishing new connections online.

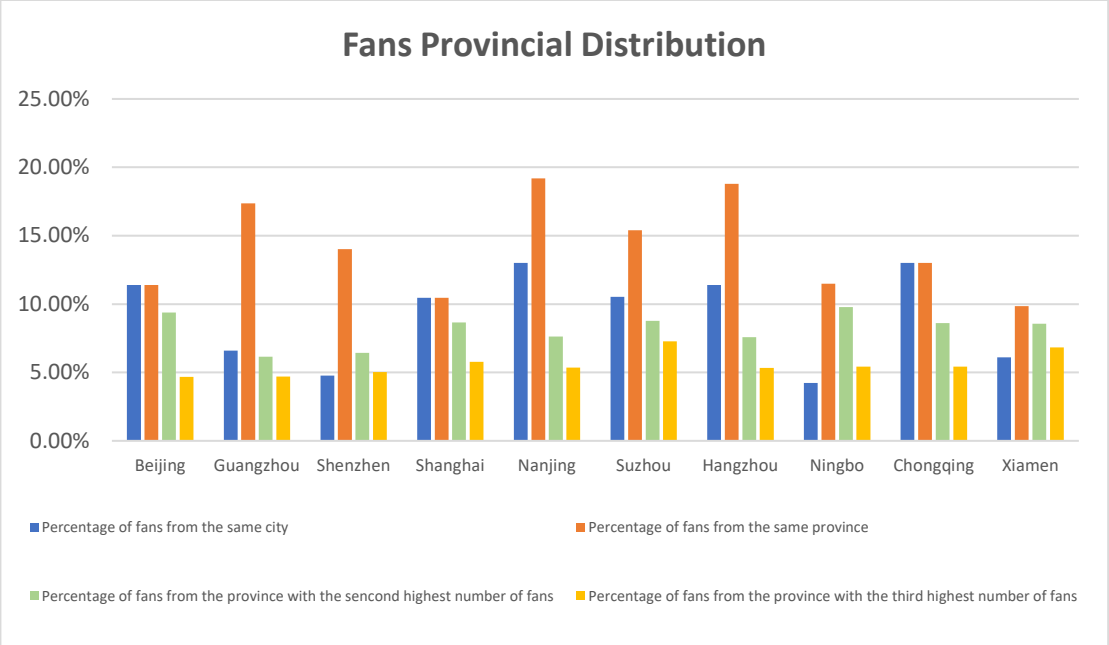

Figure 4.1 Fans Provincial Distribution

4.1.2 Borders still matter in content similarity

For each city, as Section 3.4.4 describes, a total of 4,950 pairwise similarity values were calculated between all its 100 qualified Weibo users. Using the same method, for every two different cities, 10,000 pairwise similarity values were calculated between their 100 respective users. Then all average similarity values were calculated and plotted in Figure 4.2. Each city is shown in a different color. Every ten bars represent the ten average similarity values for each city. The bars that represent the comparison within the same city are displayed in red diagonal stripes.

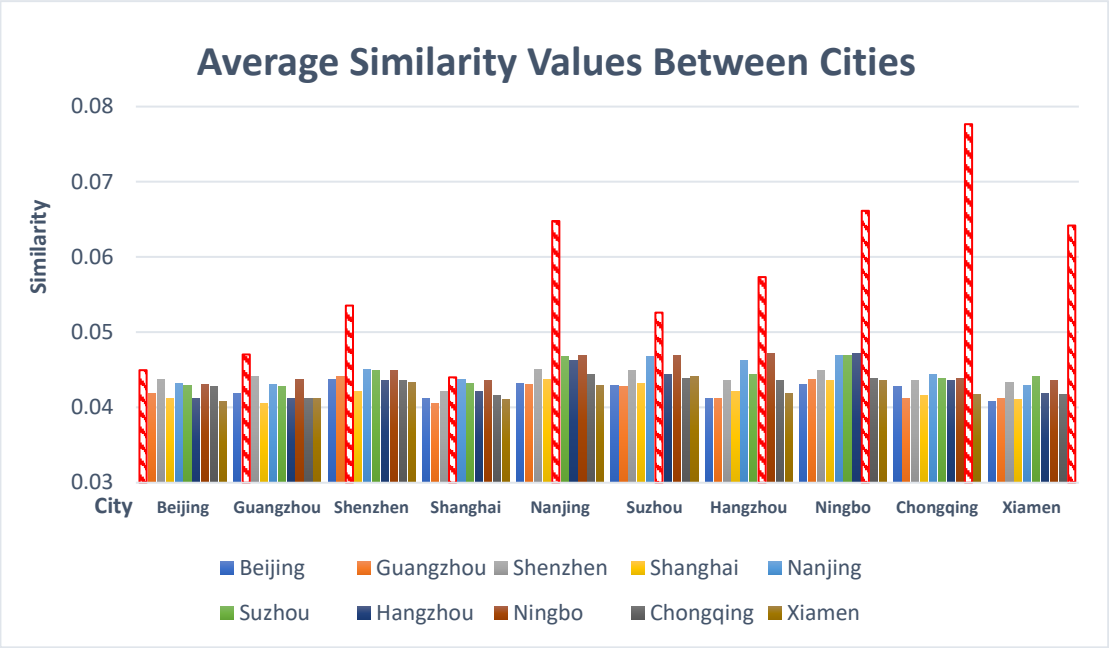

Figure 4.2 Average Similarity Values Between Cities

Obviously, for any city, the average similarity values of the users in the same city are always higher than those between different cities. For instance, when users in Chongqing are compared with users from other cities, the average similarity values are all smaller than 0.045. However, when users in Chongqing are compared with themselves, this value is 0.078, almost twice the second highest similarity value.

High similarity value means that some words that are rarely present in other users' tweets, are somehow concurrent frequently in the two compared users' tweets. These words are representations of users' interests and topical preferences. The results indicate that in the same city, residents have more common interests or attentions on some unique topics. Once beyond the borders of the city, these topics may become irrelevant to the residents of other

cities, or the intensity with which these topics are mentioned is diluted by other topics. That uniqueness of the city-based tweets goes beyond the city borders. For example, Figure 4.3 is a word cloud graph that was plotted with the top 100 high-frequency words in the corpus. The most popular topics always revolve around family, place, entertainment, etc. Then another two graphs, which excluded nationwide frequent words were created for Chongqing (Figure 4.4) and Shanghai (Figure 4.5) users to reveal the regional preference. As Figure 4.4 shows, Chongqing users' favorite topics concentrate on food, music and sports while for Shanghai users, the topics cover a broader range, which eventually leads to a low similarity value within the city.

Borders are not just some lines drawn on a map. They define the spaces where people spend most of their lives. Within the boundaries of a city, unique culture and habits are slowly taking shape. Although information technology has, to a certain extent, led to a more diverse culture, the unique characteristics of a city have not been completely eliminated.

4.2 Distance still matters

Even though the importance of distance in modern society has been diminished, this study finds that distance still plays a vital role in establishing new online ties and shaping similar topical preference.

Figure 4.3 Word Cloud of All Cities

Figure 4.4 Word Cloud of Shanghai

Figure 4.5 Word Cloud of Chongqing

4.2.1 Distance still matters in establishing new online connections

In traditional social interactions, due to geographical proximity and strong cultural and social recognition, people usually build up their social connections within the surrounding living and working groups. Even though the spatial distance that used to be a barrier to build and maintain social connections is compressed in modern digital age, the possibility of establishing connections with people close to them is still far greater than that of people who are far away.

This regularity is evident when provincial distribution of fans of each city is plotted in Figures 4.6 to 4.9. Taking Xiamen as an example (Figure 4.9), aside from the province where the city is located-Fujian, the neighboring provinces, such as Zhejiang and Guangdong, are also fan-intensive areas. Only a few fans come from remote provinces, such as Tibet and Inner Mongolia. The effect of distance decay is quite evident and can be found on all other cities.

In the traditional distance decay gravity model, besides distance, population, size, social-economic development all can affect the attractions and repulsions between two regions. They work together to influence the strength of spatial interactions. Like Nanjing in Figure 4.8, although it is clear there is high concentration of fans distribution in its neighboring provinces, it also has close connections with Guangdong Province and Beijing, which are distant but with a high level of economic development. Similarly, Chongqing (Figure 4.7) has more connections with the eastern coastal provinces than with the neighboring provinces.

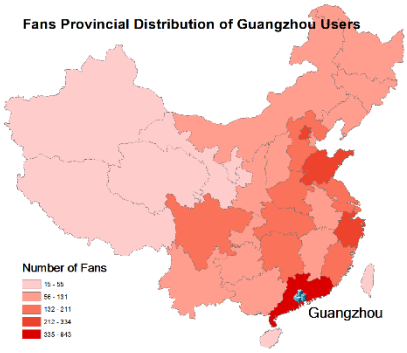

Figure 4.6 Fans Provincial Distribution of Guangzhou Users

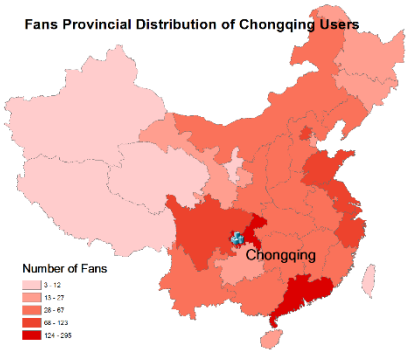

Figure 4.7 Fans Provincial Distribution of Chongqing Users

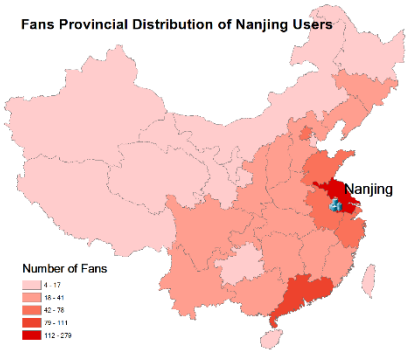

Figure 4.8 Fans Provincial Distribution of Nanjing Users

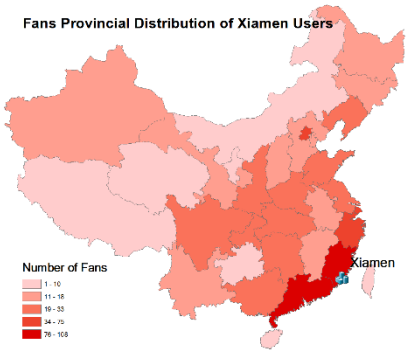

Figure 4.9 Fans Provincial Distribution of Xiamen Users

4.2.2 Distance still matters in content similarity

Then, how does the distance between cities affect the similarity values? By using ArcGIS, all ten cities are plotted on a map of China (Figure 4.10). Each line that connects two cities represents the average similarity value between them. ArcGIS split up all the average similarity values into five continuous classes with Jenks natural breaks method, which can maximize the difference between classes (Jenks, 1967). The effect of distance on content similarity is

clearly shown on this figure. As the distance between cities increases, the similarity value decreases. Take Chongqing and Xiamen as examples, not a single red line connects them with the other cities because they are far away from the others.

The content similarity changes drastically when the Yangtze River Delta is examined (Figure 4.11). All five cities in this region are relatively close to each other, and almost all the lines between them are either very high (red line) or high (purple line). Like in establishing new online connections, many factors contribute to the content similarity. However, it is fair to say that, in general, the physical distance still has a great influence on affecting people's content similarity. The closer the social media users are, the more similar their interests and topical preferences are. In the Yangtze River Delta, Shanghai is an exception, which will be explained in following section.

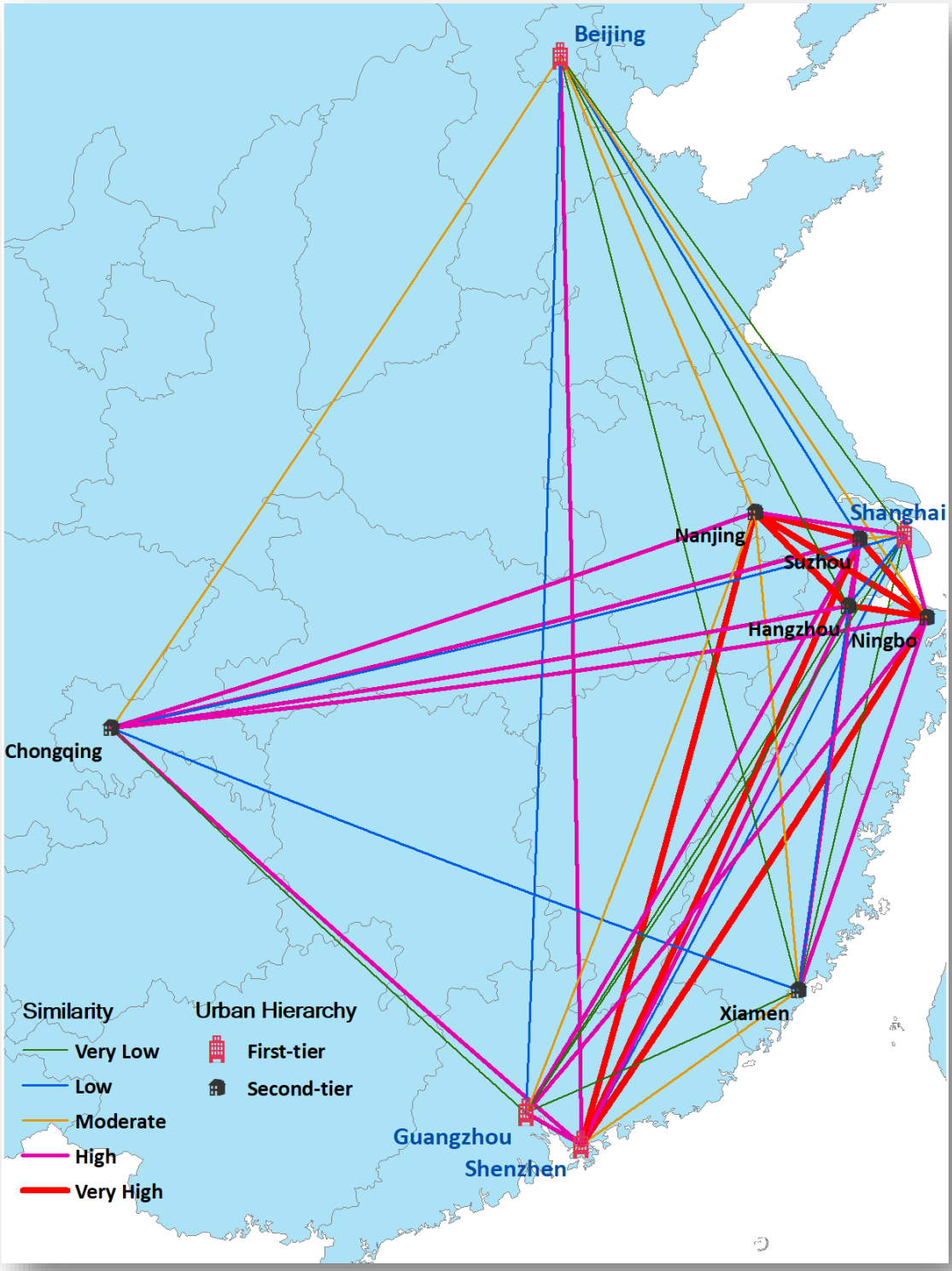

Figure 4.10 Distance and Similarity Value

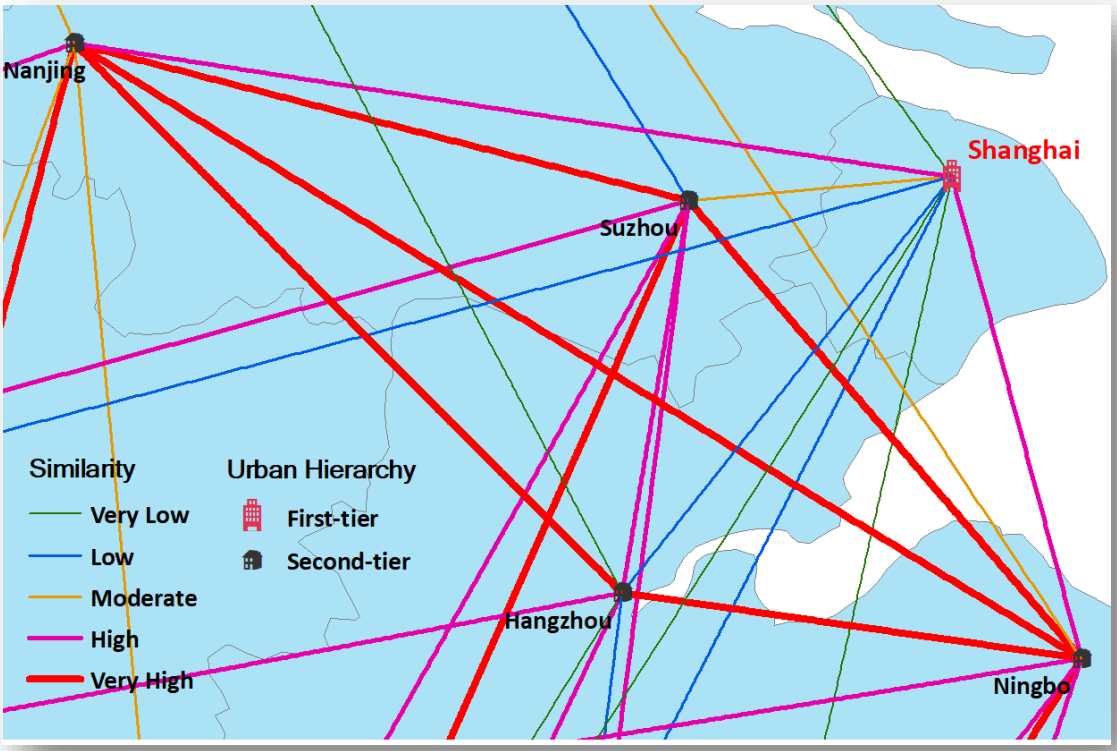

Figure 4.11 Distance and Similarity Value in Yangtze River Delta

4.3 Places still matter

As Figure 4.3 shows, many high-frequency words occurred in users' tweets are cities' names, streets' names and districts' names. A closer look reveals that the place-related nouns are an important part of the high-frequency words list in each city. Do the topics closely related to these place names lead to the high content similarity values within cities and the low similarity values between cities?

One fascinating feature of Jieba Segmentation Tool is that it can distinguish not only the different parts of speech of words, it can also categorize nouns into common nouns, person names, place nouns (including city names, country

names, streets, districts, etc.), organization names, time nouns and so on. Aside from keeping only nouns, all place-related words were removed from the corpus. Then another similarity analysis was conducted using the same method.

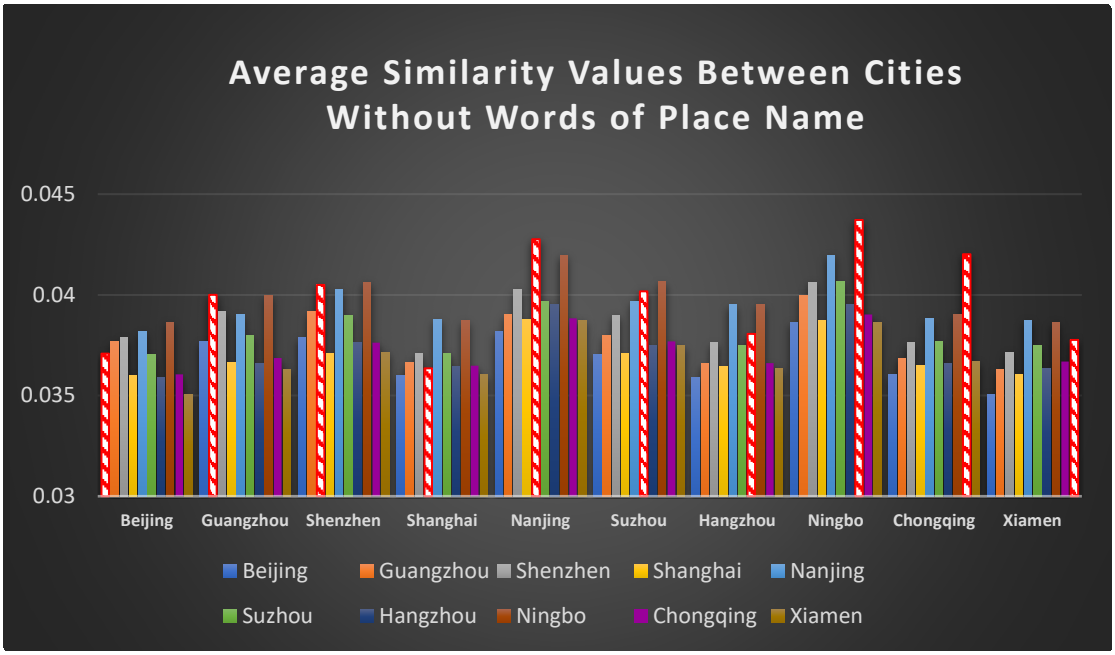

Figure 4.12 Average Similarity Values Between Cities Without Words of Place Name

The results (Figure 4.12) show that almost all similarity values have become smaller. Compared to Figure 4.2 where none of these comparisons is smaller than 0.04, now in Figure 4.12, only a dozen of similarity values are bigger than that. For Nanjing, Chongqing and Ningbo, the similarity values between the same city are still higher than those between different cities, but the gap is not that obvious. Take Chongqing as an example, the same city similarity value now is 0.042, 46% lower than the original value. And for the other cities, this gap is completely gone which means there is no distinct difference in content similarity for users located in the same city or different cities.

Places intersect with people's everyday lives in many ways. Numerous place-related topics, such as local criminal incidents, symbolic urban landscape, and acclaimed restaurants, play key roles in tightly binding people who live in the same city. On the one hand, these topics are closely related with people's daily lives; on the other hand, discussing these topics provides people with a sense of belonging and identity, especially for social media users who may not know each other offline. When these place-related topics are removed from the analysis along with the place-related words, the topics that people in a city like talking about on social media become less focused, and the subjects of common concern become rarer.

4.4 Urban System Hierarchy still matters

Compared with other cities, first-tier cities in Mainland China have experienced a longer period of economic reform and more exposure to information, innovation and multi-culture. Similarly, because of their vibrant economic and cultural activities, people living in the first-tier cities have more opportunities to establish connections with people from all over the country. They are not only exposed to the latest technologies and information, but also faced with more competition and challenges. Reflected in social media, between the first-tier cities and the second-tier cities, both the formation of new social connections and the preference of topics show different patterns.

4.4.1 Urban System Hierarchy still matters in establishing new online connections

For each city, the top twenty cities with largest number of fans were selected for analysis. The results show (Figure 4.13), for users in the first-tier cities, their fans are distributed more evenly nationwide. Taking Guangzhou as an example, about 41% of its new fans come from the top twenty cities. On the contrary, about 51% of Suzhou and Ningbo's fans come from the top twenty cities. The pattern is still observable when the number of selected top cities is downsized to 10 or upsized to 50.

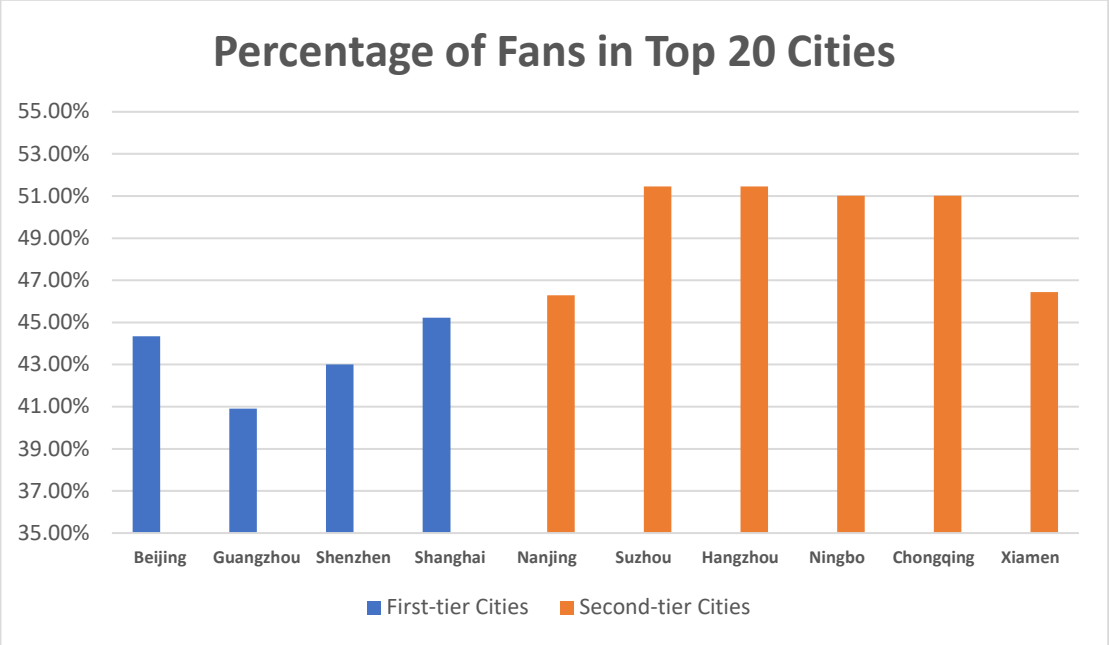

Figure 4.13 Percentage of Fans from Top 20 Cities

For users in the first-tier cities, their fans are not densely concentrated in a few economically developed or neighboring cities. As the national economic, cultural and political centers, the first-tier cities themselves attract people from all over the country, which means the proportion of migrant workers in these

cities is much higher than that in other cities. At the same time, large public events that occur in the first-tier cities can easily become national hot topics. The additional visibility and publicity inevitably lead to more online interactions. To some extent, the second-tier cities also have such characteristics, but their influence may be more pronounced in local areas.

4.4.2 Urban System Hierarchy still matters in content similarity

As Figure 4.8 shows, even though all five selected study cities in the Yangtze River Delta are close to each other, not a single red line connects Shanghai with the other four cities. Actually, there is no red line connecting Shanghai with any other selected cities. The same pattern can be found on Beijing too. Of course, one can argue that this is because Beijing is far away from the other cities. But for the other two first-tier cities, Guangzhou and Shenzhen, a similar pattern is also observed. Like Shanghai and Beijing, there is no red line connecting Guangzhou to other cities. The similarity value between Shenzhen and Guangzhou is not high either, even though they are in the same province and pretty close together. This reveals another characteristic of first-tier cities: neither the similarity value in the same first-tier city nor the similarity value between different cities is high. The situation in Shenzhen is a little special, as Section 4.1.1 explains, the city was built by immigrants from across the country. Their pre-existing connections in their hometowns work together with their

earlier experience to shape their online interests and attentions despite the fact that they have migrated to Shenzhen.

As stated in Section 4.3, all similarity values become smaller after removing place-related words. But for the first-tier cities, this decline is less dramatic than that of the second-tier cities (Figure 4.13). Except for Shenzhen, the three other first-tier cities' similarity values are reduced by less than 20% in contrast to about 35% or more for the second-tier cities. The users located in the first-tier cities are talking less about place-related topics. Originally, their interests are broader and more diverse, so when these words are removed, their same city similarity values are less affected.

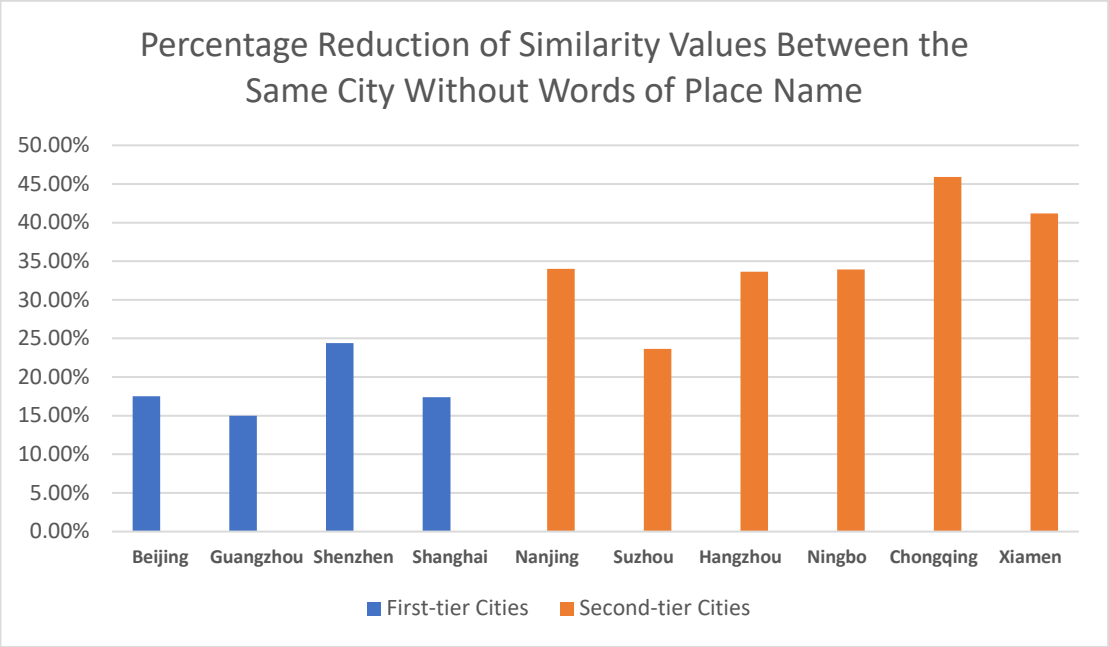

Figure 4.14 Percentage Reduction of Similarity Values

As Figure 4.5 shows, the hot topics that Shanghai users discuss in social media include economics (e.g., 'Company', 'Money', 'Resource'), education (e.g., 'School', 'University', 'IELTS'), culture (e.g., 'Book', 'Literature', 'History'), and

politics (e.g., ‘Policy’, ‘Official’). Most of these topics are less subject to geographical difference.

4.5 Summary

As a response to the second research objective, this chapter reveals the importance of geography in the era of social media by respectively investigating the role of the four key geographic concepts on both formation of new online connections and content similarity. Even without the limitation of physical distance, people still tend to establish new connections with those who are geographically close to them. Meanwhile, the culture and customs specific to each city do not dissipate with the popularity of social media.

CHAPTER FIVE

5. Conclusions

5.1 Discussion

In this study, a web crawler was developed by the Python programming language to collect profile, fans and tweets data of Sina Weibo users in ten selected Chinese cities. After applying the filtering mechanism, TF-IDF and cosine similarity algorithms were used to calculate pairwise similarity values in Weibo content between every two users. Then this study delves into the interplay of new social media connections, contents similarity and the four key geographic concepts of borders, distance, places, and urban system hierarchy. The results show that all these geographic concepts are still playing an important role in the formation of new online connections and shaping people's interests. Social media has not declared the end of geography.

This study finds that municipal and provincial borders have a profound influence on online socialization. In the literature, some researchers tried to demonstrate the importance of borders from a variety of different perspectives. For example, Liu et al. (2014) investigate the human mobility pattern by extracting check-in data from online social media platforms and finds that borders act as barriers for intra-urban and intra-province movements and communication. Eisenstein et al. (2010) use social media data to analyze the

geographic linguistic variation and reveal that certain words display strong regional affiliates. Borders are still important in distinguishing and grouping people. They not only reflect the existing difference but also reinforce the sense of "otherness" (Houtum & Naerssen, 2002). This study provides new empirical evidence about the influence of borders by focusing on online connections and contents similarity. The result shows that the virtual world, like the physical world, has not become "borderless".

The role of distance in the modern society is also investigated from many perspectives. For example, geographical proximity among firms is proved to remain important in increasing the opportunities for face-to-face communication, which is crucial for developing business relationships and acquiring market information (Agnes, 2009) (Porter, 1998). Some studies find that people would like to live closer to their important connections (Axhausen K. , 2000). Other studies focus on the relationship between distance and the frequency of telephone contact and find that the contact frequency declines as distance increases (Wellman, 1979) (Moka, Wellman, & Basu, 2007) (Lambiotte, et al., 2008). People tend to establish and keep in touch with people near to them. Using social media data, some studies find that the probability of friendship is inversely proportional to distance (Liben-Nowell, Novak, Kumar, Raghavan, & Tomkins, 2005) (Backstrom, Sun, & Marlow, 2010) (Takhteyev, Gruzd, &

Wellman, 2012) (Scellato, Noulas, & Mascolo, 2011) (Kwak, Lee, Park, & Moon, 2010).

This conclusion is validated by analyzing the provincial distribution of the fans of social media users in this study. Besides, this study provides a new perspective to demonstrate the importance of distance by investigating the relationship between proximity and contents similarity. The results show that the nearer people live, the more similar their social media contents are. The effect of distance decay has not disappeared in the social media era. Tobler's first law of geography is still useful to guide geographic research (Miller, 2004). Indeed, social media has changed people's daily lives, including the experience and perception of place (Adriana & Frith, 2010). The significance of places is reinforced by social media rather than reduced according to literature. Hjorth (2012) finds that social media help users keep contact with their kinship and facilitate a sense of place. Another study on Flicker shows that social media users who took pictures of the same place at approximately the same time are likely to know each other (Crandall, et al., 2010).

In general, the literature on the role of places is primarily qualitative, focusing on exploring new interpretations of places in the social media era. This study uses a quantitative approach to compare the contents similarity by keeping and removing place-related words and finds that places remain essential in shaping

the interests of social media users. People still frequently and meaningfully interact with places.

This study also finds that the unique culture and habits that specialize in one city become less visible when removing the place-related words from the analysis. As Poorthuis el al. (2010) state: "Information always has a geography. It is created in places, it is used in places, it is changed and repurposed in places" (p. 249). However, due to the limitation of the data granularity, this study cannot investigate the relationship between place and friendship which need to be explored in future studies.

First-tier cities, which are the largest cities in China's urban system, have a greater range of economic activities, more employment opportunities, more diverse culture and far superior access to information than lower-tier cities. These cities in turn attract people from all over the country and exert a huge influence on the behavior of their residents across China's urban system hierarchy. It can be imagined that social media users who live in the first-tier cities therefore would also exhibit different online activities patterns from users in the cities which are lower in urban hierarchy.

The new rising social media has not weakened the influence of urban hierarchy. Instead, it provides a new source of data to examine the urban hierarchy. Zook (2001) finds that global cities still play a critical role in generating internet contents and argues that the decentralized nature of the internet needs more

research and proof. Fekete (2014) uses FourSquare data to demonstrate that the consumption pattern in the United State has not changed on social media and claims the urban hierarchy is still persistent in the southeastern U.S. Wang and Zhen (2016) draws a similar conclusion by investigating search engine data. Dou et al. (2006) compare people's traditional media consumption between Guangzhou and Xi'an and find people in the first-tier cities prefer information-based topics rather than entertainment-based contents. This study identifies similar patterns and also finds that users in the first-tier cities have much broader range of interests than their counterparts in low-tier cities when comparing their similarity values.

The results also demonstrate that, compared with people in the second-tier cities, the new online connections for people who live in the first-tier cities are more evenly distributed nationally. Most existing literature focuses on developing new urban ranking scheme by using social media data, completely ignoring the study of the relationship between social media user behavior and urban hierarchies. This study fills this gap and shows that social media data can help to reveal and understand the differences among cities.

In sum, to the best knowledge of the author, this is the first study that systematically examines the role of geography in Chinese social media platforms. The results show that borders, distance, places and urban system

hierarchy still reflect real and significant difference among social media users. Social media has not rendered geography irrelevant.

5.2 Limitation

Although promising findings have been presented in this study, several limitations need to be taken into consideration for future studies. Firstly, this study assumes the location information that is presented on a user's profile webpage is real. However, due to the privacy concerns or other reasons, the authenticity of Weibo users' location information is not guaranteed. Sometimes people will migrate from one city to another city without changing their Weibo profile. Secondly, one of the drawbacks of the TF-IDF algorithm is that it completely ignores the semantics of words. For example, "L.A." can mean "Los Angles" or "Lethbridge, Alberta". But using TF-IDF, they are deemed as the same word. On the contrary, people may use "Calgary", "C-town" or "Cowtown" to address the city of Calgary, but the algorithm will calculate three distinct TF-IDF values for them, which may lead to low similarity values. Thirdly, this study has not restricted the time span of collected tweets. For each qualified user, the collected 200 tweets might be published within a month or span up to a year. In extreme cases, a heavy social media user might post hundreds of tweets in just a week, all on a hot topic, such as the Olympics, while another user might explore many topics in his 200 tweets over the course of a

year, but the similarity value between the two is low due to the length of the time span, which could be inconsistent with the facts. Finally, the demographic composition of Weibo users is not fully representative of the entire population. Therefore, the generalized pattern might be biased towards online population. Social media is still a novelty that is predominantly used by young adults, and this study, is able to capture that segment of the population. Given these possible limitations and biases, there is a need to explore other methods for social media data search, manipulation and analysis and incorporate them in future studies.

5.3 Summary

Social media has broken through the restrictions in space and time and has extended social interactions from physical space to virtual cyberspace, bringing with it a whole new social structure. Massive easily accessible data serve as a gold mine for geographers to explore. This study is a new attempt to investigate the role of geography in the social media era.

Besides contributing new empirical evidence to the long-standing debate about whether geography still matters, the web crawler developed for this study can be utilized by researchers who also intend to use Sina Weibo as their study object or data source but has limited programming skills or are not willing to be restricted by its official APIs. The data filtering mechanism in this study,

though simple and direct, is based on empirical observations and also can be referenced to clean noise data. Also, this study demonstrates the effectiveness of using similarity measure to detect and represent user interests and propose a novel perspective to reveal its regional characteristics.

For policymakers, the methods and results of this study can help pinpoint major social issues, identify people's concern and provide real feedback on new policies. For business owners, especially for small business owners who do not have enough budget for comprehensive market research and analysis, this study can help reveal consumers' regional preferences, which could lead to more effective advertisements and marketing. For ordinary social media users, this study can be useful to find like-minded friends or online communities, discover new places and fit into a new city seamlessly.

The world is undergoing a geospatial revolution. The uniqueness and popularity of social media provide a great amount of quality data to help researchers better understand the world and human beings themselves. This study reassesses and proves the importance of four geographic concepts – borders, places, distance and urban system hierarchy, in the modern digital age. The information technology and globalization has not declared the end of geography, and it is safe to say that Geography still matters.

References

Adriana, D. S., & Frith, J. (2010). Locative Mobile Social Networks: Mapping Communication and Location in Urban Spaces. *Mobilities, 5*(4), 485-505.

Agnes, P. (2009). The "End of Geography" in Financial Services? Local Embeddedness and Territorialization in the Interest Rate Swaps Industry. *Economic geography, 76*(4), 347-366.

Agnew, J. A., & Duncan, J. S. (1989). *The Power of Place: Bringing Together Geographical and Sociological Imaginations.* Unwin Hyman Publishers.

Albert, M., Jacobson, D., & Lapid, Y. (2001). *Identities, Borders, Orders: Rethinking International Relations Theory.* U of Minnesota Press.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2012). Effects of User Similarity in Social Media. *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 703-712). ACM.

Axhausen, K. (2000). Geographies of somewhere: a review of urban literature. *Urban Studies, 37*(10), 1849-1864.

Axhausen, K. W., & Gärling, T. (1994). Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport Reviews, 12*(4), 323-341.

Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. *WWW '10 Proceedings of the 19th international conference on World wide web* (pp. 61-70). ACM.

Beaverstock, J. V., Smith, R. G., & Taylor, P. J. (1999). A roster of world cities. *Cities, 16*(6), 445-458.

Becker, R. A., Caceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., & Volinsky, C. (2011). A Tale of One City: Using Cellular Network Data for Urban Planning. *IEEE Pervasive Computing, 10*(4), 18 - 26.

Bellotti, E. (2015). *Qualitative Networks: Mixed methods in sociological research.*

Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*, 167-271.

Bhattacharyya, P., Garg, A., & Wu, S. F. (2011). Analysis of user keyword similarity in online social networks. *Social Network Analysis and Mining, 1*(3), 143-158.

Blij, H. d. (2007). *Why Geography Matters: Three Challenges Facing America: Climate Change, the Rise of China, and Global Terrorism.* Oxford University Press.

Borchert, J. R. (1967). American Metropolitan Evolution. *American Geographical Society, 57*(3), 301-332.

Boyd, D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210-230.

Bradner, E., & Mark, G. (2002). Why distance matters: effects on cooperation, persuasion and deception. *Proceedings of the 2002 ACM conference on Computer supported cooperative work* (pp. 226-235). ACM.

Burt, R. (1992). *Structural Holes: The Social Structure of Competition.* Harvard University Press.

Butts, C. T. (2002). Predictability of Large-scale Spatially Embedded Networks. In R. Breiger, K. Carley, & a. P. Pattison, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers* (pp. 313-321).

Cai, Y., Cong, G., Jia, X., Liu, H., He, J., Lu, J., & Du, X. (2009). Efficient Algorithm for Computing Link-based Similarity in Real World Networks. *2009 Ninth IEEE International Conference on Data Mining* (pp. 734-739). IEEE.

Cairncross, F. (1997). *The Death of Distance: How the Communications Revolution Will Change Our Lives.*

Caney, S. (2006). *Justice beyond borders: A global political theory.* Oxford University Press.

Castells, M. (1989). *The informational city: information technology, economic restructuring and urban-regional process.*

Centonze, A. L. (1989). Quasi-economic locational determinants of large foreign headquarters: The case of New York City. *Economic Development Quarterly, 3*(1), 46-51.

Chandra, S., Khan, L., & Muhaya, F. B. (2011). Estimating Twitter User Location Using Social Interactions--A Content Based Approach. *2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing* (pp. 838-843). IEEE.

Chen, J., Geyer, W., Dugan, C., Muller, M., & Guy, I. (2009). Make new friends, but keep the old: recommending people on social networking sites. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 201-210). ACM.

Chen, X. (1991). China's city hierarchy, urban policy and spatial development in the 1980s. *Urban Studies, 28*(3), 341-367.

Chen, Z., Liu, P., Wang, X., & Gu, a. Y. (2012). Follow whom? Chinese users have different choice. *arXiv preprint arXiv:1212.0167*.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. *CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759-768). ACM.

China Internet Network Information Center (CNNIC). (2016). *User Behavior Study of 2015 China Social Networking Applications.*

China Internet Network Information Center (CNNIC). (2017). *The 39th Survey of China Statistical Report on Internet Development .*

China Internet Network Information Center (CNNIC). (2017). *User Behavior Study of 2016 China Social Networking Applications.*

Christaller, W. (1966). Central Places in Southern Germany.

Confessor, N., Dance, G. J., Harris, R., & Hansen, M. (2018, Janurary 27). The Follower Factory. *The New York TImes*. Retrieved from https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html

Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., & Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, (pp. 22436-22441).

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., & Sur, S. (2008). Feedback Effects between Similarity and Social Influence. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* , (pp. 160-168).

Cresswell, T. (2014). *Place: an introduction.* John Wiley & Sons.

Daly, A. J. (2010). *Social Network Theory and Educational Change.*

Dekker, R., & Engbersen, G. (2014). How social media transform migrant networks and facilitate migration. *Global Networks, 14*(4), 401-418.

Derudder, B., Taylor, P., Ni, P., Vos, A. D., Hoyler, M., Hanssens, H., . . . Yang, X. (2010). Pathways of Change: Shifting Connectivities in the World City Network, 2000—08. *Urban Studies, 47*(9), 1861-1877.

Dou, W., Wang, G., & Zhou, N. (2006). Generational and regional differences in media consumption patterns of Chinese generation X consumers. *Journal of Advertising, 35*(2), 101-110.

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2014). *Social Media Update 2014.* PewResearch Center.

Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution, 22*(6), 469-493.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A Latent Variable Model for Geographic Lexical. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1277-1287). Association for Computational Linguistics.

Ellison, N. B. (2007). Social Network Sites: Definition, History,and Scholarship. *Computer-Mediated Communication, 13*(1), 210-230.

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Computer-Mediated Communication, 12*(4), 1143–1168.

Eltantawy, N., & Wiest, J. B. (2011). Social Media in the Egyptian Revolution:Reconsidering Resource Mobilization Theory. *International Journal of Communication, 5*, 1207-1224.

Fekete, E. (2014). Consumption and the Urban Hierarchy in the Southeastern United States. *southeastern geographer, 54*(3), 249-269.

Findler, N. V., & Leeuwen, J. V. (1979). A Family of Similarity Measures Between Two Strings. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(1), 116-118.

Freeman, L. C. (2004). *The development of social network analysis.* Empirical Press.

Friedman, K. (1996). Restructuring the city: Thoughts on urban patterns in the information society. *The Swedish Institute for Future Studies.* The Swedish Institute for Future Studies.

Friedman, L. W., & Hershey, H. F. (2013). Using Social Media Technologies to Enhance Online Learning. *Journal of Educators Online, 10*.

Friedman, T. L. (2005). *The world is flat: A brief history of the twenty-first century.* Macmillan.

G Salton, A. W. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Gallusser, W. A. (1994). Political boundaries and coexistence. *IGU-Symposium Basle/Switzerland*, (pp. 24-27).

Ganesan, P., Garcia-Molina, H., & Widom, J. (2003). Exploiting Hierarchical Domain Structure to Compute Similarity. *ACM Transactions on Information Systems (TOIS), 21*(1), 64-93.

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779–782.

Grabowicz, P. A., Ramasco, J. J., Moro, E., Pujol, J. M., & Eguiluz, V. M. (2012). Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media. *PLoS ONE, 7*(1). doi:https://doi.org/10.1371/journal.pone.0029358

Graham, S. (1998). The end of geography or the explosion of place? Conceptualizing space, place and information technology. *Progress in Human Geography, 22*(2), 165-185.

Granovetter, M. (1977). The strength of weak ties. *American Journal of Sociology*, 1360–1380.

Granovetter, M. (2005). The Impact of Social Structure on Economic Outcomes. *The Journal of Economic Perspectives*, 33-50.

Guy, I., Jacovi, M., Perer, A., Ronen, I., & Uziel, E. (2010). Same places, same things, same people? Mining user similarity on social media. *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 41-50). ACM.

Haklay, M. (2010). How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and planning B: Planning and design, 37*(4), 682-703.

Hamm, M. P., Chisholm, A., Shulhan, J., Milne, A., Scott, S. D., Given, L. M., & Hartling, L. (2013). Social media use among patients and caregivers: a scoping review. *BMJ Open, 3*(5).

Han, B., Cook, P., & Baldwin, T. (2014). Text-Based Twitter User Geolocation Prediction. *Artificial Intelligence Research, 49*, 451-500.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques.* Morgan Kaufmann.

Harvey, D. (1989). *The condition of postmodernity: an enquiry into the origins of cultural change.*

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from

Online Social Media. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, a. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science, 41*(3).

Hepworth, M. (1991). *Information technology and the global restructuring of capital markets. Collapsing Space and Time: Geographic Aspects of Communications and Information*. HarperCollins.

Hjorth, L. (2012). Still mobile: A case study on mobility, home and being away in Shanghai. In R. Wilken, & G. Goggin, *Mobile Technology and Place* (pp. 140-156).

Houtum, H. V., & Naerssen, T. V. (2002). Bordering, Ordering and Othering. *Tijdschrift voor economische en sociale geografie, 93*(2), 125-136.

Huang, R., & Yip, N.-m. (2012). Internet and Activism in Urban China: A Case Study of Protests in Xiamen and Panyu. *Journal of Comparative Asian Development*, 201-223.

Huang, Y., Yang, C.-G., Baek, H., & Lee, S.-G. (2016). Revisiting media selection in the digital era: adoption and usage. *Service Business, 10*(1), 239-260.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56-65). ACM.

Jeh, G., & Widom, J. (2002). SimRank: a measure of structural-context similarity. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 538-543). ACM.

Jenks, G. F. (1967). The data model concept in statistical mapping. *International yearbook of cartography, 7*, pp. 186-190.

Jiang, B., Yin, J., & Zhao, S. (2009). Characterizing the human mobility pattern in a large street network. *Physical Review E, 80*(2).

Kadushin, C. (2012). *Understanding Social Networks: Theories, Concepts, and Findings.* Oxford University Press.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *Journal of the American Documentation, 14*(1), 10-25.

Khondker, H. H. (2011). Role of the New Media in the Arab Spring. *Globalizations, 8*(5), 675-679.

Kitchin, R. (1997). Social transformations through spatial transformations: from 'geospace' to 'cyberspaces'. In J. Behar, *Sociological studies of telecommunications, computerization and cyberspace,* (pp. 149-174).

Know, P. L., & Marston, S. A. (2002). *Places and Regions In Global Context: Huamn Geography.* Pearson Education INC.

Komito, L. (2011). Social media and migration: Virtual community 2.0. *Journal of the American Society for Information Science and Technology, 62*(6), 1075-1086.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web* (pp. 591-600). ACM.

Lambiotte, R., Blondel, V. D., Kerchove, C. d., Huens, E., Prieur, C., Smoreda, Z., & Doorena, P. V. (2008). Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications, 387*(21), 5317-5325.

Lee, S. H., Kim, P.-J., & Jeong, H. (2006). Statistical properties of sampled networks. *Physical Review, 73*(1).

Leinhardt, S. (2013). *Social networks: A developing paradigm.* Elsevier.

Liang, Y., Caverlee, J., & Mander, J. (2013). Text vs. Images: On the Viability of Social Media to Assess Earthquake Damage. *WWW 2013 Companion Proceedings of the 22nd international conference on World Wide Web companion*, (pp. 1003-1006).

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America, 102*, pp. 11623-11628.

Liu, Y., Kang, C., Gao, S., Xiao, Y., & Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Geographical Systems, 14*(4), 463–483.

Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014). Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data. *PloS one, 9*(1). doi:https://doi.org/10.1371/journal.pone.0086026

Longueville, B. D., Smith, R. S., & Luraschi, G. (2009). "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. *LBSN '09 Proceedings of the 2009 International Workshop on Location Based Social Networks*, (pp. 73-80).

Lu, L., & Huang, R. (2012). Urban hierarchy of innovation capability and inter-city linkages of knowledge in post-reform China. *Chinese Geographical Science, 22*(5), 602-616.

Marin, A., & Wellman, B. (2011). Social Network Analysis: An Introduction. In J. Scott, & P. J. Carrington, *THe SAGE Handbook of Social Network Analysis* (pp. 12-25).

Marriott, T. C., & Buchanan, T. (2014). The true self online: Personality correlates of preference for self-expression online, and observer ratings of personality online and offline. *Computers in Human Behavior, 32*, 171-177.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). *Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies.* US Department of Education.

Menczer, F. (2004). Combining Link and Content Analysis to Estimate Semantic Similarity. *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, (pp. 452-453).

Meyrowitz, J. (1986). *No sense of place: The impact of electronic media on social behavior.* Oxford University Press.

Milgram, S. (1967). The small world problem. *Psychology today*, 60-67.

Miller, H. J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers, 94*(2), 284-289.

Mitchell, W. J. (1996). *City of bits: space, place, and the infobahn.* MIT Press.

Mizzaro, S., Pavan, M., & Scagnetto, I. (2015). Content-Based Similarity of Twitter Users. *European Conference on Information Retrieval* (pp. 507-512). Springer, Cham.

Moka, D., Wellman, B., & Basu, R. (2007). Did distance matter before the Internet? Interpersonal contact and support in the 1970s. *Social Networks, 29*(3), 430-461.

Neal, Z. P. (2011). From Central Places to Network Bases: A Transition in the U.S. Urban Hierarchy, 1900–2000. *City & Community, 10*(1), 49-75.

Newman, D. (2006). The lines that continue to separate us: borders in our `borderless' world. *Progress in Human Geography, 30*(2), 143-161.

O'Brien, R. (1992). *Global financial integration: The end of geography.* Royal Institute of International Affairs.

Ohmae, K. (1991). *The Borderless World Power and Starategy In The Interlinked Enkonomy.* HarperCoIlins Publishers.

Olson, G. M., & Olson, J. S. (2000). Distance matters. *Human-Computer Interaction, 15*(2), 139-178.

Pan, Y., Li, D.-H., Liu, J.-G., & Liang, J.-Z. (2010). Detecting community structure in complex networks via node similarity. *Physica A: Statistical Mechanics and its Applications, 389*(14), 2849-2857.

Papachristos, A. (2009). Murder by Structure: Dominance Relations and the Social Structure of Gang Homicide. *American Journal of Sociology*, 74-128.

Partridge, M. D., Rickman, D. S., Ali, K., & Olfert, M. R. (2008). Lost in space: population growth in the American hinterlands and small cities. *Economic Geography,, 8*(6), 727–757.

Peng, H., Cambria, E., & Hussain, A. (2017). A Review of Sentiment Analysis Research in Chinese Language. *Cognitive Computation, 9*(4), 423-435.

Poorthuis, A., Zook, M., Graham, M., Shelton, T., & Stephens, M. (2010). Using Geotagged Digital Social Datain Geographic Research. In N. Clifford, M. Cope, T. Gillespie, & S. French, *Key methods in geography* (pp. 248-269).

Porter, M. E. (1998). Clusters and the new economics of competition. *Harvard Business Review, 76*(6), 77-99.

Rawashdeh, A., & Ralescu, A. L. (2015). Similarity Measure for Social Networks-A Brief Survey. *Proceedings of Modern AI and Cognitive Science Conference (MAICS)*, (pp. 153-159).

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to Recommender Systems Handbook. In *Recommender Systems Handbook* (pp. 1-35). springer.

Robins, K., & Hepworth, M. (1988). Electronic spaces: new technologies and the future of cities. *Futures*, 155-176.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of The ACM, 18*(11), 613-620.

Scellato, S., Noulas, A., & Mascolo, C. (2011). Exploiting place features in link prediction on location-based social networks. *the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1046-1054). ACM.

Schwartz, M. F., & Wood, D. C. (1993). Discovering shared interests. *Communications of the ACM - Special issue on internetworking* , 78-89.

Scott, A. J. (2010). Globalization and the Rise of City-regions. *European Planning Studies, 29*(7), 813-826.

Scott, J., & Carrington, P. J. (2011). *The SAGE Handbook of Social Network Analysis.* SAGE Publications.

Seamon, D. (1979). *A Geography of the Lifeworld. 1979.*

Shim, D. C., & Eom, T. H. (2008). E-Government and Anti-Corruption: Empirical Analysis of International Data. *International Journal of Public*(31), 298-316.

Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. *PLOS ONE, 10*(12).

Small, H. (1973). CoCitation in the scientific literature: a new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology, 24*(4), 265-269.

Stiglitz, J. E. (2007). *Making Globalization Work.* WW Norton & Company.

Stollberg, B., & Groeve, T. d. (2012). The use of social media within the global disaster alert and coordination system (GDACS). *WWW '12 Companion Proceedings of the 21st international conference companion on World Wide Web*, (pp. 703-706).

Sun, J. (2012). *Jieba Segmentation*. Retrieved from GitHub: https://github.com/fxsjy/jieba

Sutton, J., Palen, L., & Shklovski, I. (2008). Backchannels on the Front Lines: Emergent Uses of Social Media in the 2007 Southern California Wildfires. *Proceedings of the 5th International ISCRAM Conference.*

Taaffe, E. J. (1962). The urban hierarchy: An air passenger definition. *Economic geography, 38*(1), 1-14.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, 73-81.

The Merriam-Webster Dictionary. (2016). Merriam-Webster Inc.

Thiemann, C., Theis, F., Grady, D., Brune, R., & Brockmann, D. (2010). The Structure of Borders in a Small World. *Plos One, 5*(11). doi:http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0015422

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography, 46*, 234-240.

Tomas, D. (1991). Old rituals for new space: rites de passage and William Gibson's cultural model o f cyberspace. In M. Benedikt, *In Cyberspace: First steps* (pp. 31-48).

Townsend, A. M. (2001). Network Cities and the Global Structure of the Internet. *American Behavioral Scientist , 44*(10), 1697-1716.

Tuan, Y.-F. (1979). Space and place: humanistic perspective. *Philosophy in geography*, 387-427.

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The Anatomy of the Facebook Social Graph. *arXiv preprint arXiv:1111.4503*.

Wang, B., & Zhen, F. (2016). China's City Hierarchy under Internet and Its Influencing Mechanism: An Empirical Analysis Based on Baidu Search. *ECONOMIC GEOGRAPHY, 36*(1), 46-52.

Wang, M., & Ning, Y. (2005). The Internet and the Rise of Information Network Cities in China. *ACTA Geographica Sinica, 59*(3), 446-454.

Wang, Y. (2014, Aprial 3). *Meet China's Protest Archivist-Powered by the web, a former migrant worker is connecting local unrest to international audiences.* Retrieved from Foreign Policy: http://foreignpolicy.com/2014/04/03/meet-chinas-protest-archivist/?wp_login_redirect=0

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge University Press.

Watts, D., & Strogatz, S. (1998). Collective dynamics of "small world" networks. *Nature*, 440-442.

Wellman, B. (1979). The community question: the intimate networks of East Yorkers. *American Journal of Sociology, 84*(5), 1201-1231.

Wellman, B. (1997). Structural analysis: From method and metaphor to theory and substance. *Contemporary Studies in Sociology, 15*, 19-61.

Winsborough, H. H. (1960). Occupational composition and the urban hierarchy. *American Sociological Review, 25*(6), 894-897.

Xu, R., & Wunsch, D. C. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645-678.

Yang, H. F., Chen, M. L., & Zhao, Z. (2017). Analysis on Applicability of Common Chinese Word Segmentation Software in Literature Study of Traditional Chinese Medicine Text. *DEStech Transactions on Computer Science and Engineering, (cst).*

Yin, J., Lampert, A., & Cameron, M. (2012). Using Social Media to Enhance Emergency Situation Awareness. *IEEE Intelligent Systems, 27*(6), 52-59.

Yoon, S.-H., Kim, J.-S., Ha, J., Kim, S.-W., Ryu, M., & Choi3, H.-J. (2014). Link-Based Similarity Measures Using Reachability Vectors. *The Scientific World Journal*.

Yu, T., Gu, C., & Li, Z. (2008). China's urban systems in terms of air passenger and cargo flows since 1995. *Geographical Research, 27*(6), 1407-1418.

Zhao, P., Han, J., & Sun, Y. (2009). P-Rank: a comprehensive structural similarity measure over information networks. *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 553-562). ACM.

Zhong, Y., & Lu, Y. (2011). Hierarchical structure and distribution pattern of Chinese urban system based on railway network. *Geographical Research, 30*(5), 785-794.

Zhou, T., Lü, L., & Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B, 71*(4), 623–630.

Zhou, Y., Zhang, L., & Wu, Y. (2001). Study of China's Urban Centrality Hierarchy. *Areal Research and Development, 20*(4), 1-5.

Zook, M. A. (2001). Old Hierarchies or New Networks of Centrality? The Global Geography of the Internet Content Market. *American Behavioral Scientist, 44*(10), 1679-1696.