TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

MASTERARBEIT

Spatial-temporal Analysis of International Connections Based on Textual Social Media Data

Ausgeführt am Department für
Geodäsie und Geoinformation
der Technischen Universität Wien

unter der Anleitung von
Univ.Prof. Mag.rer.nat. Dr.rer.nat. Georg Gartner, TU Wien
und
Dr.-Ing. Linfang Ding, TU München

durch

Kaidi Guo

Kaisermühlenstraße 14, 1220 Wien

15th October 2017

_____
Unterschrift (Student)

MASTER'S THESIS

Spatial-temporal Analysis of International Connections Based on Textual Social Media Data

Conducted at The Department of
Geodesy and Geoinformation
Technical University Vienna

under the supervision of
Univ.Prof. Mag.rer.nat. Dr.rer.nat. Georg Gartner, TU Wien
and
Dr.-Ing. Linfang Ding, TU München

by

Kaidi Guo

Kaisermühlenstrße 14, 1220 Wien

15th October 2017

_____
Signature (Student)

Acknowledgments

First, I would like to thank my supervisor Prof. Dr. Georg Gartner for the delightful support he gave when I first conceived this topic and all the feedback on the thesis draft.

I would also like to thank Dr.-Ing. Linfang Ding, for her support and guidance throughout the process of thesis, especially the timely encouragement and help when I facing hard time

My gratitude also goes to all the friends who helped me during this two years study, and all the professors, lecturers involved in this Erasmus Cartography Programme. It has been a real honour getting to know you all.

Abstract

The development of social media has improved public discussion online. The proliferation of social media data provides easy access to a huge amount of user-generated content. Geotagged microblogs have been investigated in a variety of research fields for understanding spatial temporal pattern in subjects like human mobility and characteristics of a city. Besides geotagged microblogs, textual content can also be used to harvest geographical information. However, there is little research on this topic. Therefore, in this study, we aim to demonstrate how we can illustrate the connections between a virtual community and other states and regions by analysing microblogs.

To fulfill the research aim, we present the workflow of how to define a virtual community and harvest geographical information from microblog feeds. Then we discuss the visualization techniques we need to display such information. In order to do so, Sina Weibo was picked as a data source to demonstrate the international connections between China and rest of the world.

Through visual analysis, we find out that even in Cyberspace, the physical world still has an impact, and it can be proved that states geographically closer to China have a higher level of attention (stronger data flow). In the meantime, we can also notice the barrier of information exchange caused by real space distance has less influence in Cyberspace, a virtual community can have strong bound with faraway country. But generally speaking, the level of attention of a state is related to its economy power. And there is a different topic preference between male and female user. This study may provide a new research aspect for analysis human thought, the spread of news, international relations, cross-culture communication based on textual geoinformation contained in social media data.

Table of Contents

List of Figures

1. Introduction

With the development of social media and the growing popularity of smartphones, extracting and interpolation information from social media data has become a popular topic in a variety of domains. Extraction of information from user-generated data is no longer a fresh topic. Combined with GIS techniques, various research related to society has been carried out. The GPS function in smartphone makes it possible for nowadays user to add a location tag when they post something online, those geotagged records have become the main data source in GIS domain. However, the increasing demand of location data has a clear conflict with users concern about private. Since user-generated natural language data takes a great role in these platforms, it is useful for us to think whether we can also extract geographic information from textual data.

Traditionally, when we talk about the spatial-temporal analysis of social media data, it always related to topics like human mobility patterns, point of interest detection, important event detection and the mode of transmission. However, social media has its own value as a reflection of real world including the invisible connection and influence between states and communities. The geographical information extracted from social media data, therefore, could offer a new possibility to depict the connection between virtue communities.

This research project will respond to those questions, collecting social media data with mappable textual geographical information, with the help of visual data analysis to reveal interesting spatial patterns about the relation between virtue communities.

1.1 Purpose and Motivation

When we talk about social media data analysis, the study of public opinion and trend detection are often important points. The fact is, getting useful information from social media data has been considered to be one of the cheapest and fastest strategies for product innovation and service improvement for years. At the same time, we also cannot omit the mappable geographical information contained in social media data. Every post online has a

record of the timestamp indicating when exactly it was created, and users have well defined profiles containing data such as name, age and location (Mathioudakis and Koudas 2010). Such information could be extracted and used in different fields. In the past few years, enlightening research has been conducted. In cartography field, geographical information is of interest. Traditionally, when social media data and GIS are combined, research directions are more concentrated on analysis geo-tagged posts, from which visualization systems have been built to dig up unusual events and trend (Piller, Vossen, Ihl et al. 2011). However, the current progress achieved mainly through geotagged data, further research into harvesting geoinformation from the text is in need throughout wider topics. As a reflection of real space, new research ideas among GIS and Cartography domain is stimulated by those social media users along with their location information and textual geographical information they generated. The work of Anthony et al (2013) and Ming et al. (2012) have demonstrated the expansion of research topics involving textual information.

Although previous studies have revealed spatial patterns, there is potential for new aspects of textual geoinformation. As referred in a work conducted by Kalev et al. (2013), the text could be used to expand the mappable information. Every day, textual geoinformation is generated and shared through internet, for instance: another superhero movie from “American”; a special street snake in “Korea”; a new LGBT policy announced in “Germany”; a stunning tragedy happened in “British”. By looking into a certain group of the user, the amount of the mappable information could indicate the relation and connection between virtue communities, and when users see, share, or post, these actions, in turn, shape the data flow and the border of their community. Today, the globalization becomes faster and more natural with the help of the internet, the domestic user can hear overseas information easily, international relations has become a more interesting and daily topic for normal people than before. Multiple authors have brought the influence and value of social media on foreign policy and culture transmission into an academic area (e.g., Cheng, 2013; Wang 2012).

1.2 Research Objectives and Research Questions

To connect the research realm and bring spatiotemporal aspects into the field of social, cultural and international relations studies, this thesis broadly aims at depicting the connection between states derived from user generated content in social media, performing an analysis from geographical and temporal viewpoints. In this study, we use Sina Weibo as data source to analyse the Chinese community's level of concern for, or interested in, other states. The reason we choose Sina weibo is that almost all of the Weibo user are post in Chinese language and located inside the boundary of China (Louis et al. 2011), therefore they could be identified as representative of Chinese community. Even though the term "state" has a more abstract political and cultural meaning in Cyberspace, the term can still be bound to an actual geographic space. Therefore, those state related datasets can be presented by a map, through which we expect to see the different level of concern for other states shows some spatial patterns.

To reach the research objective, we establish the following research questions to provide a clearer study content.

Data Acquisition A "weibo" is like a tweet – a short message broadcast publicly which has less than 140 characters. Hundreds and thousands of weibos are generated every day, some of them contain textual description refers to a specific state but not all. It is not necessary and hardly possible to collect all weibo posted during the study period. Thus, we need to specify which weibo we should collect, and how we can collect them. This lead to the first research question.

- RQ1: What are the possible methods to harvest weibos that contain textual geoinformation at a state level.

To be more specific, this question could be further divided into two sub questions:

- Which states should be covered in this study?
- How to access microblogs that contain the required textual information?

Information Extraction After we obtain the actual data (weibos) and metadata (e.g., user information, the time of its creation), to analysis large volumes textual content, we need to perform some text processing tasks. This will bring us the second research question.

- RQ2: How should we process the textual content of weibo?

Visual Analysis The weibo content and the meta data like user gender, repost count, have different characteristics, and those characteristics would make certain visualization techniques more suitable than others when presenting them. Therefore, we consider the following research question.

- RQ3: What information visualization techniques should be used to present the different information we harvested from actual weibos and their metadata?

1.3 Structure of The Thesis

The thesis will be finished with the following structure.

Chapter 1 Introduction. This chapter gives a brief introduction of this project, the motivation, and the research objectives.

Chapter 2 state of the art. In this chapter, the research background, the previous work achieved, and related studies are introduced. This chapter will start with the value and characteristic of social media and different geographical information, then comes to the related technique background that includes text processing and visualization.

Chapter 3 Methodology and workflow. This chapter is split into two main sections: methodology design and workflow/process details, which addresses data acquisition, data handling and visualisation.

Chapter 4 Result analysis and discussion. In this chapter, we analyse the information we obtained by describing the spatial-temporal patterns and their possible explanations.

Chapter 5 Conclusion. In the end, the thesis is summarized and possible future work and limitations of the study are presented.

2. State of the Art

The concept of Volunteered geographic information (VGI) was first brought up early in 2007 (Goodchild, 2007). Since shortly after the concept was brought up, comes the fast-growing age of social media. Ever since that time, social media data analysis gained an important role.

In the following period, researchers in GIS domain have been driving towards many interesting directions. Helped by the statistical method and Nature Language Process (NLP), we can use social media data to the model human emotion pattern, to predict the type of a point of interest (POI), or to visualize the transmission of a breaking news.

Today, studies surrounding social media has already reached other social sciences fields. Topics aimed at revealing social structure, national interest and depict virtual communities have been established. Combined with impressive visualization methods, more and more questions have been answered with strong, short and easy-to-understand representations.

The literature review will, therefore, start with this development history of social media analysis in GIS domain, then cover the concept of virtue community. Later, the technical background of NLP commonly involved in related research as well as this project will be introduced. Finally, details about a variety of visualization methods are provided to support the latter work in chapter 3.

2.1 Social Media

2.1.1 Social Media. Characteristic and Value

It is not that easy to understand the impact of an idea or event through what we call "traditional media"— newspapers and TV programs. Because of the user—or we should call audience, does not really participate in the spreading procedure, they play a role more like a pure "receiver". Cyberspace (Gibson, 1984) (including web pages, social media, and online communities) however, is a powerful platform for collecting social discussions, personal networking, and new ideas, for the user also participate in the data generation and spreading process. Now we can trace, monitor, and analyze the spreads of social movements, protests, political campaigns, etc., via social media and weblogs (Ming, 2012). The existence and prosperity of social media have converted the invisible public opinion, public interest, and social issues into capturable social media data stream streams.

What is social media, Kaplan and Haenlein (2010) defined this term in 2010 as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content". From this definition, we can conclude that the main characteristic of social media is that the information content through the whole platform is generated almost all from the user, the platform only provides technical support. Popular platforms include blogs, microblogs (Twitter), Wikipedia like projects and social networks (Facebook, Google+). And the join of multimedia expanded this pool, platforms aim at photographs like Flicker and Instagram appeared, others like YouTube focus on videos, last.fm focus on music share. Nowadays social media even include online games (Piller et al. 2011).

The influence of social media continues to grow quickly and has become an unseparated part of people's life. Since 2007, microblog and other social media (including social networking sites) have heavily influenced the information market globally. According to a statistic published in 2016, 48% of internet user in East Asia is also social media users, and Facebook's global registered users have exceeded 2000 million, Twitter surpasses 300 million monthly active users. This situation was further intensified by the popularization of the mobile device. With a smartphone, users can share latest news or ideas with a text

message, picture, music, and video almost any time, anywhere. Thus, the original limitation of social media was extended through hardware. Receiving and exchanging of user-generated information can be conveniently achieved through social sharing. Till 2016, more than half of Facebook's daily active user is from mobile platform. According to Sina User report 2016, mobile platform users account for 90% of total monthly active user.

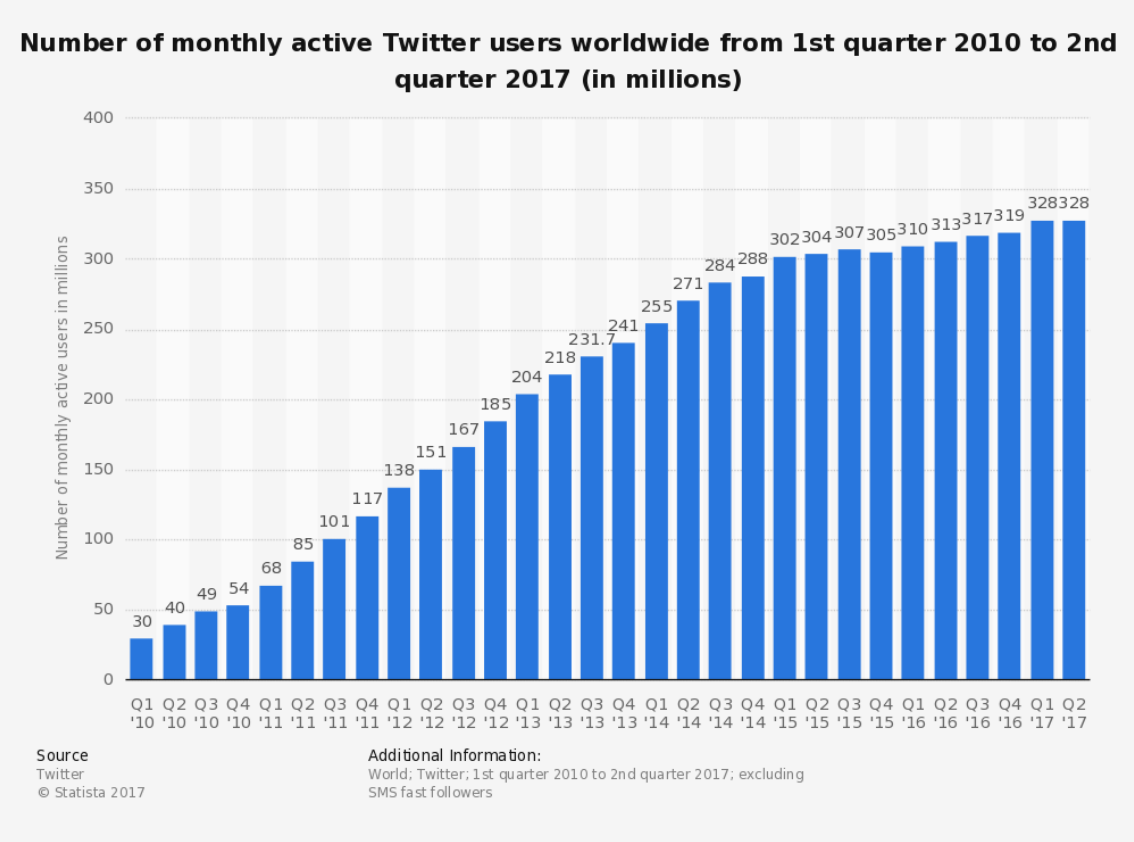


Figure 1.1 Number of monthly active Twitter users. 2010 to 2017 (Statista, 2017)

The current discussion of social media (includes such as Facebook, Twitter, Foursquare, Yelp, and Flickr) in academic area is focusing on its advantages and how we can use those advantages. As mentioned above, the mobile platform takes an important role when talking about social media. Mobile device has advantages of "5A": Anyone, Anytime, Anywhere, Anything, Anyway (Yuan, 2011). While internet support instant multimedia data sharing and searching on a global scale. This combination gives nowadays social media unique features and functions, such as low cost and high-speed information spreading with wide coverage, easy for everybody to participate, convenient communication between different social groups, possibility to enhanced community identity, multimedia integration, geoinformation

attaching and so on. Those features and advantages, in turn, attract participants from a large and comprehensive user pool compare to traditional media. Since participants of social media are real people in the physical world, the platform built real connections between individuals and individuals, individuals and communities, individuals and information.

The content on social media platform is commonly referred as User Generated Content (UGC). This term is described as: The sum of all ways in which people make use of Social Media (Kaplan and Haenlein, 2010). Based on this explanation and the common structure of social media platform (user owns a personal account as representatives of themselves, and they can share information with others in public domain), UGC of social media have two obvious features: First, user create those content as a method to present themselves in Cyberspace (Schau and Gilly, 2003), second, various forms of information could be generated and exchanged. Therefore, Kaplan and Haenlein conclude, social media classification could be conducted based on the degree of self-presentation, and the media richness. Under this classification, they stated that blog is an example of high self-presentation or self-disclosure — it contains personal information include thoughts, feelings, likes, and dislikes. And those types of information shaped a real person in a digital way. Nowadays the popularity of microblogs bypassed blogs. Different from text-based blogs, microblogs (e.g., Twitter, Weibo) limited the length of each post, this limitation makes microblog easier to use for normal people. Other than the convenience of shorter text, the user can post and share other forms of media includes video and pictures attached to personal opinion. This possibility stimulates the interest of the user to join in the public discussion.

Besides all those advantages mentioned above, social media also has a relatively better accessibility compared to traditional data – those data owned by government or media company. Because of those advantages and its large potential, social media data—especially microblogs data as concluded – has become an increasingly popular starting point for those who want to start research involve public opinion (e.g., Shirky 2011; Tsou et al. 2013; Luo 2014), social network (e.g., Hanna et al. 2011), intercultural communication (e.g., Sawyer and Chen, 2012) and so on.

2.1.2 Geographical Information in Social Media Data

Even though the fast development of microblogs has drawn quite an attention in computer science and sociology domain, there are still limited studies from a geographical perspective. The problem spatial-orientated scientists facing is that even we know most of the social media records are generated by real people in real space at an accurate real-time point, it is not easy to extract that information and connect them to the physical world. While this situation has been partially improved by the popularisation of GPS incorporated mobile device and "geo-tag" function supported by many social media platforms (Erickson, 2010).

Nowadays, almost every smartphone has embedded GPS function. Since August 2009, Twitter has allowed tweets to include geographic metadata to indicate where the tweet was generated (Twitter, 2009). Commonly there are two types of geolocation information available: Server suggested a place, this type of information is generated by asking the user to choose a specific gazetteer manually from a pre-edited suggestion list. Or location extraction (Leetaru et al. 2013). The second is to exact user location information via GPS or cellular triangulation, this type of information usually is a set of coordinates. Since place location needs to be input manually or at least need to be select from a list, this step causes trouble for the user who is traveling to other place and does not have internet connection on the way. Spatial and temporal information always intertwine, if the user will post a photo after returning to the hotel, the suggestion place list will change based on the hotel location. In contrast, that Exact Location method access user location through the geographical features of the mobile device itself, the user does not need to manually select a location. While Leetaru (2013) also emphasized that, due to the high precision (to four decimals), this method could capture a house or a shop which lead to privacy risks. So normally, the user need to enable this function first— if the platforms do support this function.

To take advantage of those data, studies involving different aspects have been done, from detecting and locating unusual activates in city (Xia et al. 2014), understanding how the characteristic of different area in a city changes through time (Cranshaw et al. 2012), how is the human mobility pattern through space (Liu et al. 2014; Hu et al.2017), to illustrating the culture pattern of material world by visualization user-generated placemark (Graham and Zook 2011) and map global emotion expressed on Twitter in real time. (Leetaru et al.2013).

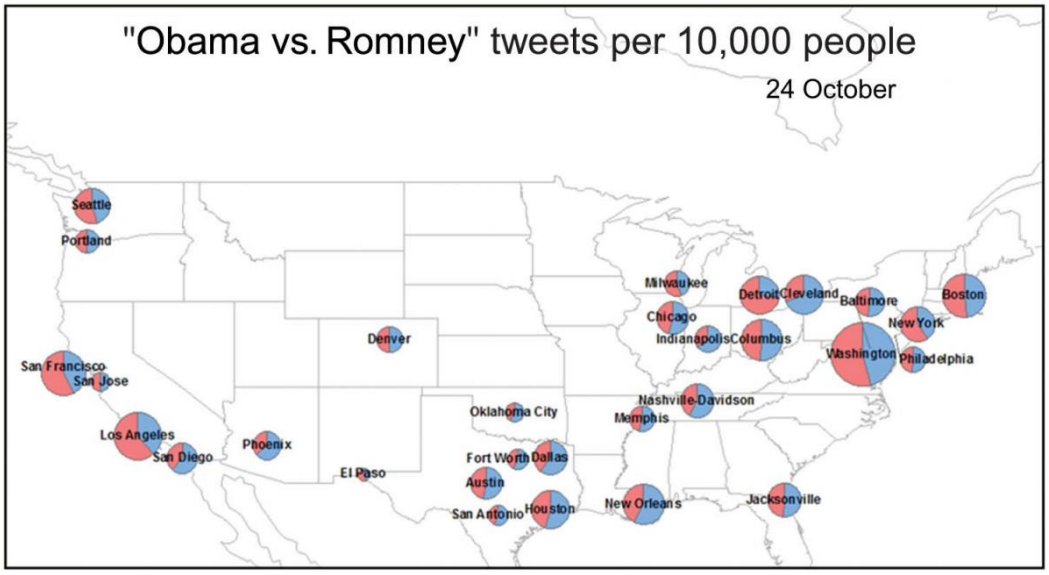
A work down by Beihang Interest Group on SmartCity (BIGSCity) demonstrates how exact location data is used in dot distribution map. This type of visualization could reveal population distribution. In this project, BIGSCity matched the extracted coordinates to their closest POI (represent as one dot), and visualized the "check-in" count of each POI. The brighter the dot, the more intense the location been "checked" through social media.

Figure 2.2: 200 thousand POI visualization in Beijing. The lighter the dot, the higher the "check-in" count. "City mood" project by Beihang Interest Group on SmartCity (2015)

Other than visualizing data with precise coordinate, generalizing coordinates into broader geographic areas is also a common step. In a case study conducted in 2012, Geo-search-enabled Twitter Tools were used during the collection of Twitter data (Ming, 2012). In this study, a set of cities were first chosen to form a no-overlapping search area. Then, tweets was collected based on inside which city range they were generated, but further details like real coordinate and user information were left behind. This geographic area based data collection method is popular for study in public opinion monitor domain. For these studies, the location is the main parameter when talking about data acquisition, while specific user information and user difference are less valuable since the core of "public opinion" is the general trend and main idea of massive data but not individual's. This "from location to data"

approach omits the differences between individuals, it is commonly used when we consider the target as a unified whole, only the distribution and statistical feature of “public” matters.

(a)



(b)

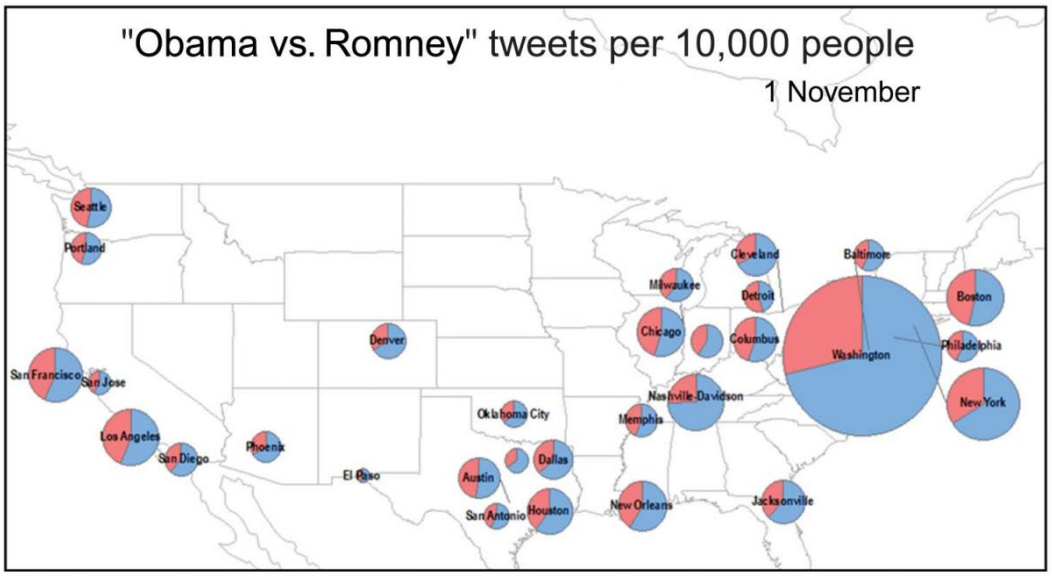


Figure 2.3: Higher-level geotagged data acquisition and visualization. Each pie chart indicates a number of tweeters generated inside a city border range (Ming 2012)

While obviously, geotagged data can also be used in an opposite “from user to location” approach. Another study conducted by Xiaoqian Hu in 2017 demonstrated the potential value of tracking an individual user through social media. In this study, urban population mobility

pattern was identified by tracking the location change using weibo posted during Chinese New Year break. Different from Ming´s study—where individual users were considered the same for data acquisition, for Hu, the series of locations changed during Spring Festival for one user have much more value. This into detail tracking and comparing ensures that the mobility pattern during Spring Festival can be visualized in one map with lines indicate the movements of each tracked user. The advantage is obvious, we can perceive data that represent location change with only one map. But this type of precise data and visualization has its own drawbacks, like cannot present precise numbers, and requires higher computing ability for hardware.

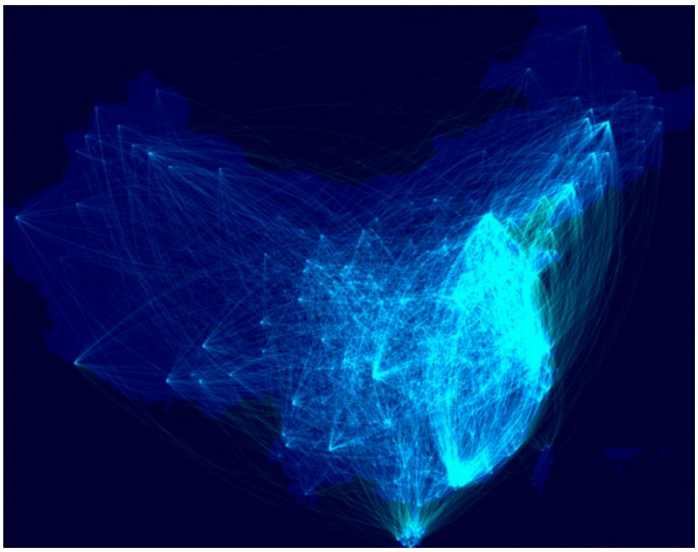


Figure 2.4: User mobility patterns in spring festival travel rush. Visualized with the complete Weibo data collected during the 2017 Spring Festival (Xiaoqian Hu 2017)

While georeferenced social media records may same have significant meaning for all type of study fields, but still they are generated by only a very small portion of the user. During the early mentioned project "Global Twitter Heartbeat" conducted by supercomputing manufacturer Silicon Graphics International (SGI), the University of Illinois, and social media data vendor GNIP (Kalev et al. 2012), just 8.2 percent of all user had either Place or Exact Location information available for their tweets. Research by Hale, Gaffney, and Graham (2012) indicated that geolocated tweets were as low as 0.7%. But it does not matter what is the real percentage of geolocated tweets, what we know is that they do not represent the full value of the geographical information contained in social media. Even with the prosperity view, we

still notice the lack of voice in relatively more sociology direction using geotagged social media data.

To address the limitation caused by the low percentage of geotagged data and expand the usability of social media data. Studies focusing on enlarging the possible harvesting pool was made.

Commonly for social media, the profile information is considered as an important and necessary part. Even though there are some social media targeting “anonymous experience” have achieved quite a good compliment, like Whisper, Secret (has been shut down on April 29, 2015) and Yik Yak, the mainstream of social function design still values the user profile. Especially in recent years, after celebrities, companies, and organizations, sometimes even local government departments opened their “official social media account” widely, the demanding and concern of realizing a real-name system for social media increased heavily.

Under this trend, nowadays systems usually set up a series of methods leading the user to complete their biographical profile page, like a small percentage indicator placed in your main home page showing how much you have leaved blank. Among all those fields in profile page, some of them contain formattable textual geographic information. Take Facebook as an example, it is obviously that “places you have lived in” field contains user input location information, and If we consider one step further, from “work and education” section, we can also extract potential geographic information, for all the possible selections are created with a data page asking for location details of the object. Finally, the content of a record itself may also mention one or more places.

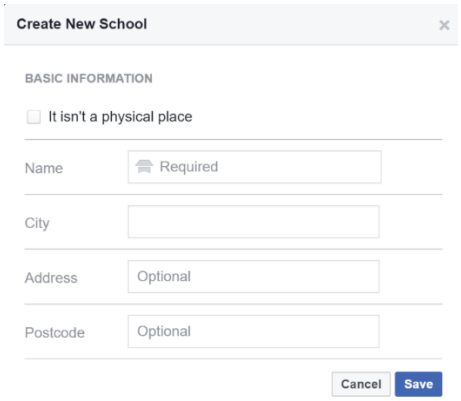


Figure 2. 5: Facebook “Create New School” page

According to “Global Twitter Heartbeat”, generally 71.4 percent of tweets have available user location field data. Compare to geotagged tweets (8.2 percent of total), to harvest geographic information from text may be a more profitable choice.

Textual information gives us the potential to answer more questions, however, this approach leads to new technique problems such as how to determine the level of details, since people refer to different places with different level, from a street name to a country. And if a platform has users from all over the world using different language talking about the same place, this situation becomes even more complicated.

For those data fields, all geographic information is presented in textual form, for example when the user moved from” New York City” to “Hamburg”, or when user post “I will travel to Japan next week”. It requires extraction and geocoding algorithms to interpolate from text into mappable coordinates. But problems occur, because “*Nearly one-third of all locations on earth share their name with another location somewhere else on the planet*” (Leetaru 2012). How we can interpolate textual location into correct, unique coordinate and avoid error and ambiguity is a necessary step, especially when the different geocoding system has their own standard? Hence, for most of the cases, research area and target language are determined with deliberate consideration to minimize the complexity.

2.1.3 Virtual Community

The massive development of social media brings the prosperity of online communities. Different from real space region, the definition and border of a state are relatively fuzzier in Cyberspace. Traditionally, “community” is bound to a geography area, but the “virtual” part of the term indicates that there is no such a physical space (Ridings and Gefen, 2004). When a user using the internet to locate and communicate with others who have similar interests, they form a virtual community (Hiltz and Wellman, 1997). And the communities are not static. “*They evolve because of cultural, environmental, economic, or political trends, external interventions, or unexpected events*” (Ahn 2011). Even though virtual community could be formed across physical state boundaries and always changes, but members with similar language, cultural background, same national interest still could be observed have an obvious closer connection with others.

These features lead us to a new perspective that we can analyze a certain virtue community on "nation" level— which includes the core members (a state's citizens who live within its physical boundaries) as well as those physically overseas members who still actively joining the public discussion in this community.

In the book "The ethnography of communication", Saville-Troike said: "*Regarding group membership, language is a key factor - an identification badge – for both self and outside perception*". Thus, the language could be seen as an important indication of identity (Donath, 1996). Traditionally, in many studies, the dominant language or official language of the research state will be used as a factor to distinguish community members. Like the previous study conducted by Ming, both of the candidates' names were searched in English, even though according to American Community Survey Report (2011), of 291.5 million people aged 5 and over, 60.6 million people (21 percent of this population) spoke a language other than English at home.

The lack of consideration of multiple languages may lead to a biased outcome if the target state-level virtue community has a complex culture component and more than one dominant languages (Main languages). In a work from Fischer (2011), different colors were used to represent the language a geotagged tweet used from 23 October 2012 to 30 November 2012. The result indicates that most countries show strong homogeneity with a single language, but people in some countries show a diversity use of languages. This situation is significant in Europe, country such as Serbia even have no dominant language. Under this situation, it is not easy for scientists to establish researches involve multiple languages and states, especially for those cross-culture topics. The methods to identify communities are heavily affected by the trend of globalization and massive immigration in recent years, study around analysis the states and boundaries remained an important topic now in Cyberspace.

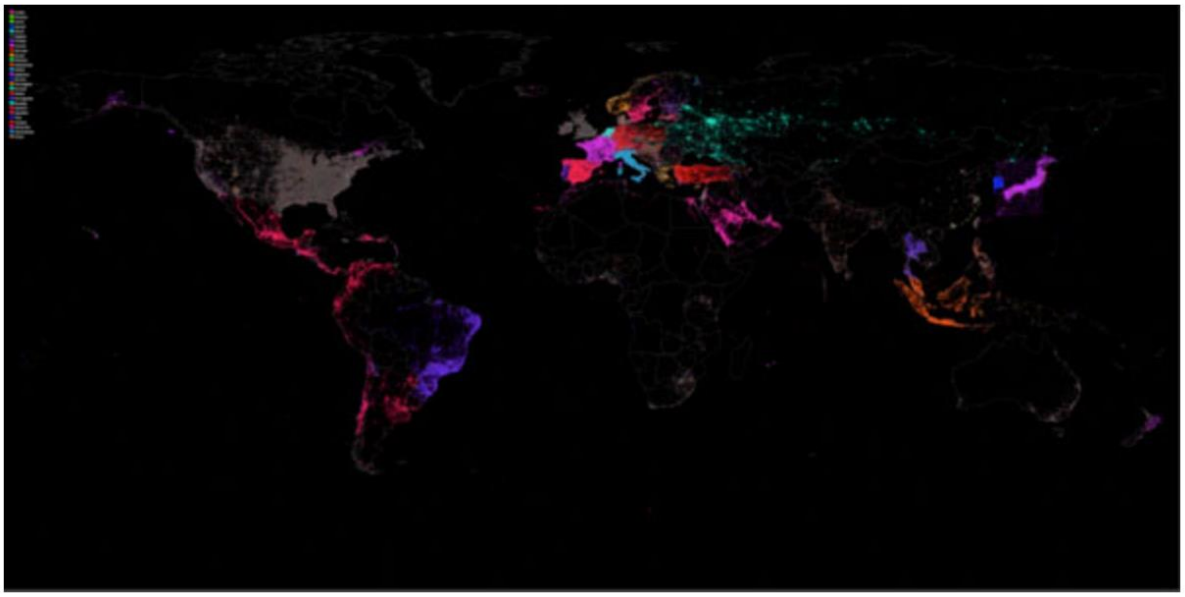


Figure 2.6: Georeferenced tweets from 23 October 2012 to 30 November 2012 colored by language (Fischer, 2011)

Thus, different approaches have been conducted to bypass the difficultly caused by multiple languages usage in social media data. The study conducted in 2012 (Stefanidis et al. 2012) offers a different way of thinking. If the main question of the study is to identify the connection between states (the people of those states), then using one keyword (state name) with several predetermined languages is enough to represent a virtual state, as well as depict the communities surround it and its projection on the physical space. In this study, Stefanidis (2012) picked Syria as a test case, collected Twitter feeds through keyword queries. Over the study period of one week, Twitters included "Syria" or its hashtag have been harvested globally. By analysis, the dataset, a simplified relation network of Syria was created. In this project, Stefanidis combined the two types of geographical information mentioned in the last section. The textual information represents "Syria" in both English as well as the formal Arabic spelling was used to query the discussion about Syria from the Twitter server, the records with any sort of geolocation information ranging from geotagging to user profile city name were documented into different real space states. His work suggests the new direction of using textual geographical information even without fully developed geocoding algorithms. A pre-select request keyword could serve as a "place of interest" while social media data related to it could be harvested and used as an indicator of its connection through the world.

2.1.4 Sina Weibo

In July 2006, Twitter was created. Three years later, Sina Weibo was launched in China. Later this year, Twitter and Facebook were blocked in mainland China. Since then, the user group of domestic microblog services such as Sina and Tencent started to grow fast (Yu 2011). During the past decade, both Twitter and Sina achieved great success. Different from Twitter, the user of Sina Weibo is concentrated within Chinese. This situation offers a distinct convenience for researchers, Twitter user comes from a variety of countries, while the millions of users of Sina Weibo are almost all located in China and post in Chinese language (Yu 2011). This feature makes it possible for us to consider Sina user group as a unified virtue community representing China.

According to the 2016 Weibo user report (Sina, 2016), the whole user pool consists of 55.5% male account and 45.5% female account. 77.8% of the total have received higher education. People under 30 years of age are the major user, accounting for more than 80%. User between 18-30 years of age reached approximately 70%. Based on the urban development level of user distribution, the user composition indicated a trend of user increasing in a second and third tie of cities. The current statistic shows a relatively equal distribution through cities from all level of development since in China (Sina, 2016), the city tier is assigned mostly based on GDP or population [1], the comprehensiveness and representativeness of Sina Weibo data are guaranteed.

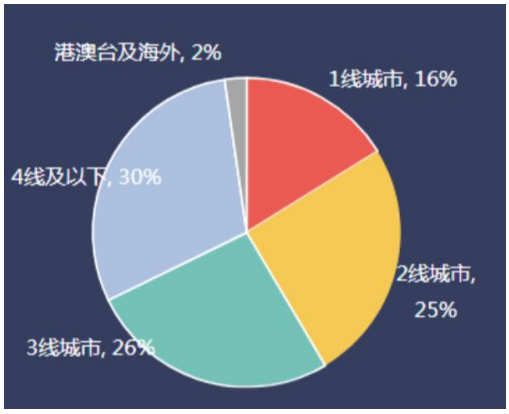


Figure 2.7: User composition of Sina Weibo (Sina Data Centre, 2016). 16% from first tie cities, 25% from second tie cities, 26% from third tie cities, 30% from fourth tie cities and below, 2% from oversea or special administrative regions (SAR)

[1] Urban legend: China's tiered city system explained. http://multimedia.scmp.com/2016/cities/

2.2 Nature Language Processing

The typical research directions involve social media data include work like topic classification, trend detection, sentiment analysis. For all of them, text processing is the very first step, hence, those process roughly fall in Nature Language Processing domain.

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things (Chowdhury 2003). In the 90s, the three trends—heavily increased texts through the internet; computers with increased speed and memory; and the arrival of the Internet, brought in a rapidly growing in NLP area (Liddy 2001). Commonly, when talking about NLP, most of the functions are mainly based on statistical algorithms relying on the textual representation of data. Those functions covered a wide range of basic demanding from retrieving texts, splitting text into parts, checking spelling to word frequency counting (Cambria 2014). But the goal of NLP is not just language "processing" but more "understanding". According to Liddy, a full NLU (Nature Language Understanding) System would be able to: 1. Paraphrase an input text. 2. Translate the text into another language. 3. Answer questions about the contents of the text. 4. Draw inferences from the text. However, even though in the past great progress have been made to accomplish tasks 1 to 3, the current study still cannot answer the last goal.

Even though NLU is hard to realize, NLP is possible. The current study surrounding text processing—including Chinese text, usually concern on the following three specific topics.

Chinese word segmentation

Unlike in English, the text consists sequences of words naturally separated by space, Chinese is an ideographic language in that there is no delimiter between words in sentences. Therefore, Xue (2003) concluded that word segmentation is treated as an inseparable part of Chinese sentence understanding. Most of the existing word segmentation systems could be classified into three categories depending on the algorism they use:

1. Segmentation based on string matching:

This method is also called the mechanical word segmentation method, it compares the target Chinese string with every word in a “big enough” pre-compiled dictionary if the dictionary finds the string contains one pre-compiled word, that is a match (successful identified a word). Based on different scanning direction, the string matching word segmentation method can be divided into forwards matching and reverse matching. According to the different length of the priority match, can be divided into the maximum (longest) match and the minimum (shortest) match; Based on whether the process is combined with other methods, it can be divided into pure word segmentation or word segmentation combined with part-of-speech tagging.

2. Segmentation based on statistical approach:

The form of a Chinese word is a stable combination of two or more Chinese characters, so the more a combination of adjacent characters appearing in the context, the more likely they form a word. Therefore, the frequency or probability of adjacent characters reflects the credibility of correct word forming. The mutual information between two characters can accordingly be calculated. When giving a string of characters, the pair of adjacent characters with the largest mutual information greater than a predetermined threshold is recognized as a word. Repeat this process until there are no more pairs of adjacent characters with a mutual information value greater than the threshold. This method only needs to count the frequency of the words in the contexts and does not need a pre-compiled dictionary, so it is called no-dictionary word segmentation method or purely statistical word segmentation.

But this method also has some limitations, there are certain adjacent character pairs with high mutual information but do not form any commonly used word, since the Chinese language also have single character words. That lead to the third method.

3. Statistical dictionary-based word segmentation:

In order to get the best of both approaches mentioned above, in practice, world segmentation is commonly conducted by combining the use of a dictionary and statistical information. Using a dictionary to recognize commonly used words, at the same time using a statistical approach to detect new words from context. Compared

with the other approaches, statistical dictionary-based word segmentation has the advantage of fast and high efficiency from string matching segmentation and also the merits of statistical segmentation like improved accuracy from new word detection ability.

More recent work on Chinese word segmentation discussed the usability of understanding based word segmentation (Guohe and Wei, 2011). The basic idea is to perform semantic analysis with word segmentation, using the syntactic and semantic information to deal with potential ambiguity. Scientists consider it the future trend, many studies have been done based on this direction mainly concern Artificial Neural Network (ANN) algorithms.

Part-of-speech tagging

Parts-of-speech (also known as POS, word classes, or syntactic categories) is the process of assigning a part-of-speech marker to each word in an input text. It is important and useful because of the large information they can tell about the word and its adjacent words (Jurafsky and Martin 2014). The tags or word classes, broadly include eight big categories: noun, verb, pronoun, preposition, adverb, conjunction, participle, and article. Nonetheless, in some projects involve language other than English, more detailed classification has been used, like the 43 tags adopted in the word segmentation standard for Chinese information processing issued by the Central Standards Bureau in 1999.

Part of speech tagging is important, according to Jurafsky, in knowing whether a word is a noun or adverb and can help us judge the information of neighbouring words. Even though we normally consider word segmentation the precondition of parts-of-speech tagging, but recent researchers brought those two steps into an equal position. Using the potential language relationship between the two tasks, the use of the joint model for word segmentation and part-of-speech tagging will lead to a great improving for both tasks (Ng and Low, 2004; Zhang et al. 2014).

Keywords extraction

Keywords extraction or Keyphrase extraction is the process to select a set of most relevant terms from a giving text, the outcome briefly summarizes the document. Extracted keywords are particularly useful because they can be interpreted individually and

independently of each other (Witten, et al., 1999). They can be used as descriptions of the documents for retrieving in a digital library (e.g., Nguyen and Kan, 2007), measuring document similarity (Habibi and Popescu-Belis, 2015), classification, abstract and other fields. For example, consider the articles have similar keywords as one group can improve the performance of K-means clustering; from all the keywords extracted from one day's news reports, you can generally know what happened that day, or extract the keywords from a group of microblogs would tell you what they are mainly discussing.

There are many existing methods to perform keyword extraction process, the most basic and simple one is frequency–inverse document frequency (TF-IDF) (Salton and Buckley, 1988). To determine whether a word is important in an article, the most obvious way is word frequency: important words tend to appear in the article more. But on the other hand, it is not necessary that words with higher frequency are more important, since some commonly used words are frequently present in more than one articles, the importance of them is certainly not as much as those that have a high frequency in only one specific article. In other words, the importance of a word is proportional to the number of times it appears in one document, but it is inversely proportional to the frequency it appears in the corpus. From the statistical point of view, the core of TF-IDF is to assign those uncommon words a larger weight, while reducing the weight of common words. Here, IDF (inverse document frequency) is the weight, TF (term frequency) refers to the word frequency. The overall approach works as follows. Given a word t_i , its importance can be represented as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Where $n_{i,j}$ equals the number of times word t_i appears in document d_j , the denominator is the total number of all words in document d_j.

$$idf_i = log \frac{|D|}{|\{j: t_i \in d_i\}| + 1}$$

$|D|$ is the total number of documents in corpus, $|\{j: t_i \in d_i\}|$represent the number of documents contain word t_i, but if this word is not pre-existed in corpus the denominator would equal to zero, so in practice $|\{j: t_i \in d_i\}| + 1$ is commonly used.

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

TF-IDF is still the most wildly used method in the present open source NLP tools..

2.3 Visualization

One feature of social media data analysis is that it cannot be easily quantified, including videos, images, Emojis, geographical information and text. Microblogs as a text-based social media type, mainly dealing with user-generated textual content. NLP work introduced in section 2.2 could help us to process the massive amount of textual data, however, we still need to figure out how should we extract, convert and present useful information contained in the textual data—which, is always of a variety of types. For instance, textual "user gender" filed could be counted and converted into quantifiable user gender ratio; Emoji frequency could be calculated and ranked; "Hometown" filed could be geocoded into mappable coordinates. The main point is that, from textual raw data, we could obtain different new types of data. To make the analysis possible, exploratory approaches and user involvement is the key point, and the answer to this problem is a visual representation (Schreck and Keim, 2013).

Thus, the following section will continue with the structure and meaning of visual data analysis, then, introduce the general feature, preference, and advantages of different visualization methods.

2.3.1 Visual Data Exploration

The arrival of the internet has brought the explosive growth of data, covering documents, social networks, financial transactions, urban environments, news multimedia and other aspects. How to present information in a more intuitive and easier way is a great challenge of big data analysis (Keim 2006). Facing the challenge, scientists have developed information visualization technology, and represent data with visual symbols to help people understand a large amount information.

Graphics, symbols, colors, and textures are easier to understand by humans compared with text and numbers. Information visualization is this typical interdisciplinary

research field aims at conducting visualization analysis with human perception and certain interactive means to enhance people 's understanding of data.

Initially, it is easy for the user to collect a huge amount of data from any other automated process, however, if the data is displayed textually, the amount of data able to be presented is very limited. And without human participation, the value of the data cannot be interpreted. Therefore, Visual data exploration (Visual data mining) aims at integrating human in data exploration process (Keim 2002; Simoff et al. 2008).

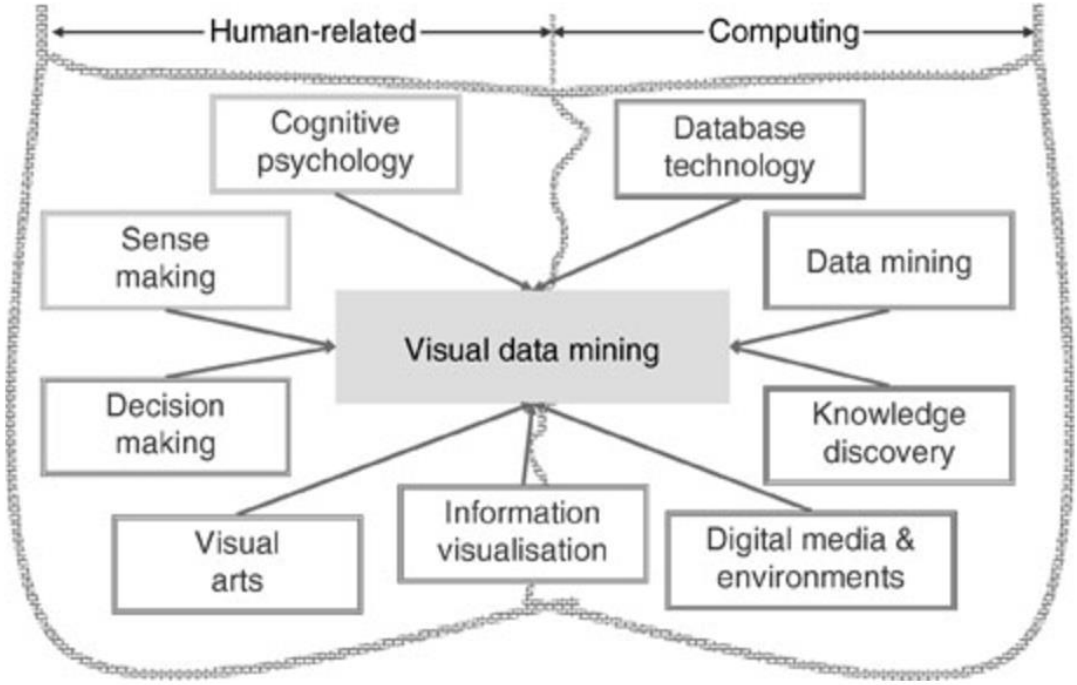


Figure 2.8: Visual data mining as a confluence of disciplines. (Simoff et al. 2008)

Keim (2002) explains the traditional three steps process in visual data exploration: 1. Overview. 2. zoom and filter. 3.details on demand.

First, an overview of data needs to be presented to the user. Thus, the user can perceive interesting patterns from it and focus on one of them. For a closer analysis of the pattern, the user needs the possibility to access details of the data. Visualization technology could be adopted in all three steps: For data overview, visualization techniques allow the user to identify interesting subsets efficiently. And when a subset is picked, it is important to keep the overview while conducting another visualization techniques on the subset. In conclusion,

the goal of visual analysis has been summarized as "*pursues a tight integration of human and machine by enabling the user to interact with the system.*" (Sacha et al., 2014)

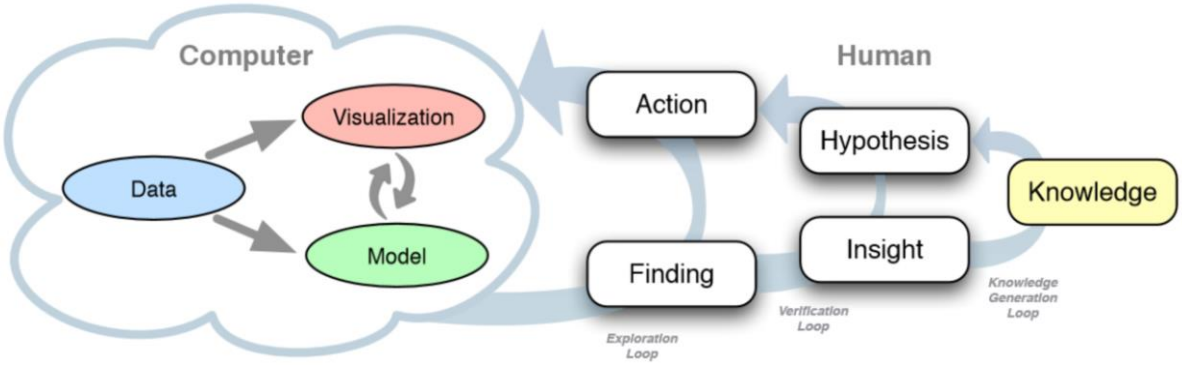


Figure 2.9: Knowledge generation model for visual analytics. (Sacha et al., 2014)

2.3.2 Classification of Information visualization techniques

Data we harvested always contains multidimensional information. To map this type of abstract into physical space clean and clear, a variety of visualization techniques are used, such as histograms, pie charts, scatter plots, line plots. Keim (2002) summarized that they can be classified from three aspects: data type, visualization techniques, interaction or distortion techniques.

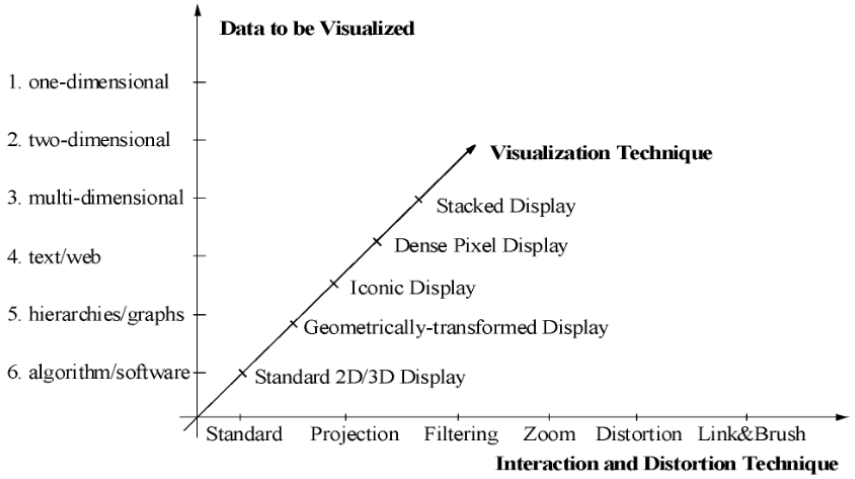


Figure 2.10: Three criteria of information visualization techniques. (Keim 2002)

There is always more than one way to present a dataset, to determine the most suitable visualization methods for different datasets in a project, we could consider from these three aspects.

When preparing data for visualization, we commonly classified them firstly by dimensions. From simple to complex, one and two-dimensional data are the most common type. One-dimensional data includes most of the statistical data, usually are mathematically expressed values (number, fraction, percent). Two-dimensional data refers to those have two distinct attributes. Coordinates, for example, the longitude value and the latitudes value are the two attributes. Accordingly, we could also define three-dimensional data. These low dimensional data could be easily visualized with 2D or 3D plots. Like in an x-y-z plot, use the intercept of each axis to indicate the value from each dimension. For specialized data like coordinates, the map could be used. The standard 2D and 3D visualization techniques include well-known charts like bar charts, pie charts, line graphs.

While multidimensional data consists of more than three attributes, therefore cannot be easily visualized with standard 2D or 3D plots. An example given by Keim (2002) are tables have ten or even hundreds of columns (attributes). Those datasets could be presented by more sophisticated methods like Parallel Coordinates— presenting all dimensions with parallel axes, and same as in a 3D plot, still use intercept of each axis to display the corresponding value. Other visualization techniques for high-dimensional data were classified into Iconic displays (e.g., Star glyphs visualization), Dense Pixel displays (e.g., Data clock), Stacked displays. Based on the pursuit of an individual project, they have their own strengths in different situations.

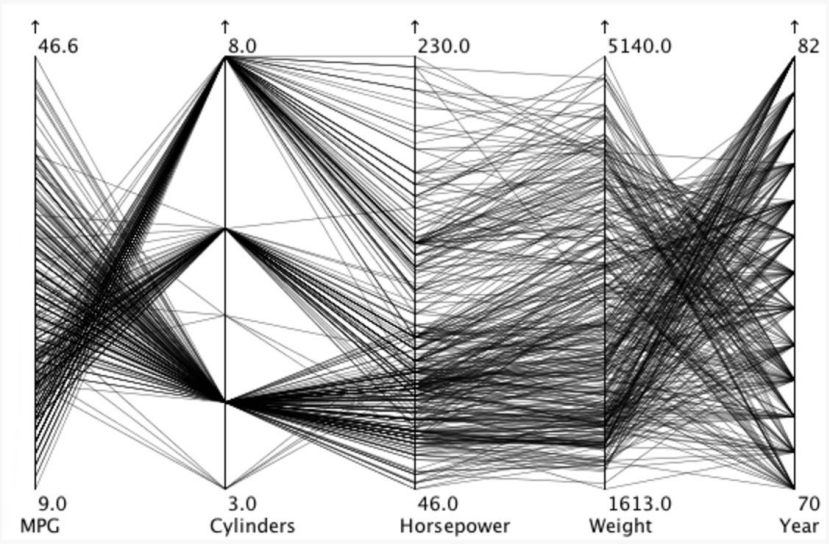


Figure 2.11 Parallel coordinates. (Kosara 2010)

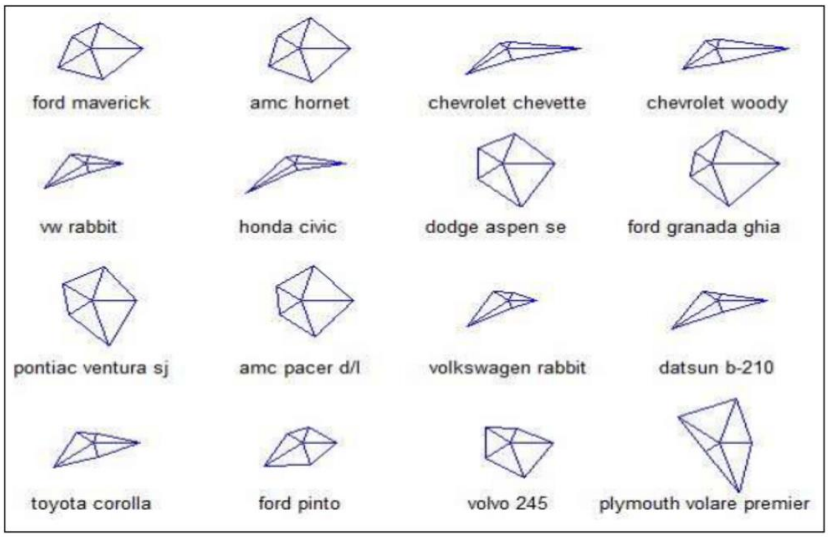


Figure 2.12: Star glyphs visualization (UCI machine learning repository)

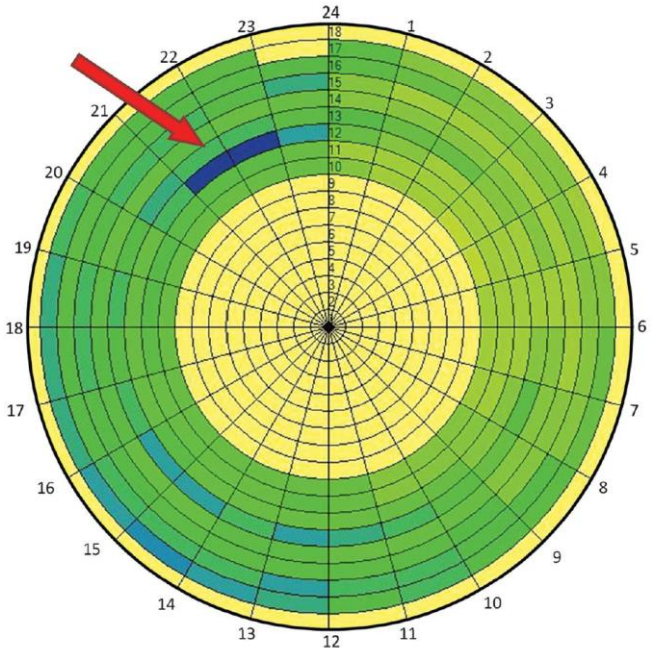


Figure 2.13: A data clock used to indicate tweet count in a month per hour (Stefanidis et al. 2012)

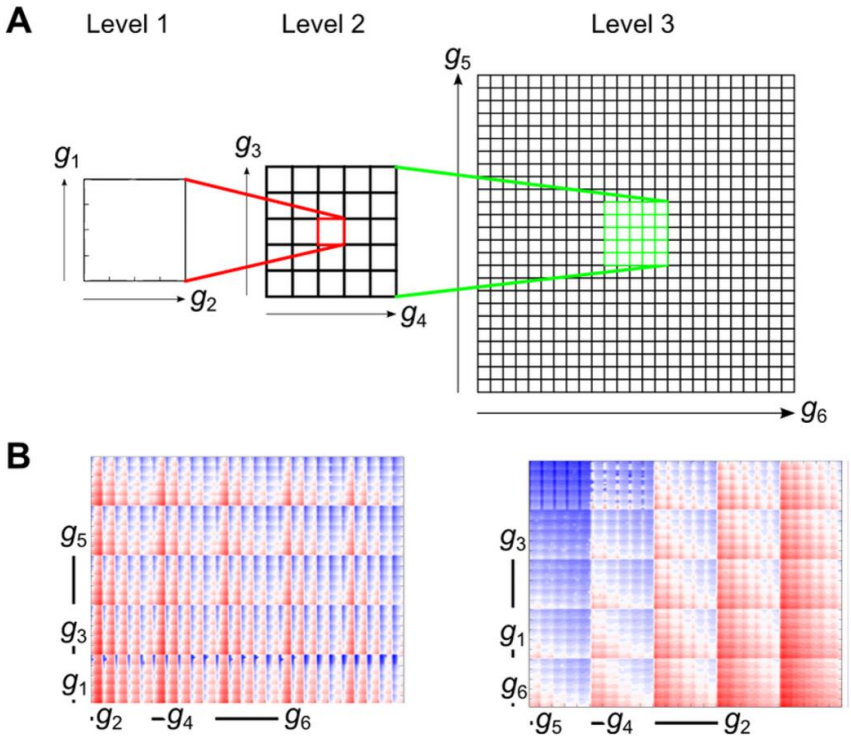


Figure 2.14: Dimension stack display and its process (Gemmell, Burrage et al. 2014)

Not all data could be described with “dimension”. Textual data, for example, contains very rich meaning, but hard to present directly. When dealing with textual content, a human could understand the meaning of each word, while visualization techniques could help to emphasise the quantifiable attributes. In most of the cases, textual data will be displayed in combinations with other statistical values. A simple transformation example is Word counting. When using word cloud to present the word frequency, the statistical value is bound to the size or color of the word.

Besides text, there is another type of important data: Hierarchies and graphs. For any social media, users could be seen as individual dots, and the “friend relation” between users could be seen as connections or edges, the whole platform is like a net. This type of data is especially important for study involving social network, community structure (e.g., Girvan and Newman, 2002), they can be visualized by network graphs.

Interaction and distortion technique is the third aspect when talk about visual analysis. From the above two dimensions, a specific dataset could be connected to a most suitable representation method. Like using a pie chart to display market share, using a map to indicate the location of an event, or using a bar chart to present the rank and number of state’s populations, using a line chart to indicate the price trend of a product. When dealing with more complex multidimensional dataset, we could use other mentioned techniques to present more information in one static visual outcome.

However, interaction techniques are also needed if we want to bring human into the visual analysis. The interaction between human and a visualization includes filtering, zoom in, zoom out, and linking. These functions could give user the ability to observe patterns and connections across datasets, and dynamically change the visualization based on the current objectives. Besides interaction, distortion technique is also a useful method when dealing with complex dataset. Changing the level of details according to the importance or current mission could bring the focal point on different part of the dataset. Like birds eye view display, the center of current view shows more details than the data far from the center. The interactive and distortion technique could improve the efficiency of human interpretation in a great level, they are therefore especially popular in this digital age..

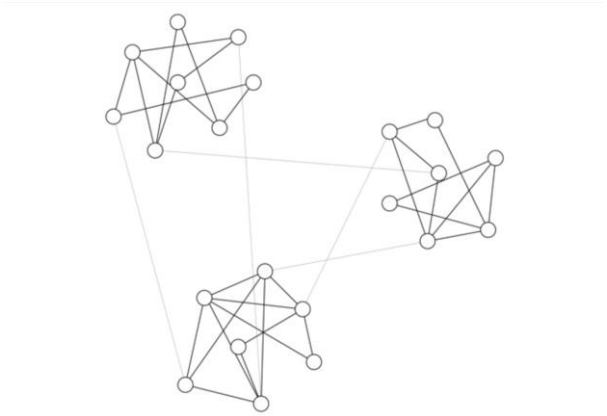


Figure 2.15: A schematic representation of a network with community structure. (Girvan and Newman, 2002)

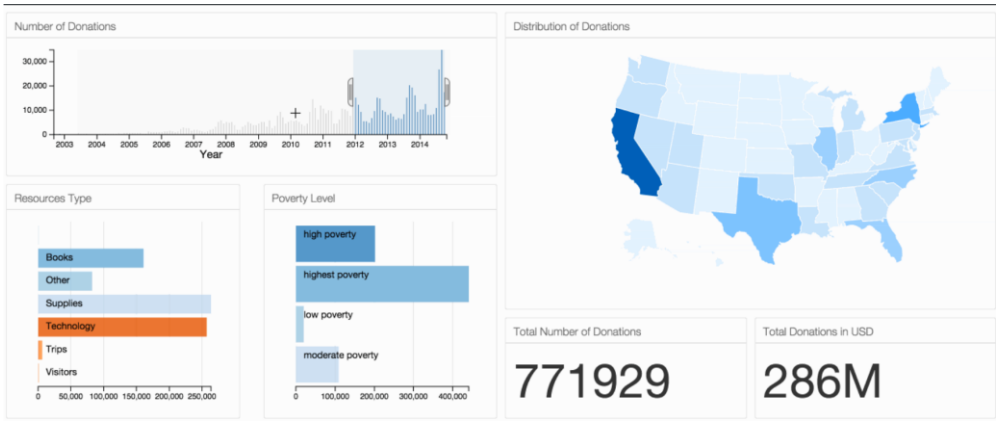


Figure 2.16: An example of Interactive Data Visualization project (Adil Moujahid, 2015)

Social media data analysis is an interesting but complicated topic. The majority of preceding literature has answered questions involving many areas. When talking about geoinformation in Cyberspace, the previous studies proposed the concept of virtue communities and how we can project invisible connections between them into real space (e.g., Stefanidis et al.2013). Hence, this additional research will continue the discussion and try to reveal the relations between Chinese social media user and the rest of the world. The next methodology part will provide the research framework and details to answer the research objectives and research questions present in first chapter.

3. Methodology

3.1 Broad Design

The research methodology was designed starting with determining the study area and target virtual community. Since Sina Weibo is the most popular microblog platform in China and is dominated by Chinese language users (Louis et al. 2011), this study picked Sina Weibo as study platform. Then a list of states was decided, and worked as the input of data acquisition method. This part addressed the first research question. The acquired social media records were the data source of the study. Basic NLP functions were then conducted through a word segmentation tool. After addressing question 2, keywords of each record were extracted and combined with attributes obtained beforehand — state name, user screen name, user gender, user follower, platform, post time, repost count. These consisted the interest of this study, and the characteristic of each attribute was considered to address the last research question. The last part consisted of an analysis of the study data.

Based on the design, the workflow of this study is depicted as follows. Details and process are included in the last section of this chapter.

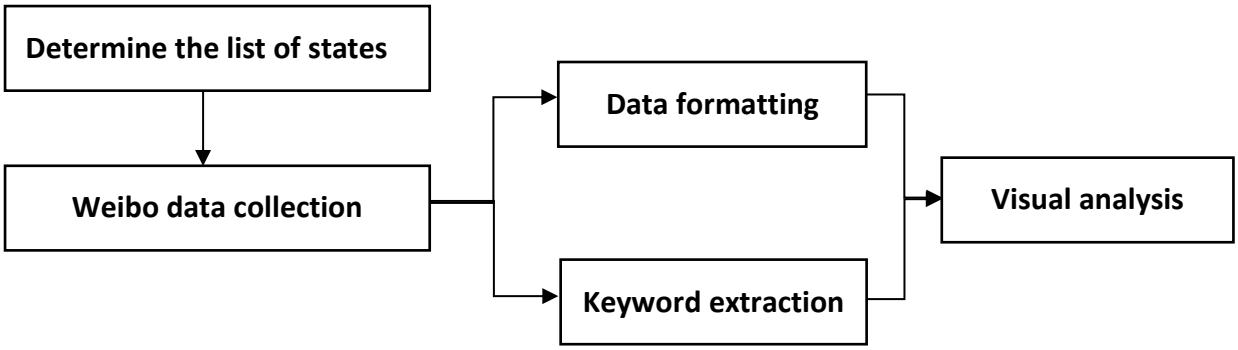


Figure 3.1: Workflow design

3.2 Data Acquisition

There are different ways to collect Sina Weibo data. For this study, two main methods, i.e., API-based and web crawler-based methods, were considered and compared.

An application programming interface (API) is a set of predefined functions that applications can use to interact with other project components, servers, enterprise, and database. With the help of APIs, it is easier to develop a program because programmers can call a method block without knowing the details about how it was realized. Next, web crawler is also a popular method to get information from the internet, it can be used to copy the content of web page for later analysis. Unlike APIs, web crawlers cannot interact with the source, and it is used mainly for automatically collecting large web data.

Sina APIs

In July 2010, Weibo open platform was formally announced and open to the public. Weibo open platform is an information subscription, sharing and communication platform based on Sina Weibo, providing assistance and communication channels with Sina's huge amount of data. Developers or website builders can create their own applications to access the information in Weibo system or acquire those data for further analyzing purpose through its Open API.

Weibo APIs are not open source, the programmer needs to apply access using a specialized key that is assigned to every registered application. The Software Development Kit (SDK) support several popular languages and a variety of platforms including Java, PHP, Python and Android SDK, IOS SDK and so on.

As of early 2017, more than 40 different type of stable, flexible functions were provided by Sina. They include the following 4 main types.

1. Weibo API

   The Read API and Write API provide functions related to microblog itself. Read API includes return the latest public weibos, return the latest weibos of one user, return the single weibo's content by its ID and so on. Write API includes post, repost, and delete function of a weibo.

2. User API

User API provides access to user data, including functions such as return user profile by user ID, return the follower's counts and weibo counts of a batch of users.

3. Comment API

Comment API provides functions like return comments of one weibo, return a batch of comments, return the latest comments of the authenticating user.

4. Search API

Search API contains three big part. First, the suggestion search API provide suggestion function when user search for something, suggestions for searching weibo or suggestions for search applications. Second, User Search API provides the ability to search for a user based on username. Third, Content search API (the only one paid feature) is a high-level search web content function.

web crawler

A web crawler, also known as a robot or spider, is a program that is capable of iteratively and automatically downloading web pages (Chun 1999). Web crawlers can help scientists and programmers to obtain data from large numbers of web pages especially when there is no API.

The basic algorithm of an “iterative” web crawler is simple: Given a seed URL or URLs, the crawler automatically downloads the page content and extracts any other hyperlinks in this page, and then downloads those referenced pages iteratively. If the the target pages are very limited and specific, the URLs could also be defined beforehand. Therefore, the main problem of building a tailored web crawler is to define which fraction of web pages you want to download, and then form the corresponding URLs.

The three steps to create a simple web crawler could be described as:

1. Analyze the request format.
2. Make HTTP requests.

3. Send requests continually, and get the data.

The existing Sina API provides a number of functions that perform parts of the most valuable data collection processes, such as public_timeline, *which* returns the latest 20 public weibos, or show_batch, which returns a batch of comments for a weibo. However, currently available functions offer a limited possibility to request weibo that contains a certain word, forcing us to develop an individual crawler program to acquire data.

3.3 Keyword Extraction

To reach the broad aim of this study, analyzing what people are talking about is essential. One way is to count the word frequency, but due to the feature of natural language, the most frequent words are always unimportant conjunctions or pronouns. To address this problem, the strategy was to extract keywords from each record first and then only calculate frequency from those extracted key words.

Currently, there are plenty of open source modules online provide basic NLP functions to developers. These include as Stanford CoreNLP[2] and Apache Lucene[3]. For Chinese NLP specifically, there are projects like ICTCLAS[4], Pan Gu Segment[5], Paoding's Knives[6], and Jieba[7].

For this study, the following aspects were emphasized:

1. Performance under microblog like short text.

   Different NLP modules use divers segmentation algorisms, this led to performance disparity for a different type of input text. Some of the modules may pioneer when dealing with long text like news report and article, while others have higher accuracy with short, informal text.

2. Supported programming languages.

[2] Stanford CoreNLP: https://stanfordnlp.github.io/CoreNLP/
[3] Apache OpenNLP: http://opennlp.apache.org/
[4] ICTCLAS: http://ictclas.nlpir.org/
[5] PanGu.Segment: https://www.nuget.org/packages/PanGu.Segment/
[6] Paoding's Knives: https://github.com/cslinmiso/paoding-analysis
[7] Jieba Segmentation: https://github.com/fxsjy/jieba

Modules usually support only several programming languages or even support only one specific language. Cross-language using is possible but not recommended considering the difficulty, especially when there are other choices.

3. Interface methods.

   Usually, to use existed module, there are three main possibilities. First is through REST API. Through REST API, uses need to use HTTP verbs (GET, POST, DELETE, DETC) to make the request, the advantage is that user does not need to download any extra files, but a stable connection to the server is needed. Second method is upload local data to the server and ask for processing. This method is easy to use and usually with a user-friendly user interface, therefore the user does not to know any programming knowledge, but the functions are less flexibility and limited. The third way is to dowload and install the open source package locally. Since the package contains all preprogramed functions, user could work offline, but some programming ability is required.

Based on the three points above, NLP component Jieba was used in this project. Jieba is one of the best open source Python-based word segmentation module for Chinese. It supports keyword extraction, no need for authentication, works offline, small and fast, and it also supports user-defined dictionary. Finally, it is open source does not cost any more to use.

As mentioned in section 2, the most widely used keyword extraction algorism is TF-IDF. Jieba also used this algorism to realize this function.

3.4 Visual Analysis

In the previous section, several attributes were mentioned: state name, user screen name, user gender, user follower, platform, post time, repost count. The statistical results could reveal a lot interest patterns and connections. Hence, to fulfil an effective visual analysis, presenting these data with suitable visualization methods is necessary.

Based on experience and intuition, we could assume some potential results. For example, there should be a huge amount of weibos were harvested under “popular states”

(e.g., a place is often mentioned because it is a famous tourist destination), while some states may only be mentioned rarely. And this different popularity of the states shall follow some spatial pattern, or consistent with certain phenomenon. Then, for one state, the number of weibos collected everyday should fluctuate, this fluctuation may correspond to breaking news. On the same time, male and female user may have a different preference towards different states and also a diverse topic preference for one county.

The attributes data we harvested have some different features, we need to consider them separately to determine the most suitable visualization techniques.

Presenting data amount related to the states

We want to emphasis two points when dealing with this information. First, we have 195 states involved in this study, and they all need to be indicated and viewed separately. Second, the contrast of different data amount related to each state should be clearly perceivable in a map.

There are two ways to present the weibo counts. First, a pie chat that use percentage to give a clear view about who takes bigger part. Second, we can use a bar chart to list all the objects from highest to lowest. For this study, the more suitable chart type is the **bar chart**. The main reason is that if we divide a pie chart into 195 pieces, it is hard to indicate the smallest slides. In contrast, a bar chart is a very practical method to visualize data with a big number of categories. Here for this study, the height of the bar indicates the size of the group—the number of weibos harvested under this state. Also, bars can be ordered based on the height, therefore provide a clear rank. The width of the bars are all the same, this ensured that even the smallest component can be presented clearly and checked independently.

Presenting the location of the states and their data amount

To present the location information, a map was added. There are two popular ways to present statistical variable on a map, one is using a choropleth map, second is using a proportional symbol map.

Here in this study, we choose **proportional symbol map** for one of its outstanding advantage— the size of the target unit does not matter. A state with small geographic area could hardly be seen on a choropleth map even when it has a large data value attached, but

using a proportional symbol map will not have this problem— a big symbol will be present over this state.

Presenting the top popular platforms

From the extracted attribute, we can count how many weibos were posted from which platform. The rank could indicate the popularity of a certain mobile device, a web browser or a third-party application. One special feature about this attribute is that, the number of different platforms is theoretically limitless. New applications could join in from any time. So we decided to use a **bar chart** to present only the top ten popular platforms.

Presenting the sex ratio

The user of different gender could have a distinct preference. Or some states are more appealing for one gender than the other one. The proportion of male and female matters. A pie chart is suitable to present the component subsets of the whole dataset, especially when the number of the category is small. Since a user sex is either male or female, a two slices **pie chart** would be clear and direct enough.

Presenting the most frequently mentioned words

To visualize the importance or the frequency of the words, word cloud and proportional dots are the two most popular techniques. The idea of word cloud is that the font size of each word indicates its importance or frequency, since larger elements draw greater attention than smaller elements, people would notice words with higher frequency first. Proportional dots method is relatively the same, use a dot to present the word frequency. Compare to plain word cloud, proportional dot method provides a more appealing visual feeling, but sometime when a word it too long, it is hard to put the text inside a small dot. In this study, we would like to present 120 most frequently mentioned words in a chart, using direct **word cloud** could keep more details.

Presenting the data amount change of each state

To depict a trend, **line chart** is always a popular choice. Any outstanding peak of each line could represent an outbreak news or equivalence. Stacked area chart could be an alternative possibility, especially could present also the part-to-whole relations. However the trend analysis for certain part would be harder compare to line chart.

3.5 Workflow

3.5.1 Test State Selection

In order to cover most of the regions on earth, at the same time avoid political bias, the list of states was mainly determined based on the member states of United Nations. Until the start time of this research, there are 193 members in total. However, the two Non-member observer states: Holy See (often informally referred to as "Vatican") and State of Palestine (commonly referred to as "Palestine"), is also important, and should not be left out.

Specifically for the Chinese community, Chinese Taipei (commonly referred to as "Taiwan") owns a controversial identity and “Taiwan” as a culturally and politically distinguishing topic always obtains a high level of attention and a huge amount of discussions. Considering this situation, “Taiwan” was included in the request list and referred in this thesis also as a “state”, though it is not recognized by United Nations.

In the end, since the broad aim of this research is understanding how Chinese community reacts to overseas states and international news, but not domestic one, “China” was removed from the request list.

Considering the particular data we were dealing with is social media feeds, a transformation of the names was conducted. Commonly used references were adopted to replace some of the unfamiliar official names. The full list is included in appendix 1.

3.5.2 Weibo Data Collection

In order to depict the interest distribution of Chinese virtual community through the analysis of weibo content, we need to collect weibo feeds that mentioned any of the listed states. As discussed in chapter 3.2, using Sina API to access the specific data we need for this project is not possible. An alternative method to obtain the data is using a web crawler to harvest data from web page itself.

Data Source

Sina Weibo provides normal user search function in all its client platforms. Take the web client as an example. With the general search function, the weibo server will return microblogs, users, as well as news articles, contained the keyword you put in. The user could

also access the high-level search setting for customized specific search. The setting box offers five search attributes: Keyword, Type (all, hot, original created, from the user I followed, authenticated user, media), Contains (all, contains pictures, contains video, contains music, contains link), Time (from when to when), User location (province, city). For every operation, the server will return up to 50 pages, 20 weibo per page result based on the input.

Figure 3.2: Setting box of Weibo high-level search function

Theoretically, the page source contains all the content we see on a page, after downloading the HTML source we could analyze and extract useful data from it. However, most of the well-developed dynamic websites nowadays use AJAX techniques.

AJAX is the abbreviation of Asynchronous JavaScript and XML. It enables applications to receive asynchronous data from web servers in the background without interfering with the display of the existing page. (W3Schools, 2009)

Sina web client adds 20 weibos in the results page through AJAX. This technique ensured user a smooth and efficiency usage, but make it harder to collect data. Because those dynamic page contents were loaded by AJAX from JSON file that the server sent, then rendered in the browser, the HTML source does not contain those weibo data. But that does

not mean that we cannot access data by this way. Since Sina weibo has more than one sites, we should consider multiple possibilities and find the easiest way to obtain JSON form data.

In recent years, to improve user experience, big companies usually will build multiple localized subsites in different regions or continents, and specially modified clients for different platforms (e.g. Android client, IOS client, WP client). Sina Weibo also starts several subsites like hk.weibo.com for Hong Kong, sg.weibo.com for Singapore, us.weibo.com for North American. In the meantime, Sina still keeping the m.weibo.cn site for smartphone browser user.

The mobile version site only realizes a small part of functions compare to PC Web version, but the general search function is kept. Much like Web version, through m.weibo.cn, the search operation will return weibos from most recent post to older, 10 weibos in each page, and when user scroll down, new pages will be loaded from JSON files— which can be accessed by HTTP request.

Create web crawler

In chapter 3.2 we have mentioned the three steps of creating a web crawler. The first step is analysis the URL, find the request format. The JSON data we harvested was received from the following example URL (input word is “德国”, Germany):

```
https://m.weibo.cn/api/container/getIndex?type=all&queryVal=%E5%BE%B7%E5%9B%BD&featurecode=20000320&luicode=10000011&lfid=106003type%3D1&title=%E5%BE%B7%E5%9B%BD&containerid=100103type%3D1%26q%3D%E5%BE%B7%E5%9B%BD&page=1
```

Figure 3.3: JSON data URL

According to the RFC 1738 published in 1994:

"...Only alphanumerics [0-9a-zA-Z], the special characters "$-_.+!'()," [not including the quotes - ed], and reserved characters used for their reserved purposes may be used unencoded within a URL."*

To ensure the correct data transfer, any other character that is not included in this stander will be encoded. Here in this example, Chinese word “德国” has been transferred into “%E5%BE%B7%E5%9B%BD”. Therefore, an easier to read version of this two URL would be:

```
https://m.weibo.cn/api/container/getIndex?type=all&queryVal=德国&
featurecode=20000320&luicode=10000011&lfid=106003type%3D1&title
=德国&containerid=100103type%3D1%26q%3D 德国&page=
1%26q%3D%E5% BE%B7%E5%9B%BD&page=1
```

Figure 3.4: Decoded URL

Obviously, the key parameters are the user input content—query value, (here is “德国”) and the page number in the end (here is 1). Then we can follow this structure to make our URLs.

3.5.3 Data Formatting

Data structure

The harvested raw data was in JSON form. For each page, the structure of the data was as follows:

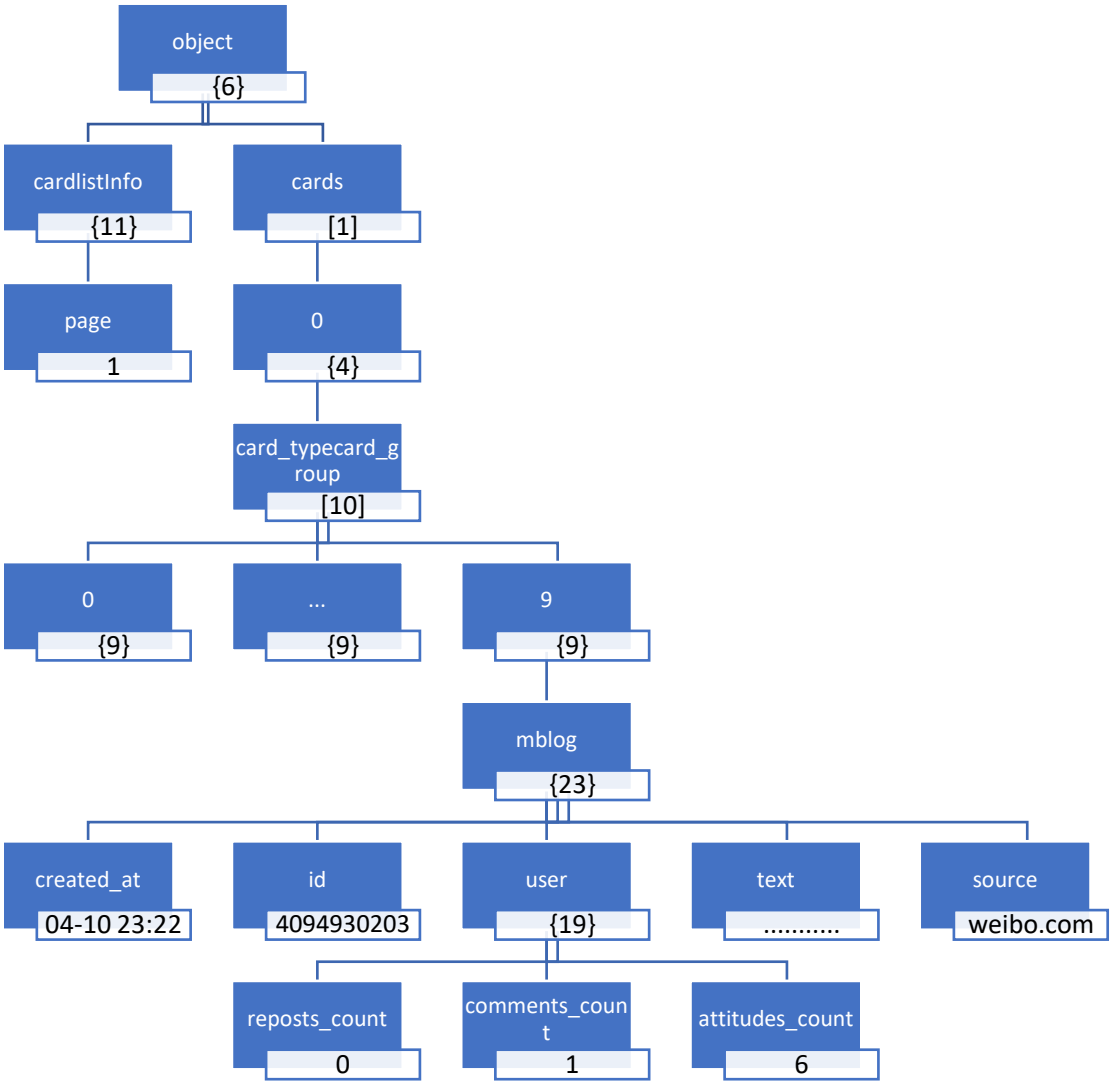


Figure 3.5: The general structure of obtained JSON data for each page

An example of each record is as follows:


```
        {
        "card_type":9,
        "card_type_name":"微博",
        "itemid":"seqid:455381132|type:2|t:|pos:1-0-1|q:不丹
|ext:&mid=4098377753366453&",
        "actionlog":{
                },
                "mblog":{
                "created_at":"04-19 22:41",
                "id":"4098377753366453",
                "mid":"4098377753366453",
                "idstr":"4098377753366453",
                        "text":"【亚东绝美风景】亚东，系藏语又称卓木，亚东，它是一
                个边境小城，一个小城竟与不丹、锡金、印度接壤。但是这里的边贸并
                没完全开放，不像樟木一样，那么热闹，但是这里的公路却比樟木好很
                多，这座漫山遍野全是杜鹃花的边陲小城亚东就在这种奇异的融合中宁
                静安祥地微笑着。 ",
                "textLength":252,
                "source":"微博 weibo.com",
                "favorited":false,
                "thumbnail_pic":"",
                "bmiddle_pic":"",
                "user":{
                        "id":5517515437,
                        "screen_name":"张蹦米光",
                        "profile_image_url":" .....kaa0.jpg",
                        "profile_url":"http:\/\/m.weibo.cn\/u\/5517515437?uid=..... ",
                        "statuses_count":377,
                        "verified":false,
                        "verified_type":-1,
                        "description":"",
                        "gender":"m",
                        "mbtype":0,
                        "urank":9,
                        "mbrank":0,
                        "follow_me":false,
                        "following":false,
                        "followers_count":102,
                        "follow_count":199,
                        "cover_image_phone":"…….jpg"
                        },
                "reposts_count":0,
                "comments_count":0,
                "attitudes_count":0,"
                "isLongText":false,
                },
```


Figure3.6: An example of a weibo record

For this project, the following attributes were collected:

- Weibo ID
- Weibo create time
- Repost count
- Weibo post source
- The user screen name
- The user ID
- The user gender
- The user follower count
- Weibo content

The JSON data is easy to read with any programming language for its organized structure. Python was used in this project.

```
itemID = thisdata['cards'][0]['card_group'][i]['mblog']['id']
itemCreat = thisdata['cards'][0]['card_group'][i]['mblog']['created_at'].encode("UTF-8")
itemRepostCount = thisdata['cards'][0]['card_group'][i]['mblog']['reposts_count']
itemSource = thisdata['cards'][0]['card_group'][i]['mblog']['source']
itemUser = thisdata['cards'][0]['card_group'][i]['mblog']['user']['screen_name']
itemUserID = thisdata['cards'][0]['card_group'][i]['mblog']['user']['id']
itemUserGender = thisdata['cards'][0]['card_group'][i]['mblog']['user']['gender']
itemUserFollower = thisdata['cards'][0]['card_group'][i]['mblog']['user']['followers_count']
itemText = thisdata['cards'][0]['card_group'][i]['mblog']['text']
```

Figure 3.7: Extract values from JSON structured metadata

Data processing

Before data can be imported into the database, there were several problems need to be solved.

- The value of create_at does not fit the standard data and time format.

- The weibo content may include non-Nature language HTML parts.
- Emoji appeared not just in the content but also username.
- Some weibo content contains series of repost @username.
- VIP user can modify their “source” tag, causing a problem for platform classification.

The first problem is caused by optimized time display. To give the user a better understanding, Sina modified the time tag with three ways. Weibo posted within a short time will be labeled with a format like “30 secs ago”, and weibo posted earlier this day will be labeled with format “Today 20:00”. Older weibo will be labeled with month-date format like “4-15 20:00”.

Figure 3.8: Three different formats of time display

To normalize the created time of each weibo, especially those with “30 secs ago” format, a calculation based on the timestamp of this request is used. When the JSON data is queried, the returned metadata will include this information state when exactly this request happened. This information is the value of "start time" tag. Use this timestamp as a foundation, the corresponding creates time for all weibo in this page could be achieved easily. Consider the future work with MySQL database, all-time value should convert into MySQL DATETIME standard form.

According to MySQL 5.7 Reference Manual:

The DATETIME type is used for values that contain both date and time parts. MySQL retrieves and displays DATETIME values in 'YYYY-MM-DD HH: MM: SS' format. The supported range is '1000-01-01 00:00:00' to '9999-12-31 23:59:59'.

Time will be normalized into “YYYY-MM-DD HH: MM: SS” format. While the metadata only has an accuracy of a minute, the second of old weibo with an original format like “4-15 20:00” will be complement with “00”.

Under the “text” tag is the full content of one weibo. Which means it may also include links, pictures, special emphasizing formats.

```
<a class='k' href='http://m.weibo.cn/k/爱她定制旅行?from=feed'>#爱她定制旅行#</a><br/>【不丹】<br/>看看不丹的饭有没有和中国很像，酸辣为主，很像川西北高原的餐饮习惯。因为蔬菜种类稀少，不丹人无辣不欢，把辣椒当蔬菜，而不是调料。<br/>深夜发吃，有没有馋到你<span class=\"url-icon\"><img src=\"//h5.sinaimg.cn/m/emoticon/icon/default/d_chanzui-ad3f4f182c.png\" style=\"width:1em;height:1em;;\"></span>
```

Figure 3.9: Content with irrelevant data

With the help of Beautiful Soup (a Python library for pulling data out of HTML and XML files), it is not hard to cut out the HTML code. But like what has been mentioned above, emoji, special Chinese punctuation, emoticons exist. In some situations, they are crucial, for example, when we need to analysis the textual emotion of a weibo, emoji could give us a direct hint. However, for this project, the use of emoji and emoticon are not considered as a highly-relent topic. Therefore, nothing specific will be conducted for them, only a rough filtration to remove some of the obvious emoticons for a clearer textual content. As for multiple @ mentions, the structure is always like “@UserTag1:……//@UserTag2:……//”. Pattern matching is used to remove this redundancy..

Sina Weibo is a highly-commercialized product, one big part of its revenue comes from “VIP account” service. One privilege of VIP user is that they can modify the “source” tag. Normally, the “source” tag is generated automatically based on the device and the client. If user post weibo through Sina Weibo mobile client, the default value is identical to the model of this device, for example: “iPhone SE”,” OPPO R9”, “Honour P9”. While, if this weibo is “shared” from another website, the value will be created based on the source, for instance: “Sina Daily News”, “Instagram”,” QQ Music”. Now with the VIP privilege, those users may alter the default value according to personal interest, like “My pure golden iPhone 6”, in this case, those records had been processed and kept only the standard part.

To help to understand the general popularity degree for a variety of hardware, a calculation based on the brand has been conducted. For this calculation, all iPhone models were considered as “iPhone”, same for others, all Galaxy models were calculated under “Samsung”, all Honour models under “Huawei”. The output was stored separately as a new attribute “source2”.

3.5.4 Keywords Extraction

A weibo contains a maximum of 140 characters, and on average the number of characters of a normal text is shorter than that. For this reason, each record had been extracted with only 5 keywords, the number was defined as a parameter (see figure 4.10).

```
jieba.analyse.extract_tags(sentence, topK=20, withWeight=False, allowPOS=())
```

Figure 3.10: Jieba keyword extraction function and its parameters

Here “sentence” is the text to be extracted. “topK” defines how many keywords with the highest TF/IDF weights will be returned. The default value is 20. “withWeight” tells whether return TF/IDF weights with the keywords. The default value is False. “allowPOS” defines filter words with which POSs are included. Empty for no filtering. The output was also stored separately as a new attribute “tags”.

3.5.5 Store Data in MySQL Database

To ensure an effective search and select, all cleaned and structured data was imported into the database. After the last step in the previous section, the cleaned weibo record is organized in one data line. When we load data from TXT file into MySQL database, each line should be recognized as one row. Since MySQL database use tab as separator, after we define the stop indicator of each line, the attributes data we organized in the previous step could be broken into columns automatically. Here the stop indicator is identical to line feed '\r\n'.

```
CREATE TABLE wbdata (
        creat_at DATETIME NOT NULL,
        keyword VARCHAR(30) NOT NULL,
        wid BIGINT NOT NULL,
        repost INT NOT NULL,
        platform CHAR(35) ,
        platform2 CHAR(35) ,
        user CHAR(45) NOT NULL,
        uid INT NOT NULL,
        follower INT NOT NULL,
        gender VARCHAR(2),
        text VARCHAR(500),
        tags VARCHAR(500)
    ) ;
```

Figure 3.11: Table information. Include 12 attributes data column

MySQL is a fast, easy-to-use open-source relational database management system (RDBMS). In the past decades, MySQL has become one of the most popular database systems. It still maintains its excellent performance even with huge amount of data. For this project, we chose MySQL out of the following common considerations:

1. MySQL is released under an open-source license. It is free to use.

2. MySQL uses a standard form of the well-known SQL data language.

3. MySQL works on many operating systems and with many languages.

4. MySQL works very quickly and works well even with large datasets.

5. MySQL supports large databases, up to 50 million rows or more in a table. The default file size limit for one table is up to 4GB.

The overall table contains all those 700 thousand records, but to achieve fast search, we need to separate table. The study period is approximately three months, preparing search based on time, records have been divided every 7 days by their created time and stored into

separate tables. For selection operated for each state, records have been stored into 195 individual tables based on keyword.

3.5.6 Visualization

There are a lot of different methods to visually display data with charts in the front end. Successful APIs exists, while for this project D3.js has been chosen for its broader control of visualization details.

D3.js is a Data-Driven-Documents visualization library for creating interactive and dynamic graphics and charts. The main character of D3 is that it allows the user to bind arbitrary data to a Document Object Model (DOM), and then apply data-driven transformations to the document. For instance, the user can use D3 to generate an HTML table from an array of numbers. Or, use the same data to create an interactive SVG bar chart with smooth transitions and interaction. For this study, D3 full fit the very simple ability we want: Fast, into detail design, easily cooperates with CSS.

To display all the data we obtained, several plots were created and linked. First is a bar chart display data flow rank, and a map shown the location of the corresponding state. The other parameters like platform information was displayed also with a bar chart. Next, a pie chart demonstrates the sex ratio of all those users, a word cloud presents words with top 120 highest frequencies, a line chart with the date as X-axis and data amount as Y-axis shows data amount changing of each state. The additional two text boxs for weibo with top 5 highest repost counts and user with top 10 highest follower counts are also added in for this web application.

The interactive design is achieved by linking all the charts as together and the chart themselves as parameter indicator. This interactive design basically means when user clicks any chart element, the selector is changed, the parameter of the chart changed accordingly. So the other chats will refresh the views based on this new input. For example, under the initial page, if user clicks “Japan” in the first histogram, the platform rank chart will display the most popular platforms calculated only from those webs which contain keyword “Japan”. Our web application is hosted locally by Flask, the data displayed on the page are queried from MySQL database through Python.

4. Result Analysis

In this part, we will discuss the results we got from this project, which can answer the aim of this research by displaying the connection between Chinese virtual community and other states. For this study, data acquisition period started in the beginning of May till the end of July, in total 94 days (1 May through 2 July 2017). Totally around 700,000 weibo have been harvested. Through the analysis, we will come to some discussions and conclusions.

Mapping the textual geolocation-based Weibos

The data amount of each state shall indicate "the level of interest" or "the level of concern". Through the visualization it is not difficult to find out that real space distance still matters. Although people can cross geographical distance with low costs through the internet, the geographically closer states still intent to show stronger connection and influence.

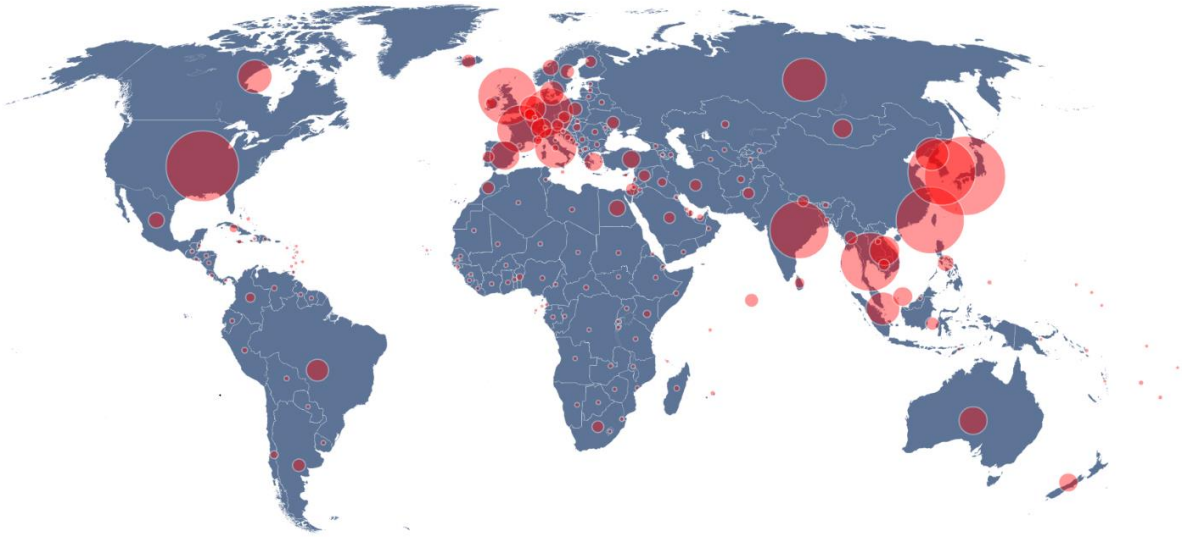


Figure 4.1: The data amount under each state name during this study period

According to the data, the most commonly mentioned 30 states are:

Japan, USA, Taiwan, South Korea, Thailand, UK, India, Germany, France, Russia, Italy, Canada, Singapore, North Korea, Spain, Australia, Netherlands, Denmark, Brazil, Switzerland, Malesia, New Zealand, Greece, Mongolia, Turkey, Philippine, Egypt.

Among them, Asian countries take almost half of the portion, especially in the top 10.

Figure 4.2: Most commonly mentioned 30 states according to the continent

The reason behind this phenomenon could be discussed from various aspects. For example, closer states intend to have a stronger impact in the physical world. A typical example is North Korea. According to the data, when mentioned North Korea, a lot of weibos were talking about military movements and politically related news. People care more about a specific neighbor state due to national interest.

While, the second reason may be popular culture emerging enhanced by culture similarity. Compare the word cloud of Japan, South Korea and India (a high culture similarity with China) with UK, Germany and France (lower culture similarity with China), we can notice the difference of topic range. For higher culture similarity states, weibo user discussed more about popular culture (e.g., Japanese animation, India movie, South Korea pop music) while for the lower culture similarity states, the discussion lean to general news or specialized interests (e.g., Football and political talks under Germany, Terrorism attack and Brexit related discussing around UK, travel, presidential election in France).

The space-time distance has less influence in Cyberspace

The previous conclusion shown how real space distance influences the virtual world. However, the connect limitation caused by geographical distance is not as heavy as in physical space, almost every state can be reached in Cyberspace.

We can notice that there is also huge data flow for far away regions, and every region in the retrieve list has been mentioned. For example, the second commonly mentioned country is American, even though the distance in real space is far. And the top 5 less mentioned countries are Saint Vincent And The Grenadines, Sao-Tome-Principe, Antigua-Barbuda, Saint Kitts And Nevis, and Comoros. But even for them, still 5, 16, 22, 22, 24 records were respectively obtained during our study period.

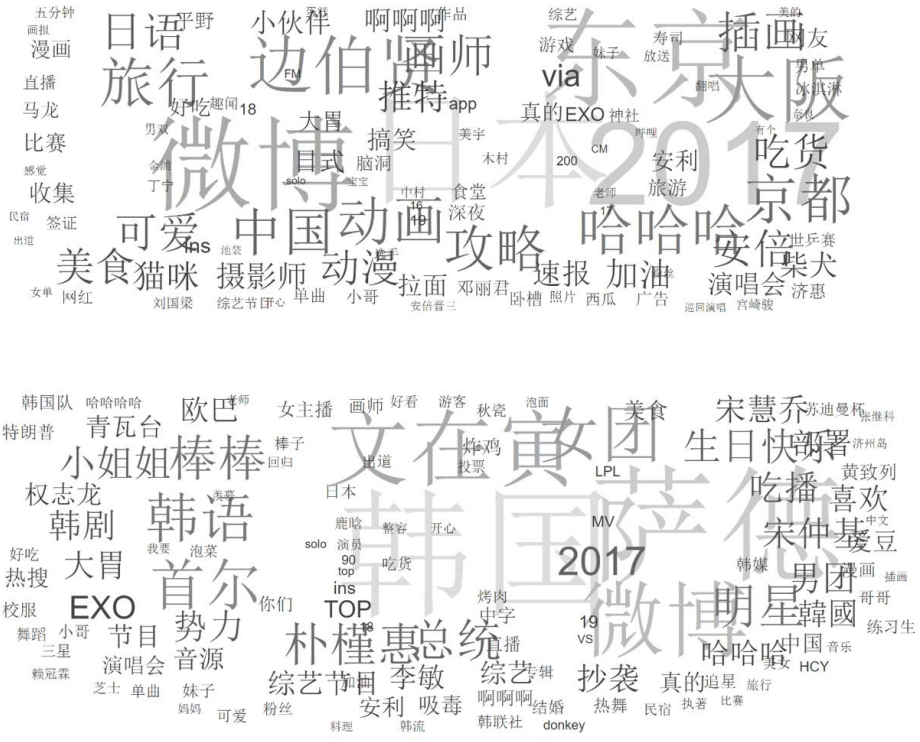


Figure 4.3: Word clouds of Japan and South Korea during the study period

Figure 4.4: Word clouds of British and Germany during the study period

The strength of connection is consistent with the level of economic development

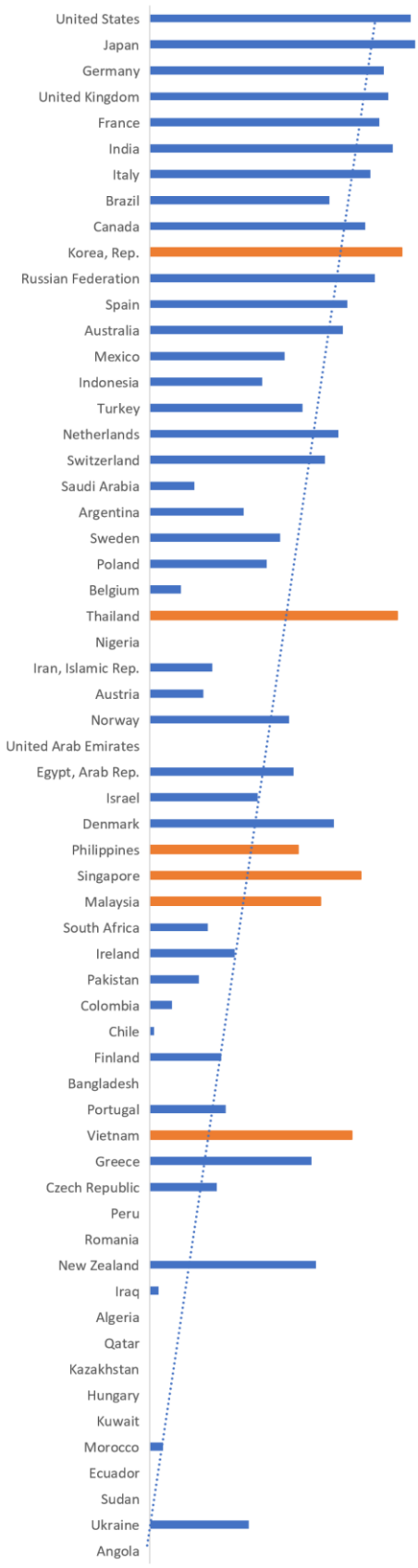


Figure 4.5: Top 60 country based on 2016 GDP and their data flow rank

Take 2016 global GDP ranking as represent of the level of economic development, and analysis the dependency between the strength of data flow and state economy. As shown in figure 31, the vertical order is top 60 states based on the global GDP ranking of 2016, the higher the country, the better the economy. The horizontal length of the bar indicates the data flow rank of the state, longer bar means higher data amount rank, no bar means the data rank of the state is not inside top 60. From figure 31, the data flow of a country generally consistent with its GDP, especially for those on top. Some outstanding bars (colored in orange) fit are states that geographically or culturally close to China (include Korea, Thailand, Philippine, Singapore, Malaysia, Vietnam).

The temporal trend was influenced heavily by news report

By monitoring the temporal data trend, we noticed that data flow concerns a state would be heavily influenced by breaking news.

Figure 4.6 depicts the data flow change of American and Japan during the study period. We can see the uneven distribution. The four synchronous drops were caused by server connection problem. While the following figure 4.7 depicts the data flow trend concerns Denmark. The red line marks the same drop as in figure 4.6, it happened on 15 May. After leaving out this drop,

an unusual peak can be seen in the middle of May, and the trend slowly decreased back to its average around the beginning of June.

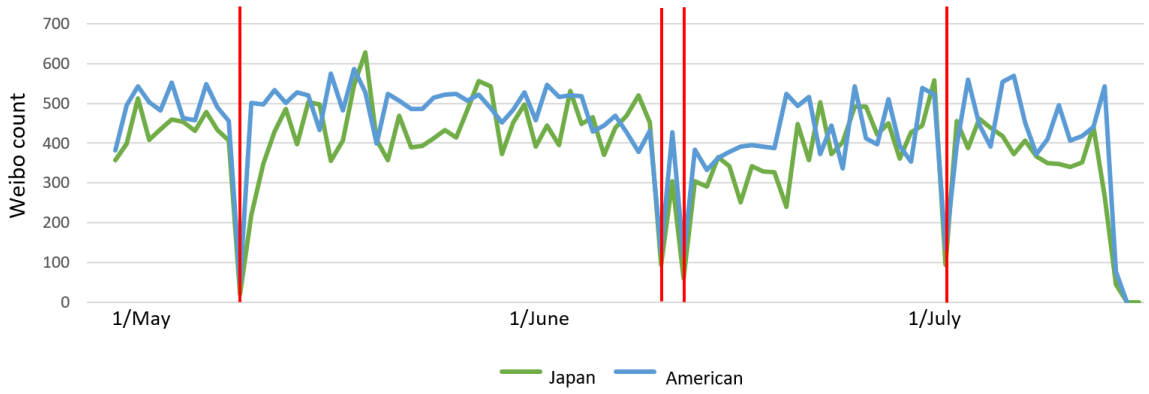


Figure 4.6: Data flow of "Japan" and "American", with four same sharp drops, possibly due to bad connection or server updates.

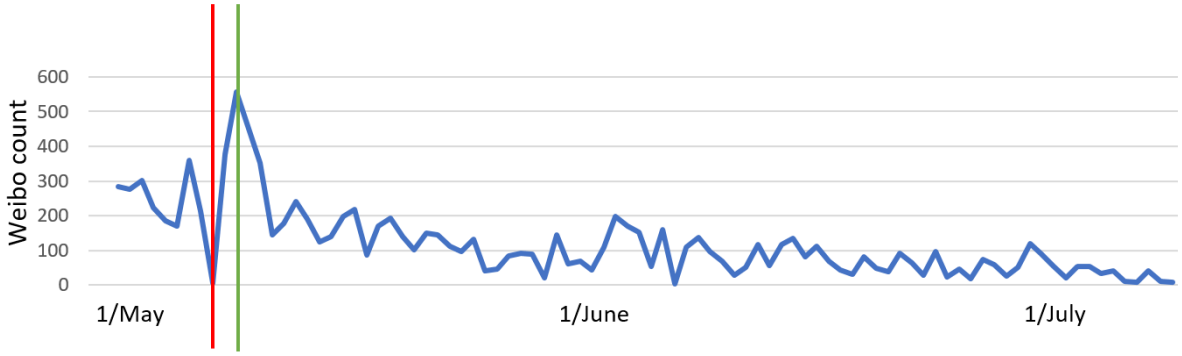


Figure 4.7: The weibo samples count which contains keyword "Denmark" from 2 May 2017 to 28 July 2017

This peak coincides with the news that Denmark is suffering from the invading of Oyster, the ambassador invites Chinese people to travel to Denmark help locals "eat this problem out", and a young pop idol Qianxi was assigned as the spokesperson of the Danish tourism image. This news broke into his fans group very quickly and been reposted a lot in a short time. From the generated word cloud we can see the same pattern. The biggest words consist of two parts of information, the first part, this idol's name, his band name, his nickname, also "spokesperson", and the second part is "Denmark", "oyster", "travel".

The efficiency of social media has made any news easy to share and trend. This feature could be used to detect breaking news and find the popular topics.

Figure 4.8: The frequency word cloud summarizing keyword tags extracted from weibo from 5 May 2017 to 15 May 2017. It shows the 120 most frequently appeared keywords in that data set. The size is proportional to frequency

Figure 4.9: The comparison of word clouds between male and female user related to Japan

Gender preferences are different

The male and female user could have a different focus when talking about a state. Figure 37 shows the male and female word clouds concern Japan. The lower female user word cloud has a clear preference of topics about vacation and food (e.g., “travel”, “noodles”, “VISA”, “cake” “watermelon”, “Tokyo”, “photo”). By contrast, male user showed a more complex topic preference involve politics, sport, and animation (e.g., “competition”, “team”, “comic book” “prime minister”, “table tennis”, “robot”).

5. Conclusion

5.1 Summary

The growing use of social media has stimulated online public discussion around international news and issues. Textual geoinformation like city name, state name pervades social media platforms. This type of information can be harvested and used to reveal international connections. In this paper, we used China as a study case, demonstrated the workflow of how we can illustrate the international relations from social media data, and how we can bound the connections to real geographic space and explain them from a spatiotemporal perspective.

To begin with, we answered the first research question about which weibos would be collected and how can we collect them. Based mainly on UN member states we first derived a list of 195 states' name in Chinese, and encoded them with their commonly used abbreviation in social media. Then, weibos which contained those states' textual reference were collected through a web crawler. After that, from the meta data we gained, attributes were extracted— state name, user screen name, user gender, user follower, platform, post time, repost count and the textual content.

The second research question was about how can we process the textual content of weibo so that we can analysis the information in those massive records. We first introduced some basic natural language processing (NLP) functions and the commonly used algorithms to realise them. Then stated the consideration when comparing different NLP tools. And with the help of the Jieba Chinese segmentation library, we extracted 5 key words from each weibo content.

Finally, we discussed the feature of each attribute and the advantage of various visual techniques, thereby answering the last research question: How to present those different datasets with the suitable visual techniques. In the end, through an integrated visual analysis, we came to some interesting conclusions.

First, the study found out that cyberspace increased the distance between people's information exchange. States in a faraway location or with weak diplomatic relations would not necessarily lead to a lower level of concern. This indicates that the existence of cyberspace reduces the influence of real space distance. But we still need to note that, although the existence of cyberspace made information exchange effortless, the physical space still has its impact on shaping a virtual community. This could be summarized from the pattern: China's geographical neighboring states which have higher cultural similarity background intend to have a higher probability of being discussed.

Second, the study found out that the level of attention is generally consistent with the state's economic situation. This result further illustrates that the role of physical space continues its impact in the Cyberspace. But this does not mean that the Cyberspace is a simple projection of the real space. Not like the generally stable relationship between two countries, the connection between virtual communities could be heavily influenced by breaking news. Denmark was picked as an example to present this situation. However, this characteristic can be used as a method to detect many important events.

5.2 Outlook

During the study, we have answered the research questions and achieved the research aim. However, there exist some alternative possibilities that may prove to be better solutions. Thus, we will discuss some of the drawbacks in our method, and consider further improvement as part of the future work.

The first thing that needs to be pointed out is that, the boundary and character of a country, or a state, is always different in Cyberspace than in real word. Especially when we are dealing with user-generated information rather than serious political definitions. We may find it hard to leave out or add in some objects that have a controversial identity. Like in this study, “Taiwan” was also included though it is not recognized by UN as a country. This is because Taiwan issue plays an important role in Chinese community that give it a special value to be considered. In contrast, other controversial political entities and regions with less importance in Chinese community were left out, like we did not consider Northern Ireland

separately from UK, nor did we add Crimea in the list. The consideration should be different with regard to which state we analyse.

Second, different social media platforms have their own policy regarding open API. Using a web crawler is not always the best choice in most of the case. The data acquisition procedure of using web crawler is not as stable as using API. This is seen with the three transmission drops in our data flow. For future work, a stable and comprehensive data acquisition process through official API is preferred.

The third point is that we performed only basic NLP functions and extracted sentence's key words in this study. We would expand the semantic study to cover sentiment analysis. This would provide us with a deeper understanding not only on how strong is the connection in virtual community, but also to which emotional dimensions is the connection.

Based on this thesis, we expect some future work to conduct this social media based international connection studies to have a focus on other state, not just the Chinese community. To perform similar study on applications that have multicultural user group like Twitter, will encounter two additional problems we did not experience here— first is how to derive a batch of users/tweets that come from a same country, second is how to handle a community that have more than one commonly used language. Once the problems are answered, we could expect a more comprehensive and unique understanding on multilateral connections, the reason behind events, trends, ideas, the cause of stereotypes and changing of national image, from a people's, not government, perspective.

6. References

Ahn J, Taieb-Maimon M, Sopan A, et al. Temporal visualization of social network dynamics: prototypes for nation of neighbors[J]. Social computing, behavioral-cultural modeling and prediction, 2011: 309-316.

Berners-Lee T, Masinter L, McCahill M. Uniform resource locators (URL)[R]. 1994.

Cambria E, White B. Jumping NLP curves: A review of natural language processing research[J]. IEEE Computational intelligence magazine, 2014, 9(2): 48-57.

Chowdhury G G. Natural language processing[J]. Annual review of information science and technology, 2003, 37(1): 51-89.

Cranshaw J, Schwartz R, Hong J I, et al. The livehoods project: Utilizing social media to understand the dynamics of a city[J]. 2012.

Donath J S. Identity and deception in the virtual community[J]. Communities in cyberspace, 1999, 1996: 29-59.

Erickson I. Geography and community: New forms of interaction among people and places[J]. American Behavioral Scientist, 2010, 53(8): 1194-1207.

Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences, 2002, 99(12): 7821-7826.

Goodchild M F. Citizens as sensors: the world of volunteered geography[J]. GeoJournal, 2007, 69(4): 211-221.

Graham M, Zook M. Visualizing global cyberscapes: Mapping user-generated placemarks[J]. Journal of Urban Technology, 2011, 18(1): 115-132.

Guohe F, Wei Z. Review of Chinese automatic word segmentation[J]. Library and information service, 2011, 55(2): 41-45.

Habibi M, Popescu-Belis A. Keyword extraction and clustering for document recommendation in conversations[J]. IEEE/ACM Transactions on audio, speech, and language processing, 2015, 23(4): 746-759.

Hanna R, Rohm A, Crittenden V L. We're all connected: The power of the social media ecosystem[J]. Business horizons, 2011, 54(3): 265-273.

Hiltz S R, Wellman B. Asynchronous learning networks as a virtual classroom[J]. Communications of the ACM, 1997, 40(9): 44-49.

Hu X, Li H, Bao X. Urban population mobility patterns in Spring Festival Transportation: Insights from Weibo data[C]//Service Systems and Service Management (ICSSSM), 2017 International Conference on. IEEE, 2017: 1-6.

Ickea I, Sklara E. TECHNICAL REPORT Visual Analytics: A Multi-faceted Overview[J]. 2009.

Jensen Schau H, Gilly M C. We are what we post? Self-presentation in personal web space[J]. Journal of consumer research, 2003, 30(3): 385-404.

Jurafsky D, Martin J H. Speech and language processing[M]. London: Pearson, 2014.

Kaplan A M, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media[J]. Business horizons, 2010, 53(1): 59-68.

Keim D A, Mansmann F, Schneidewind J, et al. Challenges in visual data analysis[C]//Information Visualization, 2006. IV 2006. Tenth International Conference on. IEEE, 2006: 9-16.

Keim D A. Information visualization and visual data mining[J]. IEEE transactions on Visualization and Computer Graphics, 2002, 8(1): 1-8.

Keim D A. Visual exploration of large data sets[J]. Communications of the ACM, 2001, 44(8): 38-44.

Leetaru K, Wang S, Cao G, et al. Mapping the global Twitter heartbeat: The geography of Twitter[J]. First Monday, 2013, 18(5).

Liddy E D. Natural language processing[J]. 2001.

Liu Y, Sui Z, Kang C, et al. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data[J]. PloS one, 2014, 9(1): e86026.

Liu Z, Chen X, Zheng Y, et al. Automatic keyphrase extraction by bridging vocabulary gap[C]//Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2011: 135-144.

Luo Y. The Internet and agenda setting in China: The influence of online public opinion on media coverage and government policy[J]. International Journal of Communication, 2014, 8: 24.

Mathioudakis M, Koudas N. Twittermonitor: trend detection over the twitter stream[C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010: 1155-1158.

Ng H T, Low J K. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based?[C]//EMNLP. 2004: 277-284.

Nguyen T, Kan M Y. Keyphrase extraction in scientific publications[J]. Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, 2007: 317-326.

Piller F T, Vossen A, Ihl C. From social media to social product development: the impact of social media on co-creation of innovation[J]. 2011.

Ridings C M, Gefen D. Virtual community attraction: Why people hang out online[J]. Journal of Computer‐Mediated Communication, 2004, 10(1): 00-00.

Sacha D, Stoffel A, Stoffel F, et al. Knowledge generation model for visual analytics[J]. IEEE transactions on visualization and computer graphics, 2014, 20(12): 1604-1613.

Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513-523.

Saville-Troike M. The ethnography of communication: An introduction[M]. John Wiley & Sons, 2008.

Sawyer R, Chen G M. The impact of social media on intercultural adaptation[J]. 2012.

Schreck T, Keim D. Visual analysis of social media data[J]. Computer, 2013, 46(5): 68-75.

Shirky C. The political power of social media: Technology, the public sphere, and political change[J]. Foreign affairs, 2011: 28-41.

Simoff S J, Böhlen M H, Mazeika A. Visual data mining: An introduction and overview[M]//Visual Data Mining. Springer Berlin Heidelberg, 2008: 1-12.

Stefanidis A, Cotnoir A, Croitoru A, et al. Demarcating new boundaries: mapping virtual polycentric communities through social media content[J]. Cartography and Geographic Information Science, 2013, 40(2): 116-129.

Tsou M H, Yang J A, Lusher D, et al. Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election[J]. Cartography and Geographic Information Science, 2013, 40(4): 337-348.

Witten I H, Paynter G W, Frank E, et al. KEA: Practical automatic keyphrase extraction[C]//Proceedings of the fourth ACM conference on Digital libraries. ACM, 1999: 254-255.

Xia C, Schwartz R, Xie K, et al. Citybeat: Real-time social media visualization of hyper-local city data[C]//Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014: 167-170.

Xue N. Chinese word segmentation as character tagging[J]. Computational Linguistics and Chinese Language Processing, 2003, 8(1): 29-48.

Yu L, Asur S, Huberman B A. What trends in Chinese social media[J]. arXiv preprint arXiv:1107.3522, 2011.

Zhang M, Zhang Y, Che W, et al. Type-Supervised Domain Adaptation for Joint Segmentation and POS-Tagging[C]//EACL. 2014: 588-597.

Zook M, Graham M, Shelton T, et al. Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake[J]. World Medical & Health Policy, 2010, 2(2): 7-33.

袁靖华. 微博的理想与现实——兼论社交媒体建构公共空间的三大困扰因素[D]. , 2010.

Monthly active user of Facebook at:

https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Monthly active user of Twitter at:

https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

We are Social:

https://wearesocial.com/special-reports/digital-in-apac-2016

Monthly active user of Facebook from mobile platform at:

https://www.statista.com/statistics/346195/facebook-global-mobile-dau/

2016 Sina annual report at:

http://tech.sina.com.cn/i/2017-02-23/doc-ifyavvsh5976843.shtml

“City mood” project by Beihang Interest Group on SmartCity (2015) at:

http://urbandataview.com/citymood.html

The 2016 Weibo user report at:

http://data.weibo.com/report/reportDetail?id=346

Twitter, 2009. “Twitter Blog: Location, location, location” (20 August), at:

http://blog.twitter.com/2009/08/location-location-location.html

UCI machine learning repository.

http://archive.ics.uci.edu/ml

Interactive Data Visualization with D3.js, DC.js, Python, and MongoDB at:

http://adilmoujahid.com/posts/2015/01/interactive-data-visualization-d3-dc-python-mongodb/

The global economy by GDP at:

https://howmuch.net/articles/the-global-economy-by-GDP

7. Appendix

1. List of states

Selected states	Official name in Chinese	Common name in Chinese (if different)
Afghanistan	阿富汗	
Albania	阿尔巴尼亚	
Algeria	阿尔及利亚	
America	美国	
Andorra	安道尔	
Angola	安哥拉	
Antigua-Barbuda	安提瓜和巴布达	
Arab	阿拉伯联合酋长国	阿联酋
Argentina	阿根廷	
Armenia	亚美尼亚共和国	亚美尼亚
Australia	澳大利亚	
Austria	奥地利	
Azerbaijan	阿塞拜疆共和国	阿塞拜疆
Bahamas	巴哈马	
Bahrain	巴林	
Bangladesh	孟加拉国	孟加拉
Barbados	巴巴多斯	
Belarus	白俄罗斯	
Belgium	比利时	
Belize	伯利兹	
Benin	贝宁	
Bhutan	不丹	
Bolivia	玻利维亚	
Bosnia	波斯尼亚和黑塞哥维那	波斯尼亚
Botswana	博茨瓦纳	
Brazil	巴西	
Brunei	文莱	
Bulgaria	保加利亚	
Burkina-Faso	布基纳法索	
Burundi	布隆迪	
Cambodia	柬埔寨	
Cameroon	喀麦隆	
Canada	加拿大	
Central-Africa	中非	
Chad	乍得	
Chile	智利	
Colombia	哥伦比亚	
Comoros	科摩罗	

Congo	刚果（布）	刚果布
Congo-Kinshasa	刚果（金）	刚果金
Costa Rica	哥斯达黎加	
Cote-D-Lvoire	科特迪瓦	
Croatia	克罗地亚	
Cyprus	塞浦路斯	
Czech	捷克	
Denmark	丹麦	
Djibouti	吉布提	
Dominica	多米尼克	
Dominican Republic	多米尼加	
Ecuador	厄瓜多尔	
Egypt	埃及	
El Salvador	萨尔瓦多	
England	英国	
Equatorial-Guinea	赤道几内亚	
Eritrea	厄立特里亚	
Estonia	爱沙尼亚	
Ethiopia	埃塞俄比亚	
Fiji	斐济	
Finland	芬兰	
France	法国	
Gabon	加蓬	
Gambia	冈比亚	
Georgia	格鲁吉亚	
Germany	德国	
Ghana	加纳	
Greece	希腊	
Grenada	格林纳达	
Guatemala	危地马拉	
Guba<-Cuba	古巴共和国	古巴
Guinea	几内亚	
Guinea-Bissau	几内亚比绍	
Guyana	圭亚那	
Haiti	海地	
Holand	荷兰	
Honduras	洪都拉斯	
Hungary	匈牙利	
Iceland	冰岛	
India	印度	
Indonesia	印度尼西亚	
Iran	伊朗	
Iraq	伊拉克	
Ireland	爱尔兰	
Israel	以色列	
Italy	意大利	

Jamaica	牙买加	
Japan	日本	
Jordan	约旦	
Kazakhstan	哈萨克斯坦	
Kenya	肯尼亚	
Kirgizstan	吉尔吉斯共和国	吉尔吉斯
Kiribati	基里巴斯	
Korea	韩国	
Kuwait	科威特	
Laos	老挝	
Latvia	拉脱维亚	
Lebanon	黎巴嫩	
Lesotho	莱索托	
Liberia	利比里亚	
Libya	利比亚	
Liechtenstein	列支敦士登	
Lithuania	立陶宛	
Luxembourg	卢森堡	
Macedonia	马其顿	
Madagascar	马达加斯加	
Malawi	马拉维	
Malaysia	马来西亚	
Maldives	马尔代夫	
Mali	马里	
Malta	马耳他	
Marshall Islands	马绍尔群岛	
Mauritania	毛里塔尼亚	
Mauritius	毛里求斯	
Mexico	墨西哥	
Micronesia	密克罗尼西亚联邦	密克罗尼西亚
Moldova	摩尔多瓦	
Monaco	摩纳哥	
Mongolia	蒙古	
Montenegro	黑山	
Morocco	摩洛哥	
Mozambique	莫桑比克	
Myanmar	缅甸	
Namibia	纳米比亚	
Nauru	瑙鲁	
Nepal	尼泊尔	
New-Zealand	新西兰	
Nicaragua	尼加拉瓜	
Niger	尼日尔	
Nigeria	尼日利亚	
North-Korea	朝鲜	
Norway	挪威	

Oman	阿曼	
Pakistan	巴基斯坦	
Palau	帕劳	
Palestine	巴勒斯坦	
Panama	巴拿马	
Papua-New-Guinea	巴布亚新几内亚	新几内亚
Paraguay	巴拉圭	
Peru	秘鲁	
Philippines	菲律宾	
Poland	波兰	
Portugal	葡萄牙	
Qatar	卡塔尔	
Romania	罗马尼亚	
Russia	俄罗斯联邦	俄罗斯
Rwanda	卢旺达	
Saint Kitts And Nevis	圣基茨和尼维斯	
Saint Vincent And The Grenadines	圣文森特和格林纳丁斯	
Saint-Lucia	圣卢西亚	
Samoa	美属萨摩亚	萨摩亚
San Marino	圣马力诺	
Sao-Tome-Principe	圣多美和普林西比	
Saudi-Arabia	沙特阿拉伯	沙特
Senegal	塞内加尔	
Serbia	塞尔维亚和黑山共和国	塞尔维亚
Seychelles	塞舌尔	
Sierra-Leone	塞拉利昂	
Singapore	新加坡	
Slovak	斯洛伐克	
Slovenia	斯洛文尼亚	
Solomon Islands	所罗门群岛	
Somali	索马里	
South Sudan	南苏丹	
South-Africa	南非	
Spain	西班牙	
Srilanka	斯里兰卡	
Sudan	苏丹	
Surinam	苏里南	
Swaziland	斯威士兰	
Sweden	瑞典	
Switzerland	瑞士	
Syria	叙利亚	
Taiwan	台湾	
Tajikistan	塔吉克斯坦	
Tanzania	坦桑尼亚	
Thailand	泰国	

Timor	东帝汶	
Togo	多哥	
Tonga	汤加	
Trinidad-And-Tobago	特立尼达和多巴哥	
Tunisia	突尼斯	
Turkey	土耳其	
Turkmenistan	土库曼斯坦	
Tuvalu	图瓦卢	
Uganda	乌干达	
Ukraine	乌克兰	
Uruguay	乌拉圭	
Uzbekistan	乌兹别克斯坦	
Vanuatu	瓦努阿图	
Holy See	圣座	梵蒂冈
Venezuela	委内瑞拉	
Verde	佛得角	
Vietnam	越南	
Yemen	也门	
Zambia	赞比亚	
Zimbabwe	津巴布韦	