Exploring Human Activity Patterns Across Cities through Social Media Data

Master's Thesis

Vincent X. Gong

Exploring Human Activity Patterns Across Cities through Social Media Data

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK INFORMATION ARCHITECTURE

by

Vincent X. Gong
born in Jia Jiang

Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
http://wis.ewi.tudelft.nl

Exploring Human Activity Patterns Across Cities through Social Media Data

Author:	Vincent X. Gong
Student id:	4328000
Email:	vincent.gong7@gmail.com

Abstract

Cities are complex systems, a particular dimension largely contributing to the complexity of cities is detected in the multiplicity of the behaviour of people who inhabit. In the urban context, these behaviours can be inferred from human activities. Understanding the distinctions of these activities between cities is crucial. Previous works carry out such studies using surveys, censuses, interviews and onsite observations, which are rather costly, requiring extensive human or other resources. With the development of Internet and information technology, the geo-referenced social media networks which link human activities to specific places have been more and more popular. It, therefore, becomes a good candidate for studying individual human activities in the city. In this work, we explore to what extent can social media data be used for studying differences of individual human activities in an urban environment, considering impacts of three aspects: the choice of social media platforms, the user roles and demographics, and the various of Point of Interests (PoIs). To perform this study, we design an experiment to compare user activities on multiple social media platforms in Rotterdam and Shenzhen. A pipeline for crawling and enriching Weibo data is developed and encapsulated as an extension to the Social-Glass framework. A novel dataset containing data from four platforms (i.e. Twitter, Instagram, Foursquare and Weibo) in two cities for two weeks is created and analysed at a disaggregate level across several dimensions. Results indicate that, in the context of user activity study, Twitter and Weibo shows distinct differences for resident regarding local, general, and neutral topics particularly for young and mid-age users on Twitter, and for teen and young users on Weibo. While, Instagram shows evident differences for non-residents concerning international, specific, daily and emotional topics, especially for young and mid-age users. In the meantime, various PoIs provide significant differences in user activity.

Thesis Committee:

Chair:	Prof.dr.ir. Geert-Jan Houben, Faculty EEMCS, TUDelft
University supervisor:	Dr. Alessandro Bozzon, Faculty EEMCS, TUDelft
	Ir. Achilleas Psyllidis, Faculty of Architecture and the Built Environment, TUDelft
	Ir. Jie Yang, Faculty EEMCS, TUDelft
Committee Member:	Prof.dr.ir. A. van Timmeren, Faculty of Architecture and the Built Environment, TUDelft

Preface

The fun of this research is that it provides an opportunity for me, as a computer science student, from the data perspective to investigate the different activity patterns of people who are living cities with distinct culture, scale, language and so on. I was born in a tiny county in China, grew up in a small and mid-level city, and once worked in the metropolis. I also worked and studied in European cities. Different experiences in various urban environments inspired me with the mutual impacts of multiple factors in the urban context on the people's life. I am glad that this study is a good start to find answers.

Several people have contributed to this thesis project from academic and engineering perspective. First and foremost I offer my sincerest gratitude to my supervisors, Alessandro Bozzon, Achilleas Psyllidis, and Jie Yang. Alessandro Bozzon has supported me throughout my thesis with his patience and knowledge. He supervised me to make a solid logic structure of the study and provided crucial feedback. Achilleas Psyllidis has guided me through the whole process with his expertise in the urban field. He helped me in constructing the research structure and solving problems with providing clear and essential comments. Jie Yang has guided me in the structure of data analysis and provided me critical hits for solving problems in data crawling, measurements, and so on. A great deal of appreciation goes to Stefano Bocconi and Christiaan Titos Bolivar. Stefano Bocconi helped in solving difficulties in integrating the Weibo Crawler into Social-Glass framework. Christiaan Titos Bolivar helped me to understand the code and solve problems of the Social-Glass framework. Moreover, I would like to thank Professor Geert-Jan Houben for providing me the opportunity to work on this topic and for proposing ideas. Likewise, I would like to thank Professor A. van Timmeren for his participation in the thesis committee. Furthermore, I also wish to thank all of the participants. Without their support I would not be able to finish this research. Last but not least, I am very obliged to my parents and the whole family for providing me the opportunity to study and enjoy the life.

Vincent X. Gong
Delft, the Netherlands
April 7, 2016

Contents

List of Figures

Chapter 1

Introduction

Cities are complex systems by definition, inasmuch as they form more than the sum of the various systems that comprise them. ([1, 2]). A particular dimension, largely contributing to the complexity of cities, is detected in the multiplicity of individual and collective behaviours of the people who inhabit. These behaviours are influenced, to a certain extent, by the cultural norms guiding different social groups in the urban environment (Comunian, Roberta. [9]). While in the urban context, human activities can be used for inferring these human behaviours.

Understanding differences in these activities between cities is critical. Many previous works proposed approaches to study the differences of human activities in cities ([45, 51, 30]). Performing such studies typically require a huge capacity of information. In carrying out such studies, surveys, censuses, interviews, and on-site observations traditionally constitute the predominant methods for collecting data about the ways people use the city. Notwithstanding their high degree of accuracy and trustworthiness, these methods are rather costly, requiring extensive human or other resources. This further encumbers the performance of longitudinal studies.

With the develop of Internet and information technology, social media networks which allow people to create, share, or exchange information, ideas in virtual communities and networks (Buettner, Ricardo [5]) have been more and more popular than ever before. On social media networks (such as Twitter [1], Instagram [2]), people send posts consisting their expressions, profiles, and timestamps. A certain percentage of posts also contain geo-referenced information, i.e. the location where they are sent. The advantage of social media data, compared with other data sources, is that they are semantically rich information linking human activities (presented on social media) to specific places in a city. It, therefore, becomes a good candidate for studying individual human activities in the city.

Given that the differences in individual human activities influenced by cultural background observed from social media have not been studied yet, in this thesis, we explore whether data from social media can be used for studying differences in individual human activities in urban environments.

[1] https://www.twitter.com

[2] https://www.instagram.com

1.1 Research Question

Several research questions have been defined to guide this research. The main research question of this work is defined as:

RQ *To what extent can social media data be used for studying differences of individual human activities in urban environment?*

We choose to take two cities, Rotterdam in the Netherlands and Shenzhen in China, to maximise the differences in cultural background. Accordingly, we select different social media (the Twitter, Instagram, and Foursquare for the city of Rotterdam, and Weibo for the city of Shenzhen) to perform this research[3], with defining three sub-research questions as below:

1. How does *Choice of Social Media Platforms* impact on studying differences of human activities in urban environments?
2. How does the *User Individual Roles and Demographics* observed from social media impact on studying differences of human activities in urban environments?
3. How do various *Venues (Point of Interests)* impact on studying differences of human activities in urban environments?

1.2 Method(ology)

The goal of this study is to investigate the influence of factors of cities, social media platforms, data observed from social media, and to what extent they can be used for studying differences in human activities. To this purpose, we explore each impact of them in an experimental study.

The first step is to select different cities. It starts with two cities, and the candidate cities should be: (a) representative of different cultures; (b) characterised, to a certain extent, by mutual features so as to be mutually comparable; (c) different in terms of social media usage and platform popularity. To this end, we choose the city of Rotterdam in the Netherlands, and the city of Shenzhen in China, considering that Europe and China have significant cultural differences ([40], [20]). The detail motivation for choosing these two cities for the experiment will be described in Chapter 3.

The choice of social media platforms in this two cities is essential, as one of the sub-research questions is to explore its impact on identifying the behaviour patterns observed from social media. The candidate platforms should be: (A) popular in the target urban context, and (B) available for collecting the geo-referenced data sent by users, i.e. the system should not be closed. Otherwise, it will be impossible to crawl data for the system.

[3] A motivation of this selection of countries, cities and social media platform is provided in section 3.

According to the criteria, we choose Twitter, Instagram and Foursquare for Rotterdam, and Sina Weibo for Shenzhen. The detail reasons for making this selection are described in Chapter 3.

After selecting suitable cities and social media platforms, to perform the experiment, it is required to collect data on each of platforms. The data is to be collected for two weeks on Twitter and Instagram in Rotterdam, and on Weibo in Shenzhen. This operation can be done through the data collecting API provided by Twitter, Instagram, Foursquare, and Weibo. To connect user activities with meaningful venues in the urban environment to study their potential patterns, the dataset will be further enriched with POI information.

The dataset is to be measured to reveal differences in the activity patterns in the two cities. User activities took place in the urban context involve several elements, i.e.: the city, social media platform, user, post, PoI, visit. To investigate the user activity quantitatively, we model those elements and measure them considering the aspects of scales (for both city and social media platform), temporal, spatial, and content. The results of data measurement are further analysed.

Techniques used in this research including data collecting through social media API, data crawler developed as an extension for SocialGlass framework [4], user modelling for data analysis, data analysis from multiple aspects, and so on. The Figure 1.1 shows the conceptual process of this experiment.

1.3 Contributions

The main contributions of this work can be summarised as follows:

1. Analysis of social media data at a disaggregated level (i.e. individual users), across several dimensions: The user, Content, Spatio-temporal aspects, and POI. In User dimension, data is analysed from demographic aspects (age, gender) and individual role of the city (resident, commuter, foreign tourist). In the second dimension, the Content is analysed from the semantic perspective, particularly the word use frequency. For the spatial dimension, the dataset is mapped to the city map, and the geometric distribution of posts is analyse within different spatial scales. For the temporal dimension, we focus on the 24-hours of the day, as well as other specified durations. POI dimension analyses the characteristics of POI categories, as well as its temporal and geometric distributions.

2. Development of pipeline for crawling Weibo data, enrich it with POI information and determine demographic features. Further, encapsulate the Monitoring function as an extension of the Social Glass framework. This component continuously collects Weibo posts and user profiles within a specified range of time and space. It also processes the data and extract several features in the background, such as point of interests enrichment, semantic extraction, home location determination.

3. Creation of a novel dataset, combining data from four different social media platforms (i.e. Twitter, Instagram, Foursquare, Weibo), in two regions (Europe

[4] http://www.social-glass.org

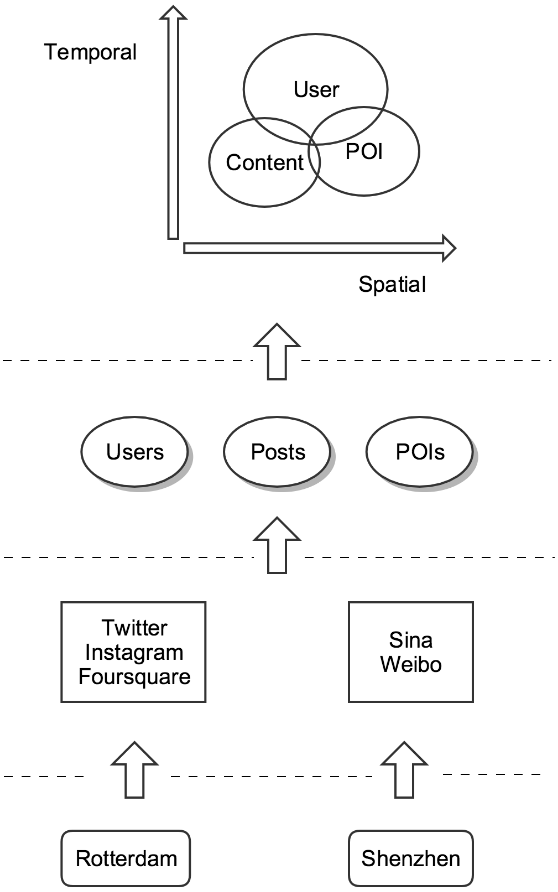


Figure 1.1: Conceptual Process of the Experiment

and China), in two different cities (Rotterdam and Shenzhen), with multiple languages (e.g. English, Dutch, Chinese) with aligned features (e.g. user profiles, geo-referenced posts, point of interests, path patterns, word use).

4. Use the dataset to explore differences of the individual human activities in different urban environment observed from social media. The findings show that, in the context of user activity study, the Twitter and Instagram is more suited for residents concerning local, general, and neutral topics particularly for teen and young users on Weibo and young and mid-age users on Twitter. While, the Instagram is more useful for non-resident concerning international, specific, daily and emotional topics, especially for young and mid-age users. In the meantime, analysis of individual user roles and demographics and various POIs provides significant differences in user activity.

1.4 Thesis Outline

This thesis is structured as follows. First, we give an overview of related work in Chapter 2, looking at works relate to Urban Studies and Cultural Studies respectively, using social media or other datasets as data sources. In Chapter 3, the experiment setting is described, including the urban context, the social media selections, and a set of models with attributes for characterising User, Post, and user Mobility for the experiment. We also introduce the measurement method. Chapter 4 present a framework for performing the experiment, showing its architecture, essential techniques, and implementation. In chapter 5, with performing the experiment in the Rotterdam and Shenzhen, we analyse the dataset from multiple aspects and show Impressive results which will be further discussed in Chapter 6. The possible future work will be in chapter 7.

Chapter 2

Background & Related Work

The subject of this work is a crossing-culture urban study, through social media. Therefore, it involves fields of cultural studies, urban studies, and social media studies, i.e. it includes performing cross-cultural studies and urban studies with the help of social media.

In this chapter, we survey several works related to the above area, both to point out the contributions of previous researchers and to place our contributions in the proper context. We organise this survey around the two main themes of our research: Cultural Studies with Social Media, and Social Media in Urban Context. In both cases, we first provide the reader with background information, followed by a discussion of previous research related to ours. We conclude by summarising how our research builds on those previous works.

2.1 Cultural Studies & Social Media

Since various definitions of culture suggest that collective human behaviour is tightly associated with human culture (Hoebel et al. [22], McGrew, William C. [32], Taylor, Walter W. [46]), and considering that location-based social media reflect human behaviour, hereby, we survey related works with regard to human behaviour study based on social media.

Gao Qi, et al. in their work [17] performs a comparison study of user behaviour in microblogging activities between U.S. people and Chinese people. This study makes use of twitter data which collected from U.S and Weibo data which generated by Chinese people to analyse the distinctions from several aspects, including microblog accessing means, semantic and syntactic analysis, sentiment analysis. The result reveals significant differences in microblogging behaviour between Twitter users and Weibo users. Authors also analyse some of these distinctions from the social science perspective and generate interesting points.

Mitchell Lewis, et al. [36] correlate the social media data with the annual survey and census data to study multiple aspects of U.S. people from the culture perspective. Authors in this study collect more than 10 million geo-referenced Twitter posts from 373 areas in the U.S. during 2011, combining with the census data from the 2011 American Community Survey 1-year estimates. Making use of this integrated dataset, authors examine the language usage culture in U.S. states and cities, and generate taxonomies, determine the happiness level across the country, perform analysis with demographic characteristics with happiness level, and correlate the language usage with urban characteristics such as education and obesity. To measure the sentiment (happiness), the author adopts the Language Assessment by Mechanical Turk (LabMT) word list[1]. This research examines the potential of using social media data in a demographic culture research from multiple aspects.

YANG DINGQI, et al. propose an approach in their work [48] to study the culture between cities based on the social media data. This method examines city culture by extracting three features including residents daily activity, mobility, and linguistic characteristics, based on location-based social networks (LBSN) data, i.e. geo-tagged tweets, Facebook data, and Foursquare check-ins. The feature of daily resident activity is extracted based on the POI category characteristics visited by the inhabitants. The function of mobility is calculated based on the similarity of people who have check-in behaviour in multiple cities. Linguistic feature compares the language using preference in the candidate city. Authors then build a culture clustering model to discover culture clusters based on those features. The system furthermore visualises the discovered couture clusters for future analysis. The experiment shows that the system using LBSN data is capable of capturing the culture elements and generating cultural maps that well correspond with the traditional cultural maps based on survey data.

Hu, Yingjie, et al. [23] proposed a novel method to identify and understand Urban areas of interest (AOI) based on geotagged pictures. Authors use Flickr as the data source and has collected data from 2004-2014 for this study. In the research, authors first extract textual tags and preferable photos from the dataset and then identify AOI through DBSCAN clustering algorithm. This research has been performed in six different cities from six different countries. Besides, the extracted AOI has been examined from spatial, temporal, and thematic perspectives.

McKenzie, Grant, et al. in [34, 33] study the relationship between temporal characteristics and types of POI people visited. As an analogy to spectral signatures, authors modelled the temporal check-in behaviour on a specific kind of POI as a semantic signature, named as a temporal signature, e.g. a user tends to perform a check-in to a particular type of POI at an absolute timestamp. Intuitively, people most likely to visit a restaurant during lunch or dinner time rather than at midnight. Authors studied these temporal signature characteristics in multiple U.S cities and performed a cross-culture study of the signature variant between cities in U.S. (Los Angeles, New York City, Chicago, New Orleans) and the city of Shanghai in China. The datasets studied in this research consists two parts: a dataset collected through Foursquare check-ins for the U.S cities and a dataset collected through Jiepang, a leading Chinese check-in service provider, for the city of Shanghai in China. Authors propose a hypothesis and validate it to understand the temporal variant in different regions. Primarily, to perform the study within cross culture cities, authors apply a dissimilarity calculation using the algorithm of earth mover's distance.

Spielman, Seth E., and Jean-Claude Thill [42] performs a cross-cultural demographic study between a U.S. city (the New York City) and a European city (Paris), through analysing a dataset describing 79 attributes of the census tracks. Their study explores the capability of some assumptions for building the functional form of spatial

[1] http://neuro.imm.dtu.dk/wiki/LabMT

relationships with multiple techniques.

Huang, Yuan, et al. [24] performs a study using tweets to understand the regional linguistic variation in the U.S. Authors collected tweets from Oct. 2013 to Oct. 2014 as the dataset. They summarise the frequencies of a set of lexical alternatives that each user has used, then extract lexical characteristics for those users. After performing several language processing algorithm for features extraction and analysis, the results shows interesting linguistic, regional variations across the U.S. and provide new insights and details about the recent linguistic patterns in the country.

2.2 Social Media & The Urban Context

Attempts to make urban planning and practices more 'scientific' with help of high technologies have been persistent since last century (Fairfield, John D.[12]; Ford, George B. [15]; LeGates et al., [27]; Light, Jennifer S., [28]). The new sources of digital data, such as location-based social media, are increasingly able to be used for performing the urban studies.(Bettencourt, Luis, and Geoffrey West., [4]; Batty, Michael, [3]). There are some related works, and valuable conclusions have been addressed in this area, which will be introduced in the sections below.

Cao, Guofeng, et al. in their work [7] perform a similar research to ours, which propose a comprehensive method and introduce an integrated framework to harness massive location-based social media data for efficient and systematic analysis from spatiotemporal perspective. The dataset used in this research is geo-tagged posts collected from Twitter in North America. Authors extract user information, geo-referenced and timestamp from the dataset, and build the path model, named as trajectories in the paper, which include user information, visiting post list, home location and radius of gyration. Furthermore, based on the path model, a comprehensive model, named as spatiotemporal data cube model, is built, which consists of three dimensions, the human dimension, the spatial dimension and the temporal dimension. Each dimension is constructed with information from a specific perspective, e.g. human dimension includes census data, spatial dimension is built with visiting points within different spatial scales, the temporal dimension has all timestamps mixed with multiple temporal scales. Authors implement an integrated framework to collect data, process data, analyse data and finally visualise them. The visualisation component mapped the three dimensions mentioned above onto a map. It supports not only one data source for mapping but multiple data sources, to provide a comprehensive view of overall movement patterns of the are of interest.

Steiger, Enrico, et al. in the paper [43] explore the semantic association between geo-referenced tweets and their respective spatiotemporal whereabouts. Authors in this study perform a semantic topic model classification and spatial auto-correlation analysis on the geo-referenced data from Twitter in London to identify specific human social activities from those tweets. They introduce a framework which consists functions of data pre-processing, semantic similarity assessment, and spatiotemporal autocorrelation analysis to fulfil this task. In the data pre-processing step, texture content from tweets go through a pipeline which performs operation of tokenisation, stop word removal, whitespace removal, punctuation removal and stemming, to generate word vectors. In the next step, authors apply Latent Dirichlet Allocation (LDA) algorithm with Gibbs sampling to produce the similarity measurement. In the last step, the local indicators of spatial association (LISA) introduced by Anselin is used for correlation analysis. The experiment shows positive correlation in social activities derived from social media data and workplace population census data.

Feick, Rob, and Colin Robertson [13] propose a novel method for exploring patterns in the spatial and thematic content of geotagged photograph (GTPs). In their study, they deal with data collected from Flickr in the city of Vancouver, from 2001 to 2012, totally 54,522 distinct geotagged pictures with 4666 contributors. The spatial pattern is generated through building tag-space, a tags similarity calculation model based on all tags extracted from those GTPs. This method processes the GTP data within steps as below. First of all, data is collected through a grid search in the whole city. It then maps those coordinates into a hexagon layer covering the city. Further, the tag space and content space is built using similarity calculation algorithms. The data is then analysed and visualised in an individual level. It examines two properties of a GTP: the GTP magnitude and GTP tag similarity. The spatial expressions are generated in the last step.

Jiang, Shan, et al. [25] introduce and demonstrate a novel framework to collect, classify and validate the POI data to estimate disaggregated land use in the city of Boston in a very high spatial resolution. Authors in this paper use machine learning to classify the POI data collected from Yahoo! Local, a typical publicly available volunteered geographic information (VGI) data, into standardised NACIS categories, and then combining the classified POI with official employment data from U.S. government. Based on these data, authors in this study build activity-based LUTE models and show agent-based urban simulation (especially for work location choice and destination choice models) and perform analysis of urban and regional economies.

Zhou, Xiaolu, Chen Xu, and Brandon Kimmons [50] attempt to identify tourist destinations in an urban context through analysing the social media dataset of geotagged pictures. The dataset is collected from Flickr. Authors focus on the feature of text tags, geo coordinates, and timestamp of each picture. By extracting and calculating these functions, authors build reverted tree to point back to the pictures and their geo-locations. With re-ranking the order of prediction places and validating with another official dataset, authors examined their novel approach in the urban environment. They make use of cloud-computing techniques to accelerate the whole process.

Kunze, Carola, and Robert Hecht [26] extract semantic information through analysing tags on a dataset of OpenOStreeMap (OSM). The dataset is collected from a city in Germany. Authors further enrich the official building footprint with the extracted tags, and in this way to enhance the capability of estimating the number of dwelling units and population.

Gao, Song, et al. [18] performs a spatial analysis in an urban context. They used Flickr as the data source and collected an enormous amount of geo-referenced pictures from Flickr in a city of U.S. Authors build a pipeline to process features extracted from the dataset, such as textual tags, title, description, taken time, geo coordinates and user ID. In this way, they introduce a novel approach to harvesting the crowd-sourced gazetteer entries from social media making use of high-performance cloud-computing.

Hemsley, Jeff, and Josef Eckert in their paper [21] study the social properties of social network users with the spatial proximity of the networks. Authors use Twitter as the data source. They collected a dataset of Twitter users and their followers. Through performing a comprehensive analysis of network density and network transitivity, authors examines the distinctions of social network and the local network in an urban context. They found that the network density and spatial clustering depends on the network size, i.e. smaller network are more social clustered, and the Twitter network are more transmittable than the local network in urban context.

Croitoru, Arie, et al. [10] study the characteristics of information propagation both in a physical urban environment and in the cyberspace through social media. Authors perform this study based on analysing the data from Twitter. The analysis of the spatial footprint, social network structure, and content both online and offline. The apply the proposed novel approach into two case studies. Typically, in the second case, authors analysed the information propagation mechanism in the Boston Marathon bombing in April 2013.

Stephens, Monica, and Ate Poorthuis.[44] performs a companion analysis in the social network between online social media and offline networks, based on the geo-tagged tweets and the following relationship collected in the U.S. The result shows that the density and the spatial clustering depends on the size of the network. The research further reveals that the Twitter networks are more efficient at transmitting information at the local level.

Shelton, Taylor, Ate Poorthuis, and Matthew Zook. [41] introduce a novel conceptual and methodological framework for analysing social media data from socio-spatial perspective. In this study, authors perform an analysis in Louisville, Kentucky, using the tweets collected from that city. By combining the conceptual approach of relational socio-spatial theory with the methods from GIS science, authors analyse spatially and mobility characteristics and examines it about the segregation notion. The result shows that the studied urban area is not separate and apart from the rest of the city.

2.3 Summary of Related Works

In the previous section, we have introduced several related work in culture studies with social media and social media in urban context. Most of those works involve multiple aspects, such as human behaviour, user mobility, data collecting, and so on. In this section, we summarise these works from perspectives of concerning fields, data sources, features, introduce the framework and cross-culture.

Research Aspects Involved

Varies fields are involved in these works. Authors in [34, 13, 18, 43, 23, 7, 48] perform spatiotemporal related studies, while works of [26, 42, 17] study from demographic and behaviour perspective. Relation network is touched by [44, 41]. Language and information transferring is studied in [24, 10]. Land use and place recommendation is discussed by [50, 25].

Data Source Explored

There are several social media networks are used as data sources. Studies of [50, 13, 18, 23] use Flickr geo-tagged data, while research works of [43, 24, 44, 10, 7, 41, 17, 48, 36] are based on geo-tagged tweets. Foursquare check-ins and Jiepang are used in works of [34, 48]. Chinese Weibo is explored in works of [17]. Other data sources include Facebook [48], OpenStreetMap [26], Yahoo Local [25].

Data Features Investigated

Check-ins, coordinate, Point of Interest, timestamp are the most common features of these studies. Besides, tags are examines in [50, 13, 26, 18, 23, 36]. Address related features such as city, region, street, is used in [34, 26, 48]. Following and follower relationship is touched in [44, 41].

Research Framework Introduced

Most of works in this area also produce frameworks, applications, pipelines or toolset, e.g. [50, 18, 43, 25, 23, 17, 48]. In generally these frameworks consists functions of data collection, features extraction, model generation, and comprehensive analysis. Some of them even support data visualisation in multiple level and scale.

Cross Culture Related

As our research is aiming at cross-culture comparison study, hereby we also list the cross-culture related works. [34, 42, 23, 17, 48]. Some of them performs cross culture study in different countries, while some of them do the comparison between different cities. The result shows that using social media it is possible to perform such comparison research.

Summary

These works have been the main inspiration of this thesis. They show that social media sources are possible to be used for culture and urban related studies. In our research, we use Weibo, Twitter, Instagram and Foursquare as data sources. We focus on user analysis, spatiotemporal analysis, demographic analysis, and do a cross culture comparison study.

Chapter 3

Experiment Design

The research questions of this study involve several elements, i.e. the urban context, social media platforms, and user activities. In this chapter, we design an experiment investigating these elements to answer research questions.

3.1 Urban Context

The research is based on multiple social media platforms in different urban environments in diverse countries. Approximately selecting two cities is essential. The criteria for selecting cities are:

1. Representative of different cultures.
2. Characterised, to a certain extent, by mutual features so as to be mutually comparable.
3. Different in terms of social media usage and platform popularity.

There are many candidate cities that match above criteria. We start with cities in Europe and China for the reason that these two regions are believed to have significant culture differences in the urban environments in many aspects ([40], [20]). We further select city of Rotterdam in the Netherlands and city of Shenzhen in China to start. The reason for choosing these two urban environments is that, despite their differences (e.g. in population size, land scale), they present similar features. For instance:

- Both are "second cities", i.e. not capitals yet major ones to each one of the two countries.
- Both cities are located on river Deltas, i.e. Rotterdam located on Rhine-Meuse-Scheldt river delta, and Shenzhen located on Pearl River Delta.
- Both have harbour facilities, which might be relevant to the activities carried out in each one of them to a great extent
- Both were built from scratch. Rotterdam is a historic city in the Netherlands which was destroyed completely during the world war II, while Shenzhen has

developed into a big city from a small fish village in past decades, and has become the first and one of the most successful Special Economic Zones (SEZ) after 1979 [14].

The basic information for these two cities is shown in the Table 3.1.

Table 3.1: General Information About City of Rotterdam and Shenzhen

	Rotterdam	**Shenzhen**
Population	624,799	10,547,400
Area (km^2)	325.79	1991.64
#Districts	14	10
GDP per capita (Euro)	36,000	22,309
Primary Economy Activity	Maritime transport, Shopping	High-tech industries, Manufacturing, Maritime transport

3.1.1 City of Rotterdam

Rotterdam, located within the Rhine Meuse Scheldt river delta in the North Sea, is a major city in the Netherlands. The city map is shown in Figure 3.1.

The port of Rotterdam is the largest cargo port in Europe and the 10th largest in the world. The river separates the city into two part, i.e. the North part and the south part, with several bridges connecting them. The shipping industry is the primary industry in Rotterdam. Nowadays, well-known companies have set offices in this city for business[1].

Rotterdam consists of 14 sub-municipalities, shown in Table 3.2. People in the Rotterdam are diverse. The city centre has, according to the recent analysis, a larger amount of higher education and higher income people than other areas. It also has more than 50% of foreign-born citizens [2].

3.1.2 City of Shenzhen

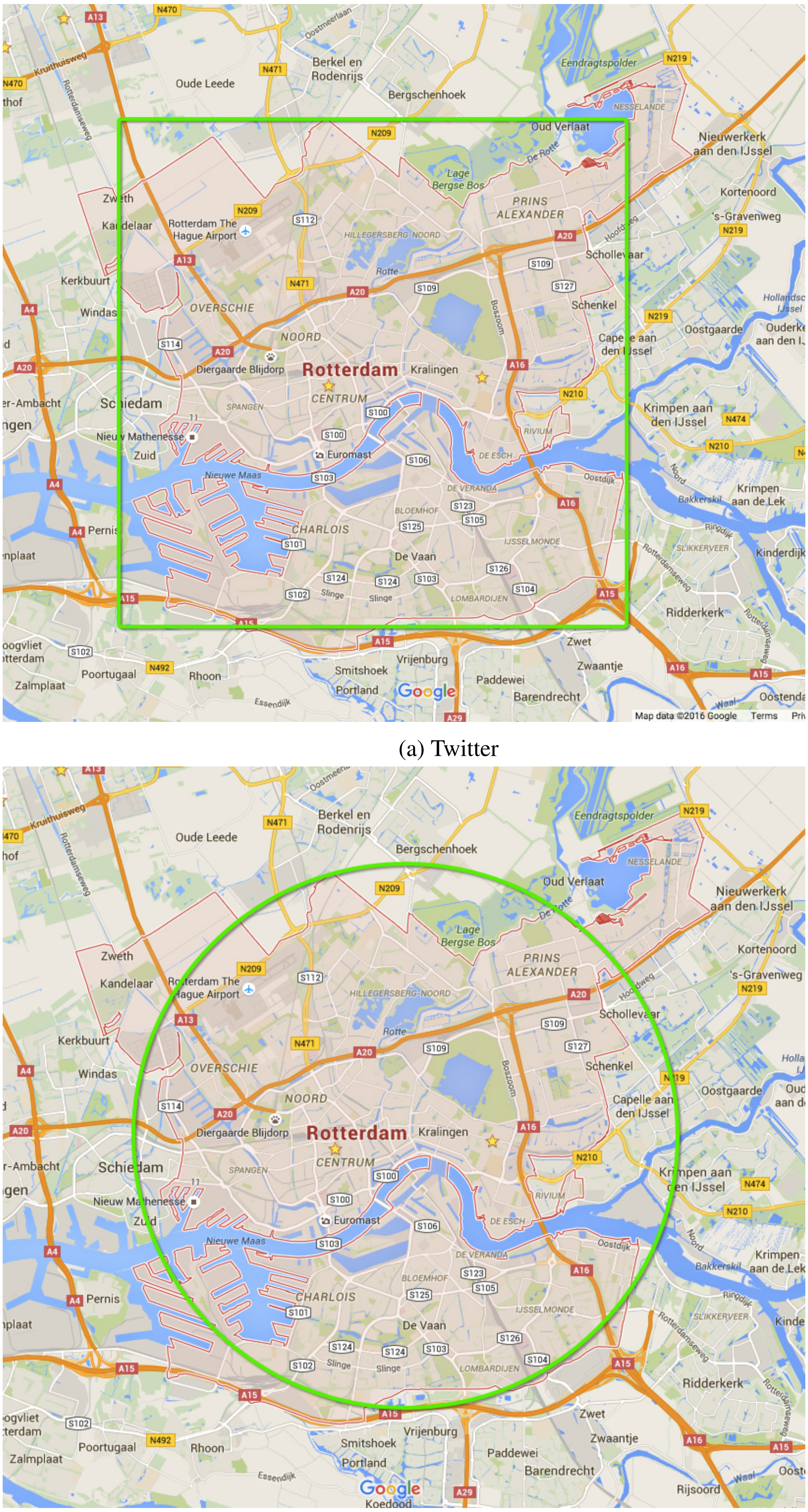
Shenzhen was only a market town called Sham Chun Hui which the Kowloon-Canton Railway passes through before 1979, but now it has become a major city in Guangdong Province, in the south of P.R.China, situated immediately north of Hong Kong Special Administrative Region. This is because of the institution of the policy of "reform and opening" establishment of the Special Economic Zones (SEZ) in late 1979. Shenzhen has become the first and one of the most successful SEZ after 1979 ([14]).

Shenzhen is a major manufacturing centre in China. It is also the home to some of China's most successful high-tech companies (Cai, Yu, et al. [6]).

Moreover, Shenzhen is a sub-provincial city which has total ten administrative districts (six administrative districts and four management new districts with a model service industrial corporation with Hong Kong 3.2). Shenzhen's population and activity have developed rapidly since the establishment of the SEZ. Now it is the largest

[1] https://en.wikipedia.org/wiki/Rotterdam

[2] http://goo.gl/3jsEd6



(a) Twitter

(b) Instagram

Figure 3.1: Rotterdam Municipality Administrative Boundaries (red) with Bounding Box/Circle (green) of Twitter and Instagram

Table 3.2: Administrative Districts in Rotterdam

District	**Area** (km^2)	**Population**
Centrum	5.57	31,860
Charlois	12.08	63,820
Delfshaven	5.79	73,430
Feijenoord	7.89	70,140
Hillegersberg-Schiebroek	13.27	42,555
Hoek van Holland	17.44	9,665
Hoogvliet	10.73	34,050
IJsselmonde	13.10	58,425
Kralingen-Crooswijk	12.77	49,995
Noord	5.37	50,340
Overschie	15.79	15,950
Pernis	1.60	4,790
Prins Alexander	20.23	92,650
Rozenburg	6.50	12,505

Source: Centraal Bureau voor de Statistiek (CBS). (2014). Wijk- en Buurtkaart 2011 & Kerncijfers (2nd update).
Retrieved from: http://www.cbs.nl/nl-NL/menu/themas/dossiers/nederland-regionaal/publicaties/geografische-data/archief/2012/2012-wijk-en-buurtkaart-2011-art.htm (Accessed on April 4, 2016).

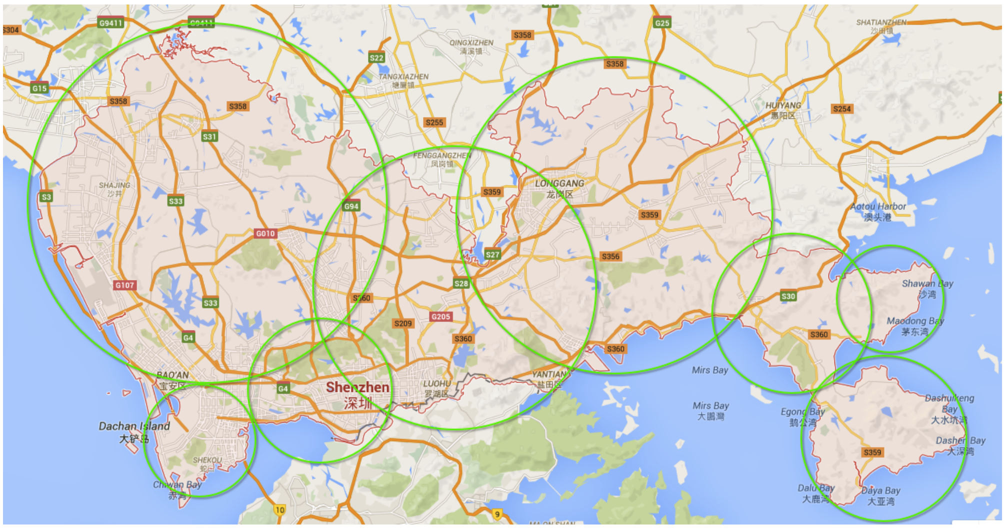


Figure 3.2: Shenzhen Municipality Administrative Boundaries (red) with Bounding Circles (green) of Weibo

migrant city in China [3]. The population structure is shown in two opposing extremes: intellectuals with a high level of education, and migrant workers with poor education [4]. Detail population and area information of districts in Shenzhen are shown in Table 3.3.

[3] http://goo.gl/EY6r3g

[4] http://goo.gl/y67jZi

Table 3.3: Administrative Districts in Shenzhen

District	Area (km^2)	Population
Nanshan District	185.49	1,108,500
Futian District	78.65	1,317,511
Yantian District	74.63	209,360
Dapeng New District	295.05	126,560
Guangming New District	155.44	480,907
Pingshan New District	167	300,800
Longgang New District	387.82	1,672,720
Luohu District	78.75	923,421
Longhua New Distric	175.58	1,379,460
Bao'an Distric	398.38	2,638,917

Source: http://www.phbang.cn/finance/data/152423.html

3.2 Social Media Platforms

It is necessary to choose proper social media platforms for this experiment to collect social media data. The criteria for choosing social media platforms, as described in Chapter 1, includes:

1. It should be popular in the target urban environment.
2. the geo-referenced posts sent by users should be available for collecting, i.e. the system is not closed, that it is possible to crawl data for research.

We then select proper social media based on the criteria above. There are plenty of social media platforms popular worldwide in the recent year[5]. We focus on the most popular ones in the Rotterdam and Shenzhen. In Rotterdam, we choose Twitter, Instagram and Foursquare, because other attractive candidates, such as Facebook and WhatsApp, are all closed systems, i.e. it is not possible to collect user behaviour data. In our solution, the Twitter and Instagram are used for sending posts, while Foursquare is used for mapping Point of Interest (PoI) with those posts. In Shenzhen, as the Twitter, Instagram and Foursquare are not available, we have to choose the Sina Weibo, which is the most popular social media platform in China, as our target. Other popular platforms such as WeChat and QQ are all closed systems and therefore not suitable for this study. In our solution, the user posts data in Shenzhen is collected from Sina Weibo and further enriched with POI through Sina Weibo's location APIs.

3.3 Models

The models for this experiment are used for characterising Users, Posts, and their Visit. We hereby introduce each of them and explain their attributes in the following sections.

[5] http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

3.3.1 Users

In the context, people refers to individuals who live in the city using social media and sending geo-referenced posts. Several attributes can characterise each user. We therefore define an individual user in **Definition** 1 and explain these attributes in the following sections.

Definition 1. *User = {id, name, gender, age, individual_role, home, radius_of_gyration, profile_picture}*

User id, name

The user id is the identification id in the specific social media platform. The name is usually a screen name used on the Internet. These two attributes are retrieved directly from the user profile on the social media platform.

User gender and age

In Weibo platform, one's gender is a field provided by the user, while in Twitter and Instagram gender is not provided. For both of the platform, user age is not provided. Therefore, to make comparative among those social media platform, it is necessary to find an approach to extract user gender for Twitter and Instagram users and age for Weibo users. User information required for determine the gender and age are generally from two elements. The first one is the user profile which contains user screen name, profile picture, address, following count and so on. The second one is the content sent with their posts. There are some works have done on this problem.

Peersman et al. [38] performs a text categorization approach for the prediction of age and gender on a corpus of chat texts collected from social media. They found that different types of features of those texts were distinct for predicting the age and gender. Mislovet et al. [35] detects the Twitter user gender using their first name on Twitter with significant data bias. C.Titos.Bolivar [47] in his work identifies the user gender and age through a face detecting service, FacePlusPlus [6], which uses user profile picture calculating user gender with a confidence and determining user age with a range, created based on the research of Zhou et al. [49].

In our study, for Twitter and Instagram users, we determine their gender and age using the FacePlusPlus service. For Weibo users, we determine their age using FacePlusPlus service, and extract their gender from the gender field in their profile, as described in Table 3.4.

Table 3.4: Method of Determining User Gender and Age

	Gender	**Age**
Twitter	FacePlusPlus	FacePlusPlus
Instagram	FacePlusPlus	FacePlusPlus
Weibo	Profile	FacePlusPlus

FacePlusPlus: http://www.faceplusplus.com/

[6] http://www.faceplusplus.com/

Home location

Home location refers to users' real home address in the city. We need this information because having one's home address we can identify city role, and perform movement analysis. Generally, this information is not requested users to provide in most of the social media platforms. In Twitter and Instagram, users are not asked to register their home address at all, while in Weibo a similar field named as "Gift Receiving Address" is asked, but not mandatory.

T.Pontes et al. [39] infer one's home address based on all tweets of one sent, assuming that the combination of all tweets location may reveal the user's routine. C. Davis Jr et al [11] propose a method using the locations of friends to predict a user home location, based on the assumption that the most users friends tend to live in the same city. However it reaches low accuracy. J. Mahmud et al. [31] uses an ensemble of statistical and heuristic classifiers to predict locations and makes use of a geographic gazetteer dictionary to identify place-name entities. C.Titos.Bolivar [47] in his work determines users' home location using a similar approach as Cheng et al. [8] proposed. Instead of a custom grid size, C.Titos.Bolivar improves the approach and uses geohashes [7] which already represents coordinates as a grid. The approach is a recursion search process. It starts with a grid search with geohashes length of 2, in order to find the geohashes with most of posts. Further it performs a grid search with geohashes length gradually increasing one each time until 8, i.e. using length of 3, 4, 5, 6, 7, and 8, respectively. Finally, reaching the geohashes length of 8, meaning the searching process will return areas with approximately of weight 40m and height 20m [8], the approach returns the centre of the area with the most of posts as the home location.

In this research, we use the same method as C.Titos.Bolivar used for determining users' home location.

User's individual role

Users living in the city have different roles. They may be residents or guests from either domestically or abroad. In our study, we determine users' role in the city to perform demography analysis.

As we described in the previous section, given that Twitter is not available in China, we choose Weibo, the Chinese Twitter, as the social media platform for the comparison research in Shenzhen. Due to the language distinctions, foreigners in China seldom use Weibo. Consequently, it is impossible to collect Weibo posts sent by foreigners in Chinese cities in an analysable scale, and therefore, only residents and commuter (non-resident from the same country) are available to be identified through Weibo data. In our study, we only consider residents and commuters for Shenzhen. For the city of Rotterdam, we include the foreign tourist for analysis. We define these three roles as below:

Definition 2. *Resident: People who live in the city.*

Definition 3. *Commuter: people who come from the same country as the visiting city.*

[7] http://geohash.org/site/tips.html

[8] https://en.wikipedia.org/wiki/Geohash

Definition 4. *Foreign Tourist: Tourists who come from the different country with regard to the studied city.*

The way to identify resident and commuter is different on various social media platforms. On Weibo, user province and city is provided by the user in the profile. It therefore can be used for identifying whether this user is a resident or a local tourist. On the Twitter and Instagram, however, such information is not available. Therefore, in this research we use home location to help us determining users' city role on Twitter and Instagram, i.e. if the user's home location is in the city, it is marked as a Resident. Otherwise, it is marked as a Commuter.

Radius of gyration

The radius of gyration is a useful metric to measure a user's travel distance in the city, which used widely ([19]). It indicates how far and how often one user travels. The radius and gyration is defined in **Definition** 5:

Definition 5.

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - r_{cp})^2} \tag{3.1}$$

r_{cp} refers to the centre of all places a user visited and sent posts. r_i is the position of a certain place which the user visited and sent a post. Thus the $r_i - r_{cp}$ is the distance between the centre of places and each places the user travelled in the city. N is the number of places the user travelled.

The radius of gyration indicates the extent of user's travelling in the city. Take the Table 3.5 as an example, which shows the average distance moved by each user towards the central of those locations. Note that different users may have a different number of visited places for this calculation, such as the User 4 in this example only visited three places while other users visited five places. The average of the radius of gyration is listed in the last row of the table, showing that the User 2 has the longer movement distance than other Users, while User 1 has the least one.

Table 3.5: Example of Radius of Gyration

	User 1	**User 2**	**User 3**	**User 4**	**User 5**
$r_1 - r_{cp}$	3	98	7	2	26
$r_2 - r_{cp}$	2	2	10	5	40
$r_3 - r_{cp}$	3	4	9	5	30
$r_4 - r_{cp}$	4	2	7	n/a	1
$r_5 - r_{cp}$	1	3	17	n/a	2
r_g	2.79	43.9	10.65	4.24	25.22

3.3.2 Posts

A social media post refers to a piece of complete information sent by a user on social media. On Twitter, it is usually called as Tweet. On Instagram and Weibo, there is no specific name dedicated to it. Therefore, in our study, we use "Post" as its name, and define it as a combination of attributes in **Definition** 6.

Definition 6. *Post = {post_id, user_id, content, latitude, longitude, PoI_id, timestamp}*

Post id, user id, content

Post id is a unique identification of a post, sending by the user of that user id. The content is the text information encapsulated in the post. A Twitter post has 140 characters limit, as well as in Weibo.

3.3.3 Point of Interests

A Point of interest, or PoI, is a particular Point location that someone may find useful or interesting, such as a hotel, a restaurant, or a bus station. When people sent geo-referenced posts on the social media platform, the latitude and longitude are contained in the posts. Mapping the coordinate to a PoI helps us not only knowing the place but also understanding the place with abundant information from PoI, such as its title, functionality category, popularity and so on. In this research, we map all the collected posts into PoIs and perform analysis based on these PoIs. To do so, we define a PoI as a collection of properties in **Definition** 7.

Definition 7. *PoI = {id, title, latitude, longitude, category}*

Point of Interests category

Each PoI belongs to a certain category with regard to its function, such as Museum, Music Venue, and so on. There is no unified PoI categorisation. Each social media platform adopts its own PoI categories. Twitter and Instagram use the Foursquare Category Hierarchy, while formulate the category hierarchy by itself. Normally the category hierarchy has several levels. Category hierarchy from different providers vary in the structure, especially in the root level or the first level, but share similarities in the second or third level, and they are generally cover all PoI functionalities. We define the PoI category hierarchy in **Definition** 8:

Definition 8. *PoI category hierarchy = {id, name, parent_id}*

J. Lingad et al. [29] leverage the Natural Language Processing techniques to identify location related words from tweets, and to determine the PoIs. C. Davis Jr et al [11] inferring the PoI of Twitter messages based on user relationships. Noulas, Anastasios, et al. [37] tackle this problem using Foursquare Venues Service, which maps location information to PoIs with the help of its venues database of millions of places from all over the world, enriches the location with information of category and popularity. In our research, we use Foursquare for Twitter posts PoI mapping. On Weibo, we use PoI mapping service provided by Weibo itself.

To explore POI category influences in this study, it is necessary to align the POI categories from Weibo with the ones from Foursquare. As the top level of categories from both sides have some conflicts and overlaps, we make the collection of those categories in the second level. Table 3.6 shows that Weibo POI categories have been represented by second tier categories from one or several top level Foursquare categories. Detail mapping relationship is listed in the Appendix.

Table 3.6: POI Categories Mapping from Weibo to Foursquare i.e. Weibo POI Categories are Represented by Foursquare POI Categories

Weibo POI Category	**Foursquare POI Category**
Travel & Accommodation	Travel & TransportResidence
Office building and Organisation	Professional & Other Places
Education	College & University
Food & Beverage	Food
Shopping	Shop & Service
Life & Entertainment	Arts & Entertainment College & University Nightlife Spot Professional & Other Places
Park & Plaza	Outdoors & Recreation
Enterprise	Professional & Other Places
Other Places	Outdoors & Recreation Travel & Transport Professional & Other Places

3.3.4 Visit

A visit action, defined in **Definition** 9 consisting three attributes, i.e. the user's id, the PoI and the timestamp, representing a user sending a post in a PoI at the timestamp.

Definition 9. *Visit = {user_id, PoI, timestamp}*

3.4 Measurement

In the context of this research, according to the research questions, the research object - i.e. user activities - involve several elements that have been modelled in above sections. We employ a set of metrics listed in Table 3.7 to measure those elements. These metrics can be divided into three categories, namely:

- **Aggregate metrics.** such as the number of Users, Posts, PoIs, showing the absolute capacity of the social media users, the number of posts generated, and the number of PoIs visited.
- **Ratio metrics.** which is the normalised aggregate metrics, or the ratio between two aggregate metrics, to measure the relative extent between them, such as the number of posts per user, the ratio of the mean of the radius of gyration with the size of the city.

- **Statistics metrics.** which reveal the insights of the data through statistical methods, such the similarity of the temporal distribution probability of the users' active time.

Moreover, several aspects are considered for choosing the metrics. First of all, the scale is an important aspect to consider for user activity measurement, given that the range of cities, namely the size and population, is different in the various countries. For example, the number of posts on social media in Rotterdam is less than in Shenzhen. However, the average number of posts per user may lead to different conclusion considering that Shenzhen has a huge more social media users than Rotterdam. Be aware of the scale distinctions of the target cities help us understand the dataset more accurately. In this study, the city and social media platform are measured through scale metrics, depicting the size and populations of the city and the number of users on the social media platform.

Temporal and spatial aspects are critical in measuring user activities. Activity, in general, took place on two dimensions, i.e. the time and space. In the context of this research, temporal and spatial metrics are employed to reveal the time and space characteristics of the user activities. For the temporal aspect, we use Gini index and 24 hours active probability to show the posting time features and use Jensen-Shannon divergence[9] to measure the similarity between two probability distributions. For the spatial aspect, the Gini index, spatial distribution and radius of gyration show the spatial characteristics.

Another aspect to consider is the content sent with the post. Each user activity in this context is binding with a piece of the message sent with the post. Investigating the content of the message helps in characterising the activities. The content is measured through statistics of top words in the social media content.

The Gini index used in this work is a measure of statistical dispersion initially used for showing the income distribution of a nation's residents. It is the most commonly used measure of inequality among values of a frequency distribution, which ranges from 0 to 1, meaning from perfect equality to maximal inequality. In our experiment, we employ Gini coefficient as an indication of the statical dispersion of the distribution of posts and POIs. In our study, we use Peter Rosenmai's approach[10] to generate the Lorenz curve and further calculate the Gini index.

The Jensen-Shannon divergence used in our work is a popular method of measuring the similarity between two probability distributions. In this study, we use xboard's library[11] to implement the calculation.

[9] https://goo.gl/RSp0GO

[10] http://www.peterrosenmai.com/lorenz-curve-graphing-tool-and-gini-coefficient-calculator

[11] https://github.com/xboard/python-libJS

Table 3.7: Metrics

Object	Metric	Description
City	Area	The size of the city.
	#Population	The population of the city.
Social Media Platform	#User	The number of Users on the platform.
	#Post	The number of Posts sent by the users on the platform.
User	%Users	The percentage of these users in all users. e.g. the percentage of male users in all users on Weibo.
	Post/User	The number of posts per user of these users.
	Sig.(Post/User)	The extent of distinctions of the posts per user of these users.
	Post/PoI	The number of posts per PoI of these users.
	Sig.(Post/PoI)	The extent of distinctions of the posts per PoI of these users.
	#PoI	The number of PoIs visited by the users.
	PoI preference	The Top three PoI categories that visited by the users.
	Avg(RoG)	The Average radius of gyration of these users.
Post	Gini.Temporal	The statistical temporal dispersion of the posts
	Temporal Distribution	The temporal distribution of the posts in 24hours of a day.
	Gini.Spatial	The statistical spatial dispersion of the posts
	Spatial.Distribution	The spatial distribution of the posts in 24hours of a day.
	Word use	Top words that used in those posts.
PoI	Gini.Temporal	The statistical temporal dispersion of the PoIs.
	Temporal Distribution	The temporal distribution of the PoIs in 24hours of a day.
	Gini.Spatial	The statistical spatial dispersion of the PoIs.
	Spatial.Distribution	The spatial distribution of the PoIs in 24hours of a day.

Chapter 4

System Architecture and Implementation

A system is required for crawling and process the social media data to perform the proposed research. In this chapter, we describe the overview of the system's requirement, architecture, and implementation, as well as the integration with Social-Glass framework.

The data we need to perform the proposed research consists of Tow part: the Weibo data in Shenzhen, and the Twitter and Instagram data in Rotterdam. For the Weibo data we need to create a Weibo Crawler to collect that, while for the Twitter and Instagram data in Rotterdam, we will use Social-Glass Framework, an existed social media data-collecting, and visualisation tool, to reach our target. At the same time, we will add the Weibo Crawling function into Social-Glass framework.

Consequently the whole task consists two main parts. The first one is the Weibo Crawler, which is responsible for Weibo data crawling and processing within a given area and a given period. Another one is to encapsulate the Weibo Crawler into an extension according to the framework interface, and integrate it into Social-Glass framework.

In this chapter we will illustrate each of these two parts in three sections: requirement, architecture and implementation.

4.1 Weibo Crawler

4.1.1 Requirement

Concerning functionality, the Weibo Crawler should be capable of monitoring all the geo-referenced posts within a given area in a given city and a specified period. After collecting those data, the system should be able to process the raw data, enriching it with PoI information, calculating user attributes, determining home location and city role, extracting path and path patterns, as well as analysing semantics. It should be easy to add new data process function to the system, to maximise the value of the data. All the information should be stored permanently in the database or JSON files for later analysis and reuse. Notably, similar to most of the social media platforms that have restrictions in data crawling, Weibo also have such kinds of limitations which are vaguer and hard to detect. To crawl Weibo data for this research, the Weibo Crawler

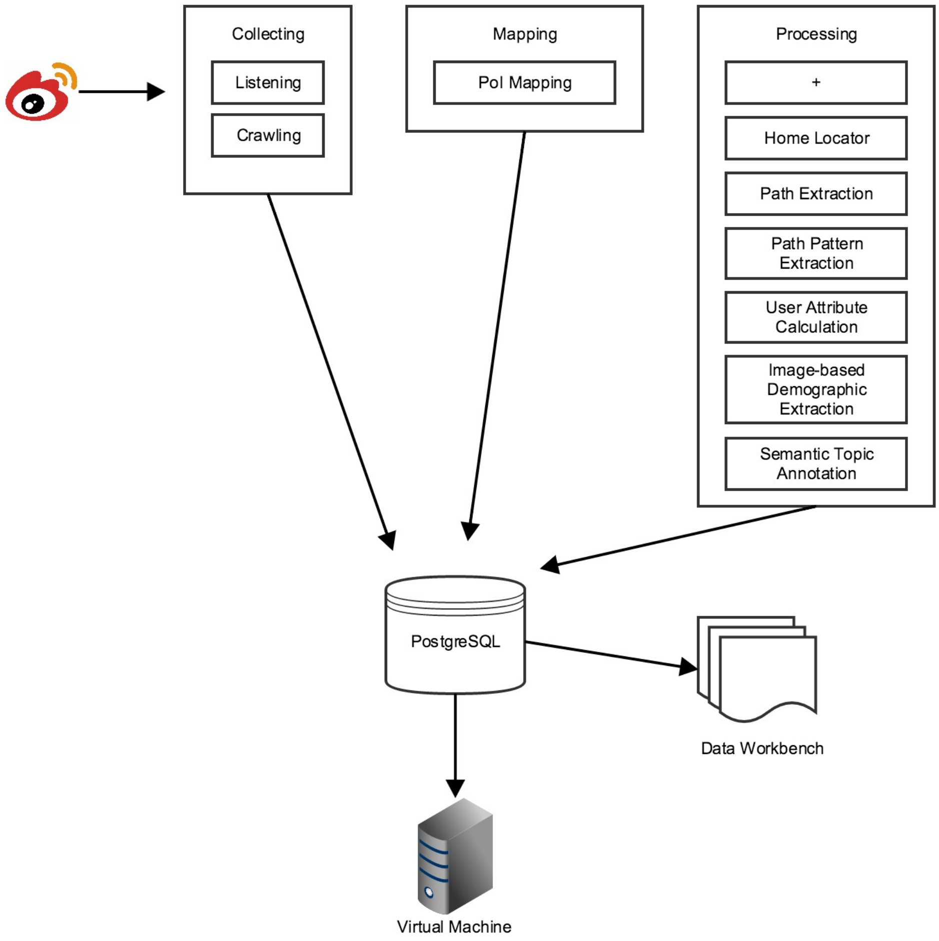


Figure 4.1: Architecture of Weibo Crawler

should be able to avoid been banned, and at the same time to crawl Weibo data as fast as possible.

4.1.2 Architecture

The architecture of Weibo Crawler, Figure 4.1, shows that the main components and their connections. Currently, it has five major components, which are Collecting, Mapping, Processing, PostgreSQL DB, and Data Workbench. All of them can be deployed on virtual machines. To facilitate adding new functions to data processing, the interface of the Processing component is well designed for plug-ins, i.e. any new features complying with the proposed interface will be easy to be added and get accessing to the data.

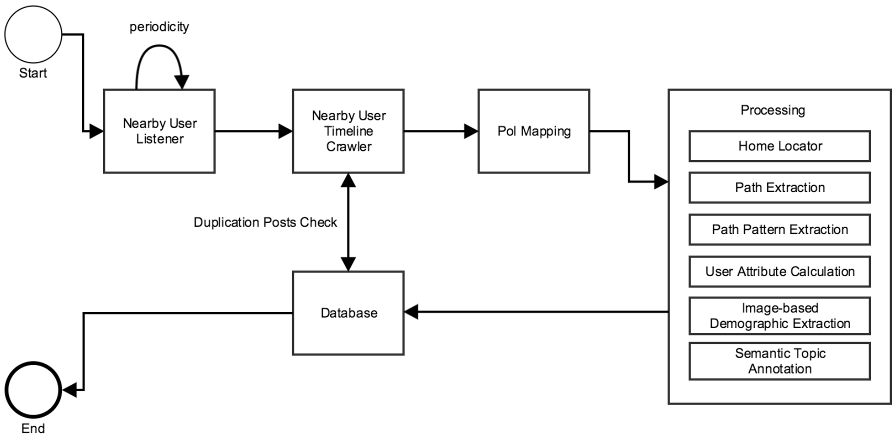


Figure 4.2: Data flow of Weibo Crawler

4.1.3 Implementation

The Weibo Crawler have five principal components. Figure 4.2 shows the data flow between these components.

The target of the Weibo Crawler system is to collect geo-referenced posts from Weibo. We, therefore, studied the Weibo API to find a proper way to crawl those data. There are two available methods to do that. The first one is to use "Places/Nearby Posts" API, which returns a particular amount of posts sent within a given location area. However, the number of returned posts is relatively small. While, the second method is to get all users who sent posts in the particular field, by the API of "Places/Nearby Users", and then crawl their history or monitor the real-time posts through "Places/User Timeline". Experiments show that the second method captures an enormous amount of posts. Since Weibo officially does not explain the rate of returning posts, we could not know the precise rate. In this study, we use the second method to crawl data as much as possible.

In the data flow, Nearby User Listener component is monitoring users who sent gen-enabled posts in the given area. It only records the User ID of those users and pass these IDs to the Weibo Crawler component. The Weibo Crawler component response for crawling all the geo-referenced posts of given users. It will filter posts according to the time and area, i.e. only those posts sent during the given range of time and area will be stored. It also performs a de-duplication operation to avoid duplicated posts. Once this is done, the Weibo Crawler component passes the collected data to the next component, the PoI mapping component, which in charge of mapping the coordinates where this post is sent to a PoI place. Further, the data enriched with PoI will be passed to the processing component package, where several processing functions or sub-components calculate and process the data. All the data is stored in a PostgreSQL database.

In this following, we illustrate each of the major components in detail.

Nearby User Listener Component

The function of this component is that crawl all users who have sent gen-enabled posts in a given area. According to the Weibo API, three parameters are required, which are the coordinate of the centre point, and the radius, with returning a list of Weibo Users in JSON. However, Weibo API only supports a relatively small radius, 11132 meters. As we know, Chinese cities such as Shenzhen is quite significant, and such small radius is not enough to cover the whole city. To solve this problem, we developed an algorithm which accepts a centre point with unlimited radius and transfers it into a set of centre posts with Weibo accepted radius.

This algorithm [1] helps in generating a set of coordinates for building smaller circles to cover a given larger circle with as low overlap as possible. The degree of overlap is configurable. It is typically used in social media crawling when the target crawling area is relatively larger than the one that the API supports. The figure 4.3 shows an example of crawling Weibo in a big area in Shenzhen through a set of small circles. The red lines outlined the whole city of Shenzhen and filled with light red background while the green circles are established for generating small circles to cover the entire city with configured overlap degree.

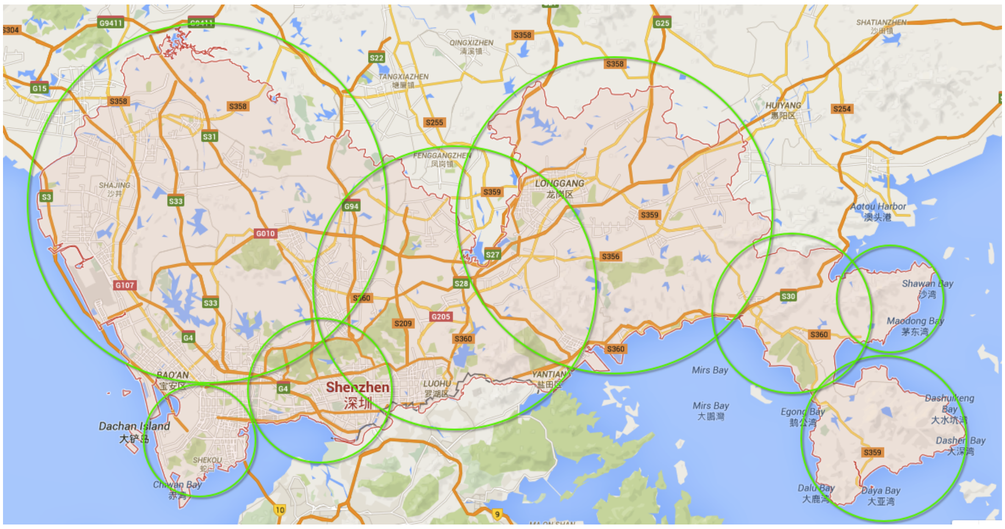


Figure 4.3: Generate Small Circles from Big Circles Covering the Big City, Shenzhen

Weibo User Timeline Crawler Component

Weibo User Crawler component receives a set of user id and crawls the geo-referenced posts of those users. The result includes all history posts per user. The component then filters posts and only keep unique ones within given time and area. It parse all posts and users from JSON and extract user profile and information to store into the database.

[1] https://github.com/vincentgong7/coordinate-generator

Point of Interests Mapping Component

Once the Weibo User Crawler Component captured a post, the PoI Mapping Component will start to map the coordinate in the post into a PoI place, through the method we describe in Chapter 3.1.1. For the Twitter and Instagram, we use Social-Glass system for mapping. Hereby, we describe the Weibo PoI mapping implementation. For Weibo posts, we use PoI service provided from Weibo itself to mapping coordinate with PoI. The corresponding API in Weibo is "place/nearby/pois", which requires the parameter of a coordinate, range, and sorting strategy. The coordinate is the latitude and longitude to be mapped to a PoI. While the range is the searching distance. The sorting strategy includes three options: weight, distance, or check-in amount. As Weibo does not explain the algorithm of calculating the weight, we would not be able to know how it works. Thus, we mapped all posts based on all three options, for later analysis.

Weibo User Home-Locator

In this research, we use the same method as C.Titos.Bolivar [47] used for determining users' home location. For Twitter and Instagram, we use it to detect home address, city, and country. For Weibo, we use it to detect home address. The city and country information is provided by the user and can be retrieved through Weibo API.

Path Extraction

The Path Extraction Component generate path based on the PoIs visited by a user in a day. It first lists all PoIs a user visited, and then split them into several list according to different days. In each list, those PoIs are ordered by visiting timestamp.

Path Patterns Extraction

As discussed in Chapter 3.4.3, in this research we use PrefixSpan algorithm to extract sub-sequences pattern in the path. The Java library SPMF [16] is used in the implementation.

Image-based Demographic Extraction

Users' properties such as gender, age, race are required to perform an analysis from the demographic perspective. According to the Chapter 3.1.2, in this research we use a face detecting service, FacePlusPlus [2], to extract these properties.

Faceplusplus is a human face detecting service that given an image it detects faces and return demographic information in a machine readable format. An example of its usage is as below. Given a profile page, it returns the detecting result in JSON.

Listing 4.1: Face++ detecting response

```
{
  "face": [
    {
      "attribute": {
```

[2] http://www.faceplusplus.com/

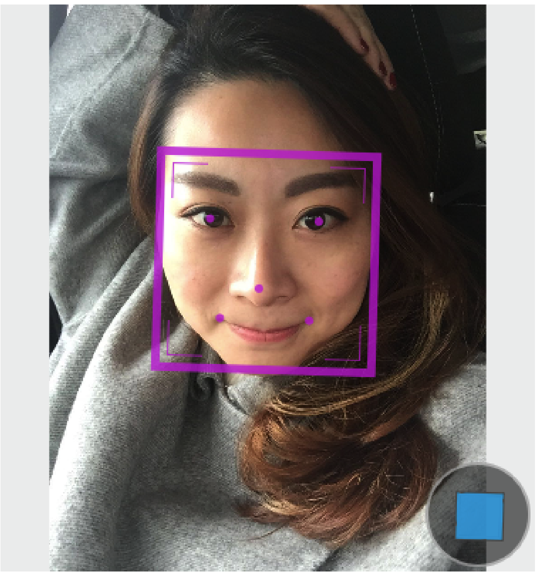


Figure 4.4: The Profile Picture Containing a Face

```
        "age": {
          "range": 5,
          "value": 22
        },
        "gender": {
          "confidence": 99.9999,
          "value": "Female"
        },
        "race": {
          "confidence": 83.8305,
          "value": "Asian"
        }
      }
    }
  ]
}
```

User Attribute Calculation

User attributes will be calculated through data collected from social media, and enriched from other processing functionalities such as Home-locator, Image-based Demographic Extractor. The properties to be computed are user age, gender, the radius of gyration, home location, city role.

Faceplusplus service helps in extracting user gender, age. Home locator determines user home address. The city role is defined differently in different social media platform. According to the Chapter 3.1.4, on the Weibo platform the city and country information is provided by the user, therefore, we use it to determine if this user's city role, i.e. if the city code and province code is equal to the Shenzhen, this user is determined as a Resident, otherwise a Local Tourist. On the Twitter and Instagram, we identify the city role with the help of home locator. Which means if a user's home address is in a particular city, for instance in Rotterdam, this user is identified as Resident. If the user's home address is not in the city but still in the country, he(she) is identified as a Local Tourist. Otherwise, he(she) is a Foreign Tourist.

The radius of gyration is calculated based on all the places a user has visited during a specified period. We implement an algorithm according to Chapter 3.1.5 to calculate the value for each user.

4.2 Social-Glass Extension

The Social-Glass system is a large-scale urban data social analytics demo from the Delft University of Technology. It has a powerful and flexible framework support collecting, analysing and visualising social media information in the city. It currently supports Twitter and Instagram. As part of this study, we are getting to integrate Weibo Crawler into the Social-Glass framework, making it capable of supporting social media platform of Weibo.

4.2.1 Architecture

The architecture of the Social-Glass framework and the integration of Weibo Crawler is shown as in the Figure 4.5. The whole structure consists six modules. Web UI in response to the requests from Web management. Listener & Crawler module holds the extensions of social platform data collectors, where Weibo extension should be added. Other modules, the Pipeline, Visualisation and Configuration, are the standard components of the system.

4.2.2 Implementation

As different social media platforms have distinct data collecting mechanism, to integrate Weibo Crawler as an extension of the Social-Glass framework, we encapsulate the Nearby User Listener and Weibo User Timeline Crawler as two extension components and integrate them into the Listener & Crawler module. In the configuration module, we added a set of configuration files dedicated to Weibo Crawler, such as APP Keys, API settings, and so on. In the database module, according to the rules of Social-Glass framework, we build a set of SQL files for generating all tables for supporting Weibo data crawling.

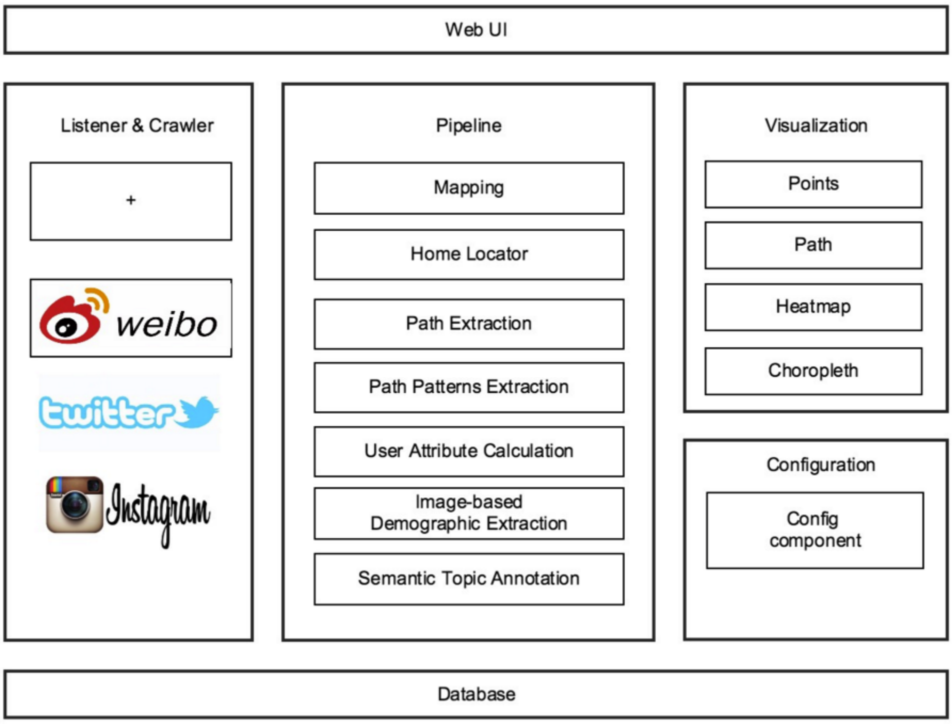


Figure 4.5: Architecture of Integrating Social-Glass with Weibo Crawler

Chapter 5

Analysis

To answer research questions, we analyse the dataset collected on four social media platforms, across two cities, and in a period of two weeks from Oct 1st, 2015 to Oct 14th, 2015. In this chapter, this dataset will be analysed from three aspects in accordance with which in sub-research questions, i.e. the choice of social media platforms, the user individual city role and demographics, and the diversity of PoI.

5.1 Analysis Structure

Dataset to be analysed in this study contains extensive information, for multiple elements and huge capacity. We employ a set of metrics from three aspects to analysis the dataset, to reveal the differences and answer research questions. To this end, the dataset is explored in a sequence accordance with three sub-research questions, where:

For the section **(1) the selection of social media platforms**, and section **(2) the individual user role and demographics**, such items are analysed:

- The total number of users on each social media platforms, and the
- normalised user intensity on each platform.
- The total number of posts on each social media platforms and the average number of posts sent by each user.
- The total number of PoIs visited by each user groups (i.e. per individual city roles, age, and gender), and the PoI intensity in the city.
- The PoI preference of each user groups.
- The average of the radius of gyration of each user groups, and the normalised value in two cities.
- The temporal analysis w.r.t. posts: Gini index, temporal distribution, and the similarity of different temporal distributions.
- The spatial analysis w.r.t. posts: Gini index, spatial distributions, and the similarity of the various spatial distributions.
- The top words used in the content of posts sent by each user groups.

For the section **(3) the various of PoIs**, analogously, items analysed are:

- The temporal analysis w.r.t. PoI categories: Gini index, temporal distribution, and the similarity of different temporal distributions.
- The spatial analysis w.r.t. PoI categories: Gini index, and spatial distributions.

5.2 Social Media Platform Influence Analysis

We start with analysing influences of different social media platforms for studying people's activities in the city of Rotterdam and Shenzhen (RQ1). In the Table 5.1, an overview of the dataset is shown, including city scale, the number of posts, users, PoIs, normalised ratios, as well as its statistical dispersion measured by Gini coefficient. The data of posts and users are collected from Twitter, Instagram and Weibo, and the data of PoIs from Rotterdam is collected from Foursquare while the in Shenzhen it is collected from Weibo.

Table 5.1: Overview of Posts, Users, and PoIs, with Interesting Figures Underlined.

	Rotterdam		**Shenzhen**
Inhabitant	624,799		10,547,400
Area (km^2)	325.79		1,991.64
	Twitter	Instagram	Weibo
#User	1,184	11,165	87,274
#Post	5,445	21,604	242,495
#PoI	1,313	1,585	6,349
%User	1.19%	11.21%	87.60%
User / Population	.00190	.01787	.00827
Post per User	4.599	1.935	2.779
Post / POI	4.147	13.630	38.194
PoI / PoI_all	n/a	n/a	0.03
Avg(RoG)	1,413.61	1,298.95	3,244.22
Avg(RoG) / City size	.094	.087	.046
Gini.Temporal	.305	.343	.263
Gini.Spatial	.771	.850	.750

[1] According to Table 3.1, Rotterdam population: 624,799; Shenzhen population: 10,547,400.
[2] The number of all captured PoIs on Weibo in Shenzhen is 207,384.
[3] In Metres.
[4] The diameter of city bounding-box for normalising the Avg(RoG), in (metres). Rotterdam: 15,000; Shenzhen: 70,000.

Observation of the City Scale, Number of Users, Posts, and PoIs

It is evident that the city scale has significant differences between two cities, i.e. the number of inhabitants in Shenzhen is around 16 times than in Rotterdam, and Shenzhen is more than six times bigger in area than Rotterdam. The bounding box/circles of social media platforms for Rotterdam and Shenzhen are shown in Figure 3.1 and

Figure 3.2 in Chapter 3. The #users and #posts in this two cities show the similar pattern, i.e. the number of users on Weibo in Shenzhen is around nine times than users on Instagram in Rotterdam, and the number of posts is more than ten times. However, when normalising the number of users on different social media platforms with the population of the city, it shows that the value for Instagram is almost twice of that on the Weibo, representing that the user intensity of Instagram in Rotterdam is higher than that of Weibo in Shenzhen. In the city of Rotterdam, figures show that Instagram is more popular than Twitter. This may because Instagram is an image-based social media platform while Twitter is a text based one, given that taking a picture is far more natural than writing something. This is further explained in the table 5.2, in which the number of identified young users and the number of identified tourists (Commuters and Foreign tourists) on the Instagram are far more than those on the Twitter. Besides, the number of PoIs visited by social media users shows that Weibo users visit more PoIs than Twitter and Instagram users.

Observation of Mean of Posts, Posts per PoI, PoI Intensity

However, the mean of posts per user shows that in average user sends more posts on Twitter than on Instagram and Weibo. However, the average of the number of posts per PoI in different social media shows different patterns, i.e. the Weibo users posts approximately 38 posts in a PoI while on Instagram and Twitter is 14 and four posts. The proportion of PoIs visited by Weibo users in Shenzhen with respect to the total PoIs we crawled in Shenzhen is around 0.03. Due to lack of the total number of PoIs crawled in Rotterdam, this proportion for Twitter and Instagram is not available in this study and has been put for the future research.

Observation of Average of Radius of Gyration

The average of the radius of gyration in the table indicates that Weibo users in Shenzhen move about three times longer than Twitter and Instagram users in Rotterdam. In contrast, when normalising the radius of gyration with the diameter of the bounding box of the cities, the result shows that Twitter and Instagram users move wider than the Weibo users on the city size.

Observation of Temporal Distribution

The Gini.temporal index shows that Weibo posts in Shenzhen are distributed more balanced than Twitter and Instagram in Rotterdam from temporal aspects. The Figure 5.1 illustrates the temporal distribution of posts in these three platforms in 24 hours of a day, with the interval of half an hour. The lines of Twitter and Instagram have similar patterns which are sharp while Weibo has a different, smoother one. Before Dutch lunch time (normally between 12:00 to 13:30, considering a slight delay on social media), the number of activities on Twitter are more than that on Instagram. After the afternoon, starting before the dinner time, there are more Instagram activities present, until the midnight. On the Weibo, in contrast, the activities are lower than other two platforms until the 19 o'clock, the Chinese dinner time. It then increases dramatically and reach the peak on the mid-night. This finding is in agreement with the result

Table 5.2: Number of Users with Demographic and Individual City Roles

		Rotterdam		**Shenzhen**
		Twitter	Instagram	Weibo
Age	Teen (0-15)	67	966	19,733
	Young (16-30)	296	2,681	13,021
	Mid-aged (31-45)	239	1,028	1,709
	Older (46+)	76	134	193
	Unknown	506	6,356	52,618
Gender	Male	589	3,542	31,865
	Female	370	4,541	54,642
	Unknown	225	3,082	767
Individual city role	Resident	453	4,753	37,503
	Commuter	401	2,769	49,771
	Foreign tourist	282	3,299	0
	Unknown	48	344	0
All		3,552	33,495	261,822

Table 5.3: Temporal Distribution Similarity of User Activities in 24-hours of a day on three platforms

	Twitter	**Instagram**	**Weibo**
Twitter	.00000	.01083	.03462
Instagram	.01083	.00000	.02746
Weibo	.03462	.02746	.00000

for China in 2015[1] that more than one-third of white-collar workers work overtime more than 5 hours per week, particularly, workers from industries of IT, Electronic and Internet work overtime more than 9.3 hours. The data shows that from around the night the usage of Weibo is increasing. Table 5.3 lists the Jensen-Shannon divergences between each temporal distribution of social media platforms. According to the data, activity temporal distributions on Twitter and Instagram are more similar, followed by Instagram with Weibo. Twitter and Weibo have the least similar distributions. Which means from the activity temporal perspective, Twitter and Instagram is closer, than Instagram with Weibo, and least close is Twitter with Weibo.

Observation of Spatial Distribution

Similarly, the Gini.spatial index shows dispersion distinctions of posts of the three platforms: the Weibo posts are distributed more balanced while the Twitter and Instagram are distributed more unbalanced. According to Figure 5.2, posts on the Weibo are distributed in almost all the area, particularly in the southern west of the city. Instead,

[1] http://article.zhaopin.com/pub/view/217834-26071.html

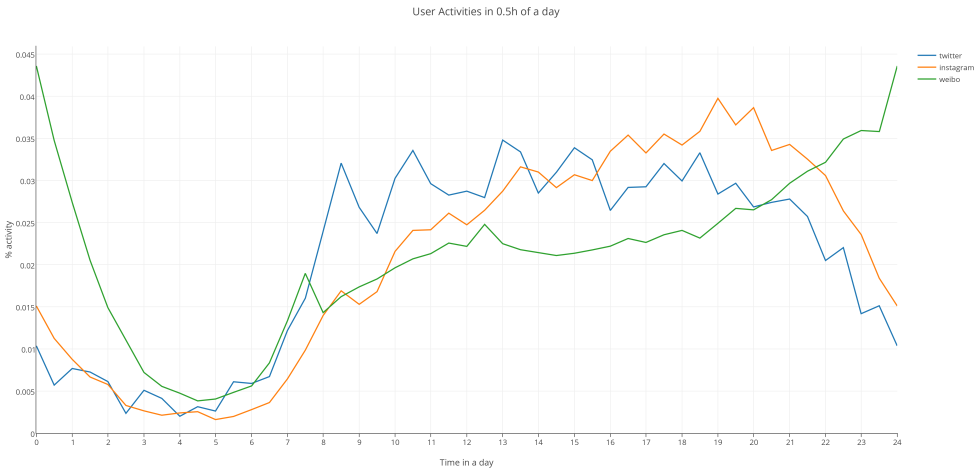
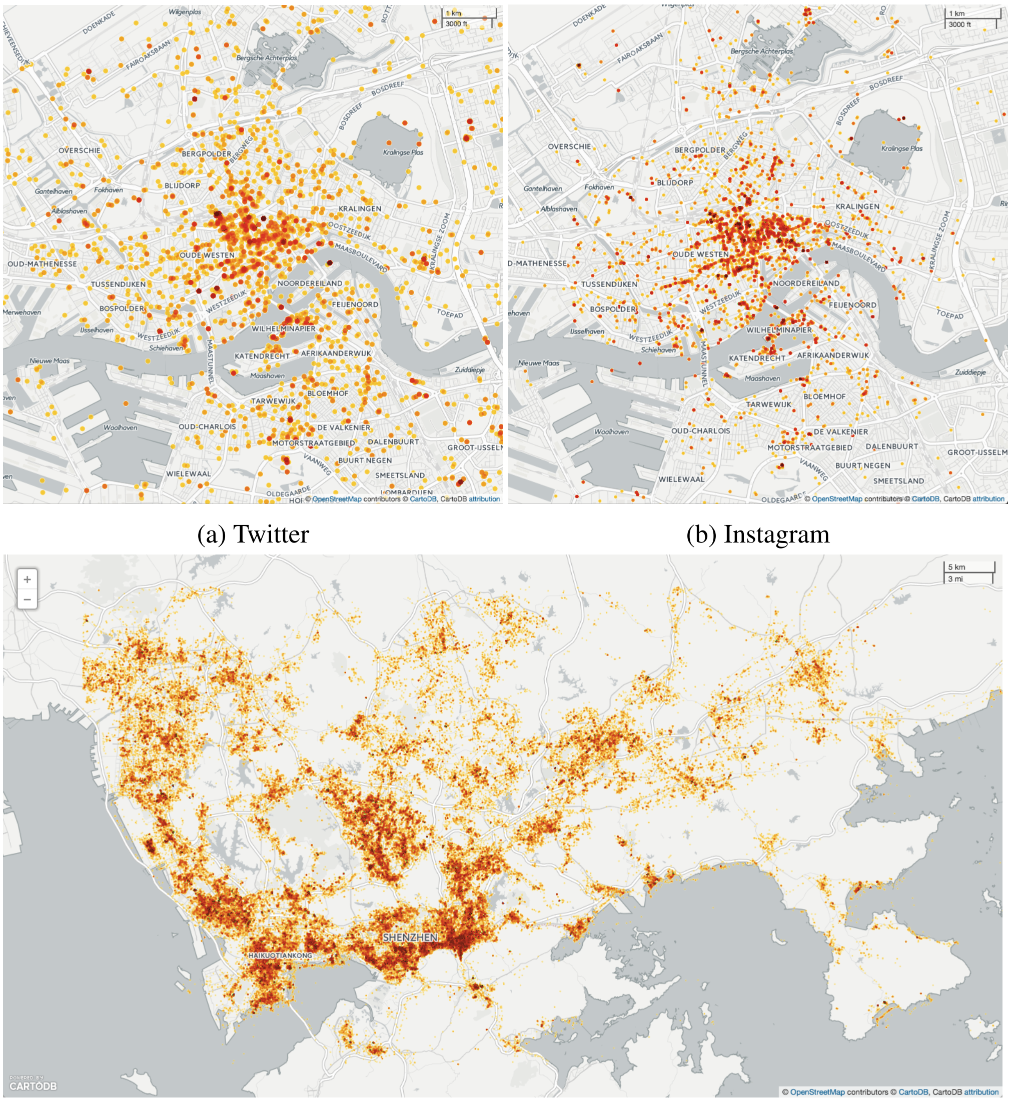


Figure 5.1: User Activities in 24-hours of a day

the posts on Twitter and Instagram are distributed around the city centre, with more in the North of the river and less in the south. Further, Instagram with higher Gini spatial index than Twitter indicates that Instagram is wider used and thus more popular than Twitter.

Observation of Word Use

To understand what they people talked about during each active period, we calculate the word frequencies after removing stop words in Dutch, English and Chinese, and list top words in the Table 5.4. On the hypothesis that users may use different words during the different period in a day (morning, noon, evening, night), we analyse the words for Twitter, Instagram and Weibo in the morning (07-10 o'clock), in the noon (12-14 o'clock), and in the evening (18-21 o'clock). For Weibo, we do an extra period to see what the users talk about in Shenzhen in the midnight (21-02 o'clock), due to the increasing number of posts during that time. The result shows that most of the top words used on Twitter posts are in Dutch, and they are more neutral and related to specific topics or places (Central, Station, Markthal, Erasmusbrug), while on Instagram there are lots of English words, and they are more emotional and concern the daily life (love, food, beautiful, fashion, art, happy, coffee, life, wonderful). An explanation of this finding may be based on the premise that more tourists are using Instagram, and the image-based social media is more used for reflecting people's daily life. In contrast, top words used on Weibo posts are mixed. From time periods perspective, the differences show: in the morning, top words on Instagram concern the daily life and emotion in the morning (morning, good morning), on Weibo the similar words are observed (Good morning, come on, smile, work). In the lunch time, Instagram posts contain meal related top words (food, lunch), while Twitter and Weibo posts do not show the similar top words.



(a) Twitter

(b) Instagram

(c) Weibo

Figure 5.2: Post Distribution, Twitter and Instagram in Rotterdam, and Weibo in Shenzhen

Observation of PoI Preference

To understand which places the users on different social media platforms prefer to visit, the popularity of PoIs across various categories is listed in the Table 5.5, with top 3 categories underlined. First of all, the distinctions of data scale in PoI popularity on different social media is due to the different amount of geo-referenced posts captured in the dataset, i.e. Weibo has the most PoI visits. According to the data, users on various social media platforms have clearly different PoI category preferences. On the Twitter, the top 3 PoI categories are, respectively, Shop & Service, Professional & Other Places, and Travel & Transport. While Instagram users visit PoIs of Shop & Service, Food, and Arts & Entertainment. Similar to Twitter, Weibo users prefer Travel

Table 5.4: Top words in posts of Twitter, Instagram and Weibo in multiple active periods

City	S.M.	Period	Top Words
Rotterdam	Twitter	07 to 10	Letsel, photo, Centraal, posted, vertraagd, minuten, Nieuwe, Brandweer, vandaag, Veiligheidsregio, Ochtend, Amsterdam, terug, Station, Zie, Geld, fail, Groene
		12 to 14	inzet,vertraagd, Breda, Centraal, terug, Geld, Maasstadweg, Gravendijkwal, Hilledijk, Amsterdam, fail, Station, Letsel, Groene, Kleiweg, Gebouwbrand, vandaag, Markthal, klein, Just
		18 to 21	inzet, vertraagd, minuten, Centraal, Breda, Zie, terug, Geld, Letsel, photo, fail, Just, Amsterdam, Station, Erasmusbrug, posted, Melding, vandaag, rijdt, contact, Kleine, brand
	Instagram	07 to 10	holland, love, morning, architecture, city, goodmorning, travel, food, day, beautiful, Netherlands, fashion, art, new, weer, nature, 2015, Good, happy, sky, coffee, super, life, wonderful, europe, amazing, autumn
		12 to 14	love, holland, food, art, architecture, lunch, like, fashion, time, city, weer, Netherlands, new, travel, instadaily, night, life, markthal
		18 to 21	netherlands, love, architecture, night, food, beautiful, art, Netherlands, travel, friends, instagood
Shenzhen	Weibo[1]	07 to 10	Good morning, song, Shenzhen, cry, hee hee, like, come on, hit list, smile, doge, crazy, sun, work, life, happy
		12 to 14	cry, Shenzhen, share, hee hee, song, laughing, like, pictures, glutton, vacation
		18 to 21	Hee hee, Shenzhen, song, like, pictures, laughing, refueling, smile, crazy, hit list
		21 to 02	Good night, cry, hee hee, Shenzhen, like,, rain tomorrow, bye, sad, moon, music, sleep, vacation

[1] Words on Weibo are translated from Chinese.

& Transport, Residence, and Professional & Other Places. The distinct PoI preference indicates that users on Instagram tend to send posts when they are having meals or visiting shops or museums. While Twitter and Weibo users tend to post in the office, or at home, or on the way. This may indicate that Twitter and Weibo are similar social media platforms. Another explanation for this finding may be based on the premise that image based social media, the Instagram, is more used in a colourful context, i.e. the Shop, the Food, and the Arts.

Table 5.5: PoI Preference across categories on different Social Media Platforms, with Top 3 categories Underlined

	Rotterdam		**Shenzhen**
PoI Category	Twitter	Instagram	Weibo
Arts & Entertainment	333	2,458	4,227
College & University	246	504	11,523
Food	542	3,147	16,311
Nightlife Spot	214	1,672	5,399
Outdoors & Recreation	279	1,102	29,867
Professional & Other Places	1,045	1,731	35,270
Residence	184	553	38,220
Shop & Service	1,058	7,552	32,322
Travel & Transport	555	1,566	56,411
Event	0	3	0

5.3 User Relationship and Demographic Influences Analysis

In the second sub-research question, we hypotheses that the use roles and demographics observed from social media plays a role in studying differences in human activities in urban environments. To investigate this premise, we analyse the dataset regarding individual user roles and demographic characteristics in this section.

5.3.1 Individual Role

Observation of the Number of Users, Posts, and PoIs

Table 5.6 shows the user activities across different individual roles that are available in the observed social media dataset in two cities. We have observed more users recognised as Residents than Commuters on both of social media platforms in the city of Rotterdam. In contrast, there are more Commuters than Residents in Shenzhen (R: 42.97%, C: 57.03%). An explanation of this may be based on the premise that Shenzhen has become the largest migrant city in China[2], which provides enormous job opportunities attracting a large number of workers who, however, living nearby Shenzhen for cheaper costs. The number of posts is also evident that residents post more than commuters in Rotterdam while commuters post more than residents in Shenzhen. Concerning the PoI visiting, the table shows that residents in Rotterdam visit more PoIs than commuters, while the number of PoIs in Shenzhen attended by residents and commuters is almost the same, with a slightly more visited by commuters.

Observation of Mean of Posts, Posts per PoI, PoI Intensity, and Radius of Gyration

While considering the number of users, the average number of posts per user on all social media platforms in two cities follow the same pattern, which the residents post

[2] http://www.goodarticle.info/Article/Business/Manufacturing/201511/592854.html

Table 5.6: User Activity Statistics across different Individual Roles that are available in the dataset through lens of social media in Rotterdam and Shenzhen, with Interesting Figures Underlined. R. = Resident, C. = Commuter.

	Rotterdam				**Shenzhen**	
	Twitter		Instagram		Weibo	
	R.	C.	R.	C.	R.	C.
#User	453	401	4,753	2,769	37,503	49,771
#Post	2,785	930	9,738	4,097	116,199	123,597
#PoI	1,044	308	1,233	658	5,192	5,551
%User	53.04%	46.96%	63.19%	36.81%	42.97%	57.03%
User / Population	.00073	.00064	.00761	.00443	.00356	.00472
Post / User	6.148	2.319	2.049	1.480	3.098	2.483
Sig.(Post/User)	2.707		.402		.435	
Post / POI	2.668	3.019	7.898	6.226	22.380	22.266
Sig.(Post/POI)	.249		1.182		.081	
PoI / PoI_all	n/a	n/a	n/a	n/a	.025	.027
Avg(RoG)	1387.908	1639.628	1386.097	1345.021	2554.033	2569.099
Avg(RoG) / City size	.093	.109	.092	.090	.036	.037
Gini.Temporal	.316	.404	.339	.362	.260	.268
Gini.Spatial	.782	.822	.851	.855	.796	.723

[1] According to Table 3.1, Rotterdam population: 624,799; Shenzhen population: 10,547,400.

[2] The translated Sig.(Post/User) is the standard deviation measuring the extent of the distinction between residents and commuters on the same social media platform on the ratio of posts per user.

[3] The Sig.(Post/PoI) is the standard deviation measuring the extent of the distinction between residents and commuters on the same social media platform about the ratio of posts per PoI.

[4] The number of all captured PoIs on Weibo in Shenzhen is 207,384.

[5] In Metres.

[6] The diameter of city bounding-box for normalising the Avg(RoG), in (metres). Rotterdam: 15,000; Shenzhen: 70,000.

more than commuters. The extent of distinctions between residents and commuters for the average number of posts per user on different social media platforms are measured with the metric of Sig(Post/User), where Twitter has the higher value (distinctions) than Instagram and Weibo. The average number of posts sent from a PoI shows that residents on Twitter sent more posts per PoI, but on Instagram and Weibo it is reverse. The metric Sig(Post/PoI) further indicates that the Instagram residents and commuters have a larger difference than Twitter and Weibo users in average posts per PoI. As regard to the mean of the radius of gyration, it is clear that commuters on Twitter and Weibo move longer than residents, but on Instagram, the residents move longer. When normalising the average radius of gyration with the diameter of the city, the results show that Twitter and Rotterdam users travel wider with the city size than Shenzhen users.

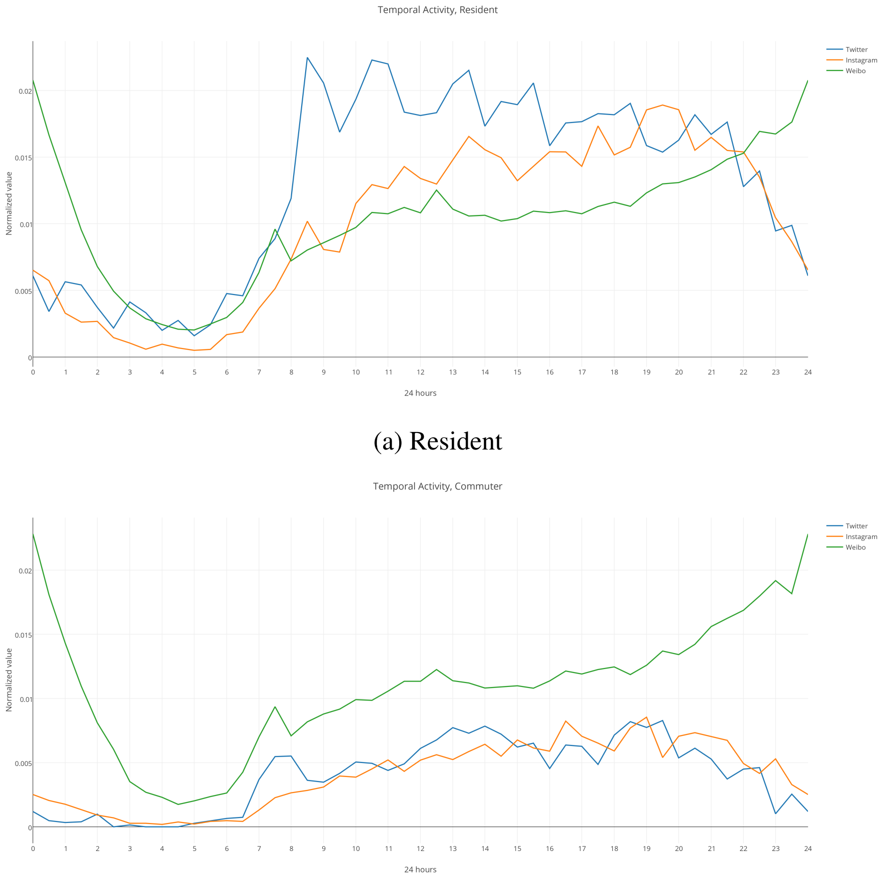
Observation of Temporal Distribution

The Gini.Temporal reports that the commuters' active time in a day is distributed slightly more balanced than that of residents. This natural based on the premise that commuters post during their commute. Figure 5.3 illustrate the temporal activities of residents and commuters in both Rotterdam and Shenzhen through the lens of three social media platforms. It shows that the major activities taken place in Rotterdam is during the day time, and decreased after around 9 pm. In contrast, activities in Shenzhen are keeping increasing slightly after 5 pm and reach the peak in the mid-night.

Table 5.7: Temporal Distribution Similarity of User Activity, Resident

	Twitter	**Instagram**	**Weibo**
Twitter	.00000	.01152	.02235
Instagram	.01152	.00000	.01427
Weibo	.02235	.01427	.00000

This in certain extent reflects the active preference of different city, i.e. Shenzhen has a more fantastic nightlife than Rotterdam. Table 5.7 and Table 5.8 report the Jensen-Shannon divergences between each temporal distribution across individual city roles revealing the similarity between them. It shows that with the respect of user activity temporal distribution, Twitter and Instagram is more similar, followed by Instagram with Weibo, and the Twitter with Weibo is least similar.



(a) Resident

(b) Commuter

Figure 5.3: Temporal Activity Distribution in 24 hours of a day in Rotterdam and Shenzhen

Table 5.8: Temporal Distribution Similarity of User Activity, Commuter

	Twitter	**Instagram**	**Weibo**
Twitter	.00000	.00422	.05472
Instagram	.00422	.00000	.04626
Weibo	.05472	.04626	.00000

Observation of Spatial Distribution

In Gini.spatial, results show that the posts of commuters are distributed more balanced than that of residents in Rotterdam. In contrast, Shenzhen shows in an inverse way. From Figure 5.4, it is evident that posts of commuters in Rotterdam are more focus to the city centre, particularly close to the central station and tram or railway stations while posts of residents are more widely spread. In Shenzhen, according to the table, distribution of posts from residents are more cluttered than commuters.

Observation of Word Use

To examine what the differences do the residents and commuters talk about, we counted the most frequent words in their posts shown in Table 5.9. The resident users on Twitter talk more about local places in the Rotterdam city (e.g. Maasstadweg, Wytemaweg, Hilledijk.Centraal, Molewaterplein, Hofplein), while commuters also talk about places outside Rotterdam city (e.g. Breda, Amsterdam). On the Weibo in Shenzhen, the words frequently mentioned by resident and commuters are almost the same. One possible reason behind this is that given the city is growing very fast, the life gap between commuters and residents in Shenzhen has become fewer and fewer.

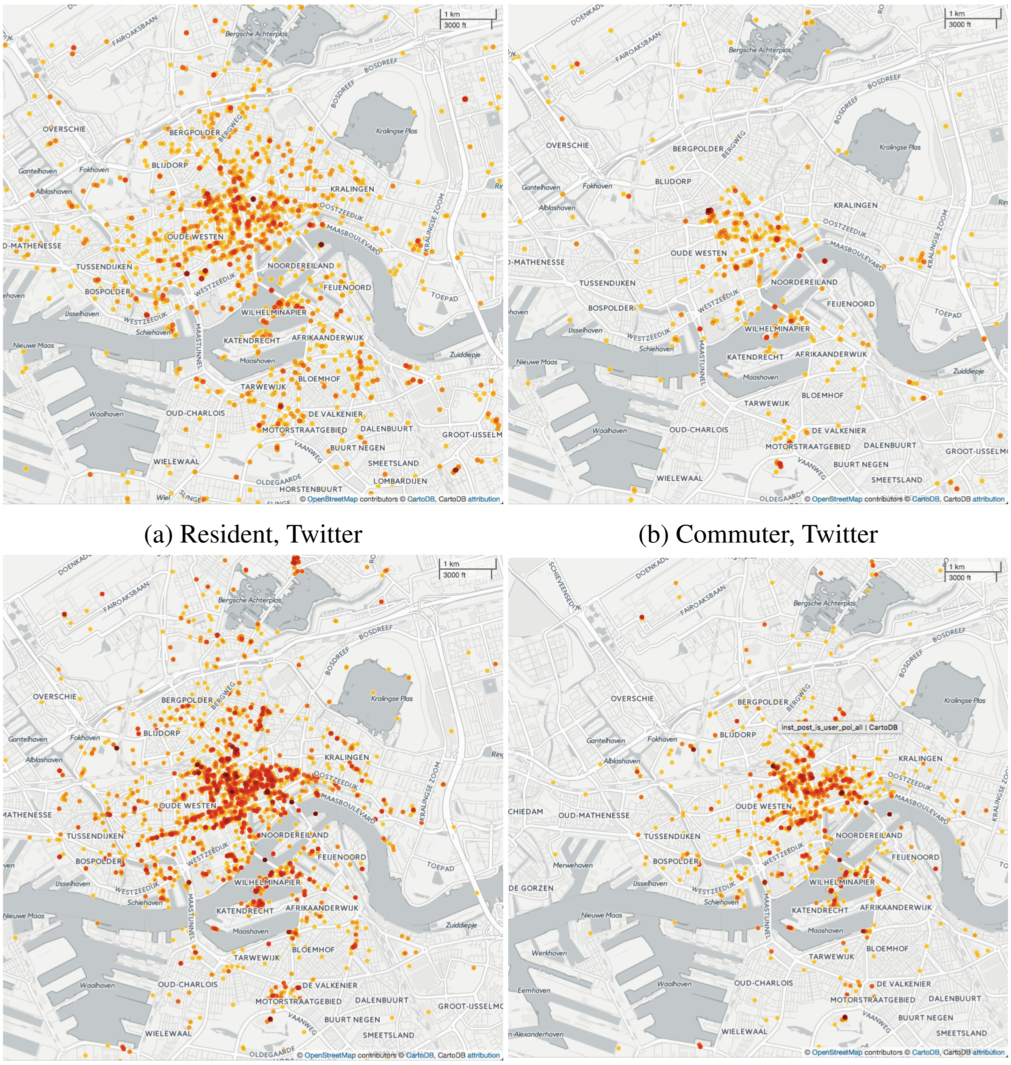
Observation of PoI Preference

Analogous to PoI preference, Table 5.10 reports the number of visits across different PoI categories by the users in various individual roles. According to the data, Twitter residents in Rotterdam prefer Professional & Other Places, Shop & Service, and Food venues, while commuters visit Travel & Transport, Shop & Service, and Arts & Entertainment the most. However, either on the Instagram in Rotterdam or on the Weibo in Shenzhen, though the number of visits of different PoI categories is different, yet the top 3 most visited PoI categories concerning residents and commuters are the same, i.e. on the Instagram it is Shop & Service, Food, and Arts & Entertainment, on the Weibo it is Travel & Transport, Residence and Professional & Other Places.

5.3.2 User Age

Observation of the Number of Users, Posts, and PoIs

Table 5.11 shows the data statics through the lens of social media in two cities across four categories of user age, i.e. Teen from 0 to 15, young from 16 to 30, mid-age from 31 to 45, and old who are older than 46. From the table, it is clear that Rotterdam has more young users than any other age groups on Twitter and Instagram. However, teen



(a) Resident, Twitter

(b) Commuter, Twitter

(c) Resident, Instagram

(d) Commuter, Instagram

Figure 5.4: Spatial Distribution of posts in Rotterdam across Roles

Table 5.9: Top Words used in social media by users of different Roles

City	S.M.	Role	Top Words
Rotterdam	Twitter	R.	inzet, Letsel, Gravendijkwal, stadweg, Wytemaweg, Groene, Hilledijk, Centraal, Station, contact, Molewaterplein, klein, Hofplein, brand, Brandweer,Veiligheidsregio, auto,opnemen, Dienstverlening, Politie, Kleine, meldkamer
		C.	minuten, vertraagd, Breda, Holland, Zuid, Centraal, terug, Zie, Geld, fail, Amsterdam, vandaag, rijdt, Station, Netherlands, Markthal, Ahoy, photo, netherlands, Kunsthal, nl, weer, Opschaling, Gebouwbrand, Just, Hotel, nieuwe, Hiring, day, Erasmus, Euromast
	Instagram	R.	love, bnw, food, nl, holland, night, fashion, weekend, www, day, city, art, architecture
		C.	netherlands, love, holland, architecture, VIP, bw, ig, night, food, 2015
Shenzhen	Weibo[1]	R.	song, laughing, crazy, smiling, rose, life, good morning, light rain, holiday, friends, good
		C.	cry, Shenzhen, songs, hee hee, good night, Wei Chen, crazy, mask, holidays, National Day, people strive

[1] Words on Weibo are translated from Chinese.

Table 5.10: PoI Preference across categories in different Individual User Roles, with Top 3 categories Underlined. R. = Resident, C. = Commuter.

	Rotterdam				**Shenzhen**	
	Twitter		Instagram		Weibo	
PoI Category	R.	C.	R.	C.	R.	C.
Arts & Entertainment	134	103	955	749	2,188	2,039
College & University	198	30	303	82	5,314	6,209
Food	354	100	1,847	521	7,660	8,651
Nightlife Spot	133	39	908	380	2,806	2,593
Outdoors & Recreation	192	54	525	208	13,588	16,279
Professional & Other Places	783	75	899	368	17,450	17,820
Residence	135	10	231	94	19,816	18,404
Shop & Service	546	239	3,268	1,317	15,596	16,726
Travel & Transport	195	275	665	339	27,475	28,936
Event	0	0	0	2	0	0

Table 5.11: Statics of User Activities on social media in two cities across different Age Group, with Interesting Figures Underlined. T. = Teen, Y. = Young, M. = Mid-Age, O. = Old.

	Rotterdam								**Shenzhen**			
	Twitter				Instagram				Weibo			
	T.	Y.	M.	O.	T.	Y.	M.	O.	T.	Y.	M.	O.
#User	67	296	239	76	966	2,681	1,028	134	19,733	13,021	1,709	193
#Post	143	801	544	248	1,634	4,692	2,040	331	57,210	38,495	4,667	594
#PoI	71	264	253	135	401	727	444	111	4,328	3,828	1,361	287
%User	9.88%	43.66%	35.25%	11.21%	20.09%	55.75%	21.38%	2.79%	56.94%	37.57%	4.93%	0.56%
User / Population	.00011	.00047	.00038	.00012	.00155	.00429	.00165	.00021	.00187	.00123	.00016	.00002
Post / User	2.134	2.706	2.276	3.263	1.692	1.750	1.984	2.470	2.899	2.956	2.731	3.078
Sig.(Post/User)	.507				.354				.144			
Post / POI	2.014	3.034	2.150	1.837	4.075	6.454	4.595	2.982	13.219	10.056	3.429	2.070
Sig.(Post/POI)	.532				1.450				5.320			
PoI / PoI_all	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	.021	.018	.007	.001
Avg(RoG)	1361.872	1340.134	1491.652	1645.412	1337.940	1325.285	1381.283	1314.921	2945.690	3259.453	3514.686	2822.315
Avg(RoG) / City size	.091	.089	.099	.110	.089	.088	.092	.088	.042	.047	.050	.040
Gini.Temporal	.524	.390	.453	.470	.366	.355	.354	.422	.266	.260	.251	.264
Gini.Spatial	.830	.801	.816	.808	.854	.857	.858	.847	.752	.752	.771	.815

[1] According to Table 3.1, Rotterdam population: 624,799; Shenzhen population: 10,547,400.

[2] The translated Sig.(Post/User) is the standard deviation measuring the extent of distinction between residents and commuters on the same social media platform with respect to the ratio of posts per user.

[3] The Sig.(Post/PoI) is the standard deviation measuring the extent of distinction between residents and commuters on the same social media platform with respect to the ratio of posts per PoI.

[4] The number of all captured PoIs on Weibo in Shenzhen is 207,384.

[5] In Metres.

[6] The diameter of city bounding-box for normalising the Avg(RoG), in (metres). Rotterdam: 15,000; Shenzhen: 70,000.

Table 5.12: PoI Preference categories across different Age Groups, with Top 3 categories Underlined. T. = Teen, Y. = Young, M. = Mid-Age, O. = Old.

	Rotterdam								**Shenzhen**			
	Twitter				Instagram				Weibo			
PoI Category	T.	Y.	M.	O.	T.	Y.	M.	O.	T.	Y.	M.	O.
Arts & Entertainment	17	66	64	25	211	587	216	29	968	679	83	57
College & University	8	30	24	10	43	115	36	6	2,569	1,856	206	19
Food	15	75	67	44	248	627	250	42	4,089	2,709	277	39
Nightlife Spot	5	42	24	15	119	428	142	16	1,455	875	95	17
Outdoors & Recreation	5	58	19	14	68	260	118	19	6,425	4,624	616	82
Professional & Other Places	11	129	91	31	149	373	181	31	8,606	5,310	794	72
Residence	3	26	34	1	55	99	38	11	8,802	6,119	811	81
Shop & Service	52	157	114	57	525	1,579	770	126	7,556	5,008	649	79
Travel & Transport	15	97	49	25	114	339	175	35	13,877	8,978	929	128
Event	0	0	0	0	2	0	0	0	0	0	0	0

users are the most ones on the Weibo in Shenzhen. This result indicates the degree of technique penetration or social media preference in different user ages in Rotterdam and Shenzhen, which calls for a separation of user age groups in relevant research. This pattern also holds concerning the number of posts sent and the number of PoIs visited by different ages in Rotterdam and Shenzhen.

Concerning the PoI visiting, the table shows that young user identified in Rotterdam visit more PoIs than other age groups followed by Mid-age users, with the least PoIs are visited by teen on Weibo and old users on Instagram. While teen users in Shenzhen visited more PoIs than other age groups followed by young users, and old users visit the least PoIs.

Observation of Mean of Posts, Posts per PoI, and Radius of Gyration

However, when normalising the number of posts with the number of users, the average posts sent per user shows that the old users across two cities send more posts than other age groups. In the meantime, Twitter has the largest distinctions with the average posts per user across age groups, followed by Instagram, and the Weibo has the least differences. The average posts per PoI also follow this pattern, according to the data. Moreover, the figures show that Weibo has larger distinctions on the differences of average posts per PoI across user groups, followed by Instagram, and the smallest is presented on Twitter. The value of the mean radius of gyration of old users on Twitter, mid-age users on Instagram and Weibo travels longer than other age groups on the platforms, respectively. When normalising the average radius of gyration about the city diameter, it shows that users in Rotterdam travel wider than users in Shenzhen.

Observation of Temporal Distribution

According to Gini.Temporal, the active time for teen and old users, are distributed more unbalanced. This holds in all three social media platforms in two cities. One possible reason for this is that the life schedule of teen and old users is more regular than young and mid-age users. Figure 5.5, Figure 5.6, Figure 5.7, and Figure 5.8 shows distribution of posting time in a day in four diagrams with age groups of teen, young, mid-age, and old, respectively. Across all four age groups, Weibo users in Shenzhen are all tend to be more active during evening and night. However, Twitter and Instagram users in different ages in Rotterdam have diverse active time slots: teen users are more active during afternoon and evening, young users are more active during evening, mid-age users are more active during the daytime, old users are more active in the afternoon and evening. Data shown in Table 5.13, Table 5.14, Table 5.15 and Table 5.16 report the Jensen-Shannon divergences between each temporal distribution across age groups revealing the similarity between them. It shows that with the respect of user activity temporal distribution, for teen and mid-age users, Twitter and Instagram is more similar, followed by Instagram with Weibo, and the Twitter with Weibo is least similar. However, this pattern does not hold for young and old users. According to the data, for young users, Twitter and Instagram is more similar, followed by Twitter with Weibo, and the least is Instagram with Weibo. For old users, Instagram and Weibo are more similar, followed by Twitter and Instagram, and Twitter with Weibo is least related.

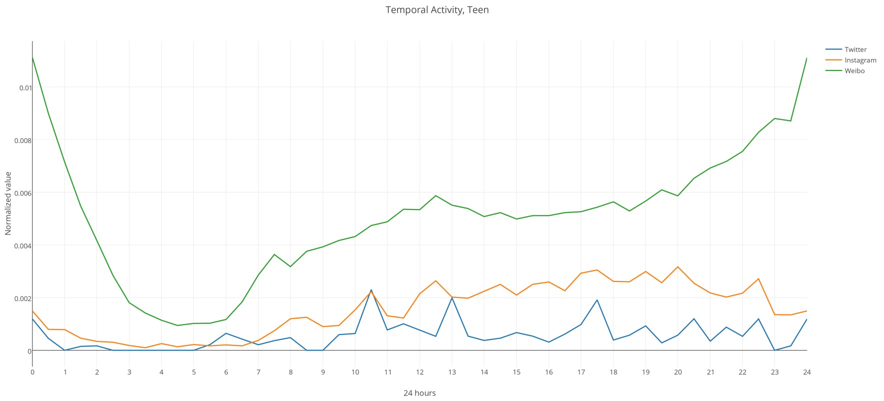


Figure 5.5: Temporal Distribution of posts across age groups, Teen

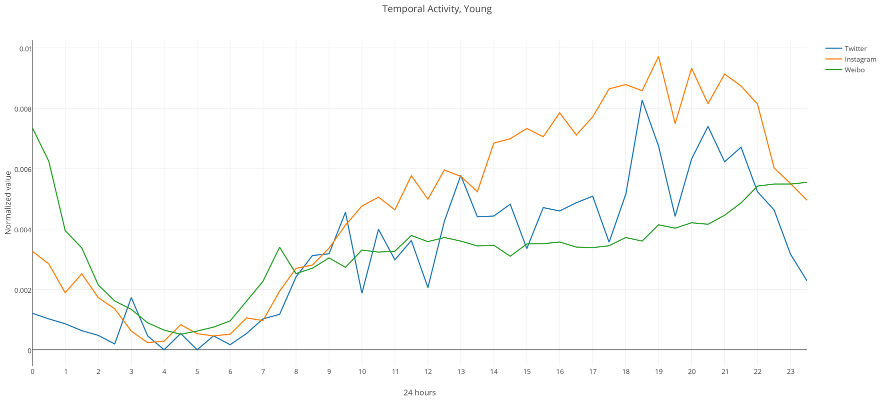


Figure 5.6: Temporal Distribution of posts across age groups, Young

Table 5.13: Temporal Distribution Similarity of User Activity, Teen

	Twitter	**Instagram**	**Weibo**
Twitter	.00000	.00891	.05286
Instagram	.00891	.00000	.02590
Weibo	.05286	.02590	.00000

Observation of Spatial Distribution

Concerning Gini.Spatial, posts from teen users on Twitter in Rotterdam are distributed more unbalanced in the specific area than other age groups. Shown in Figure 5.9, posts from teen users are distributed close to schools in Rotterdam, such as Rotterdam School of Management BV, Erasmus MC, Islamic University of Rotterdam, Codarts Hogeschool voor de Kunsten, Grafisch Lyceum Rotterdam. While old users on Weibo in Shenzhen are identified in the specific area than other age groups, as shown in Figure 5.10. Instead, there is less variation in Instagram usage about users age.

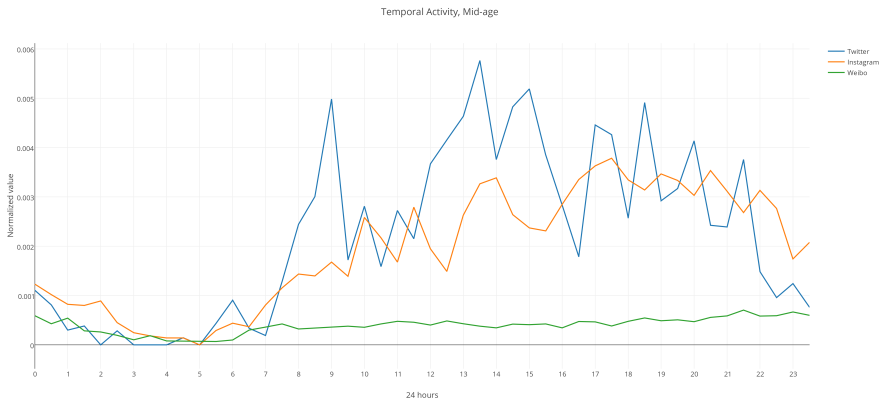


Figure 5.7: Temporal Distribution of posts across age groups, Mid-age

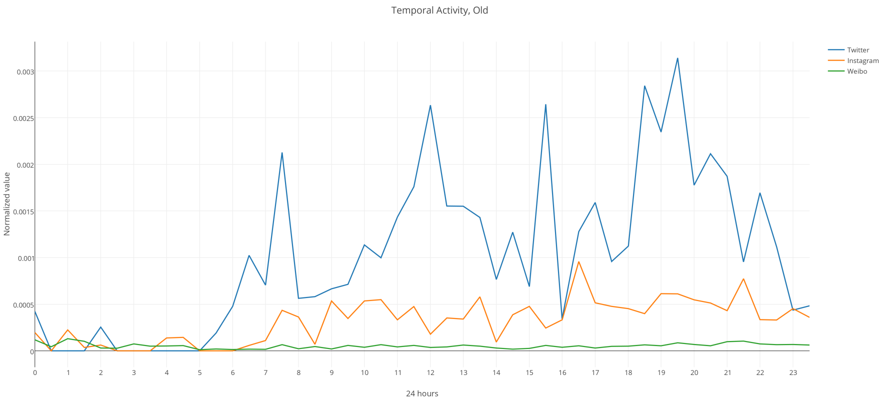


Figure 5.8: Temporal Distribution of posts across age groups, Old

Table 5.14: Temporal Distribution Similarity of User Activity, Young

	Twitter	**Instagram**	**Weibo**
Twitter	.00000	.00616	.00770
Instagram	.00616	.00000	.00895
Weibo	.00770	.00895	.00000

Observation of Word Use

Table 5.17 lists top words in the posts of users in different age groups in two cities. According to the data, both on Twitter in Rotterdam and Weibo in Shenzhen, teen and young users tend to talk about colourful and lovely things in a direct way, to express their emotion. Such as Sevinc (joy), grafisch (graphic), sunshine, Apple, drinking, crib, happy, beauty, wine, inspiration. On the Twitter, the mid-age and old users tend to use relatively more abstract words, and talk about big-scale things, such as coffee, university, society, flygtninge (refugee), cuppa. While on the Weibo in Shenzhen, the mid-age users talk more about family, work, and their life in the posts, such as treasure,

Table 5.15: Temporal Distribution Similarity of User Activity, Mid-age

	Twitter	**Instagram**	**Weibo**
Twitter	.00000	.00386	.01985
Instagram	.00386	.00000	.01428
Weibo	.01985	.01428	.00000

Table 5.16: Temporal Distribution Similarity of User Activity, Old

	Twitter	**Instagram**	**Weibo**
Twitter	.00000	.00637	.01407
Instagram	.00637	.00000	.00305
Weibo	.01407	.00305	.00000

thanksgiving, husband, wife, a lot of work, loneliness. The old users on Weibo in Shenzhen talk more about the whole family, their life, and societal problems, such as family portrait, friends, parents, demolition, Dream of Red Mansions.

Table 5.17: Top Words used in social media by users in different Age Groups

City	S.M.	Age	Words
Rotterdam	Twitter	T.	love, Home, Sevinc, sweats, Grafisch, Lyceum, sunshine, Erasmus, Residences, Apple, University, plek, bridge, local, Station, Zuidplein, Drinking
		Y.	party, Joshua, crib, streetphotography, blacknwhite, Eleonora, Zoo, Airport, Happy, Law, College, Inspiration, beauty, wine
		M.	weer, University, mooie, blogger, Keith, Beurs, Huisartsenposten, Markthal, vandaag, Museum, shopping, nieuwe, avond, Foundation, coffee
		O.	day, Centraal, weer, society, catstagram, morning, lekker, lover, vegan, Danst, onderwijs, harbor, vluchtelingen, cuppa, Groene, buiten
	Instagram	T.	ORGANICTAN, Tan, love, EUROPE, organic, Free, happy, day, Eco, weekend, time, beautiful, skin, amazing, back, like, Sonya, nofilter, Keith, cute, morning, Haring, pretty, volleybal
		Y.	love, day, gaan, avond, picoftheday, healthy, fitfam, fitdutchies, dinner, nice, photooftheday, igersrotterdam, girls
		M.	night, city, architecture, love, weekend, time, weer, day, food, art, new, dag, autumn, que, like, lunch, netherlands, travel, gezellig
		O.	day, love, travel, morning, rondomhetzwaanshals, blackandwhite, food, clouds, vsconetherlands, life, hub, igersbnw, vanaf, demand
Shenzhen	Weibo[1]	T.	love, skin, acne, baby, moisturizing, women, men, treasure, free shipping, clothes, with, delicious
		Y.	Good Morning, Day, effort, birthdays, holidays, happy, Wei Chen, Bang Bang, film, sweety, mother, home
		M.	Life, national day, really, holidays, publishing articles, treasure, finally, a lot of work, Hong Kong, Thanksgiving, dreams, success, husband, wife, worry, kids, loneliness
		O.	Perfect, pros and cons, family portrait, land, friends, relations, owners, life, plan, planning, riding, work, happiness, way of life, love, demolition, parents, Dream of Red Mansions

[1] Words on Weibo are translated from Chinese.

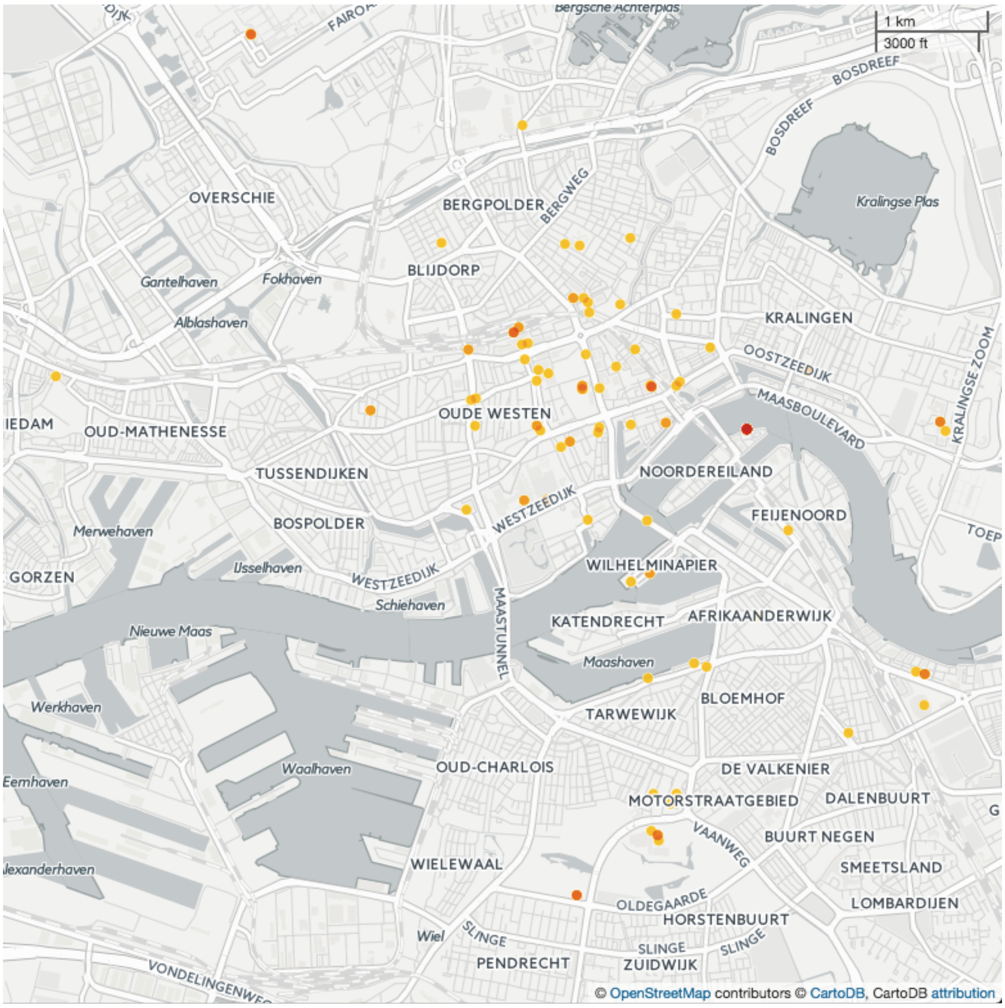


Figure 5.9: Spatial Distribution of posts of Teen users on Twitter in Rotterdam

Observation of PoI Preference

Regarding the PoI preference, Table 5.12 reports on the number of visits of different PoI categories performed by users in various age groups on the Twitter, Instagram, and Weibo. According to the data, in general, the PoI preference of users across different ages in the city of Rotterdam, no matter on Twitter and Instagram, is more diverse than that of users in Shenzhen. It is clearly that Shop & Service, Food, Arts & Entertainment, Professional & Other Places, and Travel & Transport are the most popular PoI categories in Rotterdam. Among all of them, the Shop & Service is the most popular one, which is the top 3 category in across all age groups, followed by Food and Art & Entertainment. In contrast in the city of Shenzhen, the PoI preference is more stable across all ages. The Travel & Transport, Residence, and Professional & Other Places almost occupied in the top 3 PoI categories of all ages. The distinct differences, therefore, calls for a deeper research with more detail information.

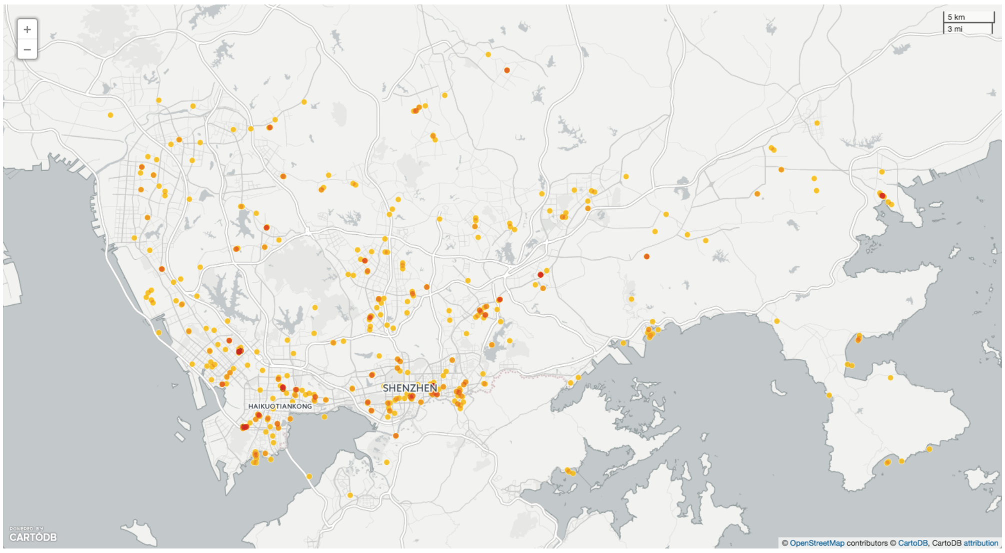


Figure 5.10: Spatial Distribution of posts of Old users on Weibo in Shenzhen

5.3.3 User gender

Table 5.18: Statics of User Activities on social media in two cities across Genders, with Interesting Figures Underlined. M. = Male, F. = Female.

	Rotterdam				Shenzhen	
	Twitter		Instagram		Weibo	
	M.	F.	M.	F.	M.	F.
#User	589	370	3,542	4,541	31,865	54,642
#Post	1,544	935	6,844	8,155	83,469	156,744
#PoI	487	367	914	943	5,111	5,759
%User	61.42%	38.58%	43.82%	56.18%	36.84%	63.16%
User / Population	.00094	.00059	.00567	.00727	.00302	.00518
Post / User	2.621	2.527	1.932	1.796	2.619	2.869
Sig.(Post/User)	.067		.096		.176	
Post / POI	3.170	2.548	7.488	8.648	16.331	27.217
Sig.(Post/POI)	.440		.820		7.698	
PoI / PoI_all	n/a	n/a	n/a	n/a	.025	.028
Avg(RoG)	1442.923	1437.745	1326.243	1322.542	3497.991	3137.883
Avg(RoG) / City size	.096	.096	.088	.088	.050	.045
Gini.Temporal	.350	.374	.338	.355	.233	.280
Gini.Spatial	.805	.818	.856	.860	.746	.753

[1] According to Table 3.1, Rotterdam population: 624,799; Shenzhen population: 10,547,400.

[2] The translated Sig.(Post/User) is the standard deviation measuring the extent of distinction between residents and commuters on the same social media platform with respect to the ratio of posts per user.

[3] The Sig.(Post/PoI) is the standard deviation measuring the extent of distinction between residents and commuters on the same social media platform with respect to the ratio of posts per PoI.

[4] The number of all captured PoIs on Weibo in Shenzhen is 207,384.

[5] In Metres.

[6] The diameter of city bounding-box for normalising the Avg(RoG), in (metres). Rotterdam: 15,000; Shenzhen: 70,000.

Observation of the Number of Users, Posts, and PoIs

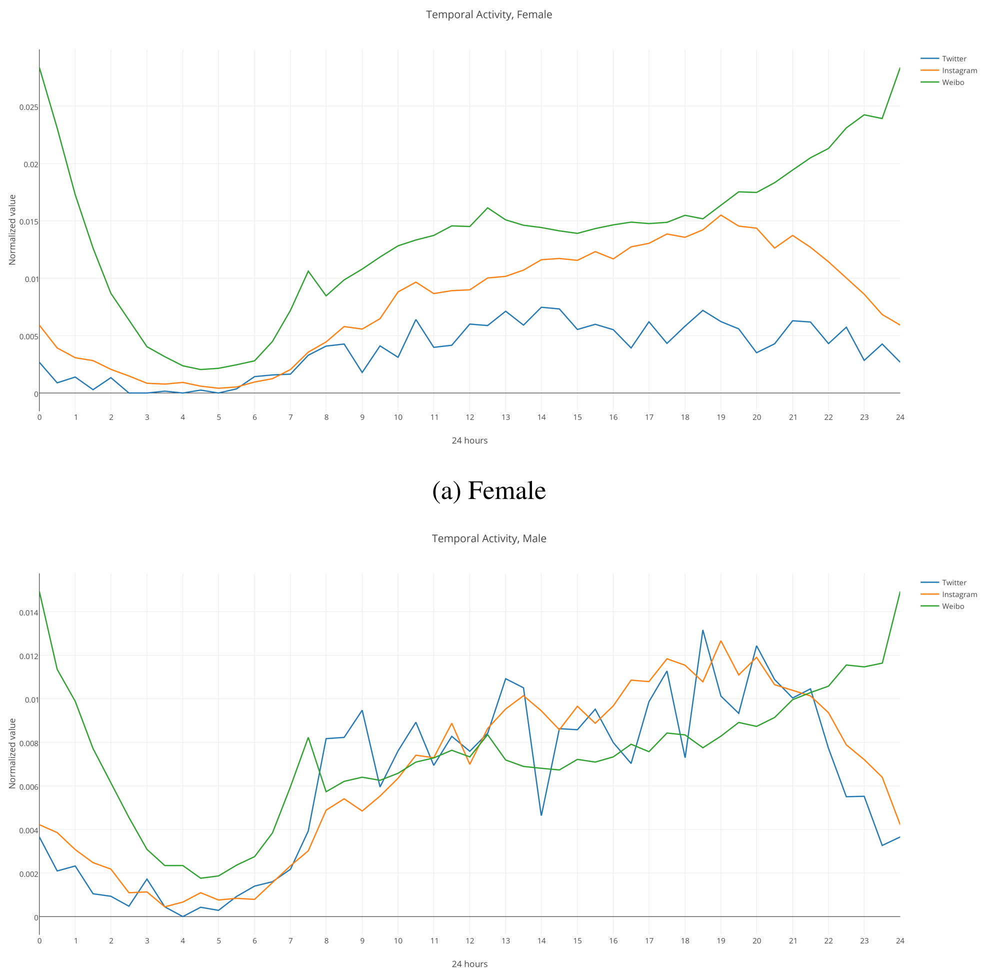
When it turns to the gender of users, according to the Table 5.18, there are more male users than female users in the city of Rotterdam, while there are more female users than male users in the city of Shenzhen. This also holds the number of posts sent by male or female users, and the number of PoIs visited by male or female users. For the number of PoIs, the data shows that male users on the Twitter visit more PoIs than Female users in Rotterdam, while Female users on Instagram and Weibo visit more PoIs than male users.

Observation of Mean of Posts, Posts per PoI, and Radius of Gyration

The average number of posts per user shows that male users in Rotterdam send more posts than female while female users in Shenzhen send more posts than male. Female users in Shenzhen send around 2.869 posts per user, which is the largest number among all. The distinctions of average posts per user on Weibo in Shenzhen is greater than in Rotterdam, followed by Instagram and Twitter. The average posts per PoI on Weibo reaches around 16 for male users and 27 for female users, which are far more than that in Rotterdam. The distinction of average posts per PoI across genders in Shenzhen is also larger than in Rotterdam. Regarding the radius of gyration, the result shows that on Twitter in Rotterdam and Weibo in Shenzhen, male users move a longer distance than female users. However, on the Instagram in Rotterdam, female users travel a longer distance than male. An explanation of this finding may be based on the premise that more travellers are using Instagram, and they accordingly visit attractions in the city. Considering the city size, the average radius of gyration on the city diameter shows that users on Twitter and Instagram move wider than users on Weibo, and Twitter users move slightly longer than Instagram users.

Observation of Temporal Distribution

Concerning the Gini.Temporal, it is even that the active time for female users is distributed more unbalanced than male users in both cities. Figure 5.11 shows the temporal distribution of posts of female and male users in 24 hours. According to the figure, female users, in general, have more deactivate period than male users (female: 03-05 am, male: 04 am). The number of posts from female and male users on the Twitter decrease at the end of the day starting from distinction time point. Female users keep active till midnight while the number of activities from male users decreases sharply from 21:30. Moreover, the "rush hour" of the activity on Twitter is also not same (female: 14h-15h, male: 18h-19h). On the Instagram and Weibo, these features are less identifiable. Table 5.19 and Table 5.19 report the Jensen-Shannon divergences between each temporal distribution across genders revealing the similarity between them. It shows that with the respect of user activity temporal distribution, Twitter and Instagram is more similar, followed by Instagram with Weibo, and the Twitter with Weibo is least similar.



(a) Female

(b) Male

Figure 5.11: Temporal Distribution of posts across Genders in Rotterdam and Shenzhen

Table 5.19: Temporal Distribution Similarity of User Activity, Male

	Twitter	**Instagram**	**Weibo**
Twitter	.00000	.00340	.01579
Instagram	.00340	.00000	.01068
Weibo	.01579	.01068	.00000

Observation of Spatial Distribution

The Gini.Spatial reports that the extent of dispersion of posts distribution from female and male users either in Rotterdam or Shenzhen are similar from the geographical perspective, as shown in Figure 5.12, Figure 5.13, Figure 5.14, and Figure 5.15.

Table 5.20: Temporal Distribution Similarity of User Activity, Female

	Twitter	**Instagram**	**Weibo**
Twitter	.00000	.02135	.07976
Instagram	.02135	.00000	.03000
Weibo	.07976	.03000	.00000

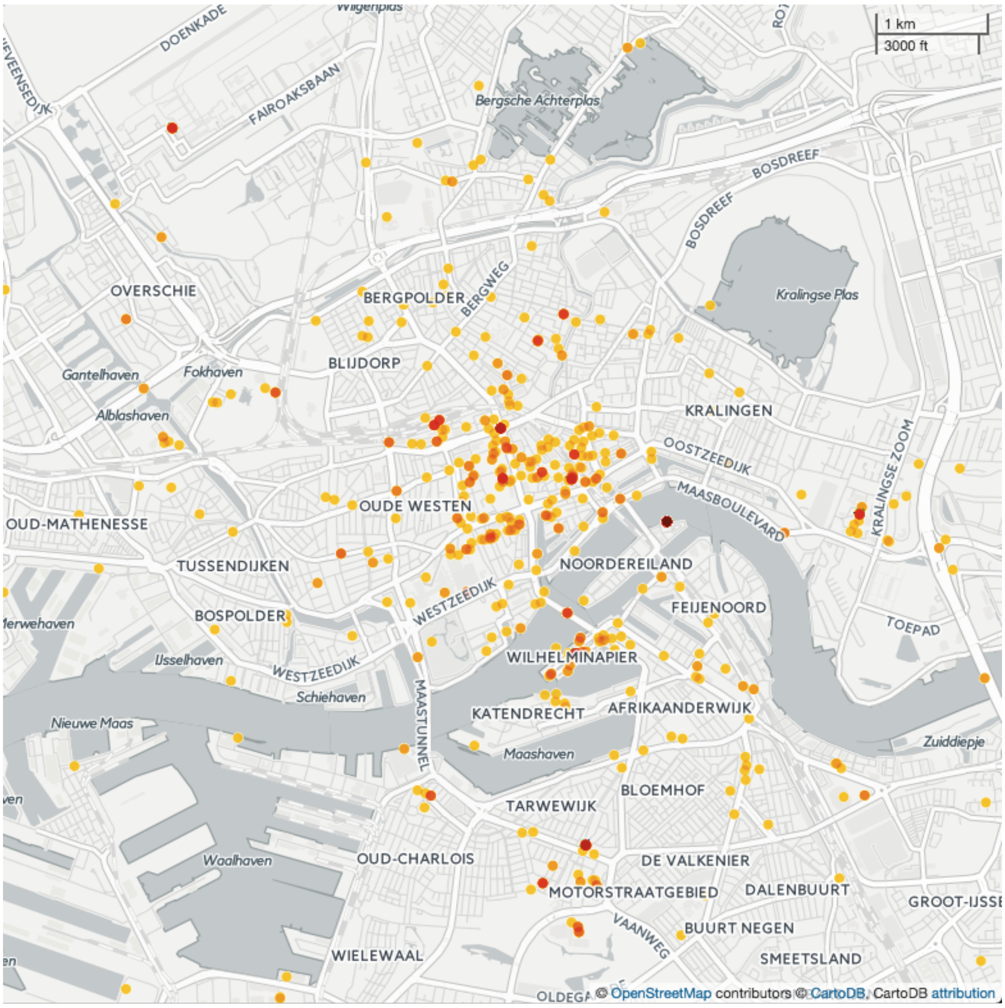


Figure 5.12: Spatial Distribution of posts across Genders, Female on Twitter

Observation of Word Use

Table 5.21 reports on the top words used by male and females on different social media platforms in the city of Rotterdam and Shenzhen. It shows that male users on Twitter have mentioned words related to "photo" or "picture" for more times in their posts, while female users posts content referring to wider topics. On the Instagram, the word use preference of male and female users do not show obvious distinctions. On the Weibo in Shenzhen, the difference is noticeable. Female users talk more about grooming and emotion while male users talk more about general topics.

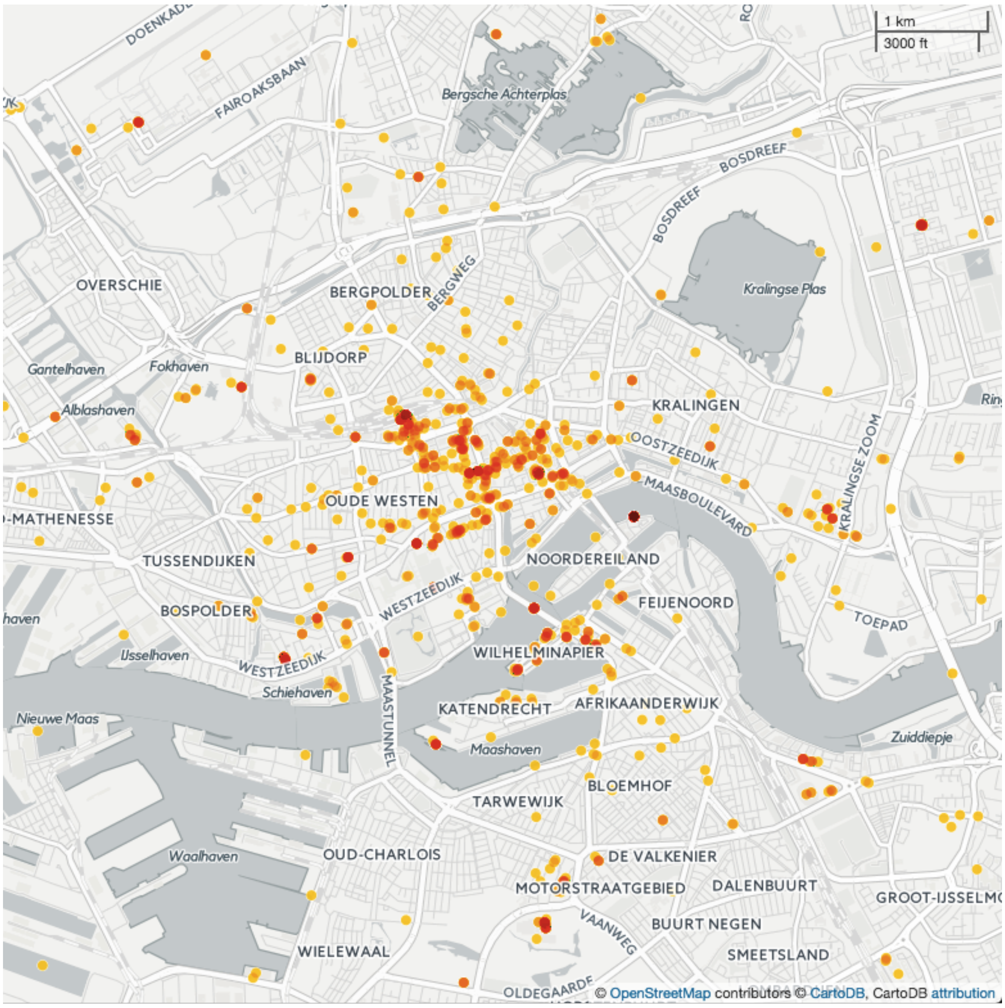


Figure 5.13: Spatial Distribution of posts across Genders, Male on Twitter

Observation of PoI Preference

Concerning PoI preference, Figure 5.22 shows the number of visits of PoIs categories by male and female users on social media platforms. Male and females users in Rotterdam have slightly more diverse PoI category preferences than users in the city of Shenzhen. Male and female users in Rotterdam prefer Arts & Entertainment, Food, Professional & Other Places, and Shop & Service, while the male and female users on Weibo in Shenzhen prefer to Professional & Other Places, Residence, and Travel & Transport. The PoI preference of male and female users in Rotterdam seems more related to more aspects of life while male and female users in Shenzhen are more focused on living. These distinctions calls for a specific research in this area with more information.

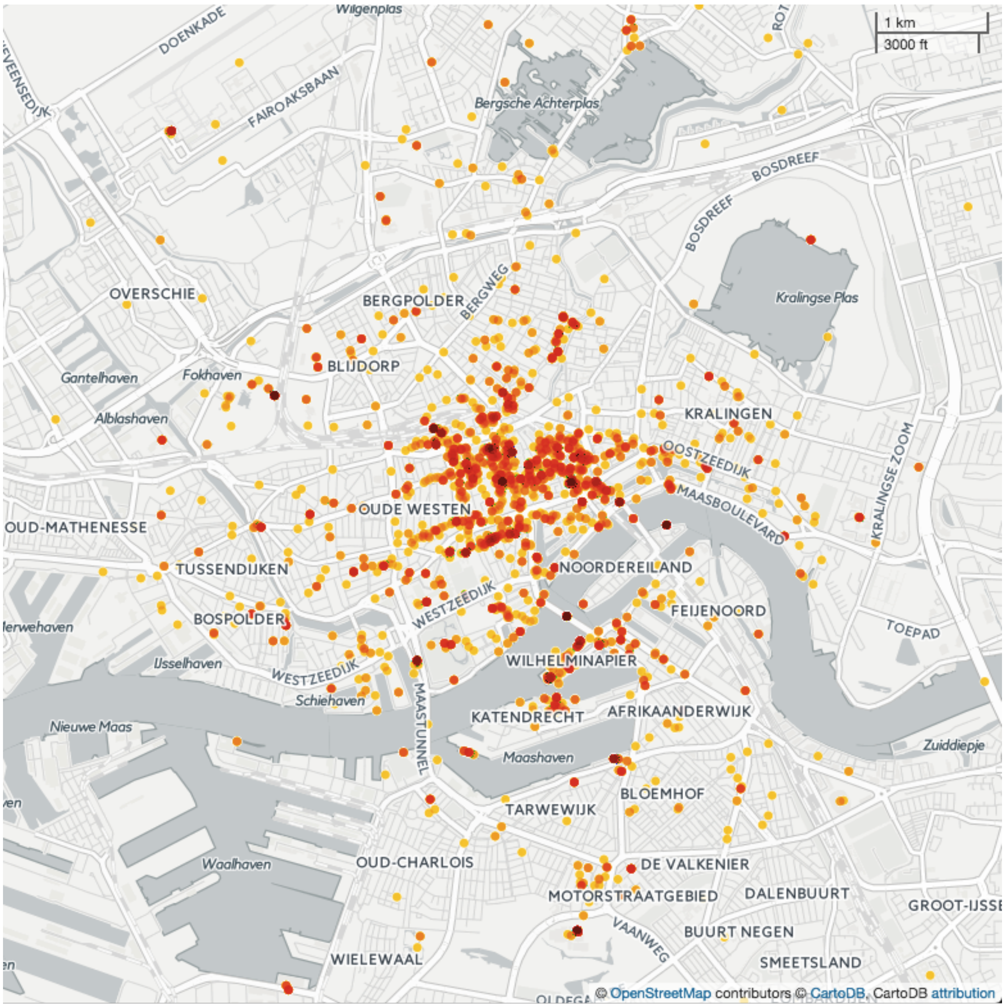


Figure 5.14: Spatial Distribution of posts across Genders, Female on Instagram

5.4 Point of Interests Influences Analysis

To answer the research question RQ3, in this section, we explore the PoI influence on studying user activity through social media in the city. In this work the PoI categories we studied are nine categories which we can get data from social media in Rotterdam and Shenzhen, namely, Food, Arts & Entertainment, Outdoor & Recreation, Professional & Other Places, Shop & Service, Travel & Transport, and College & University, Nightlife Spot, and Residence. In the previous sections, the popularity with regard to different PoI categories have been examined, in terms of the various social media platforms (Table 5.5), the users individual roles (Table 5.10), the users age groups (Table 5.12), and user genders (Table 5.22). In the following sections, the study will be focused on temporal and spatial aspects.

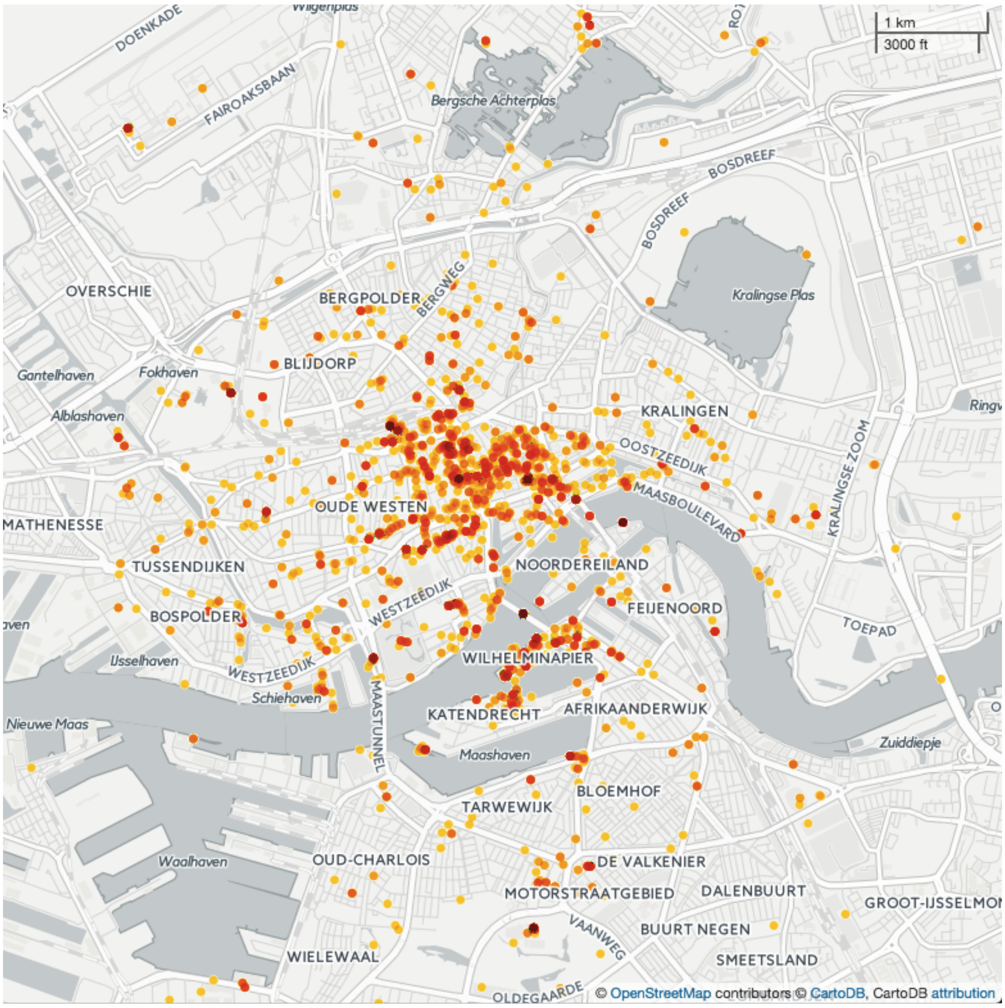


Figure 5.15: Spatial Distribution of posts across Genders, Male on Instagram

5.4.1 Point of Interests Temporal Distribution

To explore the change of PoI popularity of each category, we investigate their temporal distributions shown in Figure 5.18. In general, it is clearly that popularity of all PoI categories on Weibo in Shenzhen have less fluctuation than those on Twitter and Instagram in the city of Rotterdam.

Dispersion and Similarity of PoI Temporal Distribution

To understand the distinction of temporal distribution of each PoI category on Twitter, Instagram and Weibo, we employ Gini.Temporal index to measure their dispersion, and calculate the Jensen-Shannon divergences for each category to measure their similarity. The Gini.Temporal index for each PoI category is shown in Table 5.23, and the temporal distribution similarity is shown in Table 5.24.

According to the data from Table 5.23, for all social media platforms the temporal distribution of user activity on all PoI categories on Twitter are more unbalanced, followed by Instagram. The distribution on Weibo is relatively balanced.

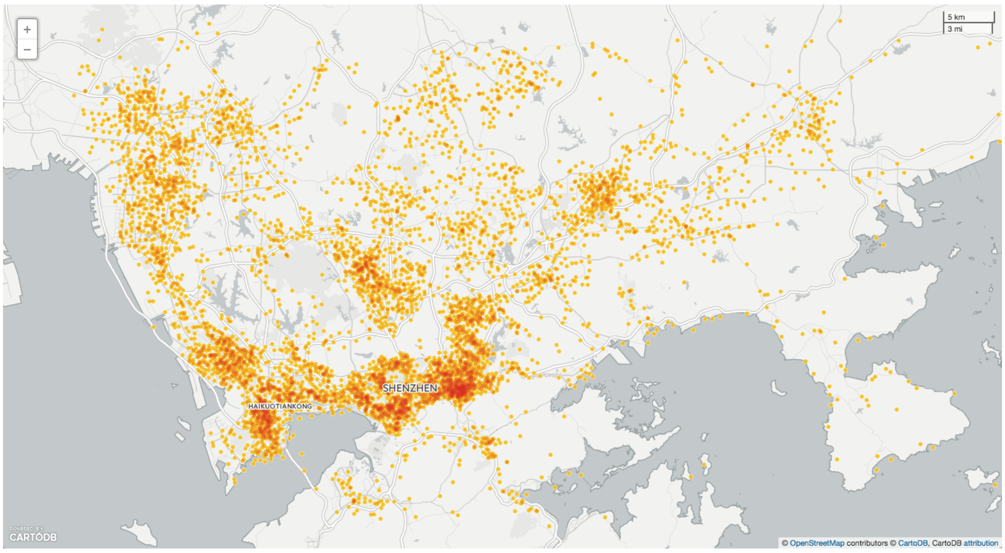


Figure 5.16: Spatial Distribution of posts across Genders, Female on Weibo

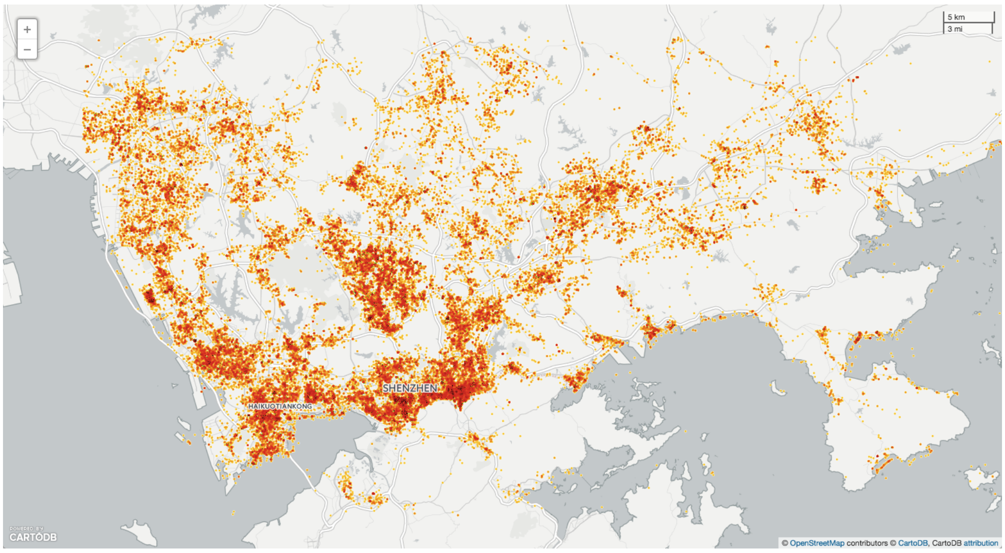


Figure 5.17: Spatial Distribution of posts across Genders, Male on Weibo

Concerning the temporal distribution similarity, Table 5.24 shows the Jensen-Shannon divergence value of Twitter vs. Instagram, Twitter vs. Weibo, and Instagram vs. Weibo, perspective. The box plot in Figure 5.19 shows the distribution of these divergence values. According to the figure, Instagram and Weibo are the most similar across all PoI categories, followed by Twitter with Instagram, and the least similar pair is the Twitter with Weibo.

With on the platform of Instagram and Weibo, the PoI category of Food is more similar than any other categories, while the College & University is the least similar categories. While, on the platform of Twitter and Instagram, the PoI category of Shop & Service is the most similar categories, and the Residence is the least similar categories. This also holds on the platform of the Twitter and Weibo.

Table 5.21: Top Words used by Female and Male users on social medias in city of Rotterdam and Shenzhen

City	S.M.	Gender	Words
Rotterdam	Twitter	M.	Atin, Bijenkorf, Instituut, Nieuwe, party, photoCity, Hague, new, night, photoRotterdam, lekker, Amsterdam, streetphotography, style
		F.	contact, love, day, auto, Piket, atin, super, weekend, Hague, today, Markthal, Happy, goed, dinner, kinderen, Law
	Instagram	M.	day, netherlands, cats, night, city, architecture, holland, love, world, time, 2015, new, weekend, art, food, great
		F.	love, night, city, weekend, weer, beautiful, happy, food, friends, amazing, morning, fun, architecture, good, fashion, best, life, art
Shenzhen	Weibo[1]	M.	Life, living, National Day, friends, microblogging, people, tomorrow, 2015, work, mobile, holidays, good morning, happy, world, typhoon
		F.	Happy, feel, good morning, skin, life, happiness, tomorrow, pay, work, the National Day holiday, light rain, lady, baby

[1] Words on Weibo are translated from Chinese.

5.4.2 Point of Interests Spatial Distribution

Analogous to the PoI spatial distribution, Figure 5.29 shows this distribution in the city of Rotterdam and Shenzhen regarding different PoI categories. According to the map, there are more PoIs from the category of Food in the city of Rotterdam, while more PoIs of Professional & Other Places in the city of Shenzhen. Considering the popularity of PoIs, Figure 5.30 shows the PoI popularity distribution in the city of Rotterdam and Shenzhen on different social media platforms. According to the Figure, in the city of Rotterdam, popular PoIs are generally distributed around the city centre, with more in the north of the river, and less in the south. In contrast, the distribution pattern is clearly different in the city of Shenzhen. According to the Figure 5.30c, the popular PoIs are distributed more in the southwest of the city, and less in the northeast.

Table 5.25 shows the Gini.Spatial index of each PoI category on Twitter, Instagram and Weibo. According to the calculation, PoIs of Arts & Entertainment on Weibo in Shenzhen is located more unbalanced, followed by on Instagram in Rotterdam. On Twitter the PoIs from this category lie relatively balanced. For the category of College & University, the most unbalanced distributed PoIs are observed on Instagram, followed by on Twitter, and least on Weibo. The same pattern is found in the category of Residence and Travel & Transport. For the Food category, PoIs are observed locating more unbalanced on Instagram, followed by Weibo, and least on Twitter. The same pattern is also found in the PoI category of Nightlife Spot, Outdoors & Recreation, Outdoors & Recreation, Professional & Other Places, and Shop & Service.

Table 5.22: PoI Preference across categories by social media users in different Genders, with Top 3 categories Underlined. M. = Male, F. = Female.

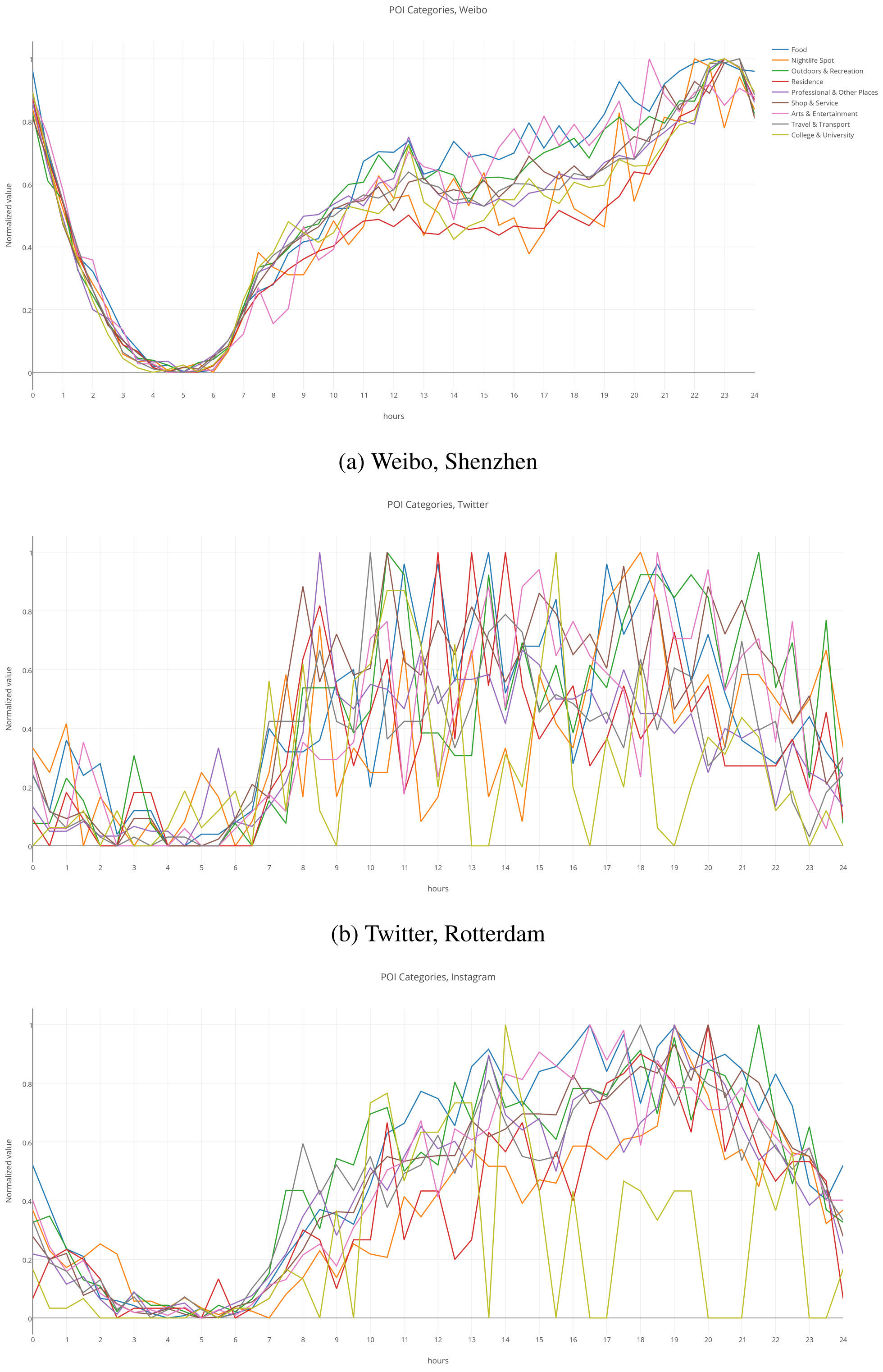
	Rotterdam				**Shenzhen**	
	Twitter		Instagram		Weibo	
PoI Category	M.	F.	M.	F.	M.	F.
Arts & Entertainment	167	84	778	984	1,518	2,664
College & University	46	43	162	204	4,121	7,327
Food	187	99	807	1,173	5,471	10,674
Nightlife Spot	77	47	497	647	1,939	3,424
Outdoors & Recreation	97	71	409	416	10,388	19,106
Professional & Other Places	186	135	582	626	12,229	22,776
Residence	31	59	192	204	13,459	24,438
Shop & Service	375	233	2,479	2,803	10,427	21,512
Travel & Transport	185	75	556	569	19,405	36,542
Event	0	0	1	2	0	0

Table 5.23: Gini index of Temporal Distribution of Visits to PoI Categories

	Rotterdam		**Shenzhen**
POI Category	Twitter	Instagram	Weibo
Arts & Entertainment	.46440	.38500	.29389
College & University	.45088	.44911	.27229
Food	.37801	.35711	.27424
Nightlife Spot	.40500	.35571	.28267
Outdoors & Recreation	.42828	.34887	.27329
Professional & Other Places	.35921	.35307	.25670
Residence	.46418	.42427	.28043
Shop & Service	.34664	.34579	.27333
Travel & Transport	.39674	.34988	.25586

Table 5.24: PoI Category Temporal Distribution Similarity

POI Category	Twitter vs Instagram	Twitter vs Weibo	Instagram vs Weibo
Arts & Entertainment	.06571	.11595	.03949
College & University	.12538	.20030	.13311
Food	.05526	.07371	.03213
Nightlife Spot	.09952	.10186	.04495
Outdoors & Recreation	.05648	.08203	.04159
Professional & Other Places	.05783	.08434	.04631
Residence	.14768	.20740	.09864
Shop & Service	.03131	.05951	.03420
Travel & Transport	.06664	.13298	.04996



(a) Weibo, Shenzhen

(b) Twitter, Rotterdam

(c) Instagram, Rotterdam

Figure 5.18: Temporal Distribution of PoI visiting in Rotterdam and Shenzhen

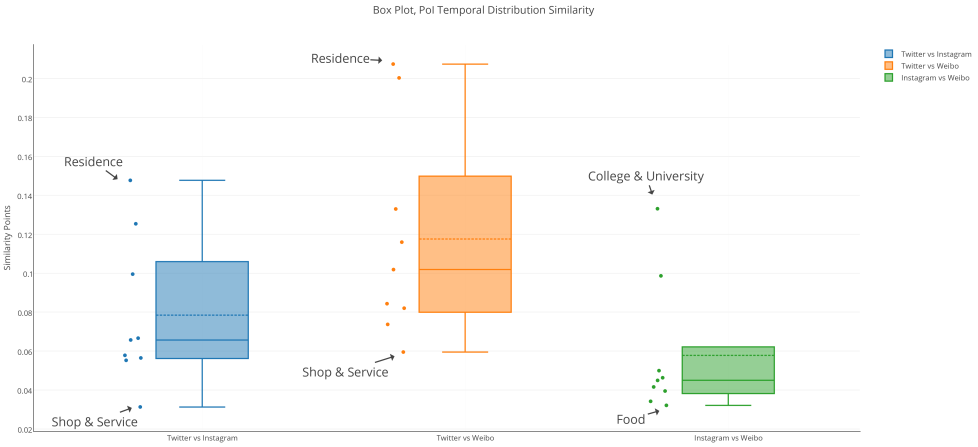


Figure 5.19: The Distribution of PoI Category Temporal Distribution Similarities

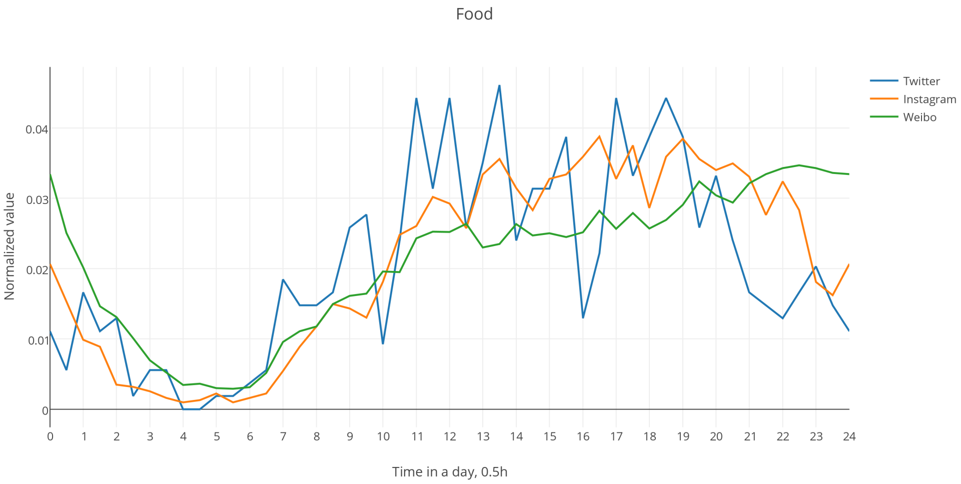


Figure 5.20: Temporal Distribution of User Activity in PoI Category, Food

Table 5.25: Gini index of Spatial Distribution of PoI Location per Categories

	Rotterdam		**Shenzhen**
POI Category	Twitter	Instagram	Weibo
Arts & Entertainment	.83317	.85323	.91964
College & University	.82489	.84554	.79819
Food	.83163	.86851	.83570
Nightlife Spot	.86468	.89106	.88774
Outdoors & Recreation	.77134	.84287	.77932
Professional & Other Places	.77889	.84602	.79230
Residence	.84460	.89335	.83549
Shop & Service	.82142	.89150	.82518
Travel & Transport	.84525	.85060	.77924

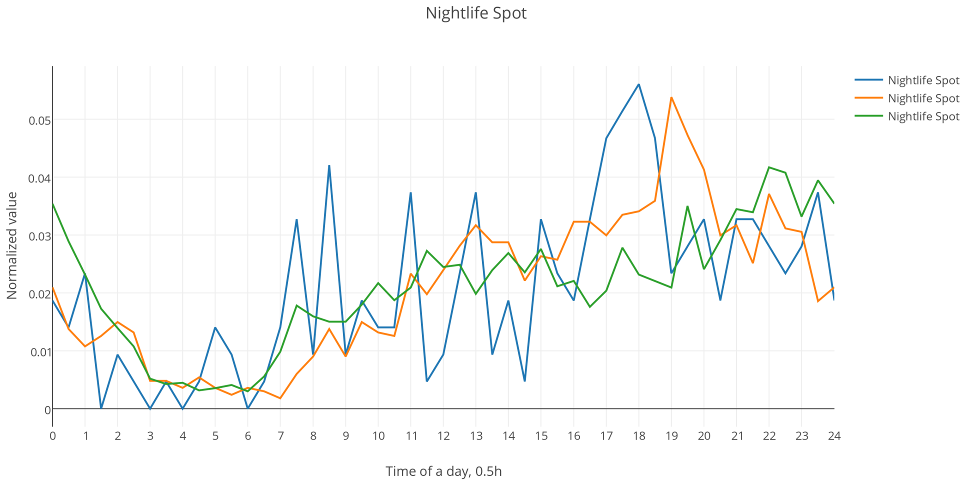


Figure 5.21: Distribution of User Activity in PoI Category, Nightlife Spot

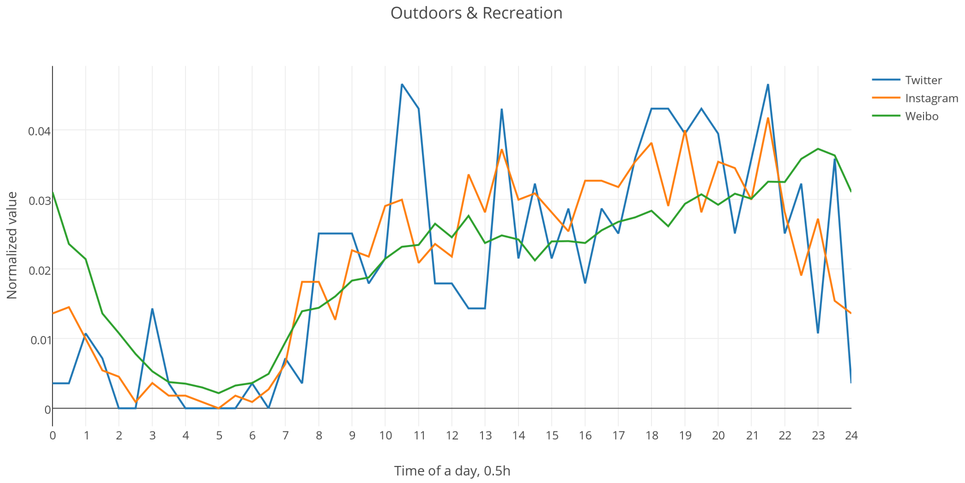


Figure 5.22: Distribution of User Activity in PoI Category, Outdoors & Recreation

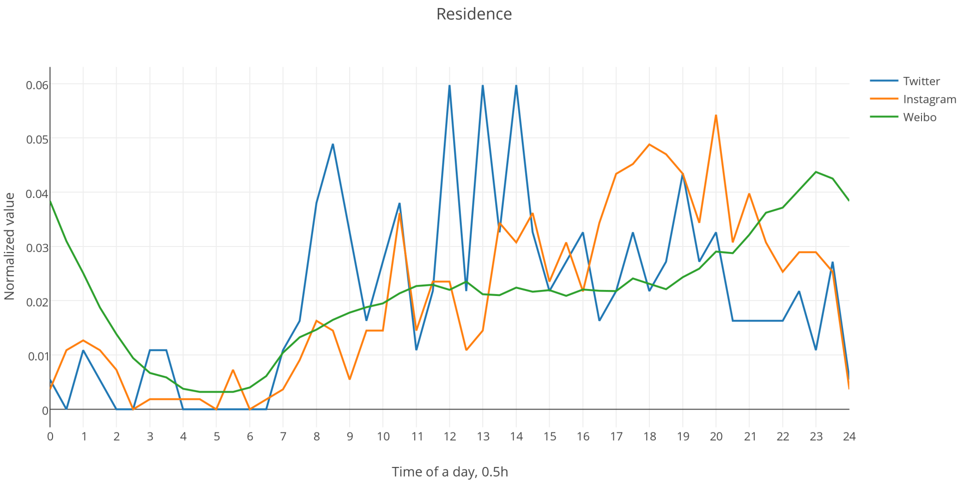


Figure 5.23: Distribution of User Activity in PoI Category, Residence

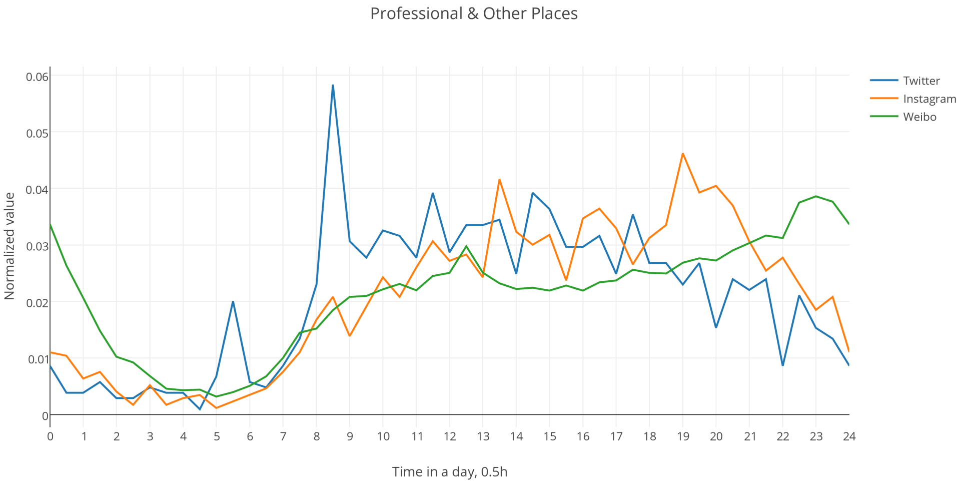


Figure 5.24: Distribution of User Activity in PoI Category, Professional & Other Places

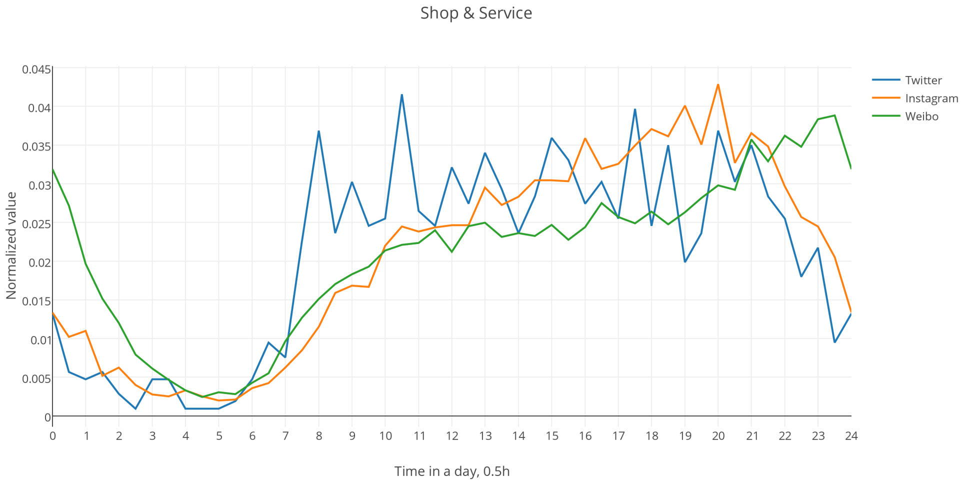


Figure 5.25: Distribution of User Activity in PoI Category, Shop & Service

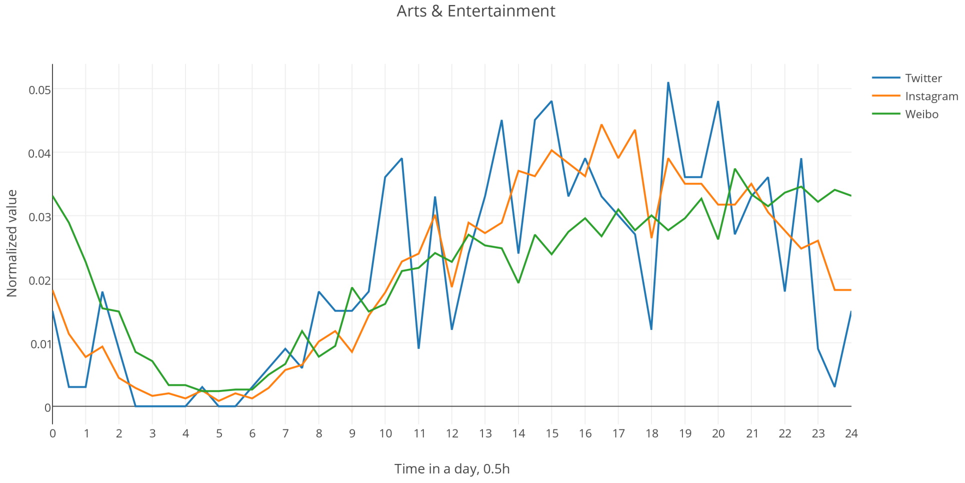


Figure 5.26: Distribution of User Activity in PoI Category, Arts & Entertainment

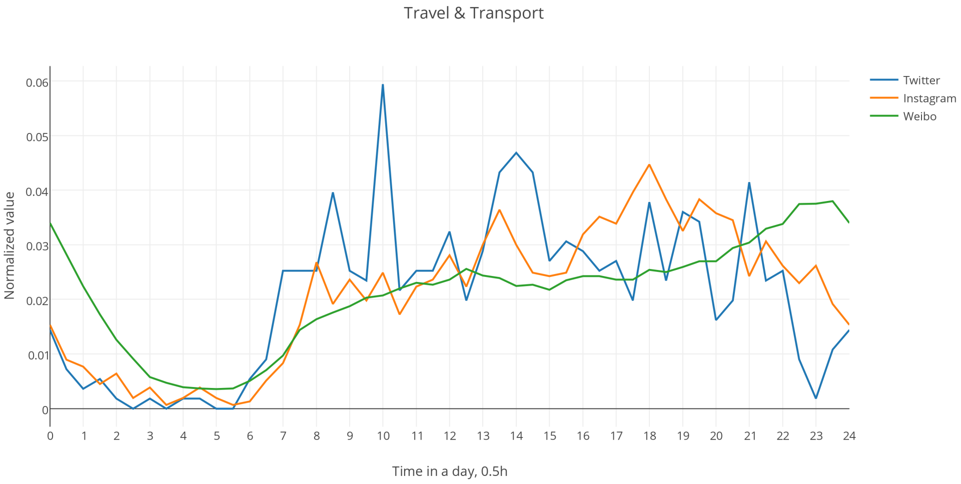


Figure 5.27: Distribution of User Activity in PoI Category, Travel & Transport

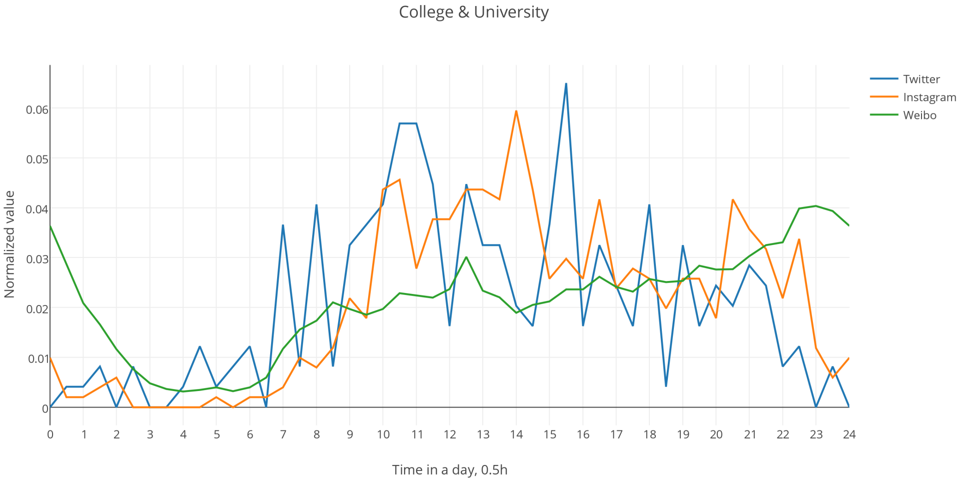
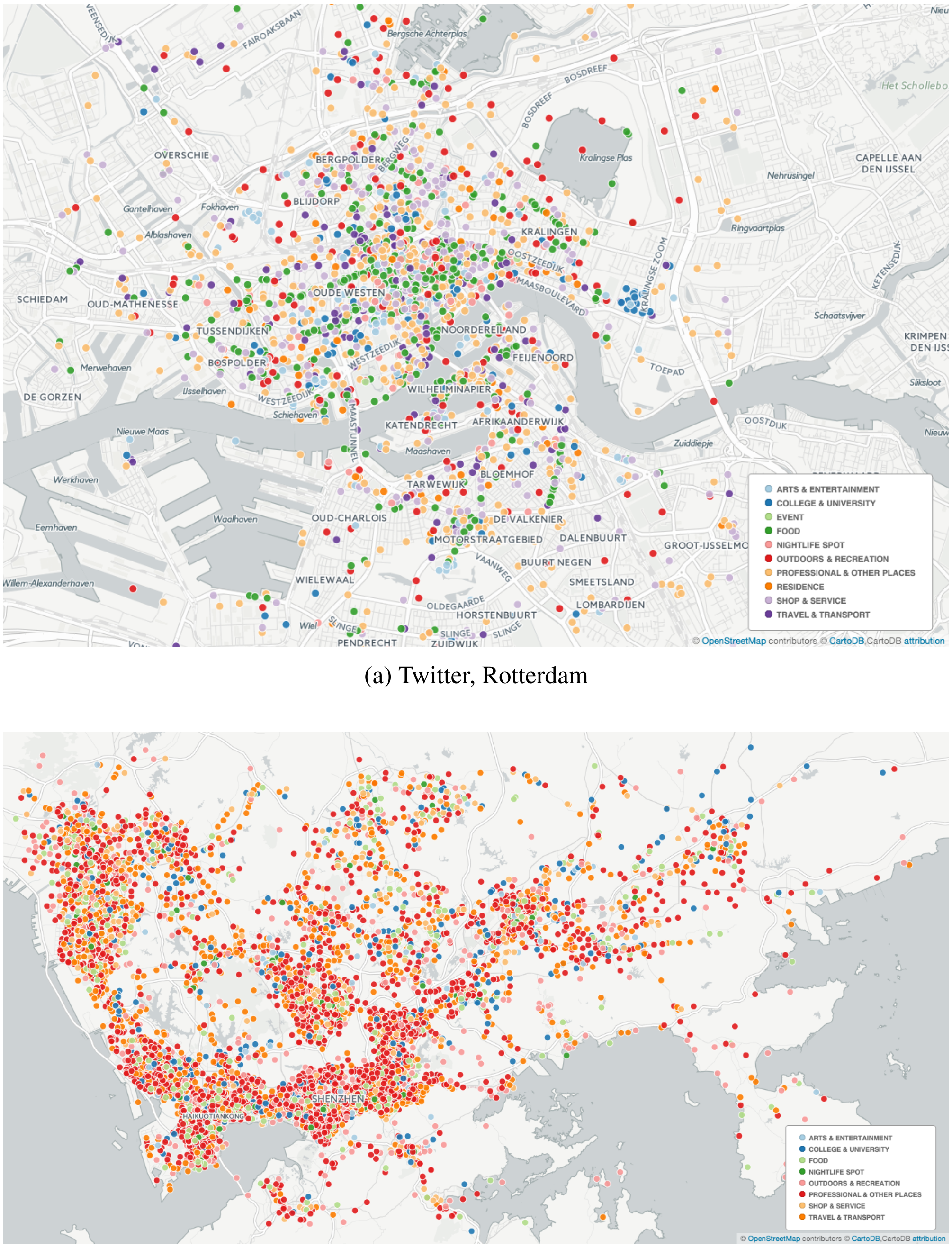


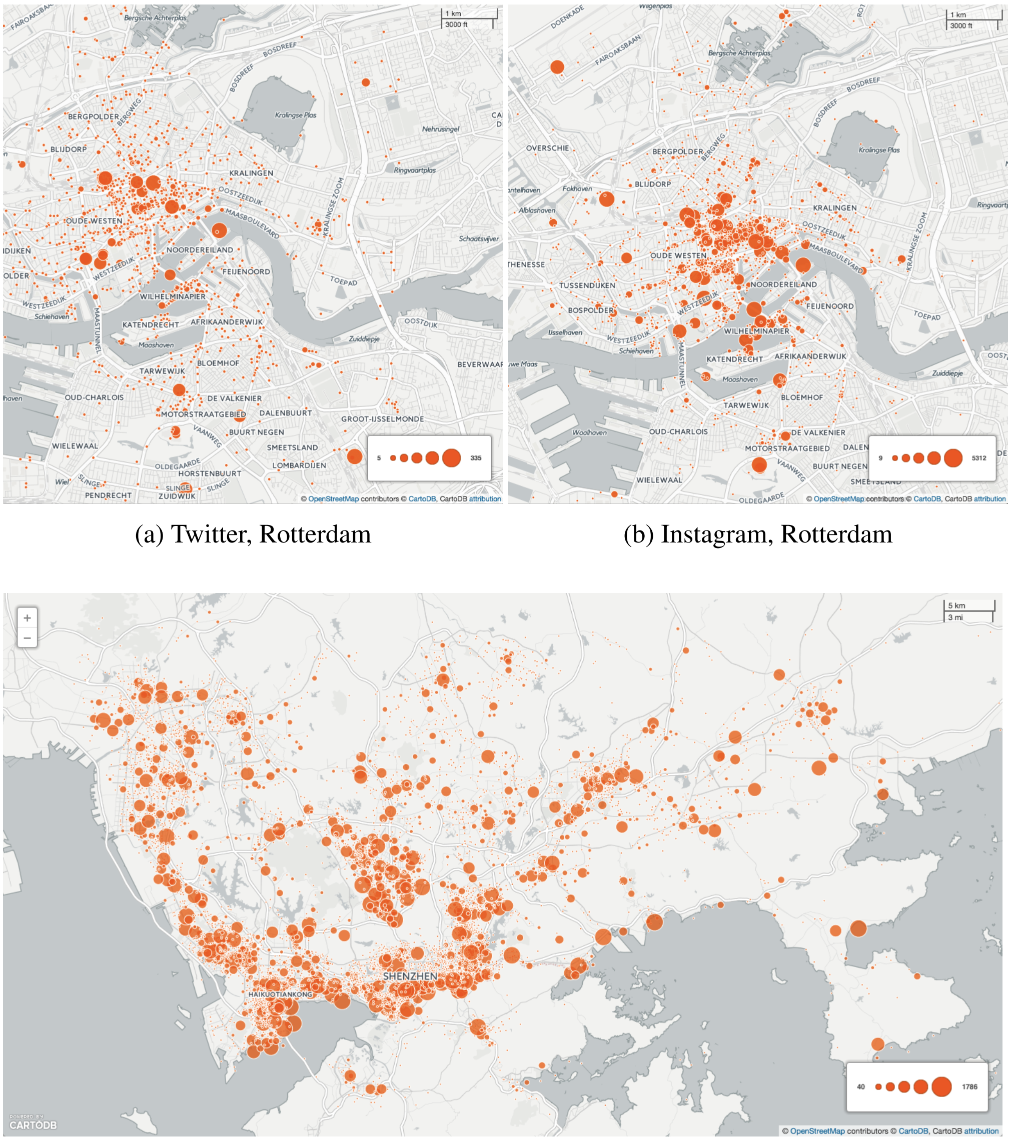
Figure 5.28: Distribution of User Activity in PoI Category, College & University



(a) Twitter, Rotterdam

(b) Weibo, Shenzhen

Figure 5.29: PoI Category Spatial Distribution in city of Rotterdam and Shenzhen



(a) Twitter, Rotterdam

(b) Instagram, Rotterdam

(c) Weibo, Shenzhen

Figure 5.30: PoI Popularity Spatial Distribution in city of Rotterdam and Shenzhen

Chapter 6

Discussion

The analysis of the dataset in the previous chapter shows several interesting findings. In this chapter, we discuss these findings and the Threats to Validity of this study.

6.1 Discussion

Concerning the choice of different social media platform in user activity study in two cities, the findings show that the number of Weibo users in Shenzhen is about nine times more than the number of Twitter and Instagram users in Rotterdam. The number of posts sent on Weibo in Shenzhen is more than ten times than Twitter and Instagram in Rotterdam. In Shenzhen, Weibo users move about three times longer than Rotterdam's Twitter and Instagram users. One possible reason for this may be based on the premise that scale of city impacts on the usage of social media.

In Rotterdam, there are more Instagram users than Twitter users, particularly, identified younger users and non-residents (commuters and foreign tourists) on Instagram are far more than those on Twitter. It is an implication that the Instagram, an image-based social media in nature, attracts more users, especially younger users.

With regard to spatial distribution, Weibo posts are distributed in all area, particularly in the southern west of Shenzhen. While Twitter and Instagram posts are around the Rotterdam city centre, with more in the North and less in the south. Weibo posts in Shenzhen are distributed more balanced than Twitter and Instagram. With regard to temporal perspective, Weibo posts in Shenzhen are more balanced than Twitter and Instagram in Rotterdam. Weibo users in Shenzhen have clearly different active mode compared with Twitter and Instagram users in Rotterdam, i.e. Weibo is less active than Twitter and Instagram until the 19 o'clock, reach the peak at midnight, while Twitter and Instagram are more active in the daytime and tend to be quiet in the night. This finding is in accordance with a report[1] saying that more than one-third of white-collar workers in China work overtime more than five hours per week in 2015, particularly, workers from industries of IT, Electronic and Internet work overtime more than 9.3 hours, which also indicates that employees in Shenzhen are bearing high pressure of competition.

The word use on different social media platform is also different. Instagram has more English words; however, Twitter has more Dutch words. Top words on Twitter

[1] http://article.zhaopin.com/pub/view/217834-26071.html

are more neutral and related to specific topics or places while on Instagram are more emotional and concern of daily life. Top words on Instagram and Weibo are temporal sensitive, i.e. evenly distinct in the morning, noon and night. However, this is not shown on Twitter. Moreover, PoI category preferences on different social media platforms are clearly different, which indicates that Instagram users tend to send posts when they are having meals or visiting shops or museums while Twitter and Weibo users tend to send posts in the office, or at home, or on the way. An explanation of this finding may be based on the premise that more tourists are using Instagram, and the image-based social media is more popular for reflecting people's daily life, which is widely utilised in a colourful context, e.g. the Shop, Food, and the Arts.

It also shows interesting findings considering user individual roles. There are more residents than commuters in Rotterdam while more commuters in Shenzhen. Accordingly, there are more posts sent by residents in Rotterdam, and more posts sent by commuters in Shenzhen. One possible reason for this might be that Shenzhen has become the largest migrant city in China[2], which provides huge job opportunities and attracts a large number of workers who, however, living nearby Shenzhen for cheaper costs.

Commuters in Rotterdam move longer than residents while in Shenzhen it is not distinct. An explanation of this finding may be based on the premise that due to the big scale of the city, residents have to move as long as the commuters travel to reach their office.

In both cities, commuters' active period in a day is slightly more unbalanced than residents. This natural based on the premise that commuters post during their commute. Particularly in Rotterdam, posts sent by commuters are more unbalanced graphically, around the central station and railways. While in Shenzhen, posts sent by residents are more unbalanced than commuters graphically. An explanation of this finding may be based on the premise that residents in city of Shenzhen are tend to live in high-density apartments due to expensive living cost, and the commuters come from different directions of the city through various transportation means, therefore, are more widespread.

Rotterdam residents tend to talk about names of local places (e.g. Maasstadweg, Wytemaweg, Hilledijk.Centraal, Molewaterplein, Hofplein) on Twitter while commuters tend to talk about names of places which are not in the city of Rotterdam (e.g. Breda, Amsterdam). However, in Shenzhen, top words use by residents and commuters are almost the same on Weibo. PoI preference of resident on Twitter in Rotterdam is distinct while on Instagram and Weibo is almost the same. One possible reason for this might be that the life gap between commuters and residents in Shenzhen has been decreased.

Interesting findings are shown concerning the user age. There are more young users in Rotterdam while more teen users in Shenzhen. The result indicates that The social media usage preference calls for a separation of user age groups in relevant researches.

Teen and old users in both cities send posts in a more unbalanced period in a day, which indicates that the life schedule of teen and old users is more regular than young

[2] http://www.goodarticle.info/Article/Business/Manufacturing/201511/592854.html

and mid-age users. Besides, in Rotterdam, users in different age groups from Twitter and Instagram have diverse active time slots.

Different age groups in both cities have distinct top words. Teen and young users on Twitter in Rotterdam and Weibo in Shenzhen tend to mention colourful and lovely things in a direct way, while mid-age and old on Twitter tend to mention big-scale things. Old users on Weibo tend to talk about the whole family, life, and societal problems while mid-age ones only focus on their family, work, and life.

Considering different gender of users in both cities, it shows several interesting findings. In Rotterdam, there are more male users than female ones, but in Shenzhen, there are more female than males. The same pattern is also shown regarding the number of posts in two cities. An explanation of Shenzhen has more female through the lens of social media may be based on the premise that Shenzhen, as a city with strong manufacturing in industry, has a huge number of female workers than male workers.

Male users on Twitter and Weibo moves longer than female; however, Instagram female users move longer than males, which may indicate that more travellers are using Instagram, and they accordingly visit attractions in the city.

The temporal analysis shows that in general the active time of female users is more unbalanced on all social networks. Particularly, female users keep active till midnight, while male users become quiet gradually after 21:30 o'clock. Male users on Twitter tend to mention things in specific fields while females are interested in general topics. A similar pattern is also shown on Weibo in Shenzhen. In regard to PoI preference, male and female users in Rotterdam have more diverse preferences than their counterparts in Shenzhen.

Moreover, with regards to PoI categories, there are more PoIs from the category of Food in the city of Rotterdam, while more PoIs of Professional & Other Places are shown in the city of Shenzhen. In the meantime, the popularity of all PoI categories on Weibo in Shenzhen have less fluctuation in temporal distribution than on Twitter and Instagram in Rotterdam. In Shenzhen, popular PoIs are distributed more in the southwest, and less in the northeast. This is in accordance with the fact that the two administrative districts in the southwest area, the Nanshan[3] and Futian[4], has the most GDP per capita[5].

The study also shows clear results concerning the extent of the differences in the ratio of Posts per User and Posts per PoI on different individual city roles, age groups, and genders in two cities. Twitter shows the clear distinction of the ratio of Posts per User in Rotterdam across individual city roles and age groups. While Instagram shows the enormous distinction of the proportion of Posts per PoI in Rotterdam across individual city roles. In the meantime, Weibo shows distinct differences w.r.t. to the ratio of Posts per PoI across age groups and genders, indicating that users in different age and gender in Shenzhen may have huge different PoI preferences.

All of above findings indicate that, in the context of user activity study, Twitter and Weibo shows distinct differences for resident w.r.t. local, general, and neutral topics, particularly for young and mid-age users on Twitter, and for teen and young users on Weibo. While, Instagram shows evident differences for non-residents w.r.t.

[3] https://en.wikipedia.org/wiki/Nanshan_District,_Shenzhen

[4] https://en.wikipedia.org/wiki/Futian_District

[5] http://www.phbang.cn/finance/data/152423.html

international, specific, daily and emotional topics, especially for young and mid-age users. In the meantime, various PoIs provide significant differences in user activity.

6.2 Threats to Validity

In this study, we investigate user activities from aspects of choice of social media platforms, individual user roles and demographics, and various of PoIs. The dataset is created with data from four social media platforms in two cities during two weeks. Inevitably there are some limitations in this whole research which may affect the results.

The main limitation is the scale of the dataset, including the type and number of cities, social media platforms, and the data collecting period. The candidate cities are chosen carefully with criteria from certain perspectives, which lead us to believe that extending the type and number of cities will result in increased differences leading to different observations. It also calls for a more general study on this topic. In the meantime, due to the limitation that Twitter, Instagram, and Foursquare are not available in China, we have to choose Weibo as the social media platform for Shenzhen. Thus, the data from different cities are collected through various social media platforms, which may introduce additional variants. The collecting period is another limitation that should be considered. In this research, we only monitored user activities in two cities in two weeks, which can be extended to a longer period to avoid data bias.

Another limitation is the completeness of the data from social media. Social media platforms used in this research all have collecting restrictions concerning crawling geo-referenced data, i.e. an only certain percentage of geo-referenced data is available for crawling through API. Therefore, the incompleteness of data affects the precision of the findings we observed from the dataset.

Last but not the least, an important limitation is introduced by the algorithms for enrich or inferring user information, such as home location determination and demographics determination. The algorithm determines users home by grid search based on the places where users sent the posts. In case, that a user never sends posts at or near home, and always sent posts around the office, the algorithm may lead to an incorrect result. Similarly, the algorithm determines user's gender and age through profile pictures provided by the user on the social media. Either user does not provide a real profile picture or the profile picture provided can not be identified properly, will lead to incorrect results. All these risks introduce limitations to our findings.

Chapter 7

Conclusions and Future Work

This chapter gives an overview of the project's contributions, and the proposals for future work.

7.1 Conclusions

In this work, we research on the user activities in the urban context through social media, to understand the impacts from the choice of social media platforms, the individual user roles and demographic, and the various of Point of Interests.

To this end, an experiment is designed to make a comparison of user activities on four social media platforms (i.e. Twitter, Instagram, Foursquare, Weibo) in two cities (i.e. Rotterdam and Shenzhen) for two weeks. The experiment is executed using the Social-Glass framework [1] to collect and enrich dataset from Twitter, Instagram, and Foursquare. Moreover, a crawler dedicated for collecting and enriching Weibo data is created, and integrated into Social Glass as a Weibo Crawler component.

The created novel dataset is carefully analysed with a set of metrics. Though there are some limitations, the analysis of the dataset provides interesting results. In general, we found that, in the context of user activity study, Twitter and Weibo shows distinct differences for resident w.r.t. local, general, and neutral topics particularly for young and mid-age users on Twitter, and for teen and young users on Weibo. While, Instagram shows evident differences for non-residents w.r.t. international, specific, daily and emotional topics, especially for young and mid-age users. In the meantime, various PoIs provide significant differences in user activity.

7.2 Future Work

Our work as an explanatory study provides several opportunities to continue the research in the future.

In this study, we designed an experiment to explore the user activities from various aspects on social media in two cities for two weeks period, using the Social Glass system and Weibo data collector to collect and enrich the data, and yield a novel dataset for analysis.

[1] http://www.social-glass.org

In the future, we aim to extend the data collecting time to a longer period, as well as engage more cities for a generalise research.

Moreover, we plan to explore the possibility of standardising the research aspects - i.e. social media platforms, individual city roles and demographics, various of POIs - as a package for performing user activity study on social media in urban context.

Besides, we plan to investigate the impacts of the package mentioned above in user activity study from the perspective of the data quality, i.e. how and to what extent can the distinctions of the data shown accordance with those aspects be used for user activity study.

Bibliography

[1] Peter M Allen. *Cities and regions as self-organizing systems: models of complexity*. Routledge, 2012.

[2] Michael Batty. *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. The MIT press, 2007.

[3] Michael Batty. Smart cities, big data. *Environment and Planning-Part B*, 39(2):191, 2012.

[4] Luis Bettencourt and Geoffrey West. A unified theory of urban living. *Nature*, 467(7318):912–913, 2010.

[5] Ricardo Buettner. Getting a job via career-oriented social networking sites: The weakness of ties. In *49th Annual Hawaii International Conference on System Sciences. Kauai, Hawaii: IEEE. do i*, volume 10.

[6] Yu Cai, Dan Dudek, Xiaoming Ma, Jianyu Zhang, Josh Margolis, Bella Liu, and Junping Ji. Research and practice on incorporating mobile sources into shenzhen carbon emissions trading system. *Journal of Renewable and Sustainable Energy*, 7(4):041502, 2015.

[7] Guofeng Cao, Shaowen Wang, Myunghwa Hwang, Anand Padmanabhan, Zhenhua Zhang, and Kiumars Soltani. A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems*, 51:70–82, 2015.

[8] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. Exploring millions of footprints in location sharing services. *ICWSM*, 2011:81–88, 2011.

[9] Roberta Comunian. Rethinking the creative city: the role of complexity, networks and interactions in the urban creative economy. *Urban Studies*, 2010.

[10] Arie Croitoru, N Wayant, A Crooks, J Radzikowski, and A Stefanidis. Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems*, 2014.

[11] Clodoveu A Davis Jr, Gisele L Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L Arcanjo. Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[12] John D Fairfield. The scientific management of urban space. professional city planning and the legacy of progressive reform. *Journal of Urban History*, 20(2):179–204, 1994.

[13] Rob Feick and Colin Robertson. A multi-scale approach to exploring urban places in geotagged photographs. *Computers, Environment and Urban Systems*, 2014.

[14] Isaac Stone Fish. A new shenzhen, 2010.

[15] George B Ford. The city scientific. *Engineering Record*, 67(May):551–552, 1913.

[16] Philippe Fournier-Viger, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Cheng-Wei Wu, and Vincent S Tseng. Spmf: a java open-source pattern mining library. *The Journal of Machine Learning Research*, 15(1):3389–3393, 2014.

[17] Qi Gao, Fabian Abel, Geert-Jan Houben, and Yong Yu. A comparative study of users' microblogging behavior on sina weibo and twitter. In *User modeling, adaptation, and personalization*, pages 88–101. Springer, 2012.

[18] Song Gao, Linna Li, Wenwen Li, Krzysztof Janowicz, and Yue Zhang. Constructing gazetteers from volunteered big geo-data based on hadoop. *Computers, Environment and Urban Systems*, 2014.

[19] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[20] Avner Greif and Guido Tabellini. Cultural and institutional bifurcation: China and europe compared. *The American economic review*, pages 135–140, 2010.

[21] Jeff Hemsley and Josef Eckert. Examining the role of" place" in twitter networks through the lens of contentious politics. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 1844–1853. IEEE, 2014.

[22] Edward Adamson Hoebel. *Anthropology: The study of man*. McGraw-Hill, 1966.

[23] Yingjie Hu, Song Gao, Krzysztof Janowicz, Bailang Yu, Wenwen Li, and Sathya Prasad. Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, 54:240–254, 2015.

[24] Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. Understanding us regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 2015.

[25] Shan Jiang, Ana Alves, Filipe Rodrigues, Joseph Ferreira, and Francisco C Pereira. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 2015.

[26] Carola Kunze and Robert Hecht. Semantic enrichment of building data with volunteered geographic information to improve mappings of dwelling units and population. *Computers, Environment and Urban Systems*, 2015.

[27] Richard LeGates, Nicholas J Tate, and Richard Kingston. Spatial thinking and scientific urban planning. *Environment and Planning B: Planning and Design*, 36(5):763–768, 2009.

[28] Jennifer S Light. *From warfare to welfare: Defense intellectuals and urban problems in Cold War America*. JHU Press, 2003.

[29] John Lingad, Sarvnaz Karimi, and Jie Yin. Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1017–1020. International World Wide Web Conferences Steering Committee, 2013.

[30] John J Macionis and Vincent N Parrillo. *Cities and urban life*. Pearson Education, 2004.

[31] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):47, 2014.

[32] William C McGrew. Culture in nonhuman primates? *Annual Review of Anthropology*, pages 301–328, 1998.

[33] Grant McKenzie and Krzysztof Janowicz. Where is also about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal semantic signatures. *Computers, Environment and Urban Systems*, 54:1–13, 2015.

[34] Grant McKenzie, Krzysztof Janowicz, Song Gao, and Li Gong. How where is when? on the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, 54:336–346, 2015.

[35] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of twitter users. *ICWSM*, 11:5th, 2011.

[36] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. 2013.

[37] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 144–153. IEEE, 2012.

[38] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2011.

[39] Tatiana Pontes, Gabriel Magno, Marisa Vasconcelos, Arpan Gupta, Jorge Almeida, Ponnurangam Kumaraguru, and Virgilio Almeida. Beware of what you share: Inferring home location in social networks. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 571–578. IEEE, 2012.

[40] David A Ralston, David H Holt, Robert H Terpstra, and Yu Kai-Cheng. The impact of national culture and economic ideology on managerial work values: A study of the united states, russia, japan, and china. *Journal of International Business Studies*, pages 177–207, 1997.

[41] Taylor Shelton, Ate Poorthuis, and Matthew Zook. Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, 2015.

[42] Seth E Spielman and Jean-Claude Thill. Social area analysis, data mining, and gis. *Computers, Environment and Urban Systems*, 32(2):110–122, 2008.

[43] Enrico Steiger, René Westerholt, Bernd Resch, and Alexander Zipf. Twitter as an indicator for whereabouts of people? correlating twitter with uk census data. *Computers, Environment and Urban Systems*, 54:255–265, 2015.

[44] Monica Stephens and Ate Poorthuis. Follow thy neighbor: Connecting the social and the spatial networks on twitter. *Computers, Environment and Urban Systems*, 2014.

[45] Alexander Szalai et al. The use of time: Daily activities of urban and suburban populations in twelve countries. *The use of time: daily activities of urban and suburban populations in twelve countries.*, 1972.

[46] Walter W Taylor. *A study of archeology*. American Anthropological Association Washington, DC, 1948.

[47] C Titos Bolivar. *City Usage Analysis using Social Media*. PhD thesis, TU Delft, Delft University of Technology, 2014.

[48] DINGQI YANG, DAQING ZHANG, and BINGQING QU. Participatory cultural mapping based on collective behavior data in location based social networks.

[49] Erjin Zhou, Zhimin Cao, and Qi Yin. Naive-deep face recognition: Touching the limit of lfw benchmark or not? *arXiv preprint arXiv:1501.04690*, 2015.

[50] Xiaolu Zhou, Chen Xu, and Brandon Kimmons. Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Computers, Environment and Urban Systems*, 54:144–153, 2015.

[51] Sharon Zukin, Scott Lash, and Jonathan Friedman. *Postmodern urban landscapes: mapping culture and power*. 1992.

Appendix A

The Basic Data Model

In the appendix we give an overview of the basic data model and its attributes built for this study. The extended data model is designed based on the basic data model to fulfil future requirements.

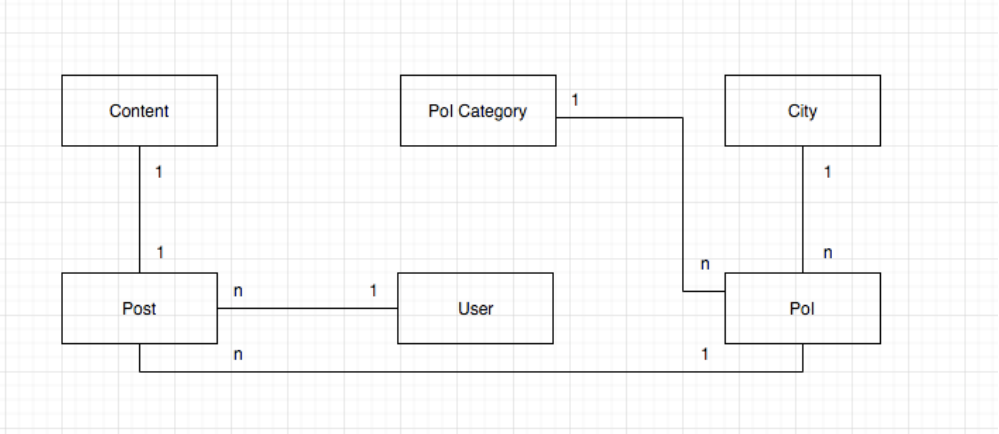


Figure A.1: The Overview of E-R model for Weibo Data Crawling

Table A.1: User

Property name	Description
user_id	the uid in Sina Weibo
name	the login name
profile_pic	the avatar profile picture for face identification
city_role	local resident or traveler
gyration	
gender	
age	
np_places	number of places visited by this user in the scope

Table A.2: Post

Property name	Description
post_id	the id of the post
user_id	the person who sent this post
content	text content of the post
source	how the post is sent: by app or web, by which mobile
lat	latitude
lon	longitude
poiid	the id of poi where this post is sent
timestamp	the time point of sending this post

Table A.3: PoI

Property name	Description
poiid	the poiid in Sina Weibo geo system
title	name of the poi (Chinese)
address	address in normal format
lat	latitude
lon	longitude
category	the category of this pos
postcode	the postcode of this poi
city	city number
country	country number
checkin_num	checkin times by people

Table A.4: PoI Category

Property name	Description
category_id	th id of this poi category
parent_id	the parent of this category
name	category name in English
name_cn	category name in Chinese

Table A.5: City

Property name	Description
city_id	the id of the city
city_name	the English name of the city

Table A.6: Content

Property name	Description
post_id	the id of post that the content belongs to.
user_id	the user who sent this post
content	the text content (in Chinses)
json	the json result from Language Technology Platform Cloud