SEMANTIC, SPATIAL AND TEMPORAL MODELLING OF GEOTAGGED SOCIAL MEDIA DATA FOR DESIRABLE REGION AND EVENT DETECTION

LIU ZHEWEI

Ph.D

The Hong Kong Polytechnic University

2021

This page is intentionally left blank

The Hong Kong Polytechnic University

Department of Land Surveying and Geo-Informatics

Semantic, spatial and temporal modelling of geotagged social media data for desirable region and event detection

LIU Zhewei

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

December 2020

This page is intentionally left blank

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

LIU Zhewei (Name of student)

This page is intentionally left blank

Abstract

The rise of Geotagged Social Media Data (GSMD) has provided new data sources and tools to investigate traditional research issues. Semantic, spatial and temporal information can be attached to GSMD, enabling human mobility patterns and urban structure to be revealed by GSMD. However, previous methods/models show limited effectiveness in analytics of GSMD, due to the complexity of GSMD's characteristics.

Given above, the research objective of this thesis is to develop novel effective methods/models to handle GSMD, from three progressive perspectives: (1) semantic modelling, (2) spatial semantic modelling and (3) spatiotemporal semantic modelling. From each perspective, new models and data-handling methods are developed for tackling specific research problems and the performances are evaluated accordingly.

For semantic modelling, a new hashtag network model is developed for topic modelling, and shows good performance on short social media texts. Statistical methods are traditionally used for topic modelling and geographical topic discovery. Nevertheless, statistical methods commonly require prior knowledge of the number of topics and large amounts of well-organized documents for training, which are inconsistent with the social media environment where prior knowledge is always lacking and short noisy texts predominate. Consequently, a new data-driving topic modelling method is proposed, where the hashtags attached to GSMD is used to construct network model and divided into semantic communities.

For spatial semantic modelling, a new scale-concerned model and a new data-driven model are proposed respectively for predicting regional desirability. The proposed scale-concerned model is an extension of traditional Hypertext Induced Topic Search (HITS) model, with consideration

of the size of the region, and predicts the regional desirability with better accuracy than previous methods. Further, a new data-driven model RegNet is proposed to predict regional desirability, using adaptive encoding-prediction structure of neural network.

For spatiotemporal modelling, a new model is developed for event detection by finding spatiotemporal irregularities. The intuition is that a social event may cause irregular geographical patterns, especially irregular human mobility and interaction patterns. The proposed model thus constructs both global and local features/indicators to characterize spatial patterns of GSMD. The social events are then detected by finding feature irregularities.

The experiments are conducted with real-world datasets and the results demonstrate the proposed models' effectiveness and outperformance over previous baseline methods.

In sum, this thesis serves as a systemic study on modelling of GSMD from several perspectives. Particularly, it focuses on the development of new models and data handling methods by combination of semantic, spatial and temporal information attached to GSMD, for the task of topic modelling, desirable region detection and event detection. The presented works in this thesis can benefit relevant urban study by providing effective and robust data handling models/methods, and also be potentially implemented as data-processing tools for tackling practical real-world problems.

List of Publications

Articles:

1 **Liu Z**, Shi W*, Zhang A, et al. Analysis of the performance and robustness of methods to detect base locations of individuals with geo-tagged social media data. International Journal of Geographical Information Science (accepted and in press)

2 Shi W, **Liu Z***, An Z, et al. RegNet: a neural network model for predicting regional desirability with VGI data. International Journal of Geographical Information Science, 2021, 35(1): 175-192 (**corresponding author**)

3 **Liu Z**, Zhou X, Shi W*, et al. Recommending attractive thematic regions by semantic community detection with multi-sourced VGI data [J]. International Journal of Geographical Information Science, 2019, 33(8): 1520-1544

4 **Liu Z**, Zhou X, Shi W*, et al. Towards Detecting Social Events by Mining Geographical Patterns with VGI Data [J]. ISPRS International Journal of Geo-Information, 2018, 7(12): 481

5 Yao Y, Shi W, Zhang A, **Liu Z**, Luo S. Examining the diffusion of coronavirus disease 2019 cases in a metropolis: a space syntax approach. International Journal of Health Geographics (accepted and in press)

6 Huang X, Lu J, Gao S, Wang S, **Liu Z**, Wei H. Staying at home is a privilege: evidence from fine-grained mobile phone location data in the U.S. during the COVID-19 pandemic. Annals of the American Association of Geographers (accepted and in press)

7 Wang A, Zhang A, Chan EH, Shi W, Zhou X, **Liu Z**. A Review of Human Mobility Research Based on Big Data and Its Implication for Smart City Development [J]. ISPRS International Journal of Geo-Information. 2021, 10(1):13

8 Chen P, Shi W*, Zhou X, **Liu Z**, et al. STLP-GSM: a method to predict future locations of individuals based on geotagged social media data[J]. International Journal of Geographical Information Science, 2019, 33(12): 2337-2362

Patents:

1. Regional desirability predicting method and device, based on HITS model (under review, Chinese patent number: 201910436416.0)
2. Regional desirability predicting method and device, based on artificial neural network (under review, Chinese patent number: 201910715730.2)

Acknowledgement

Upon completion of this thesis, memories flash back. Old things and people come flooding my mind, as if it was yesterday once more.

I would like to give my sincerest gratitude to my dear parents, who always stand by me whenever I need. Their unconditional love is my biggest comfort and strength. They are giving, caring but never requiring anything back. I wish I can be a person like them and passing down their love to others.

I am grateful for my supervisors Prof.Wenzhong Shi and Prof.Geoffery Q.P. Shen, for their insightful academic advice. During my study in HKPU, they are the ones who teach me what is research and how to think critically. Their insistence on excellence also reminds me that there is no shortcut to success, except hard work and persistence.

I would like to express my gratitude to all the colleagues in my team, both in HKPU and WHU. They are more like brothers and sisters to me, after all those 'fights' we have fought together. Big dream can only by realized by big team. Being a sole Robin Hood is never a promising path. Any progress I obtained could not happen without them.

Also, I would like to thank all my friends for their company. Life is not only about research but more about love and commitment. I am especially grateful for the company of Daniel, Harry, Tiffany, Nelson, Lily, Micah. I also would like to offer my sincere appreciation for my friends in Taibai Reading Club. Bonds with you guys let me know how friendships mean so much.

I would like to thank the blessing of fate, for my advantages that others crave for yet don't have.

Finally, I would like to thank myself. Thank you for not giving up. All the struggle in the days and sleepless nights, finally paid off. You deserve it.

Table of Contents

This page is intentionally left blank

List of Figures

This page is intentionally left blank

List of Tables

This page is intentionally left blank

Chapter 1 Introduction

1.1 Background

Geotagged Social Media Data (GSMD) refers to the social media posts with attached geographic information indicating the post location. Currently, the geotagged social media posts have been adopted by majority of the mainstream platforms, including Twitter, Instagram, Flickr, Weibo, TikTok and so on.

Geotagged social media data reveals human footprints as well as activities at specific locations, and provides researchers with new tools to study traditional research problems. The use of geotagged social media data has enabled researchers to collect granular data in a more cost-efficient way and been widely applied into various research fields, such as tourism (Chua et al. 2016, García-Palomares, Gutiérrez and Mínguez 2015, Lee and Tsou 2018, Sinclair, Ghermandi and Sheela 2018), advertising and recommendation (Lai, Cheng and Lansley 2017, Bao et al. 2015, Cai et al. 2018, Ding and Chen 2018, Gao et al. 2013) , disaster monitoring (Haworth and Bruce 2015, De Albuquerque et al. 2015, Zook et al. 2010, Peary, Shaw and Takeuchi 2012) , human mobility (Li and Yang 2017, Liu, Wang and Ye 2018, Chen et al. 2019), and urban study (Huang and Wong 2016, Jia et al. 2019, Paldino et al. 2016, Zhang et al. 2020, Longley and Adnan 2016).

Various kinds of information can be attached to geotagged social media data. The kinds of information that are commonly attached mainly include, semantic, spatial and temporal information, as shown in Figure 1.1. The attached semantic information is mostly in the form of texts, indicating the activities or user's opinions associated with the social media posts. The attached spatial information is to indicate the post location, which is mainly in the form of point

of interest (POI), i.e., a specific venue or region, rather than plain coordinates with latitude and longitude. The attached temporal information is to indicate the timestamps when the posts are made. Apart from the aforementioned information, some other kinds of information may also be attached, such as audio, picture and video.

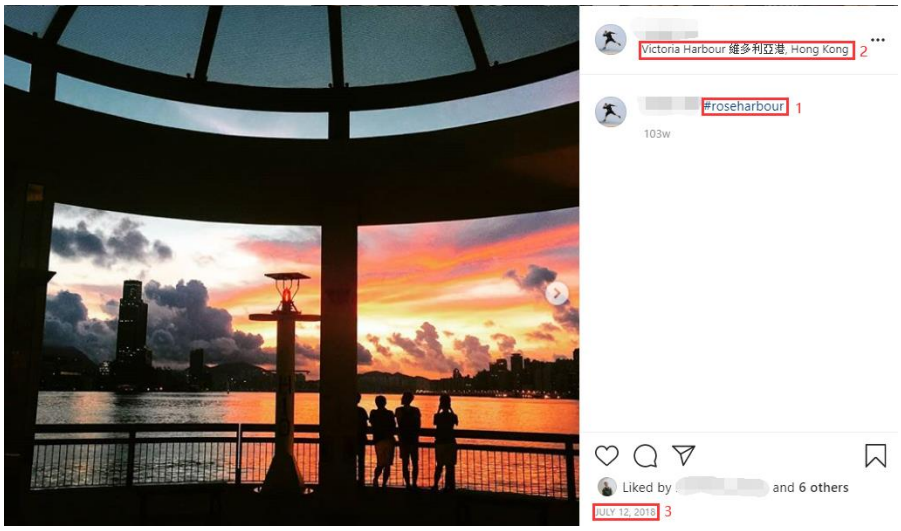


Figure 1.1 Geotagged social media data, with attached 1) semantic, 2) spatial and 3) temporal information. The figure is generated by the thesis author.

As a newly emerging data sources, research with geotagged social media data suffers from certain challenges:

- The first challenge is the issue of data availability. Currently, most geotagged social media data is collected from open APIs provided by the companies. This data collection procedure is highly uncertain and subject to the company strategies. Due to the privacy issue and commercial interest, the number of free open APIs has been diminishing, aggravating the difficulties for data collection.
- The second challenge is the issue of data bias. The users of social media are only a skewed sample of the whole population and mainly consists of younger generation.

Besides, for specific users, their daily activities and mobility patterns can not be fully represented by the social footprints. The users may post on the locations they never visited or on the contrary, visit locations without posting online, which undermines the effectiveness to reconstruct users' mobility with online footprints. How to quantify the data bias and improve the reliability of relevant research are still not fully addressed.

- The third challenge is the issue of effective data handling. Due to the sudden rise of various social media, effective methods to analyze geotagged social media data are still lacking.

This thesis focuses on addressing the aforementioned third challenge (i.e., the issues of effective data handling) by combining (1) **semantic information**, (2) **spatial information**, (3) **temporal information**, of GSMD. Specifically, this thesis aims to develop new data handling models/methods from the following three progressive perspectives: (1) **semantic modelling**, (2) **spatial semantic modelling** and (3) **spatiotemporal semantic modelling**. The previous works on GSMD analytics are given below.

1.2 Previous works on GSMD analytics

The previous works on (1) semantic modelling, (2) spatial semantic modelling and (3) spatiotemporal semantic modelling, of GSMD, are reviewed and described as below.

1.2.1 Semantic modelling: geographical topic discovery

The semantic information attached to GSMD and other kinds of spatial big data can be used to discover topics in a geographical context, which is the aim of research field geographical topic discovery (Yin et al. 2011).

Among the geographical topic discovery methods, statistical models such as Latent Dirichlet Allocation (Blei, Ng and Jordan 2003) are commonly used for geographic topic modelling. For example, Lansley and Longley (2016) explored the use of LDA method to classify geotagged Tweets into a small number of groups recorded during typical weekdays throughout 2013 in Inner London. The classification identified 20 topic groups. The classification results were further joined with analysis from the perspectives of temporal dimensions, spatial distributions, and user individual behaviors to gain insights into the content and coverage of Twitter usage across Inner London. By proposing a geographical hierarchical self-organizing map (Geo-H-SOM), Steiger, Resch and Zipf (2015) analyzed the geospatial, temporal and semantic similarities of georeferenced tweets. Each tweet's topic association and topic-word association which were found by LDA method were the semantic input components to Geo-H-SOM, a type of artificial neural network (ANN) which abstracted information from multi-dimensional primary signals and represented data properties in a two-dimensional topological connected output space. The geospatial and temporal components were also input into the neural network in order to explore clusters of high-dimensional geospatial and semantic information in output space. Zhang, Sun and Zhuge (2013) combined LDA and DBSCAN in different ways to generate three different methods for geographical topic modelling and evaluate the respective performance in terms of topic discovery and time efficiency.

Nevertheless, a disadvantage of LDA method is that it has limited effectiveness in handling social media data. This is because that LDA was originally proposed to detect topics for long and well-structured documents, while in social media context, short noisy texts are predominant and *a priori* knowledge such as a predefined count of topics is unavailable (Prateek and Vasudeva 2016).

As for the types of the model input, the spatial and textual information of spatial big data are traditionally used as input data. However, some other types of information have also been used, such as images (Rykov, Nagornyy and Koltsova 2016) and heterogeneous unstructured articles (Adams and Janowicz 2012), to facilitate the process of topic discovery. Rykov et al. (2016) used a combination of semantic clustering and image recognition to study the geospatial pattern of geotagged Instagram photos in Saint-Petersburg. The study processed the images with Google Cloud Vision API service to assign artificial tags of the recognized entities to the photo and then constructed a semantic network where each vertex represented a Google-defined tag and each arc represented a measure of similarity based on normalized co-occurrence of a particular pair of tags assigned to the same image. The semantic network was clustered into different groups to generate topics. Apart from using volunteered geographic information, Adams and Janowicz (2012) proposed a topic modelling method to extract topics from heterogeneous and unstructured data from Wikipedia articles and travel blog entries. The paper geo-referenced a document by matching the document to the set of coordinates associated with the named place found in the document and then used kernel density estimation to identify contiguous regions that were thematically differentiable from other regions.

1.2.2 Spatial semantic modelling: desirable thematic location recommendation

After identifying the semantic topics of GSMD, how these topics distribute spatially and where are the desirable locations of these topics are interesting subsequent questions. A research topic related to these questions is desirable thematic location recommendation.

The rise of GSMD and location-based social networks has given rise to novel recommender systems that seek to recommend desirable spatial-related locations (POIs or regions) to users. Technically, the traditional recommender systems mainly used two types of approaches:

collaborative filtering and content-based filtering (Jafarkarimi, Sim and Saadatdoost 2012). In Bao et al. (2015), the techniques for LBSN recommender systems were further classified into three main types: 1) content-based methods, which used information from users' profile and features of locations for recommendations (Park, Hong and Cho 2007, Ramaswamy et al. 2009); 2) collaborative filtering methods, which inferred users' preferences from historical behaviors (Horozov, Narasimhan and Vasudevan 2006, Ye, Yin and Lee 2010, Lemire and Maclachlan 2005); 3) link analysis-based methods, where link models were used to detect informative users and desirable places (Raymond, Sugiura and Tsubouchi 2011, Zheng et al. 2009).

In terms of recommended objectives, there are mainly two types of stand-alone location recommender systems: POI and region recommender systems. The types of items recommended by POI recommender systems are mainly individual venues that match user interests or querying requirements. Various kinds of spatiotemporal and contextual features have been incorporated in the POI recommender models (Cai et al. 2018). Particularly, the spatial effects are widely considered, which differentiates POI recommenders from other kinds of recommender systems. For example, as users are likely to visit venues nearby, such spatial effects were modelled as exponential relationships, probability distribution, or power law relationships (Yang, Cheng and Dia 2008, Ye et al. 2011, Kurashima et al. 2013, Liu and Seah 2015). Some works also modelled the periodicity of check-ins (Yuan et al. 2013, Gao et al. 2013), social relationships of users (Cheng et al. 2012, Gao, Tang and Liu 2012) and POI tips (Yang et al. 2013). To address the variety and complexity of the features used in POI recommendation, machine learning and statistical techniques are both used for analysis (Li et al. 2016). However, there are still some problems that may undermine the effectiveness of machine learning approaches in POI recommendation, such as data bias and computational complexity (Wan et al. 2018). A recent

new trend is to introduce deep neural network (DNN) into recommender systems. The DNNs are capable of learning high-order features and the unknown interactive relationships for a specific task (LeCun, Bengio and Hinton 2015) and have been proven effective in recommending tasks. Cheng et al. (2016) jointly trained deep neural networks and wide linear models to use the advantages of both memorization and generalization, and evaluated their methodology on Google Play with good feedbacks. He et al. (2017) presented a neural network-based collaborative filtering framework for matrix factorization and recommendation. Regarding POI recommendation, Ding and Chen (2018) developed a DNN recommender system that incorporated the joint influences of co-visiting, geographical proximity and categorical correlation.

The application of POI recommender systems can be limited when users want to find spatial areas with many venue options, which is compensated by region recommender system. In Kurashima et al. (2010), the authors extracted landmarks from geotagged photos and estimated the probability that a user visited certain regions. A similar work was done by Sun et al. (2015) where urban landmarks were identified and travel routes were recommended accordingly. In terms of calculating regional desirability, previous works mainly used presumed empirical models, such as power law distribution (Sun et al. 2015) and Hypertext Induced Topic Search (Zheng et al. 2009, Zheng and Xie 2011). However, there are several challenges inflicting current works regarding region recommendation. Firstly, among existing methods, few efforts have been made to integrate the related parameters of region spatial scale and users' redundant user visits into the model, which may undermine the effectiveness of current region recommending strategies. Secondly, as the relations between regional desirability and user online footprints were rarely revealed, there is still a lack of convincing models that can accurately

predict the interactive mechanism between user check-in and regional desirability.

1.2.3 Spatiotemporal semantic modelling: event detection

The semantic, spatial and temporal information of GSMD can be combined and jointly used for answering the question: what (semantic-related) is happening at some place (spatial-related) and some time (temporal-related)? This question is the focus of research on event detection.

Recently, the rise of various sensing techniques has greatly improved the ability of humans to detect natural and social events. However, a tricky issue regarding event detection with sensing technology is that the word ‘event’ has been widely used throughout literatures from various fields yet rarely clearly defined. Yu et al. (2020) introduced three definition of ‘event’ in a progressive way. The first definition of ‘event’ is ‘something that happens at a particular time and place’ (Allan et al. 1998). With this definition, an event can be represented from three dimensions: what, where and when. This definition is used in many fields, such as environment management and earth observation. The second definition is ‘a specific occurrence involving participants’, adding a new dimension ‘who’ into the previous event representation. The third definition is a ‘change of state’ in the monitored measurement (Kopetz 1991), and was applied in many signal processing approaches (Caudal and Nicolas 2005, Hühn 2009, Meira-Machado et al. 2009).

The sensing data sources for event detection can be categorized into four types (Yu et al. 2020): remote sensing, in-situ sensing, health sensing, and social sensing. For remote sensing, the data used for event detection mainly includes optical, thermal or radar image from remote sensing satellites. These remote sensed images provide information for various missions, such as agriculture monitoring, pollutants monitoring, and mineral inspection (Debba et al. 2005,

Manolakis et al. 2013, Adam, Mutanga and Rugege 2010). Different from remote sensing, in-situ sensing collects data from the instruments directly located in the scenes, such as various in-situ sensors and Internet of Things (IoT). These in-situ instruments enable environment monitoring with high temporal resolution or near real-time (Ajo-Franklin et al. 2019, Werner-Allen et al. 2006, Boubrima, Bechkit and Rivano 2017). For health sensing, the wearable sensors are deployed to monitor the conditions of human body. The application of health sensing includes detecting accidental fall for elder or disable people (Mubashir, Shao and Seed 2013), investigating mental health issues (Rodrigues et al. 2015), recognizing behavior during sports game (Kautz, Groh and Eskofier 2015), and collecting patient information for doctors in remote places (Gao et al. 2016, Varatharajan et al. 2018). Another data source used for event detection is social sensing, which is to utilize the crowd-sourced data or user-generated content (UGC) to monitor social event or group behavior of human beings.

Specifically, due to the rapid development of social sensing technology, the social media data or user-generated content has been a very popular data source for event detection. A challenging problem regarding event detection with social media data is to extract useful, structured representations of events from the disorganized corpus of noisy posts (Ritter, Etzioni and Clark 2012). Sudden increases in the frequency ("bursts") of sets of keywords have been popularly used for detecting new events in a data stream (Imran et al. 2018). Some other approaches include wavelet-based clustering of frequency signals, topic clustering with meta-data analysis and domain-specific approaches (Corley et al. 2013, Weng and Lee 2011). Schinas et al. (2018) grouped the event detection methods into three categories: (1) feature-pivot, i.e., detection of abnormal patterns in the appearance of features (Guille and Favre 2014, Weng and Lee 2011, Alvanaki et al. 2012, Zhang et al. 2015); (2) document-pivot, i.e., clustering documents with

similarity measures (Lee 2012, Petkos, Papadopoulos and Kompatsiaris 2012, Bao et al. 2013, Petkos et al. 2017); and (3) topic modelling, i.e., using statistical models to identify events (Hu et al. 2012, Diao and Jiang 2013, Wei et al. 2015, Zhou, Chen and He 2015). Becker, Naaman and Gravano (2011) highlighted four feature categories (temporal, social, topical, and Twitter-centric features) that can be used for modelling event features. These methods are mainly semantic-based, that is, they use text mining techniques to investigate the change in social media semantic information for event detection. Other studies have used spatial-related methods. These have addressed issues such as measuring regional irregularities with post count and user count (Lee and Sumiya 2010), using content-based methods to detect event location (Paule, Sun and Moshfeghi 2019, Sakaki, Okazaki and Matsuo 2010), estimating the influenced area of an event with kernel density estimation (Gao et al. 2018), investigating the geographic extent to find localized events (Abdelhaq, Sengstock and Gertz 2013), assessing the impact area of a natural disaster (Panteras et al. 2015), and detecting events (outliers) by finding the weighted centroids outside the spatial standard deviation ellipse (Wachowicz and Liu 2016). Compared with semantic-based methods, these spatial-related methods basically focused on identifying the spatial information of the detected event or using the change in the frequency of the posts/users in certain spatial regions to detect events.

1.3 Research scope and objectives

In this section, the research scopes of the thesis are given firstly. Then, the gaps of previous works are described. Finally, the research objectives of the thesis are specified.

1.3.1 Research scope

This thesis focuses on the issues of data handling by incorporating (1) semantic information, (2) spatial information, (3) temporal information, of GSMD. Specifically, this thesis aims to develop new data handling methods from the following three progressive perspectives: (1) semantic modelling, (2) spatial semantic modelling and (3) spatiotemporal semantic modelling. Specifically, the research scope to be investigated by this these is summarized as below:

- **Semantic modelling**: What is the **topic** of the GSMD? (**Topic Modelling**)
- **Spatial semantic modelling:** Where are the **regions of specific themes**, inferred from GSMD? Among all the regions of specific themes, which are the most **desirable** and **attractive** ones? Can we give a ranking of these regions, in terms of their desirability? (**Desirable Thematic Region Detection**)
- **Spatiotemporal semantic modelling**: **What** is happening at **some place** and at **some time,** inferred from GSMD? (**Event Detection**)

1.3.2 Previous gaps

There are some previous works focusing on analytics of GSMD from the aforementioned three perspectives: (1) semantic modelling, (2) spatial semantic modelling and (3) spatiotemporal semantic modelling. This work is never the first attempt investigating this scope. However, there are still some gaps remain unfilled by previous works, which cast doubts on the effectiveness of previous data handling methods, as discussed in the Section 1.2. The previous gaps this thesis aims to address are as below:

(1) Semantic modelling: how to effectively discover topics with short noisy social media content?

- One gap is how to effectively discover topics with short noisy social media content. Statistical methods, such as Latent Dirichlet Allocation (LDA), are traditionally used for geographic topic discovery, but, nevertheless, have limited effectiveness in handling social media data. A major reason is that statistical methods commonly require large amounts of well-organized documents as training data, which is inconsistent with the social media environment where short and noisy texts predominate; another reason is that some statistical methods require predefined parameters, such as counts of topics, which are quite arbitrary and unpredictable due to a lack of *a priori* knowledge.

(2) Spatial semantic modelling I: how to develop a model to calculate regional desirability with consideration of varying spatial scales of regions?

- Large regions may have more visits than small regions simply because large regions cover more venues. When calculating the regional desirability, the influence of spatial scales needs to be taken into consideration. This scale issue is specific to region, as venues are typically treated as points without consideration of sizes.

(3) Spatial semantic modelling II: how to develop an accurate model to calculate regional desirability when the interactions of relevant features are still unknown?

- To predict regional desirability, traditional methods mainly use empirical models, like Hypertext Induced Topic Search (HITS), which are based on intuitive hypothetical relationships and can be inaccurate in predicting desirable regions, as potential perceptual bias is introduced. This results from the fact that the hidden interactions between user check-ins and regional desirability haven't be clearly

explained or precisely modelled yet, so that the empirical models based on human empirical intuition and presumed relationship are used for rough approximation.

(4) Spatiotemporal semantic modelling: how to model features by investigating geographic patterns of GSMD and detect events by finding irregular features?

- Most research works on event detection with social media are using semantic-based methods and text mining. Platforms, such as Twitter, provide accessible streaming data, which contains real-time text information, making it possible to detect emerging social events by examining the change of semantic-based features. However, a new event detection methodology can be proposed by investigating the geographical patterns of GSMD. The intuition is that a social event will affect how the objects spatially distribute across certain regions and how they mutually interact, thus causing irregular geographical patterns, especially irregular human mobility and interaction patterns (e.g., sports games causing intense human aggregation or terrorism attacks causing sudden evacuation from certain regions). By introducing depictive measuring features with GSMD and identifying the feature irregularity, such geographical patterns, in turn, can be used to distinguish social events.

1.3.3 Research objectives

This thesis aims to address the aforementioned gaps by developing new data handling methods/models. Specifically, the research objectives of this thesis are given as below, each corresponding to one of the above gaps:

- **Objective 1-Semantic modelling:**

- To develop a new model for topic modelling, with good performance on the short social media texts, as presented in Chapter 3.

- **Objective 2-Spatial semantic modelling I:**
  - To develop a new scale-concerned model to predict desirability of thematic regions, as presented in Chapter 4.
- **Objective 3-Spatial semantic modelling II:**
  - To develop a new data-driven model to predict desirability of thematic regions, as presented in Chapter 5.
- **Objective 4-Spatiotemporal semantic modelling:**
  - To develop a new model for event detection, by finding spatiotemporal irregularities, as presented in Chapter 6.

1.4 Thesis outline

The outline of the thesis is given as Figure 1.2. Chapter 2 gives the key notations and problem definitions, which are used consistently throughout the thesis. Chapter 3 addresses the gap concerning semantic modelling by proposing a new model for topic modelling (Objective 1). Chapter 4 addresses one gap concerning spatial semantic modelling by proposing a new scale-concerned model for calculating regional desirability (Objective 2). Chapter 5 addresses another gap concerning spatial semantic modelling by proposing a new data-driven neural network model for calculating regional desirability (Objective 3). Chapter 6 addresses the gap concerning spatiotemporal semantic modelling by proposing a new model to detect events by finding spatiotemporal irregularities (Objective 4). Finally, Chapter 7 summarizes the whole thesis and gives the conclusion.

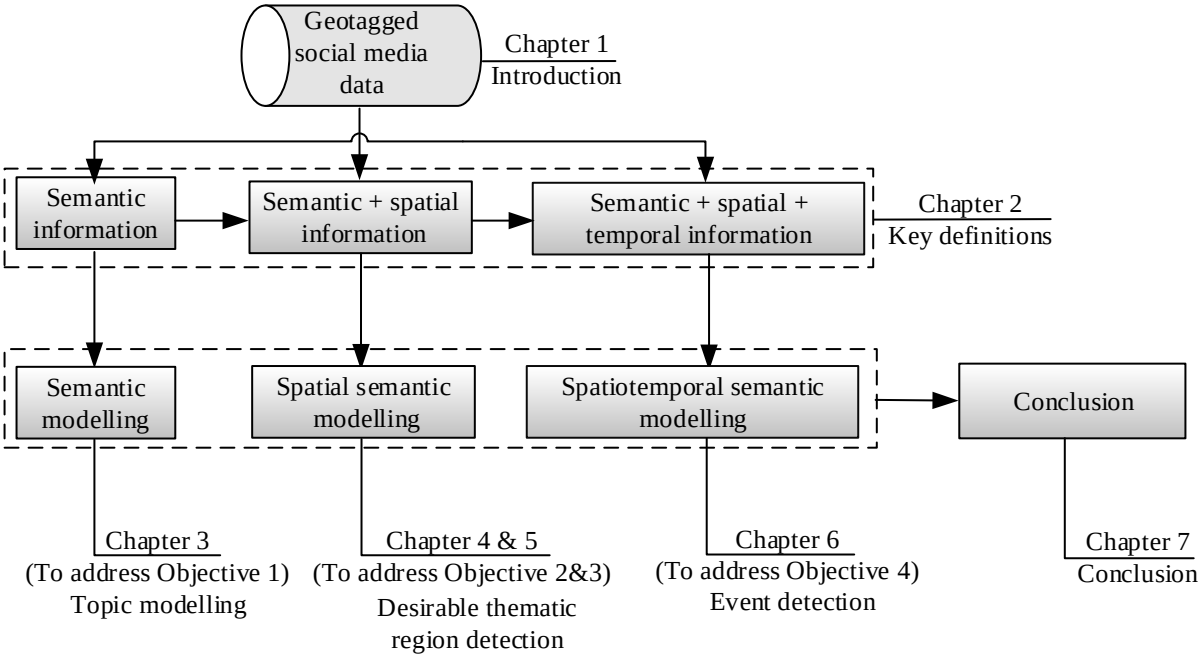


Figure 1.2 Outline of this thesis

Chapter 2 Key notations and problem definitions

The definitions related to the research questions are given in this chapter and used consistently throughout the thesis. For convenience, the key notations of the thesis are given firstly (shown in Table 2.1).

Table 2.1 Key Definitions and Notations

Variable	Description
tpc	A semantic topic. A topic set *TP* is a collection of topics, $\boldsymbol{TP} = \{tpc_1, tpc_2, \dots, tpc_n\}$.
tg	A textual hashtag. A hashtag set *HT* is a collection of hashtags, $\boldsymbol{HT} = \{tg_1, tg_2, \dots, tg_n\}$.
v	A POI with identifier, category, location and rating, $v = (vid, vcat, vloc, vrat)$.
u	A unique user in the datasets. A user set *U* is a collection of users, $\boldsymbol{U} = \{u_1, u_2, \dots, u_n\}$.
pst	A geotagged social media post, including user, visited venue, post timestamp and attached hashtags, $pst = (u, v, t, tgs)$. A post set *P* is a collection of the posts, $\boldsymbol{P} = \{pst_1, pst_2, \dots, pst_n\}$.
Rn	A Region with spatial range, set of POIs and desirability rating and theme, $Rn = (Ra, V, Rt, tpc)$. A region set R is a collection of the regions, $\boldsymbol{R} = \{Rn_1, Rn_2, \dots, Rn_n\}$

RV A multi-dimension vector consisting of user visiting to regions, $\boldsymbol{RV} = \{uv_1, uv_2, \dots, uv_n\}$.

evt An event with semantic, spatial and temporal identifier, $evt = (tpc, Rn, t)$. An event set E is a collection of the events, $\boldsymbol{E} = \{evt_1, evt_2, \dots, evt_n\}$.

Definition 1. Topic. A topic tpc is the semantic subject of social media posts and events. A topic set *TP* is a collection of topics, $\boldsymbol{TP} = \{tpc_1, tpc_2, \dots, tpc_n\}$.

Definition 2. Hashtag. A hashtag tg is a textual phrase preceded by the symbol # that indicates the potential topics of the accompanying text. A hashtag set *HT* is a collection of hashtags, $\boldsymbol{HT} = \{tg_1, tg_2, \dots, tg_n\}$.

Definition 3. POI. A POI is a venue (e.g., a hotel or a shopping place) with unique identifications. In this thesis, a POI is denoted as v_i with four attributes $v_i = (vid\ , vcat, vloc, vrat)$ where vid is a unique identifier, $vcat$ is the category, $vloc$ is the venue's geographical location, $vrat$ is the user's rating for venue v_i's desirability.

Definition 4. User. A unique user in the dataset. A user set *U* is a collection of users, $\boldsymbol{U} = \{u_1, u_2, \dots, u_n\}$.

Definition 5. Geotagged post. A geotagged social media post pst is the item published by users u on the social media platform with location information v, timestamp t and attached textual hashtag tgs, defined as $pst_i = (u, v, t, tgs)$. A post set *P* is a collection of the posts, $\boldsymbol{P} = \{pst_1, pst_2, \dots, pst_n\}$. The frequency of posts on a venue v_i can be used to indicate its popularity.

Definition 6. Region. A region is a spatial area with coverage of POIs. A region is denoted as Rn_i with four attributes $Rn_i = (Ra, V, Rt, tpc)$, where Ra is a polygon indicating the spatial

coverage of Rn_i, V is the set of POIs $V = (v_1^i, v_2^i, \dots, v_m^i)$ within the spatial range of Rn_i, i.e., $v_m^i.vloc \in Rn_i.Ra$, Rt is the numerical rating of the region Rn_i's desirability, tpc is the theme of the region. A region set R is a collection of the regions, $\boldsymbol{R} = \{Rn_1, Rn_2, \dots, Rn_n\}$.

Definition 7. Region-Visit Vector. Given a region Rn_i, a set of users *U* and check-ins *C*, a region-visit vector is denoted as $\boldsymbol{RV_i} = \{uv_1^i, uv_2^i, \dots, uv_n^i\}$, where *n*=|*U*| is the total count of users in user dataset *U*, $uv_k^i = |subC|$, where $\boldsymbol{subC} = \{e|\ e \in C\ AND\ e.v \in Rn_i.\mathrm{V}\ AND\ e.u = U.u_k\}$.

Definition 8. Event**.** An event evt is uniquely identified by three attributes $evt = (tpc, Rn, t)$, where tpc is the semantic content of the evt, Rn is the spatial range where evt takes place and t is the temporal range of evt. These three attributes jointly answer the following question: what is happening at some place and at some time. An event set E is a collection of the events, $\boldsymbol{E} = \{evt_1, evt_2, \dots, evt_n\}$.

The research scope (Section 1.3) of this thesis can therefore be respectively defined as below:

Problem 1 (semantic modelling): Given a hashtag set HT and topic set TP, the aim is to find a mapping relationship M from HT to TP, $M: HT \rightarrow TP$.

Problem 2 (spatial semantics modelling): Given a social media post set P, the aim is to detect a thematic region set R and recommend the top-k regions with the highest desirability Rt.

Problem 3 (spatiotemporal semantic modelling): Given a social media post set P and event set E, the aim is to find a mapping relationship F from P to E, $F: P \rightarrow E$.

Chapter 3 Semantic modelling: a hashtag network model for topic modelling

3.1 Introduction

In this chapter, the following question is investigated: what is the topic of the GSMD? As illustrated in Chapter 1, statistical methods, such as Latent Dirichlet Allocation (LDA), are traditionally used for geographic topic modelling, which, nevertheless, have limited effectiveness in handling social media data. A major reason is that statistical methods commonly require large amounts of well-organized documents as training data, which is inconsistent with the social media environment where short and noisy texts predominate (Prateek and Vasudeva 2016); another reason is that some statistical methods require predefined parameters, such as counts of topics, which are arbitrary and unpredictable due to a lack of *a priori* knowledge.

On the other hand, the consideration is that the semantic information, such as hashtags, attached to social media posts is sufficiently self-explanatory and the potential topics can be thus indicated. Accordingly, a new hashtag network model is developed in this chapter, to discover topics of the GSMD. The hashtags in the titles attached to the geotagged photos are extracted, based on which an undirected 'hashtag network' model is built where each hashtag is denoted as a network node and the co-occurrence frequency between hashtags is assigned as a weighting to the edge connecting the corresponding nodes in the 'hashtag network'. A greedy optimization method is then implemented to explore the communities from the hashtag network, and a common topic is assigned to the hashtags of the same community. The topics of geotagged photos are further detected by introducing the topics of attached hashtags as an indicator, which enables multiple topics to be assigned to one social media post. This developed model for topic modelling is data-driven and requires no well-organized training data or *a priori* knowledge such as topic counts, thus reducing potential biases.

3.2 Methodology

3.2.1 Proposed model: hashtag network

As noted above, LDA methods have several limitations when dealing with geotagged social media data, so a new method is developed to investigate the pattern of topics of social media contents. The assumption is that the semantic information, such as hashtags, is highly relevant to the topics of a social media post, and that hashtags of similar semantic meaning have a high probability of co-occurrence. Consequently, in a rich hashtag environment, a 'hashtag network' model is built to investigate the geographic patterns of topics. Relevant data models are defined as below:

As described in Chapter 2, let P be a collection $\boldsymbol{P} = \{pst_1, pst_2, \dots, pst_n\}$ of social media posts. Let $\boldsymbol{HT} = \bigcup_{i=1}^{n} pst_i.tgs = \{tg_1, \dots, tg_m\}$ be the hashtag union of $pst_i.tgs$ $(i = 1, \dots, n)$, i.e., if and only if $item \in pst_i.tgs$ $(i = 1, \dots, n)$, $item \in HT$.

The co-occurrence function $CRC(pst_i, tg_j, tg_k)$ is defined as follows:

$$CRC(pst_i, tg_j, tg_k) = \begin{cases} 1, & if\ tg_j \in pst_i.tgs \ \text{AND}\ tg_k \in pst_i.tgs \\ 0, & otherwise \end{cases} \quad (3.1)$$

Where pst_i is a post item from collection P, tg_j and tg_k are hashtags from HT.

A 'hashtag network' is further defined as $HTN = (V, E)$, where V is a set of vertices, each of which denotes a hashtag tg_i in HT, and E is a set of undirected edges connecting the vertices. The edge weighting is calculated as:

$$A_{jk} = \sum_{i=1}^{n} CRC(pst_i, tg_j, tg_k) \quad (3.2)$$

Where A_{jk} is the weighting assigned to the edge connecting vertices $j(tg_j)$ and $k(tg_k)$. The weighting, calculated by summing the co-occurrence frequency of the corresponding pair of hashtags, represents the connectivity between nodes.

3.2.2 Semantic community detection by dividing the hashtag network

The 'hashtag network' is to be grouped into several communities based on their connectivity, so that the vertices within the same community have a dense connection and share the same topic. The concept of Modularity (Newman 2006) is imported as a measure of the strength of network division and connectivity. Modularity Q is often used in optimization methods for detecting a community in a network and is defined as follows (Blondel et al. 2008):

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \qquad 3.3$$

Where A_{ij} is the weighting of the edge connecting vertices i and j, $k_i = \sum_j A_{ij}$ is the sum of the weightings of edges linking to the vertex i, $m = \frac{1}{2}\sum_{i,j} A_{ij}$, c_i is the community to which vertex i belongs, the δ-function $\delta(u, v)$ is 1 if $u = v$, and 0 otherwise.

To detect communities, the Louvain algorithm is employed (Blondel et al. 2008). As a greedy method inspired by the optimization of modularity, the Louvain algorithm is data-driven and does not require an *apriori* selection of the community count, so that network communities can be detected with less potential perceptual biases. The algorithm is applied to the 'hashtag network' *HTN* and several hashtag communities can be detected. Based on the semantic meanings of hashtags, a topic can be assigned to each community (Figure 3.1). The topics of social media post $pst_i = (u, v, t, tgs)$ can be identified by referring to the topics of $p_i.tgs$, enabling to provide multiple topics for one post.

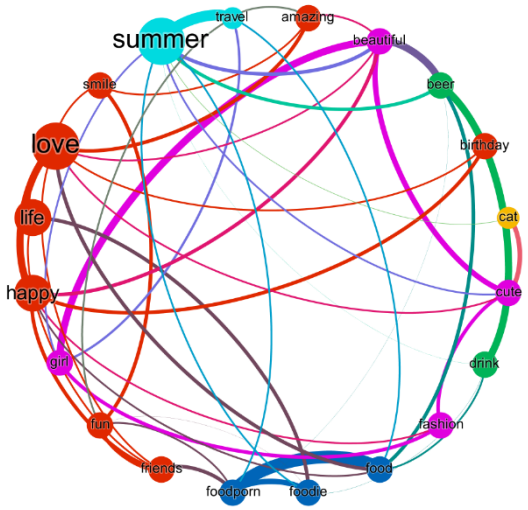


Figure 3.1 Hashtag community detection based on co-occurrence pattern and modularity, with node and edge size proportional to occurrence frequency and node color indicating community category.

3.3 Evaluation approach

The evaluation framework is illustrated in the Figure 3.2. Regarding the analytical workflow, evaluations are made at the two levels: semantic representativeness and topic identification.

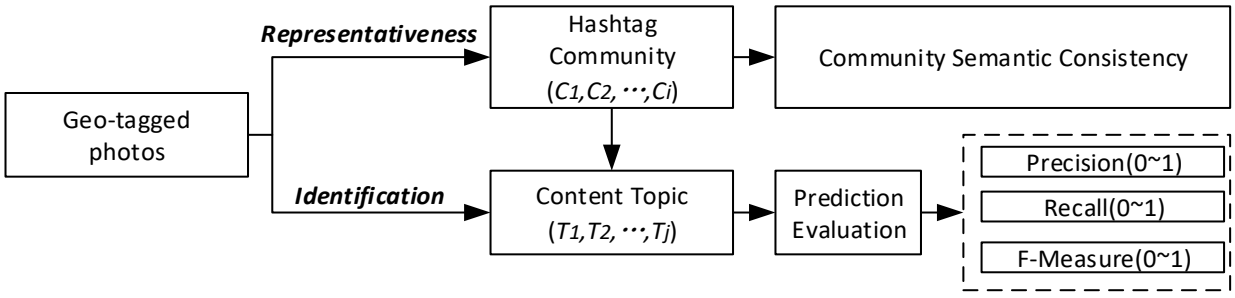


Figure 3.2 Evaluation Framework

Each aspect needs to answer the following questions:

- **Semantic Representativeness**: How well can the semantic similarity between hashtags be represented by the connectivity between vertices in a 'hashtag network'? In other words, from the semantic perspective, are the semantic communities detected internally consistent and externally different?
- **Topic Identification**: To what extent can the topics of the social media content (i.e. photos) be accurately identified by the proposed model?

By introducing quantitative measurements from these two levels, the effectiveness of the proposed geographic topic modelling method is evaluated.

Measurements for Semantic Representativeness: Google *Word2vec* is implemented to investigate the semantic similarity between hashtag communities. *Word2vec* is a group of neural network models that take a large corpus of text as training data and produces a multidimensional vector space, where each unique word in the corpus is represented by a corresponding vector. Words that share common contexts are located in close proximity to one another in the vector space and have a high cosine similarity (Mikolov et al. 2013)

The corpus of the texts attached to geotagged photos is input as training data to create vector space. Let C_m and C_n be two detected hashtag communities, the community semantic similarity (abbreviated as CSS) between C_m and C_n is calculated as below:

$$CSS(C_m, C_n) = \frac{\sum_i \sum_j sim(tg_i^m, tg_j^n)}{tg\_cnt(C_m) * tg\_cnt(C_n)} \quad 3.4$$

Where tg_i^m is the i^{th} tag in C_m, sim(tg_i^m, tg_j^n) is the cosine similarity between tg_i^m and tg_j^n in the *Word2vec* vector space, and tg_cnt(C_m) is the total count of the tags of C_m. The higher the value of *CSS* is, the more similar the hashtag communities are from the semantic perspective.

Measurements for Topic Identification: three criteria: *precision*, *recall* and *F-Measure*, are used to measure the performance of photograph topic identification. These three criteria are commonly used in pattern recognition and information retrieval. *Precision* is the fraction of correct positive predictions among all positive predictions. *Recall* is the fraction of correct positive predictions among all positive instances. *F-Measure* is the harmonic mean of *precision* and *recall*. The higher the values of *Precision*, *Recall* and *F-Measure* are, the better the topic prediction performance is.

3.4 Experiment

3.4.1 Datasets and settings

The geotagged social media photos are used in the analytical workflow. A set of geotagged photos from Hong Kong are retrieved using the Instagram API between Nov 2014 and Nov 2015. The dataset has 1,774,596 geotagged photos generated by 57,662 users (as shown in Figure 3.3). A social media post includes the post ID, user ID, the attached hashtags, the timestamp of the online post and the location indicating the place of posting (Table 3.1). All user IDs are anonymized for privacy protection.

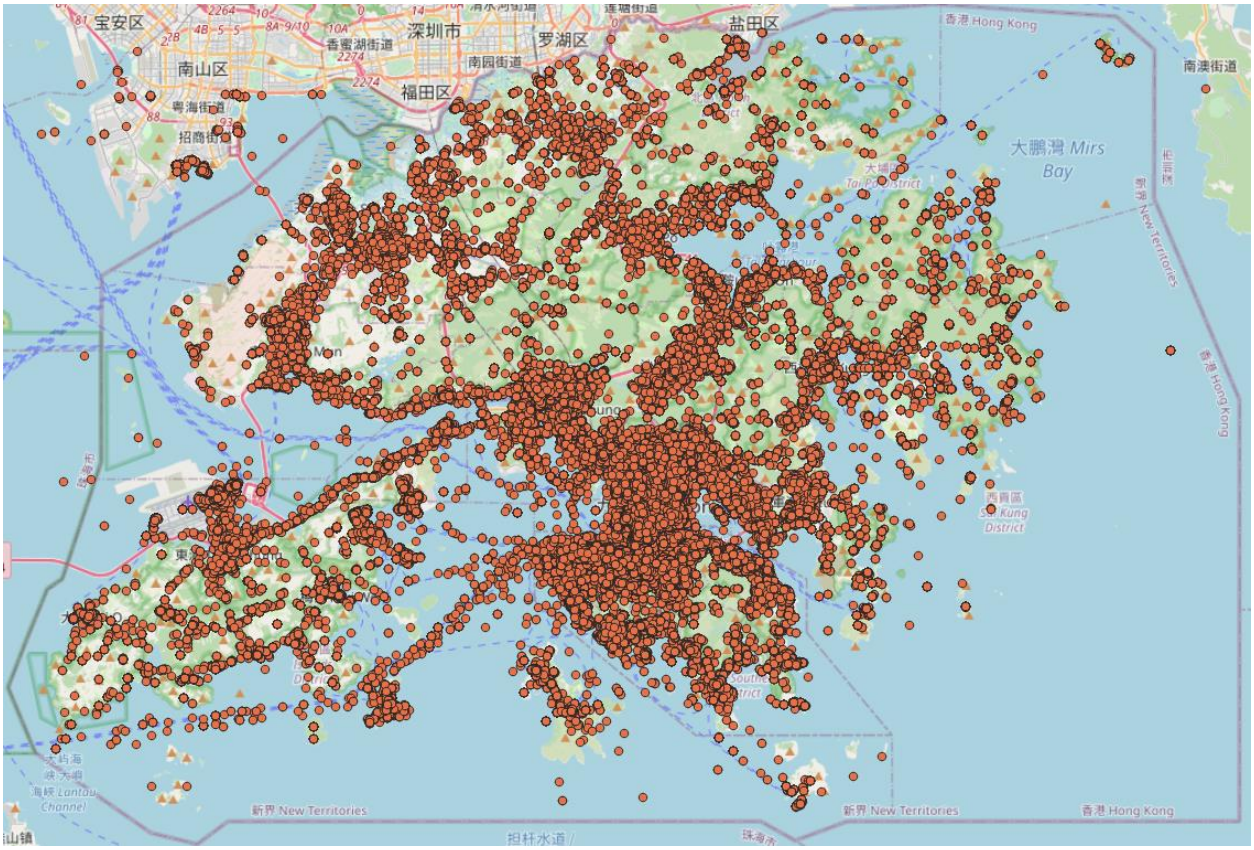


Figure 3.3 The map of original Instagram geotagged check-ins used for experiment

Table 3.1 Field description of the retrieved social media datasets

Fields	Description
pid	A string uniquely indicating a post
uid	An encrypted string uniquely indicating a user
hashtags	The attached hashtags indicating the potential post topics
stime	The time that the user publishes the online post

location	The post location

An anomaly detection method is applied to remove commercial advertising accounts. The count of photos posted by each user is first investigated. Figure 3.3 shows the proportions of users, by number of posts, with its long tail distribution. The calculation shows that the average count of posted photos per user, μ, is 30.1 and the standard deviation σ, 66.4. The three-sigma rule (Pukelsheim 1994) is introduced, which stated that, for both normally distributed and non-normally distributed variables, most cases should fall within the three-sigma intervals. Therefore, those users with photo counts outside the three-sigma intervals (i.e., $\mu + 3\sigma$, 229) is recognized as outliers and their posted photos removed from the photo dataset. After data cleaning, 1,432,733 photos, generated by 56,878 users, remained.

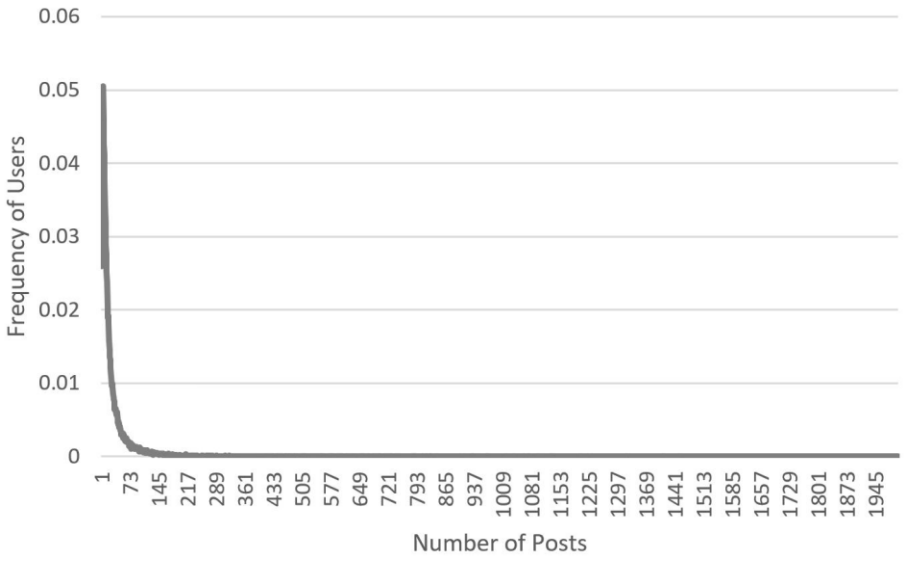


Figure 3.4 Proportion of users by number of posts

3.4.2 Semantic representativeness

The hashtag network *HTN* is built first with the experimental dataset. To strengthen the robustness of the division of *HTN*, hashtags appearing only once in the dataset are removed from

the network as such hashtags are highly probably generated by user typing mistakes and contain little semantic information. The Louvain algorithm is implemented into the *HTN* and several hashtag communities are detected. The divisions of *HTN* are visualized using a comparison word cloud of the representative hashtags in each community (Figure 3.4).

Figure 3.5 A comparison word cloud of 20 hashtag communities

Table 3.2 shows the semantic similarities between hashtag communities. For a hashtag community C_n, intra community semantic similarity (*intra-CSS*) calculated by $CSS(C_n, C_n)$ is introduced (see Equation 3.4) to measure the internal semantic consistency within the hashtag community C_n and inter community semantic similarity (*inter-CSS*) calculated by $CSS(C_n, C_m)$ $(n \neq m)$ to measure the semantic differences between communities C_n and C_m. Semantically, high *intra-CSS* can indicate that the detected community is internally consistent; significant differences between *intra-CSS* and *inter-CSSs* can indicate the detected community is externally differentiable.

Table 3.2 Comparison of the semantic similarities between communities

Hashtag Communities	intra-CSS	inter-CSS		
		maximum inter-CSS	minimum inter-CSS	average inter-CSS
Drinking&Bar	0.85	0.59	0.32	0.33
Travel&Wander	0.90	0.65	0.27	0.37
LifeStyle&Happiness	0.89	0.68	0.37	0.50
Sports&Fitness	0.86	0.56	0.32	0.37
Food	0.87	0.66	0.15	0.43
DisneyLand&Toys	0.86	0.57	0.23	0.47
Tattoo	0.91	0.34	0.22	0.24
Coffee	0.84	0.49	0.22	0.32
Pets&Animals	0.89	0.56	0.21	0.39
Shop&Luxury	0.85	0.59	0.19	0.21
Watch&SportsCar	0.91	0.37	0.26	0.29
Concert&LiveMusic	0.89	0.58	0.23	0.39
Bike&Cycling	0.92	0.45	0.23	0.25
Vegetarian	0.86	0.49	0.13	0.33

Makeup	0.86	0.54	0.26	0.36
Kids&Baby	0.91	0.49	0.26	0.30
Movie&Video	0.88	0.53	0.29	0.35
Muslism	0.85	0.48	0.19	0.30
Big Bang Group	0.88	0.51	0.23	0.30
Art&Design	0.85	0.48	0.22	0.30

3.4.3 Topic identification

The topics of geotagged photos are identified by referring to the attached hashtag topics. To evaluate the effectiveness of the topic identification, photos assigned to the most popular topic 'food' are manually examined, covering 358,471 photos (25.0% of the total dataset) in total.

As shown in Table 3.3, photo samples identified as 'food' topic are drawn in sizes N, where N = 1000, 2000, 3000, 4000, to assess the topic identification results.

Table 3.3 'Food' photo identification

Sample Number	**True Positive, TP**	**False Negative, FN**	**False Positive, FP**	**True Negative, TN**	**Total Size**
1	251	19	31	699	1000
2	520	40	59	1381	2000
3	734	68	96	2102	3000

4	1124	83	152	2641	4000

Figure 3.5 shows the *precision*, *recall* and *F-measure* calculated from Table 3.3. The fluctuations of *precision*, *recall* and *F-measure* is investigated with a varying sampling scale. As the size of a data sample grew, the values of the three measurements fluctuates and gradually stabilizes around 0.9. When N =4000, the *precision* is 0.881, *recall* is 0.931and the *F-measure,* 0.905. This means that on average, for every 10 photos identified as 'food' topic by the proposed model, there are at least 9 items that are actually food-relevant photos, and on the other hand, for every 10 food-relevant photos posted by users, there are 9 items accurately identified and assigned to 'food' topic by the method. Such a high ratio of true positive items demonstrates the good performance of the method for photograph topic modelling and identification. In Chapter 4 and Chapter 5, as the spatial cluster algorithm is implemented based on the spatial distribution of geotagged photos of specific topics, effective exploration of the thematic regions can also be demonstrated by good topic identification performance.

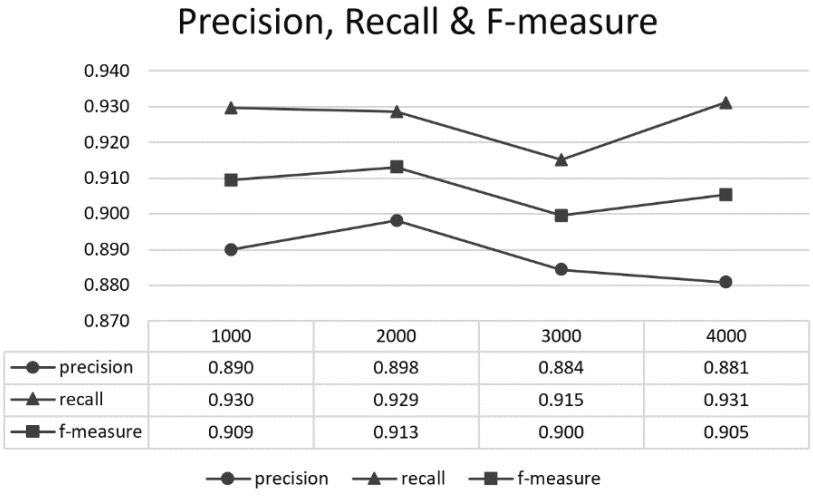


Figure 3.6 *Precision, Recall* & *F-measure* of 'food' photos

3.5 Discussion

The results in Table 3.2 indicates that the detected communities showed very high levels of *intra-CSS* (about 0.90). The top 3 *intra-CSS* scores are achieved by communities ‘Tattoo’, ‘Watch&SportsCar’ and ‘Bike&Cycling’. These three communities also achieve relatively low values for average *inter-CSS* (ranging from 0. 24 to 0.29). By studying the text content and users in details, it is found that these three hashtag communities are mainly popular among population with strong specific interest. When they post photos of relevant topics on the social media platform, the attached hashtags are always very activity-oriented and even specialized, which means the hashtags in these communities are mainly related to very specific activities, e.g., tattooing or cycling, and the chances that such hashtags co-occurred with those of other communities are relatively low. For example, the hashtag ‘solotattoo’ from the ‘Tattoo’ communitiy, may occur frequently with ‘ink’, another popular ‘Tattoo’ community hashtag, but seldom co-occurs with the hashtags from other semantic communities such as ‘Vegetarian’ or ‘Pets&Animals’. Consequently, such peculiar co-occurrence patterns with these three communities contribute to stronger internal than external community connectivity, leading finally to high *intra-CSS* and low *inter-CSS*. A similar explanation can also be applied to another relevant phenomenon. Community ‘LifeStyle&Happiness’ had a relatively high *inter-CSSs* value, indicating a strong external connectivity. A potential reason is that the hashtags in ‘LifeStyle&Happiness’ are mainly emotion-related (e.g., ‘amazing’, ‘happy’, ‘enjoy’). These emotion-related hashtags tend to co-occur with other activity-oriented hashtags to express user emotional reactions and evaluations upon the activities, leading to communities with more external connectivity and higher *inter-CSS* values. Among all communities, the *intra-CSSs* mostly achieve very high values (over 0.85) and the differences between *intra-CSSs* and average

inter-CSSs are mostly over 0.4 (the exception is ‘LifeStyle&Happiness’, where the difference is 0.39). The results verify, semantically, that the divisions of semantic communities are internally consistent and externally different.

3.6 Summary

Chapter 3 focuses on addressing the flowing gap: how to effectively discover topics with short noisy social media content.

Traditional statistical methods have limited effectiveness in handling social media data. This is mainly because that statistical methods commonly require large amounts of well-organized documents as training data, which is inconsistent with the social media environment where short and noisy texts predominate; another reason is that some statistical methods require predefined parameters, such as counts of topics, which are quite arbitrary and unpredictable due to lack of a priori knowledge.

Consequently, a new unsupervised topic modelling method is proposed by using the self-explanatory information (i.e., hashtags) attached to social media content. A ‘hashtag network’ is constructed where each network node represents a hashtag and the co-occurrence frequency between hashtags is assigned to the weight of the edge connecting the corresponding nodes. A greedy optimization method is used to segment the ‘hashtag network’ to explore semantic communities and the attractive regions are further detected by using a density-based clustering algorithm.

The performance of the proposed model is evaluated by investigating the semantic similarity between semantic communities and the accuracy of identifying the topics of photos. The results show (1) the divisions of semantic communities are internally consistent and externally different,

(2) the topics of the photos can be identified with high accuracy, which demonstrates the good performances of the proposed topic modelling method.

Chapter 4 Spatial semantic modelling I: a scale-concerned model for calculating regional desirability

4.1 Introduction

In this chapter, the following questions are studied: where are the regions of specific themes, inferred from GSMD? How desirable are these regions? Previous work on location recommendation mostly focused on recommending venues and POIs (points of interest). These studies are limited in their range of application scenarios. For example, a user may prefer to visit a region with many shops in order to compare options, rather than be simply recommended a single venue with no alternatives. Under such circumstances, POI recommender systems are of limited value as they fail to offer guiding information when users' demands are to explore thematic regions with multiple desirable venue options. A challenge is how to develop a recommending strategy which takes account of varying spatial scales. An example is that regions with large areas may have more visits than those with small areas because they have more venues. The spatial scale issue is specific to region recommendation, as venue recommendations treat the venue as points which do not have spatial sizes.

To address the aforementioned challenge, a new model is developed to calculate regional desirability, which models mutual the reinforcement between region desirability and user expertise with the consideration of region's spatial scale and user redundant visits. The workflow for calculating regional desirability is as below. Firstly, the geotagged photos (check-ins), whose topics have been identified using the method in previous chapter, are further processed by a density-based clustering algorithm to acquire the attractive regions. Secondly, derived from Hyper-Induced Topic Search (HITS) (Zheng and Xie 2011), a new model is developed to predict region desirability and user expertise with the input of attractive regions, topic-identified check-

ins and venues in terms of the varying geographic range. Besides, a mobile application prototype that implements the developed methods is also introduced, to demonstrate the effectiveness of the methodology.

4.2 Methodology

4.2.1 Thematic region detection

To discover attractive thematic regions, the DBSCAN spatial clustering method (Ester et al. 1996) is applied to the geotagged posts of one specific topic (e.g., food, shopping, tourism), which is identified with the topic modelling method described in Chapter 3. DBSCAN can identify clusters of arbitrary shape with tolerance of data noise and does not require counts of the clusters as parameters. Two parameters are needed in the DBSCAN algorithm, e.g., the radius of a cluster (Eps) and the minimum number of points (MinPts) in a cluster. A two-step procedure is devised to select values for Eps and MinPts. Firstly, the value of Eps is determined according to k-dist plot (Ester et al. 1996), which is a graphical representation of points sorted in descending order of their k-dist values. By detecting the first 'valley' visually, the points to the left of the threshold are considered as noise and the k-dist value of the threshold is used as the Eps value for DBSCAN. Secondly, the MinPts is determined using the following equation proposed by (Zhou, Wang and Li 2012), by calculating the neighborhood of every point in the dataset:

$$MinPts = \frac{1}{n}\sum_{i=1}^{n} p_i \qquad 4.1$$

Where p_i is the number of points in Eps neighborhood of point *i* and *n* is the total number of all points.

4.2.2 Proposed model: regional desirability calculation with HITS-based model

For discovered regions, a HITS-based model is used to predict user expertise and region desirability. The idea of Hyper-Induced Topic Search (Zheng and Xie 2011) model is illustrated in Figure 4.1. Users and regions are all perceived as nodes and a user check-in to one region is regarded as a directed link from user node to region node. The mutual reinforcement relationship is that a user who visits many attractive regions is more likely to be an experienced user and a region that is visited by many experienced users is more likely to be an attractive region. A hub score is assigned to a user and an authority score to a region respectively to indicate user expertise and region desirability.

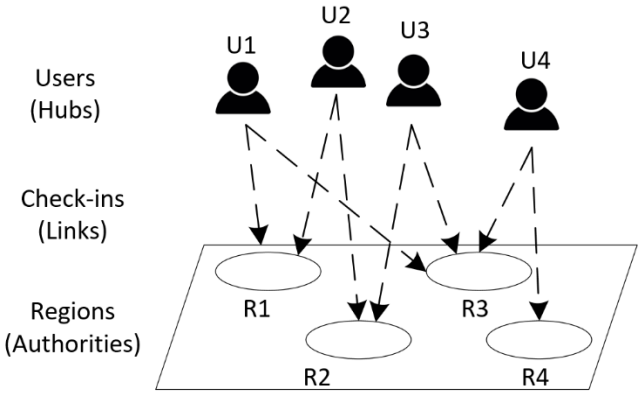


Figure 4.1 HITS-based candidate prediction model

One problem is that a region visited frequently does not necessarily have high quality and desirability, as the chances that a region is visited are probably proportional to its spatial scale (more specifically, venue count). Another problem is that a user with a specific interest may affect the calculation of region desirability and user expertise if a user visits one specific region overmuch due to personal preferences. Consequently, data models are built as below: as described in Chapter 2, *U* is the user set with records of region visiting history (inferred from *P*,

see Chapter 2), and R is a region set with records of users who have visited (the regions are clustered by the method described in Section 4.2.1). For each user $u_i \epsilon U$ and each region $Rn_i \epsilon R$, a hub score $h(u_i)$ and an authority score $a(Rn_i)$ are calculated respectively using the following equations:

$$h(u_i) = \sum_{Rn_j \,\epsilon\, u_i.R} (1 + \ln(\sigma)) * a(Rn_i) \quad 4.2$$

$$a(Rn_i) = \sum_{u_j \,\epsilon\, Rn_i.U} (1 + \ln(\lambda)) * h(u_j) \,/ n(Rn_i) \quad 4.3$$

Where $u_i.R$ are the regions that user u_i visited, $Rn_i.U$ are the users who have visited region Rn_i, σ is the number of visits of user u_i to region Rn_j, λ is the number of visits of user u_j to region Rn_i, and $n(Rn_i)$ is the count of venues of a specific category (e.g., food venues, shopping venues, etc.) in region Rn_i. By applying the log function on the number of users' visit (i.e., $\ln(\sigma)$ and $\ln(\lambda)$), we introduce a decay effect so that the increase of single user's influence decreases as the number of their visits increased. Moreover, by introducing venue count into the model, the calculated authority value can reflect the average desirability of the venues within the region so that, taking account of the impact of region spatial scale, the desirability of the recommended region is predicted.

4.2.3 Application prototype design

The proposed region detection and recommendation workflow is implemented with an android mobile application. This phase starts when users initiate a query asking for a certain count of regions matching an interest (e.g., food) within a spatial range (e.g., 5 km). Regions meeting the criteria are selected from the candidate region set. Finally, a list of regions is ranked in descending order by desirability score and recommended to users. A user query is represented as below:

$$q = [interest, location, spatial\ range, count\ ] \quad (4.4)$$

Algorithm 1: Query Response Method

Input: (1) User Interest i (2) Location loc (3) Spatial Range rng (4) Total Count of Recommended Regions cnt

Output: A ranked list of region recommendations L

Begin

Retrieve regions R' that match user interest i

Select regions R'' from R', which are within the spatial range rng from loc

for $r_i \in$ R'' **do**

$A \leftarrow r_i$.auth // Get the authority score of region r_i

end

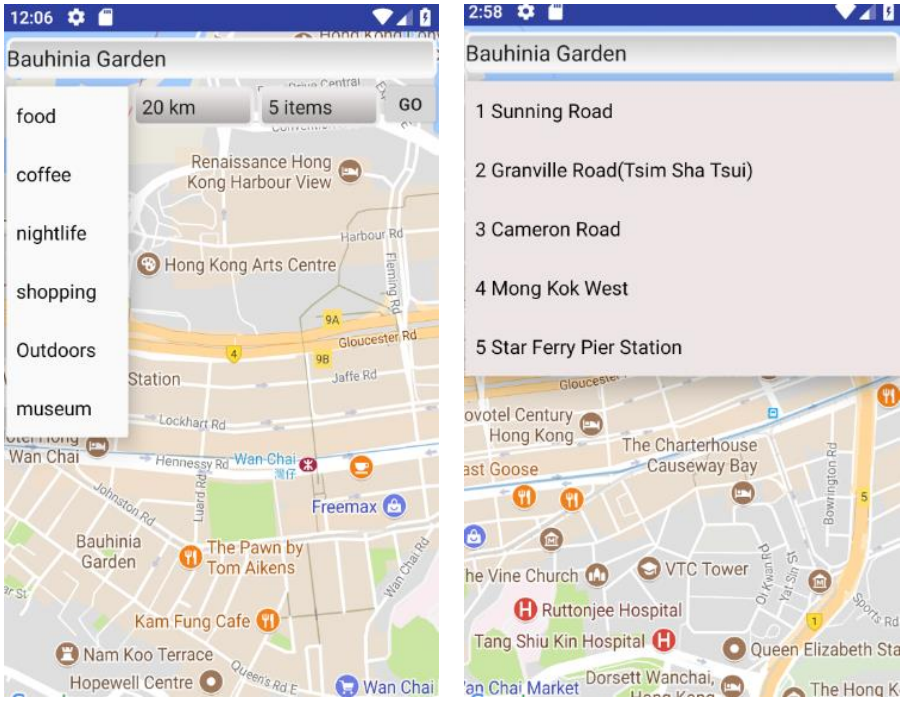
$L \leftarrow$ Rank (R'', A, cnt) // Rank the regions in R'' in terms of authority score

Return L // Return the first cnt recommended regions to the user

End

After receiving the query, the response module ranks the regions discovered in the candidate prediction module and recommends a list of regions that match the user interest within the spatial range of the location. The ranking is based on region desirability score, i.e., ranking the regions in descending order by authority score. The query response method is explained in detail in Algorithm 1.

A snapshot of the user interface of the application is shown in Figure 4.2. A typical usage scenario starts when a user logs into the application to obtain personalized recommendations. The user inputs query location, interest (e.g. food, coffee), spatial range and item count in the system. After clicking the “Go” button, a list of regions matching the request is returned on the screen (Figure 4.2 b). For the application architecture, Microsoft Access database is used to store the region attributes and spatial information and Google Maps API used for map viewing and geocoding.



(a) (b)

Figure 4.2 Application user interface. (a) home page;(b) recommendation list

4.3 Evaluation approach

4.3.1 Evaluation measurements

For evaluation of regional desirability, the following questions are answered: How effective is the recommending strategy? Do the recommended regions match user expectations?

The effectiveness of the ranking and recommending strategy is measured with two quantitative criteria: *MAE* (mean absolute error) and *nDCG* (Normalized Discounted Cumulative Gain). *MAE* is commonly used to measure the agreement between system recommendation and user rating. In the experiments, *MAE* is calculated using the following equation:

$$MAE = \sum_i^n \frac{|SR_i - UR_i|}{n} \quad 4.5$$

Where SR_i is the system rating for the i^{th} region, UR_i is the user rating for the i^{th} region and *n* is the total region count. Before calculation, *MAE*, SR_i and UR_i are both normalized to be mutually comparable. The lower the value of *MAE* is, the closer the system ratings are to the user ratings in terms of region desirability.

nDCG is a measurement of ranking quality. *nDCG* measures the gain of an item based on its position in the list, with the gain discounted at lower ranks. The *nDCG* accumulated at a particular rank position p is calculated using (Yilmaz, Kanoulas and Aslam 2008):

$$nDCG_p = \sum_{i=1}^{p} \frac{2^{rel_i - 1}}{log2(i+1)} / IDCG \quad 4.6$$

Where rel_i is the graded relevance at position *i*, and *IDCG* is the ideal *DCG* calculated by sorting all the items by their relevance. A high value of $nDCG_p$ indicates that the regions recommended by the system are of high desirability and match well with user expectations.

4.3.2 Baseline methods

The regions recommended by the proposed model are compared with those recommended by three baseline methods: **rank-by-users**, **rank-by-visits** and **rank-by-desirability**. With the rank-by-users method, the more visitors a region receives, the more attractive the region is. With the rank-by-visits method, similarly, the more check-ins a region receives, the more attractive the region is. For the third baseline method, rank-by-desirability, the desirability of a region is calculated with the following mutual reinforcement equations:

$$h(u_i) = \sum_{Rn_j \in u_i.R} \sigma * a(Rn_j) \quad 4.7$$

$$a(Rn_i) = \sum_{u_j \in Rn_i.U} \lambda * h(u_j) \quad 4.8$$

Where $u_i.R$ is the region set visited by users u_i , $Rn_i.U$ is the user set who have visited region Rn_i, σ is the number of visits of user u_i to region Rn_j, λ is the number of visits of user u_j to region Rn_i, and the desirability of each region is decided by the authority score. Compared with the proposed model in this chapter described before (Equations 4.2 and 4.3), the rank-by-desirability does **not** take the influences of user specific preferences and spatial scale of region size into consideration.

4.3.3 Ground truth

A survey of users' ratings on the recommended regions is needed to evaluate the effectiveness of recommendation. As a region spans across certain spatial ranges and covers many venues, an indicator is needed to evaluate users' general attitude towards the recommended regions as a whole. For each region Rn_i, the desirability score $Rn_i.Rt$ is calculated as:

$$Rn_i.Rt = \frac{1}{|Rn_i.V|} \sum_{v_j \in Rn_i.V} v_j.vrat \quad 4.9$$

$v_j.vrat$ (Definition 3) is the venue rating score retrieved from Foursquare platform. The user venue rating scores can be retrieved from Foursquare API. This rating score is calculated based on a variety of users' real-world explicit feedback and has been validated for metropolitan regions for accuracy and reliability in indicating venue desirability (Yang and Sklar 2016). By using Equation 4.9, the user satisfaction level for a whole region can be derived. The effectiveness of the recommendation is evaluated by comparing the ranking and scoring results with the results generated based on $Rn_i.Rt$.

4.4 Experiment

The effectiveness of the proposed model is evaluated with real-world datasets.

4.4.1 Datasets and settings

There are two kinds of data source needed in the proposed workflow: geotagged Instagram check-ins and Points of Interest (POIs). The geotagged Instagram check-ins are the same as described in Section 3.4.

The POIs dataset is retrieved via Foursquare API and 32,485 venues are collected in Hong Kong in total. A POI item includes the venue ID, venue category, venue location (longitude and latitude), and venue rating (Table 4.1). The venue rating is a numerical score (0 through 10) calculated from a wide variety of signals derived from user explicit feedback, such as: liking or disliking a venue, leaving a positive or negative tip, as well as user implicit signals, such as: whether the venue tends to have many loyal customers, the credibility and expertise of the users and so on. This rating algorithm had been validated in metropolitan areas and trusted by users for accuracy and reliability in indicating venue desirability (Yang and Sklar 2016). This rating score calculated from actual crowdsource feedback may be more capable of indicating the

venue/region desirability than the user feedback in a controlled experiment, where, in most occasions, only limited amounts of subjects will be included and inquired.

Table 4.1 Description of POIs field

Fields	**Description**
vid	A string uniquely indicating a venue
vcategory	A string indicating the category of venue
location	The post location
vrating	A numerical rating of the venue (0 through 10), which is calculated from a wide variety of signals derived from users' explicit feedback (e.g. like or dislike, positive or negative tips) and implicit feedback (e.g. customer loyalty, user credibility and expertise)

By using the *k-dist* plot and Equation 4.1, the DBSCAN method parameter is set as *eps* =71 and *minPts* =1820, and 21 'food' theme regions are detected. By using Foursquare API, 1960 venues of food category are retrieved in those 21 regions.

4.4.2 Performance evaluation

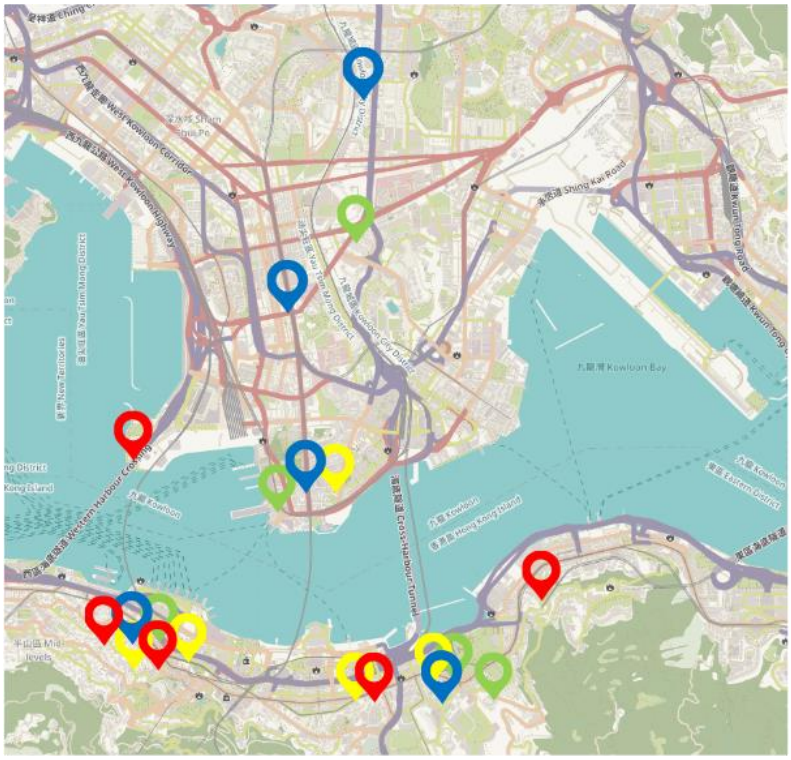


Figure 4.3 Top 5 recommended regions of 4 different themes. Green regions for 'Food' theme; Yellow for 'Drinking&Bar' theme, Blue for 'Shop&Luxury' theme; Red for 'Art&Design' theme

Figure 4.3 shows the top 5 recommended regions of different themes: 'Food', 'Drinking&Bar', 'Shop&Luxury', 'Art&Design'. Table 4.2 lists the top 10 regions of 'food' themes recommended by each method. The *rel* in $nDCG_p$ (Equation 4.6) is set based on *UR*, which means the region with higher *UR* had higher *rel*. The regions are sorted in descending order by *UR*. As there are 21 detected regions in total, the *rel* of the first region in the list is set as 21. The *rel* decreases by 1 to the next region in the sorted list and finally the *rel* of the last region is set as 1.

Table 4.2 Top 10 regions of 'food' themes recommended by each method

Rank	The proposed model	Rank-by-users	Rank-by-visits	Rank-by-desirability	Rank-by-UR (Ground Truth)
1	Star Ferry Pier Station	Lyndhurst Terrace	Lyndhurst Terrace	Lyndhurst Terrace	Star Ferry Pier Station
2	Canton Road (Tsim Sha Tsui)	Causeway Bay	Causeway Bay	Causeway Bay	Hong Kong Station-Man Cheung Street
3	Tai Hang	Cameron Road (Tsim Sha Tsui)	Cameron Road (Tsim Sha Tsui)	Cameron Road (Tsim Sha Tsui)	Statue Square
4	Hong Kong Station-Man Cheung Street	Sunning Road	Sunning Road	Sunning Road	Canton Road (Tsim Sha Tsui)
5	Kingston Street-Gloucester Rd	Mody Road	Mody Road	Mody Road	Lyndhurst Terrace
6	Austin Road West	Statue Square	Statue Square	Statue Square	Rodney Road (Admiralty)
7	Sunning Road	Canton Road (Tsim Sha Tsui)	Tung Wah	Canton Road (Tsim Sha Tsui)	Austin Road West

8	Statue Square	Tung Wah	Canton Road (Tsim Sha Tsui)	Tung Wah	Kingston Street-Gloucester Rd
9	Tung Wah	Hong Kong Station-Man Cheung Street	Southorn	Southorn	Sunning Road
10	Rodney Road (Admiralty)	Mong Kok West	Hong Kong Station-Man Cheung Street	Hong Kong Station-Man Cheung Street	Tung Wah

To evaluate the performance of recommendations in various spatial ranges, different values are assigned to p in $nDCG_p$ (Equation 4.6), as shown in Table 4.3. It shows that for all ranking methods, the values of *nDCGs* increased with the increment of p and the best *nDCGs* are achieved when p =21. On average, the proposed recommending strategy can achieve *nDCGs* with an average value as much as around 0.91, which are much greater than those of the three baseline methods (0.2-0.4), indicating that regions with higher user ratings are given more recommendation priority by the proposed strategy. Such results show that items recommended by the proposed model could well match user expectations, demonstrating the advantages of the proposed model over the several baseline methods in terms of effectively ranking and recommending desirable regions.

Table 4.3 Normalized Discounted Cumulative Gain (*nDCG*) for each ranking method

$nDCG_p$	The proposed model	Rank-by-users	Rank-by-visits	Rank-by-desirability
$nDCG_{10}$	0.905	0.225	0.220	0.221
$nDCG_{15}$	0.905	0.228	0.222	0.224
$nDCG_{21}$	0.914	0.389	0.384	0.386

The agreement between the proposed scoring strategy and user rating is further investigated. The results are shown in Table 4.4. For rank-by-users and rank-by-visits strategies, the user count and visit count are used respectively as the desirability score for each region. The *MAE* (mean absolute error, Equation 4.5) indicates that the desirability scores predicted by the proposed model achieved a better agreement with ground truth than the baseline methods.

Table 4.4 *MAE* (mean absolute error) based on ground truth

	The proposed model	Rank-by-users	Rank-by-visits	Rank-by-desirability
MAE	0.353	0.424	0.458	0.444

4.5 Discussion

In this section, some worthy questions are further discussed to provide insights into the proposed methodology: (1) what is the intuition of the proposed HITS-based model? (2) How the ground-truth regional desirability is obtained? Is it trustworthy? (3) Regarding the Algorithm 1, is it an efficient algorithm? What is the change of efficiency if the interest filter and spatial filter steps (first two lines of Algorithm 1) are swapped?

4.5.1 Intuition of proposed HITS-based model

The proposed model is a HITS-based model, with consideration of (1) the mutual reinforcement between regional desirability and user expertise, and (2) the spatial scale/size of the region. For performance evaluation, the nDCG and MAE measurements are used. In calculating *nDCG* (Equation 4.6), the highly desirable regions are given more relevance than marginally desirable regions. Highly desirable regions appearing lower in recommendation list would be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result. The proposed model clearly outperforms the other baseline methods in terms of *nDCG* (Table 4.3) because it provided a more consistent recommendations order, especially among the highly desirable regions at the top of the list, in relation to ground truth, than baseline methods (Table 4.2). This, in turn, demonstrated the method's superiority over baseline methods in effectively meeting user needs in finding regions with high desirability.

The top several regions recommended by the proposed model are all located in the downtown, mainly in Tsim Sha Tsui, Central, Mong Kok, and Causeway Bay, which are the major shopping and recreation areas in Hong Kong. For example, among all 21 recommended regions, 5 are located in Tsim Sha Tsui, 4 in Central, 2 in Mong Kok and 3 regions in Causeway Bay. Such highly concentrated spatial distribution of attractive regions, in some ways, indicates the high-density urban living environment of Hong Kong.

In addition, the proposed model is effective and useful in finding regions with small area yet high desirability, which meets the need of users to find regions where venues of superior quality are located within a reasonably accessible distance. The top 3 recommendations (Star Ferry Pier Station, Canton Road, Tai Hang) by the proposed model are all small-area regions with good quality venues attractive to users. In contrast, for the baseline methods, region size is given much

weight in deciding recommendation priorities. This is because, for the baseline methods, region desirability is strongly positively correlated with numbers of visits, which inevitably related to region size. For example, the experiments revealed that regions such as Lyndhurst Terrace, Causeway Bay and Cameron Road are highly recommended by the baseline methods, as these are large regions with large numbers of venues, which add to the amounts of visits and visitors and eventually, priority for recommendation. However, such priority ranking failed to effectively reflect user satisfaction levels with the regions, as the service quality of the venues varies from item to item and the user average rating to the whole region might be compromised by unsatisfying venues. The proposed strategy, however, takes account of both region size and user redundant visits so that region desirability is more accurately modelled.

4.5.2 Rationale of calculating the ground-truth regional desirability

In this thesis, to compute the ground truth of regions' desirability, the consideration is that the average desirability of the venues within the region can be used as the desirability of the region. This way of computing ground truth is based on the condition that the reliable desirability scores of the venues can be attained.

Some previous works use questionnaires to collect users' rating score on the locations. However, this way of collecting desirability scores of venues is limited in the following two aspects. The first issue is that only limited amounts of subjects (people) can be included and inquired, which makes the attained results inevitably biased and incomplete. Another issue is that there may be hundreds or even thousands of venues within a region, so it is unrealistic or even impossible to have the subjects' feedback on every venue as they may haven't visited or even known this venue at all.

So, another source is used to attain the desirability of venues in this thesis. The user rating score on the venue is retrieved from Foursquare API. This rating retrieved from Foursquare is crowdsourced from individuals and comprehensively calculated from a wide variety of signals derived from users' explicit and implicit feedback, such as liking or disliking a venue or leaving a positive or negative tip, as well as other signals such as whether the venue tends to have lots of loyal customers, the credibility and expertise of the users and so on (Yang and Sklar 2016). Besides, in previous study, this rating algorithm of Foursquare platform has been tested and validated in metropolitan areas and trusted by users for accuracy and reliability in indicating the venue desirability (Yang and Sklar 2016). This retrieved rating score calculated from a wide variety of actual crowdsourced feedback is more capable of indicating the venue desirability than the user feedback in a controlled experiment, where, in most occasions, only limited amounts of subjects will be inquired and incomplete venues' information can be attained. This is the theoretical consideration of why the average of the venue ratings retrieved from crowdsourced platform is used as ground-truth regional desirability.

4.5.3 Efficiency of Algorithm 1

The efficiency of Algorithm 1 is investigated. An alternative algorithm is proposed by swapping the first two lines: first filter the regions, and then create a user-specific ranking. Experiments are conducted with sample data to test the efficiency of the original Algorithm 1 and the alternative algorithm. Based on the current experiment, both ways show no difference in algorithm execution time measured by even microsecond. This may be because that the application firstly reads the whole table from database into memory and then does spatial filtering and interest filtering. In the experiment, the most time-consuming step is data input/output while doing spatial filtering and interest filtering with the data stored in memory is very fast and whichever

filter is executed first, the efficiency yields no different results. Nevertheless, one speculation is that doing interest filter firstly and then spatial filter may be more algorithmically efficient for large-scale dataset, because interest filter tends to be faster than spatial filter as calculating and comparing spatial distance requires relatively more complicated algebraic operations than comparing the venue category with user interest. Doing interest filter first can reduce the total amount of spatial filter, as opposed to the other way around, and this can save more time and be more suitable for large amount of data. Admittedly, this is so far a speculation which requires more theoretical inference and practical experiments to verify.

4.6 Summary

Chapter 4 focuses on addressing the flowing gap: how to develop a model to calculate regional desirability with consideration of varying spatial scales of regions.

Regions with large areas may have more visits than those with small areas because they have more venues. This scale issue is specific to region recommendation, as venue recommendations typically treat the data as points and thus do not consider their sizes.

Consequently, a new scale-concerned model is proposed to calculate the regional desirability with GSMD. Basically, to select and rank recommendations from a set of candidate regions, three particular aspects are considered by the proposed model. Firstly, the consideration is that larger regions tend to include more venues and consequently attract more users and visits, and such influence of region spatial area on the effectiveness of calculating regional desirability needs to be modelled. Consequently, the proposed model develops new equations (Equation 4.2 and 4.3) to quantify region desirability. Secondly, the influence of redundant visits by specific users is taken into account. The method assumption is that specific groups of users may have

particular tastes and certain venues may be frequently visited by them, however, such personal preferences can hardly represent the general opinion of the total user pool. A decay effect is thus introduced into the model so that the increase of single user's influence decreases as the number of their visits increased. Thirdly, the interactive reinforcement effects are recognized in predicting the desirability of regions and the expertise of users. This is derived from the intuitive perception that the more a region is visited by experienced users, the more attractive the region is, and equally, the more a user visits attractive regions, the more experienced the user is likely to be. The proposed recommending strategy aimed to distinguish between local experts and common users so that the attractive regions are more accurately identified.

Chapter 5 Spatial semantic modelling II: a data-driven model for calculating regional desirability

5.1 Introduction

In this chapter, a novel neural network model 'RegNet' is proposed to predict region desirability. The consideration is that traditional methods mainly use empirical models to predict regional desirability, like Hypertext Induced Topic Search (HITS) used in the previous chapter, which are quite intuitive and can be inaccurate in predicting desirable regions, as potential perceptual bias is introduced (Zheng et al. 2009, Zheng and Xie 2011, Liu et al. 2019). This results from the fact that the hidden interactions between user check-ins and regional desirability haven't be clearly explained or precisely modelled yet, so that the empirical models based on human empirical intuition and presumed relationship are used for rough approximation.

On the other hand, the recent development of artificial neural network enables to address the aforementioned challenge from a new perspective. A neural network is a collection of connected units, performing certain tasks like classification and regression (Specht 1991, Odom and Sharda 1990). An advantage of the neural network is that it is data-driven, which means, given certain training data, the neural networks are theoretically capable of universally approximating functions (Hornik, Stinchcombe and White 1989) and learning the hidden unknown interactions of high-level features (LeCun et al. 2015) without prior knowledge, which is suitable for the research problem where the hidden interactive mechanisms of relevant features are still quantitatively unknown.

Consequently, in this chapter, a novel neural network model (called 'RegNet') is proposed for predicting regional desirability. RegNet takes the pairs of user check-in history and regional

desirability score as training data. By adjusting the network weights through backpropagation algorithm, RegNet can adaptively learn the hidden interactions of high-level features and the unknown mappings from input to output, without prior knowledge. Compared with traditional empirical models, RegNet can achieve desirable region predictions with less perceptual bias and better accuracy.

Specifically, RegNet consists of two main parts: a neural network encoder structure for feature learning and a hidden-layer structure for desirability prediction. The region-visit vectors are created firstly and input into the encoder for reduction of data redundancy and computational complexity. The encoded representations are then fed into the hidden-layer structure and a score for regional desirability will be predicted as output value. Evaluations are conducted with real-world datasets and demonstrate that the proposed RegNet outperforms the popular state-of-the-art methods. Besides, how the structure of encoder affects the performance of RegNet is also examined and suggestions are given on how to choose proper sizes of encoded representations.

5.2 Proposed model: RegNet

In this part, the details for the developed model are elaborated. The rationale and techniques for feature representation are given first, followed by an elaboration of the neural network model for region desirability prediction.

5.2.1 Feature learning with autoencoder

For location recommendation with GSMD, the frequency of check-ins is commonly used to indicate the popularity of a location. However, for region recommendation, the regional spatial scale needs to be considered: as stated in Chapter 4, a large region is likely to be visited more often due to more venue amount, but such high frequency of visits doesn’t necessarily indicate

average regional desirability (Liu et al. 2019). Consequently, the region-visit vector $\boldsymbol{rv}_i$ (see Chapter 2), is divided by regional spatial scale to construct new input feature $\boldsymbol{rv}_i'$:

$$\boldsymbol{rv}_i' = \frac{\boldsymbol{rv}_i}{|Rn_i.V|} \quad 5.1$$

The dimension of feature $\boldsymbol{rv}_i'$ is $|U|$, which may be very large given big dataset and lead to several problems in neural network training: (1) the connections between input layer and hidden layer can be very complex, causing overfitting due to the relatively small amount of training data. (2) the large dimensionality of features can give rise to computational inconvenience and even, infeasibility. Consequently, in this work, an autoencoder is adopted first for feature learning. The autoencoder is a pair of transforming and reconstruction functions to learn feature representation, mostly in the form of neural network (Liou, Huang and Yang 2008, Liou et al. 2014) , with both reduction side for dimensionality reduction and reconstructing side for reconstructing the original input. By using autoencoder, the higher abstract features with reduced dimensionality can be learned for region desirability prediction, thus reducing the computational complexity and overfitting.

An autoencoder neural network includes two structures: an encoder and a decoder (shown in Figure 5.1), which can be defined as two mapping functions (Vincent et al. 2010):

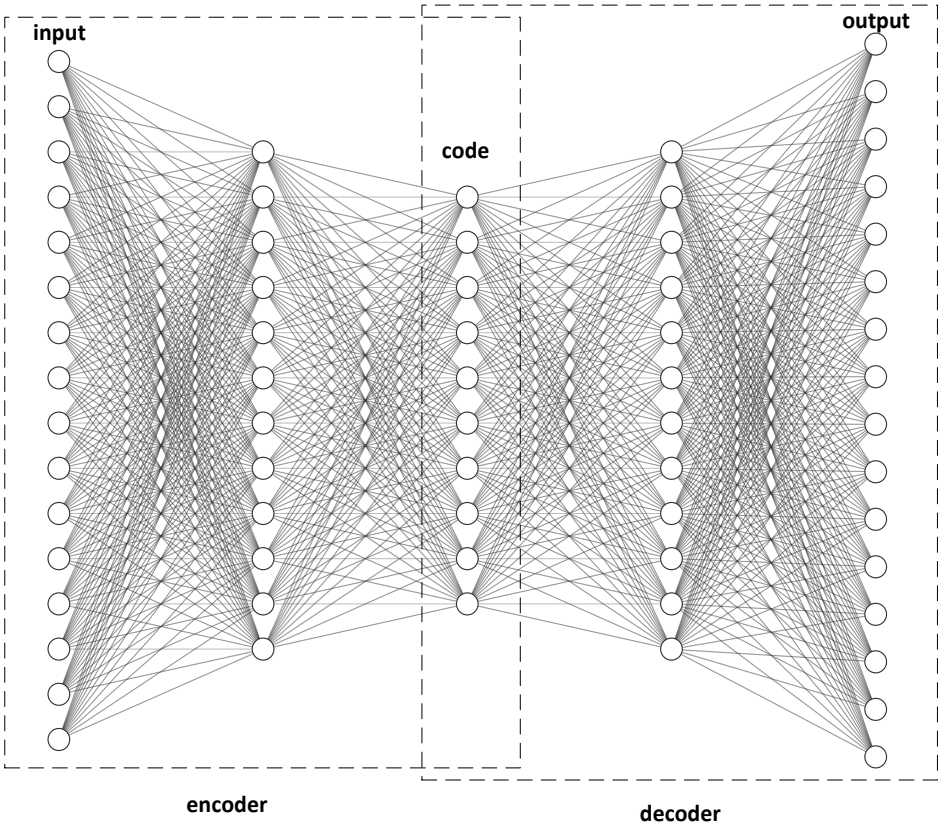


Figure 5.1 An autoencoder in the form of a fully connected neural network. The encoder transforms the input vector with multiple dimensions into a short representation, and the decoder, reversely, transforms the short representation back into vectors with the same dimension as input vector, with the aim to minimize the reconstruction errors.

Encoder: The encoder $f_\theta(\mathrm{x})$ is in the form of neural network and transforms vector into a short representation (code). The mathematical formula can be illustrated as:

$$\boldsymbol{y} = f_\theta(\mathbf{x}) = s(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \qquad 5.2$$

Where s() is a nonlinear function and the parameter $\theta = (\boldsymbol{W}, \boldsymbol{b})$ is the pair of weight matrix and bias vector.

Decoder: The representing code y is transformed back using another neural network to a

reconstructed vector, $\mathbf{z} = g_{\theta'}(\mathbf{y})$:

$$\mathbf{z} = g_{\theta'}(\mathbf{y}) = s'(\boldsymbol{W}'\boldsymbol{y} + \boldsymbol{b}') \quad 5.3$$

With parameter $\theta' = (\boldsymbol{W}', \boldsymbol{b}')$.

The training process aims to minimize the reconstruction error through backpropagation algorithm. The mean squared error is used as loss function, so the minimizing reconstruction errors is defined as (Vincent et al. 2010):

$$\arg min_{\theta,\theta'} L(\mathrm{x}, \mathrm{z}) = \frac{1}{|\boldsymbol{x}|}\left(\mathbf{x} - \mathrm{s}'\left(\boldsymbol{W}'\left(\mathrm{s}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})\right) + \boldsymbol{b}'\right)\right)^2 \quad 5.4$$

In this chapter, an autoencoder is trained first with the input and reconstruction of $\boldsymbol{rv_i}'$. After training autoencoder, the encoder from autoencoder is used independently for feature learning.

5.2.2 Regional desirability prediction with encoding-prediction neural network

The types of stand-alone location recommender systems can be classified into two: POI recommender and region recommender (Bao et al. 2015). For POI recommendation, the problems are usually defined as a binary classification, where a check-in into POI is denoted as positive and a non-check-in as negative (Ding and Chen 2018); For region recommendation, the problems are commonly defined as a problem of predicting regional attractiveness (desirability). Consequently, in this work, the regional desirability prediction problem is defined as a regression problem, where the regional desirability score is predicted based on the user visiting history, using the proposed neural network model RegNet. The structure of RegNet is presented in Figure 5.2.

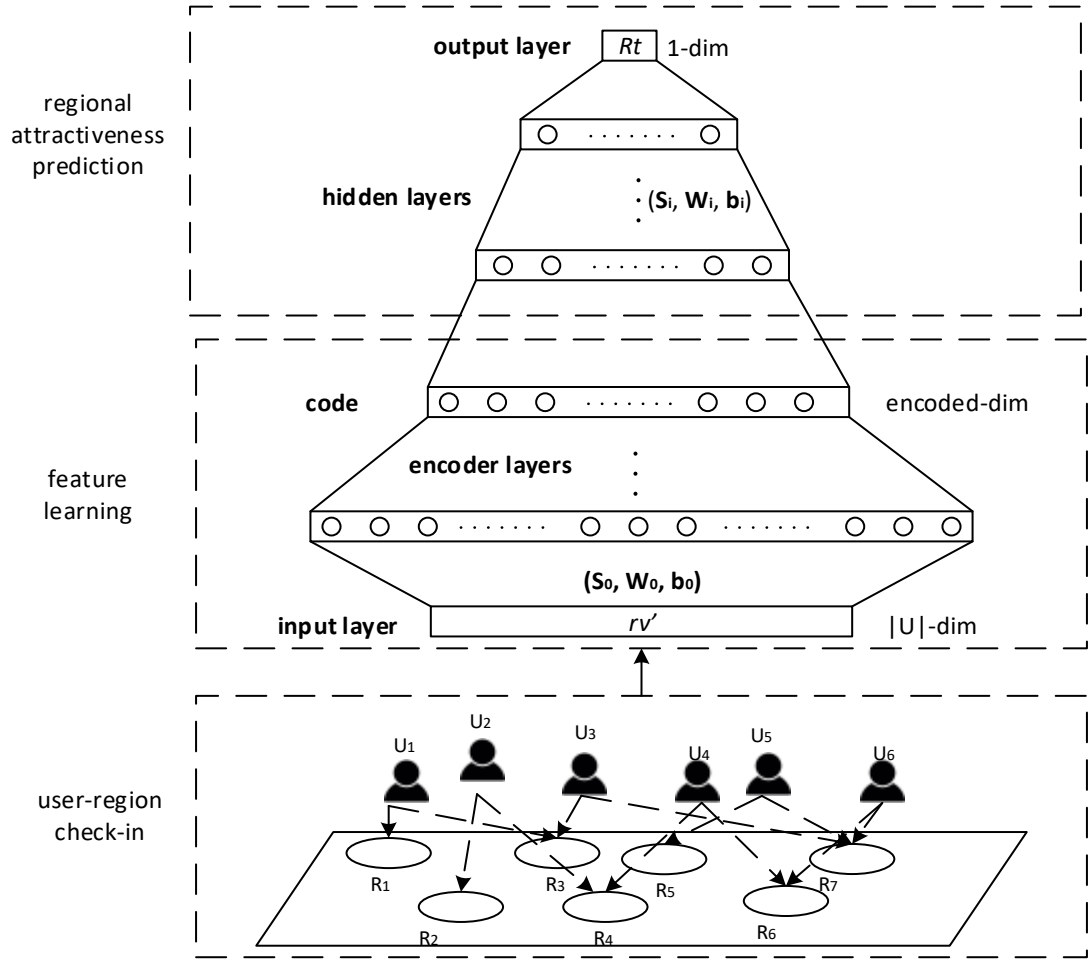


Figure 5.2 The structure of fully-connected RegNet. The scaled region-visit vector and ground-truth region rating, i.e. (*rv'*, *Rt*), are fed into RegNet pairwise for neural network training.

The RegNet consists of several levels: an input layer, a neural network encoder for feature learning, several hidden layers for prediction and an output layer. The encoder network compresses the input feature into an encoded tensor, which is then input to the hidden-layer structure for final prediction.

The workflow of RegNet can be defined as:

$$\boldsymbol{y}_i = \begin{cases} \boldsymbol{rv'}, & i = 0 \\ s_{i-1}(\boldsymbol{W}_{i-1}\boldsymbol{y}_{i-1} + \boldsymbol{b}_{i-1}), & else \end{cases} \quad 5.5$$

where $\boldsymbol{y}_i$ is the input vector of the i^{th} layer. For input layer (i.e., i=0), $\boldsymbol{y}_0$ is the network input feature $\boldsymbol{rv}'$ (see Equation 5.1). For encoder and other hidden layers, $s_i()$, $\boldsymbol{W}_i$, $\boldsymbol{b}_i$ are the activation function, weight matrix and offset vector for i^{th} layer respectively. For output layer, the $\boldsymbol{y}_i$ is the predicted score for regional desirability. To reduce computational cost, the stochastic gradient descent (SGD) is adopted as the optimizer, which can be very effective for the problems with large-scale dataset (Bottou and Bousquet 2008). As the regional desirability prediction is defined as a regression problem in this work, the mean squared error is adopted as the loss function of RegNet training:

$$Loss = \frac{1}{|R|}\sum_{Rn_i \epsilon R}(Rn_i.Rt - Y_i)^2 \quad 5.6$$

Where Rn_i is a region in R (see Chapter 2), Y_i is the predicted score for Rn_i regional desirability prediction from RegNet (see Equation 5.5). The scaled region-visit vector and ground-truth region rating, i.e. ($\boldsymbol{rv}_i'$, $Rn_i.Rt$), are fed into RegNet pairwise for neural network training, aiming to minimize the loss function by modifying ($\boldsymbol{W}_i$, $\boldsymbol{b}_i$) through backpropagation algorithm.

The desirability score for each region will be predicted. The top-k regions with the highest predicted desirability will be returned.

5.3 Evaluation approach

5.3.1 Evaluation measurements

The evaluation measurements are the same as described in Section 4.3.1.

5.3.2 Baseline methods

The baseline methods include: rank-by-users, rank-by-visits (as described in Chapter 4), HITS (Zheng and Xie 2011), HITS-based (Liu et al. 2019) (i.e., the scale-concerned model developed in Chapter 4).

5.3.3 Ground truth

The way to calculate ground truth of regional desirability is the same as described in Section 4.3.3.

5.4 Experiment

5.4.1 Datasets and settings

The datasets for experiment and evaluation consist of two sources: geotagged Instagram check-ins identified as 'food' topic in Hong Kong from Nov 2014- Nov 2015, POIs from Foursquare.

Compared with the datasets used in previous chapters (Chapter 3 and 4), the check-ins used in this chapter solely focus on the check-ins identified as 'food' topic, while the POIs dataset is exactly the same as described in Section 4.4. The basic description of the experimental datasets is given in Table 5.1. An Instagram check-in has following attributes: the check-in id, the corresponding user id, the timestamp of check-in, and the location. A POI item has following attributes: the POI identifier, the POI category, POI location and a numerical rating for POI desirability. The desirability rating is a score (0 to 10) retrieved from Foursquare platform and calculated from a wide variety of comprehensive signals derived from users' explicit and implicit feedbacks. This rating algorithm has been validated in metropolitan areas and trusted by users for accuracy and reliability in indicating the venue desirability (Yang and Sklar 2016).

Table 5.1 Description of data sources

Data sources	Count	Key fields	Others
Instagram check-ins	358,471	cid, uid, ctime, cloc	check-ins of 'food' topic; Nov 2014 – Nov2015
Foursquare POIs	32,485	vid, vcategory, vloc, vrating	

The spatial regions are needed for training RegNet, so DBSCAN method (Ester et al. 1996) is implemented on the Instagram check-ins for spatial clustering and the DBSCAN parameters Eps and MinPts are set as 7 and 18 respectively. 70% of the clustered regions are randomly selected as the training set and the remaining 30% as the test set for evaluation. For each region Rn_i, the desirability score $Rn_i.Rt$ is calculated with Equation 4.9.

For RegNet, 35 is set as the dimension of the encoded representation y (Equation 5.2), Rectified Linear Unit (ReLU) $f(x) = max\ (0, x)$ as the activation function of encoder, Softsign $f(x) = \frac{x}{1+|x|}$ as the activation function of predicting hidden layers.

5.4.2 Performance evaluation

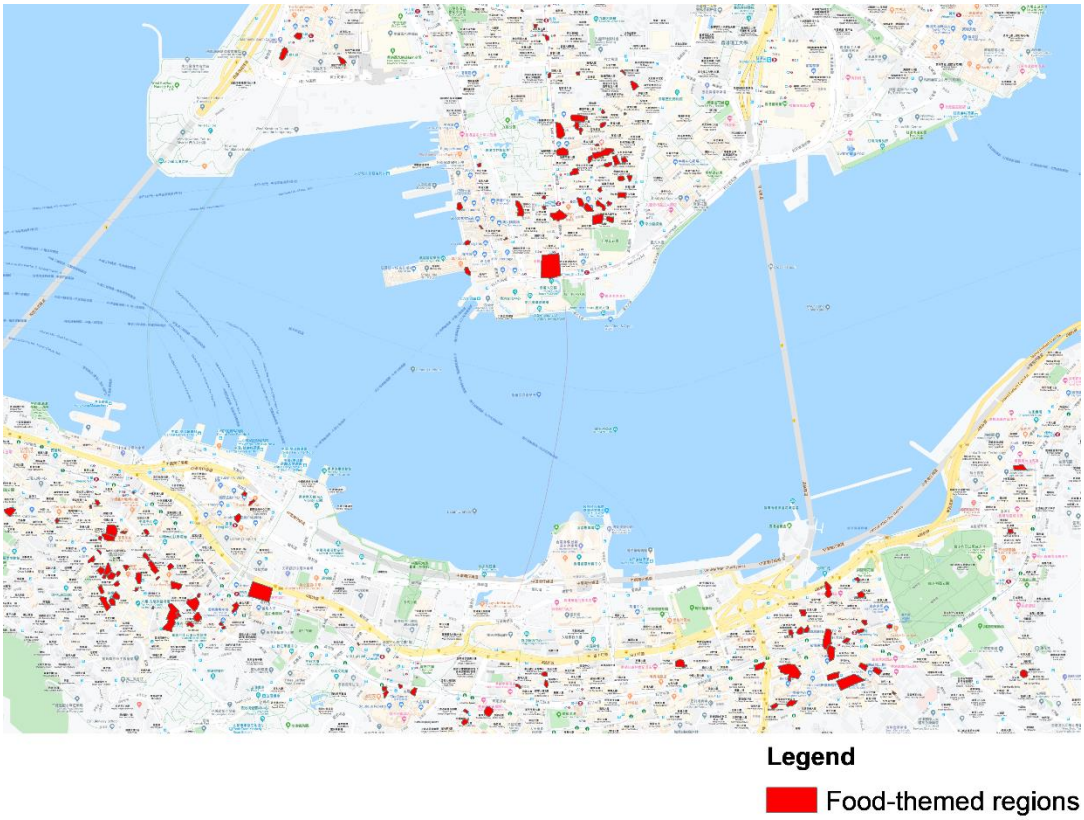


Figure 5.3 The clustered food-themed regions with Instagram check-ins across Hong Kong. The majority of the regions are located in Causeway Bay, Tsim Sha Tsui, Mong Kok and Central, all of which are the major recreation and amusement areas in Hong Kong.

The clustered regions (see Section 5.4.1) are shown in Figure 5.3. It can be seen that the clustered regions are mainly located in Causeway Bay, Tsim Sha Tsui, Mong Kok and Central. The results are consistent with Hong Kong urban situation, as all of these are the famous recreation areas, where quite a number of quality restaurants and street food are located.

For each clustered region, the desirability values are respectively calculated with the comparison methods and proposed RegNet, then compared with the ground-truth for evaluation. The

performance of the proposed RegNet is compared with the popular methods for regional desirability prediction: rank-by-users, rank-by-visits, HITS (Zheng and Xie 2011), HITS-based (Liu et al. 2019) (proposed in Chapter 4). For the rank-by-users, the regional desirability is proportional to the total count of visitors to the region. For the rank-by-visits, the regional desirability is proportional to the total count of check-ins into the region. The HITS (Zheng and Xie 2011) method is to predict regional desirability based on the mutual reinforcement relation between user's travel experience (hub score) and the desirability of a region (authority score). The HITS-based (Liu et al. 2019) method (proposed in Chapter 4) is a revised model of Zheng and Xie (2011)'s method, with consideration of regional scale and user weight decay.

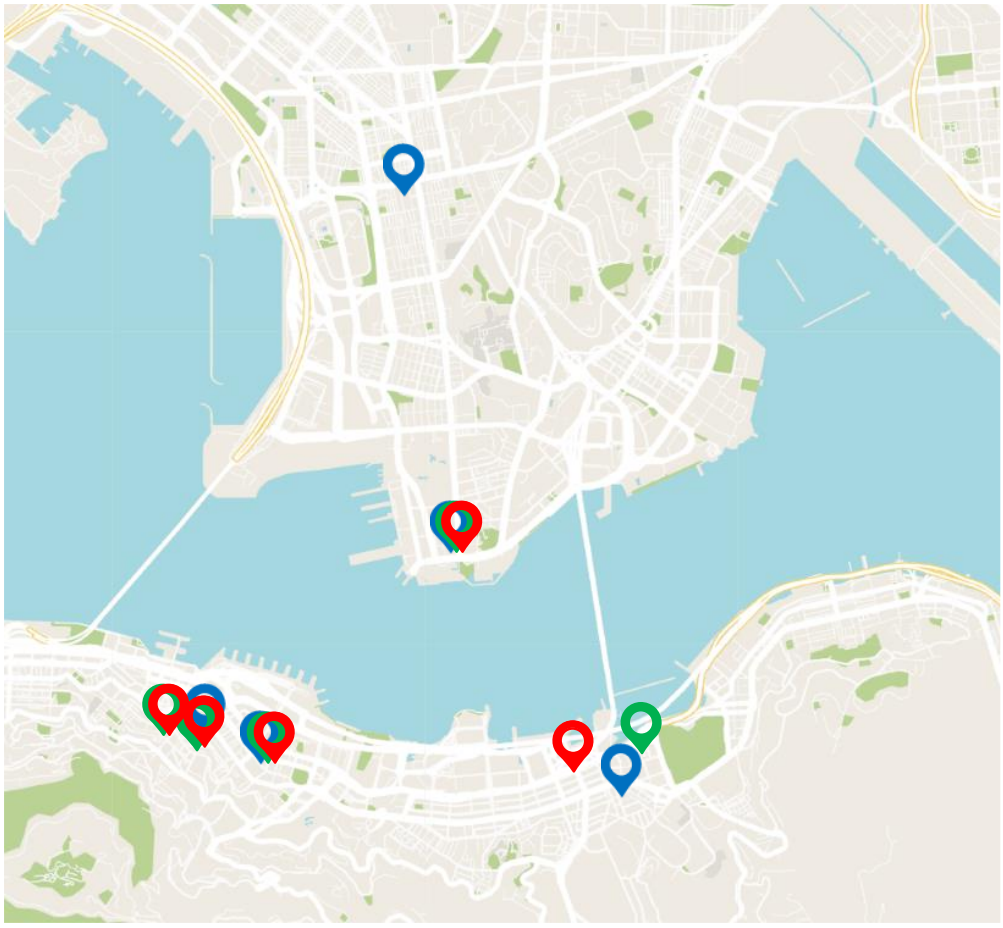


Figure 5.4 The geographical distribution of the top 5 regions with the highest desirability values respectively by HITS (Zheng and Xie 2011) (blue), HITS-based (Liu et al. 2019) (green) and proposed RegNet (red)

Top 5 regions by HITS (Zheng and Xie 2011):
1. Hennesy Road-East Point Center-Hysan Place
2. Portland Sreet-Langham Place
3. Landmark Atrium-Alexandra House
4. The peninsula
5. WinningHouse-Fortuna Building

Top 5 regions by HITS-based (Liu et al. 2019):
1. Landmark Atrium-Alexandra House
2. Aberdeen Street-Police Married Quarters
3. Towning Mansion
4. The peninsula
5. Elgin Sreet-Wai Yuen Building

Top 5 regions recommended by RegNet:
1. Landmark Atrium-Alexandra House
2. Aberdeen Street-Police Married Quarters
3. Elgin Sreet-Wai Yuen Building
4. Jaffe Road-City Mansion
5. The Peninsula

(1) (2) (3) (4) (5)

Figure 5.5 The details of the top 5 regions with the highest desirability values respectively by HITS (Zheng and Xie 2011), HITS-based (Liu et al. 2019) and proposed RegNet

The top 5 regions with the highest desirability values respectively by HITS (Zheng and Xie 2011), HITS-based (Liu et al. 2019) and proposed RegNet are geographically mapped in the Figure 5.4. The details of each regions are shown in Figure 5.5. It can be seen that regions such as Landmark Atrium-Alexandra House, Aberdeen Street-Police Married Quarters, the Peninsula are predicted with high desirability by both proposed RegNet and comparison methods.

The Table 5.2 lists several regions' normalized desirability value and ranking by each method. From Table 5.2, it can be seen the top 10 desirable regions predicted by RegNet have a good agreement with the ground truth. The region 11, 55, 75, 42 are raking at the 1st, 4th, 9th, and 10th place according to the ground truth, while predicted at 3rd, 7th, 9th, and 10th place by the RegNet.

4 out of 10 most desirable regions are predicted correctly with slight ranking inconsistency against ground truth.

Table 5.2 Normalized Regional Desirability Calculated by Each Method

Region ID	Ground truth		Rank-by-user		Rank-by-visit		HITS (Zheng and Xie 2011)		HITS-based (Liu et al. 2019)		RegNet	
	rating	rank	rating	rank	rating	rank	rating	rank	rating	rank	rating	rank
11	1.000	1	0.235	11	0.260	8	0.250	9	0.397	5	0.983	3
66	0.939	2	0.109	20	0.077	23	0.090	20	0.157	10	0.874	13
124	0.909	3	0.011	42	0.007	42	0.008	42	0.014	40	0.682	34
55	0.893	4	0.261	10	0.205	12	0.227	11	0.394	6	0.944	7
10	0.863	5	0.044	28	0.035	29	0.038	29	0.068	18	0.816	22
140	0.863	6	0.022	38	0.019	36	0.020	36	0.036	32	0.802	25
92	0.848	7	0.069	24	0.071	24	0.070	24	0.052	25	0.832	19
68	0.848	8	0.032	32	0.031	30	0.031	33	0.055	24	0.857	16
75	0.840	9	0.208	13	0.209	11	0.208	12	0.174	9	0.924	9

42	0.832	10	0.091	22	0.084	20	0.086	21	0.149	11	0.921	10
141	0.832	11	0.042	29	0.031	31	0.035	30	0.061	20	0.820	20
122	0.832	12	0.040	30	0.028	33	0.033	31	0.058	22	0.815	23
35	0.825	13	0.511	5	0.499	4	0.503	4	0.420	4	0.954	5
105	0.817	14	0.000	45	0.000	45	0.000	45	0.000	45	0.508	42

Different values are assigned to p to calculate $nDCG_p$ (Equation 4.6), as shown in Table 5.3. The results show that, as p increments, the $nDCG_p$ for both RegNet and comparison methods increases. With different p values, RegNet can stably achieve higher *nDCGs* (0.33-0.49) than the *nDCGs* of other comparison methods.

From user's perspective, it can be expected the users care more about being recommended highly desirable regions. Particularly, with p=5 and 15, RegNet achieves *nDCG*=0.329 and 0.438, which are consistently higher than comparison methods (i.e., 0.255, 0.375), indicating, for the highly desirable regions, the proposed RegNet has better predictions than other methods. From Table 5.2, it can be seen the top 10 desirable regions predicted by RegNet has a good agreement with the ground truth. 4 out of 10 most desirable regions are predicted correctly with slight ranking inconsistency against ground truth. The above results show that RegNet can achieve better ranking performance than comparison methods and demonstrate RegNet's advantage in predicting highly desirable regions.

Table 5.3 Normalized Discounted Cumulative Gain for RegNet and comparison methods

$nDCG_p$	RegNet	Rank-by-users	Rank-by-visits	HITS (Zheng and Xie 2011)	HITS-based (Liu et al. 2019)
$nDCG_5$	0.329	0.000	0.000	0.000	0.255
$nDCG_{15}$	0.438	0.205	0.228	0.219	0.375
$nDCG_{25}$	0.456	0.281	0.300	0.294	0.384
$nDCG_{35}$	0.487	0.287	0.306	0.301	0.392
$nDCG_{45}$	0.487	0.325	0.344	0.338	0.422

The agreement between the methods' ratings and user rating (normalized to be comparable) is further measured with the MAE (Equation 4.5), shown in the Table 5.4. The user count is used as the desirability score for rank-by-users method and visit count for rank-by-visits method. The results show RegNet can achieve less MAE (0.19) than the comparison methods (around 0.53), indicating that the regional desirability ratings predicted by RegNet have better agreement and less deviation with the ground truth than the comparison methods.

The RegNet is data-driven and, given enough training data, can model the hidden unknown interactions of high-level features and approximate the unknown mapping functions from input feature to output value. In the experiment, it shows the RegNet has better performance than the traditional empirical methods, in terms of both ranking quality and numerical rating agreement.

Table 5.4 MAE (mean absolute error) for each method

	RegNet	Rank-by-users	Rank-by-visits	HITS (Zheng and Xie 2011)	HITS-based (Liu et al. 2019)
MAE	0.193	0.528	0.534	0.531	0.532

5.4.3 Parameter sensitivity analysis

Region recommendation is a newly emerging research area with the recent rise of GSMD. Traditional region recommendation methods are mainly empirical models to predict regional desirability, which can be relatively intuitive and inaccurate. The underlying reasons are that the internal mechanism of the regional desirability is insufficiently known and a sound model to simulate the interaction between regional desirability and users' check-ins is still missing. On the other hand, the massive volume of GSMD provides an alternative solution. The consideration is that the mechanism of regional desirability has already be implicitly contained by GSMD, and, with proper dataset and data-driven model, the hidden patterns can be well learned without priori knowledge. Consequently, in this chapter, a new neural network model RegNet are proposed for predicting regional desirability.

The RegNet consists of an encoder structure for feature learning and several hidden layers for desirability prediction. Theoretically, the neural network has been proven to be capable of universally approximating functions (Hornik et al. 1989) and widely used for classification and regression. The user visiting history and rating are fed into the RegNet as training data and the hidden interactions of high-level features are learnt accordingly. The experiments demonstrate RegNet's advantage over traditional empirical models in predicting regional desirability and,

especially, recommending highly desirable regions.

A worthy question is how the encoder structure affects the method performance. The consideration is that the encoder is a lossy data compression, which means the feature information can be lost together with the data redundancy by dimensionality reduction with encoder. Inappropriate encoder settings may cause intense information loss and the model failing to learn the hidden patterns. Consequently, comparisons are made on how the dimension of the encoded representation (y in the Equation 5.2) affects the method performance (shown in Figure 5.6). It shows that the MAEs achieved by RegNet remain relatively stable (around 0.20) with increasing encoded dimension. The MAE achieves its minimum when dimension is 35 and increases fluctuantly as dimension continues to grow. An explanation can be made by looking into the structure of RegNet. The RegNet consists of two parts of neural network: feature learning with encoder layers, regional desirability prediction with hidden layers. RegNet will fail to learn the interaction of features with too few encoded dimensions, as hidden patterns can be lost with intense dimensionality reduction, causing the encoded feature y incapable of well representing the original input vector x (Equation 5.2), which leads to underfitting problem. On the other hand, the interaction of hidden layers in the regional desirability prediction phase can be over complicated, with large encoded dimension and limited training dataset, causing RegNet performing well with training dataset and poorly with evaluation dataset, which is overfitting problem. A good performance can only be achieved with a proper medium size of encoded dimension.

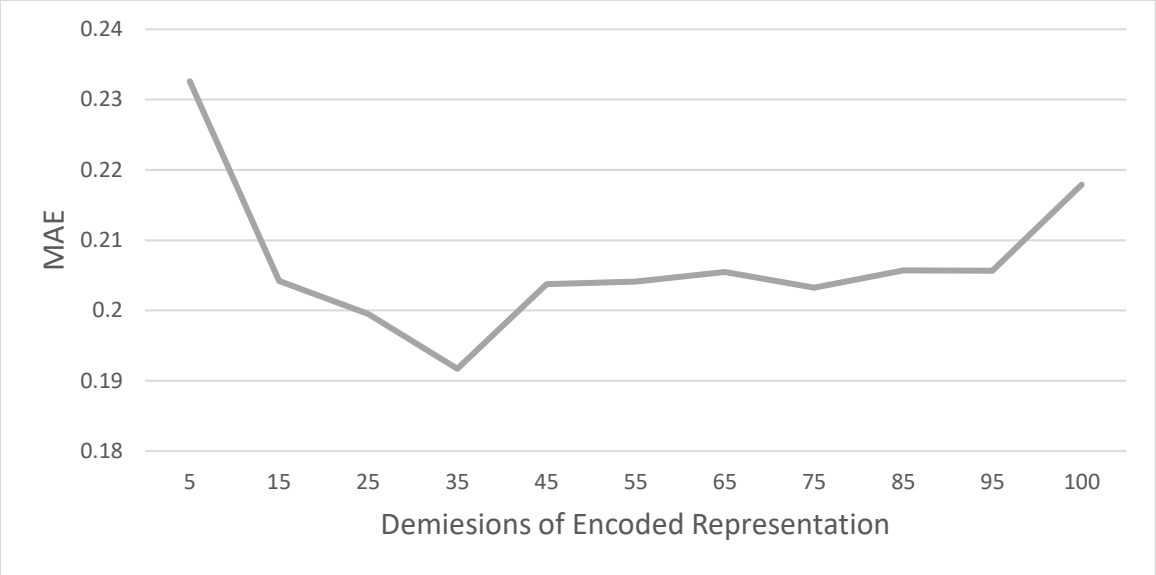


Figure 5.6 Impacts of the Encoded Dimension

5.5 Discussion

In this section, some worthy questions are further discussed to provide insights into the proposed methodology: (1) what is the intuition of proposed RegNet? Why the RegNet can address the previous gap? (2) Is RegNet a relatively standard neural network framework? If so, what is the essential novelty of RegNet? (3) Since the ground truth of regional desirability has already be obtained from some platforms (e.g., Foursquare), what's the point to calculate regional desirability with GMSD again?

5.5.1 Intuition of proposed RegNet

The motivation for this Chapter 5 is to develop a new neural network model which can achieve better performance over previous methods for predicting desirability scores of regions, as the hidden high-level features and unknown interactions between user check-ins and regional desirability can be learnt by the neural network model.

There are some previous works focusing on identifying and ranking desirable regions and this work is never the first attempt to do that. However, the consideration is that, compared with the massive works on POI recommenders, identifying and recommending desirable regions haven't been adequately investigated as some key problems remain unaddressed. One gap is that when calculating desirability scores for regions, previous works mainly use empirical models, which can be inaccurate and bring in perceptual bias. By empirical models, it means that the presumed relationships and empirical formula, such as power law distribution (Sun et al. 2015) and Hypertext Induced Topic Search (Zheng et al. 2009, Zheng and Xie 2011) are used to predict the regional desirability from user check-in history. These empirical models are developed based on human intuition and past experiences and can be inaccurate as the presumed mathematical formulas are used for prediction. The underlying reason for using empirical models is that the hidden interactions between user check-ins and regional desirability haven't be clearly quantitatively explained or precisely modeled yet, so that the empirical models with presumed relationship are used for rough approximation.

Consequently, after analyzing the inaccuracy of previous models and the underlying reason, a new neural network model is proposed to address the above challenge. An advantage of neural networks is that it is data-driven, which means, given certain training data, the neural networks are theoretically capable of universally approximating functions (Hornik et al. 1989) and learning the hidden unknown interactions of high-level features (LeCun et al. 2015) without prior knowledge, which is suitable for the research problem where the hidden interactive mechanisms of relevant features are still quantitatively unknown. So, in this work, a novel neural network model ('RegNet') is proposed for predicting regional desirability. RegNet can adaptively learn the hidden interactions of high-level features and the unknown mappings from input to output,

without prior knowledge. The performance of the proposed RegNet is evaluated and RegNet's better performance is demonstrated over traditional methods, i.e. rank-by-users, rank-by-visits, HITS (Zheng and Xie 2011), HITS-based (Liu et al. 2019) in terms of both ranking order and deviation from the ground-truth rating.

Overall, the intuition of proposed RegNet is that a new neural network model can better learn the hidden interactions of high-level features and the unknown mappings from input to output, without presumed relationships and empirical formula, thus reducing perceptual biases and achieving better performance over previous methods for predicting desirability scores of regions. Compared with previous empirical models, RegNet can achieve desirable region predictions with better accuracy, which have been demonstrated by the experiments with real-world datasets.

5.5.2 Novelty of proposed RegNet

The novelty of the proposed RegNet lies in the encoding-prediction structure of RegNet, which is a novel neural network framework and initially developed in this thesis. In this newly developed RegNet, the value of embedded parameters in the RegNet can be tuned during model training, while the developed encoding-prediction framework remains constant.

As explained in Section 5.2, RegNet are made up of two main components: a network encoder for feature learning, and a hidden-layer structure for prediction. (1) The encoder structure is trained from an autoencoder structure and capable of learning feature representation with reduced dimension, thus alleviating potential overfitting and computational complexity while preserving as much information as the original input; (2) The hidden-layer structure can learn the hidden unknown features and interactions through backpropagation, thus achieving better performance over previous methods.

This encoding-prediction framework of RegNet is a fixed and novel structure, which is initially developed in this work, while the value of embedded parameters in the network can be adjusted. The hidden-layer structure of RegNet is a network component with changeable settings such as the node count in layers, activation function, the optimizer, loss function. These parameters can be tuned in the training process. However, the encoding-prediction structure of RegNet remains constant.

This encoding-prediction network structure hasn't ever been applied or proposed in the field of calculating regional desirability. Previous works mainly use empirical models with presumed relationships, which brings in perceptual bias. While the proposed RegNet can adaptively learn the unknown interactions of high-level features without prior knowledge, which is suitable for the research problem where the hidden interactive mechanisms of relevant features are still quantitatively unknown.

5.5.3 Meaningfulness of calculating regional desirability with GMSD

The meaningfulness to model regional desirability from GMSD can be illustrated from two aspects:

(1) From a practical perspective, modelling regional desirability from social media check-ins can be applied into the situation where Foursquare POIs are not available or effective. Firstly, generating ground-truth from Foursquare are dependent on the availability of Foursquare data, which is highly uncertain and subject to the company strategies. The fact is, the POIs from Foursquare API were collected for experiments previously, however this kind of POIs is no longer available anymore as the Foursquare company has changed the settings of Open APIs, which makes collecting complete POIs datasets in a specific region impossible. Although the

Foursquare search-for-venues API is still open, only partial venues within a spatial range can be returned and the complete datasets are not publicly retrievable anymore. Only by building academic or business connection with the company, the complete POIs in a spatial range can be potentially obtained, making it difficult to generate ground-truth regional desirability from Foursquare data. Secondly, the Foursquare platform is not ubiquitously reliable across all the countries. Previous study has proven the accuracy and trustworthiness of Foursquare rating algorithm in indicating the place desirability, in metropolitan areas (Yang and Sklar 2016). However, in regions where Foursquare is unpopular or even rarely used (e.g, China Mainland, Taiwan), the plausibility of the Foursquare rating system is questionable, as the scale of crowdsourced information is limited. Consequently, if the relationship between regional desirability and check-ins can be well modeled, the developed models can be used into the above situations, predicting regional desirability with the check-ins data, when the Foursquare data is not available or effective.

(2) Besides, from a theoretical perspective, modelling the mutual relationship of regional desirability and social media check-ins can contribute to the knowledge discovery and help to answer worthy questions. Even the accurate regional desirability values can be obtained via some means, the connections between regional desirability and social media check-ins are still unknown. Is the relationship between check-ins to regional desirability quantitatively explainable? Is there an accurate depicting formula? If the relationships are still mathematically unknow, is it possible to build a black-box computation model from one side to another? These are the questions that are worth considering and can deepen our understanding about the known facts. This research can serve as a step into investigating these questions and adding more knowledge into the relationships between check-ins and regional desirability.

To sum up, from a practical perspective, modelling regional desirability from social media check-ins can be used into the situation where platform data (e.g., Foursquare POIs) is not available or effective, expanding the application scenarios of social media. Moreover, from a theoretical perspective, modelling the mutual interactions from one side to another can contribute to the knowledge discovery and add novel knowledge into the connections between check-ins and regional desirability.

5.6 Summary

Chapter 5 focuses on addressing the flowing gap: how to develop an accurate model to calculate regional desirability when the interactions of relevant features are still unknown.

Traditional methods mainly use empirical models to predict regional desirability, which are quite intuitive and can be inaccurate in predicting desirable regions, as potential perceptual bias is introduced. This results from the fact that the hidden interactions between user check-ins and regional desirability haven't be clearly explained and precisely modelled yet, so that the empirical models based on human empirical intuition and presumed relationship are used for rough approximation.

Consequently, a new multi-layer neural network model (RegNet) is proposed to address the aforementioned problems. The neural networks have been proven capable of universally approximating functions and adaptively learning the hidden unknown interactions of high-level features without prior knowledge. RegNet includes a neural network encoder structure for feature learning and a hidden-layer structure for desirability prediction. Pairs of user check-in history and ground-truth regional desirability rating are fed into RegNet as training data and regional desirability scores are calculated as the predicted values of the output layer.

The proposed RegNet is implemented with real-world datasets and the experiments prove the better performance of RegNet over baseline methods. Besides, sensitivity analysis is also made on how the dimension of the encoded representation affects the performance of the proposed model and it is found that only with a proper medium size of encoded dimension can a good performance be achieved, due to potential underfitting and overfitting issues. This research demonstrates the feasibility and effectiveness of data-driven methods (e.g., neural networks) in modelling the hidden unknow relationships and achieving a better performance for region desirability prediction over traditional empirical methods.

Chapter 6 Spatiotemporal semantic modelling: a model for event detection by finding spatiotemporal irregularities

6.1 Introduction

In this chapter, the following question is investigated: what is happening at some place and at some time, inferred from GSMD? Currently, most research on event detection with social media uses semantic-based methods and text mining. Platforms, such as Twitter, provide accessible streaming data, which contains real-time text information, making it possible to detect emerging social events by examining the change in semantic-based features.

However, a new event detection methodology can be proposed by investigating the geographical patterns of GSMD. The intuition is that a social event will affect how the objects spatially distribute across certain regions and how they mutually interact thus causing irregular geographical patterns, especially irregular human mobility and interaction patterns (e.g., sports games causing intense human aggregation or terrorism attacks causing sudden evacuation from certain regions). By introducing depictive measuring features with GSMD and identifying the feature irregularity, such geographical patterns, in turn, can be used to distinguish potential social events. Previous works mainly detected social events by word frequency and semantic analysis, and there is still a need for a comprehensive methodology to effectively detect events by investigating irregular geographic patterns.

Consequently, a major motivation for this chapter is to provide new insights into the following issues: what kinds of features can be used for detecting events, or from what aspects can a social event be differentiated with GSMD. Specifically, the novelty and point of this chapter is to detect social events by mining the geographical patterns of GSMD and using geography-based features.

To do so, a new model is proposed to construct event features by characterizing spatial patterns of GSMD. Social events are then detected by finding the feature irregularities. The detailed analytic workflow consists of three parts: (1) semantic community discovery, (2) geography-based event representation, and (3) irregular event feature detection. First, the data-driven geographic topic modelling method described in Chapter 3 is used to detect hashtag communities and identify social media topics. Second, by introducing quantitative spatial autocorrelation indicators, event features are constructed for representing the potential events in the created feature space. A time series of the univariate event feature is then generated with variable temporal granularities. Third, an outlier test is adopted to detect event feature irregularity, and the event location are identified. A detailed event description can then be obtained using the detected semantic and spatiotemporal identification. The experiment is conducted with a real-world dataset (104,000 geotagged Instagram check-ins) and the effectiveness of the proposed workflow and model is verified.

6.2 Methodology

In this section, the data model and analytic workflow of the proposed workflow are presented, followed by a detailed description of each data handling procedure. First, the topic modelling method is mentioned, which is the same as discussed in Section 4.2. Next, a novel event feature representation method is explained, based on the geographical patterns of the social media post distribution. Finally, the method for irregularity detection and location indication is described, based on the data model.

6.2.1 Analytic workflow

To define an event, the question needs to be answered: **what** is happening at that **place** and that **time**? Consequently, the semantic and spatiotemporal information is needed for event identification.

Specifically, in the proposed model, to detect social events, the following issues are considered:

- What are the social events about? In other words, semantically, how can we identify the potential topics of the social events using social media data?
- How should the social event be depicted from a spatiotemporal perspective? What is the event location and time? Is there any irregular geographical pattern (e.g., crowd aggregation, evacuation) caused by the event, and how can such patterns be represented with the geotagged social media posts?

Based on above considerations, the event detection model can be defined as follows:

$$E = W(S, G, T) \tag{6.1}$$

where E is the event detection result, S is the semantic identification to indicate the event topics, G is the geographical patterns of human mobility revealed by GSMD, T is the temporal identification of the event. and W is the analytic workflow. The underlying intuition behind the proposed workflow is that a social event may affect regular human mobility and interaction patterns. By introducing features derived from the quantitative measurements of such patterns, a new feature space can be created, and the social event can be represented and detected with the newly created features.

As represented in Equation 6.1, the semantic and geographical patterns are both considered in the model. Thus, to address the semantic and geographical issues, a threefold analytic workflow (Figure 6.1) is constructed, which consists of three interrelated modules: (1) semantic community discovery; (2) geography-based event representation; and (3) detection of event feature irregularity.

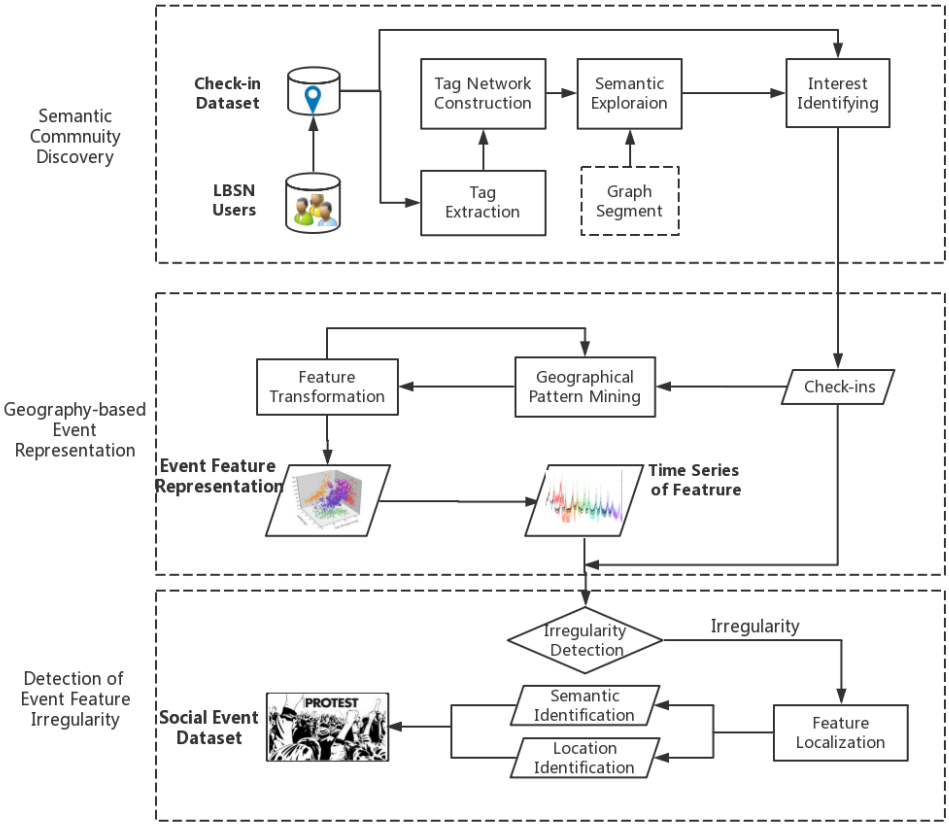


Figure 6.1 Proposed workflow of event detection by finding spatiotemporal irregularities

Semantic Community Discovery: In this step, the semantic content attached to the GSMD is investigated first to infer the user activity associated with the geotagged posts (e.g., geotagged photos). This step starts by retrieving geotagged media data from a relevant API (check-in dataset). Furthermore, the method described in Section 4.2 is adopted to detect semantic

communities of the hasthags. The topics of geotagged photos are further identified by introducing the topics of attached hashtags as an indicator, as shown in Figure 6.2 (Interest Identifying).

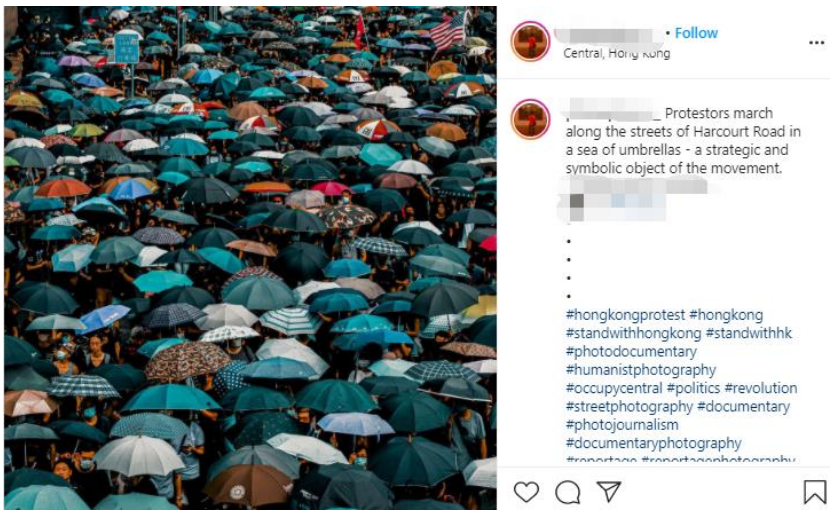


Figure 6.2 Photo sample from Instagram. The photo has a title with multiple hashtags (e.g., #occupycentral, #umbrellarevulotion, #documentary) to annotate, indicating potential topics for the photo. (source: https://www.instagram.com/p/B5-iuDHp2HN/)

Geography-based Event Representation: In this step, a new event feature representation method is developed. By introducing quantitative spatial autocorrelation indicators (Geographical Pattern Mining), a global index is constructed for representing the potential event in the created feature space (Feature Transformation & Event Feature Representation). A time series of the univariate event feature with different temporal granularity can then be yielded, by adjusting the temporal range parameter.

Detection of Event Feature Irregularity: After mapping the geographical patterns into a new feature space, in this step, an outlier test is adopted to detect the feature irregularities. The global

index is then localized for identifying event location. By combing the semantic and location identification, a detailed description of the detected event can be obtained.

The methods used for semantic community discovery is described in Section 4.2. The details of other modules are presented below.

6.2.2 Proposed model: event feature modelling by spatial patterns mining

After identifying the semantic topics of the geotagged posts, the next step is to construct geography-based features for event representation. The consideration is that a social event may lead to irregular human mobility (e.g., crowd aggregation), and such irregularity can be revealed by the geographical pattern of GSMD. Consequently, a geography-based event representation method is proposed. The spatial autocorrelation is introduced as a geographical description of the human mobility and Moran's I (Moran 1950) as the quantitative measurement. Given a set of features and an associated attribute, Moran's I and z-score can be calculated to evaluate whether and to what degree the pattern expressed is clustered.

To analyze the spatial autocorrelation, a spatial fishnet is created first and the geotagged posts are then mapped into the corresponding fishnet cells, according to the post locations. Let T be the collection $\boldsymbol{T} = \{t_1, \dots, t_m\}$ of continuous time segments t_i, and C be the collection $\boldsymbol{C} = \{c_1, \dots, c_n\}$ of the fishnet cell $c_{\mathrm{i}} = (r, num(t_k))$, where r_i is the spatial region of c_{i}, and $num(t_k)$ is the number of the posts pst_j, where $pst_j.v \in c_{\mathrm{i}}.r$ AND $pst_j.t \in t_k$ AND $pst_j.tgs$ is of specific topics.

The Global Moran's I based on the post number can be further calculated as below (Anselin 1995):

$$I_{t_k} = \frac{\mathrm{n}}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2} \quad 6.2$$

where z_i is the deviation of $c_{\mathrm{i}}.num(t_k)$ from its average value ($c_{\mathrm{i}}.num(t_k) - \frac{\sum_{j=1}^{n} c_{\mathrm{j}}.num(t_k)}{n}$), w_{ij} is the spatial weight between $c_{\mathrm{i}}.r$ and $c_{\mathrm{j}}.r$, and S_0 is the summation of all the spatial weights.

The value of Moran's I ranges from [-1,1]. The more similar values cluster, the closer *I* is to +1; the more dissimilar values cluster (similar values disperse), the closer *I* is to -1; Moran's I being +1 indicates perfect clustering of similar values, while Moran's I being -1 indicates perfect dispersion. To construct the event feature, a global event index (GEI) is developed as below:

$$GEI_{t_k} = -c \cdot ln\left(1 - \left| I_{t_k} \right|\right) \quad 6.3$$

where c is a scale coefficient for normalization. By using Equation 6.3, the GEI value will acceleratedly increase as I_{t_k} approaches +1 (-1), indicating the transition from random to geographical clustering (dispersion).

By sequentially sampling GEI in the $T = \{t_1, \dots, t_m\}$, a discrete time series will be yielded:

$$TS = \{GEI_{t_1}, GEI_{t_2}, \dots, GEI_{t_m}\} \quad 6.4$$

where each item is a transformed feature for event representation in the created one-dimensional feature space. Different temporal granularity for analysis can be obtained, by adjusting the span of time segment t_i.

6.2.3 Event detection by finding feature irregularities in time series and spatial context

In this module, the aim is to detect irregularity of event features and pinpoint the spatiotemporal identification of social event. Thus, for irregularity detection, the generalized extreme

studentized deviate (ESD) test (Rosner 1983) is adopted, which is a statistical test to detect one or more outliers in a univariate data set. The time series TS (see Equation 6.4) will be the input data set for test and the outliers of the TS will be found accordingly. The ESD test is explained in detail in Algorithm 2.

$RFunc(X)$ finds the index k that maximizes the value of $R_k = \frac{|x_k - \bar{x}|}{s}$, where x_k, $\bar{x}$, s denote the k^{th} item, mean value and sample standard deviation of dataset X, respectively. The $Lambda(i)$ are calculated as follows (Rosner 1983):

$$Lambda(i) = \frac{(n-i)t_{p,n-i-1}}{\sqrt{(n-i-1+t^2_{p,n-i-1})(n-i+1)}} \quad 6.5$$

where n is the item count of dataset X, $t_{p,v}$ is the 100p percentage point from the t distribution with v degrees of freedom and $p = 1 - \frac{\alpha}{2(n-i+1)}$, with α being the significance level for ESD test.

Apart from temporally detecting the happening of an event with the above ESD test, another important issue is where the event happened. To further identify the event location, the GEI is localized to construct a local event index (LEI), as below:

$$LEI_{t_k}(c_\mathrm{i}) = -ln\left(1 - c\left|I_{t_k}(c_\mathrm{i})\right|\right) \quad 6.6$$

where $LEI_{t_k}(c_\mathrm{i})$ is the calculated LEI for fishnet cell c_i in the time segment t_k, $I_{t_k}(c_\mathrm{i})$ is c_i's Local Moran's I (Anselin 1995) calculated from $c_\mathrm{i}.num_i(t_k)$ being the attribute value, and c is a scale coefficient for normalization. The fishnet cell with high LEI indicates particular geographical clustering or dispersion patterns of human mobility, which may be caused by special social events in that region.

By adopting the workflow described above, a potential social event can be detected, with the identification of semantic topic and the depiction of underlying spatiotemporal patterns.

Algorithm 2: Generalized Extreme Studentized Deviate Test

Input: (1) Observation dataset X (2) Upper Bound Count r

Output: A list of detected outliers L

Begin

for $i = 1$ **to** r **do**

$(R_k, k) \leftarrow RFunc(X)$

$\lambda_i \leftarrow Lambda(i)$

if $R_k > \lambda_i$ **do**

$L.\text{add}(X.\text{item}(k))$ // The k^{th} item in X is added to L

end

X ← X.item_remove(k) // Remove the k^{th} item from X

end

Return L // Return the list of detected outliers

End

6.3 Experiment

6.3.1 Datasets and settings

Similar to previous chapters, the check-in dataset used in this chapter is retrieved from Instagram API, which covered 127,630 geotagged check-ins in Hong Kong from Dec 6, 2014 to Jan 4, 2015, generated by 26,420 users. A social media post includes the post ID, user ID, the attached

hashtags, the timestamp of the online post and the location indicating post place (Table 6.1). All the user IDs are anonymized for privacy protection.

Table 6.1 Description of geotagged social media posts

Fields	Description
pid	A string uniquely indicating a post
uid	An encrypted string uniquely indicating a user
hashtags	The attached hashtags indicating the potential post topics
stime	The time that the user publishes the online post
location	The post location

Some of the user accounts are for commercial advertising and might post redundant photos in specific venues. Consequently, an anomaly detection method is conducted to remove these outliers. The count of photos posted by each user is first investigated. The calculations shows that the average count of posted photos per user μ is 4.8 and the standard deviation σ is 8.8. The three-sigma rule is introduced, which states that, for both normally distributed and non-normally distributed variables, most cases should fall within three-sigma intervals. Therefore, in this section, three-sigma intervals are used to set the threshold for filtering. The users whose photo numbers are out of the three-sigma intervals (i.e., $\mu + 3\sigma$, 32) are recognized as outliers and their posted photos are removed from the photo dataset. After data cleaning, 104,366 photos generated by 25,996 users remained. For the parameter setting, the span of time segment t_i is set as 24 hours and the size of fishnet cell c_{i} is set as 100*100 m.

6.3.2 Case study

The experimental results are reported and the performance of the proposed workflow is evaluated accordingly. Evaluating event detection techniques is a very challenging and complex problem, and there is no commonly accepted standard yet (Weiler, Grossniklaus and Scholl 2017). Theoretically, the results of event detection can be evaluated by quantitative and qualitative analysis. In practice, quantitative evaluation commonly requires a thorough ground truth data set, which is always unavailable and thus undermines the effectiveness and even feasibility of this evaluation approach. Therefore, in the experiment, the qualitative evaluation is adopted.

After completing the semantic community discovery, several hashtag groups are detected, each of which shares a common topic. To demonstrate the usefulness of the proposed workflow, two hashtag groups are chosen for case studies.

It is found one hashtag community whose semantic topic is mainly about the umbrella movement (See Figure 6.3). The umbrella movement is a political movement that emerged during the Hong Kong democracy protests of 2014. This movement lasted for 79 days, and many downtown streets are occupied during the event, which ended up in a police clearance operation. By studying the spatiotemporal patterns of the corresponding geotagged check-ins, the event can be investigated from a new perspective.

Figure 6.3 Word cloud for representative hashtags of the 'Umbrella Movement' community

The GEI is calculated with the geotagged check-ins that are identified as the 'Umbrella Movement' topic. Table 6.2 lists the distribution of time series TS of GEI. Each GEI in time series TS represents the corresponding event feature for that time segment (i.e., that day). By implementing ESD test upon the TS, the irregular event feature is detected. The visualized results are shown in Figure 6.4. It can be seen that, at the beginning, the GEI value increases in an oscillatory manner and achieves its maximum as 37.43 at Dec 11, 2014, which is detected as the irregularity of the event features by the ESD method, and then decreases to its normal level after that.

Table 6.2 Time series of GEI with geotagged check-ins identified as the 'Umbrella Movement'

Date	**Dec 6**	**Dec 7**	**Dec 8**	**Dec 9**	**Dec 10**	**Dec 11**	**Dec 12**	**…**	**Jan 4**
GEI	3.13	3.72	1.48	13.18	8.13	37.43	6.48	…	0.61

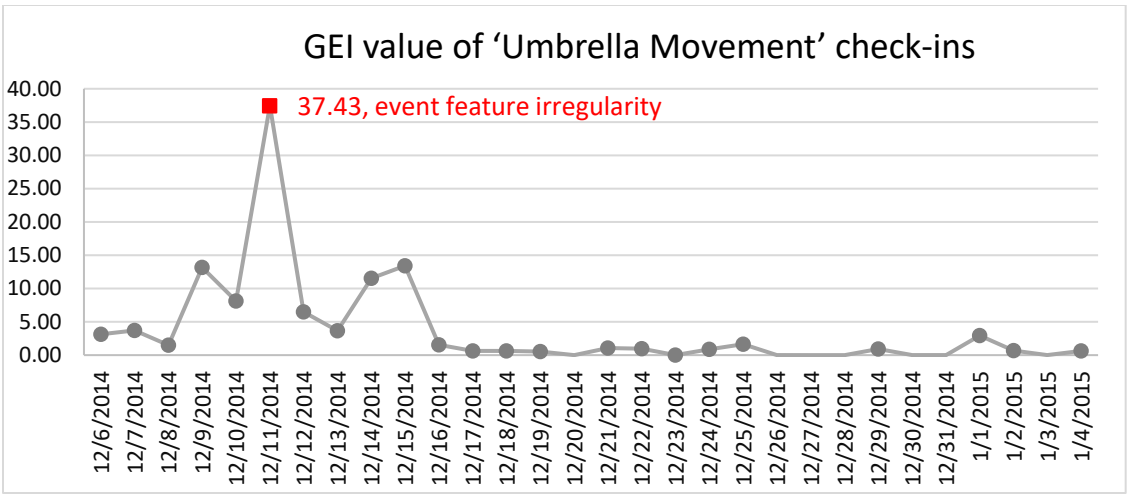


Figure 6.4 Detected feature irregularity as the 'Umbrella Movement' event on Dec 11, 2014

Moreover, to identify the potential event location, the LEI (see Equation 6.6) is further calculated for Dec 11, 2014. The spatial distribution of LEI is shown in Figure 6.5. It is found that the regions with a high LEI value are mainly around Admiralty, where the key governmental departments (i.e., the Hong Kong Central Government Offices, Legislative Council Complex and the High Court of Special Administrative Region) are located, and the place with highest LEI is at [lat: 22.28, lon: 114.17], where the LEI are achieved at 125.46, which, according to Equation 6.6, indicates significant human aggregation in that region and the regions nearby.

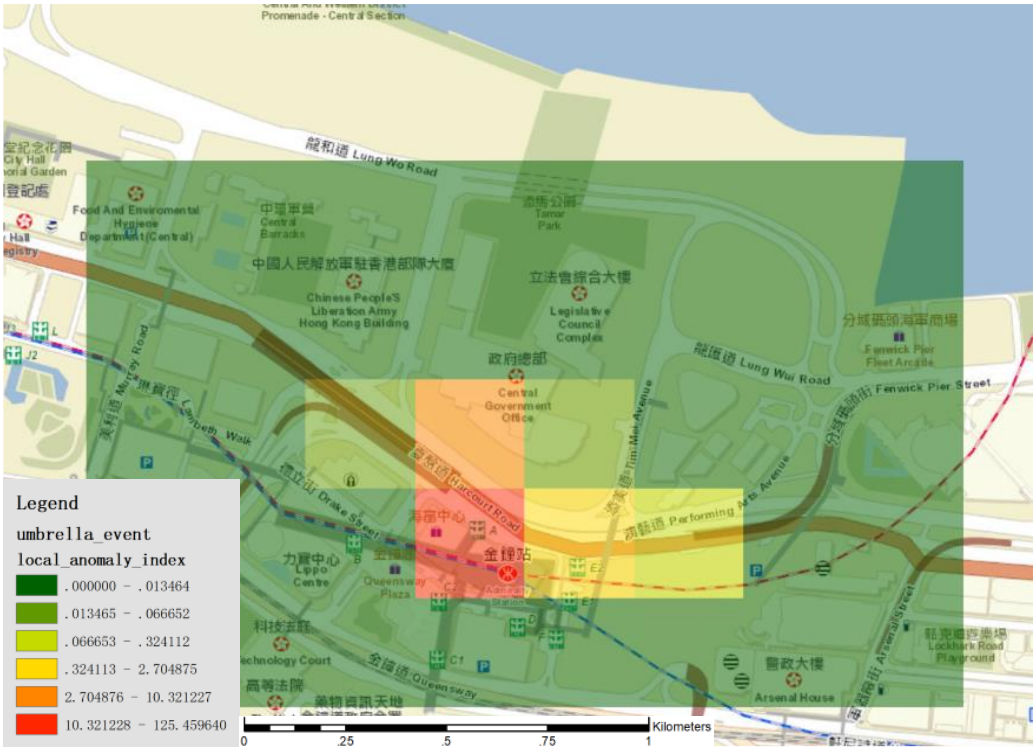


Figure 6.5 Admiralty region with the high *LEI* value of 'Umbrella Movement' check-ins, indicating significant human aggregation in that region and the regions nearby

To determine the underlying reason causing this human movement pattern, the background news is searched with reference to the semantic identification 'Umbrella Movement' and detected spatiotemporal pattern 'Admiralty, Dec 11, 2014' (see Equation 6.1). By researching the local news, the relevant event, i.e., Admiralty Site Clearance, is detected (see Figure 6.6). On Thursday Dec 11, 2014, the Hong Kong police executed the first site-clearance operation to end the main sit-in of the 'Umbrella Movement' in Admiralty with the arrest of 247 people. The police released the announcement about the site-clearance before the operation. Therefore, many citizens came to the Admiralty site to witness this social event on that day, causing significant human aggregation in Admiralty, which consequently is indicated by the *LEI* spike in the corresponding regions (Figure 6.5).

Figure 6.6 Site-clearance operation by Hong Kong police in Admiralty at Dec 11, 2014 (source: https://www.voachinese.com/a/central-occupy-hk/2554501.html)

To further demonstrate the effectiveness of the proposed workflow, another case study is given. It is found another hashtag community whose semantic topic is mainly about fitness. The word cloud for the hashtags of that group is shown in Figure 6.7.

Figure 6.7 The word cloud for representative hashtags of 'fitness' community

GEI is calculated with the geotagged check-ins identified as the 'fitness' topic (see Table 6.3) and the event feature irregularity is detected with the ESD test (see Figure 6.8). The event irregularity is detected at Dec 7, 2014, when the GEI value achieves its maximum as 32.47.

Table 6.3 Time series of GEI with geotagged check-ins identified as 'Fitness'

Date	**Dec 6**	**Dec 7**	**Dec 8**	**Dec 9**	**Dec 10**	**Dec 11**	**Dec 12**	**…**	**Jan 4**
GEI	1.20	32.47	1.77	0.86	0.34	0.35	0.76	…	0.01

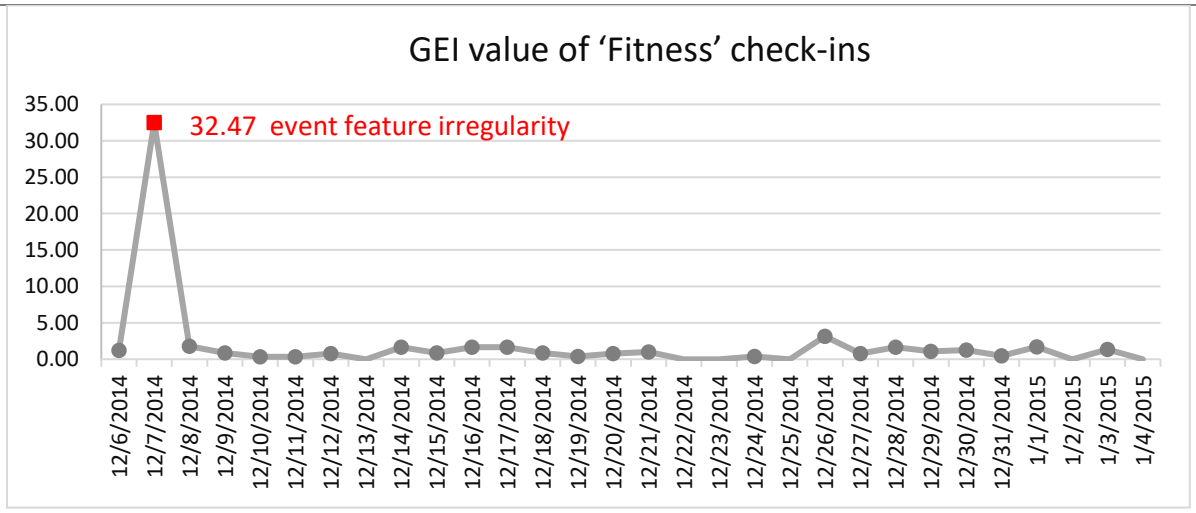


Figure 6.8 The detected feature irregularity of the ‘fitness’ event on Dec 11, 2014

The location information is needed to pinpoint the social event; therefore, the LEI is further calculated. Figure 6.9 shows the spatial distribution of LEI calculated from the ‘fitness’ geotagged check-ins on Dec 7, 2014. A finding is that some areas in Chek Lap Kok had unusually high LEI, compared with other regions. The area with the highest LEI is at [lat: 22.32 lon: 113.94], where the LEI are achieved at 189.23. This area is actually the AsiaWorld-Expo, which has the biggest purpose-built indoor-seated entertainment arena in Hong Kong.

Figure 6.9 The Chek Lap Kok region with the high LEI value of ‘fitness’ check-ins, indicating significant human aggregation

The underlying event related to this irregularity of GEI and LEI is searched, by combing the semantic identification ‘fitness’ and spatiotemporal pattern ‘AsiaWorld-Expo, Dec 7, 2014’ (see Equation 6.1). After searching the relevant background knowledge, the Color Run Hong Kong event is identified. This event is the first Color Run event in Hong Kong and held on Dec 7, 2014

at the AsiaWorld-Expo (Figure 6.10). This is one of the most popular single day events for the Color Run in the Asian region, and there are approximately 16,000 fitness fans and runners participating in this event, bringing about immense human aggregation to the AsiaWorld-Expo and that is reflected by the detected irregularity of the calculated GEI and LEI (Figure 6.8 and Figure 6.9).

Figure 6.10 Color Run event in Hong Kong AsiaWorld-Expo at Dec 7, 2014 (source: https://hk.ulifestyle.com.hk/activity/detail/100465/the-color-run-hong-kong-%E6%9C%80%E5%BF%AB%E6%A8%82%E7%9A%845%E5%85%AC%E9%87%8C%E8%B7%91)

6.3.3 Urban structure understanding

The workflow can also be used for studying the urban structure. Studying the urban structure on a large scale has traditionally been a challenge, which requires considerable labor and may result

in a partial depiction of reality. The consideration is that the geotagged check-ins that residents generate reveal the patterns of human mobility and aggregation in an urban context, which can be used for studying the structure and composition of a city. An application scenario is to study human mobility during a festival. How people move and interact within the urban region during a festival can reveal the distinctly functional areas of the city and residents' perception about them. Consequently, in the experiment, the GEI and LEI for a festival event are calculated. One detected hashtag community is semantically about 'Christmas' (Figure 6.11).

Figure 6.11 The word cloud for representative hashtags of 'Christmas' community

To investigate the geographical patterns of human mobility and aggregation during Christmas, the LEI is calculated with the 'Christmas' geotagged check-ins on Dec 25, 2014. In Figure 6.12, it can be seen, across the whole urban area of Hong Kong, the regions with high LEI are mainly located within three regions, Tsim Sha Tsui, Lan Kwai Fong and Causeway Bay. This finding is consistent with the urban structure of Hong Kong. All three regions are, in Hong Kong, major recreational regions where residents visit for shopping and entertainment during the holidays. Specifically, the regions with highest LEI are around Lan Kwai Fong, which is the most well-known area in Hong Kong for drinking, clubbing and dining. This can be explained by the fact

that, in Lan Kwai Fong, various street performance and celebration activities are held every Christmas, which would attract many tourists and local residents to visit thus causing extreme human aggregation in that region.

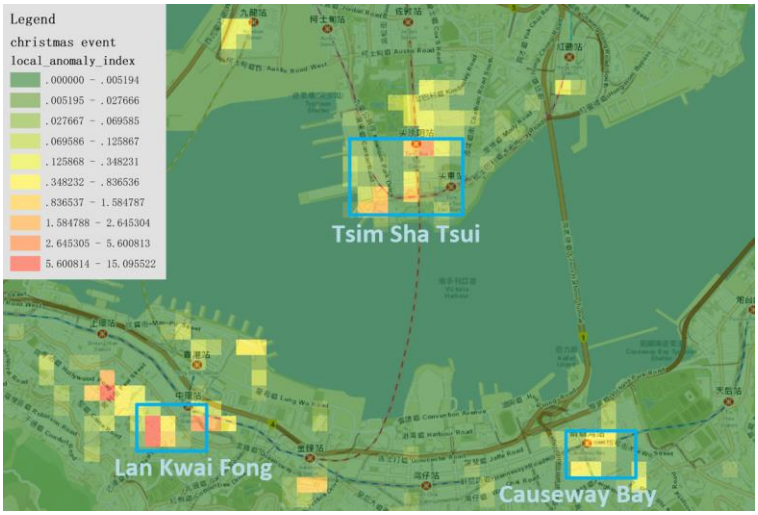


Figure 6.12 The *LEI* value across Hong Kong urban area, calculated from the 'Christmas' check-ins. The high-value regions are mainly in Tsim Sha Tsui, Lan Kwai Fong and Causeway Bay, which are all major recreational regions in Hong Kong.

6.4 Discussion

In this section, some worthy questions are further discussed to provide insights into the proposed methodology: (1) what is the intuition of the proposed workflow? (2) what is the originality and novelty of the proposed workflow?

6.4.1 Intuition of the proposed workflow

The major intuition of the proposed workflow is to take advantage of new geographical features for event detection, which means using the change ("outliers") of the geographical patterns revealed by VGI to differentiate potential events.

Event detection with social media is not a novel topic and massive relevant works have been done. Some previous methods have shown good effectiveness. The kinds of events referred to in this chapter (i.e., political protest, ColorRun activity) can possibly be detected as well, by adopting some of previous methods. However, the points of this section lie in providing new insights into the following issues: what kind of features can be used for detecting events or, from what aspects, can a social event be differentiated with VGI data. Previous methods are mostly semantic-based and detect the "bursts" of certain semantic signals to identify events. While the consideration in this chapter is that a social event will also affect how the objects spatially distribute and mutually interact, thus causing irregular geographical patterns, and by introducing depictive features with VGI, such geographical irregularities can be quantified and used to distinguish social events. Based on above consideration, a comprehensive workflow is designed where the event feature is represented by investigating the geographical patterns (e.g., spatial autocorrelation) of VGI and the social events are detected thereby.

After searching the relevant literature, some previous methodologies that incorporated spatial-related components are indeed found. However, these works still detected events mainly by finding burst of semantic signal, while the spatial-related components mainly served as a complementary role to identify the spatial information of the detected event, such as detecting event location and assessing the influenced area rather than be used, per se, as major features to differentiate social events. Another previous study that might be potentially related to this work is Lee and Sumiya (2010) where the authors measured the regional regularities with tweet and user count, which is relatively simple and intuitive. While in this chapter, the consideration is that social events can not only lead to "bursts" of certain semantic signals and posts/users count, but also cause structural change of the geographical patterns like how objects spatially distribute

and interact with each other. So, in this chapter, rather than investigating the numeric change of post/user count, features are constructed to indicate the structural change of objects' geographical distribution pattern. Specifically, the clustering patterns (e.g., spatial autocorrelation) are selected as a tentative attempt and designed the event features accordingly. The experiments are conducted and demonstrates the effectiveness of the design.

6.4.2 Novelty of the proposed workflow

The main novelty of the proposed workflow is to construct event feature based on the geographical patterns revealed by VGI and differentiate events by detecting feature irregularities. The originalities are mainly related to the second and third modules (i.e., Geography-based Event Representation, Detection of Event Feature Irregularity) of the proposed workflow in Figure 6.1, where both global and local indicators (features) are constructed by investigating the geographical patterns of VGI data and the feature irregularities are detected.

While designing the analytical workflow, the problem-solving principle is "Entities are not to be multiplied without necessity" (Occam's razor). The intention of the proposed model is to make each component clearly straightforward and effective, rather than seemingly "sophisticated" or "fancy". In this chapter, the intention is not to invent some novel topic modelling or outlier detection algorithms but to create new geography-based features for event representation and detection.

As mentioned above, the point of this chapter study is to provide new insights into the following issues: what kinds of features can be used for event detection or, from what aspects, can a social event be differentiated with VGI data. Specifically, geography-based features are constructed based on the consideration that a social event might cause irregular geographical patterns and

such geographical irregularities could be used, in turn, to differentiate social events. By experiments with sample datasets, this study demonstrates the feasibility to detect social events with geography-based features and could be complementary to the current semantic-based event detection method. This can be a potentially promising direction as relatively fewer relevant works have been done before, compared with the massive volume of the semantic-based event detection methods.

On the other hand, some parts of the proposed methodology can be further refined to be more "sophisticated" and "effective". One worthwhile question is what other geographical patterns and transformation formula can be used to create event feature so that the social events can be more differentiable in the new feature space. In this chapter, the clustering patterns (e.g., spatial autocorrelation) is picked and logarithmic formula used for event transformation. This is an initial attempt but proves the feasibility of detecting events with the geography-based feature. And the current work may be extended into a comprehensive analytic framework where different kinds of geographical pattern indicators and feature transformation operations are tested and incorporated.

6.5 Summary

Chapter 6 focuses on addressing the flowing gap: how to detect events by investigating irregular geographic patterns?

Most research works on event detection with social media are using semantic-based methods and text mining. Platforms, such as Twitter, provide accessible streaming data, which contains real-time text information, making it possible to detect emerging social events by examining the change of semantic-based features. However, a new event detection methodology can be

proposed by investigating the geographical patterns of GSMD. The intuition is that a social event will affect how the objects spatially distribute across certain regions and how they mutually interact thus causing irregular geographical patterns, especially irregular human mobility and interaction patterns (e.g., sports games causing intense human aggregation or terrorism attacks causing sudden evacuation from certain regions). By introducing depictive measuring features with GSMD and identifying the feature irregularity, such geographical patterns, in turn, can be used to distinguish potential social events.

Consequently, a new event detection method is proposed by finding the spatiotemporal irregularities of the GSMD. Basically, to detect social events from geotagged social media data, three aspects are considered by the data model. First, the consideration is that the semantic meaning should be investigated to identify the topics for social events. Therefore, a data-driven topic modelling method (the method proposed in Chapter 3) is adopted to detect hashtag communities and identify potential topics for the posts, in the rich-hashtag environment. Second, besides semantic identification, spatiotemporal identification is also considered by the model to pinpoint the events. This is derived from the intuition that across an area, the activities of the same category may take place at different places and times, which means that only by combing semantic and spatiotemporal identification can a particular event be specified. Consequently, in the model, a comprehensive workflow is implemented that included both global and local indexes to indicate the temporal and spatial information of the events, respectively. Third, to differentiate social events, a novel geography-based event representation method is developed. Previous works mainly detected events by investigating the semantic feature of the streaming social media data, while the consideration is that social events may also lead to irregular geographical patterns of human mobility, and such geographical irregularity, in turn, can be used

to detect social events. Specifically, in this study, a one-dimensional event feature is constructed based on the quantitative measurement of crowd aggregation (spatial autocorrelation) and the event is detected by finding the feature irregularity. By experimenting with sample data, several events (e.g., site-clearance operation and Color Run) are detected. It shows that, as a matter of popularity, those events attract a large number of citizens to the site, causing unusual human aggregation in corresponding areas and, subsequently, event feature irregularity. Such irregularities are captured and finally, the events are detected, using the proposed method.

Chapter 7 Conclusions

7.1 Research summary

7.1.1 Research scope and previous gaps

The rise of geotagged social media data (GSMD) has provided researchers with new tools to study traditional research questions. This thesis focuses on the semantic, spatial and temporal information attached to GSMD and develops new data handling methods from three progressive perspectives: (1) semantic modelling, (2) spatial semantic modelling and (3) spatiotemporal semantic modelling.

The scopes of this thesis are as below:

- **Semantic modelling**: What is the **topic** of the GSMD? (**Topic Modelling**)
- **Spatial semantic modelling:** Where are the **regions of specific themes**, inferred from GSMD? How **desirable** are these regions? (**Desirable Thematic Region Detection**)
- **Spatiotemporal semantic modelling**: **What** is happening at **some place** and at **some time,** inferred from GSMD? (**Event Detection**)

There are other previous research works studying similar scopes as this thesis. However, some challenging **gaps** still remain unaddressed by previous works:

(1) Semantic modelling: how to effectively discover topics with short noisy social media content?

- One gap is how to effectively discover topics with short noisy social media content. Statistical methods have limited effectiveness in handling social media data. A major reason is that statistical methods commonly require large amounts

of well-organized documents as training data, which is inconsistent with the social media environment where short and noisy texts predominate; another reason is that some statistical methods require predefined parameters, such as counts of topics, which are arbitrary and unpredictable due to a lack of a priori knowledge.

(2) Spatial semantic modelling I: how to develop a model to calculate regional desirability with consideration of varying spatial scales of regions?

- Regions with large areas may have more visits than those with small areas because they cover more venues. When calculating the regional desirability, the influence of spatial scales needs to be taken into consideration. This scale issue is specific to region, as venues are typically treated as points without consideration of sizes.

(3) Spatial semantic modelling II: how to develop an accurate model to calculate regional desirability when the interactions of relevant features are still unknown?

- Traditional methods mainly use empirical models to predict regional desirability. These empirical models are intuitive and inaccurate in predicting desirable regions, as potential perceptual bias is introduced. This is mainly because that the hidden interactions between user check-ins and regional desirability haven't be clearly explained and precisely modelled yet, so that the empirical models based on human empirical intuition and presumed relationship are used for rough approximation.

(4) Spatiotemporal semantic modelling: how to model features by investigating geographic patterns and detect events by finding irregular features?

- Most research works on event detection with social media are using semantic-based methods and text mining. However, a new event detection methodology can be proposed by investigating the geographical patterns of GSMD. The intuition is that a social event will affect how the objects spatially distribute across certain regions and how they mutually interact thus causing irregular geographical patterns, especially irregular human mobility and interaction patterns (e.g., sports games causing intense human aggregation or terrorism attacks causing sudden evacuation from certain regions). By introducing depictive measuring features with GSMD and identifying the feature irregularity, such geographical patterns, in turn, can be used to distinguish potential social events.

7.1.2 Contributions of this thesis

Consequently, this thesis develops new methods from the above perspectives, to tackle the aforementioned challenges. The contributions of this thesis are summarized as below:

(1) Semantic modelling:

- A new hashtag network model is developed for topic modelling, with good performance on short social media texts. By dividing the network into different communities, the hashtags are clustered to different topics. This workflow is data-driven and requires no well-organized training data or priori knowledge such as topic counts, thus reducing potential perceptual biases.

(2) Spatial semantic modelling I:

- A new scale-concerned model is proposed for calculating regional desirability. The proposed model is based on Hypertext Induced Topic Search (HITS) with the consideration of spatial scales of the regions.

(3) Spatial semantic modelling II:

- A new data-driven model RegNet is proposed for calculating regional desirability. The RegNet is a multi-layer neural network framework. Given training data, the proposed RegNet is data-driven and can model the hidden unknown patterns between user check-ins and regional desirability without prior knowledge and presumed relationships, thus alleviating the inaccuracy and presumed bias introduced by intuitive empirical models and achieving better prediction results.

(4) Spatiotemporal semantic modelling:

- A new model is developed for event detection by characterizing spatial patterns of GSMD and finding spatiotemporal irregular patterns. The workflow first uses topic modelling to detect the hashtag communities with GSMD semantic data. Then, the proposed model constructs both global and local features/indicators to characterize spatial patterns of GSMD. An outlier test is then implemented and a local feature map is generated, to spatiotemporally identify the potential social events.

The experiments are conducted with real-world datasets and demonstrate the effectiveness of the proposed models.

The work of this these is summarized as Table 7.1.

Table 7.1 Summary of thesis

Logic flow of thesis	**Research scopes**	**Previous gaps**	**Thesis contributions**

Sematic modelling	What is the topic of the GSMD? (Topic Modelling)	How to effectively discover topics with short noisy social media content?	A new hashtag network model for topic modelling, with good performance on short social media texts
Spatial semantic modelling	Where are the desirable regions of specific themes, inferred from GSMD? (Desirable Thematic Region Detection)	**I**: How to develop a model to calculate regional desirability with consideration of varying spatial scales of regions?	A new scale-concerned model for calculating regional desirability
		II: How to develop an accurate model to calculate regional desirability when the interactions of	A new data-driven model RegNet for calculating regional desirability

		relevant features are still unknown?	
Spatiotemporal semantic modelling	What is happening at some place and at some time, inferred from GSMD? (Event Detection)	How to model features by investigating geographic patterns and detect events by finding irregular features?	A new model for event detection by characterizing spatial patterns of GSMD and finding spatiotemporal irregular patterns

7.2 Limitations and future work

Admittedly, analytics of GSMD in this thesis has certain limitations. The limitations and potential future work are list as below:

(1) Semantic modelling:

- Regarding the evaluation approach, the evaluation of the proposed topic modelling method is not thorough. Although Chapter 3 has verified the proposed model's effectiveness by investigating the semantic similarity within and between the hashtag communities, a comparative study of the proposed model against other existing topic modelling methods is still lacking. Further efforts may be made towards a comprehensive evaluation of the proposed model, by including comparative study.

(2) Spatial semantic modelling:

- Regarding the interpretability of the proposed model, RegNet suffers from black box issue. As the RegNet approximates the mapping function from user check-ins to regional desirability, it does not necessarily give knowledge on the structure of the function being approximated or how to do parameter tuning. The network weights are not capable of inferring the form of approximated function. Further studies are needed so that the internal mechanisms between regional desirability and user online footprints can be formulaically explained.
- Regarding the size of thematic regions, how to select a proper region size remains challenging. To train RegNet, spatial regions need to be clustered. A small region size may cover insufficient POIs and check-ins, making the regional rating and user check-ins unrepresentative, while a large region size may cause limited amount of total clustered regions, leading to lack of training data and overfitting. How to choose a proper region size and balance the trade-off can be a potential direction for future work.
- Regarding detecting the thematic regions, the boundaries of the regions are manually defined to match the approximate geographical distribution of the check-ins, causing inaccurate prediction of region desirability and user expertise. Research to develop an improved region boundary outlining method would be of future benefit.
- Regarding the optimal size of encoded dimension of RegNet, whether the current optimal value is a local or global optimum remains to be further investigated.

- Regarding the experiment datasets, there may be spatial inconsistency between the GSMD check-ins and POIs. This thesis clusters thematic regions with geotagged check-ins and calculate the ground-truth desirability of regions by summating the desirability of ratings. Theoretically, each check-in can be assigned to a POI. However, in practice, the check-in dataset and POIs dataset came from different platforms, and there is no accurate way of mapping check-ins to POIs. Since a mutual reinforcement process between users and POIs is used for inferring desirability and expertise, such mismatches might negatively affect the effectiveness in predicting the region desirability.
- Regarding generalization of the conclusions, due to the issue of data availability, this thesis is related to only one city, Hong Kong, and the generalization is limited. More cities, of different sizes can be analyzed when data becomes available.

(3) Spatiotemporal semantic modelling:

- Regarding the geographic patterns considered in the proposed event detection workflow, the categories of geographical patterns that are used to construct event features are still relatively simple and lacking in variety. In this study, the clustering patterns (e.g., spatial autocorrelation) and logarithmic formula are used for event transformation. This is an initial attempt, however, it proves the feasibility of detecting events using geography-based features. In the future, it is hoped that the current work can be extended into a comprehensive analytic framework where different kinds of geographical pattern indicators and feature

transformation operations are tested and incorporated so that social events can be more effectively differentiated.

- Regarding the evaluation approach, the evaluation methods are still limited in their effectiveness. In the experiments, there are small parts of the irregularities that could not be meaningfully interpreted due to a lack of relevant background information. However, these only accounts for a very minor portion of the total detected events. Most of the detected irregularities, when combined with the semantic and spatiotemporal identification (Equation 6.1), could be clearly interpreted and assigned to the corresponding social events. Due to the unavailability of a thorough ground truth data set (event set E in Chapter 2), the current evaluation of the detection results is conducted using qualitative methods, and quantitative evaluations are absent. An authoritative ground truth dataset can be very helpful in achieving plausible quantitative evaluations.
- Regarding the dynamics of an event, our approach can so far give a static description of the spatiotemporal span of the event (grid cells for spatial scale and time slices for temporal scale), yet can not explore how the event evolves spatiotemporally, which may be an important direction for future work.

(4) Bias of GSMD

- Regarding the bias and representativeness issue, analytics of GSMD inevitably suffers from data bias issues because social media users are a skewed sample of the entire population, mainly consisting of specific groups of people and the younger generation. How to quantify and alleviate the influence of data bias and

improve the effectiveness of the proposed models could be a potential future research direction.

References

Abdelhaq, H., C. Sengstock & M. Gertz (2013) Eventweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment,* 6**,** 1326-1329.

Adam, E., O. Mutanga & D. Rugege (2010) Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review. *Wetlands Ecology and Management,* 18**,** 281-296.

Adams, B. & K. Janowicz. 2012. On the geo-indicativeness of non-georeferenced text. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Ajo-Franklin, J. B., S. Dou, N. J. Lindsey, I. Monga, C. Tracy, M. Robertson, V. R. Tribaldos, C. Ulrich, B. Freifeld & T. Daley (2019) Distributed acoustic sensing using dark fiber for near-surface characterization and broadband seismic event detection. *Scientific reports,* 9**,** 1-14.

Allan, J., J. G. Carbonell, G. Doddington, J. Yamron & Y. Yang (1998) Topic detection and tracking pilot study final report.

Alvanaki, F., S. Michel, K. Ramamritham & G. Weikum. 2012. See what's enBlogue: real-time emergent topic identification in social media. In *Proceedings of the 15th International Conference on Extending Database Technology*, 336-347.

Anselin, L. (1995) Local indicators of spatial association—LISA. *Geographical analysis,* 27**,** 93-115.

Bao, B.-K., W. Min, K. Lu & C. Xu. 2013. Social event detection with robust high-order co-clustering. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, 135-142.

Bao, J., Y. Zheng, D. Wilkie & M. Mokbel (2015) Recommendations in location-based social networks: a survey. *GeoInformatica,* 19**,** 525-565.

Becker, H., M. Naaman & L. Gravano. 2011. Beyond trending topics: Real-world event identification on twitter. In *Fifth international AAAI conference on weblogs and social media*.

Blei, D. M., A. Y. Ng & M. I. Jordan (2003) Latent dirichlet allocation. *Journal of machine Learning research,* 3**,** 993-1022.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte & E. Lefebvre (2008) Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment,* 2008**,** P10008.

Bottou, L. & O. Bousquet. 2008. The tradeoffs of large scale learning. In *Advances in neural information processing systems,* 161-168.

Boubrima, A., W. Bechkit & H. Rivano (2017) Optimal WSN deployment models for air pollution monitoring. *IEEE Transactions on Wireless Communications,* 16**,** 2723-2735.

Cai, L., J. Xu, J. Liu & T. Pei (2018) Integrating spatial and temporal contexts into a factorization model for POI recommendation. *International Journal of Geographical Information Science,* 32**,** 524-546.

Caudal, P. & D. Nicolas (2005) Types of degrees and types of event structures. *Event arguments: Foundations and applications***,** 277-300.

Chen, P., W. Shi, X. Zhou, Z. Liu & X. Fu (2019) STLP-GSM: a method to predict future locations of individuals based on geotagged social media data. *International Journal of Geographical Information Science,* 33**,** 2337-2362.

Cheng, C., H. Yang, I. King & M. R. Lyu. 2012. Fused matrix factorization with geographical and social influence in location-based social networks. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Cheng, H.-T., L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai & M. Ispir. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, 7-10. ACM.

Chua, A., L. Servillo, E. Marcheggiani & A. V. Moere (2016) Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tourism Management,* 57**,** 295-310.

Corley, C. D., C. Dowling, S. J. Rose & T. McKenzie. 2013. Social sensor analytics: Measuring phenomenology at scale. In *2013 IEEE International Conference on Intelligence and Security Informatics*, 61-66. IEEE.

De Albuquerque, J. P., B. Herfort, A. Brenning & A. Zipf (2015) A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International journal of geographical information science,* 29**,** 667-689.

Debba, P., F. Van Ruitenbeek, F. Van Der Meer, E. Carranza & A. Stein (2005) Optimal field sampling for targeting minerals using hyperspectral data. *Remote Sensing of Environment,* 99**,** 373-386.

Diao, Q. & J. Jiang. 2013. A unified model for topics, events and users on twitter. ACL.

Ding, R. & Z. Chen (2018) RecNet: A deep neural network for personalized POI recommendation in location-based social networks. *International Journal of Geographical Information Science,* 32**,** 1631-1648.

Ester, M., H.-P. Kriegel, J. Sander & X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, 226-231.

Gao, H., J. Tang, X. Hu & H. Liu. 2013. Exploring temporal effects for location recommendation on location-based social networks. In *Proceedings of the 7th ACM conference on Recommender systems*, 93-100.

Gao, H., J. Tang & H. Liu. 2012. Exploring social-historical ties on location-based social networks. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Gao, W., S. Emaminejad, H. Y. Y. Nyein, S. Challa, K. Chen, A. Peck, H. M. Fahad, H. Ota, H. Shiraki & D. Kiriya (2016) Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. *Nature,* 529**,** 509-514.

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

García-Palomares, J. C., J. Gutiérrez & C. Mínguez (2015) Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography,* 63**,** 408-417.

Guille, A. & C. Favre. 2014. Mention-anomaly-based event detection and tracking in twitter. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 375-382. IEEE.

Haworth, B. & E. Bruce (2015) A review of volunteered geographic information for disaster management. *Geography Compass,* 9**,** 237-250.

He, X., L. Liao, H. Zhang, L. Nie, X. Hu & T.-S. Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, 173-182. International World Wide Web Conferences Steering Committee.

Hornik, K., M. Stinchcombe & H. White (1989) Multilayer feedforward networks are universal approximators. *Neural networks,* 2**,** 359-366.

Horozov, T., N. Narasimhan & V. Vasudevan. 2006. Using location for personalized POI recommendations in mobile environments. In *International Symposium on Applications and the Internet (SAINT'06)*, 6 pp.-129. IEEE.

Hu, Y., A. John, F. Wang & S. Kambhampati. 2012. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Huang, Q. & D. W. Wong (2016) Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science,* 30**,** 1873-1898.

Hühn, P. (2009) Event and eventfulness. *Handbook of narratology,* 19**,** 80.

Imran, M., C. Castillo, F. Diaz & S. Vieweg. 2018. Processing social media messages in mass emergency: Survey summary. In *Companion Proceedings of the The Web Conference 2018*, 507-511.

Jafarkarimi, H., A. T. H. Sim & R. Saadatdoost (2012) A naive recommendation model for large databases. *International Journal of Information and Education Technology,* 2**,** 216.

Jia, T., X. Yu, W. Shi, X. Liu, X. Li & Y. Xu (2019) Detecting the regional delineation from a network of social media user interactions with spatial constraint: A case study of Shenzhen, China. *Physica A: Statistical Mechanics and its Applications,* 531**,** 121719.

Kautz, T., B. H. Groh & B. M. Eskofier. 2015. Sensor fusion for multi-player activity recognition in game sports. In *Workshop on Large-Scale Sports Analytics, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Kopetz, H. 1991. Event-triggered versus time-triggered real-time systems. In *Operating Systems of the 90s and Beyond*, 86-101. Springer.

Kurashima, T., T. Iwata, T. Hoshide, N. Takaya & K. Fujimura. 2013. Geo topic model: joint modeling of user's activity area and interests for location recommendation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 375-384. ACM.

Kurashima, T., T. Iwata, G. Irie & K. Fujimura. 2010. Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 579-588. ACM.

Lai, J., T. Cheng & G. Lansley (2017) Improved targeted outdoor advertising based on geotagged social media data. *Annals of GIS,* 23**,** 237-250.

Lansley, G. & P. A. Longley (2016) The geography of Twitter topics in London. *Computers, Environment and Urban Systems,* 58**,** 85-96.

LeCun, Y., Y. Bengio & G. Hinton (2015) Deep learning. *nature,* 521**,** 436.

Lee, C.-H. (2012) Mining spatio-temporal information on microblogging streams using a density-based online clustering method. *Expert Systems with Applications,* 39**,** 9623-9641.

Lee, J. Y. & M.-H. Tsou. 2018. Mapping spatiotemporal tourist behaviors and hotspots through location-based photo-sharing service (Flickr) data. In *LBS 2018: 14th International Conference on Location Based Services*, 315-334. Springer.

Lee, R. & K. Sumiya. 2010. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*, 1-10. ACM.

Lemire, D. & A. Maclachlan. 2005. Slope one predictors for online rating-based collaborative filtering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, 471-475. SIAM.

Li, D. & Y. Yang. 2017. GIS monitoring of traveler flows based on big data. In *Analytics in smart tourism design*, 111-126. Springer.

Li, S., S. Dragicevic, F. A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth & A. Stein (2016) Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing,* 115**,** 119-133.

Liou, C.-Y., W.-C. Cheng, J.-W. Liou & D.-R. Liou (2014) Autoencoder for words. *Neurocomputing,* 139**,** 84-96.

Liou, C.-Y., J.-C. Huang & W.-C. Yang (2008) Modeling word perception using the Elman network. *Neurocomputing,* 71**,** 3150-3157.

Liu, Q., Z. Wang & X. Ye (2018) Comparing mobility patterns between residents and visitors using geo-tagged social media data. *Transactions in GIS,* 22**,** 1372-1389.

Liu, Y. & H. S. Seah (2015) Points of interest recommendation from GPS trajectories. *International Journal of Geographical Information Science,* 29**,** 953-979.

Liu, Z., X. Zhou, W. Shi & A. Zhang (2019) Recommending attractive thematic regions by semantic community detection with multi-sourced VGI data. *International Journal of Geographical Information Science***,** 1-25.

Longley, P. A. & M. Adnan (2016) Geo-temporal Twitter demographics. *International Journal of Geographical Information Science,* 30**,** 369-389.

Manolakis, D., E. Truslow, M. Pieper, T. Cooley & M. Brueggeman (2013) Detection algorithms in hyperspectral imaging systems: An overview of practical algorithms. *IEEE Signal Processing Magazine,* 31**,** 24-33.

Meira-Machado, L., J. de Uña-Álvarez, C. Cadarso-Suárez & P. K. Andersen (2009) Multi-state models for the analysis of time-to-event data. *Statistical methods in medical research,* 18**,** 195-222.

Mikolov, T., K. Chen, G. Corrado & J. Dean (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moran, P. A. (1950) Notes on continuous stochastic phenomena. *Biometrika,* 37**,** 17-23.

Mubashir, M., L. Shao & L. Seed (2013) A survey on fall detection: Principles and approaches. *Neurocomputing,* 100**,** 144-152.

Newman, M. E. (2006) Modularity and community structure in networks. *Proceedings of the national academy of sciences,* 103**,** 8577-8582.

Odom, M. D. & R. Sharda. 1990. A neural network model for bankruptcy prediction. In *1990 IJCNN International Joint Conference on neural networks*, 163-168. IEEE.

Paldino, S., D. Kondor, I. Bojic, S. Sobolevsky, M. C. Gonzalez & C. Ratti (2016) Uncovering urban temporal patterns from geo-tagged photography. *PloS one,* 11**,** e0165753.

Panteras, G., S. Wise, X. Lu, A. Croitoru, A. Crooks & A. Stefanidis (2015) Triangulating social multimedia content for event localization using Flickr and Twitter. *Transactions in GIS,* 19**,** 694-715.

Park, M.-H., J.-H. Hong & S.-B. Cho. 2007. Location-based recommendation system using bayesian user's preference model in mobile devices. In *International conference on ubiquitous intelligence and computing*, 1130-1139. Springer.

Paule, J. D. G., Y. Sun & Y. Moshfeghi (2019) On fine-grained geolocalisation of tweets and real-time traffic incident detection. *Information Processing & Management,* 56**,** 1119-1132.

Peary, B. D., R. Shaw & Y. Takeuchi (2012) Utilization of social media in the east Japan earthquake and tsunami and its effectiveness. *Journal of Natural Disaster Science,* 34**,** 3-18.

Petkos, G., S. Papadopoulos & Y. Kompatsiaris. 2012. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, 1-8.

Petkos, G., M. Schinas, S. Papadopoulos & Y. Kompatsiaris (2017) Graph-based multimodal clustering for social multimedia. *Multimedia Tools and Applications,* 76**,** 7897-7919.

Prateek, M. & V. Vasudeva. 2016. Improved topic models for social media via community detection using user interaction and content similarity. In *Intelligence, Social Media and Web (ISMW FRUCT), 2016 International FRUCT Conference on*, 1-7. IEEE.

Ramaswamy, L., P. Deepak, R. Polavarapu, K. Gunasekera, D. Garg, K. Visweswariah & S. Kalyanaraman. 2009. Caesar: A context-aware, social recommender system for low-end mobile devices. In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, 338-347. IEEE.

Raymond, R., T. Sugiura & K. Tsubouchi. 2011. Location recommendation based on location history and spatio-temporal correlations for an on-demand bus system. In *Proceedings of*

the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 377-380. ACM.

Ritter, A., O. Etzioni & S. Clark. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1104-1112.

Rodrigues, J. G., M. Kaiseler, A. Aguiar, J. P. S. Cunha & J. Barros (2015) A mobile sensing approach to stress detection and memory activation for public bus drivers. *IEEE Transactions on Intelligent Transportation Systems,* 16**,** 3294-3303.

Rosner, B. (1983) Percentage points for a generalized ESD many-outlier procedure. *Technometrics,* 25**,** 165-172.

Rykov, Y., O. Nagornyy & O. Koltsova (2016) Semantic and geospatial mapping of instagram images in saint-petersburg. *power,* 2607**,** 75.

Sakaki, T., M. Okazaki & Y. Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, 851-860. ACM.

Schinas, M., S. Papadopoulos, Y. Kompatsiaris & P. Mitkas (2018) Event detection and retrieval on social media. *arXiv preprint arXiv:1807.03675*.

Sinclair, M., A. Ghermandi & A. M. Sheela (2018) A crowdsourced valuation of recreational ecosystem services using social media data: An application to a tropical wetland in India. *Science of the total environment*, 642**,** 356-365.

Specht, D. F. (1991) A general regression neural network. *IEEE transactions on neural networks,* 2**,** 568-576.

Steiger, E., B. Resch & A. Zipf (2015) Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science,* 30**,** 1694-1716.

Sun, Y., H. Fan, M. Bakillah & A. Zipf (2015) Road-based travel recommendation using geo-tagged images. *Computers, Environment and Urban Systems,* 53**,** 110-122.

Varatharajan, R., G. Manogaran, M. K. Priyan & R. Sundarasekar (2018) Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm. *Cluster Computing,* 21**,** 681-690.

Vincent, P., H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol & L. Bottou (2010) Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research,* 11.

Wachowicz, M. & T. Liu (2016) Finding spatial outliers in collective mobility patterns coupled with social ties. *International Journal of Geographical Information Science,* 30**,** 1806-1831.

Wan, L., Y. Hong, Z. Huang, X. Peng & R. Li (2018) A hybrid ensemble learning method for tourist route recommendations based on geo-tagged social networks. *International Journal of Geographical Information Science,* 32**,** 2225-2246.

Wei, W., K. Joseph, W. Lo & K. M. Carley. 2015. A bayesian graphical model to discover latent events from twitter. In *Ninth international AAAI conference on web and social media*.

Weiler, A., M. Grossniklaus & M. H. Scholl (2017) Survey and experimental analysis of event detection techniques for twitter. *The Computer Journal,* 60**,** 329-346.

Weng, J. & B.-S. Lee. 2011. Event detection in twitter. In *Fifth international AAAI conference on weblogs and social media*.

Werner-Allen, G., K. Lorincz, M. Ruiz, O. Marcillo, J. Johnson, J. Lees & M. Welsh (2006) Deploying a wireless sensor network on an active volcano. *IEEE internet computing,* 10**,** 18-25.

Yang, D., D. Zhang, Z. Yu & Z. Wang. 2013. A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, 119-128. ACM.

Yang, S. & M. Sklar. 2016. Detecting Trending Venues Using Foursquare's Data. In *RecSys Posters*.

Yang, W.-S., H.-C. Cheng & J.-B. Dia (2008) A location-aware recommender system for mobile shopping environments. *Expert Systems with Applications,* 34**,** 437-445.

Ye, M., P. Yin & W.-C. Lee. 2010. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 458-461. ACM.

Ye, M., P. Yin, W.-C. Lee & D.-L. Lee. 2011. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 325-334. ACM.

Yilmaz, E., E. Kanoulas & J. A. Aslam. 2008. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 603-610.

Yin, Z., L. Cao, J. Han, C. Zhai & T. Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on World wide web*, 247-256.

Yu, M., M. Bambacus, G. Cervone, K. Clarke, D. Duffy, Q. Huang, J. Li, W. Li, Z. Li & Q. Liu (2020) Spatiotemporal event detection: a review. *International Journal of Digital Earth***,** 1-27.

Yuan, Q., G. Cong, Z. Ma, A. Sun & N. M. Thalmann. 2013. Time-aware point-of-interest recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 363-372. ACM.

Zhang, F., J. Zu, M. Hu, D. Zhu, Y. Kang, S. Gao, Y. Zhang & Z. Huang (2020) Uncovering inconspicuous places using social media check-ins and street view images. *Computers, Environment and Urban Systems,* 81**,** 101478.

Zhang, L., X. Sun & H. Zhuge. 2013. Location-driven geographical topic discovery. In *2013 ninth international conference on semantics, knowledge and grids*, 210-213. IEEE.

Zhang, X., X. Chen, Y. Chen, S. Wang, Z. Li & J. Xia (2015) Event detection and popularity prediction in microblogging. *Neurocomputing,* 149**,** 1469-1480.

Zheng, Y. & X. Xie (2011) Learning travel recommendations from user-generated GPS traces. *ACM Transactions on Intelligent Systems and Technology (TIST),* 2**,** 2.

Zheng, Y., L. Zhang, X. Xie & W.-Y. Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web*, 791-800. ACM.

Zhou, D., L. Chen & Y. He. 2015. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *Twenty-ninth aaai conference on artificial intelligence*.

Zhou, H., P. Wang & H. Li (2012) Research on adaptive parameters determination in DBSCAN algorithm. *Journal of Xi'an University of Technology,* 28**,** 289-292.

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