

# What Makes the City Pulse

Yang Li

B.Sc.

A Dissertation submitted in fulfilment of the  
requirements for the award of  
Master of Science (M.Sc.)

to the



Dublin City University

School of Computing

CLARITY: Centre for Sensor Web Technologies

Supervisor: Prof. Alan F. Smeaton

October 2013

# Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Master of Science is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Yang Li

ID No.: 57100951

Date:

# Abstract

The topics of this thesis are event detection and social network analysis in social media. Our work centres on Geo-tagged User Generated Content (UGC) in Twitter, such as Twitter data generated from the metropolitan area of Dublin Ireland over a one month period of time. In this thesis we address the problem of how to detect small scale unexpected events using UGC both in real-time and retrospectively. We proposed a language-text joint modeling algorithm to cope with the large volume and unstructured nature of UGC. We also demonstrate our discovery of interesting correlations between a Twitter user's social communities and their mobility patterns. Finally a set of features are proposed for carrying out Twitter user's account type classification, for the purpose of irrelevant contents filtering. This thesis includes several experimental evaluations using real data from users and shows the performance of our algorithms in event detection and provide evidence for our discoveries.

# Acknowledgements

I would like first to give sincere thanks to my supervisor Prof. Alan Smeaton who gave me so much valuable guidance not only on my research directions but also on my research methodologies. Every time I have discussions with him, I am impressed by his profound knowledge and his innovative guidance. I have deep gratitude to him for all he has done for me !

Many thanks to my office mates Dr. Zhengwei Qiu, Dr. Eoin Hurrell, Dr. Graham Healy, Dr. Niamh Caprani and Dr. Adam Bermingham who accompany me in my work every day. To work with them has been really great ! During my M.Sc., I also got much expertise from Dr. Neil O'Hare. Without their knowledge and suggestions, I could not solve the problems I met within my research. Special thanks to Dr. Hyowon Lee who provided me with guidance on human interface design work. I also appreciate the help for my experiments from Dian Zhang, Dr. David Scott, and Jinlin Guo.

I can not fully express my gratitude to my Mum and Dad who support me all the time. Their endless care and unselfish love encouraged me to conquer whatever difficulties I come across in my life and research. I am such a lucky guy because I have my beloved wife Shimiao Cheng in company with me during the most critical period. I owe many thanks to her for her encouragement, support and understanding. I also would like to give this reward as a gift to my upcoming daughter Eva.

The work of my M.Sc is supported by Science Foundation Ireland under grant number 07/CE/I1147 (CLARITY CSET) and by an IBM PhD fellowship award 2011-2012. Many thanks to them ! I also wish to thank Dr. Giusy di Lorenzo (IBM Ireland) who has been of great help during my research internship work in the IBM Smart City Research Lab, Ireland.

# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgements</b>	<b>ii</b>
<b>Table of Contents</b>	<b>iii</b>
<b>List of Figures</b>	<b>vi</b>
<b>List of Tables</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Introducing Social Media as a New Way of Sensing . . . . .	3
1.1.1 Potential Application of Social Media as an Event Detection Tool . . . . .	5
1.2 Introduction to Twitter Community Analysis . . . . .	6
1.3 Other type of Sensors . . . . .	7
1.4 Hypothesis and Research Questions: . . . . .	8
1.5 Thesis Structure . . . . .	11
<b>2 Background and Related Work</b>	<b>13</b>
2.1 Introduction . . . . .	13
2.2 Event Detection in Social Media . . . . .	16
2.3 Social Network Community Analysis and Population Demographics . . . . .	21
2.3.1 Homophily . . . . .	21
2.4 Summary . . . . .	22

<b>3</b>	<b>Event Detection in Social Media</b>	<b>24</b>
3.1	Introduction . . . . .	24
3.2	Twitter and Geotagging . . . . .	26
3.3	Discovering Socio-geographical Boundaries . . . . .	28
3.4	Geo-Social Event Detection . . . . .	30
3.4.1	Measuring Geographical Regularities . . . . .	31
3.4.1.1	Regular Twitter Activity . . . . .	31
3.4.1.2	Locations as bags-of-words . . . . .	33
3.4.1.3	Modeling The Language of Locations . . . . .	33
3.4.1.4	Reliability of Our Language Models . . . . .	34
3.4.1.5	Semantic Irregularity . . . . .	35
3.4.2	Event detection . . . . .	35
3.5	Experiments and Results . . . . .	36
3.5.1	The Twitter Dataset . . . . .	36
3.5.2	Evaluation Measures . . . . .	37
3.5.3	Evaluation . . . . .	38
3.6	Analysis And Future Work . . . . .	40
3.6.1	Twitter on the Way . . . . .	40
3.6.2	Event Detection Through Bluetooth Devices . . . . .	42
3.7	Summary . . . . .	43
<b>4</b>	<b>Social Community And Population Demographics Correlation</b>	<b>45</b>
4.1	Introduction . . . . .	45
4.2	Community Based Profile . . . . .	46
4.2.1	The Homophily Phenomenon . . . . .	47
4.2.2	Twitter Community Detection . . . . .	47
4.2.3	Mobility Patterns . . . . .	50
4.2.4	Homophily in the Relationship Between User Community and Mo- bility Patterns . . . . .	51

4.2.5	Experiments and Results . . . . .	52
4.2.5.1	The Dataset . . . . .	52
4.2.5.2	Evaluation . . . . .	52
4.3	Ranking Twitter User’s Influence . . . . .	53
4.4	Analysis And Future Work . . . . .	58
4.4.1	Correlation Between Social Relationship and Socio-economic Back-ground . . . . .	59
4.5	Summary . . . . .	60
<b>5</b>	<b>User Tweeting Behaviour Analysis</b>	<b>61</b>
5.1	Introduction . . . . .	61
5.2	Geographical Behaviour Analysis . . . . .	62
5.3	Temporal Dynamic Analysis . . . . .	67
5.4	Geographical Distribution vs. Population Density . . . . .	70
5.4.1	Summary . . . . .	75
<b>6</b>	<b>Twitter Source Classification</b>	<b>76</b>
6.1	Introduction . . . . .	76
6.2	Feature Selection . . . . .	77
6.3	Experiments and Results . . . . .	78
6.3.1	Experimental Setup . . . . .	78
6.3.2	Performance Evaluation . . . . .	79
6.4	Summary . . . . .	80
<b>7</b>	<b>Discussions and Future Work</b>	<b>82</b>
7.1	Main Contributions . . . . .	84
7.2	Future Work . . . . .	85
	<b>Bibliography</b>	<b>86</b>

# List of Figures

1.1	Traditional Media vs Social Media . . . . .	4
1.2	Bluetooth sensing devices . . . . .	7
2.1	Twitter community tweets about some major events . . . . .	15
3.1	Bounding box covering research target area . . . . .	25
3.2	Geo-social partitioning of Dublin into 25 clusters . . . . .	30
3.3	Population distribution of Dublin in Small Areas . . . . .	30
3.4	Twitter occurrences in hourly bins . . . . .	32
3.5	UGC projection to unit length distance . . . . .	37
3.6	Elbow method for K-means clustering results . . . . .	39
3.7	Twitter bound . . . . .	42
3.8	Tweet occurrences along the Naas Road . . . . .	42
3.9	Video footage of the street . . . . .	43
4.1	Forming a community . . . . .	49
4.2	A Twitter community . . . . .	50
4.3	Tweeting activity in 25 different zones, vertical axis numbers are user ids, darker blue shading represent higher levels of tweeting activity . . . . .	51
4.4	Twitter community . . . . .	52
4.5	Twitter Community 1 . . . . .	56
4.6	Twitter Community 2 . . . . .	57



5.1	Location distribution and power law fit . . . . .	65
5.2	Comparison between tweet geolocations and city roadmaps . . . . .	67
5.3	Distribution of tweet volumes. . . . .	70
5.4	Population (age group 16-59) density of Phoenix Park area . . . . .	72
5.5	Population (age group 16-59) density of Small Areas . . . . .	73
5.6	Twitter activity in Small Areas . . . . .	74
5.7	Unique number of Twitter ids in Small Areas . . . . .	74
6.1	Classification performance for individual features. . . . .	79
6.2	Classification performance for individual features on unbiased dataset. . . . .	80

# List of Tables

3.1	Most popular Geotagged Tweets source . . . . .	28
3.2	Accuracy at different level of MRR . . . . .	40
4.1	Percentage of tweets of type conversation . . . . .	46
4.2	Classification results using the partition feature . . . . .	53
4.3	Top 5 Klout score owners in the world . . . . .	55
4.4	Users PageRank score vs. Klout score in different Communities . . . . .	58
5.1	Percentage of users tweeting activities in different zones . . . . .	63

# Chapter 1

## Introduction

If you live in an urban area, then how well do you know the city in which you live? How do you keep yourself informed of the multiplicity of social, economic, cultural and environmental events that happen around your city every day? Many of us may pride ourselves on being in touch with what is happening in our environments, on having our finger on the pulse of our “home town”, but many more would perhaps like to be better informed and there are always things going on that you don’t know of, or things that happen in an unanticipated, unscheduled way. Especially, this kind of information is of huge interest to people like city managers, who want to collect information from every possible corner of the city in an automatic and easy way. These requirements from city managements are incorporated into the Smart City Project.

The **Smart City project**, introduced by IBM [56], is a *city which functions in a sustainable and intelligent way, by integrating all its infrastructures and services into a cohesive whole and using intelligent devices for monitoring and control to ensure sustainability and efficiency*. The solution to realising the smart city goal is to build several smart systems for specific requirements. Such smart systems are expected to have the functionalities of being able to automatically detect unexpected events and generate alerts to users in almost real-time together with analysis of the event contexts. Such systems should also be able to carry out analysis of past-events. We also expect such systems to be able to help users to understand city dynamics in a much simpler and effective way.

Such smart systems in a city may be represented by video surveillance, on the basis of closed circuit television (CCTV), already in position and of widespread use in our cities. Video surveillance has been the traditional and conventional approach to capture events in urban environments. Technological advances in the last decade have led to a large number of distinct research topics related to video surveillance, including crowd density estimation [21, 65, 42, 15], crowd behaviour monitoring and face recognition [26], and modeling and identification of group motion. However, despite substantial progress made in recent years, numerous challenges still remain, such as physical challenges because CCTV can be hard and expensive to install, coverage can be less than 100% of the area of the city, data analysis relies heavily on human involvements, surveillance can be sensitive to environmental factors such as rain, snow, darkness, even spiders on the camera lens. Most especially, context analysis through video processing remains a challenging task in terms of accuracy and processing efficiency.

The forces that shape the dynamics of a city are multifarious and complex. Cultural perceptions, economic factors, municipal borders, demography, geography, and resources all shape and constrain the texture and character of local urban life. It can be extremely difficult to convey these intricacies to an outsider; one may call them well-kept secrets, sometimes only even partially known to the locals. When outsiders, such as researchers, journalists, or city planners, do want to learn about a city, it often requires hundreds of hours of observation and interviews. And although such methods offer a way to gather deep insights about certain aspects of city life, they simply do not scale, and so can only ever uncover a partial image of the inner workings of a city.

Social media as a new form of sensing technology has been adopted by some of the above smart systems in recent years and has attracted many researchers into the field who are interested in analysis of city dynamics. However, social media is different from traditional media as an information source in many aspects. In order to assure reliability and to match the performance of traditional media, there are still many challenges to be solved, for example: how can we correctly interpret these new multimedia contents? Can we efficiently extract useful information from this immense stream of information without getting over-

whelmed by non-relevant contents? The current research area of social media analysis is trying to answer these two questions. Twitter in particular has been the most studied social media by researchers, because of its popularity and public availability. In this research we use Twitter as our main social media research target. Although we focus explicitly on Twitter in this thesis, in the future work, we will incorporate other types of social media, such as Facebook and Foursquare for the purpose of event detection and social network analysis.

## **1.1 Introducing Social Media as a New Way of Sensing**

The increasing use of ubiquitous devices by social media users, including publishing their live status through GPS-embedded smart devices, creates a new sensing paradigm in which social media users are part of a distributed sensory organism of the city. This opens up unique opportunities to measure urban dynamics in all of its facets, from social events to demographics.

The various forms of social media include Twitter, Facebook, Foursquare, Flickr, YouTube, etc. These new ways of communication have attracted tens of millions of users, who would like to share what they see and to express their feelings at anytime and from anywhere. These User Generated Contents (UGC) are information rich, they usually include text descriptions, geo-locations, time and a large amount of meta-data associated with the user who generates the content. Such characteristics present huge potential for real-time event detection and population demographic analysis. Each individual who is a social media user and also a participant in an event can be considered as a sensor, and for these sensors to be able to generate a live report they only need a network connection which is already in position in many modern cities. This presents a huge advantage over CCTV systems in terms of accessibility and resistance to environmental affects.

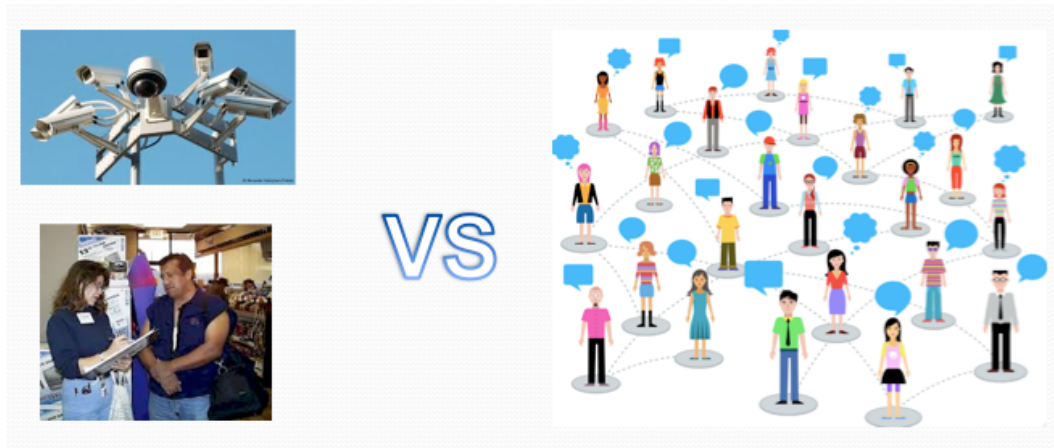


Figure 1.1: Traditional Media vs Social Media

For example, during the 2010 snowstorm in Scotland, people were better informed about blocked traffic, about motorists stuck in snow, about impassible roads, etc. in real-time through Twitter than through the official agencies (police and other governmental agency). Such updates were a lifeline for many of the travelers, some of them caught for almost 20 hours on the M80 motorway. However, many such data sources can be noisy or contain misinformation, as was the case in the recent riots in London, and the challenge is to identify the reliability and significance of a piece of information and then inform the stakeholders appropriately. Others, such as journalists, city planners, essential service providers (the police force, fire and rescue services) have crucial vested interests in having early sight of events in a city as well as the ability to track these events as they evolve. For such agencies it would be very useful to record key characteristics of events after the fact, from factual details as to their effect on city infrastructural resources to less tangible impacts such as public response or feeling. It is important to be able to record such factors either for posterity or in order to be able to better equip our city and its inhabitants to respond appropriately the next time such phenomena occur. The objective of utilizing social networks is to identify the events as and when they occur, and together with novel visualization tools we are able to create real-time representation measuring the contents and scale of events, and also record them for some future analyses to identify the cause-effect patterns of such events. The analyzed results generated from such a system will form a reference for the city

management that can be used to improve their current service facilities or to make smarter decisions for future event planning activities.

Beyond a more comprehensive overview and analysis of events, both in real-time and retrospectively, for prediction, this holistic approach is vital in potentially uncovering inaccessible and hidden information about events that might never otherwise come to light. For example, during a large concert, congestion at a particular traffic light is causing traffic chaos in a given area and the possibility of easing this congestion by diverting traffic at various places can be visualized if the police can visualize the effect of changing traffic patterns in a real-time manner. It is important to have cause-effect analysis of various actions so that the traffic controllers can make optimal decision.

Consider another example using the City Bikes schemes available in cities such as Lyon, Toyama, Santander, Dublin and Luxembourg. Online information provides only the current status of bicycle or free-slot availability. However, we have no idea how long this information is valid for, at what times it is more reliable than others, and for which sites in which cities this information changes too fast to be reliable. A recent tweet about a bicycle stand could have helped another user to make the right decision on where to collect, or drop off, a bicycle. There is added value to the information being generated that is not being exploited. In strategic terms, city planners may wish to expand bicycle sites or to install new ones. On what information can they make these decisions? Currently, this must be based on guesswork but with proper analyses, they can support their decision making with accurate descriptions of fine-grained usage of these resources.

### **1.1.1 Potential Application of Social Media as an Event Detection Tool**

We believe that a system for detecting and tracking events, based on social media sensors, can be useful in very different scenarios. In particular, we see the following customer groups and use cases:

- Police forces, fire departments and governmental organizations will want to increase their situational awareness picture about the area they are responsible for.

- Journalists and news agencies will want to be informed instantly about breaking news events.
- Private customers that have an interest in what is going on in their area. Here, the particular nature of Twitter and its adoption by a younger, “trendy” crowd suggests applications along the lines of, e.g., a real-time New York City party finder, to name just one possibility.

## 1.2 Introduction to Twitter Community Analysis

Social network community are group of users who are related through direct or indirect social relationships, such as following in Twitter case. The most important characteristics of these users are that they would like to share what they see, what they hear, and their point of view towards certain things. In the context of an event occurrences, such as a house fire, these users would like to share this piece of information with their friends within his/her social community, then their friends may spread this piece of information within their own communities. This information spreading within communities may cause a chain reaction, such a chain reaction may be another different clue of events detection too. As well as a way to broadcast messages across communities of users, Twitter is also a new way of direct communication between people, particularly between friends. Friendship between Twitter users is defined through the following-follower feature of Twitter. Exchanges of tweets directly between users can be realised through mentioning another Twitter user’s name, specified by prefixing a Twitter username with an @ symbol as in @exampleuser. This means that the message or tweet is either a direct message to another user, or mentions another Twitter user’s name. By analyzing the topology of follower-following relationships and conversations between users, we can derive social communities among users.

However what is the reason for one user to consider another user as his/her friend? In Java et al.’s work [30], they show that having a common topic of interest is one of the motivations behind the formation of these communities. Their work also shows that some users who act as a information source and who constantly publish tweets can be



influential figures inside their communities. But it is reasonable for us to expect that there might be other factors which may also cause the formation of such communities, such as a user's social status or mobility patterns. A smart system created based on the analysis of community structures in Twitter may have potential for urban dynamic analysis. Such techniques overcome the traditional costly and time-consuming public survey methods.

### 1.3 Other type of Sensors

Bluetooth sensing is a different type of sensing technique for smart crowd monitoring. Every Bluetooth device is identified by a unique number, called its MAC address. This unique number can be applied to identify each individual in a crowd based on the unique Bluetooth device (usually a smartphone), that each user carries with them; this can avoid repeated counting in cases of crowd number estimation. Also Bluetooth sensing devices are much cheaper than CCTV devices; such devices are illustrated in figure 1.2.



(a) Dreamplug device



(b) Bluetooth antenna

Figure 1.2: Bluetooth sensing devices

Bluetooth communication is through radio signals. With properly installed antenna the sensing range can be 150 meters, and the signal is not sensitive to weather conditions. A Bluetooth sensing technique presents great advantages over video surveillance in terms of

cost and energy consumption. However how to carry out good implementations of this new techniques is still non-trivial. This research will later briefly examine the potential of Bluetooth as a type of sensor for crowd monitoring.

## 1.4 Hypothesis and Research Questions:

The correct interpretation of social media information, especially real-time semantic analysis, imposes challenging problems to social media analysis due to the sheer volume of data in a short time and its noisy contents, such as being short in length, and with informal language format. In addition, a large part of the data are irrelevant or meaningless due to the nature of social media.

Social media data mining is still heavily relying on text retrieval. Text retrieval is a large branch of information retrieval and traditional text-based searching principles have been well founded since they started in the early 1960s. The task of text-based retrieval is to match the user query against a set of free-text records which are organized as documents like newspaper articles, web pages, video manuscripts and so on. The very successful technologies in text retrieval like term weighting [5], the Vector Space Model [61], the Language Model [57], PageRank for assigning importance based on links [54], to name a few, are adopted in many applications. Furthermore, text retrieval has been proved to be efficient on a large scale by current Web search engines such as Google <sup>1</sup>, Yahoo! <sup>2</sup>, Baidu <sup>3</sup>, Bing <sup>4</sup>, etc., in which text-based retrieval is the fundamental basis. However in the case of social media information retrieval, such traditional methods can't be applied directly because of the reasons mentioned above. In order to understand such a noisy content, we are looking for auxiliary ways to help us understand the contents of social media. "Content without context is meaningless", referenced from Jain et al.'s work [29], shows that context could be one of the ways which may help us to better interpret contents. In terms of event occurrences, "where, when and who" will be a good context for us to understand

---

<sup>1</sup><http://www.google.com>

<sup>2</sup><http://search.yahoo.com>

<sup>3</sup><http://www.baidu.com>

<sup>4</sup><http://www.bing.com>

any events related to social media content. In this research, we will concentrate on time-stamped, geolocation-embedded Twitter content, which represent the context of "where and when", and we will limit our target to Dublin Twitter users. This makes our analysis more challenging in terms of even more sparseness in data as compare to London or New York. Previous research has shown that most Twitter users have a certain level of consistency in their tweeting contents [48]. Through our observations, we found that users in our dataset have favourite locations where they like to always visit and send tweets, such places can be presumed to be where they work or live. Prompted by this intuition, we derive our first research question in this thesis.

**(RQ1:)** Is there some consistency in user's tweeting activities in certain areas of the city over time, such as regular users appearance and topic of interests?

Twitter allows users to follow other users or follow back, users form communities using this method, several interesting characteristics have been discovered within these communities and attracted a lot of research attention. One of these features is "Homophily". *Homophily* is a phenomenon showing that people's social networks "are homogeneous with regard to many sociodemographic, behavioral, and intrapersonal characteristics" [44]. In the context of Twitter, homophily implies that a twitterer follows a friend because she is interested in some topics that the friend is publishing about, and the friend follows back because she finds they share similar topical interests. Several users are grouped together because of this reason, these groups are called social media online communities, Many studies have shown that such communities are a major source of information spreading and etc. But do people form their communities just because their common interests? Are there any other factors which brings these users together? Are these users connected to each other also because they have a similar lifestyle? All of these interesting questions lead us to our second research question which is to be solved in this research:

**(RQ2:)** Do users within the same community also have similar mobility patterns (because of homophily phenomenon)?

A single user may have multiple intentions or may even serve different roles in different communities [29]. When we view the topology of our derived communities, certain users

appear to be the centre or main connections of different communities. These people should be our major concerns when we want to understand other users' behaviours, these people may be the information source, and other people's reactions may be affected by these people, etc. This leads us to another research question:

**(RQ3:)** Are users who have the most number of friends connections really more influential with high Klout4.3 score figure in Twitter?

The popularity and open structure of Twitter has attracted a large number of automated programs, known as bots, which appear to be a double-edged sword to Twitter. Legitimate bots generate a large amount of benign tweets delivering news and updating feeds, while malicious bots spread spam or malicious contents. More interestingly, in the middle between human and bot, there has emerged the cyborg referring to either bot-assisted human or human-assisted bot [18]. Information generated from these "users" is not related to what we are interested in, in order to avoid being overwhelmed by these irrelevant information, and to build a more reliable system, we need a way to filter our information, to target the origin of the information would be an easier solution.

**(RQ4:)** How can we filter out non related twitter accounts in order to enhance our system's performance?

These four research questions help us to formulate an overall hypothesis for our work, namely that "Social Media as a new way of sensing technology can work as an extension of traditional media for urban city dynamics interpretation". This hypothesis reflects the notion that we will use social media analysis technologies in our work for mining live dynamics of the urban city, such as live event detection and population demographics analysis. However, this does not mean we will give up traditional media and completely rely on social media information to accomplish the above tasks. On the contrary, social media sensing technology will be assimilated with other technologies in our work, these other technologies include video and audio surveillance, and Bluetooth sensing technologies. We use the word "extension" in our hypothesis with the meaning that social web sensing technologies can be brought into the process of event analysis in urban city dynamics interpretation and achieve satisfactory performance in event detection and population demographics analysis.

## 1.5 Thesis Structure

The above proposed four research questions and overall hypothesis are addressed in the following chapters in the thesis. The thesis expands the research questions with an overview of current research methodologies on event detection and social community analysis in social media, together with some interesting phenomena discovered from our observations. Then the development of new algorithms and the modeling of research problems are described in detail as well as the demonstration of our experimental results and application performances.

Chapter 2 gives a brief background description of state-of-the-art methodologies in social media event detection research and social media community analysis. The prevailing social media content language modeling and social media community analysis techniques are discussed to illustrate the potential benefit of social media applied to smart city management. In addition, the difference between social media and traditional multimedia are compared in terms of unexpected event detection and understanding of demographic structure. We also talk about twitter as our research target and we analyze the corresponding difficulties induced respectively.

Chapter 3 starts with an explanation of our proposed method for detection of small scale unusual event, based on geo-social regularities of Twitter user behavior, and gives details about the experiments we carried out for testing the reliability of our language models built for each of 25 partitions of Dublin city, We answer **RQ1** based on the analysis of our experimental results. We also briefly explained the future work we are planing to do for the next stage of this work.

Chapter 4 briefly talks about the work we carried out on Twitter social communities. First, we demonstrate our discovery of certain levels of the homophily phenomenon in users' mobility patterns who are within the same Twitter social community. Then we analyze the influence of users within our derived Twitter social community. Small scale experiments were carried out to support our discoveries. We answer **RQ2** and **RQ3** based on the analysis to our experimental results. At the end of the chapter, we again briefly explain the

future work we are planning to do in the next stage for identifying the correlations between a Twitter user's social relationship and their social economical background.

Chapter 5 demonstrates our observations on our users' geographical and temporal tweeting behaviours. We find that although Twitter users are more active in their favourite locations in terms of tweet generation, such as their home, workplace or leisure places, they do contribute a significant amount of tweets from random locations, and these tweets are of particular interest to us for event detection tasks. By studying the temporal tweeting patterns of our users, we can identify groups of users with similar patterns and be able to roughly estimate their social status. In the end of the chapter, we talk about our discovery of interesting correlations between Twitter users tweeting activities and population densities in the Dublin city areas.

Chapter 6 explains our experiments implementing state-of-the-art methodologies for user account type classification to our dataset, and together with some results and analysis.

Chapter 7 ends this thesis with some conclusions as well as future avenues for other research.

## Chapter 2

# Background and Related Work

### 2.1 Introduction

In this chapter we present some of the background work to our research and we provide some pointers to previous work related to our study. Social media sites (e.g., Twitter, Facebook, and YouTube) have emerged as powerful means of communication for people looking to share and exchange information on a wide variety of topics. Our interest is in cases where people use social media sites for reporting real-world events as they are happening and also forming social communities. In this research we focus on Twitter<sup>1</sup> as a social media site because of its prominence and the easy availability of data through public and open APIs.

**Background: Twitter** is a popular website with more than 500 million registered users as of Jun, 2013. Twitter’s core function allows users to post short textual messages, or *tweets*, which are up to 140 characters long. Several features play important roles on Twitter. Specifically, Twitter users can use the *hashtag* annotation format (eg.,#dublinGAA) to indicate what their 140 character posted messages are about or to capture other aspects or characteristics related to the message. In addition, Twitter allows several ways for users to directly converse with each other and to interact with other users by referencing each other in messages using the @ symbol. A *reply* is a public message from one user that

---

<sup>1</sup>[www.twitter.com](http://www.twitter.com)

is a response to another user's message. Replies start with the replied-to user @username (e.g., "@whitey252 I'm there where are you?"). A *mention* is a message that includes other username in the text of the message (e.g., "like the tour @onedirection"). Although the length of a tweet is limited to 140 characters, Twitter allows url links to be attached to each tweet, these url links provide linkage to other web contents such as news or advertisements. This link feature is well adopted by users as an extension to their Twitter contents, such as "@rtenews Educate Together to run two new schools <http://www.dublinbus.ie/en/News-Centre/General-News/Townsend-Street-Diversions/>". Twitter does not have direct support for pictures, but allows users to use the link feature to attach a url link to a third party image application, such as Flickr, Instagram etc. This picture link feature plays an important role in enriching the multimedia contents of Twitter.

Much research work has been carried out and reported in the literature utilizing the characteristics revealed by these Twitter features, including the realtime detection of live events, which is where our interest lies. Twitter messages can be constructed to reflect useful event information for a variety of events of different types and scale. In particular they can be used for unplanned events which is more challenging than planned events. For example, Twitter users live broadcast the protests in Iran [23] and the Mumbai blasts [2]. Some examples of these can be seen in Figure 2.1. Some studies on the content of Twitter around these events, carried out retrospectively can be used to demonstrate the evolution and progression of events over time [13].



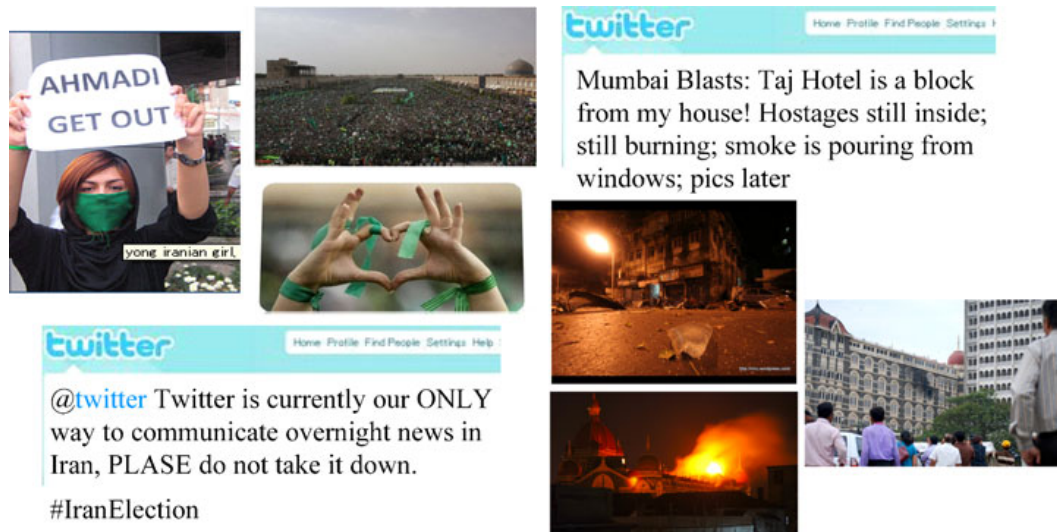


Figure 2.1: Twitter community tweets about some major events

Twitter users react to live events through tweets so fast that sometimes Twitter spreads news prior to the traditional news media [33, 60]. Often, Twitter users post messages in anticipation of an event, which can lead to early identification of interest in these events. Additionally, Twitter users often post information about local, community-specific events (e.g., a local flood), where traditional news coverage at a regional or national level is low or non-existent. It is this type of event with which we are concerned in this research.

The huge potential of utilizing Twitter as a new type of event sensing media has been adopted by many large projects. These projects incorporate Twitter as an assistant to traditional event surveillance technologies in urban city areas. These include the *IBM Smart City Project*<sup>2</sup> being carried out in Dublin, Ireland, and the *SAGACITY project*<sup>3</sup>. One important objective of the SAGACITY project is to design a generic and modular architectural framework for harvesting social media data, and mining and linking events and activities which can be detected from this social media data. This platform will support adding new modules for integrating new data sources, analysis and event mining tools. Twitter is integrated into the SAGACITY project as one of these seeking event mining tools. In the next section, we look more closely at how event detection in social media has been carried out

<sup>2</sup>[http://www.ibm.com/smarterplanet/us/en/smarter\\_cities/overview/](http://www.ibm.com/smarterplanet/us/en/smarter_cities/overview/)

<sup>3</sup><http://cordis.europa.eu/fp7/ict/>

by other researchers in the field.

## 2.2 Event Detection in Social Media

Event detection has long been a research topic across many application areas and using any sources of data or information [71]. The topic detection and tracking (TDT) and news event detection task carried out as part of the annual TREC benchmarking activity at the National Institute of Standards and Technology (NIST) [6] was a notable collective effort to discover and organize news events from a continuous stream of text information (e.g. newswire, radio broadcast). The TDT track of TREC pre-dated the emergence of Twitter and so social media wasn't used as a data source, but the TDT work represents one of the first coordinated approaches to event detection.

This early work leveraged natural language processing tools, such as named-entity extraction for online news event identification. Such tools work well on well-structured text like newspaper articles and TV broadcasts, but do not perform well over social media contents due to their heterogeneous and noisy nature. To tackle this problem, other methods have been proposed by researchers. **Twitterstand** [62] gathers and disseminates breaking news from Twitter, and uses an online clustering method to cluster similar Twitter messages, and a naive Bayesian classifier to deal with the noisy nature of Twitter contents. Sakaki et al. [60] classify Twitter contents using a Support Vector Machine [19] based on proposed features. Event detection in textual news documents has also been studied in depth. Looking at text stream data from social blogs and email, Zhao et al. [73] detect events using textual, social, and temporal document characteristics. However compared to Twitter, social blogs and email have much better quality content in terms of document length, language quality and overall linguistic coherence. Various methods have been proposed in recent research which carry out real-time event detection tasks using social media, Twitter in particular, such as streaming first story detection [55] and breaking news detection [33]. **Twitcident** [3, 4] enables filtering, searching, and analyzing Twitter information streams during incidents as they are happening. The system listens to a broadcast network which provides

information about incidents. Whenever a new message comes in, it searches for related tweets which are semantically extended in order to allow for effective filtering. Users may also make use of a faceted search interface to dive deeper into these tweets. Other work [9, 63, 10] also reports work carried out for real-time event detection from Twitter based on temporal and textual features of tweets on the Twitter social network.

These previous works successfully developed detection of breaking news or live events in Twitter streams. Their methods are sensitive to large scale events such as the Presidential inauguration in the USA. This is because they rely on the fact that the target events are able to generate significant boosts from among the main stream of Twitter. Our research concerns events that do not occur in a global setting but in a small city area, Dublin city in our case, and are thus of a much smaller scale, much more local and focussed, such as local floods or a local party.

In recent years, along with the popularity of geolocation enabled smart devices, social media contents have more multimedia information integrated, such as GPS coordinates. Social media users are able to not only report events in real-time, but also to provide the locations of where the events occurred. So, event detection techniques should be extended to incorporate these new features of social media information in addition the functionality of the detection platform should be expanded with the capability to measure the scale of the target events in terms of space and temporal duration. Several studies have been done in geo-tagged social media mining for event detection. For analysis of geo-social characteristics with blogging sites, Moriya et al. [49] developed a system that estimates images, impressions, or the atmosphere felt by bloggers about a region from texts, in relation to geographic information provided by blogs, and the resultant analysis displayed the results on a digital map. This work is similar to our approach in terms of geo-social analysis, but our targeted media are much noisier in terms of textual contents as mentioned above. Rattenbury et al. [59] use a Scale-structure Identification method to extract place and event semantics for tags based on GPS metadata of images in Flickr<sup>4</sup>, however, in our work we did not attempt to aggregate social media documents.

---

<sup>4</sup><http://www.flickr.com/>

In X. Liu et al.'s work [39], they proposed a framework which uses multi-modality features such as text, time, visual features and "owner" metadata to correlate Flickr images with events. Also in their other work [40, 41], they present a method combining semantic inferencing and visual analysis for finding automatically media (photos and videos) illustrating events. This previous work achieved good results in event detection on Flickr, but the detection is for events which are mainly retrospective, detecting events from past data whereas the aim of our research is real-time event detection.

Other methodologies used in the literature for event detection involve identification of bursts in the time and frequency domain. In [24], the authors apply *Discrete Fourier Transformation* (DFT), which converts the signals from the time domain into the frequency domain. A burst in the time domain corresponds to a spike in the frequency domain. In their later work, they used a Gaussian Mixture model to estimate the time period of which the event burst happen. *Wavelet analysis* is used to build signals for individual words, event detection with clustering of such signals is carried out in Twitter data [69] and Flickr data [16]. In Weng and Lee's work [69], they address the challenge of constructing a signal for each word occurring in Twitter messages using wavelet analysis, thereby making it easy to detect bursts of word usage. Frequently recurring bursts can then be filtered by evaluating their auto-correlation. The remaining signals are cross-correlated pairwise and clustered using a modularity-based graph partitioning of the resulting matrix. Due to the quadratic complexity of pairwise correlation, they rely on heavy pre-processing and filtering to reduce their test set to approximately 8,000 words. As a result, they mainly detected large sporting events, such as soccer world cup games, and elections.

Sakaki et al. [60] present an approach that gathers tweets for target events that can be defined by a user via keywords. The authors apply classification and particle filtering methods for detecting events, e.g., earthquakes in Japan. However their targeted events are too specific, their users have to describe the events with event specific keywords, and this leads to a bad extensibility to their system. What all this work represents is a considerable effort in building and using ontologies in the task of event detection in social media. Mostly, ontologies have been useful assets in the detection task, but their drawbacks are in the large

effort needed to construct them, and the fact that there isn't a single best way to use them.

Other interesting work has also been carried out by researchers in the field related to the relationships between events and social media. In Liu et al.'s most recent work [38], they addressed the problem of organizing media data by events. In their work, they reported the study of both feature selection and handling missing value in the scope of event based media categorization.

Identifying events in real-time on Twitter remains a challenging problem. Particularly in our case, our research target location is much smaller compared to larger areas like New York or London in terms of Twitter volumes, and our target events are also at a small scale, such as local floods, traffic accidents etc. This makes creating a system for detecting these patterns and events even more difficult. The methodology proposed in this research is broadly related to 3 categories: topic modelling, text retrieval and text classification.

**(a) Topic Modelling** One important part of our event detection method used in this research is identification of geographical topic of interests. We model a location by its social media users' topic of interests. Topic modeling is a classic problem in text mining. The most representative models include PLSA [25] and LDA [12]. Wang et al. [68] use an LDA-style topic model to capture both the topic structure and the changes over time. In these studies, they do not consider the location information of the documents, so they do not focus on geographical topics. In [67], Wang et al. propose a Location Aware Topic Model to explicitly model the relationships between locations and words, where the locations are represented by predefined location terms in the documents.

Mei et al. [46] proposed a probabilistic approach to model the subtopic themes and spatiotemporal theme patterns simultaneously in weblogs, where the locations need to be predefined. However, in geographical topic discovery, we do not know the locations or regions of interest beforehand. If we directly use the administrative region partitions, it would be difficult to discover topics whose corresponding regions are not aligned well with the pre-segmented regions. In [45], Mei et al. proposed a model called NetPLSA to combine PLSA with a graph-based regularizer, where adjacent nodes in a document similarity

graph should have similar topic distribution. The techniques proposed in these previous works will be used for testing the performance against our method.

**(b) Text Retrieval** Later in this thesis we describe how we build formal mathematical models using a language modelling approach, to model the distribution of words among social media data. Once the language models are built from the social media contents, the next step is event related content retrieval. The retrieval model implemented in this research is an enhanced version of the model described in [47], which combines the language modeling [58] and inference network [20] approaches to information retrieval.

**(c) Text classification** Another important task of event detection in social media is text classification. This classification clusters texts with similarity in semantics and furthermore identifies the topic of the clustered contents. Existing work on classification of short text messages integrates messages with meta-information from other information sources such as Wikipedia and WordNet [8, 27]. In classification of short texts such as Twitter, Sriram et al. [64] use a small set of domain-specific features extracted from a user's profile and text. Their proposed approach effectively classifies the text to a pre-defined set of generic classes such as news and events. Sakaki et al. [60] carried out semantic analysis using Support Vector Machines [19] based on proposed features; we proposed a set of 4 features for tweet type classification: presence of personal pronouns, emphasis on words, presence of slang words, and presence of non-ASCII keywords. In Hila et al. [11]'s work, they proposed features to classify whether a tweet is event-related or not. Some of these proposed methods will be used in our research.

Apart from event detection, researchers are also interested in the social networks which are generated between social media users. Research has shown strong correlations between users' online social media communities and their social status in the real world. This is another research topic of this research. In the next section, we will provide a review for the work reported in this literature.

## 2.3 Social Network Community Analysis and Population Demographics

**Background** Social networks represent the links between a set of entities connected to each other via different types of relationships. In the case of Twitter, which is a recently-emerged way of communication between people, users share status between each other, particularly between friends. Friends in Twitter setup their connections by *following* each other, and communicate by *mentioning* the name of their friends. Twitter messages going back and forth between people are similar to making phone calls, but for different purposes, such as friends sharing their feelings through Twitter in their friends circle. In this section we will review some research work studying the topological and geographical properties of Twitter's social networks, this work is closely relate to our study.

A fundamental property of social networks is that people tend to have attributes similar to those of their friends. There are two underlying reasons for this:

- First, the process of social influence leads people to adopt behaviours exhibited by those they interact with; this effect is at work in many settings where new ideas diffuse by word-of-mouth or imitation through a network of people.
- Second, people tend to form relationships with others who are already similar to them. This phenomenon, which is often termed selection, has a long history of study in sociology.

@ Aristotle "People love those who are like themselves" @ Plato "Similarity begets friendship" @ Lazarsfeld & Merton "Birds of a feather flock together" [53].

### 2.3.1 Homophily

*Homophily* is the principle that a contact between similar people occurs at a higher rate than among dissimilar people [44]. This phenomenon is mainly studied by sociologists. Recently, this idea was introduced by researchers to the analytics of social networks in

social media such as blog posts and Twitter. In a survey of bloggers, Nardi et al. [50] describe different motivations for "why we blog". Their findings indicate that blogs are used as a tool to share daily experiences, opinions and commentary. Based on their interviews, they also describe how bloggers form communities online that may support different social groups in the real world. Lento et al. [35] examined the importance of social relationship in determining if users would remain active in a blogging tool called Wallop. A user's retention and interest in blogging could be predicted by the comments received and continued relationship with other active members of the community. Users who are invited by people with whom they share pre-existing social relationships tend to stay longer and active in the network. Moreover, certain communities were found to have a greater retention rate due to existence of such relationships.

Mutual awareness in a social network has been found effective in discovering communities [37]. Weng et al.'s work on Twitter user relations reported that two users who follow reciprocally share topical interests by mining their 50 thousands links [70]. Kwak et al.'s work found a non-power-law follower distribution [33] in Twitter. In Java et al.'s work [31], they identified different types of user intentions and studied community structures. Other research has used the social network derived from exchange of telephone calls made between users to successfully predict these users' social economical status. In chapter 4 of this thesis, we will use the social communities derived from social networks in Twitter to predict the users' mobility patterns.

## **2.4 Summary**

In this chapter we presented a high-level knowledge background for event detection from social media information sources. An overview of aspects for social network community analysis was also discussed in this chapter together with related work. As a new form of multimedia, social media has its own characteristics compared to traditional media such as broadcast TV, in modality, content quality, information diversity, etc. We take Twitter as our research target and we analyzed the corresponding difficulties induced respectively.



In the next chapter we will further elaborate on how semantic concepts contribute to understanding events in social media and in particular how the combinations of irregularity in social media user's behaviour can be interpreted as events. In the rest of the thesis, we will provide details of our work in developing our approaches to event detection and dealing with such issues as semantic inconsistency detection and information filtering, as well as enhancement of language modeling. We also give details to our analysis on correlations between social networks of Twitter users and Twitter user mobility patterns. During the description, state-of-the-art technologies will be compared and our further work plans discussed within details of experiments and evaluation.

## Chapter 3

# Event Detection in Social Media

### 3.1 Introduction

Recent advances in technology have enabled social media services such as Twitter to support space-time indexed data, and internet users from all over the world have created a large volume of time-stamped, geo-located data. Such spatio-temporal data has immense value for increasing situational awareness of local events, providing insights for investigations and understanding the extent of incidents, their severity, and consequences, as well as their time-evolving nature.

Event detection using geo-tagged Twitter data has attracted much research interest, such as Sakaki et al. [60]. In this work, they present an approach that analyzes tweets related to natural disasters, such events can be defined by Twitter users using specific keywords in Twitter contents. They apply classification and filtering methods for reporting the status of natural disasters, such as predicting the center of an earthquake and predicting the trajectory of a typhoon. Their method presented convincing results, but their target events are too specific, therefore their system does not have good scalability. In Lee et al. [34]'s work, they proposed a geo-social event detection method based on the geographic regularity which reflects a geographic region's usual status through crowd behaviour observable on Twitter. The methodology they proposed performs well for detecting large scale events, such as festivals which happened in Japan. Both of the above event detection methods rely

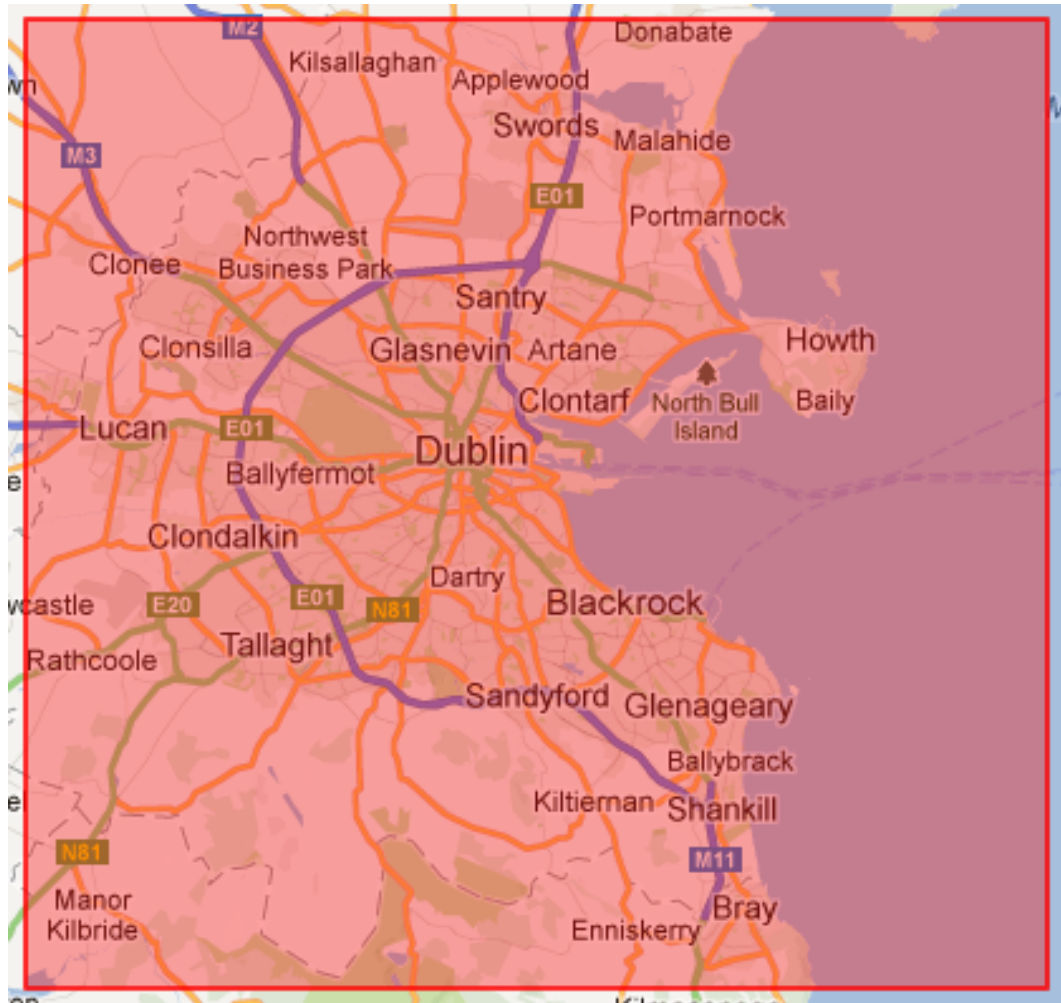


Figure 3.1: Bounding box covering research target area

on the fact that the target events are able to attract the attention of a large number of Twitter users, and are able to cause a significant increase or decrease on the main stream of Twitter. The type of events we discover in our work are often on a rather small-scale and localized, they happen at a specific place in a given time period, such as a house fire, traffic jams, or local flood. We are interested in working on a relatively small data-set, Dublin city in this case, which is only approximately 1% of the size of New York in terms of the number of tweets generated per day. Figure 3.1 shows the boundary of our target Dublin city area. Because of the problems mentioned above, the event detection task is very challenging. In this work, we present a geo-social event detection method based on geographical regulari-

ties of local crowd behaviour through Twitter. For the purpose of the detection of unusual socio-geographic events, we first decide what the usual status of local crowd behaviour in a geographical region is in terms of Twitter. After mapping the geo-tagged tweets onto relevant locations on a map, we focus on the following points: First, a sudden increase or decrease in the number of tweets happening in a geographical region which can be an important clue to an unusual event happening. Secondly, the increasing number of non-regular visitors in a geographical region for a short time period may indicate the occurrence of a local event. Thirdly, inconsistency in Twitter content in the region can be another important clue of unusual events happen in the region. Our aim is to build an alert system based on the detection results. This system can raise alarm in almost real-time for what is going on in the city, in other words, "sensing the city pulse". Potential users of such a system are:

- Police forces, fire departments and governmental organizations to increase their situational awareness picture about the area they are responsible for.
- Journalists and news agencies to instantly be informed about breaking events.
- Private customers that have an interest in what is going on in their area.
- Event organizers that want to understand event participant behaviour.

## **3.2 Twitter and Geotagging**

*Twitter*, as described in the Introduction, is a micro-blogging service which allows users to share 140 character messages, also known as statuses and tweets. Users are automatically shown the tweets of other users who they "follow". They can also keep track of conversations by searching for topics or usernames of interest. Status updates can be either publicly available or restricted to a user's connections. Users can make status updates on the Twitter website, or using one of many applications that interface with Twitter. Twitter has many mobile users, including some who use GPS-enabled devices to geotag their tweets. It is also possible to allow Twitter to access browser location information to geotag a user's tweets, however the tagged locations are only at a high level, such as Dublin city, or southern city.

Because our research target area is at a city level, tweets sent through browsers are not considered in this work. According to popular digital advertising website *eightytwenty*<sup>1</sup>, around 5% of tweets generated in Ireland are geotagged, we would expect higher ratio in Dublin city area in terms of the higher ratio of smart phone users. Particularly in our dataset, we collected 387,800 tweets over one month period from Dublin area, that is about 13,000 per day, this is a significant amount considering the Twitter population size of Dublin. Application developers have two options for attaching geotags to tweets: they can include the latitude and longitude of the tweet, or they use Twitter's reverse geocoding function to include a description of a place, for example at the neighbourhood level. Our analysis makes use of those tweets which are tagged with the user's coordinates. There is a Twitter-specific syntax which will later be taken into account in building language models for different part of a city. Tweets can contain mentions of usernames, specified by prefixing a username with an @ symbol as in "@RTE1". Tweets can be tagged with a topic or other annotation, by prefixing a tag with a hash to make a "hashtag" e.g. *#twitterapi*. Twitter users can also "re-tweet" an other user's status updates to relay a message to their own followers, by prefixing a message with the "RT username:", or by clicking a "re-tweet" button, which results in the tweet's metadata showing it as a retweet. Retweets are just a duplicate of the original tweet, so we do not consider these either. Table 3.1 lists the five most commonly occurring sources of geotagged tweets, sampled within a month period. A source is the service such as a website or application from which the user sent the tweet. Some services have the purpose of providing information about the location of the user at the time the tweet was issued. For example, Foursquare allows users to "check-in" at a venue to win points. A check-in results in the creation of a tweet containing location information such as "I'm at the Auld Dubliner (Dublin)". For example, Foursquare<sup>2</sup> is the most popular location-oriented Twitter application, and has been used for analysis of user spatio-temporal behaviour [51], but other location-based services provide similar functionality, such as Instagram<sup>3</sup> which allow users to upload pictures etc. But through our observations, the majority of the tweets generated

---

<sup>1</sup>[www.eightytwenty.ie](http://www.eightytwenty.ie), date accessed 01-06-2013

<sup>2</sup><https://foursquare.com/>

<sup>3</sup><http://instagram.com/>

through these check-in services are automatically generated contents, they do not provide useful information in their contents, such as the Foursquare generated tweets given above, so tweets from these sources are not considered in this work.

Table 3.1: Most popular Geotagged Tweets source

Device	% of tweets
Iphone	67%
Android	25%
Windows	4%
BlackBerry	2.4%
Mobile Web	1.4 %

### 3.3 Discovering Socio-geographical Boundaries

To detect unusual local events for a given large area, Dublin city area in this case, we first need to determine how to partition the target city area into sub-areas by establishing socio-geographic boundaries. In order to configure socio-geographic boundaries conveniently, we adopt a clustering-based space partition method that can reflect a geographical distribution of a dataset and better deal with heterogeneous regions differently. Some research work divided the target area into equally sized grids with different granularity [16]. There are a few reasons we chose not to use this approach. Firstly, the adequate cell size is very difficult to determine. For instance, if we split a region into excessively small cells, most suburban areas will consume considerable unnecessary monitoring costs, even though the probability of tweet occurrence is generally very rare. Secondly, since this approach does not consider the geographical distribution of tweets, the balance over the target region becomes inefficient and consequently results in poor detection performance. On the other hand, partitioning on the basis of administrative districts also has a weakness since we cannot determine whether crowd activity regions are strongly relevant or almost dependent on the administrative districts. In addition, if two neighbouring districts are strongly connected to each other in terms of social crowd activities, simply splitting them into two different groups will not be a good choice.

In detail, we adopt the K-means clustering method [43] based on the geographical occurrences of our dataset. The K-partitioned regions are demonstrated in different colours onto a unit graph, as shown in Figure 3.2. As a result, we achieve an appropriate socio-geographic boundary setting for the target region by distributing the actual occurrences of tweets. The city of Dublin is partitioned into 25 regions empirically, 25 is decided using elbow method as shown in Figure 3.6, experimental results proved that 25 is the best solution for our case. By comparing the partition results to the actual population distribution of the Dublin city area according to the Central Statistical Office data, as in Figure 3.3, the partition results are acceptable. Hot spots can easily be identified, such as city center area, where there are high population density and high volume occurrence of tweets. In addition, some low population areas which with high volume occurrence of tweets can be identified, such as Dublin Airport and the Phoenix Park. Each polygon area in Figure 3.3 is called a Small Area. *"Small Areas are areas of population comprising between 50 and 200 dwellings created by The National Institute of Regional and Spatial Analysis (NIRSA) on behalf of the Ordnance Survey Ireland (OSi) in consultation with CSO. Small Areas were designed as the lowest level of geography for the compilation of statistics in line with data protection and generally comprise either complete or part of townlands or neighbourhoods. There is a constraint on Small Areas that they must nest within Electoral Division boundaries. Small areas were used as the basis for the Enumeration in Census 2011. Enumerators were assigned a number of adjacent Small Areas constituting around 400 dwellings in which they had to visit every dwelling and deliver and collect a completed census form and record the dwelling status of unoccupied dwellings. The small area boundaries have been amended in line with population data from Census 2011".*<sup>4</sup>

---

<sup>4</sup><http://www.cso.ie/en/census/census2011boundaryfiles/>

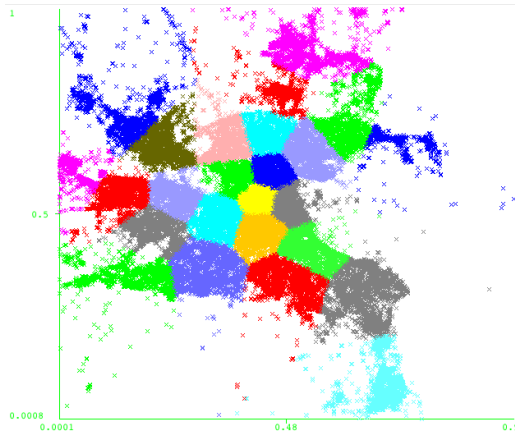


Figure 3.2: Geo-social partitioning of Dublin into 25 clusters

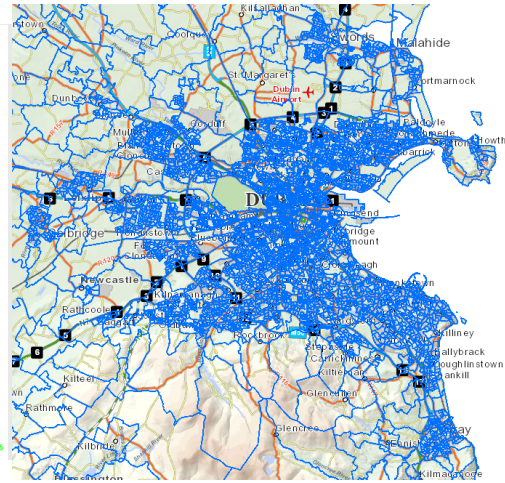


Figure 3.3: Population distribution of Dublin in Small Areas

**Major Assumption** Every Twitter user has his/her topic of interests, and favourite locations that he/she always likes to visit and send tweets, such as places they work or live. Therefore for each location there is some consistency in Twitter activity, such as regular users appearance and topic of interests over time.

### 3.4 Geo-Social Event Detection

Our assumption is that we can recognize local events from any inconsistency in Twitter user behaviour in the location. In this section, we address the process of our proposed event detection method in detail. We first describe a platform which is designed to realize the proposed method based on the two critical functions: (1) geographic regularity construction and (2) final event detection. In the following subsections, we will explain the process separately from the configuration of socio-geographic boundaries to the final detection.

We present a summary of the status of these user behaviours:

1. The number of tweets sent in the zone in a given time window
2. List of regular users (visitors) in a given location
3. Language model for representing the semantic consistency of all of the tweets from



a location

### **3.4.1 Measuring Geographical Regularities**

This section explains how we set up the measurements of regularity and how we derived our city partitions.

#### **3.4.1.1 Regular Twitter Activity**

**Number of Tweets (NT)** Within each partition of the city, there are a number of tweets generated over time. In our work we analyze weekday and weekend days differently. This is because certain partitions have different activities for weekday vs weekend day. Examples would be partitions covering business areas, such as industrial estates, which will be relatively quiet during weekdays, but partitions covering shopping areas will be much more active during the weekend. The regularity of the total amount of tweets is calculated using the average of each day during a month period, with  $\pm 1$  standard deviation. The results are assigned into hourly bins, any number outside the 1 std will be considered as an unusual activity, as shown in Figure 3.4.

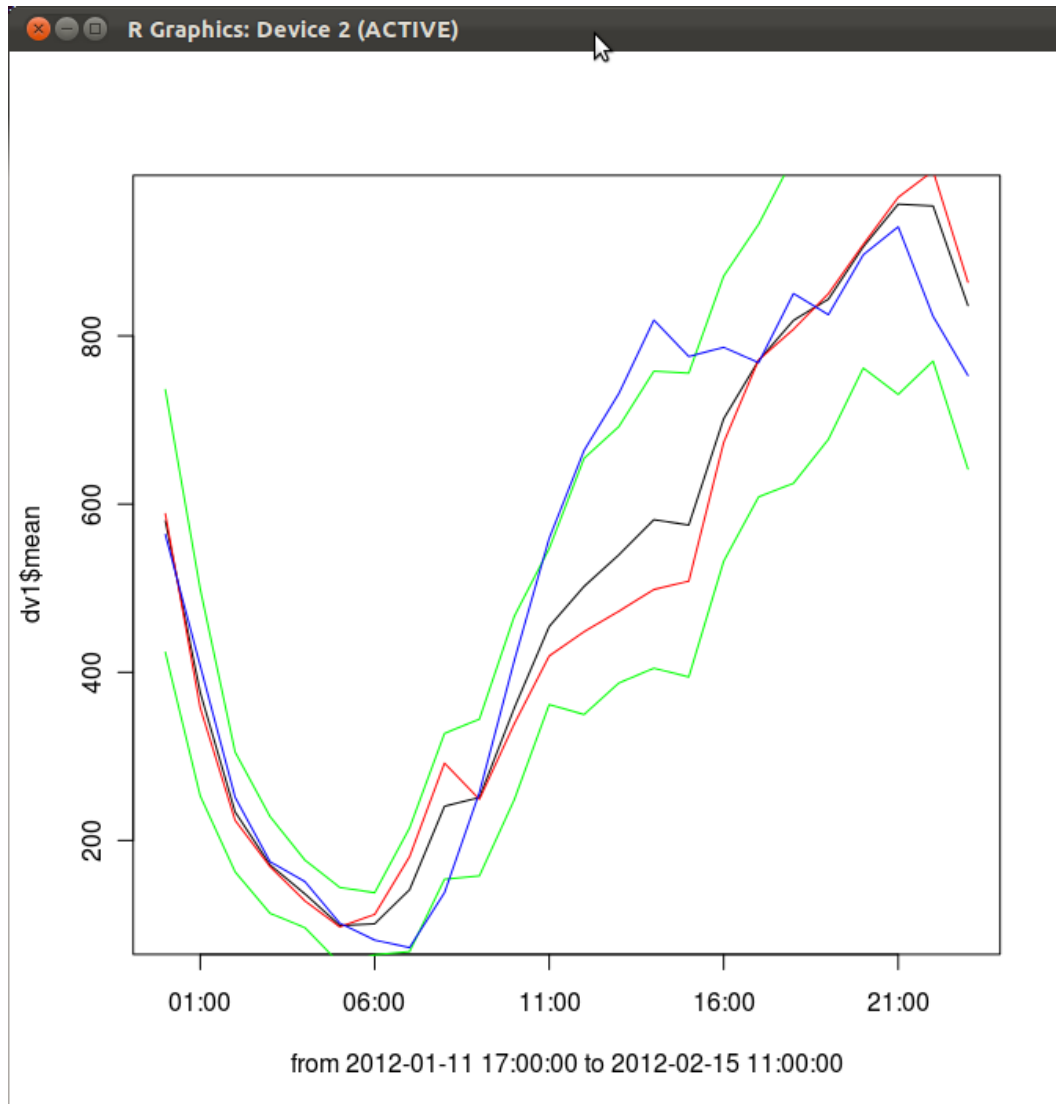


Figure 3.4: Twitter occurrences in hourly bins

**Number of Users (NU)** For every partition we keep a list of regular Twitter users, these users are constantly active inside the partition. If there are many unseen Twitter users (visitors) sending tweets in the partition, we consider this as another clue of irregular Twitter activity.

**Semantic Regularity (SR)** Measurement of semantic regularities is explained in the following sections.

### 3.4.1.2 Locations as bags-of-words

For each tweet in our data collection, we keep the following sources of information: a Twitter ID, a pair of geographical co-ordinates (latitude and longitude), tweet text, a timestamp of when was the tweet was generated, a set of tags (mention tags and hash tags), the source from which application the tweet was generated. After we identify the partitions of the city area, we place each tweet into the partition, then use all of the texts in each partition to derive a language model that represents the semantic consistency of the location. In order to preserve the semantics of the tweet contents we do not apply any stop-word filtering, special characters are removed, such as "#" and "@". Although special characters such as emoticons can be useful for sentiment analysis, we do not consider them in this work, this may be a subject for further analysis in future work.

### 3.4.1.3 Modeling The Language of Locations

We use the language modeling approach as described in Ponte and Croft [58] to build individual models of locations for each of the partitions in the city. For each location, i.e. for each of the city partitions, we estimate a distribution of terms associated with that location. We can then estimate the probability that a new tweet was issued from a given location by sampling from the term distribution for that location. The locations can then be ranked by the probability that they “generated” the tweet. More concretely, given a set of locations  $L$ , and a tweet  $T$ , our goal is to rank the locations by  $P(L|T)$ . Rather than estimate this directly, we use Bayesian inversion:

$$P(L|T) = \frac{P(T|\theta_L)P(L)}{P(T)} \quad (3.1)$$

where  $L$  is the model of the location. Assuming independence between terms:

$$P(T|\theta_L) = \prod_i P(t_i|\theta_L) \quad (3.2)$$

The probability of a term, given a location,  $P(T_i|\theta_L)$ , is estimated using Dirichlet smoothing [72]:

$$P(t|\theta_L) = \frac{c(t, L) + \mu P(t|\theta_C)}{|L| + \mu} \quad (3.3)$$

where  $\mu$  is a parameter set empirically,  $t, L$  is the term frequency of a term  $t$  for location  $L$ , and  $|L|$  is the number of terms in location  $L$ . In this work we assume the prior probability of the locations,  $P(L)$ , is distributed uniformly. We ignore  $P(T)$ , since it is the same for all locations, and thus does not affect the ranking. The locations can be ranked directly by the probability of having "generated" the tweet, or they can be ranked by comparing the model yielded by the tweet to the mode of the location using Kullback-Leibler (KL) divergence. In this thesis we use both methods for ranking locations. When ranking by KL divergence, we let  $\theta_T$  be the language model for the tweet  $T$  and  $L$  be the language model for the location  $L$ . Then the negative divergence from the query language model to the document language model is:

$$KL(\theta_T|\theta_L) = \sum_i P(t|\theta_T) \log \frac{P(t|\theta_T)}{P(t|\theta_L)} \quad (3.4)$$

where  $t$  is a term. The KL divergence is smoothed according to:

$$KL(\theta_T|\theta_L) = \sum_i P(t|\theta_T) \log \frac{P(t|\theta_T)}{\alpha P(t|\theta_L)} + \log(\alpha) \quad (3.5)$$

where:

$$\alpha = \frac{\mu}{\mu + |L|} \quad (3.6)$$

In this work we use the Lemur Toolkit [7] for building our language models and carrying out our experiments.

#### 3.4.1.4 Reliability of Our Language Models

As mentioned above, we generate language models for each partition of the city using tweets generated from that area. We rely on these language models to detect irregularities in semantics, but how reliable are they? How accurate can our language models represent the consistency in the topics of interest in the location? Can we maximize the reli-

ability of the language model? To address the above questions, we carried out experiments to test our language models by trying to predict the generation of location of tweets, then using the prediction accuracy to test the language model's reliability. We explore different ways to improve the accuracy of the prediction such as doubling or eliminating the weight for hash tags and mention tags. Experimental results show that removal of hash tags do not improve the accuracy of the prediction, however double mention tags does improve the accuracy. Details are explained in the experiments and results section.

#### **3.4.1.5 Semantic Irregularity**

In order to identify semantic irregularity in each partition, we setup a small time window of 1 hour, for instance. If there are a number of tweets which are generated in this partition in the time window, then these contents are compared to all of the language models defined for each of the 25 partitions, and the similarity is measured using KL-divergence. If the actual partition is not in the top ranked predictions, then we consider these tweets are semantically irregular, and should be used as another clue for the occurrence of unusual events.

#### **3.4.2 Event detection**

In our research, we want to detect geo-social events that result in unusual Twitter user behaviour. For this, we define that a socio-geographic boundary is under an unusual condition when its indicators, Number of Tweets (NT), Number of Users (NU) and Semantic Regularity (SR) satisfy the following equation:

$$F = \alpha NT + \beta NU + \gamma SR \quad (3.7)$$

In equation 3.7,  $F$  is a measure for the scale of an unusual event,  $\alpha$ ,  $\beta$ , and  $\gamma$  are coefficients for normalizing the measurements of each regularity. If  $F$  is over a fixed threshold, we predict that it is an indication of an unusual event happening. The value of  $F$  is determined empirically using historical event related data.

## 3.5 Experiments and Results

In this section we describe the experimental setup for testing the reliability of our language models through predicting the locations where a tweet was made. We will first discuss the Twitter data-set we used, followed by the evaluation measures adopted. The results are tested against the true location of where the tweet was generated.

### 3.5.1 The Twitter Dataset

To evaluate the proposed models, we crawled geo-tagged Twitter messages through the Twitter Streaming API <sup>5</sup>. We set up a bounding box which covers the Dublin area with north east corner of NE (53.489679, -5.946350), and south west corner of SW (53.174765, -6.502532) as shown in Figure 3.1. We crawled over a time period starting from 24/Jan/2013 to 19/Mar/2013. We kept only English tweets with exact geo-locations attached to the message. Our dataset ended up with 387,800 tweets in total. The first task was to map each geographical co-ordinate to a location on the map of Dublin city. In order to do this we implemented K-means partitioning. Each of these universal geographical co-ordinates (UGC) was projected onto unit length distance to the SW corner of the bounding area, the geo distance was calculated using the Haversine formula <sup>6</sup>, then normalized into unit distance. For example: (53.3297976,-6.2581027) was projected into a point as (0.4372,0.5627), as shown in Figure 3.5; then this distance was used to calculate the partitions of the city area.

---

<sup>5</sup><https://dev.twitter.com/docs/streaming-apis/parameters#locations>

<sup>6</sup>[http://en.wikipedia.org/wiki/Haversine\\_formula](http://en.wikipedia.org/wiki/Haversine_formula)

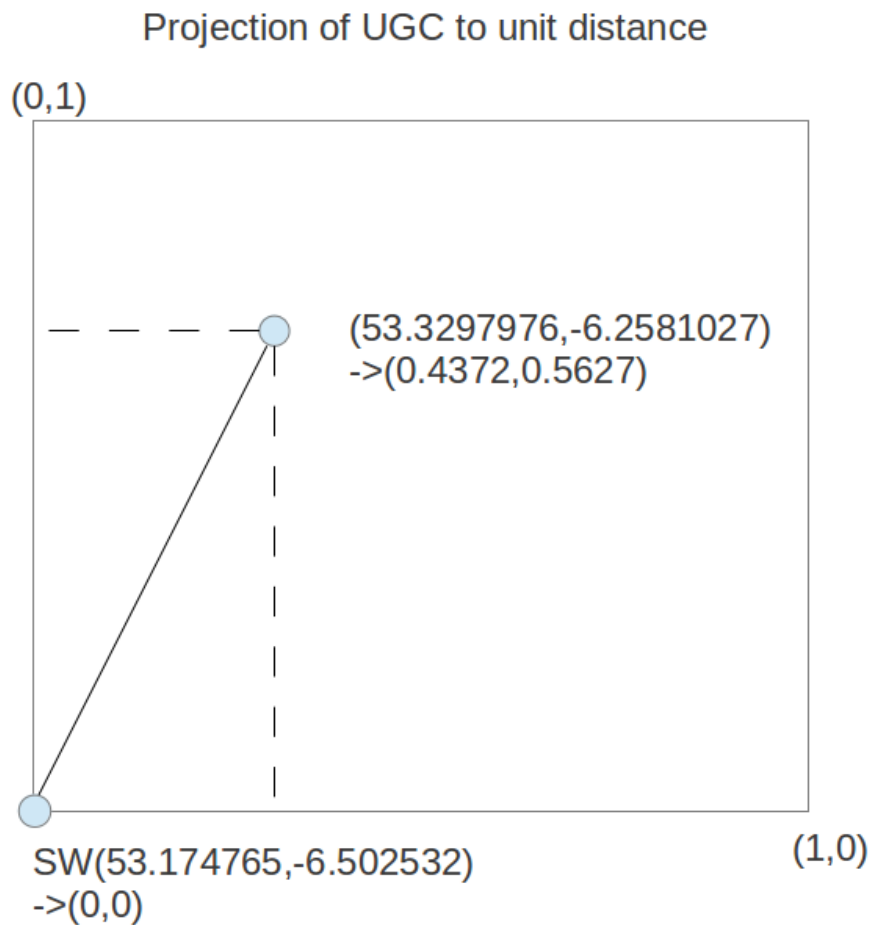


Figure 3.5: UGC projection to unit length distance

### 3.5.2 Evaluation Measures

The main metric that we use for the evaluation and for tuning parameters on our training data is location accuracy (**Acc**), which calculates the percentage of correct predictions over all test examples. We also analyze using additional measures of prediction quality, namely **Mean Reciprocal Rank (MRR)**. MRR is a statistical measure for evaluating any process that produces a list of possible responses to a sample of queries, ordered by probability of correctness. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks

of results for a sample of queries  $Q$ .<sup>7</sup>

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \quad (3.8)$$

MRR measures the ability of the system to find the actual location of a tweet among its top recommendations. The Lemur toolkit returns a list of predicted partitions, the top one of the returned list is considered as accurate, and very likely there are different partitions with a similar topic of interests and also close in space. So it is reasonable for us to assume that if the predicted partition is close in space with the actual locations, we can also consider it as accurate, but the results are calculated using MRR as mentioned above.

### 3.5.3 Evaluation

In order to test how well our language model can represent the consistency of the partition, we carry out an evaluation by comparing the predicted location to the actual location. The prediction accuracy is computed using 10-fold cross validation. We use the Weka 3 toolkit<sup>8</sup> to carry out our experiments. The targeted city area is partitioned into 25 zones using the elbow method<sup>9</sup> as illustrated in Figure 3.6.

---

<sup>7</sup>[http://en.wikipedia.org/wiki/Mean\\_reciprocal\\_rank](http://en.wikipedia.org/wiki/Mean_reciprocal_rank)

<sup>8</sup><http://www.cs.waikato.ac.nz/ml/weka/>

<sup>9</sup>[http://en.wikipedia.org/wiki/Determining\\_the\\_number\\_of\\_clusters\\_in\\_a\\_data\\_set](http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set)



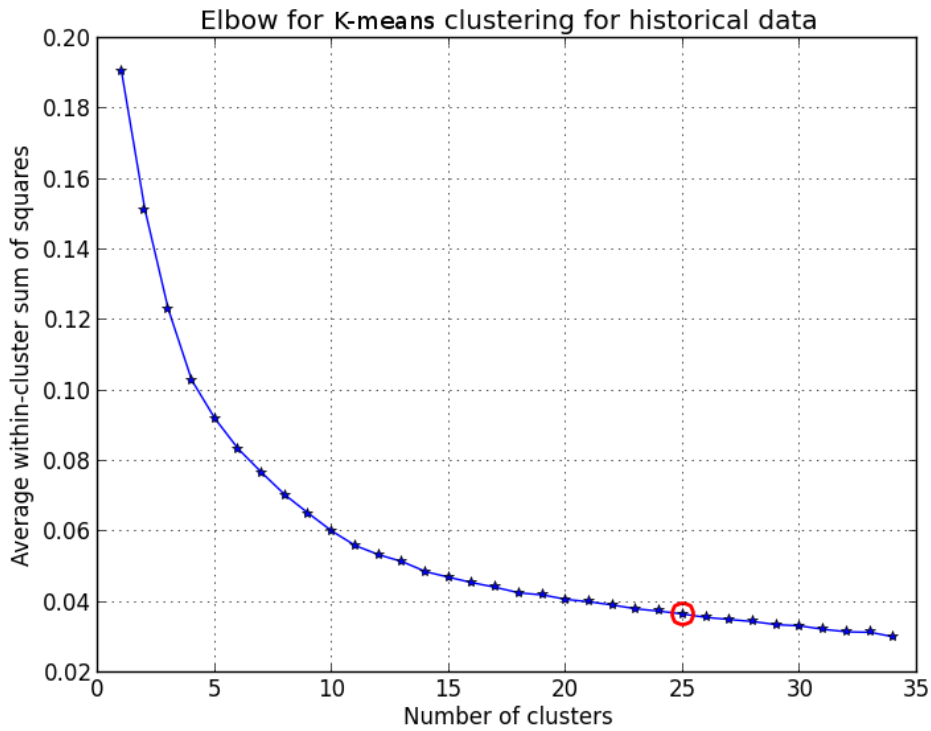


Figure 3.6: Elbow method for K-means clustering results

We performed no stopword removal, and each tweet text was stemmed by the Krovetz stemmer, with default Dirichlet smoothing with  $\mu = 2500$ . We also performed some manipulation of the features in the tweets text, including removal/double the occurrences of hashtags and mentions tags. Double or removal of entities from text contents are commonly used techniques in text retrieval research. In this thesis, we considered hashtags and mentions tags as such entities, and performed above experiments to see if the same techniques can be applied in Twitter case. Double of entities involved adding a same entity after each one, and removal of entities involved deleting of each entity from tweet content. The results are shown in Table 3.2.

As can be seen from the results, removal/double of hashtags from the text body, makes almost no difference to the prediction accuracy, this may be because the hashtags are completely independent from the Twitter text contents. However removal of mentions tags

Table 3.2: Accuracy at different level of MRR

type	Acc	MRR
with all tags	0.3124	0.4110
no mention tag	0.1481	0.2396
double mentiontag	<b>0.3347</b>	<b>0.4290</b>
double hashtag	0.3171	0.4166
no hashtag	0.3110	0.4085

caused a 20% drop in the prediction results, and doubling mention tags achieved the best accuracy at 42.9% MRR. Mention tags plays an important role in the contents of the tweets in building language models. 42.9% shows good validation of our assumption that there is good consistency in our partitioning of the city areas.

### 3.6 Analysis And Future Work

Based on our experimental results in this initial part of our work, we find that with our identified city partitions, the language models generated from the contents created inside each of the partitions provide good consistency for defining the regularity of each partition of the city. This also answers our RQ 1 introduced in Hypothesis 1.4: "Is there some consistency in user's tweeting activities in certain areas of the city over time, such as regular users appearance and topic of interests?." The actual event detection using the proposed model is beyond the scope of this thesis, and will be tested in future work and more complicated text-driven models will be used in our next experiments. Our results will be used to compare with some state-of-art event detection methodologies. In the following subsections we will briefly explain our plans for some future work to be carried out based on our collected data-set.

#### 3.6.1 Twitter on the Way

Content without context is meaningless, according to Jain et al. [29]. Context can help us understanding content. Much research uses "closing in space" as a context for understanding Twitter contents similarity, in [34], tweets generated within a 200 metres radius

are used as their context and clustered together. However, it is not reasonable to assume that people sitting in their cars driving on the highway will report the same event as people who sit in the pub, who are both within a 200 metres radius. From our observations of the tweet occurrences in the Dublin city area, there are many clear travel and commuter routes which can be identified from the tweets' geo-locations from Figure 3.7. The M1 motorway heading north, the M50 ring road, the Naas road, Navan Road, etc. are all quite clear in the map. So we can see that a considerable amount of Twitter messages are sent on the Twitter users' traveling routes. These tweets may not be generated by the driver, but by someone seating in the vehicle, which could be a private car or public transport, and it is highly likely that these tweets are related to what is happening on the route, such as traffic. Obviously "being on the same road" can be a much better context than "closing in space" in the case of an event occurrence related to the route. Much information can be revealed from these tweets as these users may be heading towards the same events, if they are talking about similar events or performers, or they are all stuck in traffic, not only because they tweet about traffic, but also because they send more tweets than usual in a short traveling distance. Because of the above analysis, we decided to extract all the tweets which were generated along the route, and see if there are any semantic or temporal relations.

Figure 3.8 is an illustration of extracted tweets sent on the route of the Naas Road. A few samples of the contents are:

- @BeckyCroke I am I'm on my way
- My first match in the Aviva and not a single fecking try scored!! #depressed #TooManyPenalties
- Traffic is mental :-( I hate Monday
- En route home see you bright and early @EmmaLCarey @aoifehannon1 :):!

By simply reading the content, we can see these tweets were generated by commuters and a match spectator on his/her way home from the event, so it is not necessary for these

tweets to have specific keywords such as "traffic". The purpose of this task is to make the task of our extraction of traffic-related tweets, easier.

In this thesis, we also explore other popularly implemented crowd monitoring and event detection techniques, such as Bluetooth tracking. Each Bluetooth device can be identified by a unique serial number and detected in a short range by Bluetooth sensors. Bluetooth function is widely available in almost every mobile device, much research has successfully carried out crowd monitoring research using Bluetooth tracking technology [66]. We briefly introduced our implementations of Bluetooth tracking for our event detection task in the next section.

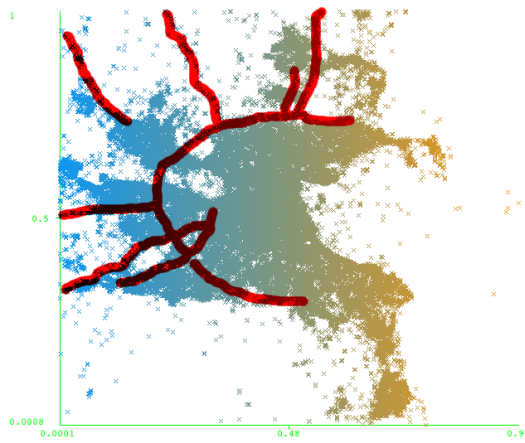


Figure 3.7: Twitter bound

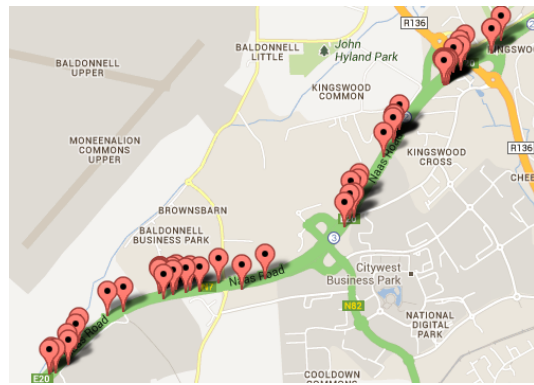


Figure 3.8: Tweet occurrences along the Naas Road

### 3.6.2 Event Detection Through Bluetooth Devices

Bluetooth device detection data is a type of accurate spatio-temporal information applicable for event detection, Bluetooth can be applied to measure the size of a crowd or to predict the crowds' trajectory if it is in motion. In [66], Mathias et al. used Bluetooth technology for the mobile mapping of spectators of the Tour of Flanders 2011 road cycling race, and achieved good results in identifying the most popular spectating locations along the race route. Each Bluetooth device is represented by a unique MAC address (a 48-bit identifier of the mobile device), which can be considered as an individual person. Based on estimated crowd size, we can identify an unusual gathering of crowds such as

a demonstration etc. by measuring and detecting unusual Bluetooth occurrences. During our research period, we set up a device in a Dublin city center location, just beside a busy road, and collected Bluetooth information from pedestrians and vehicle passing by. We collected video records as groundtruth for the actual size of the crowd. As shown in Figure 3.9, the Bluetooth data was logged simultaneously, as in the following sample *Timestamp : 01/08/2013, 9 : 43 : 15, 00102EE80C6E, MOBILE, Jenny*

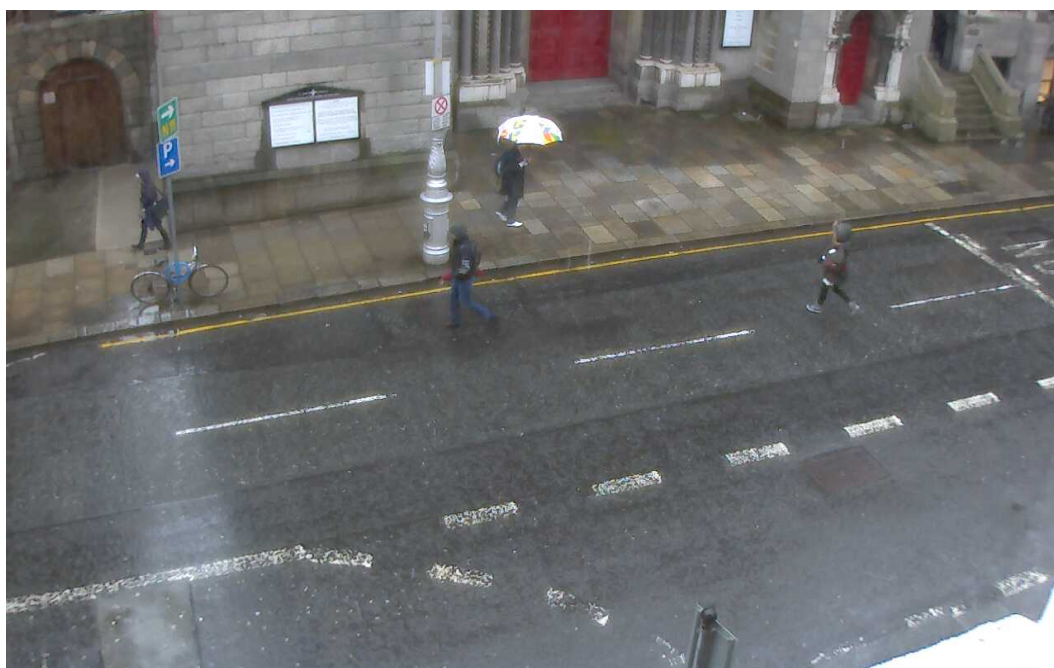


Figure 3.9: Video footage of the street

The total amount of data collected was for about a one-month period, logging both Bluetooth and video data 24 hours a day, 7 days a week. In our future work, we will try to detect any unusual gathering of crowds during that period by analysing the Bluetooth occurrences.

### 3.7 Summary

This chapter explains our proposed method for detection of small scale unusual events, based on geo-social regularities of Twitter user behaviour and the experiments we carried out for testing the reliability of our language models built for each of 25 partitions of Dublin

city, We answered **RQ1** based on the analysis to our experimental results: "Is there some consistency in user's tweeting activities in certain areas of the city over time, such as regular users appearance and topic of interests?". We also briefly explained the future work we are planning to do in the next stage for the event detection. The event detection is beyond the scope of this thesis, and will be carried out in later work.

Based on our observations of Twitter message distributions and population distributions, we found interesting correlations between Twitter messages and their locations where they were generated and population distribution. This raises some research questions such as how Twitter users are distributed across the city, and what age groups contribute the most event-related information through Twitter. These would be interesting questions to solve and will be addressed in the next chapter.

## Chapter 4

# Social Community And Population Demographics Correlation

### 4.1 Introduction

As well as a way to broadcast messages across communities of users, Twitter is also a new way of direct communication between people, particularly between friends. Exchanges of tweets directly between users can be realised through mentioning another Twitter user's name, specified by prefixing a Twitter username with an @ symbol as in @exampleuser. This means that the message or tweet is either a direct message to another user, or mentions another Twitter user's name. In this work, we refer to such tweets with mention tag(s) as part of a *conversation*. The number of mentioned usernames is the number of people involved in the conversation.

In our dataset, we have 387,800 tweets in total, these tweets were collected through Twitter API by predefining a bounding box which covers the whole Dublin metropolitan area over a collecting time span of one month. Table 4.1 shows the percentages of tweets that are of different types of conversation in our overall dataset. As we can see from the table, nearly 60% of tweets are of the type conversation mentioning one or more other Twitter users. Especially in Twitter, friendship between users can be formed by the Twitter-specific

Table 4.1: Percentage of tweets of type conversation

No of mentions	Percentage of overall
no mentions	40.73%
mentions 1 user	47.25%
mentions 2 users	8.69%
mentions 3 users	2.12%
more than 4 users	1.17%

follower-following feature. In order to explore this further and to see how this could be used in other applications, we decided to combine friendship and conversation between users in order to identify communities between Twitter users. In order to find such communities, we propose a definition for defining “close friends” on Twitter. We explore the homophily phenomenon and analyse the influential figures within our derived communities. The goal of this chapter is to study the topological characteristics of Twitter, in particular those geo-tagged tweets coming from Dublin city. Through the analysis of Twitter’s follower-following topology, which is different to the geographic topology of geo-tagging those tweets, and conversations between friends, we are aim to address the following 2 research questions:

- RQ2: Do users within the same community also have similar mobility patterns (because of the homophily phenomenon)?
- RQ3: Are users who have the most friend connections, really more influential (for example with high Klout (see section 4.3) scores), in Twitter?

## 4.2 Community Based Profile

In this section we introduce our method for community detection in Twitter, and explore the presence of the homophily phenomenon within Twitter communities. We use a user’s mobility patterns to examine this phenomenon. We also carry out some small scale experiments to test our major hypothesis: "Twitter can be used as a new way to interpret population demographic analysis", and we provide the results accordingly.



### 4.2.1 The Homophily Phenomenon

*Homophily* is a tendency that “a contact between similar people occurs at a higher rate than among dissimilar people” [44]. There are two underlying reasons for this:

- First, the process of social influence leads people to adopt behaviours exhibited by those they interact with; this effect is at work in many settings where new ideas diffuse by word-of-mouth or imitation through a network of people.
- Second, people tend to form relationships with others who are already similar to them. This phenomenon, which is often termed *selection*, has a long history of study in sociology.

Or to put this more simply . . .

- @ Aristotle "People love those who are like themselves"
- @ Plato "Similarity begets friendship"
- @ Lazarsfield & Merton "**Birds of a feather flock together**"

Various studies have demonstrated the homophily phenomenon within Twitter communities. Weng et al. have reported that two users who follow reciprocally share topical interests by mining their 50,000 follower-following relationships [70]. Kwak et al. [33] studied the reciprocated relationships between Twitter users and they found a certain level of homophily in degree correlation, which is where *users of certain popularity follow other users of similar popularity and they reciprocate*. Here, we are trying to detect and show the homophily phenomenon among Twitter user communities from the point of view of users' mobility patterns.

### 4.2.2 Twitter Community Detection

One of Twitter's special characteristics is that Twitter allows a user, A, to “follow” updates from other members who are added as “friends”. An individual who is not a friend of

user A but “follows” her updates is known as a “follower”. Thus friendships can either be reciprocated or one-way. One challenging task of community detection is to deal with this reciprocity since it may even become circular where A follows B, B follows C and C follows A, but there are no other follower relationships among these three users.

**Reciprocity** Top users, as determined by their number of followers on Twitter, are mostly celebrities and mass media personalities like sports stars, actors or musicians. Most of these individuals do not follow their followers back. In fact Twitter shows a low level of reciprocity where 77.9% of user pairs with any link between them are connected in one-way only, and 22.1% have reciprocal relationships between them as observed in our dataset. We call these latter followers *r-friends* of a user as they reciprocate a user’s following. Previous studies have reported much higher reciprocity on other social networking services: 68% on Flickr [14] and 84% on Yahoo! 360 [32]. Moreover, 67.6 % of users are not followed by any of their followers in Twitter at all. We conjecture that for these users, Twitter is rather a source of information rather than a social networking site. User profile identification will be explained in detail in Chapter 6, and this can provide some insight into this question.

To tackle the challenge of interpreting Twitter user relationships, we further restrict our definition for community detection to keep only users who have *conversations* between each other within the timespan of our data collection, one month in our case. A conversation is defined as: one user directly sends tweet(s) to another user or mention another user’s name in his tweets. Our communities can then be detected under the following 3 rules:

- A and B are two Twitter users who are active in Dublin (generate at least one tweet within Dublin area recently).
- A and B have a bidirectional relationship, A follows B, also B follows A.
- There is/are conversation(s) between A and B (A send B a message directly through reply or mention B in a message using @ tag)

By following these 3 rules, we can filter out users who are not really friends, such as a regular user and the celebrity he follows. In order to form a larger community in which users

have direct and indirect connections to others, we use the transitive relation, for example, if A and B are friends, B and C are also friends, then A and C are friends too. Figure 4.1 gives examples of how these communities can be formed.

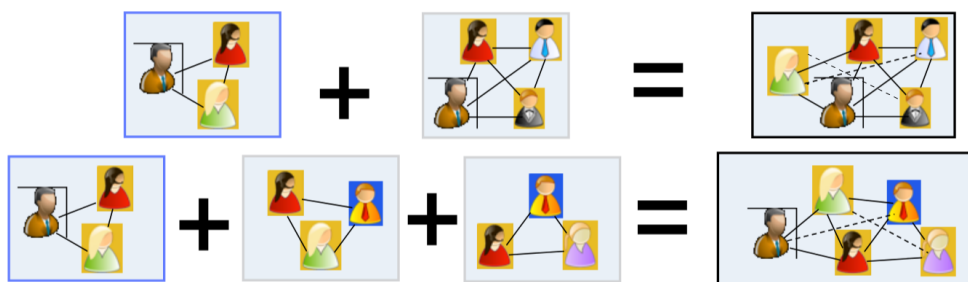


Figure 4.1: Forming a community

From this follower information, a uni-directional weighted graph can be built where the number of tweets from A to B or from A mentioning B are considered as the weight of the edge. Figure 4.2 gives an example for an identified community. Each node in the graph represents a Twitter user, the width of the edge represents the frequency of communication between 2 users, and the size of the node represents the degree of Twitter activity of the user.

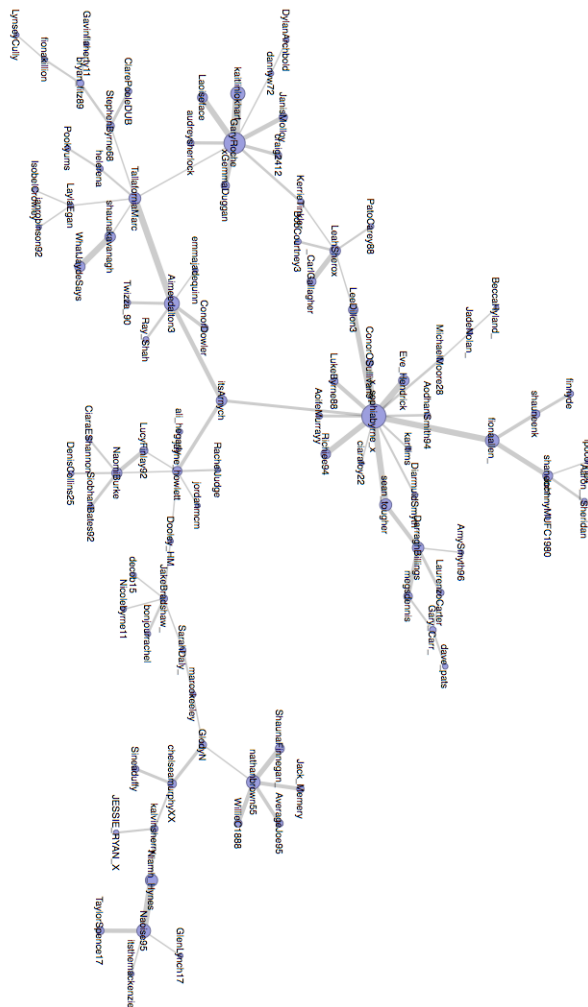


Figure 4.2: A Twitter community

### 4.2.3 Mobility Patterns

In our research work in Chapter 3, after we project Twitter users’ tweeting locations onto our 25 zones, we observed interesting spatial patterns, namely that Twitter users tend to send their tweets from across a range of different zones in the city, though there are always 1 or 2 “favourite” zones which contain most of the tweets sent by a user. This pattern is very common across different users in our dataset. People have certain locations where they spend most of their time such as their home or workplace, and prompted by this intuition, we assume these favourite tweeting locations are the users’ working, living, or leisure places.

The way these users visit (generate tweets in) their favourite locations are considered as the users' mobility pattern. In this work, we only consider the spatial pattern, which is the number of times that the user sends his/her tweets from inside the zone. Previous work has shown that Twitter users in the same community have topics of interest in common. We assume this homophily phenomenon is also shown in users' mobility patterns, so we form our hypothesis as "Users in the same Twitter community will have similar mobility patterns", this is closely related to our RQ2: "Do users within the same community also have similar mobility patterns (because of the homophily phenomenon)".

#### 4.2.4 Homophily in the Relationship Between User Community and Mobility Patterns

In order to prove our hypothesis, we examine whether users from the same community can also be classified as similar based on features shown in their mobility patterns. The number of tweets sent in a zone is considered as a feature. For each user there are 25 features, which are used for our classification task. Figure 4.3 shows an example of 5 different users (represented by their user ids), with their tweeting activity across our 25 different zones in the city.

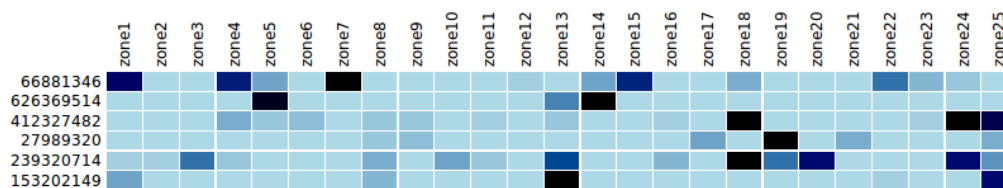


Figure 4.3: Tweeting activity in 25 different zones, vertical axis numbers are user ids, darker blue shading represent higher levels of tweeting activity

The classification is achieved using a Support Vector Machine (SVM), and the implementation is through LibSVM<sup>1</sup>.

<sup>1</sup><http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

## 4.2.5 Experiments and Results

In this section we describe the experimental setup for testing the homophily phenomenon in user's mobility patterns of the same Twitter community. We will first discuss the Twitter dataset we used, followed by our evaluations. Experiments are conducted with available implementation of SVM classifier in WEKA using 10-fold cross validation, a linear kernel and all other parameters are set using default values.

### 4.2.5.1 The Dataset

We used a small portion of the overall data, which is drawn from a subset of just 239 users chosen because these users belong to 5 different communities as shown graphically in Figure 4.4, where the communities are labelled C1, C2, ... C5.

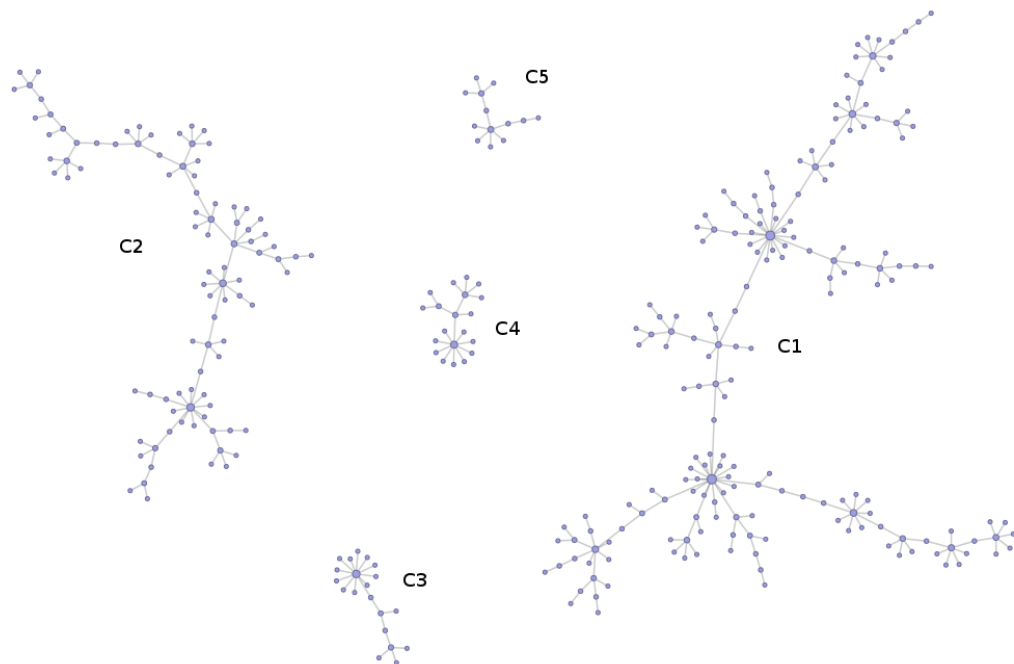


Figure 4.4: Twitter community

### 4.2.5.2 Evaluation

Table 4.2 gives the results of our classification.

Table 4.2: Classification results using the partition feature

Class	Precision	Recall	F-Measure
C1 (123 users)	63.5%	87.8%	73.7%
C2 (77 users)	60%	42.9%	50%
C3 (13 users)	50%	7.7%	13.4%
C4 (15 users)	25%	6.7%	10.5%
C5 (11 users)	59.8%	62.3%	58.3%
Weighted Avg.	59.8%	62.3%	58.3%

With such a simple experimental setup, we achieved 62.34% accuracy in our classification. The best performance is from the largest community group (C1). However the classification performance on community C5 is much better than on C3 or C4, even though they have almost the same number of users. Further analyse of the reasons behind the focus of future work.

By looking at the derived Twitter communities graph from our experiments shown in Figure 4.4, we can see that there are certain users who appear to be in the centrepoint of the whole community. These users have more connections than other users, and they appear to be the connecting points of small sub-communities; we call these users *pivot users*.

While the results of these experiments are in themselves interesting, they also lead to another research question: are these pivot users really influential in the Twitter world? In the next section, we use some of the most popular methods from the literature to measure these users influence, and we give some brief analysis.

### 4.3 Ranking Twitter User’s Influence

Measuring influence and social networking potential on Twitter has been discussed in various previous published work as well as in numerous blogs and online media [28, 70]. Related scientific work on Twitter analysis includes approaches which measure influence by not only taking followers and interactions into account, but also by analysing topical similarities with the help of a ranking method similar to PageRank [70], the ranking algorithm which exploits the topology of a linked graph to reward highly connected “nodes”, first proposed and used in the Google search engine. An interesting aspect of this work is that

in the analysed sample of Singapore-based users, a high reciprocity (e.g. mutual following relationship) was found. Huberman et al. [28] suggest that Twitter actions and thus influence are crucially influenced by “hidden networks” which consist of closer relationships between network nodes than a mere follower/following relationship.

Due to Twitter’s openly available API, there are numerous rating services that, on the one hand, calculate a score for individual users, and, on the other hand, compare scores of Twitter users to create a rating and ranking of users. A very popular and commonly used online rating service in this sector is **Klout**<sup>2</sup>, which determines user performance and influence on Twitter.

**Klout score** is a very popular and commonly used online rating service [1] which determines user performance and influence on Twitter, Facebook and LinkedIn. The service works with numerous partners who integrate Klout scores in their products (e.g. the Klout score:influence of people. Social CRM platform Radian6). Klout measures, as it states on its website, a user’s overall online influence with a score ranging from 1 to 100, with 100 being the highest amount of possible influence. Klout analyses more than 25 variables, also offering the possibility to combine the scores from all three analysed platforms, Twitter, Facebook and LinkedIn. The complex algorithm used to calculate the score is not published and cannot be reconstructed, but Klout states that it sees influence as the “ability to drive people to action”, thus making replies and retweets the most important factors. According to the calculated score, Klout places the user in a so-called influence matrix with 16 possible classifications created from the combination of eight attributes. Table 4.3 gives the 3 top Klout score owners in the world.<sup>3</sup>




---

<sup>2</sup>[www.klout.com](http://www.klout.com)

<sup>3</sup><http://toplibertarian.com/twitter/klout/>



Table 4.3: Top 5 Klout score owners in the world

Rank/Score	Name	Description
1 / 91.8	reason  @reason	Reason is the monthly magazine and website of “free minds and free markets.” Follow @reason247, a newsfeed for people who care about freedom.
2 / 89.7	Ron Paul  @RonPaul	Former US Congressman from Texas.
3 / 86.6	Senator Rand Paul  @SenRandPaul	Proud to represent the Commonwealth of Kentucky in the United States Senate.

The problem of ranking nodes based on their topological inter-dependence in a network is similar to ranking web pages based on their connectivity or links. Google uses the PageRank algorithm to rank web pages in their search results [54]. The key idea behind PageRank is to allow propagation of influence along the network of web pages, instead of just counting the number of other web pages pointing at the web page. In this section we rank our Twitter users by our form of the PageRank algorithm based on the following relationships between users, this "following" relationships include following and message exchange in terms of Twitter. As a working collection, we take 2 large communities from our dataset, such communities are formed by a number of users who have connections defined by our 3 rules, these 3 rules are described in the Reciprocity section, examples of these communities are shown in Figure 4.4. Figure 4.5 shows a part of community 1, which contains 183 users, and figure 4.6 shows community 2, which contains 93 users.



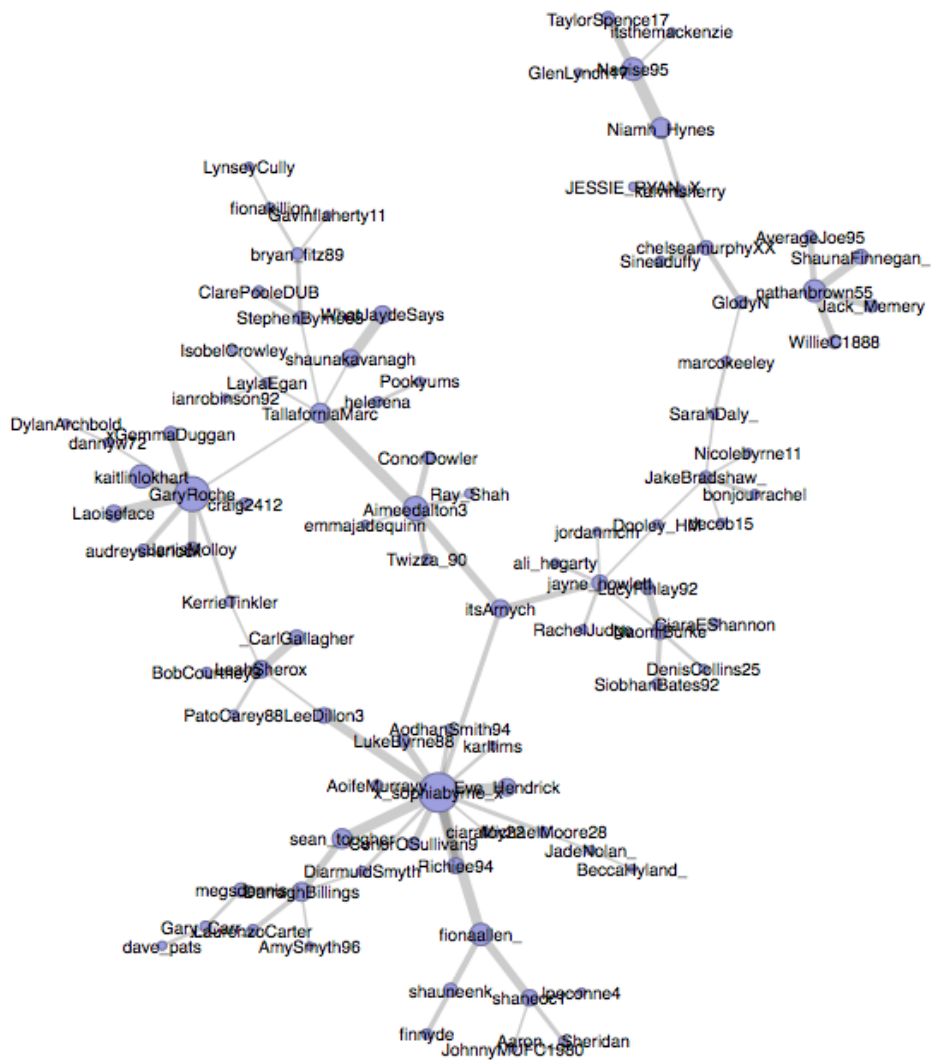


Figure 4.6: Twitter Community 2

We apply the PageRank algorithm to calculate scores for each user, such score measures the relative importance of this user within his/her community. We also use the Klout API<sup>4</sup> to retrieve each user's Klout score. Tables 4.4a and 4.4b show a listing of the top 5 PageRank scorers and their Klout scorers from the 2 selected communities.

<sup>4</sup><http://klout.com/s/developers/docs>

Table 4.4: Users PageRank score vs. Klout score in different Communities

(a) Community 1: Users PageRank score vs. Klout score

User name	Pagerank score	Klout score
Glanraff	0.0125	35.179
leinsterrugby	0.009	79.658
IrelandHandball	0.009	36.677
darraghdoyle	0.0080	78.544
samanthamumba1	0.005	81.802

(b) Community 2: Users PageRank score vs. Klout score

User name	Pagerank score	Klout score
Naoise95	0.0179	40.054
GarryRoche	0.0154	37.760
TallaforiaMarc	0.008	50.242
fionaallen	0.008	48.573
x_sophiabyrne_x	0.007	49.705

From this list, we can see that some users have high Klout score and also high PageRank score, such as “leinsterrugby” and “darraghdoyle”. “leinsterrugby”, as is suggested by its name, is an account aimed at sending information to rugby fans in Dublin, allowing them to share their feelings about Leinster rugby by tweeting to this account; “darraghdoyle” is a very active Dublin Twitter user, his tweets topics cover various categories, and he has many active followers. We can also see that some users have low Klout scores but high PageRank scores, such as “GarryRoche” and “Naoise95”. These 2 users do not have many followers or following, but within their friends’ circles, users are relatively active in terms of the number of tweets generated per day.

#### 4.4 Analysis And Future Work

Based on our experimental results in this initial part of our work, we find the following:

- There are some levels of homophily in the users mobility patterns where those users belong to the same Twitter community. For example, by using the users’ mobility patterns (tweeting activity in 25 different zones), we successfully identified the actual Twitter communities for 62.34% of the users. We can conclude that users from the

same Twitter community not only have topics of interest in common, but also similar tweeting activities. We will further expand this idea to users' socio-economic status in future work.

- We also found that users who appear to have the most connections in the community are not necessarily the most influential figures in Twitter world. For example, "GarryRoche" and "Naoise95", these 2 users do not have many followers or following, they are not very influential in Twitter world in terms of Klout measure, but within their friends' circles, these users are relatively active in terms of their daily tweet exchanges. We also found that some very high Klout scorers do not have much interaction with their friends in their circles, such as "samanthamumba1", who is a celebrity born in Dublin, she has very high Klout score but low PageRank score. This is because even though she has a large number of followers, she only publishes her own status, and barely interacts with her followers.

These results also answered our RQ 2 and RQ 3 research questions, introduced in the Hypothesis section 1.4 earlier in the thesis: RQ2: "Do users within the same community also have similar mobility patterns (because of homophily phenomenon)?" and RQ3: "Are users who have the most number of friends connections really more influential with high Klout4.3 score figure in Twitter?", as mentioned above. In the following sections we will briefly explain our plans for some future work to be carried out based on our collected data-set.

#### **4.4.1 Correlation Between Social Relationship and Socio-economic Background**

In Oliver et al.'s work [53], they identify social communities through telephone call exchanges between users. They use the derived community to predict the user's socio-economic status. Because tweet exchanges between Twitter users are kind of similar to making phone calls, in our future work, we aim to find correlations between Twitter users and their community structures, and their socio-economic status. Furthermore we will aim to be able to make predictions as to the socio-economic status of these users too.

## 4.5 Summary

In this chapter, we briefly talked about the our work on Twitter social communities. We found certain levels of the homophily phenomenon in users' mobility patterns who are within the same Twitter social community. We also analyzed the influence of users within our derived Twitter social community. We answered **RQ2** and **RQ3** based on the analysis of our experimental results. We also briefly explained the future work we are planing to do in the next stage for identifying the correlations between Twitter user's social relationships and their socio economical background.

We found interesting correlations between Twitter message distributions and population distribution. This raises some research questions such as how Twitter users are distributed across the city, and what Twitter user age groups contribute the greatest amount of event-related information through Twitter? These would be interesting questions to solve which will be addressed in the next chapter.

## Chapter 5

# User Tweeting Behaviour Analysis

### 5.1 Introduction

In our dataset, collected for this thesis, each tweet is timestamped and embedded with its geolocation. For each user, aggregating this time and location information over time reveals a lot of information about his/her temporal and geographical behaviour. *As members of society, while we would like to believe that our movement and mobility patterns have a high degree of freedom and variation, at a global scale human mobility exhibits structural patterns which are subject to geographic and social constraints [17].*

Based on our observations from our dataset, we found that such types of periodic behaviour commonly appear among the individuals in our dataset. In order to make our analysis task simple, we project all individual geolocations of tweets into 25 different zones, as introduced in Chapter 3. Then for each user, there are a number of different zones in which the user will generate tweets over a one-month period. We noticed that for most users there are 1 or 2 zones which are each user’s “favourite”, which are those zones which contain a large proportion of the overall tweets for each user among our 14,533 users. On average, 73% of tweets were sent from the favourite 1 or 2 different zones of each user.

One would expect that people exhibit strong periodic behaviour in their movement as they move back and forth between their homes and workplace [22, 36]. Prompted by this intuition, and together with our observations mentioned above, we assume that for each

user the number of zones which contain most of a user's tweets, are indeed his/her home, workplace, or favourite leisure locations. It is obvious that Twitter users are more active when they are at home or at leisure locations. However, we also found that users tweeting behaviour patterns demonstrated in each different zone follow a power law figure. From this we can infer that users not only tweet actively at their home or workplace, but also they are active in random locations. We expect that tweets sent from random locations will contain unexpected information different from user's regular topic of interests. Through the analysis of these random tweets, we are able to build a better context for helping us to better understand the tweet contents, and extend our event detection models.

By studying our users' temporal tweeting behaviours, we found groups of users who have typical timing characteristics in their tweeting activities. For each group of users, we roughly estimated their social status based on their temporal tweeting characteristics. We also observed interesting correlations between the users' tweeting activities across different zones of the city and the city's population demographics. Based on these correlations, we are able to understand the fabric and operation of the city from a different point of view, not visible previously.

In this chapter, we will study the main two aspects of users' tweeting behaviour: geographic behaviour (where do we tweet?) and temporal dynamics (when do we tweet the most?), and correlations between tweeting activity and city demographics.

## **5.2 Geographical Behaviour Analysis**

In this section, we demonstrate the geographical distributions of our users' tweeting locations. One would expect that people exhibit strong periodic behaviour in their movement as they move back and forth between their homes and workplace. We observed this pattern in our users' tweeting locations. We use the partitions generated from Chapter 3, for which the Dublin city area was partitioned into 25 different zones based on the geolocations embedded in the tweets collected over a one-month period of time. We identified 5,875 unique users from our dataset. These users generated 95% of the total number of tweets, which



narrowed down our total number of tweets to 368,476 tweets. The reason for us to drop tweets generated by the rest of the 5,658 users is because of their inactiveness, where most of those users only generate 1 or 2 tweets within a month.

We observed strong periodic behaviour in the distribution of locations where tweets were sent from. In table 5.1, we can see that almost 44 % of users sent tweets from less than 3 of 25 different zones across the city.

Table 5.1: Percentage of users tweeting activities in different zones

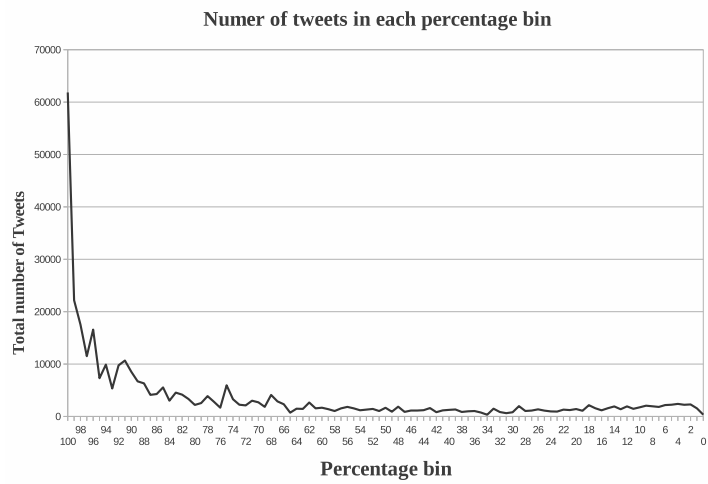
Different number of zones	% of overall users
1	21.8%
2	22.7%
3	18.8%
4	13.7%
5-25	23%

It is reasonable for us to assume these locations are the users' homes, workplaces or leisure places, and it is common sense that Twitter users are more active in the above locations. But we also notice that there are 23% of users who not only generate tweets in their favourite zones but also across different seemingly random zones. The tweets sent from these random locations are of particular interest, especially for our event detection task introduced in Chapter 3. When we go back to the event detection task in Chapter 3, understanding the noisy contents of Twitter is a non-trivial task, but having good information on the context in which tweets are sent may help us to address this problem. If people only send tweets while in their favourite locations, their contents can be expected to be similar, and even predictable. Thus if we want to find irregular, unexpected event-related content, tweets sent from their non-favourite locations should be what we are looking for. Therefore, we consider Twitter messages generated from such random locations as examples of good context for event detection.

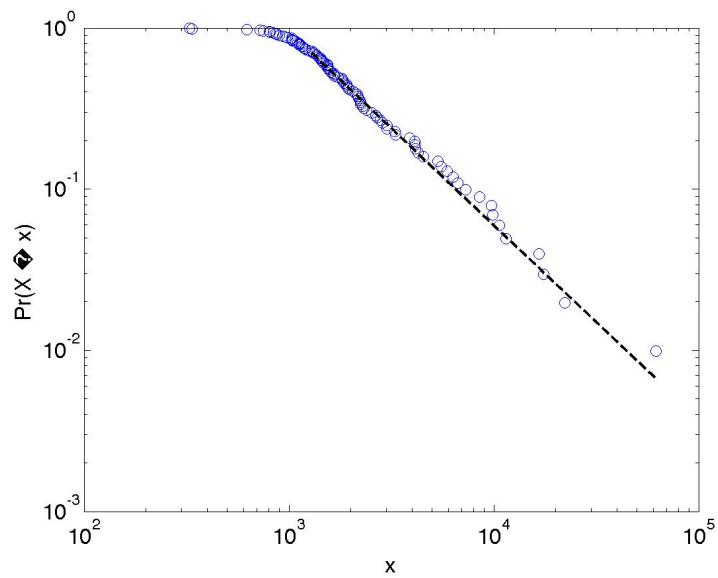
However, will we be able to find enough information from these contexts in order to make automatic event detection possible? To answer this question, we construct a location distribution for each user, for example: user1 sent 10% of tweets in zone1, with 200 tweets in total, and 90% in zone5 with 1,800 tweets in total, thus zone5 is user1's favourite lo-

cation; user2 sent 100% of tweets in zone10 with a total of 1,000 tweets, thus zone10 is user2's favourite location, etc. Then we make 100 percentage bins, one for each 1%. So in the 100% bin, we will have 1,000 tweets in total from user2, and in the 40% bin, we will have 400 tweets in total from user1, etc. The results are shown in Figure 5.1a.

As we can observe from the graph, there are a significant number of tweets generated at random locations distinct from the users' favourite locations, our next experimental results show that the distribution is well-fitted into a long tail distribution, as shown in Figure 5.1b. In statistics, a *long tail* of some distribution of numbers is the portion of the distribution having a large number of occurrences far from the "head" or central part of the distribution.



(a) Location distributions in percentage bins

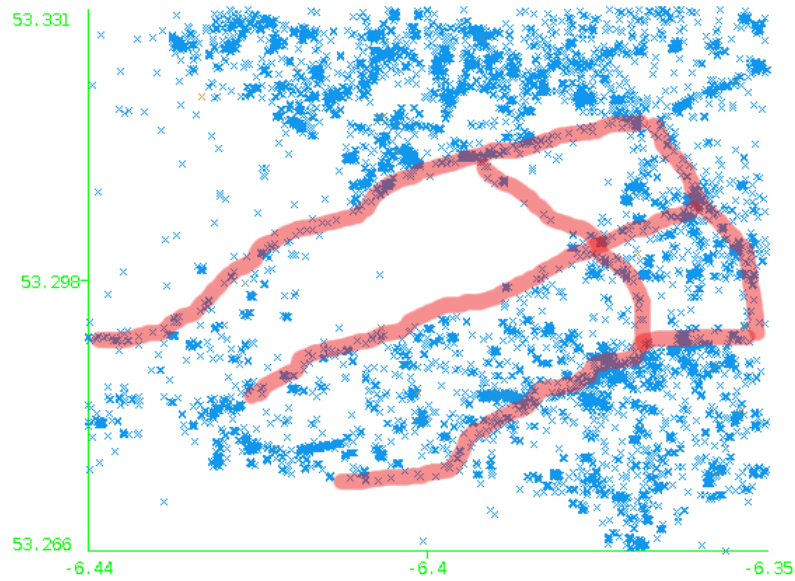


(b) Fitting of power law distribution

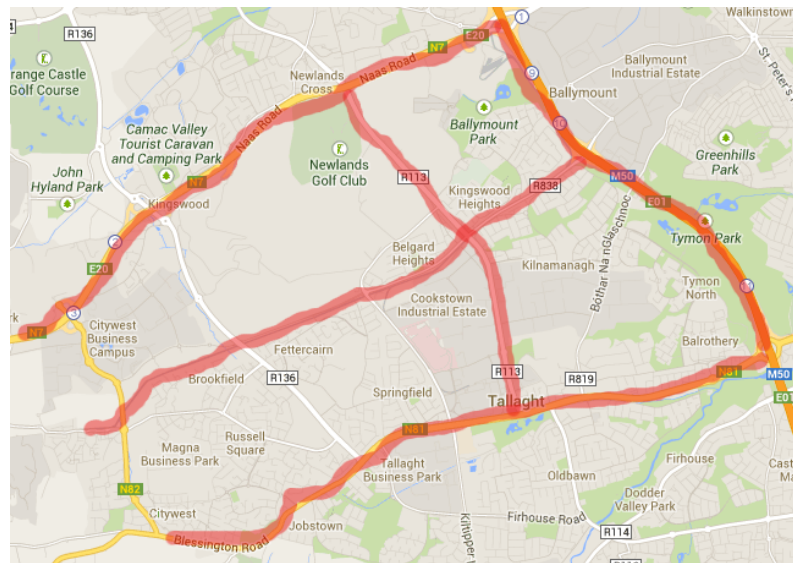
Figure 5.1: Location distribution and power law fit

So we can conclude that even though Twitter users do have certain favourite locations from where they send most of their tweets, there are still significant amounts of information that can be found outside their favourite locations. Therefore we are able to find enough information from these random tweeting locations to support our event detection task. More interestingly, when we take a closer look at the geolocations of our tweets in a small region,

we observe that these locations match closely to actual road locations, as compared in Figure 5.2a and Figure 5.2b where the lines marked in red have very close match. This match happens on some major roads in other part of the city areas too. This would be a good example for the context we use for our events detection, because what people see on the road is most likely to be related to traffic or transport-related events which may cause traffic conditions.



(a) Plot of tweets: geolocations over a city area



(b) Highlights of road maps of a city area

Figure 5.2: Comparison between tweet geolocations and city roadmaps

### 5.3 Temporal Dynamic Analysis

The volume of tweets generated by different users over time exhibits unique characteristics. These characteristics potentially represent, in some way, the user's daily living patterns. Through studying Twitter users temporal tweeting behaviours, we hope to group users with

similar daily life patterns, and roughly estimate the social status of these grouped users. Each tweet in our dataset is timestamped, for example "Thu Jan 24 10:57:03 +0000 2013". Based on these timestamps, we aggregate the tweets into hourly bins for each 24 hours for weekdays and weekends. The reason that we analyze user's tweeting behaviours during weekdays and weekend days differently is because we observed significant differences between them as most people work during the week and relax at weekends. Figure 5.3a shows 2 different trends from user's tweeting patterns for weekdays and weekends in terms of the average number of tweets generated in each hour. We can see that users are much less active during the weekend than weekdays, the boosting time of the volume of tweets starts much later in the weekend, which is 2pm as compared to 8am during weekdays. Particularly, in the weekdays trend, the most active tweeting time is between 9pm to 12pm. We speculate that this is because people active in the night are the most active contributors of tweeting activity.

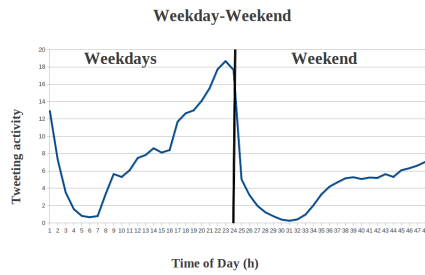
We cluster our users by their temporal tweeting features. For each user, there are 48 features, each feature represents the average number of tweets a particular user generated within an hour, across a timespan of a one month period. The first 24 features are for weekdays, and the other 24 features are for weekend days. In our experiments, we only consider users who sent more than 100 tweets in our one-month data collection timespan. 100 was chosen empirically based on observations on our dataset. This cut down the total number users in our experiments to 805. We used the built-in EM algorithm clustering from WEKA to run these experiments. We divided all of our users into 10 different clusters. Within each cluster, we can detect users who have noticeably unique characteristics in their temporal tweeting patterns, as shown in the following graphs.

Figure 5.3b shows a group of very active users, they are 10 times more active than average in terms of hourly tweeting volume. Because their temporal tweeting patterns are similar to the overall trend, we consider these people as general Twitter users, who are just more active than others.

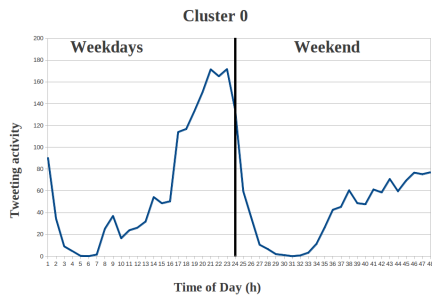
By contrast, clusters 1 Figure 5.3c and 6 Figure 5.3d show completely different patterns. For example in cluster 1, users are mostly active across the whole day. Yet in cluster 6, lunch

time, between 12pm to 2pm, is their most active tweeting time, with another small boost of activity for these users starts at 9pm and stops at around 11 pm. We could infer that these people are typical office workers, their tweeting times are mostly during their lunch break, and after dinner, and they don't stay out late at night socialising because they have to get up early for work in the morning.

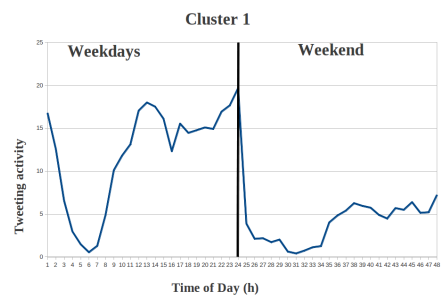
In Figure 5.3e, the pattern shows the activity of a property sales agency, this account is used for broadcasting property advertisements. As we can see, this user is constantly generating Twitter messages, no matter what time it is, this feature can be typically used to identify non-human tweeting accounts.



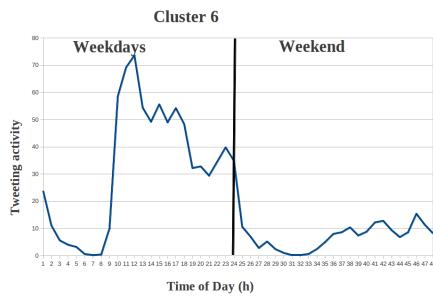
(a) Overall tweeting behaviour



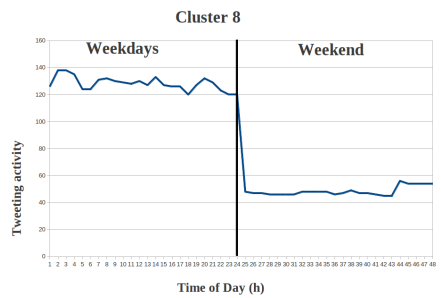
(b) Tweeting distribution for Cluster 0



(c) Tweeting distribution for Cluster 1



(d) Tweeting distribution for Cluster 2



(e) Tweeting distribution for Cluster 8: Advertisement accounts

Figure 5.3: Distribution of tweet volumes.

## 5.4 Geographical Distribution vs. Population Density

In this section we analyze the relationship between Twitter activities and population densities in the Dublin area. The derived relations reveal some interesting structures of the city,



such as tourists hotspots. The analysis of Geographical distributions of tweets. This lets us view the city and its demographical fabric in a different way to other views.

In our experiments, the population and area boundary data is taken from the Irish Central Statistics Office (CSO), where each area boundary is defined by the CSO as a *Small Area* (SA) [52] as follows: *Small Areas are areas of population comprising between 50 and 200 dwellings created by the National Institute of Regional and Spatial Analysis (NIRSA) on behalf of the Ordnance Survey Ireland (OSI) in consultation with the CSO.* Small Areas were designed as the lowest level of geography for the compilation of statistics in line with data protection legislation and generally comprise either complete or part of townlands or neighbourhoods. There is a constraint on Small Areas that they must nest within Electoral Division boundaries and cannot straddle these boundaries. Small areas were used as the basis for the enumeration of the population in the most recent census in 2011. Enumerators were assigned a number of adjacent Small Areas constituting around 400 dwellings each, in which they had to visit every dwelling to deliver and collect a completed census form and record the dwelling status of unoccupied dwellings. The Small Area boundaries have been amended in line with population data from Census 2011.

We only take a portion of the total population in each Small Area, whose ages are between 16 to 59 years old, because these are the major age groups of Twitter users according to the report from Royal Pingdom <sup>1</sup>. The population densities for each Small Area are calculated by taking the number of the population aged 16-59 (male and female) divided by the total area size calculated based on the geographical coordinates of the boundary. As in equation 5.1:

$$density = \frac{population}{\frac{|(lat_1 lng_2 - lng_1 lat_2) + (lat_2 lng_3 - lng_2 lat_3) + \dots + (lat_n lng_1 - lng_n lat_1)|}{2}} \quad (5.1)$$

For example, the SA boundary which covers the Phoenix Park area is shown in figure 5.4.

---

<sup>1</sup><http://royal.pingdom.com/2012/08/21/report-social-network-demographics-in-2012/>



Figure 5.4: Population (age group 16-59) density of Phoenix Park area

The area marked in light blue has a total residence of 574 people of age 16-59. The total area size calculated using the above equation is  $6.96 \text{ km}^2$ , so the population density of this small area is  $574/6.96(\text{km}^2)=0.000082 \text{ per km}^2$ . Figure 5.5 demonstrates the Small Area population densities of the age group 16-59 over the Dublin area. Figure 5.6 shows the Twitter activities from different areas of the city over a one month period. We can observe that some areas have low population density but have a high volume of Twitter activities. We consider these locations as Twitter user hotspots. Figure 5.7 which shows the distributions of the number of unique Twitter id numbers in different Small Areas. By comparing the 3 figures, we can visually identify some outlier zones, such as Dublin Airport, Phoenix Park, the Trinity College, UCD and DCU campuses, the Dundrum shopping centre, etc. Dublin Airport, the Phoenix Park and Heuston train station are popular areas, which have very small numbers of residents, but have a lot of visitors or travelers. College campuses have both high Twitter activities and high numbers of Twitter users. This is because college students are among the most active Twitter groups. The Temple Bar area of the city centre and the

Stephen's Green Park in the middle of the city have high Twitter activities as well, and have more Twitter users because they are the hot spots for tourists. Through the comparison between Twitter user activities and population densities we can gain some understanding of the city fabric from a different point of view.

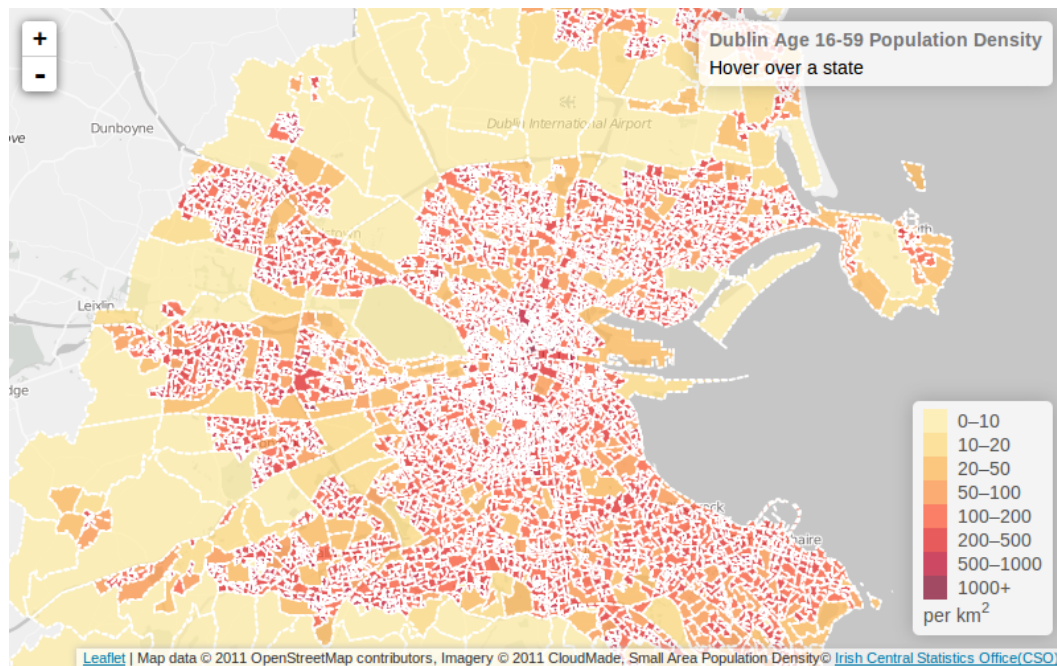


Figure 5.5: Population (age group 16-59) density of Small Areas

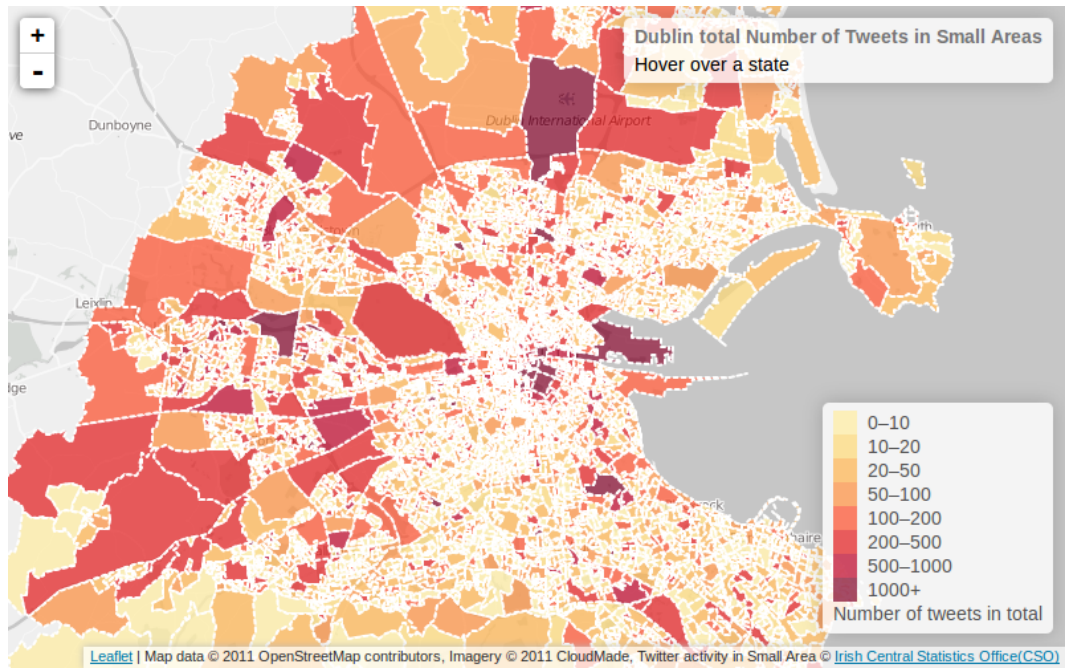


Figure 5.6: Twitter activity in Small Areas

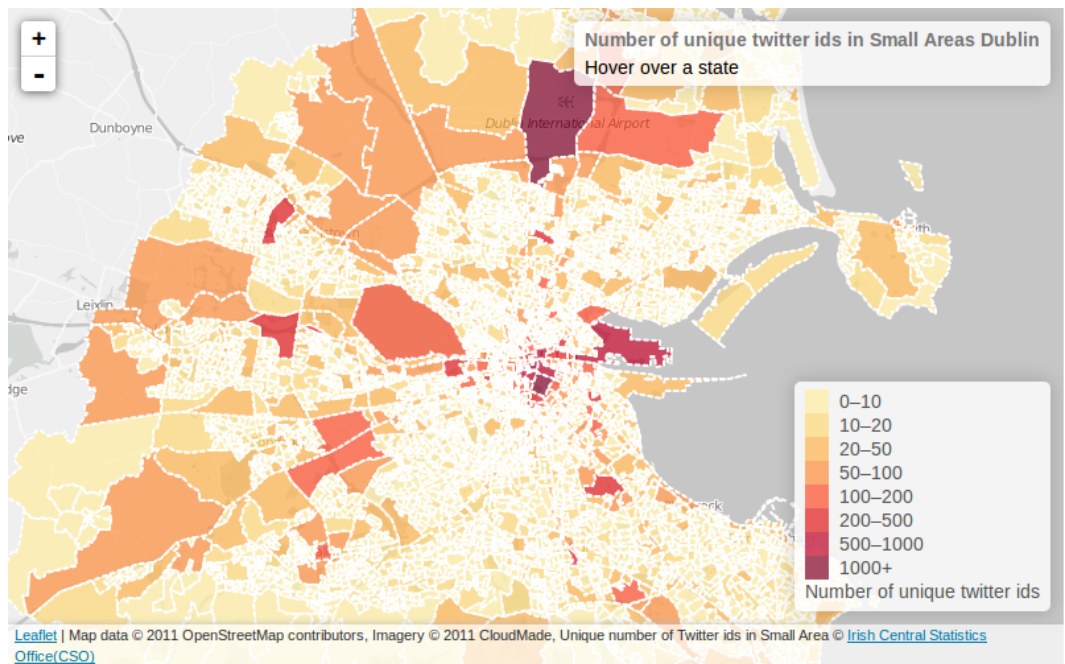


Figure 5.7: Unique number of Twitter ids in Small Areas

### **5.4.1 Summary**

In this chapter, we developed our observations of our users' geographical and temporal tweeting behaviours. We found that although Twitter users are more active in their favourite locations in term of tweet generation, such as their home, workplace or leisure places, they contribute significant numbers of tweets from random locations, and these tweets are of particular interest to us for our event detection tasks. By studying the temporal tweeting patterns of our users, we can identify groups of users with similar patterns and be able to roughly estimate their social status. We also discovered some interesting correlations between Twitter users tweeting activities and population densities in the Dublin city areas. These correlations show that we can understand the city fabric from a different angle through Twitter activities.

## Chapter 6

# Twitter Source Classification

### 6.1 Introduction

In our work we have employed Twitter, working as a form of online sensor for the task of detecting events in a real, physical city. We consider Twitter users' tweets as sensor readings and Twitter users as sensors. Even though this is a simplified model of how to use social media data, this data is very noisy and not very reliable. The noise inherent in social media data like tweets not only comes from the sensor readings themselves, because Twitter users use informal language in their tweets, but also from the sensors themselves. Twitter users have different intentions when they tweet [30], such as acting as information providers who constantly publish news or advertisements, or sometimes acting as information seekers who are looking for news or celebrity gossip, etc. Based on this notion, we divide our users into 2 different types:

- non-personal users (such as information sources)
- personal users (information seeker or people shares their everyday life).

Non-personal users usually have different motivations from personal users. The former normally publish news, advertisements or political opinions in a clear form. In our relatively small scale experiments where we target the task of detecting unexpected events, these tweeters are not of interest to us because it is very unlikely that these users will tweet about

a house fire on the street, or complain about a traffic jam at a junction on their way home, etc. In general, the information provided and the behaviour demonstrated by these non-personal accounts are not related to our event detection task.

In order to avoid getting overwhelmed by this unrelated noise generated from these accounts and to improve the performance of our system, a filter is required. The goal of this work is to build such a filter which is able to automatically determine whether a Twitter user account is personal or non-personal based on the authorship profile. This profile is derived based on the author's previous tweet contents. For our classification task, we proposed a set of features with a focus on the users' previous Twitter content, such as the presence of slang words, sentimental symbols, etc. Experimental results show that our classification method provides acceptable accuracy for this task, and so the classifier will be used in our future work. This uses this filter to remove information from unrelated accounts in order to improve the performance of our event detection language models.

## 6.2 Feature Selection

Selecting a subset of relevant features for building robust machine learning models<sup>1</sup> is a major research problem. Hence in our work we use a greedy strategy to select the feature set which generally follows the definitions of the 2 classes. We extracted a set of 4 features as follows:

1. Percentage of tweets containing personal pronouns (I, you, he, she, it, we, they). The rationale behind this is that a large number of tweets containing personal pronouns in the tweet set are considered as a strong indication that this account is a personal account.
2. Emphasis on words based on uppercase letters, and the usage of repeating characters in a word (eg. "veeery"). Like the first feature, the rationale is that this is considered as a strong indicator of a personal account.

---

<sup>1</sup>[http://en.wikipedia.org/wiki/Feature\\_selection](http://en.wikipedia.org/wiki/Feature_selection)

3. Presence of slang words is determined by a lookup in a wordlist which consists of about 100 most commonly used slang words in Ireland obtained from the Web. Presence of slang words is also indicative of a personal account.
4. Finally, we also capture the presence of non-ASCII keywords, such as smily faces (Twitter allows users to attach UTF-8 symbols in tweets), etc. or occurrences of “haha” or other popular sentimental symbols in different forms, such as “hahahaha”, and “:D” etc. this feature is commonly adopted by personal users to express their sentiment.

## **6.3 Experiments and Results**

### **6.3.1 Experimental Setup**

Our Twitter dataset was collected from within the Dublin area over a one-month period, containing a total of 384,000 tweets, created by 14,533 different Twitter users. We selected the most active 500 users from our dataset, where the least amount of tweets generated by these users over the one month period was 150 tweets. Two annotators were assigned to manually label these 500 users into one of two categories: personal user (PU) or non-personal user (NU). Their agreement on the categorisation was measured using Cohen’s Kappa coefficient yielding a result of (0.62), indicating substantial agreement between the annotators. To ease annotation, the annotators were shown 10 randomly selected tweets from the user that they were asked to categorise (at least 5 words in length, mention tags, hash tags and ascii symbols are included) from each user’s tweet set.

Experiments were conducted with an available implementation of the Naive Bayes classifier in WEKA<sup>2</sup> using 5-fold cross validation. For each user’s tweet set, we removed the stopwords except for personal pronouns (I, you, he, she, it, we, you, they).

---

<sup>2</sup><http://www.cs.waikato.ac.nz/ml/weka/>



### 6.3.2 Performance Evaluation

Figure 6.1 shows our classification results of using each of the four individual features and then all features taken together, to train the classifier. We can see that each one of our proposed features provides good performance according to three different performance measures namely Precision, Recall and F-Measure, used in earlier work in this thesis. All of the results are above 86%, but using 4 features together gave consistent performance across all 3 measures. In particular, using the personal pronoun feature scores the highest. This is a reasonable expectation because personal users tend to publish their personal status using first person pronouns such as “I” and “we”, very frequently. However we can’t ignore the fact that our personal account set is much larger than the non-personal account set among the users in our dataset, in the ratio of 468:32 in our case. In later work, we will expand the negative set to include more non-personal accounts.

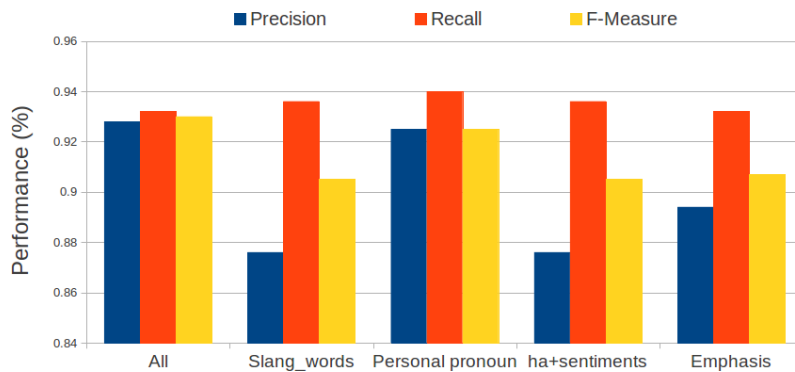


Figure 6.1: Classification performance for individual features.

To avoid the bias on the dataset, a separate experiment was carried out using unbiased

dataset, which contains randomly selected 32 out of 468 personal accounts and 32 non-personal accounts. The results are shown in Figure 6.2:

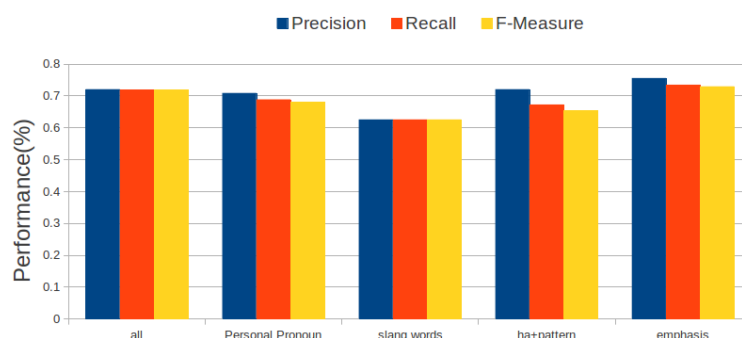


Figure 6.2: Classification performance for individual features on unbiased dataset.

The overall performance is on average 70%, which is also acceptable.

## 6.4 Summary

In this chapter, we briefly introduced our work on building a classifier for identifying Twitter users in terms of their type, either personal or non-personal users, by using an analysis of the content of their historical tweets. Using such a classifier, we can now filter out possible non-event-related information (tweets) generated from non-personal accounts. Experimental results show that our proposed features can provide accurate classifications. Currently we are working on a small set of users, 500 in this case, and our 2 categories are very unbalanced with only 32 out of 500 users in the non-personal category. In future work, we will incorporate more non-personal accounts to our dataset. Most importantly, we will use the

filtered information to improve our language modelling based event detection technique.

## Chapter 7

# Discussions and Future Work

In this thesis, we tackled some of the problems in the very challenging yet popular area of automatic social media content interpretation.

Aiming to find a better way to comprehend city dynamics through social media sensors, we first focused on an exploration of the consistencies across Twitter users' behaviour. This was in an attempt to learn more about our users and looked at things such as their topics of interest. We gauged these interests from Twitter user-generated content over a period of time, and we then concentrated on the derivation of language models from such topic of interests. Combining these language models with geographical information, we built a mapping from semantic consistency to locations.

We then ran a series of experiments which showed some level of such consistency across these. As a result, our event detection task can now be based on observing whatever inconsistencies that may arise from Twitter content based on analysing the combination of semantics of the Twitter content, and locations from which the tweets are issued. Thus unexpected content, not forming part of the language model and arising from an unusual location, can be an indicator of an event in the real world.

Secondly, we demonstrated the observed correlations between social communities formed by social media relationships and populations demographics. Our experimental results proved our proposed concept that social relationships between users can infer some of their social status, such mobility patterns, etc. Also we proposed an intuitive method for clas-

sifying Twitter users into types as personal or non-personal users by using a set of simple features.

Our four research questions were examined in turn when applying the above tasks to event interpretation and social community profile analysis. The corresponding research questions which drove our investigation into these tasks are now revisited as follows:

- **(RQ1:)** Is there some consistency in user's tweeting activities in certain areas of the city over time, such as regular users appearance and topic of interests?
- **(RQ2:)** Do users within the same community also have similar mobility patterns (because of homophily phenomenon)?
- **(RQ3:)** Do users who have the most friends connections have really more influential (With high Klout4.3 score) figure in Twitter?
- **(RQ4:)** How can we filter out these non related twitter accounts in order to enhance our system's performance?

Generally speaking, the research question (RQ1) is raised for the task of real-time event detection in Twitter, while (RQ2) and (RQ3) deal with social media community based profile analysis. The derivation of social media user behaviours from (RQ1) forms the basis for (RQ2) and (RQ3) and the answers for (RQ2) and (RQ3) are also supportive back to the task of (RQ1). The answer to (RQ2) shows that there are certain levels of homophily in the users' mobility patterns where these users belong to the same Twitter community. (RQ3) suggests that although Twitter users are more active in their favourite locations in term of tweet generation, such as their home, workplace or leisure places, they contribute a significant amount of tweets from random locations, and that these tweets are of particular interest to us for our event detection tasks.

By studying the temporal tweeting patterns of our users, we can identify groups of users with similar patterns and be able to roughly estimate their social status. (RQ4) is proposed to deal with the issue of Twitter account type classification for the purpose of enhancing the performance of our system. Trying to answer these research questions, different algorithms

are developed and demonstrated to be effective in Chapter 3, Chapter 4, and Chapter 5, which are the main contributions of this thesis. With these analysis and experimental results, we believe that the semantics of events can be maximally interpreted to provide an efficient tool for city planners to quickly grasp the city pulse, and even aiding individual information seekers for accurate information they are looking for.

## **7.1 Main Contributions**

A location-text joint modelling algorithm was introduced in Chapter 3 for the purpose of real-time event detection in Twitter. Although, event detection in social media, such as Twitter, has been studied for a few years, the area is still relatively new, and there are still many challenges to be solved, our proposed method gives a different way of cracking this issue. To the best of our knowledge, the method we introduced in Chapter 3, has never been used before. In chapters 4 and 5, we explained the correlations between social relationships and population demographics discovered through our observations. These results show that we can use social media as a tool to view city dynamics from a different angle. In chapter 6, we proposed a novel Twitter user account profile classification method, our introduced set of features produce very good classification performance. These main contributions tackled the four research questions we just revisited. Semantic Web technologies have been employed in all of our contributions at different levels of abstraction. Since not one single technology, either Multimedia Retrieval or Semantic Web, can successfully fulfill the task of event detection in social media, Semantic Web technologies have been assimilated in our contributions to address the research questions together with traditional Multimedia Retrieval technologies like supervised machine learning, unsupervised machine learning, etc. As answers to the research questions, the contributions of this thesis have supported our hypotheses formulated at the beginning of thesis, that is, “Social Media as a new way of sensing technology can work as an extension of traditional media for urban city dynamics interpretation”.

## 7.2 Future Work

The future work section is rewritten as: Our algorithms and models have shown their merits to some extent in fulfilling event semantic interpretation tasks. But not all of them are free of limitations. Especially, our event detection model has not been put into practice, however our proposed location based language modeling technique provides good consistency for defining the regularity of each partition of the city. Also, our correlation analysis between social media relationships and social status is only small scale, but we found strong indications of homophily phenomenon in Twitter user's mobility patterns, who are within the same Twitter social community. Our proposed features for Twitter account type classification also show promising results. In future work, we are planning to explore other event detection paradigms implementing different modern technologies such as Bluetooth mentioned in Chapter 3, to build a more complex system for our event detection task.

# Bibliography

- [1] Klout is available at <http://klout.com>, date accessed: 2013-06-01.
  
- [2] Mumbai attack <http://www.telegraph.co.uk/news/worldnews/asia/india/3530640/Mumbai-attacks-Twitter-and-Flickr-used-to-break-news-Bombay-India.html>, date accessed: 2013-06-01.
  
- [3] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. Semantics+ filtering+ search= twitcident. exploring information in social web streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 285–294. ACM, 2012.
  
- [4] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. Twitcident: fighting fire with information from social web streams. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 305–308. ACM, 2012.
  
- [5] Akiko Aizawa. An information-theoretic perspective of TF\*IDF measures. *Information Processing & Management*, 39(1):45–65, 2003.
  
- [6] James Allan. Introduction to topic detection and tracking. In *Topic detection and tracking*, pages 1–16. Springer, 2002.
  
- [7] James Allan, Jamie Callan, Kevin Collins-Thompson, Bruce Croft, Fangfang Feng, David Fisher, John Lafferty, Leah Larkey, Thi N Truong, Paul Ogilvie, et al. The



LEMUR toolkit for language modeling and information retrieval. *The Lemur Project*. <http://lemurproject.org> (accessed 25 January 2012), 2003.

- [8] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering short texts using Wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM, 2007.
- [9] Hila Becker, Mor Naaman, and Luis Gravano. Event identification in social media. In *WebDB*, 2009.
- [10] Hila Becker, Mor Naaman, and Luis Gravano. Beyond Trending Topics: Real-World Event Identification on Twitter. In *ICWSM*, 2011.
- [11] Hila Becker, Mor Naaman, and Luis Gravano. Selecting Quality Twitter Content for Events. *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, 11, 2011.
- [12] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022, 2003.
- [13] Alex Burns and Ben Eltham. Twitter free Iran: An evaluation of Twitter’s role in public diplomacy and information operations in Iran’s 2009 election crisis. 2009.
- [14] Meeyoung Cha, Alan Mislove, and Krishna P Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World Wide Web*, pages 721–730. ACM, 2009.
- [15] Antoni B Chan, Z-SJ Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008.
- [16] Ling Chen and Abhishek Roy. Event detection from FLICKR data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 523–532. ACM, 2009.

- [17] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.
- [18] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on Twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30. ACM, 2010.
- [19] Corinna Cortes and Vladimir Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.
- [20] W Bruce Croft, Howard R Turtle, and David D Lewis. The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 32–45. ACM, 1991.
- [21] Anthony C Davies, Jia Hong Yin, and Sergio A Velastin. Crowd monitoring using image processing. *Electronics & Communication Engineering Journal*, 7(1):37–47, 1995.
- [22] Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.
- [23] Lev Grossman. Iran protests: Twitter, the medium of the movement, <http://www.time.com/time/world/article/0,8599,1905125,00.html>. accessed on 17 june 2009. *Time Magazine*, 17, 2009.
- [24] Qi He, Kuiyu Chang, and Ee-Peng Lim. Analyzing feature trajectories for event detection. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 207–214. ACM, 2007.

- [25] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in information retrieval*, pages 50–57. ACM, 1999.
- [26] Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank. A survey on visual surveillance of object motion and behaviors. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34(3):334–352, 2004.
- [27] Xia Hu, Nan Sun, Chao Zhang, and Tat-Seng Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 919–928. ACM, 2009.
- [28] Bernardo Huberman, Daniel Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *Available at SSRN 1313405*, 2008.
- [29] Ramesh Jain and Pinaki Sinha. Content without context is meaningless. In *Proceedings of the international conference on Multimedia*, pages 1259–1268. ACM, 2010.
- [30] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we Twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.
- [31] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: An analysis of a microblogging community. In *Advances in Web Mining and Web Usage Analysis*, pages 118–138. Springer, 2009.
- [32] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Link Mining: Models, Algorithms, and Applications*, pages 337–357. Springer, 2010.

- [33] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web*, pages 591–600. ACM, 2010.
- [34] Ryong Lee, Shoko Wakamiya, and Kazutoshi Sumiya. Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, 14(4):321–349, 2011.
- [35] Thomas Lento, Howard T Welsler, Lei Gu, and Marc Smith. The ties that blog: Examining the relationship between social ties and continued participation in the wallop weblogging system. In *3rd Annual Workshop on the Weblogging Ecosystem*, page 12. Citeseer, 2006.
- [36] Zhenhui Li, Bolin Ding, Jiawei Han, Roland Kays, and Peter Nye. Mining periodic behaviors for moving objects. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1099–1108. ACM, 2010.
- [37] Yu-Ru Lin, Hari Sundaram, Yun Chi, Jun Tatemura, and Belle Tseng. Discovery of blog communities based on mutual awareness. In *in: Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem*. Citeseer, 2006.
- [38] Xueliang Liu and Benoit Huet. Heterogeneous features and model selection for event-based media classification. In *Proceedings of the 3rd ACM conference on International Conference on Multimedia Retrieval (ICMR)*, pages 151–158. ACM, 2013.
- [39] Xueliang Liu, Benoit Huet, and Raphaël Troncy. EURECOM@ MediaEval 2011 Social Event Detection Task. In *MediaEval*, 2011.
- [40] Xueliang Liu, Raphaël Troncy, and Benoit Huet. Finding media illustrating events. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 58. ACM, 2011.
- [41] Xueliang Liu, Raphaël Troncy, and Benoit Huet. Using social media to identify events. In *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, pages 3–8. ACM, 2011.

- [42] BPL Lo and SA Velastin. Automatic congestion detection system for underground platforms. In *Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on*, pages 158–161. IEEE, 2001.
- [43] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.
- [44] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444, 2001.
- [45] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web*, pages 101–110. ACM, 2008.
- [46] Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th international conference on World Wide Web*, pages 533–542. ACM, 2006.
- [47] Donald Metzler and W Bruce Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.
- [48] Matthew Michelson and Sofus A Macskassy. Discovering users’ topics of interest on Twitter: a first look. In *Proceedings of the fourth workshop on analytics for noisy unstructured text data*, pages 73–80. ACM, 2010.
- [49] Keita Moriya, Shiori Sasaki, and Yasushi Kiyoki. A dynamic creation method of environmental situation maps using text data of regional information. In *DEIM Forum*, pages B1–6, 2009.
- [50] Bonnie A Nardi, Diane J Schiano, Michelle Gumbrecht, and Luke Swartz. Why we blog. *Communications of the ACM*, 47(12):41–46, 2004.

- [51] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *ICWSM*, 11:70–573, 2011.
- [52] Irish Central Statistic Office. Census 2011 boundary files <http://www.cso.ie/en/census/census2011boundaryfiles/>, date accessed: 2013-06-01.
- [53] Ramirez Nuria Oliver and Qiankun Zhao. A method of characterizing a social network communication using motifs, May 10 2012. WO Patent App. PCT/EP2012/058,600.
- [54] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.
- [55] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.
- [56] IBM Smart Planet. On a smarter planet, we want to change the paradigm from react to anticipate, <http://www.ibm.com/smarterplanet/>, date accessed: 2013-06-01.
- [57] John C Platt. AutoAlbum: Clustering digital photographs using probabilistic model merging. In *Content-based Access of Image and Video Libraries, 2000. Proceedings. IEEE Workshop on*, pages 96–100. IEEE, 2000.
- [58] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.
- [59] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from FLICKR tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 103–110. ACM, 2007.

- [60] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web*, pages 851–860. ACM, 2010.
- [61] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
- [62] Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, and Jon Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51. ACM, 2009.
- [63] Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. Event Detection and Tracking in Social Streams. In *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.
- [64] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in Twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM, 2010.
- [65] SA Velastin, JH Yin, AC Davies, MA Vicencio-Silva, RE Allsop, and A Penn. Automated measurement of crowd density and motion using image processing. In *Road Traffic Monitoring and Control, 1994., Seventh International Conference on*, pages 127–132. IET, 1994.
- [66] Mathias Versichele, Tijs Neutens, Stephanie Goudeseune, Frederik Van Bossche, and Nico Van de Weghe. Mobile mapping of sporting event spectators using bluetooth sensors: Tour of flanders 2011. *Sensors*, 12(10):14196–14213, 2012.
- [67] Chong Wang, Jinggang Wang, Xing Xie, and Wei-Ying Ma. Mining geographic knowledge using location aware topic model. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 65–70. ACM, 2007.

- [68] Xuerui Wang and Andrew McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.
- [69] Jianshu Weng and Bu-Sung Lee. Event Detection in Twitter. In *ICWSM*, 2011.
- [70] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.
- [71] Yiming Yang, Tom Pierce, and Jaime Carbonell. A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 28–36. ACM, 1998.
- [72] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in information retrieval*, pages 334–342. ACM, 2001.
- [73] Qiankun Zhao, Prasenjit Mitra, and Bi Chen. Temporal and information flow based event detection from social text streams. In *AAAI*, volume 7, pages 1501–1506, 2007.