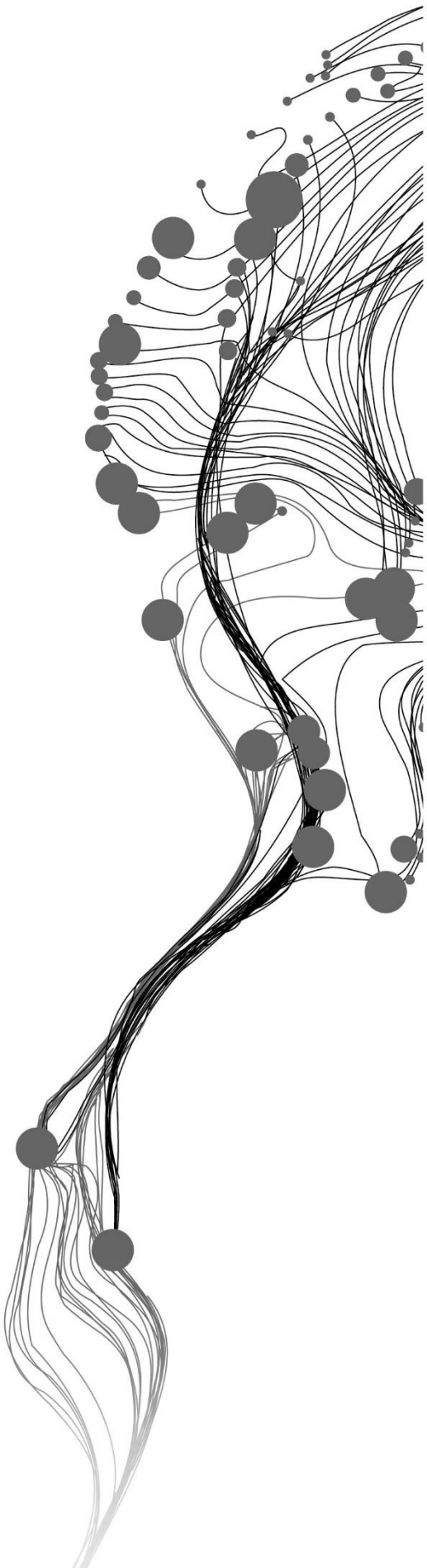
AN APPROACH TO LOCALNESS ASSESSMENT OF SOCIAL MEDIA USERS

YAN ZHANG
25 February, 2019

SUPERVISORS:
Dr. F.O.Ostermann
Dr. C.P.J.M. van Elzakker



AN APPROACH TO LOCALNESS ASSESSMENT OF SOCIAL MEDIA USERS

YAN ZHANG
Enschede, The Netherlands, 25 February, 2019

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Dr. F.O.Ostermann
Dr. C.P.J.M. van Elzakker

THESIS ASSESSMENT BOARD:
Prof. Dr. M.J. Kraak (Chair)

Dr. J. Krukar (External Examiner, Institute for Geoinformatics, Spatial Intelligence Lab, WWU Münster)

ABSTRACT

To make the utmost of people's local knowledge in smart city construction, identifying both people who related to one city and the relationship between people and city is necessary. Social media data is a potential data source of local knowledge, and identification of users who are related to one city is the precondition to extract local knowledge from social media data. Localness, which is defined as result of accumulating life experiences in a local environment, represents the relationship between people and cities and indicates potential local knowledge of people. Localness types mean the different possibilities of people's localness, and localness types include long-term resident, temporary or short-term resident, seasonal resident, non-local commuter, visitor and tourist.

The aim of this study is to design an approach to assess localness of social media users. User features are devised mainly from three perspectives to reflect localness. Temporal features answer how long one user stayed in the city and when the users were in the city, spatial features describe users from activity scope, activity concentration and tourism interest aspects, and social features represent the connection between social media users and local society. Other useful features include user-defined location in profile and language. User features and thresholds compose conditions for each localness type. More reliable features are selected as strong conditions. Conditions are combined to select users step by step from strict to loose selection conditions and all selected users are assigned one localness type as the output of the approach.

Twitter data in the London region are used as a case study to implement the approach. About 86% of Twitter users' localness can be assessed. Long-term residents and temporary/short-term residents account for 29% and 22% respectively among all assessed users. Compared with ground truth data, the accuracy of localness assessment is 69%. Localness assessment of long-term residents is the best of all the localness types, but seasonal residents and visitors have a relatively lower performance based on F1-measure. In the implementation, the assessment based on strict selection conditions shows better accuracy.

It's difficult to find clear relationship between features and localness types due to limited information from social media data, especially for spatial features. The selection of strong conditions and the thresholds used in strong conditions have a great influence on the assessment result. The approach is generic for geo-social media data and cities, and it works well in the case study. However, more ground truth data is needed to fit data sets and devise more reliable conditions.

Keywords

Localness, Social media user, Local knowledge, Twitter, London

ACKNOWLEDGEMENTS

My greatest thanks go to my first supervisor Dr. F.O.Ostermann and my second supervisor Dr. C.P.J.M. van Elzakker for their effort on encouragement, guidance and commenting through the whole research. Your rigor and speciousness not only helped me to complete the thesis, but also pushed me to grow up. I also thank all ITC staff members and all my classmates for their help and goodwill.

My deep gratitude towards of my parents for the great supports from them. My special thanks to my boyfriend, my best friend, my sister and anyone who trusted me, encouraged me and comforted me in the past year. I also thank my friends in ITC for their food and companion. Special gratitude for anyone who have ever given me a hug in the past year.

Thanks for myself for coming out of the darkness and the loneliness and beginning to grow up.

Yan Zhang
Enschede, the Netherlands.
February 2019.

TABLE OF CONTENTS

1. Introduction 1
   1.1. Background 1
   1.2. Research identification 2
      1.2.1. Research Objectives 3
      1.2.2. Research Questions 3
   1.3. Thesis structure 3
2. Localness studies 5
   2.1. Localness Definition 5
   2.2. Local knowledge and localness 6
   2.3. Localness of Social Media Users 7
   2.4. Localness and Mobility 8
3. Individual localness and types 10
   3.1. What is Individual Localness? 10
   3.2. Relationship between Localness and Mobility 11
   3.3. Localness Properties 12
   3.4. Localness Types 13
4. Localness assessment approach 17
   4.1. Data Collection and Filter 17
   4.2. User Feature Extraction 18
      4.2.1. Temporal feature 18
      4.2.2. Spatial feature 20
      4.2.3. Social feature 22
      4.2.4. Other feature 24
      4.2.5. Summary of user features 25
   4.3. Rule-based Localness Assessment 25
      4.3.1. Conditions 25
      4.3.2. Sequence 28
5. Case Study-Twitter Data in London 30
   5.1. Study Area and Data 30
      5.1.1. Study Area 30
      5.1.2. Dataset Description 31
      5.1.3. Sampling and Data Filtering 32
      5.1.4. Target Localness Types 36
   5.2. Feature extraction 37
      5.2.1. Feature extraction implementation 37
      5.2.2. Results 39
   5.3. Localness assessment 43
   5.4. Validation of localness assessment result 46
      5.4.1. Ground Truth 46
      5.4.2. Validation and discussion 48
6. Discussion 54
   6.1. Localness definition and types 54
   6.2. Localness assessment approach 55
7. concLusion and recommendations 58
   7.1. Conclusion 58
   7.2. Recommedations 60

REFERENCE 62

APPENDIX 67

LIST OF FIGURES

Figure 3.1: Concepts relationship in Chapter 3 10
Figure 3.2: The relationship between localness and mobility 11
Figure 3.3: Localness Properties 13
Figure 4.1: Overall process of localness assessment approach 17
Figure 4.2: Flowchart of temporal feature extraction 20
Figure 4.3: Flowchart of spatial feature extraction 22
Figure 4.4: Flowchart of social feature extraction 24
Figure 4.5: Localness assessment sequence 28
Figure 5.1: Map of the study area 30
Figure 5.2: Filters in data pre-processing 34
Figure 5.3: Box plot of tweeting frequency 36
Figure 5.4: Histogram of tweeting frequency 36
Figure 5.5: Flowchart for spatial feature extraction in the case study 39
Figure 5.6: Histogram of overall duration of Twitter users 40
Figure 5.7: Histogram of short duration of Twitter users 40
Figure 5.8: Histogram of the maximum interval of Twitter users 41
Figure 5.9: Histogram of average visit time of Twitter users 41
Figure 5.10: Histogram of tourist attraction proportion of Twitter users 42
Figure 5.11: Histogram of ellipse area 42
Figure 5.12: Histogram of core point proportion 43
Figure 5.13: Histogram of local follower proportion of Twitter users 43
Figure 5.14: Pie Chart of percentage of localness type 45
Figure 5.15: Pie Chart of percentage of localness steps 45
Figure 5.16: Scatter plot of maximum interval and duration of users with unknown localness 46
Figure 5.17: Spatial Distribution of typical users' tweet points 48
Figure 5.18: Confusion Matrix of localness assessment result 50
Figure 5.19: Confusion Matrix of each step in localness assessment 51

LIST OF TABLES

Table 2.1 Differences between local knowledge and professional knowledge....6
Table 3.1: Localness type description.... 16
Table 4.1: User features and feature description.... 25
Table 4.2: Conditions of user features for localness types.... 26
Table 4.3: Condition combinations used in step 3 of localness assessment.... 29
Table 5.1: Attributes in the data set.... 32
Table 5.2: Description of Dataset.... 33
Table 5.3: Percentiles of Ellipse Areas.... 42
Table 5.4: The number of assessed users for each localness type in each step.... 44
Table 5.5: Maximum interval and duration of users with unknown localness.... 46
Table 5.6: User number of each localness type.... 49
Table 5.7: Confusion Matrix of a binary case.... 49
Table 5.8: Evaluation measures for each localness type.... 50
Table 5.9: Ground truth of users assessed as unknown.... 52

1. INTRODUCTION

1.1. Background

In the last few decades, the tide of constructing smart cities has swept the globe and cities have become more efficient with the aid of Information Communication Technology (ICT) (Silva, Khan, & Han, 2018). A smart city architecture should offer users digital, efficient and reliable services and it usually consists of five components: application plane, sensing plane, communication plane, data plane and security plane (Habibzadeh, Soyata, Kantarci, Boukerche, & Kaptan, 2018). In these components, applications are designed to meet the social needs and the rest of smart city architecture is determined by the technical requirements of applications. Typical applications includes smart environment, smart home and smart building, smart surveillance and smart transportation (Habibzadeh et al., 2018). Although the wide application of technologies is the foundation of realizing a smart city, the social infrastructure such as the awareness and commitment of citizens has significant influence on the quality of life (QoL) in the city (Silva et al., 2018), and the QoL should be one of the main attributes of smart cities (Mohanty, Choppali, & Kougianos, 2016). In ideal conditions, people in smart cities should have a higher level of qualification, creativity, open-mindedness and more participation in public life (Giffinger & Gudrun, 2010). The intelligence of smart cities should reside in the combination of technologies and human brains (Nam & Pardo, 2011). But to use the intelligence of people who live in smart cities, there are some questions: which kind of information from people can be used in smart cities, where can we get the information and who can be the supplier of the information?

Local knowledge of citizens can be considered as important information which can contribute to smart cities. Local knowledge is the information directly related to the local contexts or settings, including the knowledge of "specific characteristics, circumstances, events, relationships and the important understandings of their meaning" (Corburn, 2003, p. 421). In urban contexts, local knowledge is helpful to solve some social problems in some fields, such as public health, environment monitoring, community governance and urban planning (Black & McBean, 2016; Brabham, 2009; Díez et al., 2018; Kim, 2016; Scott, 2015). However, due to the intangibility of local knowledge, obtaining itis not easy for the municipal politicians, urban planners and anyone who needs the knowledge related to one local context.

Social media data can be one possible data source of local knowledge. According to Kaplan & Haenlein (2010, p. 61), social media are a group of Internet-based applications build on the Web 2.0, and people can create and exchange user-generated content with them. Social media platforms have millions of users and the users create content any time any place. Due to the open and public nature of social media, it is possible to collect massive data from social media and some social media provide APIs for people who want to use the data on platforms such as Twitter and Flickr. Georeferencing and geocoding make citizens become human sensors to provide geographic information about local activities and life (Goodchild, 2007). Social media that contain both users' comments and geographic information can be called geosocial media (Zhang & Feick, 2016). From geosocial media data, local knowledge can be discovered from the combination of what people posted and the locations where these contents are related to. For instance, Konsti-Laakso (2017) used Facebook data to reveal security issues and events in a neighbourhood. After the determination of the data source, there is another problem: local knowledge points to the information related to specific areas, but how to identify the people who can provide the information from millions of users of social media?

The identification of people related to one area and the relationship between people and the area are necessary for local knowledge extraction from social media data. A common assumption used in social media studies is that users can be considered as local people from where they posted tweets, but in fact, it is highly possible that users post some comments during travelling to other places (Johnson, Sengupta, Schöning, & Hecht, 2016). So, only the appearance of one person in one area cannot prove that the person is local in the area. It is just the evidence that the person has a relationship with the area. The relationship between one person and one area indicates how much local knowledge the person can provide. Undoubtedly, compared to short-time visitors and tourists, long-term residents can often provide more reliable local knowledge about the city Human mobility leads to a diversity of this relationship, especially in global cities. People may visit another area for various purposes such as education, health care, work, leisure and so on. Once they visit one area, they build a relationship with that area and this relationship may change over time. To meet the needs of globalized production, people with specialized skills tend to cluster in a limited number of cities, and the populations of global cities are changing frequently (Sassen, 2001). Those global cities usually have large populations and more resources which means that they have a stronger motivation to construct smart cities and gather local knowledge from people who have a relationship with these cities.

In previous research, the relationship between people and one area was identified by binary classification and the diversity of this relationship was ignored (Andrienko, Andrienko, Fuchs, & Jankowski, 2016; Grace et al., 2017; Ostermann et al., 2015). In such a binary classification, social media users are classified as local or non-local. Identification based on the location field in the user profile is the simplest way, but this information is not very reliable (Hecht, Hong, Suh, & Chi, 2011). They reported that over one-third of Twitter users provided fake locations, words not related to their location or did not enter anything in the location field and some users entered multiple locations. All these noises decrease the credibility of location information in the user profile, so this information is not enough for identifying the relationship between social media users and areas. Another commonly used method is classifying by temporal criteria and authors usually use a simple filter to separate local people from visitors. For example, the filter used in Andrienko's work (2016) is that if the time span between the first and the last tweets is longer than 100 days the user can be considered as a local person, and Li, Goodchild, & Xu. (2013) used 10 days as the threshold in the filter. Johnson et al (2016) summarized four common criteria for local people identification: whether one user stays at least n-days in the region, where most comments of the user are posted, where is the home location of the user and location information in the user's profile. But they only used single criteria when they identified local people and they did not consider the diversity of the relationship between people and one area either.

"Localness" is a noun which refers to the quality or state of being local (Meriam Webster, 2019). The state of being local in one area for one person is one representation form of the relationship. The purpose of this relationship identification is extracting local knowledge, so to fit this purpose the term "localness" should be redefined based on the generic definition and on local knowledge. Localness should be an attribute of people to represent the relationship between people and areas. Considering the diversity of the relationship, the localness in this study should have the ability to indicate potential local knowledge of people and distinguish groups of people based on the local knowledge they may have. After such a localness definition and taking social media data as the source of local knowledge, an approach to localness assessment of social media users is needed to identify the users based on the potential local knowledge they have.

1.2. Research identification

This section is to specify the research objectives of this study and the corresponding research questions of each objective. The research objectives are built up based on the research problem, as described in the previous section, and the objectives are achieved by answering the research questions following them.

One localness assessment approach for social media users will be the outcome of this study. First, the existing localness definition and related criteria will be reviewed and be used as the basis of new localness definition and assessment. Localness will be redefined from a local knowledge perspective, and to represent the relationship between people and areas localness types will be specified. In the approach, users are represented by their features and conditions for each localness types will be designed. The localness assessment is based on the comparison of user features and these conditions. To evaluate the approach, it will be implemented in one case study to check the shortcomings and application effect of this approach.

1.2.1. Research Objectives

The overall objective of this study is to design an approach to assess the localness of social media users and implement the approach in one global city.

To achieve the overall objective, it is split into four sub-objectives:

1. To evaluate existing localness definitions and assessment criteria
2. To define individual localness and specify localness types
3. To design an approach to assess the localness of social media users
4. To implement and evaluate the approach using real-world data in a global city

1.2.2. Research Questions

Research questions related to sub-objective 1:

1. How do related works define localness?
2. Which are the criteria used in related works to assess the localness of individuals?

Research questions related to sub-objective 2:

1. How to define localness of individuals?
2. How to conceptualize different types of individual localness?

Research questions related to sub-objective 3:

1. Which user features can be used to determine the localness type of users?
2. How to assess the localness of a social media user based on user features?

Research questions related to sub-objective 4:

1. To what extent can the approach assess the user's localness correctly?
2. What are the application conditions and the limitations of the approach?

1.3. Thesis structure

The research consists of four objectives and the thesis is organized by the sequence of objectives.

After the background and research identification in Chapter 1, the first objective is addressed in Chapter 2 by a literature review of existing localness definitions and criteria.

As for the second objective, localness will be defined based on local knowledge, and the properties and types of localness are specified in Chapter 3.

The description of localness is applied in Chapter 4 to assess the localness of social media users. The user feature extraction in the approach is based on the localness properties and available information in social

media data. Localness assessment in the approach is based on user features and localness types. Some conditions are designed for each localness types, and whether one user features can meet conditions of one localness type will determine whether one user can be assigned as this localness type.

To implement and evaluate the approach, Chapter 5 presents a case study using Twitter data in Greater London. For the data set, the user features are calculated, and the approach designed in Chapter 4 will be used. The process of approach implementation and the result will be interpreted and discussed, and the evaluation of the case study results is based on a small sample which includes the labelled localness of users.

Chapter 6 contains the overall conclusion and discussion of this study and the recommendation for further research.

2. LOCALNESS STUDIES

In this chapter, the first research objective will be addressed: reviewing and evaluating existing localness definitions and assessment criteria. First, localness definition in different fields are reviewed, especially localness used in user-generated-content. Second, to define localness based on local knowledge, definitions of local knowledge are summarized. Third, researches use localness or local people identification are reviewed, and existing assessment criteria are evaluated. Fourth, to describe localness systematically, one commonly used concept "mobility" is used in this thesis and mobility studies are reviewed in the last section of this chapter.

2.1. Localness Definition

Localness as a relatively common term has been used in several fields. In the field of economy, localness is used as a statement related to the local market and it is comparable to globalness or internationalization. For example, Persky and Wiewel (1994) used localness as an indicator to demonstrate the trend of being more local in global cities. Swoboda, Pennemann and Taube (2012) used localness and globalness to analyse how consumers perceive retail brands. Schmitt, Dominique, and Six (2018) proposed five criteria to assess the degree of localness of food production. In computer science, Tu, Su, and Devanbu (2014) used localness to represent a property of source code meaning that the code contains local regularities. Ballatore, Graham, and Sen (2017) defined a localness indicator as the ratio between the number of local Google search results and the total number to reveal the unevenness of digital geography. In politics, Gschwend, Shugart, and Zitte (2009) use the term "localness" to imply the phenomenon that some legislators have a subjective local focus and please local constituents.

Although these localness usages or definitions are proposed by authors in different fields, all of them try to represent the relationship between their object of studies and the local environment. Usually, they have a concept or situation as an opposite to show lesser degrees of localness such as globalness or the situation with less local connections.

Compared to the applications of localness in the above fields, the term "localness" is being used more in the analysis of user-generated content (UGC), especially in the analysis of Volunteered Geographic Information (VGI). Hecht et al. (2010) paid attention to the "localness" of participation across entire UGC repositories and introduced spatial content production models to describe the proportion of locals in repositories of Flickr and Wikipedia. Tahara and Ma (2014) used the term "localness" as the main indicator to extract regional terms from linguistic features of tweets in ae method of local Twitter search. In their work, the value of localness is the product of four indicators: the frequency of appearance of one regional term, whether users in other areas post the term, the number of users in the target area posting the term and the number of days users in the target area post the term. Regional terms with high localness values are used in the local area feature vector to identify local users. This is an application of localness in social media user identification, but the localness is relative to areas instead of users. Sen et al (2015) treated localness as a property of VGI and localness in this work was used to demonstrate how much VGI about one place was originating from this place. To study geographic content biases in VGI, the authors collected Wikipedia editions in 79 different languages and examined the data from two relationships: 1.the geographic articles and the locations of their editors; 2. the relationship between the geographic articles and the location of sources cited in the articles. In Johnson's work (Johnson et al., 2016), localness is a common assumption in social media. Under the localness assumption, people who post contents with geographic information in

social media will be considered as locals to the region of corresponding geographic information. Johnson also examined this assumption based on some existing local user identification criteria and localness was also treated as the indicator in this examination which was calculated as the relative proportion of users who were classified as locals. Following Johnson's work, Kariryaa, Johnson, Schöning, & Hecht (2018) defined localness more clearly by identifying the definition of "local" in earlier works. They present three definitions of localness: a person is local only to the region where he lives, a person is local to the region where he votes, and a person is a local to a region if he has enough knowledge about the region. Huang and Wang (2016) thought that the localness of users depends on the answer to the question whether the user is a local resident in one city and they identified locals using the local attractiveness of venues.

In summary, so far, localness is treated as an abstract noun to show how local the study objects are. The study objects can be a whole, like a VGI repository and a group of social media users. Localness is then the proportion of local elements in the whole. The study objects can also be individuals, like a social media user and a regional term. In such a case, localness means the relationship between the individual and one area, i.e. whether one user is local to one area and to what extent one term can represent one area.

In this thesis, the localness of individuals represents the relationship between people and one area. It is defined to extract local knowledge, and the diversity of the relationship can be embodied in the localness properties and localness types which I will describe in Chapter 3.

2.2. Local knowledge and localness

As mentioned in section 1.1, local knowledge of citizens is an important information source which can contribute to smart city construction. One motivation of this study is to find social media users who might have some local knowledge about one area and identifying localness of users should be one way to achieve that. Local knowledge and the localness are closely related based on the same local environments: the localness should be the indication of some types of local knowledge and the accumulation of local knowledge is one requirement of localness changing. Therefore, knowing the exact meaning of local knowledge is necessary for the definition of localness.

The term "local knowledge" has been defined by different authors. Lindblom and Cohen (1979, p. 12) characterized local knowledge as "common sense, casual empiricism, or thoughtful speculation and analysis" . Geertz (1983, p. 75) defined local knowledge as an organized body of thought which is "practical, collective, strongly rooted in a particular place and based on immediacy of experience". Corburn (2003, p. 421) defined local knowledge by comparing local knowledge with professional knowledge and that comparison is shown in Table 2.1. He indicated that both geographically located community members or the members in specific context groups can hold local knowledge, and local knowledge can come from the tactile and emotional experiences in their lives.

Table 2.1 Differences between local knowledge and professional knowledge

	Who holds	How to gather	How to be credible	How to be tested
Local Knowledge	Community members: geographically located or contextual to specific groups	Life experience	Tactile and emotional experiences	Public forums: public narratives, community stories, street theatre
Professional Knowledge	Members of a profession, discipline, or research institution	Experimental methods and disciplinary tools	Scientific discussion	Peer review

Therefore, local knowledge about one specific area comes from people's life experiences in the area and this is the basis of localness definition in this thesis.

2.3. Localness of Social Media Users

The localness assessment problem in using geosocial media data is a relatively new problem and only attracted attention from some authors in recent years as follows:

Some studies just want to focus on the content from local people but do not try to solve the local people identification problem and used a simple way to filter the locals. For example, Andrienko et al (2016) separated the local people from visitors using a simple filter including the time span of the tweets. Ostermann et al (2015) used the overall duration and a distinct day number of tweets to identify residents and filtered out the tourists using a 30 day time window. Kumar, Bakhshi, Kennedy, & Shamma (2017) used the same way to classify the locals and tourists.

Except for the traits of user posting behaviours, other data can also be useful in this field. Grace et al (2017) came up with the "Social Triangulation" method to identify the local citizens using the local organizations they follow on Twitter, assuming that if one user follows more local organizations he is more likely to be assigned as a local. Huang, Wang, & Tao (2017) used the online check-in data for venues in social media to identify local people and nonlocal people, and assumed that local people visit more venues in the city. Huang, Wang, & Zhu (2017) proposed a framework called "Diversified Local Users Finder" to identify the set of local users from check-in data. They estimated users' home locations with check-in traces by an unsupervised framework, and then computed diversity scores by geographical distance between users' home locations, and local users were identified by maximizing their diversity scores. Tahara and Ma (2014) proposed a method for local Twitter user identification based on linguistic features of tweets. They constructed local area vectors and user vectors based on the extracted regional terms and then calculated the Cosine similarity of local area vectors and user vectors. Local users are the users of whom the vector is similar to the local area vector.

Johnson et al (2016) verified the localness assumption and summarized four common criteria (n-day, plurality, geometric median and location field) which had been used to determine localness in other works. They found that about 25% of the users who are not local posted tweets in the study city. In the paper, they only used the four criteria separately and compared the results. They used four datasets to test the four criteria and the comparison of the results showed that each single criterion did not have a stable performance in the four datasets and that there is a substantial disagreement between the four criteria. Considering the criteria of localness in parallel, the authors suggested that a combination of multiple criteria may have a more robust performance. Following Johnson's work, Kariryaa et al. (2018) indicated that the definition of people localness should contain information about where people currently live, where people currently vote and which places people are familiar with. They used existing criteria mentioned in Johnson's work to test the new definitions and ground truth data used in the tests are collected from Twitter's ad platform.

From this literature review of localness assessment of social media users, we may conclude that there are some significant shortcomings in the existing methods. First, existing localness assessment methods only focus on the identification of either local or non-local. The relationship between people and areas is simplified to only two choices and any diversity is ignored. So, the existing methods cannot distinguish, for example, long-term residents, short-term residents, visitors and any other potential situations of user localness. Second, in the existing methods only single criteria were applied. According to Johnson et al. (2016a), the results of using single criteria are not reliable and the combination of multiple criteria can lead to a more robust performance. Third, existing methods did not take full advantage of all characteristics of

social media users. Some user characteristics may be useful but were never taken into account, such as the temporal distribution of posting behaviours, the spatial distribution of locations, local social networks, and so on.

In Chapter 4, an approach to localness assessment will be designed, with the aim to make up, to a certain extent, for the above shortcomings of existing methods.

2.4. Localness and Mobility

Human mobility is one of the reasons of the diversity of localness. To some extent, the diversity of localness can measure a population's heterogeneity, which is partly caused by human mobility. At the individual level, mobilities are the "means to combine goals in space" , and these spatial behaviours combined over time can demonstrate the life course trajectories of individuals (Bell & Ward, 2000). Both human mobility and localness represent the relationship between humans and space: the former focuses on the location changes over time from the individual perspective, while the latter can be considered as a person's specific states at certain times from the location perspective. Because of the complexity of the relationship between humans and space, human mobility has various forms and the localness also has different types. So, a review of mobility studies can be helpful for the understanding of localness.

Bell and Ward (2000) compared permanent migration and temporary mobility in key concepts (usual residence and return) and three dimensions (duration, frequency and seasonality). Then they used the boundaries in space and time to classify all population movements into detailed mobilities such as commuting, seasonal work and so on. This is the first literature comparing permanent migration and temporary mobility in a systematic way. Williams and Hall (2000) linked tourism and migration and specified some migration forms, such as labour migration, retirement migration and so on. Montanari (2005) proposed an approach to analyse the mobility flows at territorial level in phases of local development related to the social and economic development of the territories like detailed mobility forms related to labour. Williams et al. (2012) illustrated the pattern of population centralization and decentralization in urban contexts by analysing the migration in Portsmouth UK, and paid attention to the fluidity of urban populations which manifested the importance of temporary population movements, but the analysis of mobility forms was limited by the secondary data used. King (2012) reviewed the migration theory from a geographic perspective and summarized a typology of migration which is a relatively comprehensive typology and can be the basis of mobility form analysis. Novy (2018) proposed a pentagon of mobility from the place consumption perspective to illustrate the diverse tourism mobilities.

Although the emphases in these works are dissimilar, the classifications are consistent, and the main factors considered in the mobility form conceptualization are time, purpose and path. Permanent mobility and temporary mobility are the most common classifications using the time factor. Temporary mobility indicates that people will stay in the destination of the movement for varying durations and return thereafter, while permanent mobility means a change of usual residence and the last relocation (Bell & Ward, 2000). Migration can be considered as one result of permanent mobility and commuting is a special case of temporary mobility. The second mobility classification is production-related versus consumption-related movement. This classification is based on the purpose of the mobility: production-related mobilities occur for an economic contribution while the latter occurs to get a good or service. The distinction between them is fuzzy because production and consumption are concurrent in most cases. A path indicates the geographical relationship between origin and destination of movement, and the distance or geographical scale of movement is the main indicator (King, 2012; Williams et al., 2012).

Literature presents mobility forms based on different topics, so they all have advantages and limitations and none of them are comprehensive. Some mobility forms are put forward based on the specific purpose of mobilities, such as business, healthcare, second home or visits to relatives and this is the mainstream of the mobility form identification. In other studies, the forms are identified by temporary or spatial traits like (temporary) migration and internal/international migration. Moreover, all the identified forms are the result of a qualitive analysis of mobilities and lack a clear definition and threshold, which means that they cannot provide distinct mobility forms covering all the mobilities without overlap.

To find a systematic way to identify mobility forms, it is necessary to know the properties of mobilities. Karamshuk et al. (2011) reviewed the progress in the field of human mobility and classified the related findings along spatial, temporal and social properties. The spatial properties are related to the travel distance of people, the temporal properties pertain to the time and frequencies of the visit and the social properties are related to the social interaction between persons. These three properties can also connect to the mobility forms mentioned before. Therefore, the properties can be used as the three fundamental directions to identify mobility forms.

Since both localness and mobilities are a representation of relationships among space, time and humans, they will have similar properties and forms. The relationship between localness and mobilities will be explained in detail in the next chapter.

In this chapter, existing works related to localness definition and localness assessment of social media user are reviewed, and local knowledge definition and mobility studies are summarized to support the localness definition, properties and localness types in next chapter. In Chapter 3, the localness will be defined and described based on local knowledge.

3. INDIVIDUAL LOCALNESS AND TYPES

In this chapter, the second research objective will be addressed: defining individual localness and specifying localness types. Figure 3.1 shows the relationship among the concepts used in this chapter. Local knowledge and concepts about mobility have been defined clearly in existing works and they are used in the localness definition. To describe localness, the relationship between mobility and localness is explained first, and then the properties of localness are illustrated based on the localness definition and mobility properties. The localness type conceptualization is based on the localness definition and some mobility forms, and localness properties are used to describe localness types.

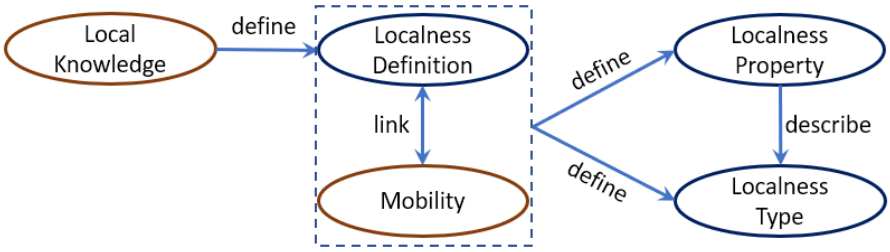


Figure 3.1: Concepts relationship in Chapter 3

3.1. What is Individual Localness?

As mentioned in section 2.2, local knowledge generated based on life experiences. People develop local knowledge over time in a given environment based on their experience and they hold local knowledge individually and dynamically (FAO, 2004). Not only do the original inhabitants of an area have local knowledge, but also all the people who have experience related to the area. For instance, migrants or visitors may hold local knowledge about this area. Because people develop and possess local knowledge individually, people in different conditions like age, gender, educational background, occupation, or socio-economic status have different kinds of local knowledge. Local knowledge can be broad (covering many aspects) and/or deep (knowing a lot about a single aspect), and the amount and types of local knowledge of one person are affected by a lot of factors which all of them are related to the generation of local knowledge.

The generation of local knowledge is closely related to how long people stay in one area and local knowledge is attached to the physical area where they have activities. For one city, tourists and long-term residents are likely to have different types of local knowledge, and commuters not living in this city and visitors staying in the city only one weekend also have different impressions of the city. In addition to the temporal perspective, the validity of local knowledge has obvious spatial boundaries. For instance, the information about an unsafe area in one city is useless for the people in another city unless they plan to visit that city. Moreover, based on different experience in the same area and period, people may generate local knowledge from different perspectives. One example can be that visitors with health care purposes will pay more attention to local medical information, while tourists will only possess local knowledge about local tourist attractions in which they are interested, and about other limited information related to their journeys.

In summary, life experiences in local environments is the source of local knowledge, and individuals accumulate their local knowledge through familiarity with the local environment from different perspectives based on different life experiences. Therefore, I define individual localness based on the local knowledge as result of accumulating life experiences in a local environment.

This definition should be elaborated on from three angles. First, localness is related to local knowledge, but different from the latter. Local knowledge is a concept that was developed to describe a kind of knowledge of individuals and this knowledge is about specific areas. Localness is a concept that describes the relationship between individuals and areas, and the relationship indicates the potential local knowledge of individuals from an area perspective. Localness can be helpful to identify the people who might have specific local knowledge, but this identification is the selection of individuals instead of local knowledge. Localness only indicates that some people are more likely to have specific local knowledge but cannot guarantee that. Localness can be a representation form of how people connect to one specific area and it can also be an attribute of individuals in the discussion of the relationship between people and one specific area. Second, to fit the mobile and globalized world, this definition assumes that one person can possess local knowledge about all areas he or she has ever stayed in, while the common explanation of local knowledge emphasizes knowledge only about the area people live. So, in this definition, the "local environment" may be variable for one specific individual, which means that one person can have a localness attribute for different areas and the value of such a localness attribute will depend on the life experiences in each of these areas. Third, to keep the definition generic, the local environment can be areas at different scales and the choosing of the localness scale depends on the objectives of different studies. Intuitively, a city will more easily make an impression on people than a bigger administrative region or a smaller unit like a postal district, and a city is the common choice when a person introduces his location to others. In addition, given the background smart and global cities in this study, choosing the city localness scale is appropriate. So, localness in the rest of the thesis means the localness at city scale.

3.2. Relationship between Localness and Mobility

To conceptualize localness types systematically, I will first link localness to concept "mobility" mentioned in section 2.4. Both localness and mobility are a representation of the relationship between humans and space. Considering the change of the human-space relationship over time, the temporal dimension is also an integral part of this relationship. Figure 3.2demonstrates how humans, space and time determine localness, as well as the relationship between localness and mobility.

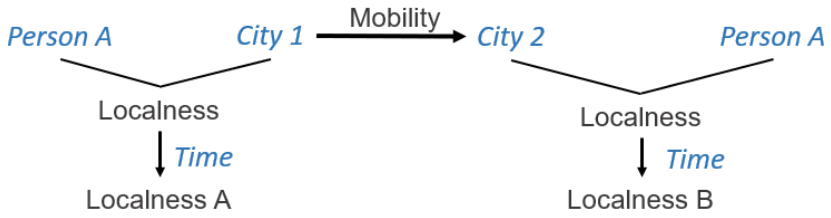


Figure 3.2: The relationship between localness and mobility

As for time factor, localness of this person for this city can change if he/she will stay there in the future or visit the city more times, because persons will accumulate more life experiences when they stay in one city for longer accumulated times. Mobility means change of the space factor and mobility between cities is the precondition for people to have life experiences in another city. Compared to localness relative to the original city, persons visiting a new city will have a different localness relative to the new city due to the life experiences in the city during the visit period. As for human factor, Figure 3.1 demonstrate the situation of one person, while different persons have different life experiences in one city then localness of persons are also different. Thus, localness as an attribute of persons can represent the relationship between persons and cities, and the change of any factor among humans, space and time can change the localness.

Mobility focuses on how people move, other than how people interact with one city. One person and one movement between two cities can determine one mobility. Except for permanent migration, every mobility has a return time and a visit time, and these time factors indicate the return mobility or the next mobility in the trajectory of one person. So, mobility represents the relationship between humans and space by describing the process of change of a person's location.

The relationships between localness and mobility are the following. First, localness can be considered as one attribute of persons and it represents a person's state, while mobility shows the process of movement and can be considered as actions of people instead of one feature of people. Second, mobility of people is the precondition for the situation that one person changes his/her localness for different cities by accumulating life experiences in those different cities. Third, except for permanent migration, the visit time in each mobility is consistent with the time persons are staying in one city for one trip. If one person is not a permanent resident migrating to one city during a period, every time he/she visits the city means one mobility of him / her to this city, and the more he/she visits it, the more life experiences about this city he/she has.

3.3. Localness Properties

Localness as a comprehensive result should have the ability to describe the life experiences of individuals from many aspects. The time one person staying in one city is one factor of an individual's localness, and his/her activities in the city, including visiting places inside the city, and the interaction with other people living in the city are also important aspects of individual localness. As mentioned in section 2.4, the properties of mobility include three aspects: spatial, temporal and social. Like the properties of mobility, localness of individuals can also be analysed from these three aspects.

These properties should be adjusted to fit the definition of localness. Mobility studies pay more attention to how people move between different cities while localness studies should concern about how people generate life experiences in one city. For temporal properties of localness, information about the duration of visits is needed and only the numeric result of visit frequencies is not enough. To get insight into how people stay in one city, the temporal distribution of people staying here should be described in detail, for instance through some statistics of interval and visit times of individuals. In addition, localness is closely related to local knowledge, but some types of local knowledge will decay over time, such as the location of some restaurants or recreation places. For example, one person might have lived in one city as a resident and had possessed some local knowledge about the city, but then he/she migrated to another city and did not stay in the original city for a long time anymore. In this case, the person's localness relative to the original city not only depends on the life experiences about the city but also the time elapsed after leaving the city will have an influence on it. Thus, there should be a temporal property to describe the last time one person stayed in one city and the duration of the period since he/she left the city will be important factors of an individual's localness. For the spatial properties, localness studies should focus on the geographical distribution of one person's activities inside the city instead of the travel distance of movement, and both the scope and the concentration of activity locations are the indication of localness. For social properties, the interaction with other people cannot describe social property of localness comprehensively. Localness studies should concern more attention to the overall social network of people and the common interests of people, because localness concerns more about final state of the relationship between people and city instead of how people build the relationship.

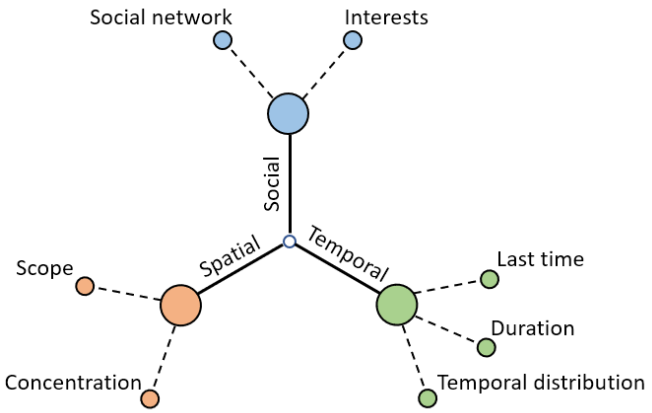


Figure 3.3: Localness Properties

Figure 3.3 shows the properties of localness. Localness should reflect the life experiences of people in a local environment from spatial, temporal and social perspectives and it is a comprehensive result to assess how local one person is. From a temporal perspective, localness indicates when a person stayed in this city, how long he stayed here, and what was the temporal distribution of his/her visit time if he/she visited the city more than once. From a spatial perspective, localness describes whether one person ever stayed in this city, the geographical scope of his life in the city and the degree of concentrated activities. From a social perspective, localness describes how one person interacts with people who live in the city. Social networks contain all friends, relatives and anyone the person knows. Interests can result in online or offline groups even if people do not know every member in the group personally and people in the same group may have similar activities, especially in the Web 2.0 era. Moreover, interests can reflect the socioeconomic status of people and the purpose of visits like education and health care which are an indication of which kinds of local knowledge they have.

3.4. Localness Types

Localness types mean the different possibilities of people's localness. Localness types are the result of a localness assessment of people and the types are also candidate values of localness attributes of people. The purpose of this section is to conceptualize types of localness. To simplify the localness types, I will ignore the situation that the last time people stayed in the city is so long ago that their local knowledge about the city became obsolete.

Intuitively, long-term residence is a necessary type of localness. Long-term residents in one city usually have usual resident in this city and almost all their activities happen in the city. From a localness perspective, people can get used to one city if they stay in the city for one year, because one year is enough for people to integrate into life in the city, explore the city and have regular life in the city. So, in this thesis one year is an import condition to identify long-term residents. Their activities have obvious centres which are their home location, work place or any place they visit at a higher frequency. They usually have visited more places in this city and constructed stronger social local networks than the following localness types.

Besides long-term residents, mobility to one city is a precondition for people who can possess local knowledge about this city. Furthermore, visiting a city can give people chances to gain life experiences in the city, and the longer one person stays there the more life experiences he/she may gain. As mentioned in section 3.2, visit time of mobility is as the same as the time people staying in one city and the total time one

person stayed in one city is an indication of life experiences of the person. So, mobility forms based on visit time provide an important reference for localness type conceptualization. Bell and Ward (2000) listed population mobilities in a time-space table. After dividing by time, mobilities are at four temporal levels: within one day (shopping, commuting), within one week (visit, excursions, health care and business travel), within one month (study and vacations) and within one year (seasonal work). These temporal mobilities can be a reference of localness type conceptualization.

For the mobility forms within one day, some other authors also mentioned similar forms in their works, such as shopping and leisure (Montanari, 2005). But most of these mobilities are parts of daily life for every person staying in one city and cannot be used to identify people with different life experiences accumulation levels (i.e. localness). However, given the city localness scale, commuting from other cities is a special kind of one-day mobility form. Some people may reside in one city but work in another city. Most of their activities about their daily life may happen near their home location in another city and they can only focus on the work-related activities which can lead to a relatively small activity scope near their work place in the city. So, these people may have different perception about the city compared to people who both live and work in the city. Another trait of these people is that they usually stay in their working city by daylight and on weekdays. The life experiences about the city and local social network they have depends on the time they live as commuters. Based on the above traits of non-local commuters, it is treated as one localness type.

Mobility forms with a short-term visit time may have different purposes such as health care, business, friend/relative visiting and so on. People visit the city to achieve their goals, and most of them will focus on their core activities. Compared to resident localness types, visitors do not have a usual residence and daily life experiences in the city. How long visitors stayed in the city is based on their visitor purposes and 30 days is a time limit used in existing works to distinguish visitors from residents(Girardin, Calabrese, Fiore, Ratti, & Blat, 2008; Vikas Kumar, Bakhshi, Kennedy, & Shamma, 2017; Sun, Fan, Helbich, & Zipf, 2013). In these works, the authors did not distinguish visitors from tourists, and they used one of these two terms to indicate all people who had short-term visits to one city. It's noteworthy that visitors may go to one city more than once based on their purposes, but they will not stay in the city for more than month in each visit. Both one-time visitor and more-than-once visitor are treated as visitor localness type in this thesis.

Among these short-term mobilities, tourism is a special one and some authors reported that tourism mobility has a significant influence on the processes of urban change and many other mobility forms (Novy, 2018; Williams & Hall, 2000). Tourists also have distinct characteristics that distinguish them from visitors with other purposes. Obviously, tourists usually pay attention to the tourist attractions in the city, stay in one city for less time. The number of tourists is large but the tourism population changes at any time. Lau et al. (2006) reported that for the tourists who spent several nights in Hong Kong they usually stay in the city for less than seven days. Combined Lau's finding with the mobility forms in the work of Bell and Ward (2000), I propose a visit time of tourists of seven days. Based on the above characteristics of tourists, treating tourists as a separate localness type is meaningful, especially for tourist cities.

Visit time of the other mobilities is from weeks to months and there are two possibilities in this situation: mobility for work or study and seasonal migration. Some people may migrate to another city for work or study, and they will live in the city as temporary residents for several months until they finish their work or study. Some people may move to the city and start life there, but they should be treated as short-term residents until they stay in the city long enough. Compared to visitors, these people will have some basic life experiences about city life, and they will accumulate this life experiences over time. Compared to long-term residents, these temporary or short-term residents usually have a smaller activity scope and a weaker local

social network because they do not have enough time to explore more places and know more local people in the city.

The other situation is seasonal migration. The reasons for seasonal migration might be vacation, seasonal work or a second home (Bell & Ward, 2000; King, 2012; Williams & Hall, 2000). People will go to another city seasonally and stay there for weeks or months and then they will return instead of staying in the city for a long time and becoming long-term residents. Seasonal residents will have moderate local social networks, stay in the city for months but less than one year and have an intermediate activity scope. Longitudinal data sets, including the movement or activity data for a group of persons for years, could be used to identify seasonal residents. Because of the special patterns of seasonal residents and the widespread existence of seasonal mobilities, the seasonal resident is one localness type that should not be overlooked.

Mobilities only focus on how people move between cities instead of how people live in one city and perceive the city, so the localness type conceptualization only based on the mobility will not be complete. For example, the localness of indigenous residents is not related to any mobility form. In addition, there should be a special localness type for people without life experiences in one city. Some people may never have been to a city but may know something about the city, and other people may even have never heard about it. All these people do not have any life experiences with the city, therefore, there should also be a localness type named "no experience".

To make sure the comprehensiveness of the localness type conceptualization, the types should be distinguished based on the properties of localness. Intuitively, spatial and temporal information about people's activities in one city can be used for localness assessment because the information is about measurable, closely related activities and can be divided meaningfully. Once a person visits a place in the city, there must be some corresponding spatial and temporal information related to where he/she visits, when he/she goes there and the summary of all the spatial and temporal information of activities can reflect the patterns of people's life. People with different localness types will have different life patterns in the city. Social properties are also useful in localness determination as long as they can be represented as comparable data. Local social network of one person is one of the results of life experiences accumulation, which can reflect the connection between people and local environments from a social perspective. Interest as the other aspect of social properties can indicate the source of different local knowledge types and can be the basis of a detailed typology for local knowledge discovery.

Table 3.1summarizes the localness types that can be distinguished. The description along localness properties elaborates localness types and enables the individual distinction by localness.

Table 3.1: Localness type description

Localness Type	Localness Type Description Based on Properties		
	Temporal	**Spatial**	**Social**
Long-term resident	>12 months	Wide activity scope, home & work place	Strong local social network
Temporary or short-term resident	1-12 months	Moderate activity scope, home & work place	Moderate local social network
Seasonal resident	1-3 month for each year	Moderate activity scope, home & work place	Moderate local social network
Non-local commuter	Periodic, working hours and work days	variable activity scope, work place	Variable local social network
Visitor	<30 days for one visitor, may visit more than once	variable activity scope, activity centre depends on visit purpose	Variable local social network
Tourist	<7 days, weekend & holiday	Small activity scope, near tourist attractions	Weak local social network
No experience	Have never been there	-	-

In chapter 3, I proposed the generic localness and localness properties, and for each localness potential I defined a localness type. The localness definition, properties and types can be used in any study related to local knowledge discovery and people identification by life experiences in cities. As mentioned in section 1.1, social media data is one possible data source of local knowledge and the localness assessment of social media users is helpful for local knowledge discovery from social media data. Therefore, in Chapter 4, I will design an approach to assess the localness of social media users.

4. LOCALNESS ASSESSMENT APPROACH

In this chapter, the third research objective will be addressed: designing an approach to assess the localness of social media users. Figure 4.1 demonstrates the overall process of the approach. The approach is divided into three steps: first, social media data will be collected as input data of this approach and the data will be cleaned by filters; second, user features of four categories will be extracted from the data; third, localness types will be assigned to each user based on the result of a comparison between user features and the localness type conditions designed in the approach. The output of following this approach is the allocation of localness types to social media users.

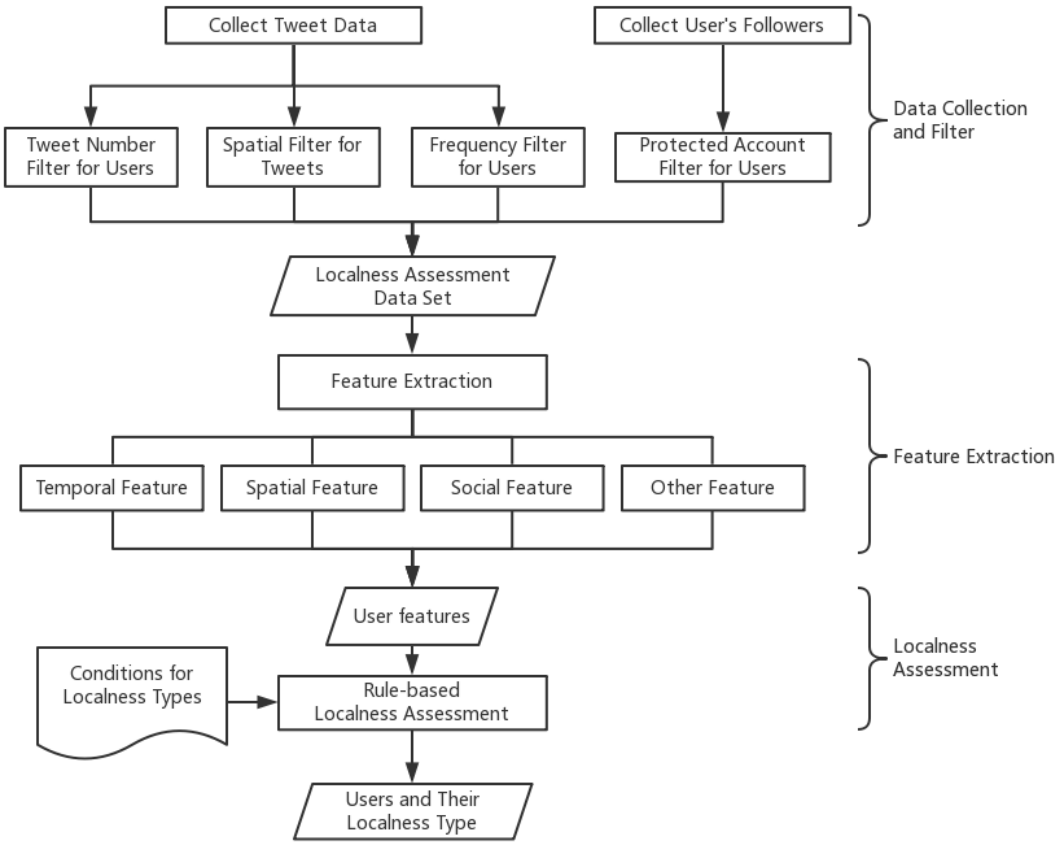


Figure 4.1: Overall process of localness assessment approach

4.1. Data Collection and Filter

To collect social media data, the use of Application Programming Interfaces (APIs), as provided by social media platforms, such as Twitter API or Flicker API, is to be preferred. When collecting social media data by APIs, time and space filters can be used and the data will be in a consistent format like JSON. There are also some libraries for accessing the APIs in common programming languages like tweepy[1] in Python. For the purpose of user localness detection, the time period of a dataset should be longer than the longest

[1] http://www.tweepy.org/

duration with which target localness types have been defined, or long enough to find out about the visit patterns of users. Otherwise, the identification will be not credible. In other words, if the target localness type is e.g. long-term resident, the dataset should contain data for at least a one-year period and if the target is a seasonal resident the dataset should cover at least two years to find the seasonal visit patterns as mentioned in section 3.4. The spatial filter should be used in data collection to make sure that all the social media contents in the dataset are located in the target area (in this study: a city), and using a spatial filter also means that the dataset will only contain geo-tagged postings and exclude other social media data without spatial information.

There are three additional data sources used in the localness assessment if all features will be extracted. The first one is the locations of tourist attractions in the target city and this will be used to measure the proportion of activities near to tourist attractions. The second one is the followers' information of social media users which will be the data source to measure the connection between target users and local people. Here, APIs provided by social media will be used again and using the names or IDs of target users as input, the names or IDs of one user's followers will be the output. The third one is the list of local organization accounts on the social media platform, and this will be used to measure the connection between users and local society.

After data collection, not all the data in the set will be useful for the study at hand. This means that the dataset will have to be cleaned. First, due to some mistakes in social media API's, the coordinates may not be located in the study area, so the spatial filter should be used again in this data cleaning step to remove those outliers. Second, to get enough information for each user, there should be a lower limit for the number of geo-tagged tweets. In this step, users who have less than three distinct post points will be discarded given the requirement of spatial feature extraction in next section. Third, in order to filter out the social media accounts which are not controlled by real persons or post excessive contents which are irrelative to user's life experiences, the frequency of tweeting behaviour will be calculated, outliers of frequency will be detected by box plot and the accounts with frequency outliers will be discarded. The last step in data cleaning is about the additional data source of the users' followers. Some users protect their account against this kind of data collection. This means that the social network information is not accessible for these users and, therefore, they should also be removed from the dataset.

4.2. User Feature Extraction

Based on the properties of localness mentioned in section 3.3, three types of information can be helpful to identify the localness of social media users: temporal, spatial and social. These three types of information will be used to derive a set of features for the further user localness assessment. Besides information related these three properties, some other information from social media user are also useful in localness assessment, and the information is classified into other features in this approach.

4.2.1. Temporal feature

Temporal features are used to describe social media users from a temporal perspective and there are five temporal features: duration, maximum interval, average visit time, night posting proportion and weekend posting proportion. With these features, how long one user stayed in the city and when the users were in the city can be answered.

- Duration

Duration is the time difference between the first and last post of one user in the dataset. Longer durations indicate that users stay in the study area for longer times. The duration value does not always mean that

users stay in one city every day of this time period. Therefore, information about the visit interval and visit time is also needed.

- Maximum interval

Interval can be seen as the time difference between two adjacent tweets after ranking the tweets by their creation date and time. The maximum interval is the longest time period when one user did not have any posting behaviour. I assume that intervals longer than 60 days are called as "long interval" and long interval indicate that one user was not in the city during the period according to common sense. In ideal conditions, people with localness types like long-term resident, temporal or short-term resident and non-local commuter will have smaller maximum intervals, visitors who went to the city more than once should have larger one and the maximum interval of seasonal residents will be longer than half a year at least.

-Average visit time

Visit time means the time from arrival to departure. If the maximum interval of one user is shorter than "long interval", users never left the city during the period based on the long interval assumption, so the duration is their visit time. With a lack of further information, the first posting time is treated as the arrival time and the last posting time is treated as the departure time for these users. However, for more-than-once visitors, the time of each visit may be variable and the time difference between two visits can be big, so durations of more-than-once visitors are broken by long intervals. For these visitors, the fist posting time is the first arrival time, and the first departure time is the time of the last posting before the first long interval. By that analogy, the duration is partitioned into visits and long intervals by pairs of arrival time and departure time, and the average visit time is the mean value of the total visit time. The average visit time of visitors should be less than 30 days as mentioned in section 3.4.

- Night post proportion

This feature means the proportion of posts at night. It can be used to distinguish users who mainly visit a city by daylight, like non-local commuters.

- Weekend post proportion

This feature refers to the proportion of posts during the weekend. It can be used to mine weekly activity patterns of users and find out about users who mainly visit one city during the weekend or on weekdays. Both the weekend and night posting proportion are the main features to identify non-local commuters and are not very useful for other localness types because people with other localness types can post social media messages at any time on any weekday.

Figure 4.2 shows the summary of above temporal feature extraction process. In these processes, timestamp of posts is used which include both date and time when one post was created. To avoid problems caused by time zone in some city, there is an additional step to unify timestamp by the same time zone. Six and seven in day of week represent Saturday and Sunday respectively.

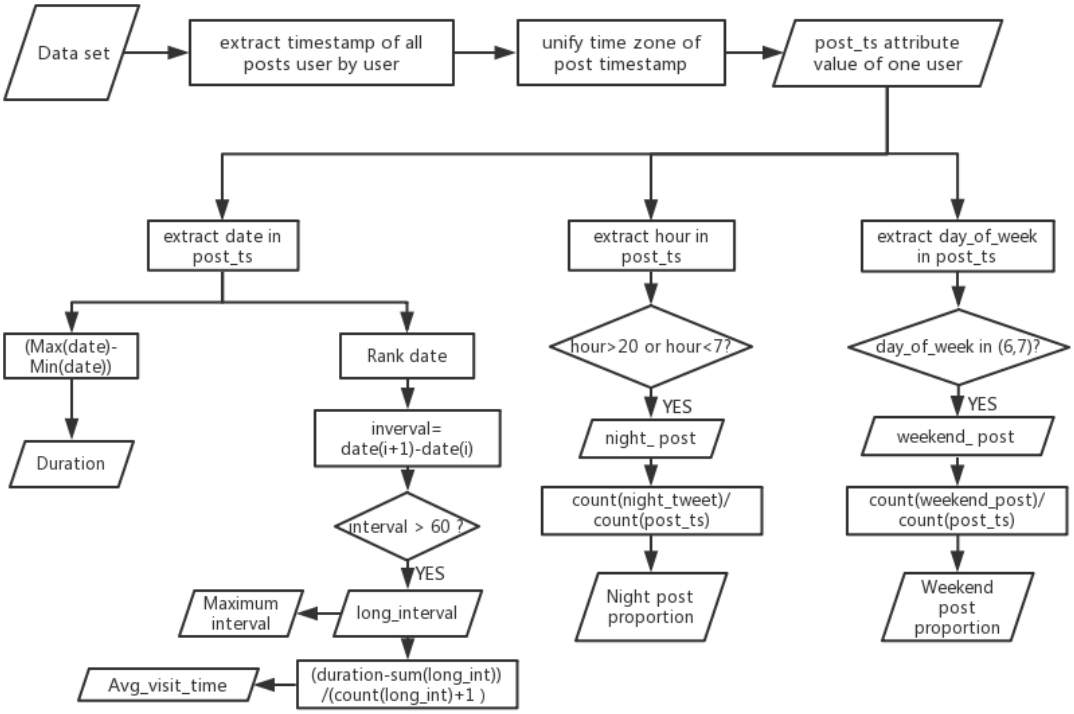


Figure 4.2: Flowchart of temporal feature extraction

4.2.2. Spatial feature

Three spatial features will be extracted to describe social media users from activity scope, activity concentration and tourism interest perspectives.

- Area of standard deviational ellipse

The standard deviational ellipse is a useful graphic representation to show the average location and distribution of a point set (Yuill, 1971). The ellipse is used to show the scope of user activity locations in this thesis. Main parameters of the ellipse are calculated using the following formulas:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}((X-\bar{X})\cos\theta-(Y-\bar{Y})\sin\theta)^2}{N}}$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}((X-\bar{X})\sin\theta-(Y-\bar{Y})\cos\theta)^2}{N}}$$

In the formulas, X and Y represents two values in point coordinates respectively, $\bar{X}$ and $\bar{Y}$ are the mean value of coordinates, θ is rotation angle of axes, and N is the total number of post points. σ_x and σ_y as results of the formulas are the length of two semi-axes. Area of ellipse is calculated using the following formulas:

$$Area = \pi\sigma_x\sigma_y$$

In Yuill work, he also pointed out that area of the ellipse can be used as a measurement to examine the concentration and dispersion of spatial data, but the interpretation should come from the comparison with

other data. Therefore, to use the ellipse area in localness assessment, percentiles of area values will be calculated first to be used as the reference values in the comparison.

- Concentrated-location proportion

People may have more activities near some typical locations such as the home location of residents and the work place of non-local commuters. More posts near the typical locations mean that the users' activities are more concentrated, and the post points may concentrate to any shape (not only a circle). The number of typical locations of users is variable and it depends on the users' daily route and tweeting habits. Therefore, the number of location clusters is also variable. To find the typical locations of user activities, density-based clustering can be useful. Ester, Kriegel, Sander, & Xu (1996) proposed an efficient algorithm called "density-based spatial clustering of applications with noise" (DBSCAN), which can find the clusters with an arbitrary shape and the cluster number is not required in the input parameters of this algorithm. In this method, point density is defined as the point number within a certain distance (eps), and if the density of one point is larger than a certain value (MinPts) the point is a core point. If the distance between two core points is smaller than eps, the two core points are in the same cluster. If the density of one point is smaller than MinPts and the distance to a core point is smaller than eps, the point is the border point of the cluster containing the core point. If one point does not belong to any cluster, the point is a noise point.

Using this density-based clustering method, user activity locations can be clustered based on proximity, and each cluster represents one area near one typical location of users. In ideal conditions, for resident localness types, there should be at least two clusters: near home location and near work place; for non-local commuters, the post points near work place form just one cluster. The proportion of core points among all points is the concentrated-location proportion in this study, and it shows the concentrative degree of one user's activity locations.

- Tourist attraction proportion

To identify tourists more accurately, the locations of tourist attractions in the city will be used as base locations, and the post locations near any tourist attraction will be labelled as attraction points.

The tourist attractions are shown as points in online maps in the collection of tourist attraction locations and the coordinates are also based on those points. However, most tourist attractions are venues and tourists may post at any location near to or inside the tourist attractions. To determine whether one post point is related to a tourist attraction, the distance from the post point to every attraction point is calculated and a minimum distance is selected to be a threshold. The threshold is devised to be 500 m to get more nearby locations and the point within this distance of any tourist attractions will be treated as an attraction point. The proportion of posts near to tourist attractions is the result of the total number of attraction points divided by the number of all post locations.

Figure 4.3 shows the summary of above spatial feature extraction. In the process, core point means concentrated point and tourist attraction coordinates will be gathered from online map such as google based on the names of attractions

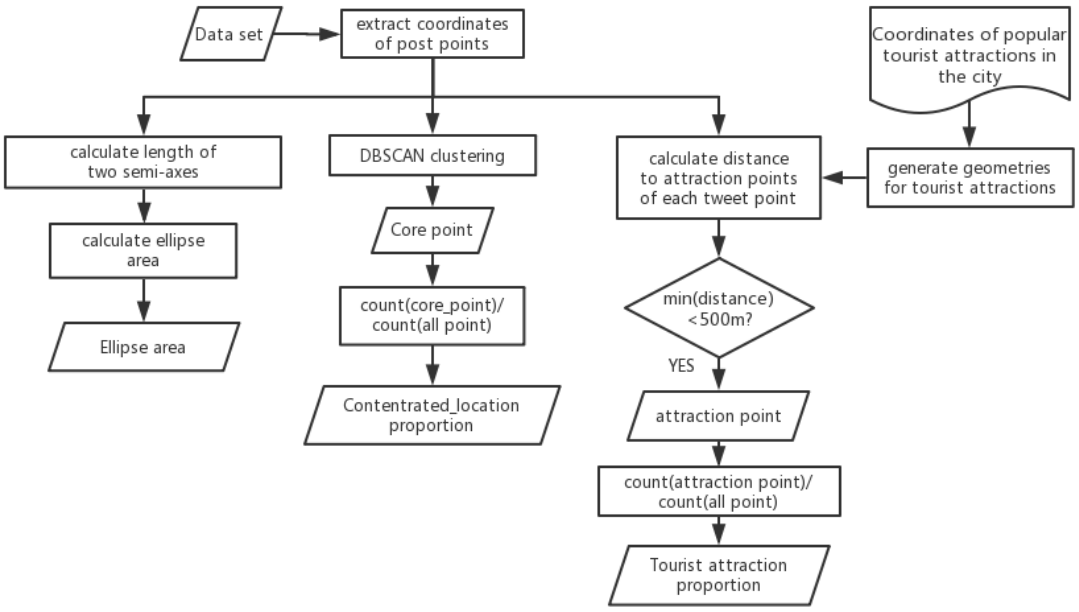


Figure 4.3: Flowchart of spatial feature extraction

4.2.3. Social feature

Social features represent the connection between social media users and local society. Local proportion in social network and followed local organization display the connection from local people and local organization perspectives respectively. User interests indicate the aspects of local society which users pay more attention on.

- Local proportion in social network

The connection between one user and local society can be represented as the proportion of local followers or followings in his/her social network on one social media platform. Users are very likely to follow each other if they know each other offline, so either followers or followings can be useful for the feature extraction in generic localness assessment approach. User follower will be used in the rest of the thesis.

There are two options to identify "local" followers: the first one is based on the localness of users and the second one is based on the location information provided by users. The localness types which can be treated as "local" depend on the study requirements. For example, local people only refer to the long-term residents in some studies while all the resident localness types can be included in the result of local people identification in other studies. Local followers as social media users also have a localness attribute relative to the city, and many of them are also waiting for a localness assessment in a particular study. In that case, this unknown information cannot be used as the data source of this feature extraction although the local followers or followings identification based on their localness is a relatively reliable way.

In the second option of local follower identification, location information in the user profile can be useful although it is not very reliable. As mentioned in section 1.1, not all the social media users enter their location information in their user profile, and some users may provide a fake location, irrelevant comments or multiple locations in the location field. But one advantage of this information is that among the users who

did input their location, they tend to disclose their location at the city scale (Hecht, Hong, Suh, & Chi, 2011). To avoid the bad influence of location information loss and other situations in which the users' real location is not available, one relative proportion will be used instead of using the location information of single users directly. This relative proportion will be the result of the number of followers who can be located in the city using the location information directly divided by the number of those who can be located outside the city using the location information in their location field.

- Followed local organization

The number of local organizations followed by social media users indicates the connection between one user and the organizations in the city and the more local organizations a user follows the stronger connection with a target area exists (Grace et al., 2017). The precondition of using this feature is to match the social media account with real local organizations which can be found from various data sources such as community directories, online searches and so on. According to Grace's work, the local organizations can be identified in eight categories: citizens' associations, civic services, emergency services, schools, bars, entertainment, restaurants and media. A search of local organizations is the crux in this feature extraction, and a useful local organization list should cover as much local organizations as possible. But this process is time consuming and the identification of those local organizations is beyond the scope of this thesis. So, the feature will not be considered in the rest of this thesis, although its effect has been proved and it's useful for generic localness assessment.

- User interest

The interests of users can be found out through the contents or hashtags of social media posts because interests are closely related to the topics which are mentioned in their posts. This feature can indicate the motivation of visit which is crucial to tell those localness types apart and it is the most important factor in the subdivision of localness types if the user localness assessment is based on the detailed and specific local knowledge and this feature. A reliable extraction of user interests is beyond the scope of this thesis, so this feature will not be used in the rest of the thesis but it's an important user feature which can be used in the generic localness assessment.

Figure 4.4 shows the summary of above social feature extraction. Only feature local proportion in social network are described in detailed because it will be used in the rest of this thesis as mentioned before. Local organizations will be found in users' followings and the extraction of user interests is beyond the scope of this thesis.

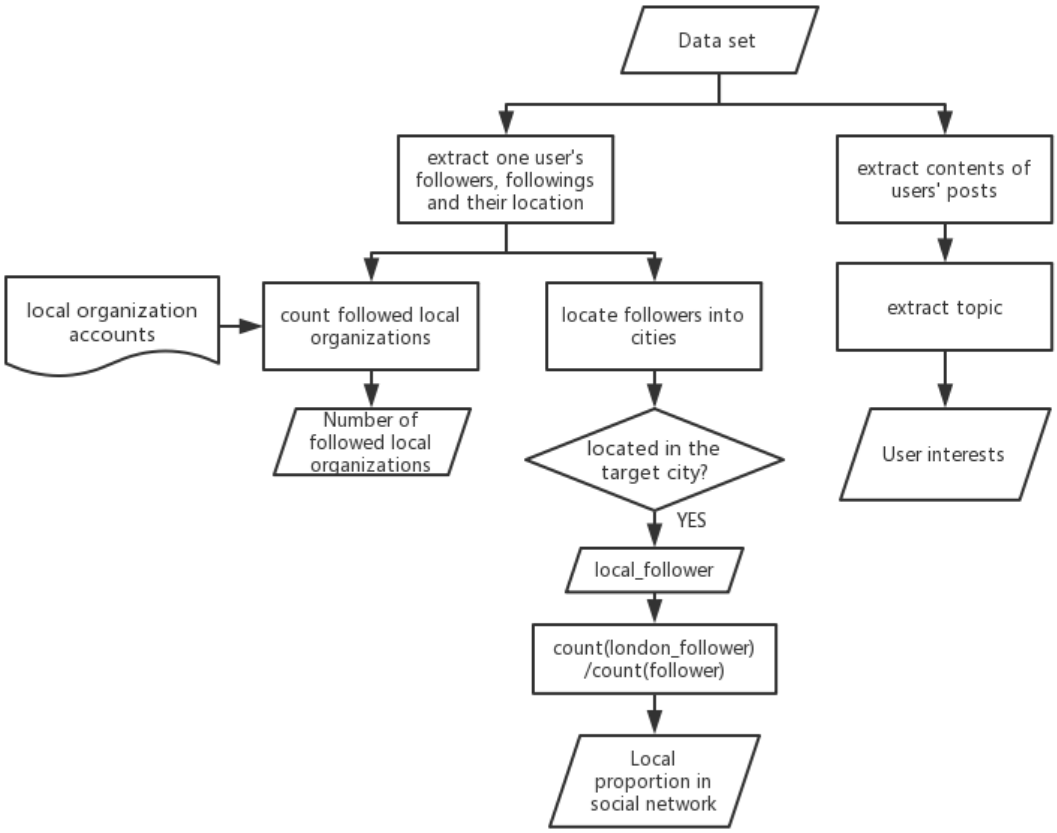


Figure 4.4: Flowchart of social feature extraction

4.2.4. Other feature

- User location in user profile

As the information provided by user on their own initiative, user location in their profile is an important indication and evidence to identify the relationship between one user and the city. Because the information in users' location field is not reliable, the relationship between users and city will not be concluded only based on this information, but this information is still a notable feature of users.

- Language

Some social media platforms are widely used in the world, so users will post content in different languages on the same platform. Some users may declare their interface language in their profile and their language preference can also be inferred easily based on the language of the users' postings. Countries have one or more official languages and there is usually one official language in one city. Therefore, language can be a feature in the localness assessment at city scale.

4.2.5. Summary of user features

Table 4.1 shows the summary of user feature and the brief description of each feature based on above sections.

Table 4.1: User features and feature description

Feature Category	User Feature	Feature Description
Temporal feature	Duration	Time difference between the first and last post of one user
	Maximum interval	Longest interval between two tweets
	Average visit time	Average time of all the visits which are defined by pairs of arrival time and departure time
	Night content proportion	Proportion of postings at night
	Weekend content proportion	Proportion of postings during the weekend
Spatial feature	Area of standard deviational ellipse	Scope of user activity locations
	Concentrated-location proportion	Proportion of core points based on DBSCAN
Social feature	Local proportion in social network	Proportion of followers/followings who declare that they live in the city
	User interest	Users' interests summarized from mentioned topics
	Follow local organization	Number of local organizations followed by social media users
Other feature	User location in profile	Location information in social media user profile
	Language	Language preference of users on social media platforms

4.3. Rule-based Localness Assessment

Localness properties are used to describe the localness types in Table 3.2 and user features are specified based on information available in social media along localness properties. To build a relationship between user features and localness types, some conditions will have to be devised. When assess localness of social media users, an assess sequence is designed to make full use of all user features and to make sure more users can be assessed.

4.3.1. Conditions

Table 4.2 shows the relationship between user features and localness types. There are three kinds of relationship: blank, strong, weak.

Table 4.2: Conditions of user features for localness types

Feature Category	User Feature	Long-term resident	Temporary or short-term resident	Seasonal resident	Non-local commuter	Visitor (once)	Visitor (more than once)	Tourist
Temporal feature	Duration	**≥12 months**	**1-12 months**	**≥ 2 years**	**≥ 30 days**	**<30 days**	**-**	**≤7 days**
	Maximum interval	**< 2 months**	**< 2 months**	**≥ 6 months**	**< 2 months**	**-**	**2-6 months**	-
	Average visit time	-	-	**1-3 months**	-	-	**<30 days**	-
	Night content proportion		-		**<30%**		-	
	Weekend content proportion		-		**<30%**		-	
Spatial feature	Ellipse area	≥70th	20th-70th	20th-70th	20th-70th	<20th	20th-70th	-
	Concentrated-location proportion	≥50%	20%-50%	20%-50%	20%-50%	-	20%-50%	-
	Tourist attraction proportion			-				**≥50%**
Social feature	Local proportion in social network	≥40%	10%-40%	10%-40%	10%-40%	<10%	10%-40%	<10%
Other feature	User location in profile	**Relate to the city**	-	-	-	-	-	-
	Language	-	-	-	-	-	-	-

Blank cells in the table mean that one feature is not useful for one specific localness type or variable. For some localness types (i.e. long-term residents, temporal or short-term residents, non-local commuters, visitors and tourists), the average visit time equals to duration. Therefore, there is no need to repeat this information. Visitors and tourists who only visit the city once, stay in the city for a short time, so the maximum interval of them is meaningless for identify them. For tourists, the tourist attraction proportion can provide much more reliable information than the other two spatial features. The other two features are closely related to the distribution of tourist attraction locations in the city for tourists which means that only the tourist attraction proportion will be enough to represent the spatial feature of these users.

The second kind of relationship between features and localness types is a strong relationship and these relationships are shown in **bold**. This relationship means that these features are more reliable for identifying the corresponding localness types. All the relationships between temporal features and localness types are strong because more posting behaviours are considered during temporal feature extraction compared to other features, and a timestamp of postings is generated automatically by social media platforms which means they provide real information. The tourist attraction proportion is also a strong feature for tourists because it is calculated based on the precise location of posts and tourist attractions and takes all posting locations into account at the same time.

The remaining relationships are weak relationships, as shown in non-bold text, containing social, other and spatial features except for the tourist attraction proportion. Users may disclose their locations in social media by place names or coordinates, only coordinates can provide precise locations and generate posting points for spatial feature analysis. However, only a small proportion of postings have coordinates, so a lot of spatial information of users' postings is missing. In addition, the amount of available spatial information highly

depends on the social media usage habits and frequency of users. So, if users do not like to reveal their locations or only use social media occasionally, temporal and spatial information of most of their activities will not be appeared in social media data. Therefore, using the limited spatial information of users' activities to estimate the activity scope and the degree of activity concentration is not very reliable. As for the social feature, the identification of local followers is based on the content in the location field of the user profile which is not reliable. The language feature can be used to identify outliers if they have ever posted something in a not official language in the city. However, since users may use local language to post contents and English is the dominant language on social media platforms, the language feature is not supportive to identify most of the users, especially when the target city is in an English-speaking country.

All the thresholds mentioned in Table 4.2 are designed based on ideal conditions and these imply that social media users record their activities by social media as much and detailed as possible and the dataset used in the localness assessment covers all the social media data of users for long enough time. Features and corresponding thresholds compose conditions for localness assessment, and strong relationships mean strong conditions and weak relationships will be weak conditions. Based on these conditions, typical situation of each localness type is described in the following paragraphs.

Long-term residents will stay in the city for more than one year with smaller intervals (30 days) of being away. Because they will have a larger activity scope than users with other localness types, the area of the standard deviational ellipse should be larger than the 70^{th} percentile of all the area values. More than half of the long-term residents' activities will happen near their typical locations and more than 40% of their followers are local people.

Temporal or short-term residents will stay in the city for at least one month with small intervals (30 days) and the corresponding ellipse area will be larger than the 20th percentile and smaller than the 70th percentile of the area. The proportion of concentrated activities will be more than 20% and less than 50%, because they may spend more time in other places rather than their home or work place to explore and get familiar with the city. The local proportion in their network will be larger than 10% and less than 40%.

For seasonal residents, the overall duration of their stays should be longer than 2 years to clarify their yearly movement patterns. Their maximum interval will be longer than half a year and their visit time will be shorter than 3 months based on the description of this localness type. The value range of their spatial and social features are as the same as for the temporal residents.

Non-local commuters will stay in the city for more than one month with smaller intervals (30 days) and duration condition "one month" is devised to make sure that enough information can be collected. Because these users only visit the city by daylight on weekdays, the night content proportion and weekend content proportion are lower. I assume that both night and weekend proportion values are less than 30%. The value range of their spatial and social features is the same as for the temporal residents.

Visitors who only go to the city for one time, will stay in the city for less than 30 days and the ellipse area should be smaller than the 20th percentile of area. The proportion of their local followers will be less than 10%. For the visitors who visit the city several times, the average visit time should less than 30 days and the maximum interval between two visits should be larger than 60 days. The value range of their spatial and social features is the same as for the temporal residents.

Tourists will stay in the city for less than or equal to 7 days and most of their activities will happen near tourist attractions and they will have less than 10% of local followers.

The no-experience users cannot be identified based on the data collection in section 4.1, because these users never visit the city and there will be no data about them in the dataset.

4.3.2. Sequence

When the features are used to assess localness of social media users, it's reasonless to expect that all the users meet all the conditions of one localness type at the same time, especially for the weak conditions. It' s possible that users meet some conditions of one type but meet other conditions of another type. Obviously, the reliability of localness type of users who meet all the conditions of one localness type is higher than those who only meet one or two conditions, and the users who meet a common condition used in multiple localness types cannot be assigned as any localness type. Thus, there should be a sequence to assess localness of social media users to make sure that users who meet more and more reliable conditions can be assessed first and all users are assessed base on distinguishable condition combinations from strong to weak. The distinguishable condition combinations mean that the combination of these conditions should have the ability to distinguish this localness type and any other types when only part of conditions of one localness type are used to assess. All localness of users should be assessed based on condition combinations instead of single conditions to make full use of user features. The combinations of weak conditions are not allowed because they are inherently unreliable. Figure 4.5 shows the sequence used in localness assessment and detailed explanation is in the following paragraphs.

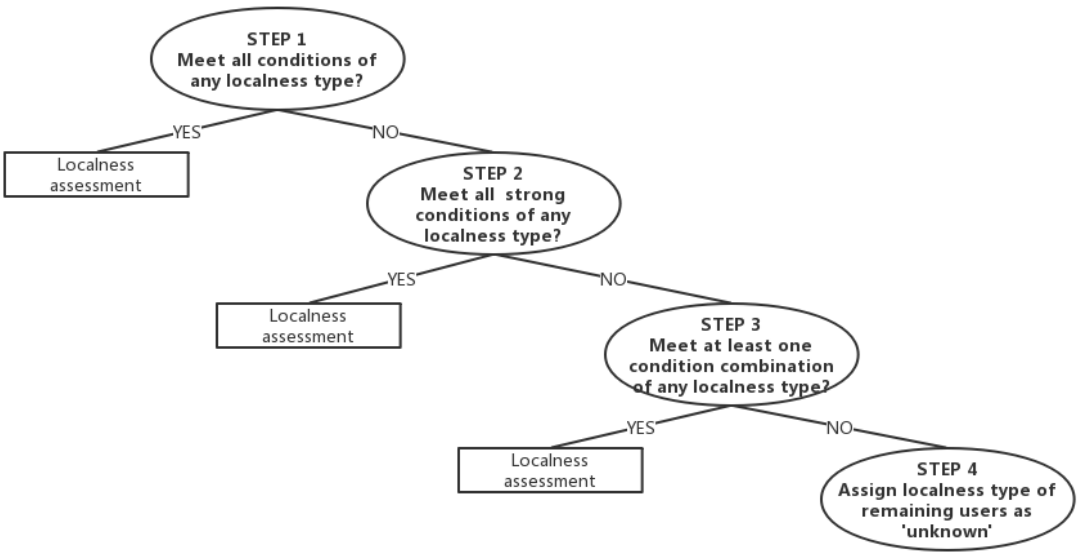


Figure 4.5: Localness assessment sequence

STEP 1:

In the first step, users who meet all the conditions of one localness type (except for once-visitor) will be selected and the localness type will be assigned to the users. One-time visitor means they only went to the city once and not went with tourism purpose. This localness type only has one strong condition (duration < 30 days) and the range of this condition overlaps with tourist's condition (duration ≤ 7 days). So, users who meet the conditions of this type will be assessed in step 3 after the selection of tourists to distinguish one-time visitors and tourists. All available features of these users are utilized in localness assessment and the result from this step is the most reliable result based on this approach.

STEP 2:

As a compromise of the last step, the users who meet all strong conditions of localness types will be selected and assessed localness. Using strong conditions can distinguish all the localness types but the information about weak conditions will not be considered, such as the ellipse area and concentrated-location proportion of users.

STEP 3:

As a further compromise of step 2, users who meet at least one condition combination of one localness type will be selected and assigned as the type. There is at least one strong condition in each combination and each condition combination can distinguish one localness type from others. The condition combinations used in this step are shown in Table 4.3. For non-local commuters, more-than-once visitors and tourists, they will not be assessed in this step because all distinguishable condition combinations include all strong conditions which are used in step 2.

Table 4.3: Condition combinations used in step 3 of localness assessment

Localness Type	Condition Combination
Long-term resident	1. Duration ≥12 months & max_interval<2 months & ellipse area ≥70th 2. Duration ≥12 months & max_interval <2 months & concentrated proportion ≥50% 3. Duration ≥12 months & max_interval <2 months & local follower proportion≥40% 4. User location related to the city & ellipse area ≥70th 5. User location related to the city & concentrated-location proportion ≥50% 6. User location related to the city & area ≥ local follower proportion ≥40%
Temporary or short-term resident	1. one month ≤ duration <12 months & 20th ≤ ellipse area <70th 2. one month ≤ duration <12 months & 20% ≤ concentrated proportion <50% 3. one month ≤ duration <12 months & 10% ≤ local follower proportion <40%
Seasonal resident	1. Duration≥2 years & max_interval ≥ 6 months & 20th ≤ ellipse area <70th 2. Duration≥2 years & max_interval≥6 months & 20%≤concentrated proportion<50% 3. Duration≥2 years & max_interval≥6 months & 10%≤local follower proportion<40%
Non-local commuter	Should meet all strong conditions
One-time visitor	1. Not tourist & duration <30 days & ellipse area <20th 2. Not tourist & duration <30 days & local follower proportion<10%
More-than-once visitor	Should meet all strong conditions
Tourist	Should meet all strong conditions

STEP 4:

All remaining users will be assigned as unknown for their localness type because none of the condition combinations is met based on their features and they cannot provide enough information to prove that they should be classified as any localness type based on available data.

In Chapter 4, I designed an approach to assess the localness of social media users. The input of the approach are social media posts, user followers/followings information and some additional information (tourist attraction location and local organization accounts). The localness type of each user, which is the output of this approach, is determined by the comparison of conditions and user features step by step. To test the approach, I will use Twitter data in Greater London to implement the approach and to analyse the implementation process and result as case study in Chapter 5.

5. CASE STUDY-TWITTER DATA IN LONDON

In this chapter, the fourth research objective will be addressed: implementing and evaluating the approach that has been proposed in Chapter 4. Twitter data in the London region are used as a case study to address this objective. First, I will describe the study area and the original dataset. Thereafter, the sampling strategy and data filters are explained in detail. Second, the user features, as designed in the localness assessment approach will be calculated, and then the localness of users will be assessed step by step, based on user features and conditions in the approach. Finally, in order to evaluate the results of the approach, the ground truths are labelled and compared with the results. Based on the comparison, there is some discussion about the approach implementation.

5.1. Study Area and Data

5.1.1. Study Area

Greater London is located within the London region of England and is sub-divided into 33 government districts: The City of London and 32 London boroughs, as shown in Figure 5.1. London is a major international financial and business centre in the world, and it is a typical global city (Sassen, 2001). The population of London in 2017 is 8.825 million (GLA, 2017) and the diversity of population in London is significant. The nationality of 24% of London's residents is not British (ONS, 2017b) and 40% of the residents is not born in the UK(ONS, 2017a). Migration indicators of London show that there were 168 thousand long-term international immigrants and 98 thousand emigrants in 2017 and 59 thousand short-term international migrations in 2016 based on the International Passenger Survey(GLA, 2018).

Figure 5.1: Map of the study area

Besides the districts within the Greater London region, there are 17 districts sharing boundaries with the region. These districts are also served by the London underground and telephone service and a relatively high percentage of the employed population in these districts works in London. According to 2011 Census data, 18% of the people who worked in London commuted from areas outside London, like these adjacent

districts. These people can be considered as non-local commuters according to the localness type classification that was presented in section 3.4.

London as a global and tourism city attracts visitors and tourists from home and abroad. There were about 296 thousand overseas visitors and more than 66 thousand domestic visitors who visited London per day in 2015 (GLA, 2015). According to global city indicators in the New York City Global City database, 15.2 million foreign tourists and 11.1 million domestic tourists visited London in 2012 (GLA, 2012).

The UK is a country with a high social media penetration. About 66% of the population in the UK were active social media users in 2018 (We Are Social, 2018) and 81% of 1091 respondents in one survey about adults' media use in 2015 reported that they use social media at least once a day (Ofcom, 2018).

Based on this population diversity, the attraction to non-local commuters, visitors and tourists, and the high penetration of social media, London was considered to be a suitable city for this localness assessment case study.

5.1.2. Dataset Description

Twitter is a worldwide online social networking service which enables registered users to read, post and interact with short messages known as "tweets". Using this social medium, users can express their opinions, record life experiences and interact with their friends by text, image, video and hyperlinks. Twitter is one of the most popular social networks in the world and it is well-known in the UK. It has 321 million active users worldwide as of the fourth quarter of 2018 (Twitter, 2018) and the number of Twitter users in the UK is about 17.1 million in 2018 (eMarketer., 2018). Among all the social media users in the UK, 89% of them reported prompted awareness of Twitter in one survey in 2015 (Harris Interactive., 2015). But not all the social media users in the UK use Twitter frequently: among all the respondents in one survey in January 2018, only 17% of them use Twitter every day, 31% of them use Twitter every week and 55% of them reported that they never use Twitter (We are Flint, 2018).

Twitter data in Greater London will be used in this case study. The original data have been collected by the GIP department of the Faculty ITC of the University of Twente since 2017 by means of Twitter's streaming API and they were stored in a PostgreSQL database. The only parameter that was used in the data collection is a bounding box defined by the longitude and latitude in WGS84, which is consistent with the coordinate reference system that is used by the Twitter API. The bounding box was defined as '-0.489,51.28,0.236,51.686', which comprises a larger than the real extent of London. The dataset contains attributes of both tweets and users as shown in Table 5.1. As for tweets, the dataset has the following attributes: tweet ID, tweet text, created timestamp, name and ID of located place, longitude and latitude of tweets. As for the users, the dataset has the following attributes: user ID, user name, user description, user location, number of followers, number of tweets.

Table 5.1: Attributes in the data set

Twitter User			Tweet		
Attribute Name	**Data Type**	**Meaning**	**Attribute Name**	**Data Type**	**Meaning**
user ID	text	string representation of the unique identifier of users	**tweet ID**	text	string representation of the unique identifier of tweets
user name	text	name of users	**tweet text**	text	content of status update
user description	text	user-defined string to describe their account	**created timestamp**	timestamp	time and date when tweets were created
user location	text	user-defined location for this account's profile	**place name /ID**	text	the place where tweets are associated
number of followers	integer	total number of followers in this account at the query moment	**longitude of tweets**	numeric	geographic location of tweets as reported by the users or client application
number of tweets	integer	total number of tweets in this account at the query moment	**latitude of tweets**	numeric	geographic location of tweets as reported by the users or client application

The dataset contains 23.45 million tweets of 819 thousand users in total. At the moment of running the first query in this research, the earliest tweet in the dataset was posted on July 11, 2017 and the last tweet on December 1, 2018. Therefore, the overall time interval of the used case study dataset is 505 days.

Users disclosed their location by place names and place IDs when they posted tweets and all tweets with place name/ID are geo-tagged tweets. Users may use a detailed place name like "The O2" which is a music venue in London, or a more overall place name like "London, England". Some users may disclose their location by coordinates except for place names, and 12.56% of tweets have coordinates among all geo-tagged tweets in the data set. The coordinates are represented as longitude and latitude attributes of the tweets. Using the coordinates, the precise locations of users when they posted the tweets are available. However, if users use place names to reveal their locations, geographic information tools can convert place names to geometries. Most of the geometries will be polygons, because many place names do not refer to specific locations but to areas with larger geographic coverages, such as district, city or region. For these tweets, the exact location points of users are not available. Therefore, only tweets with coordinates will be used in the spatial feature calculations, in order to make sure that the tweeting locations are precise enough and all the tweets can be used in the temporal feature calculations, so as to be able to make full use of the information of tweeting behaviours.

5.1.3. Sampling and Data Filtering

A sample of the dataset is used in the case study, and the reasons and processes of getting this sample are described in this section.

Undoubtedly, the more users there are in the dataset, the more potential situations of users can be covered. However, if the entire dataset is used, even a selection query with only simple conditions face time out problems and so many similar or more complex queries will be used in the case study. In addition, considering the rate limiting of Twitter APIs which only allows limited number of requests per window, if I collect followers' information of all users in the original dataset, this additional data collection may take several months. Due to the lack of processing time, I created a smaller dataset using random sampling and the process of sampling is shown in the following code:

```
create table user_sample1 as (
with tw_user as
(select user_id from fo.london_twitter_v2 group by user_id )
select * from tw_user order by random() limit 10000);
```

In the code, "fo.london_twitter_v2" is the name of the table which contains the entire original dataset and "tw_user" contains all the users in the original dataset. The random() function will return a random value between 0 and 1 for each record. Users are ranked according to the value returned by the random function, and the ID's of the top 10,000 users are stored in the table "user_sample1" which is the result of user sampling that will be used in the rest of this case study. After this random selection of users, the tweets of all the selected users are stored into table "tweet_sample1" which is the data source for the temporal and spatial feature extraction.

Table 5.2 shows the characteristics of the original dataset and the sample dataset. The sample dataset contains 304 thousand tweets of 10 thousand users in total, and 12% of the tweets has coordinates. The sample dataset is consistent with the original dataset as far as the average number of tweets per user, the proportion of tweets with coordinates and the time interval are concerned. This sample dataset will be used in the rest of the case study instead of the large entire dataset.

Table 5.2: Description of Dataset

	Original dataset	**Sample dataset**
User number	819 thousand	10 thousand
Tweet number	23.45 million	0.304 million
Tweets per user	28	30
Proportion of tweets with coordinates	12.56%	12.04%
Date of the earliest tweet	2017-07-11	2017-07-11
Date of the latest tweet	2018-12-01	2018-12-01
Time interval	507 days	507 days

After the sampling, data were cleaned to make sure that all data is useful, and all features of each user can be extracted. Figure 5.2 demonstrates the filters used in data cleaning and the detailed procedures will be explained in the following paragraphs.

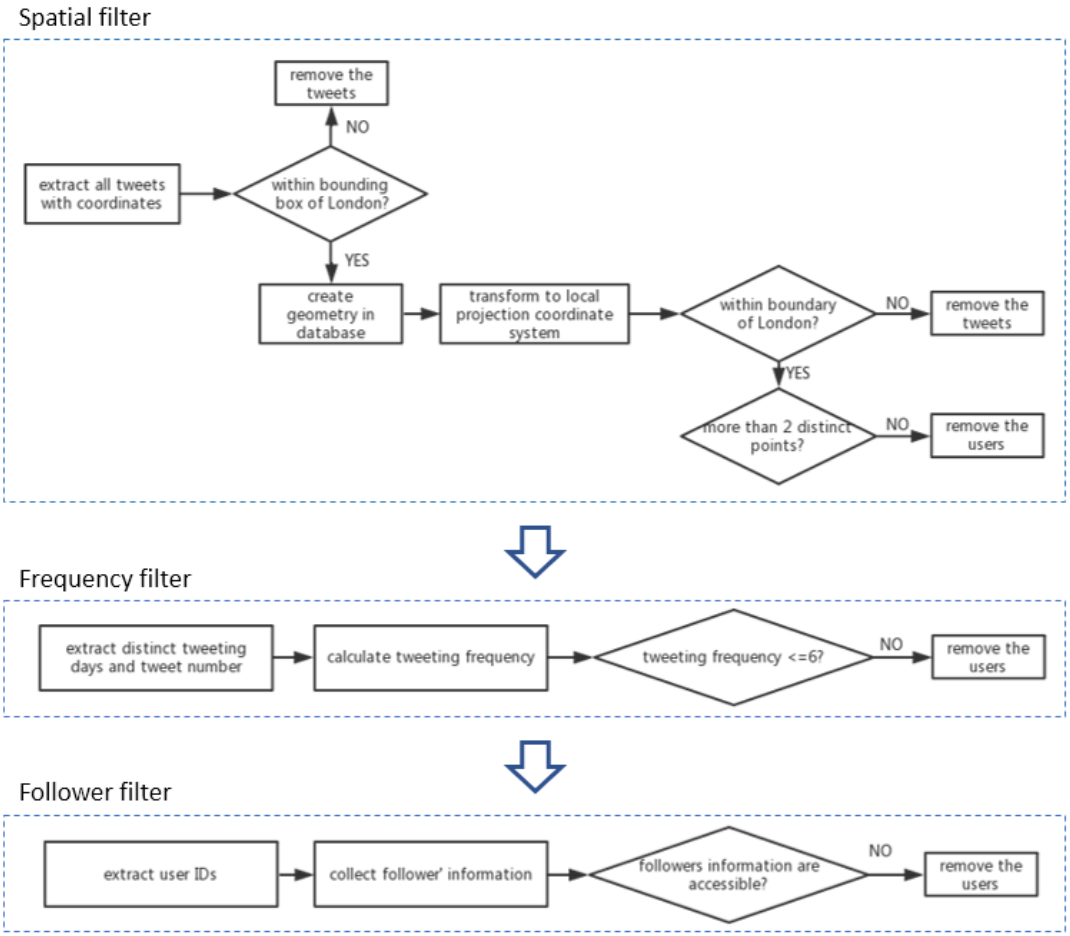


Figure 5.2: Filters in data pre-processing

Pre-processing of spatial data and spatial filtering are carried out first because the users who did not provide more than two precise locations in the dataset will be discarded, which will significantly decrease the time cost of further processing. Using PostGIS in PostgreSQL, one tweet point is generated for each tweet based on the value of the latitude and longitude attributes. Considering the distance calculation in the spatial features, the coordinate reference system of tweet points will be transformed to the British National Grid which is a local projection system in the UK and uses meter as unit, rather than using the original geographic coordinate system with degree as unit.

The first part of the spatial filter is to identify the tweets within the London case study area. Due to mistakes in the Twitter API, there are some tweets which are located outside the bounding box and some of them cannot be transformed to a local projection system because they are outside the scope of the local projection system used in this case study. So, a bounding box which covers all the Greater London area is used to filter out these tweets. Then all the tweets which are left are converted to tweet points and the coordinate reference system of the points is transformed to the British National Grid. The bounding box does not match with the boundary of London, so a shapefile of that boundary is imported and stored in the database as a table "london_boundary". The boundary table uses the same coordinate system as the tweet points, and the filter uses the st_within() function in PostGIS. One attribute called "within_london" is added to represent whether one tweet is within the study area and the tweet points outside the London boundary will be assigned as FALSE. The following SQL code is used in the filter:

```
alter table sample10t_tweet_latlon add column within_london boolean;

update sample10t_tweet_latlon set within_london=FALSE
where tweet_lat > 51.85 or tweet_lat <51.13 or tweet_lon <-1.17 or
tweet_lon >0.99;

select
AddGeometryColumn('s6037348','sample10t_tweet_latlon','the_geom',4326,'PO
INT',2,false);
update s6037348.sample10t_tweet_latlon as tw
set the_geom = ST_SetSRID(ST_MakePoint(tw.tweet_lon, tw.tweet_lat),4326)
where within_london is not FALSE;

select
AddGeometryColumn('s6037348','sample10t_tweet_latlon','the_geom_27700',27
700,'POINT',2,false);
update s6037348.sample10t_tweet_latlon
set the_geom_27700 = st_transform(the_geom,27700)
where within_london is not FALSE;

update sample10t_tweet_latlon
set within_london='y'
from london_boundary as b
where st_within(sample10t_tweet_latlon.the_geom_27700, b.geom);

update sample10t_tweet_latlon
set within_london='n'
where within_london is null;
```

The second part of the spatial filter is to filter out users based on the number of tweet points. Given the requirement of parameter calculation for a standard deviational ellipse, users should have at least three distinct tweet points and the filter is also defined to find all the users who meet this requirement. In the dataset, 3085 users have ever used coordinates to locate themselves among the sample of 10,000 users. A new table "sample10t_user_spatial" is created to calculate and store spatial user features, and only the information of the users who have more than two distinct tweet points is stored. Only 946 users remain after applying the spatial filter. The following code is used in this part of the spatial filter:

```
create table sample10t_user_spatial as(
select user_id, count(distinct the_geom_27700)
from sample10t_tweet_latlon
where within_london=TRUE
group by user_id having count(distinct the_geom_27700)>2);
```

A frequency filter is implemented to remove the users with a high posting frequency because these users provide more useless contents which are not related to their life experiences or may not be controlled by real persons. The tweeting frequency of one user is the result of dividing the number of his/her tweets by the number of distinct tweeting days. Based on the box plot of tweeting frequency as shown in Figure 5.3, frequency values larger than 3 are outliers. According to the histogram of tweeting frequency shown in Figure 5.4, there is a clear decrease in the number of users who post more than six tweets. When Twitter plays a role in recording activities in the user's life, six tweets per day are enough to achieve that even for a tourist who visit several tourist attractions within one day. So, in this case study, the threshold used in this

frequency filter is six, and users whose tweeting frequency is larger than this threshold are discarded. After applying this filter, 929 users were left in the sample dataset.

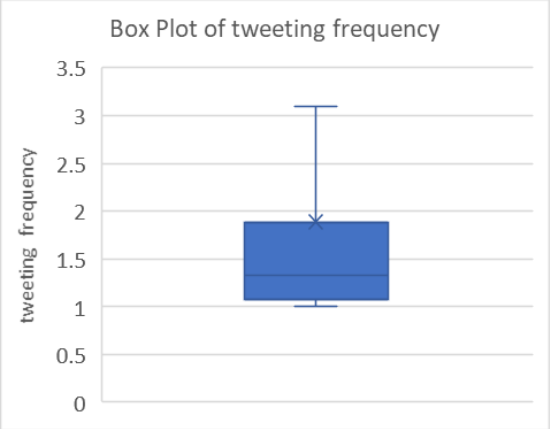


Figure 5.3: Box plot of tweeting frequency

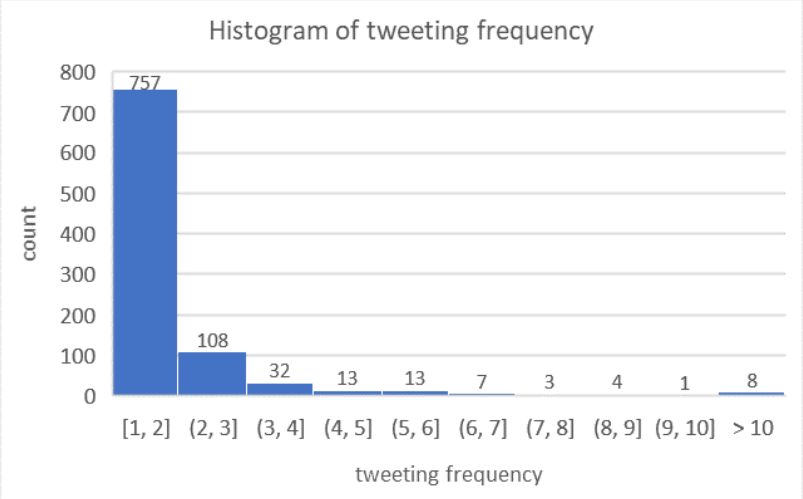


Figure 5.4: Histogram of tweeting frequency

Due to the rate limiting of Twitter as mentioned before, the collection of user followers is a time-consuming process. So, this step is carried out after implementing the spatial and frequency filters. Tweepy is a python library to access the Twitter API, and the Twitter API provides the functions which can return a collection of followers' IDs for one specific user and get the information of followers based on their user IDs. Using Tweepy and the functions in the Twitter API, a list of follower IDs is generated for each user, and location information of each follower, as included in the user profile, is also collected. Some accounts are protected from data collection by the Twitter API, so the followers' information is not accessible, and, therefore, the social feature cannot be extracted for these users. These users will be identified and removed by the last filter. After applying this filter, there were 861 users left and all the information of these users and their tweets compose the final data set used in further analysis.

5.1.4. Target Localness Types

Due to the limitations of this dataset, not all localness types mentioned in Table 3.2 can be identified reliably, and, therefore, the localness assessment in this case study cannot cover all localness types. The home locations of the non-local commuters are not in the target city, and it is very likely that the tweets they posted outside the city comprise the majority. Strong conditions of non-local commuters are the proportions of

night and weekend tweets. However, some users just prefer to post a tweet by daylight and on working days and this tweeting habit information should be gained based on a comparison between user tweeting behaviours in the target city and outside the target city. Without this tweeting habit information, it is difficult to make sure whether a low proportion of night and weekend tweeting is the result of tweeting habits or an indication of the non-local commuter localness type. Therefore, non-local commuters will not be considered in the rest of the localness assessment in this case study. These users may be identified as long-term residents or temporary/short-term residents. In addition, since the night and weekend tweet proportions are only strong conditions for non-local commuters, these two user features will not be considered in the feature extraction part.

Therefore, the target localness types in the case study are long-term resident, temporary or short-term resident, seasonal resident, visitor and tourist.

5.2. Feature extraction

In this section, user features are extracted based on the flowcharts in section 4.2 using the data set after data cleaning.

5.2.1. Feature extraction implementation

- Temporal feature

Because the night and weekend tweet proportions are not used in the case study, only the dates of the tweets are extracted, and detailed times are ignored. The feature "duration" is extracted as the time difference between the first tweet and the last tweet. The "interval" is the time difference between two adjacent tweets after ranking the tweets by their timestamps. The threshold for a long interval is 60 days as mentioned in section 4.2.1. The interval with the largest number of days is extracted as the feature "maximum interval". These long intervals split duration into visit slices, and the average of the time interval of visit slices is extracted as the feature "avg_visit_time". The following python code is used in interval extraction:

```
for i in range(len(tw_ts_list)-1):
  interval_list.append(tw_ts_list[i+1]-tw_ts_list[i])
  interval_day_list.append(interval_list[i].days)
  if interval_day_list[i]>60:
    two_month_list.append(interval_day_list[i])
sorted_list=sorted(interval_day_list)
max_interval=sorted_list[n-2]
```

In this code, tw_ts_list stores the timestamps of all tweets after ranking them from the earliest time to the latest time. Interval_list is created to store the time difference between two adjacent tweets. Two_month_list stores the intervals which are longer than 60 days, and, after sorting, maximum intervals are extracted and stored into the list max_interval.

- Spatial feature

As for the feature ellipse area, the lengths of the two semi-axes of the standard deviational ellipses and the area of the ellipse are calculated using the formulas mentioned in section 4.2.2.

In the calculation of the concentrated-location proportion, there are two important parameters: eps and MinPts. In DBSCAN, point density means the number of points within a certain distance (eps), and if the density of one point is larger than a certain value (MinPts) the point is a core point. Users have more activities

near to typical locations by walking and about 80% of the walking trips are less than one mile (Yang & Diez-Roux, 2012). In the case study, to cover most locations which are accessible from typical locations by walking, eps is assigned as one mile (about 1600 meters).

MinPts is closely related to the definition of typical locations because typical locations mean that there is at least a specific number of points near these locations and MinPts points out this specific number. The lower MinPts, the more clusters and the more typical locations. To find more typical locations of users, the MinPts should be as small as possible. However, in daily life, if one location is one typical location of one user, the location should be visited by this user for at least three times to make sure the user didn't go there accidentally. Therefore, MinPts in this case study is two, which means one specific tweet point can be considered as one core point and represent a typical location if there are two or more than two tweet points near to this tweet point.

After the determination of parameters, all points are assigned as one cluster index or noise (-1), and the concentrated-location proportion is the result of the number of non-noise points divided by the total number of all points. The following python code is used in this feature calculation:

```
clustering = DBSCAN(eps=1600, min_samples=2).fit(tweet_point)
point_in_cluster=0
for i in clustering.labels_:
  if i!=-1:
    point_in_cluster+=1
core_point_pc=round(100*point_in_cluster/len(clustering.labels_),2)
```

To calculate the tourist attraction proportion, additional information on local tourist attractions is needed. 20 Popular tourist attractions in London were selected based on a list of top attractions in the official visitor guide website of London[2] and coordinates of these tourist attractions were collected in Google Maps by searching for the tourist attraction names. The tourist attractions and their coordinates were stored in the database and point geometries were generated based on the coordinates. The complete tourist attraction names and their coordinates are listed in the Appendix. The following SQL code is used to select tweet points near to tourist attractions:

```
with distance as (
select
t.user_id,t.tweet_id,min(st_distance(t.the_geom_27700,a.the_geom_27700))
as dist
from sample10t_tweet_latlon as t, london_attraction as a
where t.user_id in (select user_id from final_features)
group by t.user_id,t.tweet_id)
select user_id, count(dist) as attraction_points
from distance
where dist<500
group by user_id
order by count(dist)
```

[2] https://www.visitlondon.com/things-to-do/sightseeing/london-attraction/top-ten-attractions

- Social feature

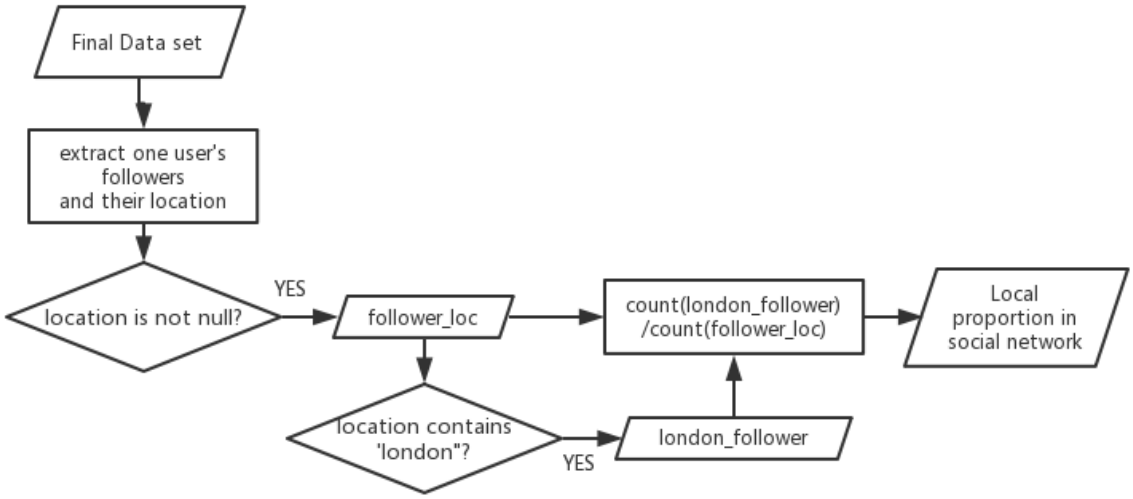


Figure 5.5: Flowchart for spatial feature extraction in the case study

Figure 5.5 shows the procedures in social feature extraction. The only social feature used in the case study is the local proportion in the social network. As mentioned in section 4.2.3, followers of one user should be located in one city using the location information in their profile. As I did not have sufficient time available to do more, the feature extraction was simplified to use the city name to filter out local followers instead of using any geoparsing tool to convert text description of locations to geographic identifiers. If one follower mentioned the term "London" in his/her location field, the follower will be considered as a local follower. Since it is impossible to check whether a follower is local or not for the followers who did not provide any location information in their profile, only the followers whose location field is not null were considered in the case study. The result of this feature is the proportion of the number of London local followers divided by the number of followers whose location field is not null.

-Other feature

Language feature is not used in the case study and for user location in profile, location information is only extracted and stored into the final feature table. The locations will be checked by term 'london' simply in SQL queries and no result will be displayed in next section.

5.2.2. Results

-Temporal feature

Duration and its corresponding threshold compose strong conditions for all localness types except for more-than-once visitors and Figure 5.6 shows how the durations of users distribute. Among all 861 users in the case study dataset, 39% of them have a duration (i.e. time difference between first and last tweet) of more than one year. The duration of 17% of the users is less than one month and 91 users show a duration of less than 7 days. For the durations of less than one-month, a zero value means that users stayed in the city for less than 24 hours and there is an obvious decrease after 7 days as shown in Figure 5.7. After 30 days, the duration has a relatively even distribution (see Figure 5.6), and there is no clear change near the threshold one year (365 days). So, the duration threshold between long-term residents and temporary or short-term residents may lead to some mistakes in localness assessment results. Because the dataset only covers about

16 months, the duration condition of seasonal residents cannot be met. To try to identify some seasonal residents, I assume that if durations of users are large than 15 months (one year and one quarter) some similar seasonal movement can be found in some users' tweeting behaviours in the quarter.

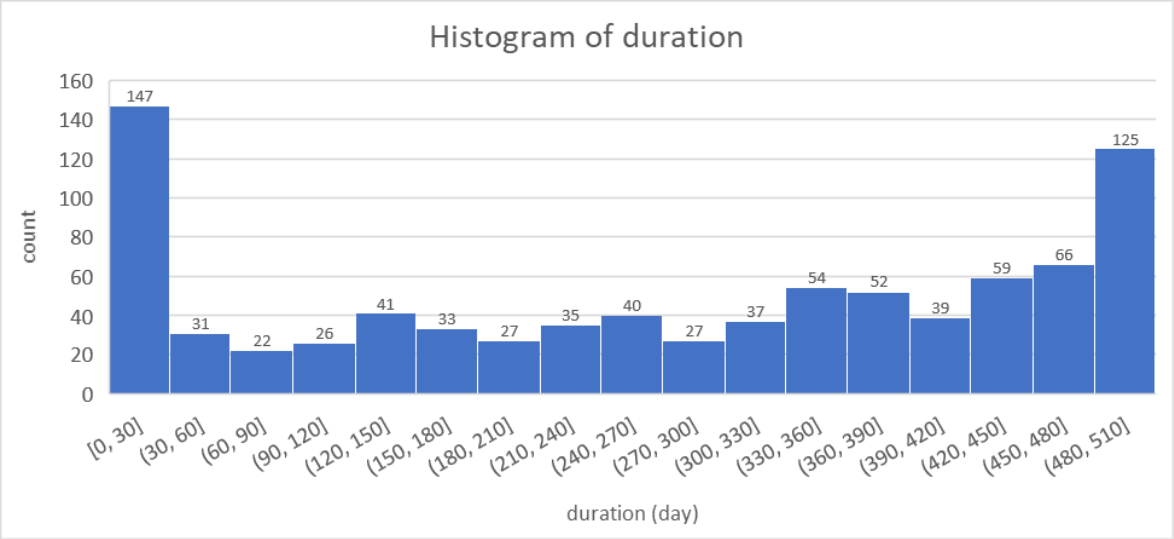


Figure 5.6: Histogram of overall duration of Twitter users

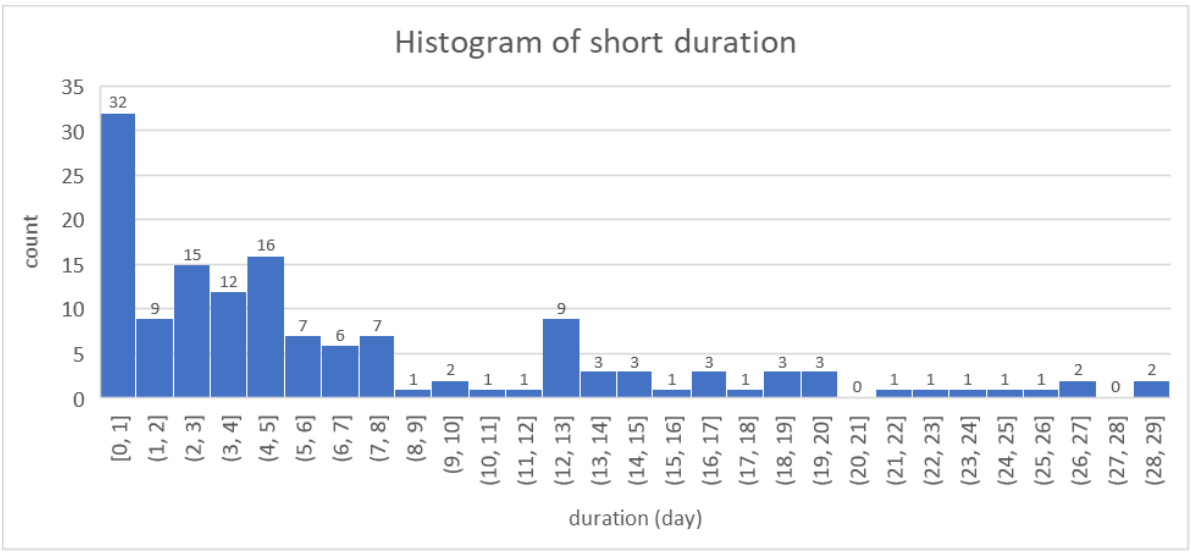


Figure 5.7: Histogram of short duration of Twitter users

Maximum interval (maximum time interval between two adjacent tweets) and corresponding threshold compose strong conditions for all localness types except for one-time visitors and tourists. As shown in Figure 5.8, there is no clear break that can be used to distinguish more-than-once visitors, whereas the threshold was set at 60 days in section 4.3.1. 58% Of the users' maximum interval is more than 60 days and all these users will be excluded from long-term residents and temporary or short-term residents.

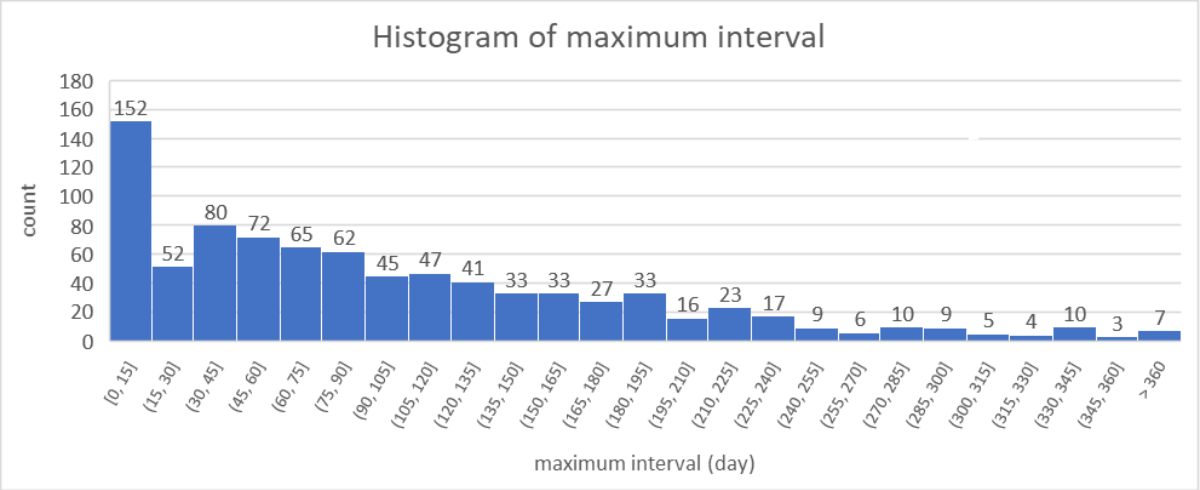


Figure 5.8: Histogram of the maximum interval of Twitter users

Conditions related to the average visit time are only used to distinguish seasonal residents and more-than-once visitors, and the thresholds are 1-3 months and 30 days respectively. There is no clear break near both thresholds as shown in Figure 5.9, so using this feature as the only strong conditions is not reliable.

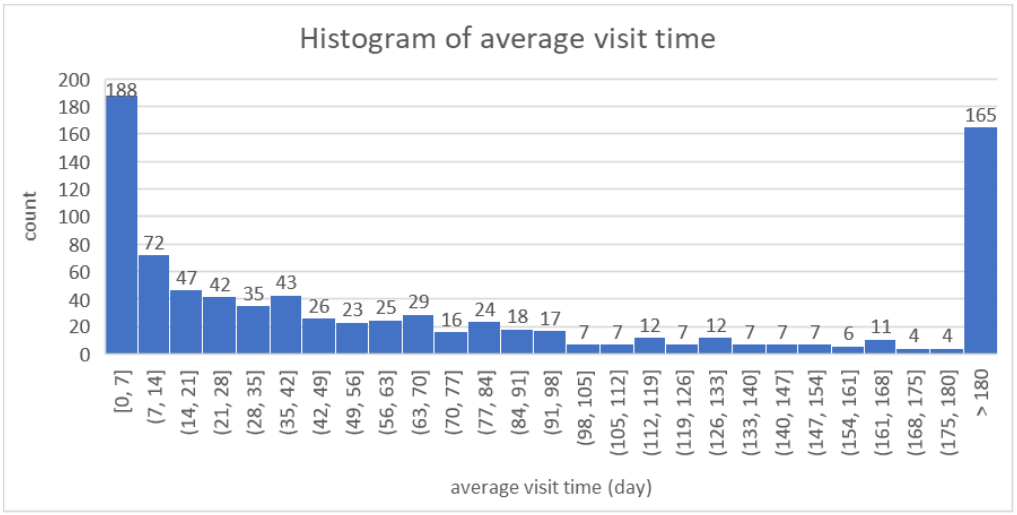


Figure 5.9: Histogram of average visit time of Twitter users

-Spatial feature

The tourist attraction proportion is the proportion of tweet points near to any tourist attraction and it is the only spatial feature used in strong conditions. As shown in Figure 5.10, this feature shows a decrease and the number of users has a relatively sharp drop near 50%, which is the threshold used to identify tourists.

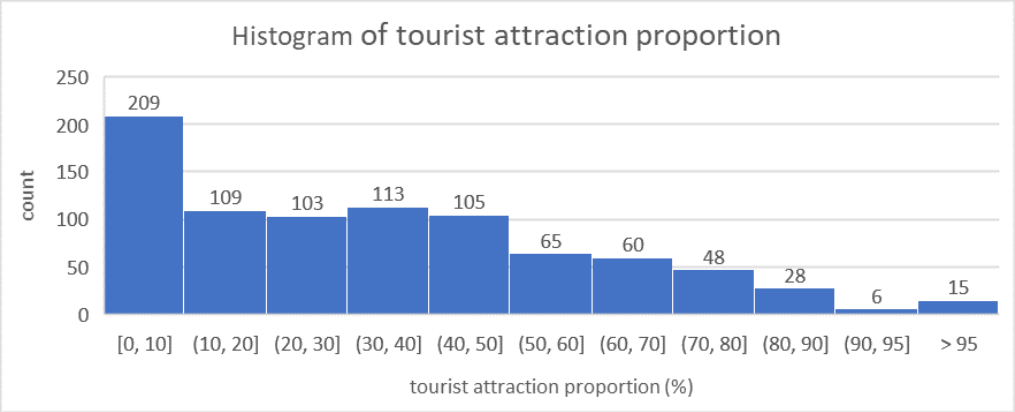


Figure 5.10: Histogram of tourist attraction proportion of Twitter users

The ellipse area and core point proportion are features of weak conditions only, because they are calculated based on unreliable data. Figure 5.11 and Figure 5.12 show the distribution of them respectively. For the ellipses, 42% of the units have an area of less than 10 square kilometres, which may be caused by fewer tweet points or concentrated activity locations. Table 5.3 shows the percentiles of the ellipse areas, and the value of percentiles will be used in localness assessment as thresholds of this feature as mentioned in section 4.3.1. For core point proportion which is called as concentrated-points proportion in section 4.2.2, only the users who can be identified typical locations are shown in Figure 5.12. In this case study, 54% of users have at least one typical location. As can be seen in Figure 5.12, the distribution of this feature is relatively even except for the first two bars. As a feature with a weak condition, it is difficult to identify clear thresholds for localness assessment.

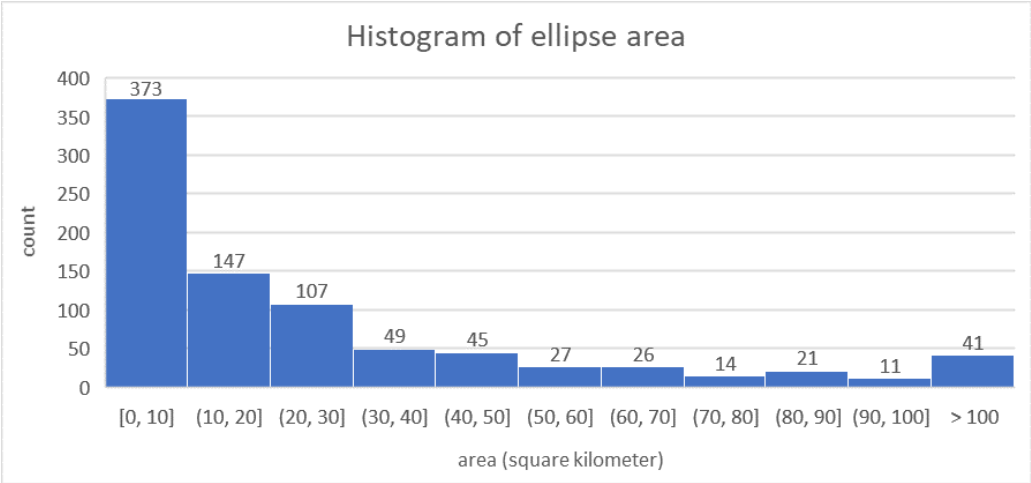


Figure 5.11: Histogram of ellipse area

Table 5.3: Percentiles of Ellipse Areas

Percentile	min	10th	20th	30th	40th	50th	60th	70th	80th	90th	max
Area(km^2)	0	0.14	1.39	2.62	4.18	6.66	9.83	14.07	21.36	35.58	159.21

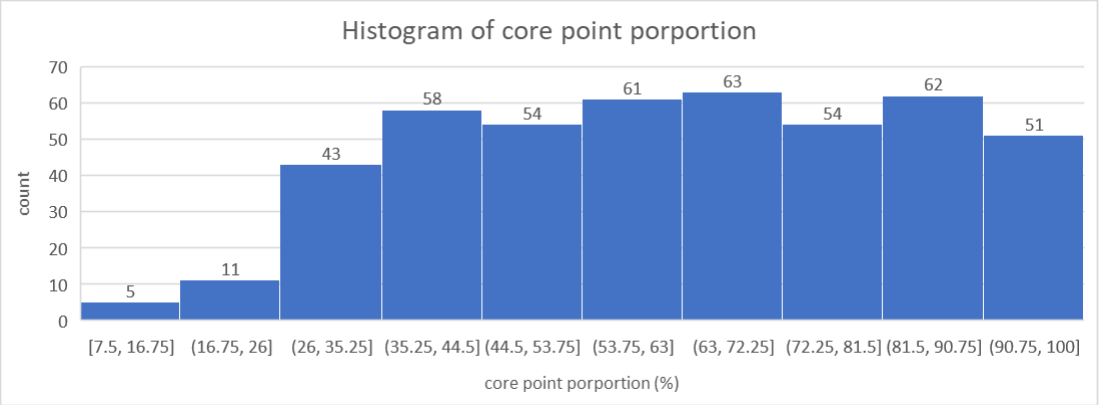


Figure 5.12: Histogram of core point proportion

-Social feature

Local follower proportion is the only social feature used in the case study and it is a feature only with a weak condition too. This feature is extracted based on the location information in the user profile. Without any geoparsing process, the local followers may miss the users who reveal their location in a more detailed way such as specific home address in the city but didn't mention the city name. Besides, the users who do not enter words related to real locations will also be included in the calculation. So, with increased denominator (local followers) and decreased numerator (total followers with location), the result of this feature should be smaller than true value. As shown in Figure 5.13, the values show a decreasing trend, and only a few of the users show more than 60% local followers.

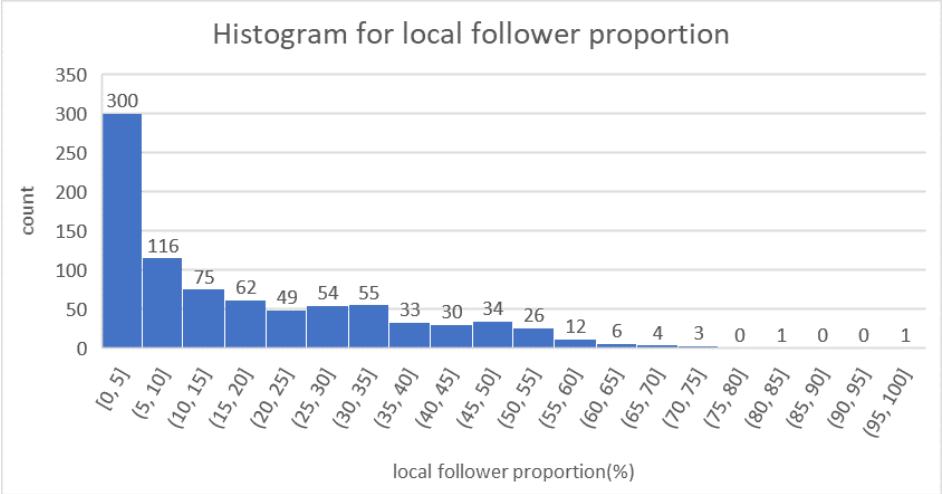


Figure 5.13: Histogram of local follower proportion of Twitter users

5.3. Localness assessment

All the features calculated in Section 5.2 were stored in one table of one PostgreSQL database, and the features were compared with the thresholds that were presented in Table 4.2 to check whether one user meets the conditions of one localness type with the WHERE clause of SQL queries. When the thresholds are used in the localness assessment, one month is converted to 30 days and one year is converted to 365

days. Twitter users are selected based on the conditions of one localness type and that type is assigned to them.

For example, the following code is used to select the users who meet all the conditions of a long-term resident:

```
select * from final_features
where duration>=365 and max_interval<60 and area_km>=14.07 and
core_point_pc>=50 and london_fol_pc >=40 and user_loc like '%london%';
```

In the code, core_point_pc represents concentrated-location proportions, London_fol_pc represents the proportion of local followers and 14.07 is the 70th percentile of all the ellipse area values. After running the query, five of the users was selected which means that five users in the dataset met all the conditions of long-term residents.

In this section, the localness of Twitter users is assessed step by step based on the sequence of localness assessment as presented in Section 4.3.2. Table 5.4 shows the result of the localness assessment in this case study. According to the table, 123 users cannot be assigned any specific localness type, which accounts for 14.29%. Each localness type has some corresponding users which demonstrate the rationality of this approach to a certain extent.

Table 5.4: The number of assessed users for each localness type in each step

Localness	Step 1	Step 2	Step 3	Step 4	total number
long-term resident	5	60	185	-	250
temporary or short-term resident	1	111	79	-	191
seasonal resident	0	16	24		40
one-time visitor	-	-	83	-	83
more-than-once visitor	2	124	-	-	126
tourist	46	2	-	-	48
unknown	-	-	-	123	123
total number	54	313	371	123	861

Figure 5.14 shows the percentage of users by each localness type. In the figure, long-term residents and temporary or short-term residents account for 29% and 22% of the total number of users respectively, and the result also shows that about half of the Twitter users are residents of London. The relatively high proportion of seasonal residents plus visitors and tourists is also within expectations because London is a global city

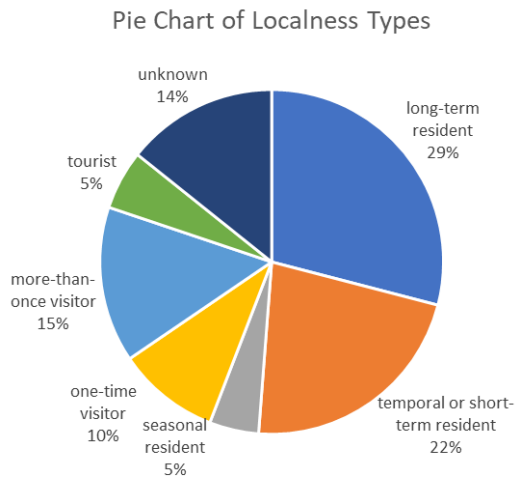


Figure 5.14: Pie Chart of percentage of localness type

As shown in Figure 5.15, the percentage of user assessed in each step are degressive from step one to step three and 43% of users are assessed in step three which are more than assessed users in the first two steps. This is consistent with expectation because from step one to step three the conditions are becoming looser. Only 6% of the users meets all the conditions of one localness type, and this might be due to unreliable features and corresponding weak conditions or unreasonable thresholds in conditions. Conflicts between conditions are the reason of unknown type, and 14% of users cannot be assessed using approach, which also indicates that there exist some unreasonable conditions in the approach.

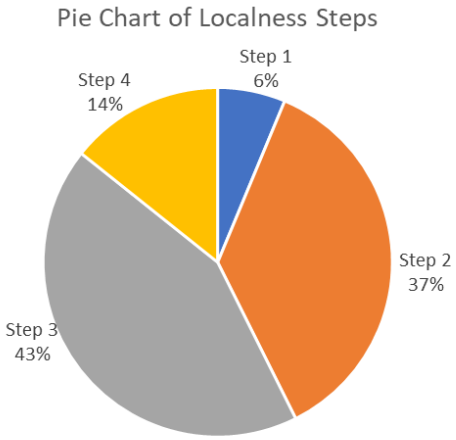


Figure 5.15: Pie Chart of percentage of localness steps

Figure 5.16 shows the distribution of the maximum interval plotted against the duration of users whose localness type cannot be assessed. According to this figure, only four users have a maximum interval and a duration of less than 30 days and most of those users have larger maximum intervals and large durations. The detailed numbers of these two features are shown in Table 5.5 and the four users with smaller duration and maximum are not included in the table because they are acceptable special case and cannot reflect problems of the approach. The maximum intervals of most users with an unknown localness are more than 60 days which is one strong condition to filter long-term, temporary or short-term residents out, but their duration meets the duration condition of these two localness types. These users cannot be assigned as a more-than-once visitor because their average visit time is longer than or equal to 30 days. So, the users

shown in **bold** cannot be assigned to any localness type due to the condition conflicts between duration condition and maximum interval condition.

There are three possible reasons to explain the situation: first, these users may have left London for a few months so that their tweets posted during the period were not collected; second, these users may not use geographic information in all of their tweets, and tweets without any geographic information are not collected either, so the tweets in the dataset show that they did not post tweets in the city for a long time; third, users may simply do not want to tweet in such a long period.

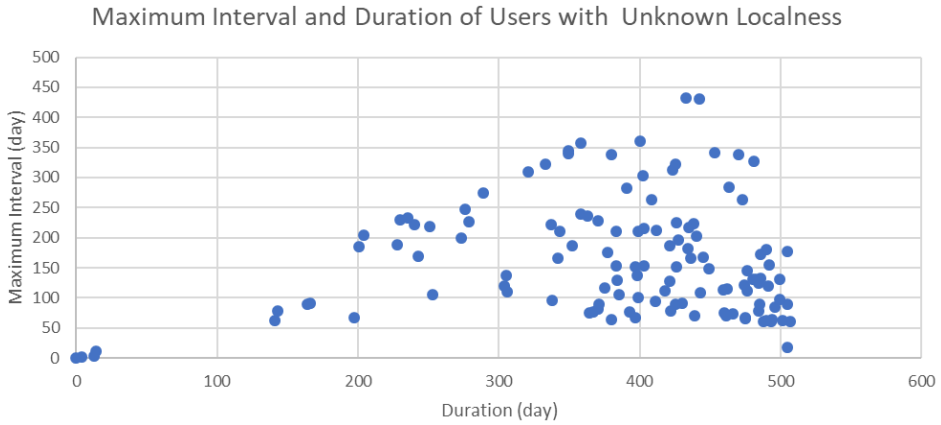


Figure 5.16: Scatter plot of maximum interval and duration of users with unknown localness

Table 5.5: Maximum interval and duration of users with unknown localness

	Max_interval <60	60 ≤Max_interval <180	Max_interval ≥ 180	Total number
Duration ≥450	1	***30***	6	37
365≤duration<450	0	***27***	***21***	48
30≤Duration <365	0	***13***	***21***	34
Total number	1	70	48	119

5.4. Validation of localness assessment result

5.4.1. Ground Truth

To validate the localness assessment results, I selected some users and labelled their localness manually to serve as ground truth. From all users in the final dataset, I selected 275 users randomly using SQL and the random() function in the same way as the sampling in Section 5.1.3 was done. For each selected user, I searched for all tweets of him/her, ranked the tweets by their timestamps, read the text in each tweet and generated a tweet map based on tweet points. The information on which I based the manual localness determination came from the tweet texts, user locations in the social media profiles, locations of the user's followers, and from the temporal and spatial distribution of the tweets. The localness labelling mainly followed the description of localness types in Section 3.4, and there were no strict conditions used in the manual localness labelling. The following paragraphs show typical users for each localness type, as examples of manual localness labelling, and typical users mean that their situations should be common cases in each localness type.

The typical long-term resident stayed in the city for 504 days and posted 407 tweets. The longest period without any tweet is 27 days. He uses London so as to show his location and 33% of his followers mention the term "London" in their location field which indicates a relatively strong local social network. In his tweet text, he mentioned office, work, meeting friends, celebrations for festivals, foods and many other things about daily life. His tweet points are widely distributed in the study area and some typical locations can be found in the area with a higher point density as shown in Figure 5.17(a).

The typical user with a temporal or short-term resident localness stayed in London for 76 days without any longer interval break and posted 122 tweets. His location information indicates that he comes from America and he posted tweets about American politics. He posted his last geo-tagged tweet in London in September 2017 which can be considered as his departure time. The spatial scope of his/her activities smaller than the scope of the typical long-term resident and typical locations can also be found as shown in Figure 5.17(b).

The typical seasonal resident has two visit periods in London. The first one is from July 2017 to September 2017, and the second visit period is from July 2018 to October 2018. In her tweets, she mentions home, work, food and tourist attraction, and some of her tweets are not in English. Either she or any of her followers mention the term "London" in the user profile. Her activities are located in a relatively wide area but most of them are near to typical locations as shown in Figure 5.17(c).

The typical one-time visitor stayed in London for 13 days and posted 12 tweets. Almost all of his tweets are about food and tourist attractions and he mentions summer vacation in some of the tweets. He discloses his location as Philadelphia and only 1% of his followers mentioned the term "London" in their user profile.

The typical more-than-once visitor stayed in London for 7 days according to the dates when he posted tweets, but the time difference from the earliest tweet to the last tweet is 425 days. The overall duration can be sliced by two long intervals. The first interval is about three months, and the second interval is 11 months. He stayed for less than 5 days in each visit and mentions work and employers in some of his tweets. Both one-time visitors and more-than-once visitors are labelled as visitor based on the localness type description in Section 3.4, and the spatial distribution of visitors is variable as shown in Figure 5.17(d) and Figure 5.17(e).

The typical tourist stayed in London for three days and posted 27 tweets, not in English. Almost all her tweets are about tourist attractions and she also posted tweets near tourist attractions. In Figure 5.19(f), the yellow points are his tweet points and the red points represent popular tourist attractions in London.

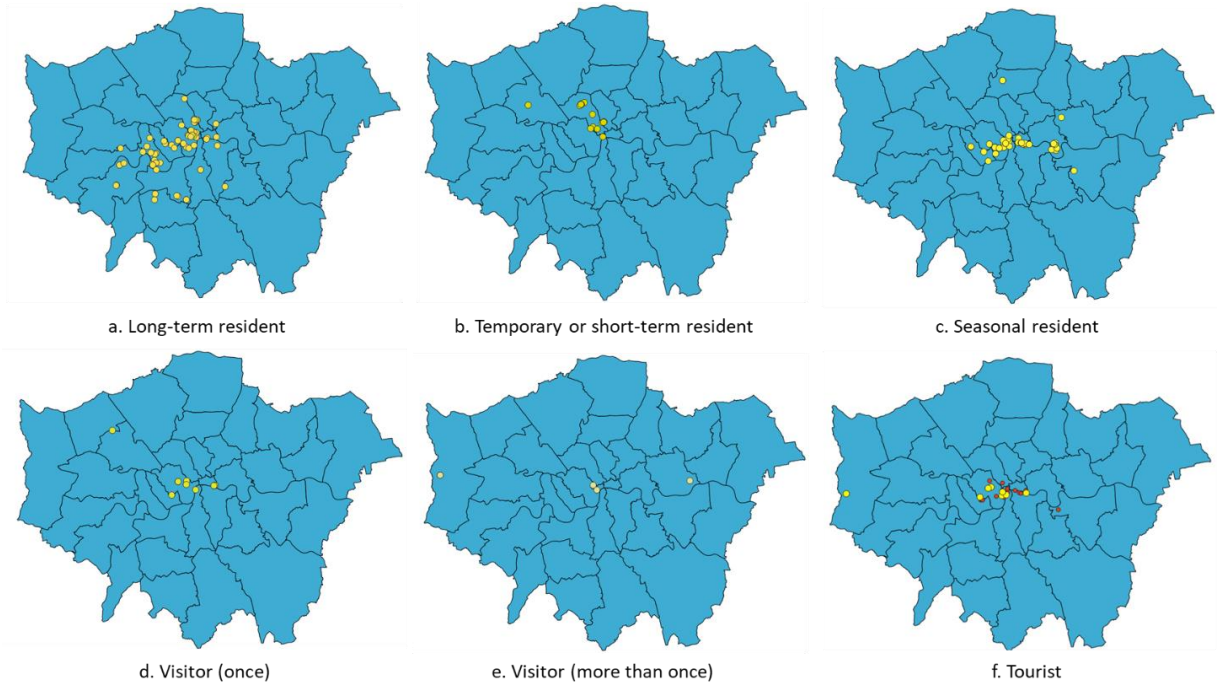


Figure 5.17: Spatial Distribution of typical users' tweet points

For the above typical users, their characteristics are obvious enough for manual localness labelling, but not all the users in the dataset can be identified clearly. Users may only post some comments or record their mood without any information to indicate the relationship between them and the city. Some users only posted a few tweets which are not enough to obtain supportive information for localness labelling. The localness labelling is based on all the available information from the dataset instead of the user features used in the localness assessment approach. Therefore, the localness ground truth is reliable to validate the results of the localness assessment approach implementation.

5.4.2. Validation and discussion

Table 5.6 shows the user number of each localness type in the ground truth set and in the test set. The test set contains the users whose localness was both assessed in the case study and labelled as ground truth localness in the meantime. Because the classifiers in the localness assessment approach cannot classify the localness of 46 of the 275 randomly selected users for the validation and they are labelled as unknown, so these users will not be considered in the comparison of localness types resulting from ground truth and applying the localness assessment approach. Therefore, only 229 localness data are used in the validation.

In the ground truth, long-term residents and temporary or short-term residents account for 73% of the total, whereby temporary or short-term residents are significantly less predicted by the approach compared with the ground truth. Only two users are labelled as a seasonal resident because it is difficult to separate long-term residents with a long interval break from seasonal residents. There are only six seasonal residents in the test set, so the validation for seasonal residents may be not very convincing. Visitors and tourists account for 15% and 11% respectively in the ground truth data and there are more visitors and fewer tourists in the test set compared with the ground truth. One reason for this situation may be that temporary or short-term residents with relatively long maximum intervals were classified as visitors while they simply used Twitter at a lower frequency.

Table 5.6: User number of each localness type

Localness Type	Long-term resident	Temporary or short-term resident	Seasonal resident	Visitor	Tourist
Ground Truth	76	92	2	34	25
Test Set	83	61	6	66	13

A confusion matrix is a tool that is often used in the evaluation of classification models. In essence, localness assessment is a classification based on the user's localness. To describe the performance of classifiers used in the localness assessment approach, confusion matrixes are created. Table 5.7 shows the confusion matrix of a binary classification case. In the table, condition positive and condition negative show the real positive and the real negative case in columns, whereby the cases are manually checked. Predicted positive and predicted negative show the prediction results of classifiers in rows. Numbers in the grey cells are the numbers of cases which are predicted correctly.

Table 5.7: Confusion Matrix of a binary case

	Predicted positive	**Predicted negative**
Condition positive	True positive	False negative (Type II error)
Condition negative	False positive (Type I error)	True negative

Powers (2011) summarized commonly used evaluation measures used in confusion matrix interpretation and four of them are used in this thesis: accuracy, recall, precision and F1-measure (F1 score). The following formulas show how these measures are calculated:

$$\text{Accuracy} = \frac{True\ Positive + True\ negative}{Total\ number}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{F1 measure} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Accuracy represents the proportion of correct predictions, and this measure is used to evaluate the overall performance of the classifiers. Recall means how many actual positives are predicted as positives. In the case study, recall means how many users with one specific localness type as ground truth are assessed as this very localness type. Precision means how many predicted positives are actual positives. In the case study, precision means among the users who are assessed as one specific localness type how many of them are correct compared with ground truth. F1-measure considers both precision and recall and is a balance between precision and recall which can avoid the effect of imbalanced data. In the case study, much more users are assessed as long-term residents or temporary/short-term residents in the ground truth than visitors

and tourists as was shown in Table 5.6. Therefore, the F1-measure is also necessary to evaluate the classifiers used in the approach proposed in this thesis.

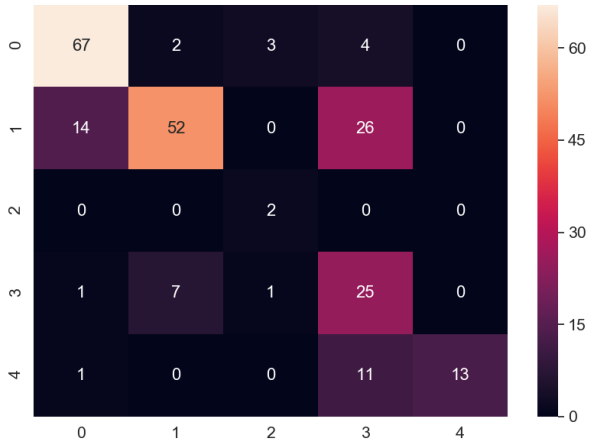


Figure 5.18: Confusion Matrix of localness assessment result

Figure 5.18 shows the confusion matrix of the localness assessment results. Rows in the matrix show ground truth and columns show prediction results. Number zero to four along axes represents long-term resident, temporary or short-term resident, seasonal resident, visitor and tourist respectively. As calculated on the basis of the confusion matrix, the accuracy of the localness assessment results in the case study is 69.43%. Table 5.8 shows the recall, precision and F1-measure of each localness type.

Table 5.8: Evaluation measures for each localness type

	Precision	**Recall**	**F1-measure**
Long-term resident	0.81	0.88	0.84
Temporary or short-term resident	0.85	0.57	0.68
Seasonal resident	0.33	1.00	0.50
Visitor	0.38	0.74	0.50
Tourist	1.00	0.52	0.68

Overall, the localness assessment of long-term residents is the best of all the localness types, but seasonal residents and visitors have a relatively low performance based on F1-measure. From the recall perspective, the prediction of the seasonal localness type is the best and all the real seasonal residents are predicted correctly. However, there are only two seasonal residents in the ground truth, so this recall is not convincing. From the precision perspective, the predictions of long-term residents, temporary or short-term residents and tourists are much better than the other two. The precision of tourist prediction is perfect, which means that all the users who are predicted as tourists are real tourists based on the ground truth.

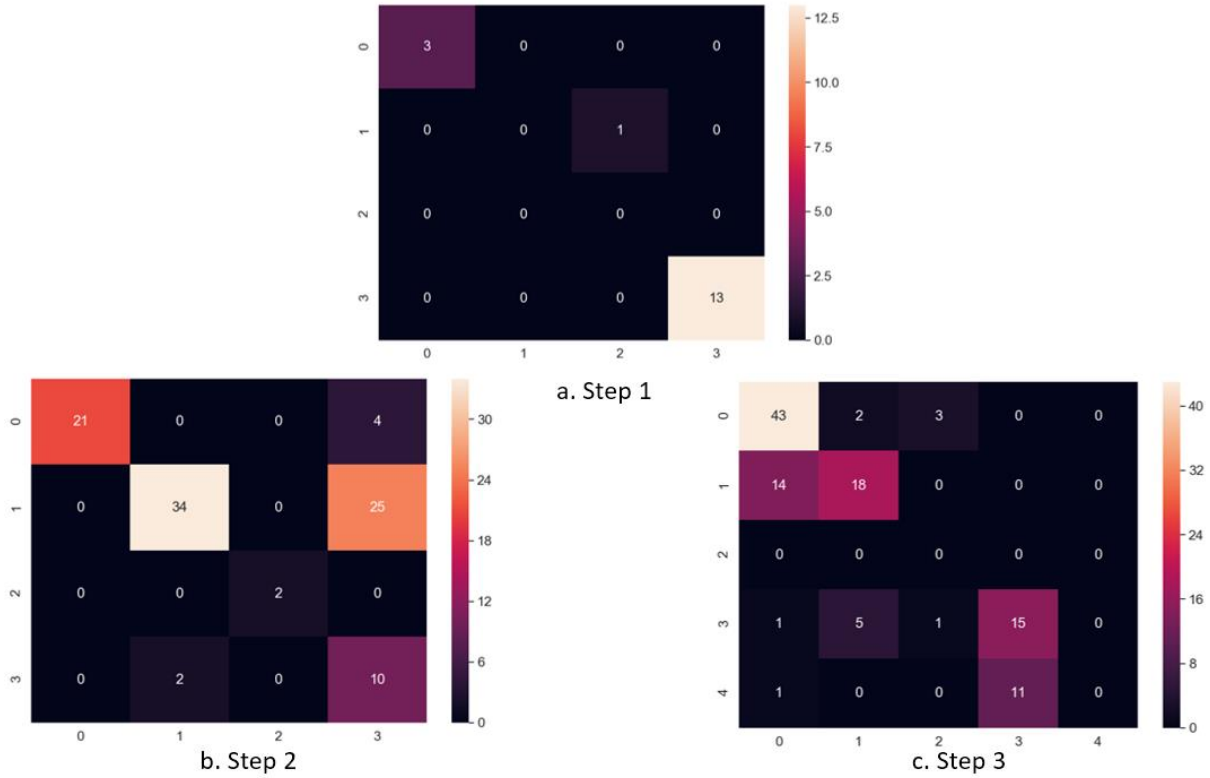


Figure 5.19: Confusion Matrix of each step in localness assessment

Figure 5.19 shows confusion matrixes of the first three steps in the localness assessment (see Section 4.3.2). Not all localness types appear in each step because some localness types appear neither in the predicted result nor in the ground truth. In step 1, seasonal residents are missing and number 2 and 3 represent visitor and tourists respectively. In step 2, tourists are missing so there are only four classes in Figure 5.19(b). The other axis labels are the same as Figure 5.18.

The overall accuracies of the three steps are 94.11%, 70.41% and 66.67% respectively. The result of the first step has the highest accuracy as expected, because all features and conditions are considered in the first step. However, the result of the second step does not show significant advantages of the combination of all strong conditions when it is compared with the result of the third step, which indicates that the determination of strong conditions may have some problems. Combining the confusion matrix of the overall approach and the confusion matrices of the three steps can be helpful for locating mispredictions into steps.

As can be seen in Figure 5.19, 26 temporary or short-term residents are incorrectly assessed as visitors: one of them occurs in step 1 and 25 of them occur in step 2. Conversely, 7 visitors are incorrectly predicted as temporary or short-term residents: two of them occur in step 2 and 5 of them occur in step 3. All the mispredictions could result from the prediction of more-than-once visitors because the duration condition of one-time visitors can tell temporary or short-term residents apart. All weak conditions of both temporary or short-term residents and more-than-once visitors are the same. Strong conditions of temporary or short-term residents are about duration and maximum interval, while the conditions of visitors focus on the maximum interval and average visit time. The core difference between these two localness types is maximum interval: the maximum interval for the former should be shorter than two months while the maximum interval for the latter should be longer than two months and shorter than six months. More mispredictions for these two types indicate that thresholds in strong conditions can mislead the localness assessment, especially the thresholds for the maximum interval. Moreover, sometimes it is difficult to distinguish visitors

from temporary residents, especially for the users who visit the city frequently and the temporary users who post tweets at low frequencies.

14 Long-term residents are incorrectly assessed as temporary or short-term resident and two long-term residents are incorrectly predicted as temporary or short-term resident. All these mispredictions occur in step 3. The duration of these users is shorter than one year, but the maximum interval is more than two months. So, based on these strong conditions of long-term residents these users should be excluded. However, there is another strong condition that is applied in long-term resident identification: location in the user profile. Common point of these mispredictions is that they all mention the term "London" in their location field, and they can be identified as long-term resident if they meet any weak condition of this localness type on the basis of meeting this strong condition. Therefore, choosing location information in the user profile as one feature with a strong condition is not wise. In addition, the same reason leads to the misprediction of one visitor and one tourist to two long-term residents.

11 Tourists are assessed as one-time visitors in step 3, as inferred from the duration conditions of these localness types. The durations of these users are less than 7 days and meet the duration condition of tourists, but the tourist attraction proportions are smaller than 50%, so they are excluded as tourists. However, the fact is that not all tourists post tweets near attractions and the attraction list used in the case study only contains 20 tourist attractions. Therefore, 50% is a relatively high threshold for tourist attraction proportions and it may lead to mispredictions.

4 Users are predicted as seasonal residents while three of them are long-term residents and one of them is a visitor. These mispredictions occur in step 3. It is difficult to discover clear seasonal movement patterns in the identification of seasonal residents because the dataset only covers about 16 months. Based on this fact, both seasonal residents in the ground truth and the localness assessment results are not reliable. Without seasonal movement pattern as a strong condition, long-term residents or visitors who meet the duration condition of seasonal residents and post geo-tagged tweets rarely could be classified as seasonal residents.

The last kind of mispredictions occurs in step 2: 4 long-term residents are classified as visitors. The maximum interval of all these users is more than two months, which is the main reason excluding them from long-term residents. In step 2, weak conditions are not used and users whose average visit time is shorter than 30 days and maximum interval is less than six months are assigned as visitors.

As for the users whose localness is assigned as unknown, the ground truths of their localness are shown in Table 5.9. The main reason of this unknown localness is the conflict between duration conditions and maximum interval conditions as mentioned in Section 5.3. To assess localness of these users, thresholds used in the maximum interval condition of long-term residents, temporary or short-term residents and more-than-once visitors should be increased or changed to fuzzy thresholds to avoid this condition conflict. It turns out that the unknown localness for one tourist in the table came from a mistake in the feature calculations.

Table 5.9: Ground truth of users assessed as unknown

Localness Type	Long-term resident	Temporary or short-term resident	Seasonal resident	Visitor	Tourist
Number	21	6	1	17	1

In this chapter, the localness assessment approach has been applied to Twitter data from London, and the results of the approach implementation were compared with ground truth. The evaluation showed that the approach can assess the localness of the majority of users, and the assessment accuracy is 69.43%. In the next chapter, the problems of the approach and the case study will be discussed partly based on the contents of this chapter.

6. DISCUSSION

This chapter answers the last research question: what are the application conditions and limitations of the approach? In this chapter, the limitations in the localness description will be explained and problems in the approach are specified.

6.1. Localness definition and types

Localness is a term commonly used in many fields as mentioned in section 2.1. In this thesis, localness has been defined as the result of accumulating life experiences in a local environment and it is also a representation form of the relationship between people and cities. The logic used in the localness definition is that local knowledge is generated from life experiences and the accumulation of life experiences indicates the relationship between one person and one city.

However, although having related life experiences is a precondition of generating local knowledge it cannot guarantee local knowledge generation. Many life experiences are repeated and that contributes little to local knowledge. Local knowledge abstraction from life experiences is highly depending on a person's characteristics. So, more life experiences do not mean more local knowledge, but more life experiences are just a precondition of more local knowledge. Local knowledge is more unobservable than life experiences, and that is why my localness definition uses life experiences rather than local knowledge directly.

The result of accumulating life experiences (i.e. localness) can only indicate and represent the relationship between people and cities but cannot determine the relationship. A strong relationship means that people's activities are closely related to one city and people experience the city widely and deeply. The relationship is determined by the will of persons. A person decides how long he will stay in one city, how many places he visits and how many local people he wants to contact, and so on. Life experiences are the result of these decisions, and localness is the integration of these life experiences. Therefore, localness can only represent the relationship to a certain extent.

The relationship between people and one city is represented as localness types in this thesis. Localness types are conceptualized based on the localness definition and on mobility forms and this kind of conceptualization has some limitations from the perspective of the relationship representation.

First, the type selection is greatly influenced by the temporal dimensions of life experiences accumulation and by the classification of mobility forms. The assumption used here is that the longer people stay in the city, the more life experiences they will have. The assumption is reasonable in most cases but there are some special situations. For example, if the available data for localness assessment only cover two years, a long-term resident may repeat his daily routine without extra activities, while a temporary resident may spend a lot of time in visiting local venues, restaurants and tourist attractions. So, this temporary resident will have generated more life experiences in the time period of the dataset.

Second, the type selection considers people's purposes of visiting the city, but it is difficult to clearly tell all the purposes apart. The non-local commuter is selected as a separate type because these people's life experiences focus on work and are less about other parts of everyday life. But non-local commuters can have more activities in the working city if they want, and, therefore, it is difficult to distinguish

them from long-term residents. Another typical example is about the localness type visitor. People can visit one city for various purposes, although they usually have a main goal among all of their purposes. If they visit the city for work and stay in the city for several months, they will be temporary residents because they have more experience about daily life in the city than experience about work and if they go to one city for visiting friends and spend more time on tourism, they can be treated as tourists. All in all, the current localness types do not consider all these special cases but only pay attention to common cases.

Third, the classification of localness types in this thesis covers all possibilities from a temporal perspective and consider the purposes of visits in the meantime. But other factors can also be used to develop a localness taxonomy. Age, gender, socioeconomic status and any other characteristic which can influence people's lives in one city significantly, are all potential factors in the development of a localness taxonomy because they can lead to an obvious difference in the people's life experiences.

Fourth, when I proposed the localness types, I used temporal thresholds to distinguish types, such as more than one year for long-term residents, more than one month for temporary residents, less than 7 days for tourists. These thresholds are determined from a life experience perspective and they are not consistent with some official standards. For example, the definition of a long-term resident by the European Commission specifies that long-term residents should live legally in an EU state for at least five years (European Commission, n.d.), while the threshold of long-term residents in my localness type is only one year. The reason for this one-year threshold is that I assume that one year is enough for one person to form his routine life in one city and get familiar with the city. Similar reasons can explain temporal thresholds for other localness types. These thresholds are not fixed and strict for each localness, and they can be changed based on the requirement of target people group identification in other localness assessment studies.

6.2. Localness assessment approach

Compared with studies related to localness assessment, the value of this localness assessment approach has three aspects. First, the result of this approach shows the relationship between people and cities as six localness types, which covers more possibilities of the relationship than the results of existing works (only local or non-local). Second, twelve user features from three aspects are considered in localness assessment and each assessment is based on at least two user features, which makes the approach more robust. Third, localness is defined based on local knowledge and life experiences, so the approach can support local knowledge discovery from social media data.

The localness assessment approach is designed to assess the localness of social media users based on geo-social media data. The approach may be used to assess the user's localness in any city based on any geo-social media dataset, as long as user features can be extracted from the dataset and the time period of the dataset meets the duration condition of the target localness type(s). Temporal features can be extracted from the timestamp of social media contents, spatial features can be extracted from the geographic information attached to what users posted and social features can be collected from user followers and following accounts. Each geo-social medium has its own characteristics and target users, but functions for posting timestamps, geographic information and social networks are commonly used in geo-social media. Available user features based on different geo-social media might be different due to characteristics of the social media platforms. For example, local organizations may create accounts on Twitter but not on Flickr, so social features related to local organizations may not available when using the Flickr dataset to assess the user's localness.

User features are selected from temporal, spatial and social perspectives, in line with the localness properties. They are accessible in all the geo-social media datasets and can display characteristics of user activities from different angles. But people do not record all their activities on social media and only a small percentage of their postings are geo-tagged. So, the social media data of one user may only show a limited number of activities and which activities are geo-tagged depends on the user. This fact leads to some problems. First, users may only post geo-tagged content when they execute some special activities, such as visiting new places. In such a case, records of most activities will be missing in the dataset. Second, social media users may use a social medium at a lower frequency, or not at all, and still do not leave the city in one time period. In addition, users may provide fake information in their user profiles on social media, such as fake locations. All these problems are reflected in the user features and may lead to wrong results in the localness assessment.

With respect to the localness assessment conditions, some are selected as strong conditions because the features used in these conditions are more reliable. However, because of the problems of social media data as described above, the features are only relatively reliable, and the features cannot always reflect well the reality of users' lives, even though more information is considered when these features are extracted. Conditions are classified as strong and weak, and this may have a great influence on the results because of the assessment sequence that was presented in section 4.3.2. The reasons for selecting strong conditions are explained in section 4.3.1, but only the amount of information used in feature extraction cannot determine whether one feature is more important or not. The strong conditions should be more important from a localness assessment perspective but may not be limited by the data. Based on the outcomes of the case study, selecting location information in the user profile as one feature with strong conditions has an obvious influence on the localness assessment results and may lead to some wrong results, which indicates that removing this feature from strong conditions to decrease the influence of this feature can improve the approach.

Features and thresholds compose conditions, and the determination of thresholds can also lead to some problems. First, the thresholds are determined based on the localness type descriptions and the specific numbers used in the thresholds are tentative values. For the feature duration, maximum interval and average visit time, the thresholds are absolute values based on the definition of localness and common sense. For other features, relative values are used in other conditions to avoid the effect of differences between datasets. Before the threshold determination, the results of each feature were explored first to make sure that the thresholds were suitable for social media data and not only based on the theoretical assumptions. However, the thresholds used in the approach that is proposed in this thesis are still tentative values, and they may not fit the features of users who should be classified as one specific localness type in reality. Therefore, the thresholds may be modified based on ground truth data to fit the reality and the dataset. Second, all the condition combinations that are used in the localness assessment are meant to distinguish one localness type from others and one user can only have one localness type. The thresholds divide all potential values of each feature into slices without any overlap, but it is possible that one user is in a fuzzy area between two localness types when considering all the user features. A strict division can also lead to many wrong results in the assessment, especially for the users whose behaviours are not accorded with the typical situations of localness types. The approach may possibly be improved by finding the thresholds which lead to more assessment failures and modifying them based on the ground truth. Another improvement can be to design fuzzy areas between localness types by letting the possible value ranges of features overlap and selecting the users who have multiple localness types, and then checking the users manually.

In validation part, ground truth data is labelled manually based on all available information of users in data set. But the information is limited and can only reflect a few of users' life experiences. Moreover, there is no strict standards used in the ground truth labelling, which means it's very likely to make mistakes due to subjective judgment in the process of manual labelling. Kariryaa et al.(2018) used Twitter advertisement platform to gather ground truth. This method is limited by the Twitter ads platform on user selection and use-cost and the result of ground truth is highly depended on how people define "local". Given these limitations, the method was not used in this thesis. Therefore, more reliable ground truth collection method can be a good motivation for further research, and the validation result based on ground truth in the thesis is not entirely trustworthy.

Because of the problems in feature selection and threshold determination, machine learning may improve the approach significantly. Using machine learning, each feature can have a weight when it is used in localness assessment, and the weights can be obtained from a training set. The classification conditions can be calculated automatically based on the training set, and then the thresholds in conditions will fit one specific data set well. However, the precondition of using machine learning is a reliable and sufficient ground truth. Due to the time limitation of this thesis research, I did not have the time to sufficiently label the true localness of users. In addition, the reliability of a ground truth from manual labelling is also a problem. Therefore, machine learning was not used in this thesis.

In this chapter, I reviewed the description of localness that was introduced in Chapter 3 and the approach designed in Chapter 4 and I explained the reasons for the problems that emerged. Due to time and data limitations of this thesis research, some improvements of the approach could not be implemented, and, therefore, they will be presented as recommendations for further research in the next chapter.

7. CONCLUSION AND RECOMMENDATIONS

7.1. Conclusion

This study reviewed existing localness definitions and criteria and defined the localness of individuals by means of distinguishing specific localness types to represent the relationship between people and cities. An approach was designed to assess the localness of social media users based on conditions and user features extracted from social media data, and this approach was tested through assessing the localness of Twitter users in London. The results of the case study show that the approach can assess the localness of most users correctly.

In this study, the research questions that were listed in Chapter 1 were answered as follows:

Sub-objective 1: Evaluate existing localness definitions and assessment criteria

Research question 1: How do related works define localness?

In studies related to UGC, localness is defined as the proportion of local elements in a VGI repository or in a group of social media users, and localness is also used to represent how local one person or one regional term is.

Research question 2: Which are the criteria used in related works to assess the localness of individuals?

Existing criteria can only classify people into local or non-local and all these criteria are used separately. To filter out non-local people, simply time spans with a specific number of days are used. For example, if the time difference between the earliest and the last social media post of one user is less than 30 days, the user will be treated as a visitor. Other user features that are used to identify local people are, for example, the number of local organizations one person follows, the number of local venues one person has ever visited, and the number of regional terms someone includes in his social media contents. But the studies using these features have not provided clear criteria and only proved that people following more local organizations, visiting more local venues or mentioning more regional terms are more likely to be local. Johnson et al. (2016) summarized four criteria: n-day, plurality, geometric median and location field. The first one is the same as the time span, the second means that people should have more geo-tagged social media in the city where they are treated as local, the third is the inference of the people's home location and the last one is using the location field of social media users directly. All existing criteria can only classify people into local or non-local but ignore the potential relationship between people and the area. The criteria are used separately, and a combination of criteria has not been used in existing works.

Sub-objective 2: Define individual localness and specify localness types

Research question 1. How to define localness of individuals?

Localness of individuals is defined as the result of accumulating life experiences in a local environment. It describes the relationship between individuals and areas, and this relationship indicates the potential local knowledge of individuals from an area perspective. Localness in this thesis is at the city scale.

Research question 2. How to conceptualize different types of individual localness?

Localness types are conceptualized based on the localness definition, so the main basis is the life experiences one person has. Mobility forms also contribute to the localness type conceptualization as a consequence of the relationship between localness and mobility. Considering both localness

definition and mobility forms, the temporal dimension of life experience accumulation and visit purposes are the main criteria for specifying localness types. Finally, six localness types were specified in this study: long-term resident, temporary or short-term resident, seasonal resident, non-local commuter, visitor and tourist.

Sub-objective 3: Design an approach to assess the localness of social media users

Research question 1. Which user features can be used to determine the localness type of users?

Three types of information can be helpful for determining the localness types of users, based on the localness properties: temporal, spatial and social. They reflect the result of life experiences accumulation from three different and necessary angles. Temporal features contain duration, maximum interval, average visit time, night post proportion and weekend post proportion. Duration indicates how long one user stays in one city and is calculated based on the earliest and the latest social media posting of that user. Maximum interval suggests the longest time period during which one user did not post anything. The average visit time is the average of all visit times if one user visits one city more than once. Night and weekend post proportions indicate when one user is posting social media contents in one city from within a day and a week respectively. Spatial features contain the area of the standard deviational ellipse, concentrated-location proportion and tourist attraction proportion. The ellipse area represents the scope of the users' activities, the concentrated-location proportion indicates how concentrated the users' activities are, and the tourist attraction proportion reflects how many activities of one user happened near tourist attractions. Social features contain the local proportion in the social network, user interest, and number of local organizations followed. The local proportion in the social network is the proportion of one user's local followers or followings, user interest indicates potential life experience topics, and followed local organizations suggest a connection between users and local society. Other features contain the location field in the user profile and language. The location information that is shown in the user profile may reflect the user's location, but this is not very reliable. Finally, language can identify non-local users if they do not use the official local language in one city.

Research question 2. How to assess the localness of a social media user based on user features?

As shown in Figure 4.1, after data collection and cleaning, user features will be extracted from social media data, and then user features will be compared with conditions of localness types step by step to determine the localness of users.

Features and thresholds compose conditions for each localness type. For example, "duration is more than one year" is one condition for long-term resident identification. Not all features are used in each localness type identification because some features are meaningless for some localness types. For example, the night posting proportion is not considered in most localness assessments, because most users except for non-local commuters stay in the city both during the day and at night and, therefore, this feature is not helpful for telling most localness types apart. Some features make more use of user information and are more reliable, and, therefore, these features and corresponding thresholds compose so-called strong conditions, while other features are regarded as weak conditions.

Conditions are combined to select users and all selected users are assigned one localness type. To assess more users, users are selected step by step (see Section 4.3.2) from strict to loose selection conditions, because not all users can meet all the conditions of one localness type at the same time and users may still be selected as long as they are distinguishable by any acceptable combination of conditions. Users are selected first if they meet all conditions of one localness type and then they are assigned to that localness type. Then all the users who meet all strong conditions of one localness type will be selected and assigned to that localness type. After that, condition combinations which can distinguish one

localness type from others are used to further classify users. Finally, the remaining users are classified as "unknown" in case the approach cannot assess the localness of these users.

Sub-objective 4: implement and evaluate the approach using real-world data in a global city

Research question 1. To what extent can the approach assess the user's localness correctly?

Based on the confusion matrix of the overall approach application in the London case study, the accuracy of the localness assessment is 69.43% compared with the ground truth. This is an acceptable result considering the limitations of social media data and the reliability of the ground truth. The approach functions best for the long-term resident identification, whereas the localness assessment of season residents and visitors is not reliable based on F1-measures.

Research question 2. What are the application conditions and the limitations of the approach?

The approach can be used to assess user localness in any city based on any geo-social media dataset, especially for global cities, as long as user features can be extracted from the dataset and the time period of the data set meets the duration condition of the target localness type(s).

The main limitations of this approach are that the selection of strong conditions and the determination of thresholds are based on theoretic assumptions which may not fit specific datasets well. The selection of strong conditions has a great influence on condition combinations because each condition combination should have at least one strong condition. Thresholds divide the range of features without any overlap, and strict distinctions between localness types may mislead the identification of users whose features are near to the thresholds.

7.2. Recommedations

There are some recommendations for further research based on ideas that came up during the research period and on the limitations of this study.

1. The localness taxonomy can be improved by considering more special cases of social media users and developing more detailed localness types. For example, visitors can be classified based on their main purposes and non-local students can be considered as other types because of their obvious distinct movement patterns related to school schedule. Localness can also be conceptualized from any other perspectives as long as they are related to life experience generation.
2. To make the utmost of geo-social media data, geoparsing tools can be helpful. With geoparsing tools, place names used in geotags of social media data can be converted to precise locations and used in spatial feature extraction so that more records of user activities locations can be considered in the localness assessment. Geoparsing tools can also convert user location descriptions to unambiguous geographic identifiers, and they can make this location field more useful and more reliable in both the local follower proportion calculation and user location identification, as disclosed by themselves in their user profiles.
3. Natural Language Processing (NLP) tools can be useful to extract information from social media contents. With NLP tools, important terms mentioned in social media content can be identified and user interests can be summarized from these terms and can be used as one user feature.
4. More additional data can enable more user features or make features more reliable. Limited by the time for this study, social media accounts of local organizations were not collected and only 20 popular tourist attractions were considered in related feature extraction. If a reliable list of local

organization accounts is available, the feature followed local organizations can be used in localness assessment as one important social feature. If more locations of tourist attractions would have been used in the case study, more activities related to leisure or tourism could be identified and related features would be more reliable.

5. The approach is tested using Twitter data in London area in this thesis, but it is designed as a generic approach which can be applied in any city using any geo-social media data set. Therefore, the approach can be implemented to other cities using data and other geo-social media data set to test the application effect of the approach.
6. Given the limitations of feature selection, strong condition selection and threshold determination, machine learning may improve the approach significantly. Feature selection and strong condition selection can be replaced by feature weights and these weights can be calculated based on a training set, which makes the weights reliable. Threshold determination can also rely on model training, which can make the localness assessment model fit the entire dataset better.

LIST OF REFERENCES

REFERENCE

Andrienko, N., Andrienko, G., Fuchs, G., & Jankowski, P. (2016). Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. *Information Visualization*, *15*(2), 117–153. https://doi.org/10.1177/1473871615581216

Ballatore, A., Graham, M., & Sen, S. (2017). Digital Hegemonies: The Localness of Search Engine Results. *Annals of the American Association of Geographers*, *107*(5), 1194–1215. https://doi.org/10.1080/24694452.2017.1308240

Bell, M., & Ward, G. (2000). Comparing temporary mobility with permanent migration. *Tourism Geographies*, *2*(1), 87–107. https://doi.org/10.1080/146166800363466

Black, K., & McBean, E. (2016). Increased Indigenous Participation in Environmental Decision-Making: A Policy Analysis for the Improvement of Indigenous Health. *International Indigenous Policy Journal*, *7*(4). https://doi.org/10.18584/iipj.2016.7.4.5

Brabham, D. C. (2009). Crowdsourcing the Public Participation Process for Planning Projects. *Planning Theory*, *8*(3), 242–262. https://doi.org/10.1177/1473095209104824

Corburn, J. (2003). Bringing Local Knowledge into Environmental Decision Making - Improving urban planning for communities at risk. *Journal of Planning Education and Research*, *22*(3), 420–433. https://doi.org/10.1177/0739456X03253694

Díez, J., Gullón, P., Vázquez, M. S., Álvarez, B., Martín, M. del P., Urtasun, M., … Franco, M. (2018). A community-driven approach to generate urban policy recommendations for obesity prevention. *International Journal of Environmental Research and Public Health*, *15*(4), 1–15. https://doi.org/10.3390/ijerph15040635

eMarketer. (2018). Number of Twitter users in the United Kingdom (UK) from 2012 to 2018 (in million users). Retrieved February 10, 2019, from https://www.statista.com/statistics/271350/twitter-users-in-the-united-kingdom-uk/

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). Portland, Oregon: AAAI Press. Retrieved from www.aaai.org

European Commission. (n.d.). Long-term residents. Retrieved February 25, 2019, from https://ec.europa.eu/home-affairs/what-we-do/policies/legal-migration/long-term-residents_en

FAO. (2004). What is local knowledge? Retrieved October 26, 2018, from http://www.fao.org/docrep/007/y5610e/y5610e01.htm

Geertz, C. (1983). *Local knowledge*. New York: Basic Books.

Giffinger, R., & Gudrun, H. (2010). Smart cities ranking: an effective instrument for the positioning of the cities? *ACE: Architecture, City and Environment*, *4*(12), 7–26. https://doi.org/10.5821/ace.v4i12.2483

Girardin, F., Calabrese, F., Fiore, F. D., Ratti, C., & Blat, J. (2008). Digital Footprinting: Uncovering Tourists with User-Generated Content. *Pervasive Computing*, *7*(4), 36–43. Retrieved from https://dspace.mit.edu/handle/1721.1/52693

GLA. (2012). Resources of Global City Comparison Indicators - London Datastore. Retrieved February 25, 2019, from https://data.london.gov.uk/dataset/resources-of-global-city-comparison-indicators

GLA. (2015). Daytime Population, Borough - London Datastore. Retrieved February 25, 2019, from https://data.london.gov.uk/dataset/daytime-population-borough

GLA. (2017). GLA Population and Household Projections - London Datastore. Retrieved February 25, 2019, from https://data.london.gov.uk/dataset/projections/

GLA. (2018). Migration indicators - London Datastore. Retrieved February 25, 2019, from https://data.london.gov.uk/dataset/migration-indicators

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Grace, R., Kropczynski, J., Pezanowski, S., Halse, S., Umar, P., & Tapia, A. (2017). Social Triangulation: A new method to identify local citizens using social media and their local information curation behaviors. *Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management*, (May 2017), 902–915.

Gschwend, T., Shugart, M. S., & Zittel, T. (2009). Assigning Committee Seats in Mixed-Member Systems - How Important is "Localness" compared to the Mode of Election?, 1–34.

Habibzadeh, H., Soyata, T., Kantarci, B., Boukerche, A., & Kaptan, C. (2018). A Survey of the Sensing, Communication, and Security Planes in Smart City System Design. *Computer Networks*. https://doi.org/10.1016/j.comnet.2018.08.001

Harris Interactive. (2015). Awareness of Twitter in the United Kingdom (UK) from March 2013 to April 2015. Retrieved February 10, 2019, from https://www.statista.com/statistics/303617/awareness-of-twitter-in-the-uk/

Hecht, B., & Gergle, D. (2010). On the " Localness " of User-Generated Content. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10* (p. 229). New York, New York, USA. https://doi.org/10.1145/1718918.1718962

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (p. 237). New York, New York, USA: ACM Press. https://doi.org/10.1145/1978942.1978976

Huang, C., & Wang, D. (2016). Exploiting spatial-temporal-social constraints for localness inference using online social media. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, 287–294. https://doi.org/10.1109/ASONAM.2016.7752247

Huang, C., Wang, D., & Tao, J. (2017). An Unsupervised Approach to Inferring the Localness of People Using Incomplete Geotemporal Online Check-In Data. *ACM Transactions on Intelligent Systems and Technology*, *8*(6), 1–18. https://doi.org/10.1145/3022471

Huang, C., Wang, D., & Zhu, S. (2017). Towards Diversified Local Users Identification Using Location Based Social Networks. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM '17*, 115–118. https://doi.org/10.1145/3110025.3110159

Johnson, I. L., Sengupta, S., Schöning, J., & Hecht, B. (2016). The Geography and Importance of Localness in Geotagged Social Media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 515–526. https://doi.org/10.1145/2858036.2858122

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/J.BUSHOR.2009.09.003

Karamshuk, D., Boldrini, C., Conti, M., & Passarella, A. (2011). Human mobility models for

opportunistic networks. *IEEE Communications Magazine*, *49*(12), 157–165. https://doi.org/10.1109/MCOM.2011.6094021

Kariryaa, A., Johnson, I., Schöning, J., & Hecht, B. (2018). Defining and Predicting the Localness of Volunteered Geographic Information using Ground Truth Data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (pp. 1–12). New York, New York, USA: ACM Press. https://doi.org/10.1145/3173574.3173839

Kim, S. (2016). The workings of collaborative governance: Evaluating collaborative community-building initiatives in Korea. *Urban Studies*, *53*(16), 3547–3565. https://doi.org/10.1177/0042098015613235

King, R. (2012). Geography and Migration Studies: Retrospect and Prospect. *Population, Space and Place*, *18*(2), 134–153. https://doi.org/10.1002/psp.685

Konsti-Laakso, S. (2017). Stolen snow shovels and good ideas: The search for and generation of local knowledge in the social media community. *Government Information Quarterly*, *34*(1), 134–139. https://doi.org/10.1016/J.GIQ.2016.10.002

Kumar, V., Bakhshi, S., Kennedy, L., & Shamma, D. A. (2017). Adaptive city characteristics: How location familiarity changes what is regionally descriptive. *UMAP 2017 - Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 131–139. https://doi.org/10.1145/3079628.3079665

Kumar, V., Bakhshi, S., Kennedy, L., & Shamma, D. A. (2017). Modeling Characteristics of Location from User Photos. In *Proceedings of the 2017 ACM Workshop on Theory-Informed User Modeling for Tailoring and Personalizing Interfaces - HUMANIZE '17* (pp. 1–6). New York, New York, USA: ACM Press. https://doi.org/10.1145/3039677.3039683

Lau, G., & McKercher, B. (2006). Understanding Tourist Movement Patterns in a Destination: A GIS Approach. *Tourism and Hospitality Research*, *7*(1), 39–49. https://doi.org/10.1057/palgrave.thr.6050027

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *Cartography and Geographic Information Science*, *40*(2), 61–77. https://doi.org/10.1080/15230406.2013.777139

Lindblom, C. E., & Cohen, D. K. (1979). *Usable knowledge: Social science and social problem solving*. Yale University Press.

Meriam Webster. (2019). Localness. Retrieved February 10, 2019, from https://www.merriam-webster.com/dictionary/localness

Mohanty, S. P., Choppali, U., & Kougianos, E. (2016). Everything you wanted to know about smart cities: The Internet of things is the backbone. *IEEE Consumer Electronics Magazine*, *5*(3), 60–70. https://doi.org/10.1109/MCE.2016.2556879

Montanari, A. (2005). Human mobility, global change and local development. *Belgeo*, (1–2), 7–18. https://doi.org/10.4000/belgeo.12391

Nam, T., & Pardo, T. A. (2011). Conceptualizing smart city with dimensions of technology, people, and institutions. *Proceedings of the 12th Annual International Digital Government Research Conference on Digital Government Innovation in Challenging Times - Dg.o '11*, 282. https://doi.org/10.1145/2037556.2037602

Novy, J. (2018). 'Destination' Berlin revisited. From (new) tourism towards a pentagon of mobility and place consumption. *Tourism Geographies*, *20*(3), 418–442. https://doi.org/10.1080/14616688.2017.1357142

Ofcom. (2018). How often do you visit social networking sites? Retrieved February 10, 2019, from https://www.statista.com/statistics/271919/frequency-of-social-networking-in-the-uk/

ONS. (2017a). Population by Country of Birth - London Datastore. Retrieved February 25, 2019, from https://data.london.gov.uk/dataset/country-of-birth

ONS. (2017b). Population by Nationality - London Datastore. Retrieved February 25, 2019, from https://data.london.gov.uk/dataset/nationality

Ostermann, F. O., Huang, H., Andrienko, G., Andrienko, N., Capineri, C., Farkas, K., & Purves, R. S. (2015). Extracting and Comparing Places Using Geo-Social Media. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *II-3/W5*, 311–316. https://doi.org/10.5194/isprsannals-II-3-W5-311-2015

Persky, J., & Wiewel, W. (1994). *The Growing Localness of the Global City Author*. *Source: Economic Geography* (Vol. 70). Retrieved from https://www.jstor.org/stable/pdf/143651.pdf?refreqid=excelsior%3Ad54496f21db902aae6d9385f2d03482d

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. Retrieved from http://dspace.flinders.edu.au/dspace/%0Ahttp://www.bioinfo.in/journalcontent.php?vol_id=115&id=%0Ahttp://www.bioinfo.in/contents.php?id=51

Sassen, S. (2001). *The global city : New York, London, Tokyo*. Princeton University Press.

Schmitt, E., Dominique, B., & Six, J. (2018). Assessing the degree of localness of food value chains. *Agroecology and Sustainable Food Systems*, *42*(5), 573–598. https://doi.org/10.1080/21683565.2017.1365800

Scott, D. N. (2015). 'We Are the Monitors Now': Experiential Knowledge, Transcorporeality and Environmental Justice. *Social and Legal Studies*, *25*(3), 261–287. https://doi.org/10.1177/0964663915601166

Sen, S. W., Ford, H., Musicant, D. R., Graham, M., Keyes, O. S. B., & Hecht, B. (2015). Barriers to the Localness of Volunteered Geographic Information. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 197–206. https://doi.org/10.1145/2702123.2702170

Silva, B. N., Khan, M., & Han, K. (2018). Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities. *Sustainable Cities and Society*, *38*, 697–713. https://doi.org/10.1016/J.SCS.2018.01.053

Sun, Y., Fan, H., Helbich, M., & Zipf, A. (2013). Analyzing Human Activities Through Volunteered Geographic Information: Using Flickr to Analyze Spatial and Temporal Pattern of Tourist Accommodation (pp. 57–69). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34203-5_4

Swoboda, B., Pennemann, K., & Taube, M. (2012). The Effects of Perceived Brand Globalness and Perceived Brand Localness in China: Empirical Evidence on Western, Asian, and Domestic Retailers. *Journal of International Marketing*, *20*(4), 72–95. https://doi.org/10.1509/jim.12.0105

Tahara, T., & Ma, Q. (2014). Searching for local Twitter users. *Lecture Notes in Computer Science*, 89–96.

Tu, Z., Su, Z., & Devanbu, P. (2014). On the Localness of Software. https://doi.org/10.1145/2635868.2635875

Twitter. (2018). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 4th quarter 2018 (in millions).

We are Flint. (2018). Distribution of Twitter users in the United Kingdom (UK) in January 2018, by frequency of use. Retrieved February 10, 2019, from https://www.statista.com/statistics/611306/frequency-of-twitter-use-in-the-united-kingdom-uk/

We Are Social. (2018). Total number and the share of population of active social media and mobile social media users in the United Kingdom (UK) in January 2018.

Williams, A., Foord, J., & Mooney, J. (2012). Human mobility in functional urban regions: understanding the diversity of mobilities, *6701*(March 2015), 37–41. https://doi.org/10.1080/03906701.2012.696961

Williams, A. M., & Hall, C. M. (2000). Tourism Geographies Tourism and migration: New relationships between production and consumption. *Tourism Geographies*, *2*(1), 5–27. https://doi.org/10.1080/146166800363420

Yang, Y., & Diez-Roux, A. V. (2012). Walking Distance by Trip Purpose and Population Subgroups. *American Journal of Preventive Medicine*, *43*(1), 11–19. https://doi.org/10.1016/J.AMEPRE.2012.03.015

Yuill, R. S. (1971). *The Standard Deviational Ellipse; An Updated Tool for Spatial Description*. *Source: Geografiska Annaler. Series B, Human Geography* (Vol. 53). Retrieved from https://www.jstor.org/stable/pdf/490885.pdf

Zhang, S., & Feick, R. (2016). Understanding Public Opinions from Geosocial Media. *ISPRS International Journal of Geo-Information*, *5*(6), 74. https://doi.org/10.3390/ijgi5060074

APPENDIX

LIST OF TROUTIST ATTRACTION LOCATIONS IN LONDON

Tourist attraction	Longitude	Latitude
British Museum	-0.126946	51.51956
National Gallery	-0.12831	51.50911
Tate Modern	-0.099668	51.509
Natural History Museum	-0.176443	51.49691
Southbank Centre	-0.116259	51.50621
Somerset House	-0.117148	51.51123
Science Museum	-0.174469	51.49805
Victoria and Albert Museum	-0.17218	51.49678
Royal Museums Greenwich	-0.005246	51.48103
National Portrait Gallery	-0.128165	51.5096
BODY WORLDS London	-0.133866	51.51042
Buckingham Palace	-0.141869	51.50153
London Eye	-0.119532	51.50352
Tower of London	-0.075918	51.5083
Borough Market	-0.09105	51.50553
Madame Tussauds	-0.155283	51.52295
The View from The Shard	-0.086533	51.50446
SEA LIFE London	-0.119483	51.50161
Big Ben	-0.124545	51.50131
Palace of Westminster	-0.124788	51.49952