TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

DIPLOMARBEIT

Modeling Individual's Familiarity of Places Using Social Media

Ausgeführt am Institut für

Geoinformation und Kartographie

der Technischen Universität Wien

unter der Anleitung des Betreuers

Univ.-Prof. Mag. rer. nat. Dr. rer. nat. Georg Gartner

und unter Mitwirkung von

Dr. Haosheng Huang

durch

Wangshu Wang

Matrikelnummer 1129622

Zwerchäckerweg 75, 1220, Wien

_____ _____

Ort, Datum Unterschrift

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

MASTER THESIS

Modeling Individual's Familiarity of Places Using Social Media

Institute of Geoinformation and Cartography

Vienna University of Technology

Supervisors:

Univ.-Prof. Mag. rer. nat. Dr. rer. nat. Georg Gartner

and

Dr. Haosheng Huang

Submitted by

Wangshu Wang

Student number 1129622

Zwerchäckerweg 75, 1220, Wien

_____ _____

Place, Date Signature

DECLARATION

This thesis is a presentation of my original research work. I certify that this thesis contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

_____ _____

Place, Date Signature

ACKNOWLEDGEMENTS

I would like to express my gratitude to Prof. Georg Gartner and Dr. Haosheng Huang for their constant support, instructions, guidance and encouragement during the research.

I am very grateful to Dipl.-Ing. Stefan Peters, the coordinator of this Master's program, for his very kind help during this 2 years master study.

Last but not least, a special thank goes to my parents, for their unconditional support all the time.

ABSTRACT

Different people have different prior knowledge of a certain environment, which makes it hard to provide a geo-related service (e.g. navigation) that fits all. In order to provide a service adapted to individual's spatial knowledge, a model of individual's prior knowledge of the environment is needed. The growing popularity of location aware social media provides a unique opportunity to study individual's spatial knowledge. With Foursquare being one of the most popular location-based social media, this thesis focuses on modeling individual's familiarity of places by using Foursquare data. To achieve this overall goal, two objectives are set.

The first objective is to identify individual's meaningful places. To derive meaningful places from the check-ins on Foursquare, which represent the frequency of visits to a place, an appropriate clustering algorithm is required. A comparison of four existing clustering algorithms (SLINK, K-means, DBSCAN and EM Algorithm for Gaussian Mixture Model) was then conducted. DBSCAN turned out to have the best overall performance.

To attain the second objective, which is to model individual's familiarity of places, information indicating the affected responses of an individual, e.g., the text descriptions and photos along with a check-in, was added to weight each check-in and eventually to measure the familiarity. To evaluate the modeling framework, an online survey was carried out. The results demonstrated the possibility to model individual's familiarity of places using social media.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF FORMULAS

Chapter 1: Introduction

1.1 Background and Motivation

As human understanding of the environment begins with places (Shemyakin, 1962), deriving meaningful places (i.e., places that are associated with certain activities and meanings, see Sec. 2.2) and individual's familiarity (i.e., how familiar are these places to an individual, see Sec. 2.3) to them from different source of geographic data is always an interesting topic to researchers. The blooming of social media in recent years brings us to a new information era. The advancement of mobile devices and location technologies, makes it possible for people to have ubiquitous access to geographic information nearly at anytime in anywhere (Bhattacharya, 2009). Thus, location-based social media are quickly growing in popularity, which provide a new source of tremendous geographic data and a unique opportunity to study individual's knowledge of places.

Many researches have been done on deriving meaningful places and modeling people's prior spatial knowledge using different sources of geographic data, such as GPS trajectories (Zhou *et al.* 2007, Nurmi 2009). However, people's familiarities with these places are not mentioned. There were also lots of studies on extracting and interpreting geographic information using social media (Keßler *et al.* 2009, Noulas *et al.* 2011), but nearly no research has been found to focus on individual's meaningful places using social media. Therefore, a study of the possibility and method to model individual's familiarity of places is needed.

1.2 Research Objectives

This thesis aims to develop an approach to model individual's familiarity of places by using social media data. The novelty of this research is that it is the first one to identify individual's meaningful places using social media data and consider that person's familiarity with these places. With Foursquare[1] being one of the most popular location-based social media, it serves as the data source in this thesis.

In order to find out a possible approach to model individual's familiarity of places, two main objectives should be achieved.

- The first objective is to identify individual's meaningful places.
- The second objective is to model individual's familiarity of these places.

1.3 Organization of the Thesis

This thesis is divided into six main chapters. The first chapter introduces the topic of this thesis and shows an overview of the work. The second chapter describes human spatial cognition and specifies the meaning of keywords in this thesis, including "meaningful place" and "individual's familiarity of places". A literature review on discovering meaningful places using other sources of geographic data is also in the same chapter. The state of the art of volunteered geographic information and place Identification using social media is in Chapter 3. The forth chapter consists of an introduction and an experimental comparison of four place identification algorithms. The workflow of building a familiarity model of a user is presented in Chapter 5, with

[1] https://foursquare.com/

a second experiment to evaluate the model. The last chapter concludes the whole thesis and provides the outlook for future studies.

Chapter 2: Spatial Cognition and Discovering Meaningful Places

2.1 Spatial Cognition

Spatial cognition is defined as "the knowledge and internal or cognitive representation of the structure, entities and relations of space; in other words, the internalized reflection and reconstruction of space and thought" (Hart and Moore, 1973, p. 248). Thus, spatial cognition is considered as spatial knowledge and the affective responses aroused by the given environment. The affective responses include feeling, attitude and other emotional characteristics (Hart and Conn, 1991).

For the acquisition of spatial knowledge, Piaget (1967) states that the representation of space comes up from the "coordination" and "internalization" of actions. This means it is the interaction in space that serves as the crucial part for the acquisition of spatial knowledge. From a geographic perspective, human beings begin to build their spatial knowledge when they are first exposed to an area. A preoperational level of spatial comprehension is reverted, at which their cognitive representations are mainly the fundamental topological properties of space. With the increasing interaction with the environment, the progression proceeds to projective relations, which can be expressed as perspectives or points of view. Further, the concrete operational stage is reached, in which coordinate structures are comprehended. At the formal operational level, which is the final one, an individual's spatial knowledge can be better represented by general and mixed metrics (Golledge and Stimson, 1997).

Though there are several theories of the development of spatial knowledge, the basic components of spatial knowledge can be integrated into three parts. The first one is a "declarative" component, including knowledge of "objects and/or places together with meanings and significances attached to them". It is called as "landmark" or "cue" knowledge in many theories and regarded as the minimum requirement for object and pattern recognition and discrimination (Shemyakin 1962, Golledge and Stimson 1997). The next one is known as "relational" or "route" component, which includes spatial relationships among objects or places. Concepts like "proximity" and "sequence", which leads to the development of hierarchical networks and knowledge chunking, are developed here. The third one is called "procedural knowledge". This knowledge is made up of sets of procedural rules (normally expressed in "If… then…" statements) and is needed to develop "object and procedure association" for processes such as the development of locomotive ability (Golledge and Stimson 1997, p. 163).

When asking about how human spatial knowledge is developed, many researchers hold the same view that it is a process from "landmark" to "route" to "survey knowledge". Shemyakin (1962) first argues that spatial knowledge acquisition is a consecutive process from landmark recognition to the paths linking the landmarks, and at last to the overall understanding of relational characteristics of areas.

Siegel and White (1975) succeeded Shemyakin's theory and further elaborated it into a model with several stages. The first stage is landmark recognition. Then routes develop between the landmarks, with the properties of route knowledge processes from topological to metric. Later on, on the foundation of metric relationships, sets of landmarks and routes are coordinated into clusters. Finally, based on the metric

properties within one cluster and the ones across clusters, a general “coordinated frame of reference” is developed, which produces survey knowledge.

Previous theories suggest that at early stages of spatial knowledge acquisition, there is a lack of metric information. While Montello (1998) claims that “There is no stage in which only pure landmark or pure route knowledge exists: Metric configurational knowledge begins to be acquired on the first exposure” (p. 146). Alternatively, linguistic or pre-linguistic sets are used to store the nonmetric information and communicating spatial knowledge as attributes along with metric information.

Together with the nonmetric information of spatial knowledge, the affective responses are also able to be represented by language. Therefore, study the semantics can help reverting people’s spatial knowledge. Ulrich (1983) suggests several hierarchical levels of affective responses. The broad categories, e.g., liking and disliking, are placed at the first level and detailed ones are generated as more interaction with the space occur. He also implies that once entered the hierarchy in one general type, it tends to go further to the detailed feelings of that type rather than to cross over to another one.

2.2 The Notion of Place

What is a “place”? When looking it up in a dictionary, the first item refers to “an area with definite or indefinite boundaries; a portion of space”. When asking people about it, the answers vary.

In human geography, Tuan (1977) defines place as “a pause in movement” (p. 138). He further explains that for certain biological needs, human beings pause at a locality, which is then become “a center of felt value” (p. 138) by the pause. Krämer (1995)

consider it as an entity with experience, whereas the explanation of place from Relph (1976) is a physical setting with its supported activities and the meanings attributed to it.

In environmental psychology, Aitken *et al.* (1989) interpret place as a location integrated of society, culture and nature. Hart and Conn (1991) claim that place is seen as the focus of human intentions with meaning and action.

In computer science, the definition of a place given by Kang *et al.* (2005) is “a locale that is important to a user and which carries a particular semantic meaning” (p. 58). Nurmi (2009) consider place as “a spatial area that are linked with activities and associated with meanings” (p. iii).

Though the definitions of place in different fields are not all the same, there is one point they all emphasized: meaning. What makes a place meaningful then? In reality, people think of places with their meanings, like “home”, “hospital” and “university”. A same place may have different meanings to different people. For example, I associate the meaning “shopping” to a shop, while the employees there give another meaning ”working” in addition to simply “shopping”. It is the semantic memory, that is, the memory preserves only the general significance of remembered experience, that defines the meaning of a place. Hence, we can say, it is the experience associated with the place makes it meaningful. Studies in environmental psychology (Aitken 1989, Golledge and Stimson 1997) also point the meaning of a place to the activities and experience that people have with the place. Thus, in this thesis, “meaningful place" refer to a place that is associated with certain activities and meanings, while “individual’s meaningful place" corresponds to a place that is

associated with certain activities and meanings by that person, which may not be the same as the conventional meaning.

2.3 Individual's Familiarity of Places

Individual's familiarity of a place refers to how familiar is a place to an individual. Tuan (1977) suggests that the familiarity of a place is affected by the intensity and extensity of the activities there. As been put forward by Mehrabian and Russell (1974), there are strong psychological and emotional links, which depend on the range of experience, between people and places. Thus, the familiarity of a place can be inferred as the extensity of experience there, e.g., the duration of stay, the intensity which can be represented by the affected responses aroused by that place, the kind of activity as well as the frequency of visits. Apart from the direct experience in a place, the indirect sources, e.g., textual descriptions, photos, videos and maps also influence individual's familiarity of a place (Appleyard, 1970). In general, the influencing factors of an individual's familiarity to a place includes the experiences and affected responses to the place of both direct and indirect contact, the kind of activity, duration of stay in that place and frequency of visits to it.

2.4 Discovering Individual's Meaningful Places

With the advanced location technology, like Global Positioning System (GPS) and Global System for Mobile Communications (GSM) positioning, place identification becomes an eye-catching topic. Marmasse and Schmandt (2000) used loss of GPS signals to detect buildings in their work on the “comMotion" system. A place is identified when GPS signal disappears and then reappears within a certain radius. An improved approach by Ashbrook and Starner (2002) used a variant of the k–means clustering algorithm to determine places. However, these early works on place

identification only takes physical location into consideration. The term “place” here is more likely to refer to location without specific meanings.

To bridge physical location that computers work with and places that people talk about, several researches has been addressed to discover individual’s meaningful places. Zhou *et al.* (2007) did an experiment to discover individual’s meaningful places using GPS trajectories. They carried out their experiment by asking the subjects to carry a GPS-enabled mobile device to record the GPS trajectories during their daily life. At the same time, a diary was also required from the subjects to log their meaningful places, which would be the ground truth to evaluate the algorithm’s performance later. After a temporal preprocessing which eliminates GPS readings with speed greater than 0 and the ones within a small distance of the previous reading, the DJ-Cluster algorithm (Zhou, 2004) was performed on the GPS data for place identification. This research proposed a general experimental framework on discovering individual’s meaningful places and its result indicated that clustering algorithm could be applied to discover places that are meaningful to users.

A thorough analysis on place identification was done by Nurmi (2009) in his PhD thesis. He introduced different location systems with their characteristics and a general process of place identification, mainly focus on GPS data. A detailed evaluation and comparison of existing place identification algorithm was also conducted on twelve GPS datasets. He also pointed out that people’s information needs of a place depend on their existing knowledge of the place. However, after discovering if a place was meaningful to an individual or not, he didn’t go further to the familiarity of that place.

In summary, though there are some studies on identifying individual's meaningful places, no research on individual's familiarity of places has been done yet. However, people's familiarity of places can be very useful in providing geo-related services, such as navigation and tourist guide. Proven by scholars, a more familiar stimuli needs less cognitive work than the less familiar ones (Zajonc 1968, Winkielman *et al.* 2003). It is then more comfortable to use a service provided according to user's previous spatial knowledge. In addition, such services based on user's familiarity of places help them to gain new spatial knowledge easier. Therefore, a model of individual's familiarity of places is needed, in which the influencing factors of an individual's familiarity to a place should be concerned.

Chapter 3: From Social Media to Meaningful Places

The past decade has witnessed the outbreak of social media. Popular platforms like Facebook[2], Flickr[3], Foursquare and Twitter[4] allow users to post short messages, photos, links, and videos in a highly connected social environment (Naaman, 2011). On these platforms, geographic information is shared either explicitly as check-in or implicitly as location coordinates along with user posts. With the huge amount of geographic data easily accessible, place identification from social media data becomes a trending topic (Chen et al. 2009, Keßler *et al.* 2009).

3.1 Volunteered Geographic Information

The term "Volunteered Geographic Information" (VGI) is first brought up by Goodchild (2007) as "using the Web to create, assemble, and disseminate geographic information provided voluntarily by individuals" (p. 1). In the same paper, he gave some examples like Wikimapia[5], Flickr, OpenStreetMap[6] and so on. According to this definition, the geographic information in social media like Foursquare and Twitter can also be filed into this category.

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

[3] http://www.flickr.com/

[4] https://twitter.com/

[5] http://wikimapia.org/

[6] http://www.openstreetmap.org/

Most social media platforms use the enabled GPS receiver in mobile devices to get the current geographic location and then code and store the location as geotags. In some social media platforms, e.g., Flickr, geotags can also be added by user manually choosing the geographic location of the photo. Inside the geotags, geographic locations are often represented as latitude and longitude pairs.

3.2 Foursquare

Foursquare was launched by Dennis Crowley and Naveen Selvadurai in Austin, Texas in March 2009. Its web site, API, and batch processing are almost all written in Scala. For the search geo-indexing, Google's s2 library[7] is used to store cellids within its search index. PostGIS[8] and geonames.org[9] dataset are used for reverse geo-coding (Foursquare, 2013a).

The users of foursquare can be divided into business users and customer ones. From the business side, foursquare encourage companies and shops to set up their own venues and make promotions. While from the customer side, Foursquare allows its users to check in at different venues when they are there and pushes a check-in notification to their friends at the same time. A check-in is a footprint of the user's current location. Along with a check-in, a text message called "shout" and photos can also be sent (Figure 1). Concerning privacy, users can always set the check-ins to private so that their friends are not able to see the private check-ins, but in the user's history, they are also counted. Users can interact with their friends as well as the

[7] http://code.google.com/p/s2-geometry-library/

[8] http://postgis.net/

[9] http://www.geonames.org/

venues, e.g., their friends can like their check-ins or comment on it and they can get promotions from the venues. Except friends and venues, users can also play with the system itself. The system provides personalized recommendations based on the users' location and their friends' check-ins. It also gives virtual rewards in the forms of "points", "badges", and "mayorships" when a user checks in and reaches certain level. Points are calculated based on the importance of the check-in, while particular badges are given for particular reasons, for example if a user checked in at the same place three times a week, a "Local" badge is awarded. Mayorships are hold by the user having the most check-ins in a venue in the past 60 days. Other features like place lists and tips are also available in foursquare.

Figure 1. An example of a check-in containing a shout and a photo

As of September, 2013, foursquare claims to have over 40 million customer users and over 4.5 billion check-ins, with millions more every day (Foursquare, 2013a). Foursquare API makes it possible for researchers to get access to this considerable

geographic information to study urban space and user's spatial behavior. To gain a better knowledge of urban space, venues of different categories have been investigated by the check-ins there at different time (Komninos, 2013). For understanding user's spatial behavior, geo-related API endpoints which are associated to a certain user could be useful. Table 1 presents the related API endpoints with their request parameters and response fields, for detailed explanation of all the parameters and response fields, please refer to foursquare documentation (Foursquare, 2013b).

Table 1. Useful Foursquare API endpoints for understanding user's spatial behavior

Foursquare API endpoints	**Parameters**	**Response fields**
users/USER_ID	USER_ID (ID of the user to get details for)	User (user's profile information)
users/USER_ID/checkins	USER_ID (only self is supported for now), limit, offset, sort, afterTimestamp and beforeTimestamp	Check-ins (A count and items of chec k-ins)
users/USER_ID/mayorships	USER_ID (ID of the user to get mayorships for)	Mayorships (A count and items of objects which currently only contain compact venue object)
users/USER_ID/photos	USER_ID (only self is supported for now), limit and offset	Photos

users/USER_ID/tips	USER_ID (ID of the user to get tips from), sort, ll(Latitude and longitude of the user's location), limit and offset	Tips (A count and items of tips)
users/USER_ID/todos	USER_ID (ID of the user to get todos for) , sort and ll (Latitude and longitude of the user's location)	Todos (A count and items of todos)
users/USER_ID/venuehistory	USER_ID (only self is supported for now), afterTimestamp, beforeTimestamp and categoryId (Limits returned venues to those in this category)	Venues (A count and items of objects containing a beenHere count and compact venues)
checkins/CHECKIN_ID	CHECKIN_ID, signature (a signature appended when check-ins are sent to public feeds such as Twitter)	Check-in (A complete checkin object)

As Foursquare is mainly check-in based, most information that reveals a user's geographic location and affected response is within the check-ins. Table 2 shows information available from a single check-in.

Table 2. Parameters and description of a check-in

Parameters of a check-in	Description
id	The identity of this check-in
createdAt	The create time of this check-in
type	The object type in Foursquare (check-in)
shout	A text message created by the owner along with this check-in
timeZoneOffset	The difference, in minutes, between UTC and local time
user	The owner of this check-in, with his ID and some information
venue	Information of the venue been checked in, including ID, name, location, categories and some other basic information.
likes	All the users who like the checked in venue
like	True or false. The owner's likeness status of this check-in
photos	Photos information and count
posts	Posts information and count
comments	Comments information and count
source	The device created this check-in

To make use of Foursquare API, apps should connect with it via OAuth 2.0[10]. First, developers are required to register their app to obtain its foursquare API credential. If

[10] http://oauth.net/2/

the app needs connecting with Foursquare users, i.e., make requests to Foursquare on the behalf of a user, an access token is required. Different ways to gain access tokens on different platforms are available in (Foursquare, 2013c). Once an access token is obtained, developers can add the access token to the GET or POST request to use the API endpoints.

3.3 Place Identification Using Social Media

Previously, GPS coordinates and data from GSM cell identifiers served as the main data source for understanding urban space and place identification, which has already been addressed in Chapter 2. With the growing of location-based social networks and media, this new source of location data caught researchers' eyes.

A lot of researches focused on place semantics. Rattenbury *et al.* (2009) investigated the feasibility of automatically determining place semantics from distributed individual tags, where Flickr tags were applied. Hollenstein and Purves (2010) used tags of Flicker images to study the descriptions of city center and people's understanding of regions.

Another interesting topic was to identify city landmarks and neighborhood boundaries based on the enormous georeferenced data. Chen *et al.* (2009) leveraged geo-tagged photos to generate landmarks of an area. They identified landmarks by clustering a set of geotagged photos so that the photos in each cluster are geographically close and also have similar tags. An iterative version of the k-means clustering method was applied to achieve this. They started with all photos in a single cluster. By measuring the spread of the cluster represented by the geographic distance of the photos in it and the statistical tag variance measured by the Term Frequency-Inverse Document Frequency (TFIDF) of the tags at each iteration, a new

cluster center was added into the cluster with the largest spread. Concerning social behaviors of different photographers might bias the result severely, they added user frequency that takes the number of unique photographers in the clusters into account.

Keßler *et al.* (2009) proposed a bottom-up approach to build gazetteer based on geotagged photos. They first used a crawling algorithm to retrieve geotagged photos with certain annotation. Then, a Delaunay triangulation was performed within point clouds to find clusters. A threshold was given to split the formed graph of points and edges into clusters in the end. Their result demonstrated the possibility to derive meaningful places from geotagged photos. Similarly, Flickr (Flickr, 2008) itself also did some very nice practices on deriving neighborhood boundaries from geotagged photos. Instead of Delaunay triangulation, the algorithm they used was Alpha Shape (Edelsbrunner *et al.*, 1983).

Human activities and city dynamics was also an area that researchers concentrated on. Noulas and his colleagues (2011) used a Foursquare dataset collected from Twitter messages containing Foursquare check-ins to model human activities and geographical areas. They chose London and New York as the experimental area and divided the two cities into small areas according to their social importance reflected by nearby places and attached social activities in Foursquare. An interesting pattern of the city neighborhoods and human behavior was generated after performing clustering algorithm on the dataset. Similarly, Cranshaw *et al.* (2012) clustered such data into areas of particular social activity in cities and discovered that check-in data could reveal subtle changes in the local social patterns.

To recap, Social media has been utilized in large-scale place identification and group behavior understanding, however, individual's meaningful places was neglect. With location-based social media (e.g., Foursquare) logging an individual's traces, it is a powerful data source to identify individual's meaningful places non-intrusively. Moreover, according to the investigation of Foursquare in Section 3.2, the influencing factors of an individual's familiarity to a place (e.g., the affected responses, the kind of activity, and frequency of visits) can be represented by Foursquare user's history and some parameters in a check-in. Therefore, it is possible to build a model of individual's familiarity of places using Foursquare.

Chapter 4: Identification of Individual's Meaningful Places

Foursquare is getting popular with its unique features. To a certain place, the number of check-ins there describes user's frequency of visits to that place. Furthermore, user's check-in history on Foursquare records his trajectory. Therefore, it is leveraged as the data source to identify individual's meaningful places in this thesis. With the raw check-ins, an appropriate clustering algorithm is needed to group them in order to discover individual's meaningful places. Seeing the fact that there are many clustering algorithms already, a comparison of existing algorithms will be conducted.

In this chapter, the clustering algorithms will be introduced first. Then the experiment carried out to compare them will be presented. This section will end with a discussion of the algorithms' performances.

4.1 Clustering Algorithms

The existing algorithms for clustering can be classified into hierarchical clustering methods, partitioning clustering methods, density-based clustering methods and model-based clustering methods.

4.1.1 Hierarchical Clustering Methods

The main idea behind hierarchical clustering is that nearby objects are likely to be more similar than the objects far away. The output of this method is a dendrogram with nested clusters of various sizes represented by its nodes and data points represented by its leaves (Heller and Ghahramani, 2005). There are two types of hierarchical clustering methods: agglomerative and divisive. Agglomerative method is a "bottom up" approach which merges pairs of clusters as the hierarchy moves up while the divisive method is a "top down" approach which begins from one cluster

and splits it as the hierarchy moves down. There is no input parameters needed for hierarchical clustering method, but a termination condition is usually needed to stop the merging or division.

The most commonly used hierarchical clustering method is Single-Link, which suffers from a chaining effect, i.e., if there is a chain of points between two actual clusters then the two clusters may be considered as one (Sibson,1973). Due to this effect, there have been lots of studies on avoiding this problem and improving the performance of the algorithm (Rokach and Maimon, 2005.).

4.1.2 Partitioning Clustering Methods

From an initial partitioning, partitioning clustering methods iteratively relocate data points between clusters until an optimal partition is attained. Usually, it is a locally optimal solution. An exhaustive enumeration process can help to achieve the global optimality. However, this is not feasible, so certain greedy heuristics are used for the iterative optimization (Äyrämö and Kärkkäinen, 2006). The most commonly used algorithm of this kind is K-means clustering algorithm.

K-means

As one of the simplest and most famous clustering algorithm, K-means algorithm classifies a given data set to fixed k clusters. It starts with k centroids chosen randomly. Then, the distance between each point of the given dataset to the centroids is calculated and each point is assigned to the nearest cluster. When the first step is done, the mean value of each cluster need to be calculated and used as the new centroid. Then, we repeat the previous step with the newly generated centroids until no more changes of the centroids could be done. This iterative approach aims to minimize a squared error function (MacQueen, 1967):

Formula 1. $E^2 = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2$

Where k is the number of clusters in the data set, S_i denotes the set of points belonging to the i_{th} cluster, μ_i is the mean value of S_i and x_j is a data point in cluster S_i.

The pseudo code of K-mean algorithm can be found below:

Input: D (data set), k (number of cluster)
Output: clusters
Set μ_i to random value
While E^2 changes
    For each x
        Calculate distance and assign to cluster with the nearest centroid
    End
    For i=1 to k
        Calculate mean (S_i)
        If mean(S_i) != μ_i
            Replace the centroid
    End
End

Despite the popularity of K-means algorithm, there are some significant drawbacks of it. First, it needs a specified cluster number "K", which is hard for most of tasks to know in advance. Besides, it is sensitive to the initial configuration. Different configurations of initial centroids will result in different clusters. As it uses the mean values to be the new centroids, it is also very sensitive to noise. K-means algorithm tends to find clusters in spherical shape, while performs poor in detecting clusters in other shapes (MacQueen, 1967).

4.1.3 Density-based Clustering Methods

Density based clustering algorithm tend to discover dense regions in a data space. Because the clustering criterion is based on density connectivity but not simply distance, it can discover arbitrary shape. The most widely used density based clustering algorithm is DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

DBSCAN

DBSCAN is an algorithm proposed by Martin Ester *et al.* (1996). They aimed to provide an algorithm with less input parameters, with the ability to discover clusters with arbitrary shape and efficient on large databases. The basic idea of DBSCAN is that every point in a cluster should have a neighborhood of a given radius with at least a minimum number of points in it. Based on this notion, two input parameters should be provided, that is the radius "Eps" and the minimum number of points "MinPts". The choice of the distance function between two points determines the neighborhood shape, e.g., Manhattan distance in 2D space generates rectangular neighborhoods and Euclidean distance yields circular neighborhoods.

Before introducing the algorithm, some items should be defined.

1. The neighborhood of a point p with radius Eps is defined as:

Formula 2. $$N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\}$$

   with $dist(p,q)$ represents the distance between two points p and q calculated by the defined distance function.
2. Directly density-reachable: If a point p is directly density-reachable from a point q, then they satify

1) $p \in N_{Eps}(q)$ and

2) $|N_{Eps}(q)| \geq MinPts|$.

3. Density-reachable: If there is a chain of directly density-reachable points between q and p, then q is density-reachable from p, i.e., a chain of points $p_1, \ldots, p_n$, where $p_1 = p, p_n = q$ and p_{i+1} is directly density-reachable from p_i.
4. Density-connected: If there is a point in between points p and q that from both of them it is density-reachable, then p and q are density-connected.
5. Cluster: Cluster C is a non-empty subset of database D that if one point is in a cluster, then all the density-reachable points from it are in the same cluster and all the points in a cluster are at least density-connected. The conditions are written as:

   1) $\forall p, q$, if $p \in C$ and q is density-reachable from p, then $q \in C$,

   2) $\forall p, q \in C$, p and q are density-connected.

6. Noises are those points not belong to any clusters.

The algorithm starts with a randomly chosen point. If it is a core point, DBSCAN retrieves all its density-reachable points and end up with a cluster. If it is not a core point, it will be marked as noise first and will be discovered later if it is a border point or noise as the procedure goes on to visit other points in the database.

The pseudo code of DBSCAN can be found below:

Input: D (data set), Eps and MinPts
Output: clusters
ClusterId = 1
For i = 1 to D.size do
    If P(i). mark = unclassified then

```
        If ExpandCluster(D, P(i), ClusterId, Eps, MinPts) = true then
            ClusterId = ClusterId +1
        End if
    End if
End for

ExpandCluster (D, P, ClusterId, Eps, MinPts)
seeds = D.regionQuery (P, Eps)
If seeds.size < MinPts then
    P. mark = noise
    Return false
Else
    For i = 1 to seeds.size
        seeds(i).mark = ClusterId
    End for
    seeds.delete(P)
    While seeds != empty do
        currentP = seeds.first
        neigborhood = D.regionQuery(currentP, Eps)
        If neigborhood.size >= MinPts then
            For i = 1 to neigborhood.size do
                If neighborhood(i).mark ∈ {unclassified, noise} then
                    If neighborhood(i).mark = unclassified then
                        seeds.append (neighborhood(i))
                    End if
                    neighborhood(i).mark = ClusterId
                End if
            End for
        End if
        seeds.delete(currentP)
    End while
```

Return true

End if

As one of the most commonly used algorithm, DBSCAN can identify noise, can find arbitrary shapes of clusters and is insensitive to the ordering of the points in the database. However, if the ordering of the points is changed, the membership of the border points of clusters might change, because if two clusters are very close so that some points belong to both, they will be assigned to the cluster first discovered. While using DBSCAN, it is always not easy to determine the parameters. Moreover, for data sets with large differences in densities, this does not perform very well. Algorithm like OPTICS (Ordering Points To Identify the Clustering Structure) was then brought forward to overcome this weakness of DBSCAN (Ankerst, 1999).

4.1.4 Model-based Clustering Methods

Model-based clustering methods assume that the given data set is generated by a mathematical model and attempt to optimize the fit between them. After recovering the model, we use it to define clusters. They identify not only the clusters but also the characteristic descriptions of the groups and the natural distribution of data (Rokach and Maimon).

Expectation- Maximization Algorithm for Gaussian Mixture Model

A Gaussian Mixture Model (GMM) is defined as a parametric probability density function represented as a weighted sum o f Gaussian component densities. It can be represented by the equation below:

Formula 3. $p(x|\theta) = \sum_{i=1}^{M} w_i g(x|\mu_i, \sum_i)$

Where, x is a D-dimensional continuous-valued data vector (i.e., an observation or a measurement), M denotes the number of components, w_i,i=1,…,M, is the weight of each component, which satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$ and $g(x|\mu_i, \sum_i)$, i=1,…,M, is the component Gaussian density. Each component is represented by a D-variate Gaussian function below:

Formula 4. $g(x|\mu_i, \sum_i) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)' \sum_i^{-1}(x-\mu_i)\right]$

Where, μ_i and $\sum_i$ stands for the mean and covariance respectively (Reynolds, 2009).

The Expectation-Maximization (EM) Algorithm is a general approach to find maximum likelihood parameters in estimates in problems where some variables were unobserved (Dempster *et al.*, 1977). If the data set has n multivariate observations and we assume the observations x_j(j=1,…, n) are independent and identically distributed according to the distribution f with parametersθ, then the likelihood can be written as:

Formula 5. $L(x_j|\theta) = \prod_{j=1}^{n} f(x_j|\theta)$

For an easier calculation, the log-likelihood function is more widely used, which is:

Formula 6. $L'(x_j|\theta) = \sum_{j=1}^{n} \log [f(x_j|\theta)]$

For GMM, the log-likelihood function is:

Formula 7. $L'(x_j|\theta) = \sum_{j=1}^{n} \log \sum_{i=1}^{M} w_i g(x_j|\mu_i, \sum_i)$

However, this is a non-linear function of the parameters, therefore, the direct maximization cannot be achieved. Starting from some initial model, the EM Algorithm estimates the parameters of the model iteratively between two steps, an expectation step and a maximization step.

First, in the "E" step, we estimate the parameters and compute the conditional expectation. In the following "M" step, we determine the parameters that maximize the log-likelihood from the "E" step. Then, the model with the newly determined parameters becomes the initial model and the same process starts again. Such iteration improves the true likelihood and it stops when it reaches certain threshold (Fraley and Raftery, 2002).

As being conceptually simple and easy to implement, EM algorithm is quite powerful, but the convergence can be slow sometimes. Lots of studies were made to improve its speed, such as an acceleration of it based on classical quasi-Newton optimization techniques by Lange (1995) and some others using generalized conjugate gradient method (Collins, 1997).

4.2 Evaluation Framework

To evaluate the performance of the clustering algorithms described in Section 4.1, we did an experiment with 12 Foursquare users and applied the algorithms on their check-in histories.

4.2.1 The Experiment

Subjects

Our original intention was to look for people with different daily activities, for example, students, working groups and retirees, which could cover the major groups of people in our society. Because aged people were not as into social media as young people, we haven't got retirees as my participants. We ended up with 12 participants living in different cities in Asia and Europe. Their ages ranged from 23 to 50, with most of them at their 20s. 9 of them were females and 3 of them were males. Students served as the major group of my experiment, with 4 of them took 1/3 of all the participants. The rest of the participants were all employees who differed widely in their occupations. 2 of them were working in financial field and the other 6 were working in the areas of bio technology, education, health, information technology, logistic and social work respectively.

Data Collection

Since the temporal aspect and other information of the check-ins were not in our consideration in this experiment, only pure geographic location was needed. For collecting the location data of the users' check-ins, a Foursquare API endpoint named "venuehistory" was employed. When requesting, "venuehistory" returns a list of all venues visited by the specified user, with visiting times included (Foursquare, 2013d).

We sent participants the link of retrieving “venuehistory” provided by Foursquare API documentation[11] and asked them to return all the responded information to us. Figure 2 shows the overview of a response of “venuehistory”. Figure 3 illustrates the detailed information of an item of the checked in venues, with “beenHere” counts the number of times the specified user has been to this venue and venue information, in which only the location information is under our concern.


```
{
    meta: {
        code: 200
    },
    notifications: [
        {
            type: "notificationTray",
            item: {
                unreadCount: 0
            }
        }
    ],
    response: {
        venues: {
            count: 54,
            items: []
        }
    }
}
```


Figure 2. Overview of a response of “venuehistory”

[11] https://developer.foursquare.com/docs/explore#req=users/self/venuehistory

```
{
    beenHere: 2,
    venue: {
        id: "4a1123a1f964a5200c771fe3",
        name: "Wien Westbahnhof",
        contact: {
            phone: "51717",
            formattedPhone: "51717",
            twitter: "unsereoebb"
        },
        location: {
            address: "Europaplatz 2",
            lat: 48.19678913440571,
            lng: 16.337978839874268,
            postalCode: "1150",
            cc: "AT",
            city: "Wien",
            state: "Wien",
            country: "Austria"
        },
        categories: [],
        verified: true,
        stats: {
            checkinsCount: 85383,
            usersCount: 11279,
            tipCount: 101
        },
        url: "http://www.oebb.at"
    }
},
```


Figure 3. Detailed information of an item in the response of “venuehistory”

Except the check-in data, participants were also required to provide us a list of meaningful places based on their check-in history. For each meaningful place they report, we asked them to write down how they name the place (i.e., the meaning of the place), and also the check-in points included in the place. The meaning of a place refers to how do that person understands the place and the activities associate to it, for example, “home” and “university campus”. As one meaningful place may have several check-in points in it, the check-in points’ names are also required. An example can be the shops in a shopping center. Users checked in at several different shops but

they considered the shopping center as one meaningful place. Table 3 illustrates the structure of the meaningful places list. This list will be the ground truth to evaluate the performance of the algorithms later.

Table 3. Structure of the meaningful places list

Discovered place 1:	(Meaning:)	Check-in point name 1	Check-in point name 2
Discovered place 2:	(Meaning:)	Check-in point name 3	Check-in point name 4

To help the participants to recall their checked in points, a web page[12] with all the venues he had checked in was also provided (Figure 4).

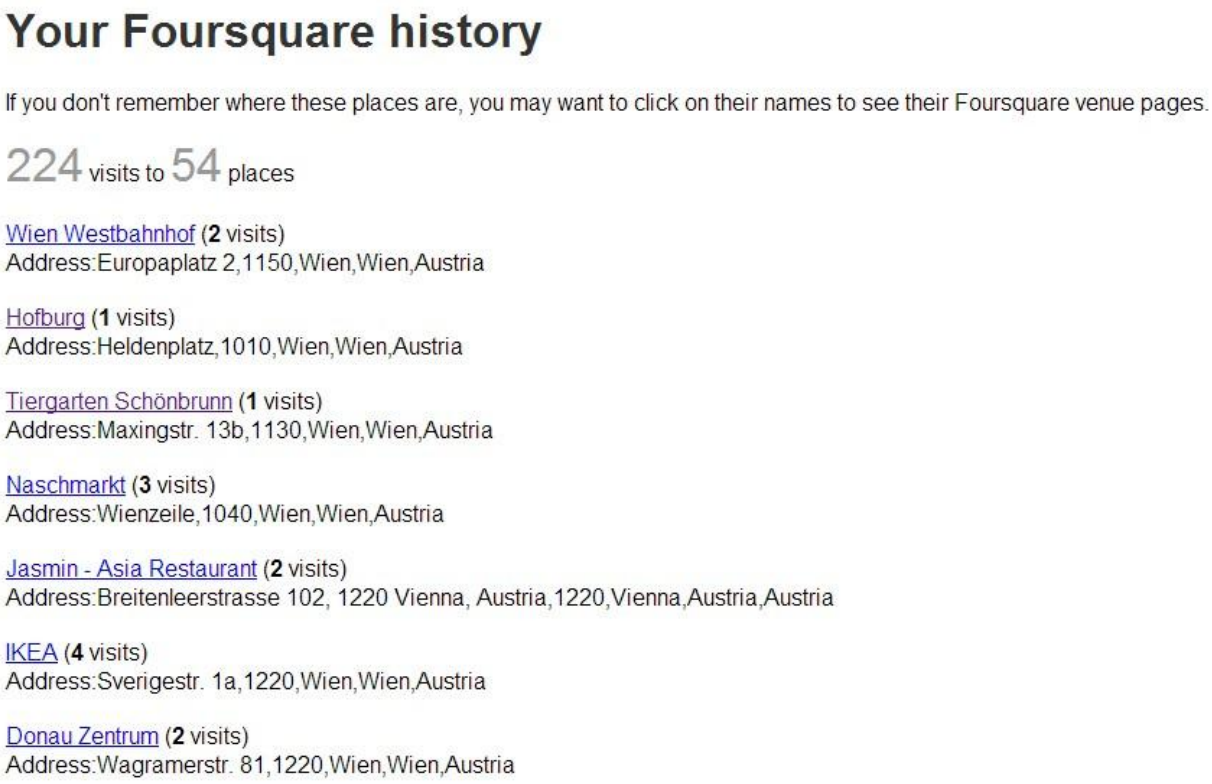

Figure 4. A webpage showing participant's check-in history in experiment 1

[12] http://web.student.tuwien.ac.at/~e1129622/foursquare/f_withoutmap.html

On the webpage, helpful information included the venue's name, address, how many times he had been there and a link to the venue's Foursquare page. The participants were told to click on the links to see more detailed information of the venue, where a map was also provided. Table 4 summarizes the participants' check-in history and their provided meaningful places lists.

Table 4. Summary of the participants' check-in history and their provided meaningful places

Participants	Number of the Total Check-ins	Number of the Checked in Venues	Meaningful Places
Participant 1	51	29	6
Participant 2	113	19	6
Participant 3	125	62	3
Participant 4	52	19	7
Participant 5	74	41	10
Participant 6	25	12	5
Participant 7	76	24	10
Participant 8	103	14	8
Participant 9	105	35	10
Participant 10	224	54	18
Participant 11	47	32	8
Participant 12	67	22	11
Total	1062	363	102

Interview Preparation

After obtaining user's check-in history, we rearranged the data into a file with each item represents one check-in. Each item contains three attributes: the venue name, its latitude and longitude.

The main source of errors in place identification is from the positioning technology (Nurmi, 2009). While in our study, the check-ins on Foursquare were chosen by users consciously, so we didn't consider the errors of GPS here. The mistakes made by the users were hard to be unified and cleaned, thus, they were not deliberated neither. We used the location data directly for clustering.

Then, we ran the above four algorithms with identical parameter settings on the 12 data sets. For the implementation of EM algorithm and K-means algorithm, the statistical computing software environment R was chosen. The R language is widely used in statistical analysis and many user-created package extended the capabilities of it. To perform the EM algorithm, a package called "mclust" was leveraged. "mclust" is an R package for normal mixture modeling via EM, model-based clustering, classification, and density estimation mainly written by Fraley and Raftery (2006).

After performing the algorithm on a data set and requesting summary, the most suitable model and number of clusters can be seen. The clustering result is under the value "classification" (Figure 5).

```
> fit <- mclustBIC(mydata)
> summary(fit, data=mydata)

classification table:
 1  2  3  4
41  1  1  8

best BIC values:
   EEE,4    EEE,5    EEV,4
617.0842 615.1739 613.9619
> mySummary = summary(fit, data=mydata)
> mySummary$classification
 [1] 1 1 1 1 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 3 1 1 4 1 4 4 1 1 1
[37] 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1
```

Figure 5. An example of a data set after performing EM algorithm

For K-means algorithm, it is always hard to define a suitable “K” number. Here, we set “K” according to the cluster numbers suggested by the EM algorithm for the same data set. K-means algorithm is embedded in R already, without any additional package needed.

As to another two algorithms, SLINK and DBSCAN, both of them need a distance function to calculate the distance between two data items. Since they are two locations on the earth, a geographic distance function should be employed to calculate the great-circle distance between them. Though there are some packages providing geographic distance functions in R, the output distance matrix seems to be not exactly suitable for the later implementation of the above algorithms. For the convenience of the second experiment, we wrote them in JavaScript.

Based on empirical tests, the distance threshold to cut off the dengrogram into clusters in SLINK was set to 100 meters and so as the Eps value in DBSCAN. Another parameter in DBSCAN, the MinPts, was given to 3.

After running all the algorithms on each dataset, we stored them as text files in the same form as the meaningful places list for further comparison.

Interview

To understand the confusion between participants provided lists and algorithms' results, an interview with each participant was conducted at the end of the experiment. During the interview, we discussed the results of each algorithm and gathered their advices.

4.2.2 Evaluation of Results

Each item in the meaningful places list is actually a cluster of check-in points. They are denoted by UC, representing "User provided cluster". The total number of UC is abbreviated as NUC. The clusters resulted from the algorithms are called AC in general, representing "Algorithm identified cluster". NAC is the abbreviation for the number of AC in general.

Comparison of User Provided Clusters and Algorithm Identified Clusters

In the evaluation phase, each cluster is then centered by the most significant point inside it, i.e., the most checked in location. If several locations have the same check-in times, the mean value of these locations will be chosen as the center of the cluster.

While matching the user provided clusters and algorithm identified clusters, a cluster will be considered as "correct" when the following criteria is satisfied:

1. If all the check-in points in an AC cluster are exactly the same as the ones in a UC cluster, then the AC cluster is marked correct.

2. If the center of an AC cluster is the same as the center of a UC cluster, then the AC cluster is marked correct.

Based on the experiences in previous place identification study (Zhou 2007, Bhattacharya 2009), the granularity that different users define a place may not be the same. Therefore, "spurious" places are also introduced in this thesis. Spurious places are defined as the algorithm discovered places that can be accepted as one single place to the user, but are better potentially merged with another place (Bhattacharya, 2009). For instance, a university campus may have several buildings in it that a student has checked in. Each building is discovered as one place, which is acceptable to that student but he prefers to consider the university campus as a whole. Spurious places for each algorithm were identified by the users during the interview.

Evaluation metrics

We used precision and recall to measure the accuracy of a place identification algorithm. A "tolerance factor" is used to represent the spurious places. As used by Nurmi (2009) to evaluate the algorithms, F1-score, the harmonic mean of precision and recall, was employed here too. The definition of precision, recall, tolerance factor and F1-score are:

Formula 8. $$\text{precision} = \frac{\text{correct}}{\text{NAC}}$$

Formula 9. $$\text{recall} = \frac{\text{correct}}{\text{NUC}}$$

Formula 10. $$\text{tolerance factor} = \frac{\text{spurious}}{\text{NAC}}$$

Formula 11. $$\text{F1} - \text{score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

4.3 Results

Table 5 shows the overall performances of all the algorithms.

Table 5. Comparison of the algorithms

Algorithm	Correct	Spurious	NAC	NUC	Precision	Recall	Tolerance Factor	F1-score	Precision + Tolerance Factor
SLINK	66	36	128	102	0.516	**0.647**	0.281	**0.574**	0.797
K-means	45	9	67	102	**0.672**	0.441	0.134	0.533	0.806
DBSCAN	53	24	84	102	0.631	0.520	**0.286**	0.570	**0.917**
EM	36	8	67	102	0.537	0.353	0.119	0.426	0.656

The results indicate that most algorithms do not have a good balance between precision and recall. For example, K-means has the best precision but the second worst recall, while SLINK has the highest recall value and F1-score but the lowest precision value. In general, EM has the poorest performance among them. If the precision and the tolerance factor are summed up, which means the rate of meaningful places among all the places discovered by the algorithm, DBSCAN performs very well. With a score of 0.917, nearly all the places discovered by it are actually meaningful to the users. Another interesting finding is that the performances of the algorithms vary strongly on different data sets, i.e., different distribution of the data points. This can be observed from the detailed results of each algorithm, which will be addressed in section 4.3.1.

4.3.1 Detailed Results of Different Clustering Algorithms

As the algorithms' performances differ on data sets with diverse data distributions, it is necessary to have an investigation on the data distribution of different type of users before we discuss the performance of the algorithms.

User Types

The data distribution affects the performance of the algorithms principally. To analyze the performances of different algorithms, users are better to be classified into different types according to the spatial distributions of their check-in data. Three types of users can be noticed by viewing their check-in patterns.

The first type is the users with their check-ins concentrated in some venues. There are not many Foursquare venues covered in their check-in history, but each venue has been visited for a lot of times. User 8 serves as an example. 14 venues were visited in her 103 total check-ins, with the most checked in place counted to 54 visits. From the interview, we got the information that she didn't have many activities except her routine. She checked in when she went to work, shopping or visiting friends on weekends.

The second type is the users with scattered check-ins concentrated in several areas, that is, their check-ins are not concentrated in certain venues like the first type, instead, they are concentrated in several areas while the check-ins are distributed in scattered venues in each area. User 11 seldom had a check-in venue for more than 3 visits, but there was a clear pattern that these venues were located densely in several areas. She revealed that they were the areas around her friend's home, university and places she went traveling. Not being a fun of social media, she was only remaindered to check in when she was with friends who were checking in. She also expressed a reluctance to check in at a same venue for many times. "New places are

more attractive to me", she said. As she didn't go to a large number of places and forgot to check in at most of them, she resulted to have check-ins condensed in several areas, while in each area, they were scattered.

The third type is the users with scattered check-ins all over their reachable space. Both user 3 and user 5 said they were motivated to check in when they were at a new venue that was not checked in before, because they wanted to share it with their friends. User 3 said keep checking in at a same place made her feel dull. User 5 also mentioned the reason for him to check in was to track his traces and help him remember his trips, therefore, there was only few check-ins in his city of residence.

SLINK

In the experiment, SLINK detected many insignificant places. That was why it achieved the best recall, but suffered the poorest precision as a whole. SLINK did not work well on the third type of Foursquare users, which was the ones with scattered distributed check-ins. An extreme example was participant 3 (Table 6). The participant just listed 3 meaningful places, however, SLINK detected 17. Low precision was inevitable even if all the meaningful places were detected. On the contrary, it resulted very well on the first type of users, i.e., the highly concentrated ones. Participant 8 and 10 could both demonstrate that. Although deliberating F1-score as the overall performance, SLINK ranked the highest, it was not the most suitable algorithm among them to derive meaningful places, because even summed up the precision and tolerance factor, the overall percentage that meaningful places took in its discovered place was still far from satisfactory.

Table 6. The Performance of SLINK

	Correct	Spurious	NAC	NUC	Precision	Recall	Tolerance Factor	F1-score	Precision + Tolerance Factor
Participant 1	4	2	12	6	0.333	0.667	0.167	0.444	0.5
Participant 2	4	7	11	6	0.364	0.667	0.636	0.471	1
Participant 3	2	6	17	3	0.118	0.667	0.353	0.201	0.471
Participant 4	7	3	11	7	0.636	1	0.273	0.778	0.909
Participant 5	4	2	9	10	0.444	0.4	0.222	0.421	0.666
Participant 6	3	1	5	5	0.6	0.6	0.2	0.6	0.8
Participant 7	6	2	9	10	0.667	0.6	0.222	0.632	0.889
Participant 8	8	0	9	8	0.889	1	0	0.941	0.889
Participant 9	7	7	16	10	0.438	0.7	0.438	0.539	0.876
Participant 10	16	2	20	18	0.8	0.889	0.1	0.842	0.9
Participant 11	3	2	5	8	0.6	0.375	0.4	0.462	1
Participant 12	2	0	4	11	0.5	0.182	0	0.267	0.5
Total	66	36	128	102	0.516	0.647	0.281	0.574	0.797

K-means

From the results of all the algorithms (Table 5), we noticed that K-means reached the best precision in general. However, it ranked only the third on both recall and F1-score, which suggested that it failed to detect lots of meaningful places and the overall performance were not good enough compared with other algorithms. Implied from Table 7, it brought out quite good results on the second user type, which are the ones with scattered check-ins concentrated in several areas, e.g.,

participant 2 and 11. A fine performance was also yielded on the data of the first user type, however, it also performed poorly on the third type of users.

Table 7. The Performance of K-means

	Correct	Spurious	NAC	NUC	Precision	Recall	Tolerance Factor	F1-score	Precision + Tolerance Factor
Participant 1	3	0	6	6	0.5	0.5	0	0.5	0.5
Participant 2	4	0	4	6	1	0.667	0	0.8	1
Participant 3	1	0	2	3	0.5	0.333	0	0.4	0.5
Participant 4	2	1	4	7	0.5	0.286	0.25	0.364	0.75
Participant 5	2	2	7	10	0.286	0.2	0.286	0.235	0.572
Participant 6	4	4	9	5	0.444	0.8	0.444	0.571	0.888
Participant 7	5	2	9	10	0.556	0.5	0.222	0.527	0.778
Participant 8	3	0	3	8	1	0.375	0	0.545	1
Participant 9	5	0	5	10	1	0.5	0	0.667	1
Participant 10	8	0	9	18	0.889	0.444	0	0.593	0.889
Participant 11	4	0	4	8	1	0.5	0	0.667	1
Participant 12	4	0	5	11	0.8	0.364	0	0.5	0.8
Total	45	9	67	102	0.672	0.441	0.134	0.533	0.806

DBSCAN

If comparing DBSCAN's performance in different types of users, similar to SLINK, it was suitable to users with highly concentrated check-ins but not the ones with scattered distributed check-ins, which was not surprising (Table 8). However, if compared with other algorithms, it produced the best result on users with scattered distributed check-ins. Better than SLINK was that it did not generate too many insignificant places which leads to a quite good precision. Furthermore, when the tolerance factor is summed with precision, it yielded a good outcome of detecting meaningful places. Nearly all the places it discovered were actually meaningful. Considering F1-score, it was also better than the others and only a little bit worse than SLINK. Among all the algorithms, it balanced the best between precision and recall.

Table 8. The Performance of DBSCAN

	Correct	Spurious	NAC	NUC	Precision	Recall	Tolerance Factor	F1-score	Precision + Tolerance Factor
Participant 1	2	2	6	6	0.333	0.333	0.333	0.333	0.666
Participant 2	4	4	8	6	0.5	0.667	0.5	0.572	1
Participant 3	2	3	7	3	0.286	0.667	0.429	0.4	0.715
Participant 4	6	2	8	7	0.75	0.857	0.25	0.8	1
Participant 5	2	2	4	10	0.5	0.2	0.5	0.286	1
Participant 6	2	1	3	5	0.667	0.4	0.333	0.5	1
Participant 7	4	1	5	10	0.8	0.4	0.2	0.533	1
Participant 8	5	0	5	8	1	0.625	0	0.769	1
Participant 9	7	4	12	10	0.583	0.7	0.333	0.505	0.916

Participant 10	14	0	14	18	1	0.778	0	0.875	1
Participant 11	4	5	10	8	0.4	0.5	0.5	0.444	0.9
Participant 12	1	0	2	11	0.5	0.091	0	0.154	0.5
Total	53	24	84	102	0.631	0.520	0.286	0.570	0.917

EM Algorithm for Gaussian Mixture Model

Table 9 illustrates the performances of EM algorithm on all the data sets. Among three types of users, it produced the best result on the second type. In spite of that, when compared with other algorithms, its performances were relatively poor.

Table 9. The performance of EM Algorithm for Gaussian Mixture Model

	Correct	Spurious	NAC	NUC	Precision	Recall	Tolerance Factor	F1-score	Precision + Tolerance Factor
Participant 1	4	0	6	6	0.667	0.667	0	0.667	0.667
Participant 2	3	0	4	6	0.75	0.5	0	0.6	0.75
Participant 3	1	0	2	3	0.5	0.333	0	0.4	0.5
Participant 4	2	1	4	7	0.5	0.286	0.25	0.364	0.75
Participant 5	2	3	7	10	0.286	0.2	0.429	0.235	0.715
Participant 6	3	1	9	5	0.333	0.6	0.111	0.428	0.444
Participant 7	2	3	9	10	0.222	0.2	0.333	0.210	0.555
Participant 8	3	0	3	8	1	0.375	0	0.545	1
Participant 9	4	0	5	10	0.8	0.4	0	0.533	0.8
Participant 10	5	0	9	18	0.556	0.278	0	0.371	0.556

Participant 11	4	0	4	8	1	0.5	0	0.667	1
Participant 12	3	0	5	11	0.6	0.273	0	0.375	0.6
Total	36	8	67	102	0.537	0.353	0.119	0.426	0.656

Moreover, Gaussian distribution of the clusters is already a very strong assumption. Far away points may be considered as in one cluster because they fit the Gaussian distribution. All the points in Figure 6 belonged to a same cluster in the experiment, because a Gaussian distribution was fitted. But in reality, the one called “Taiping Lake” was more than 100km away from the other points and did not share a same meaning with them to the participant. This caused its low score on all the indexes.

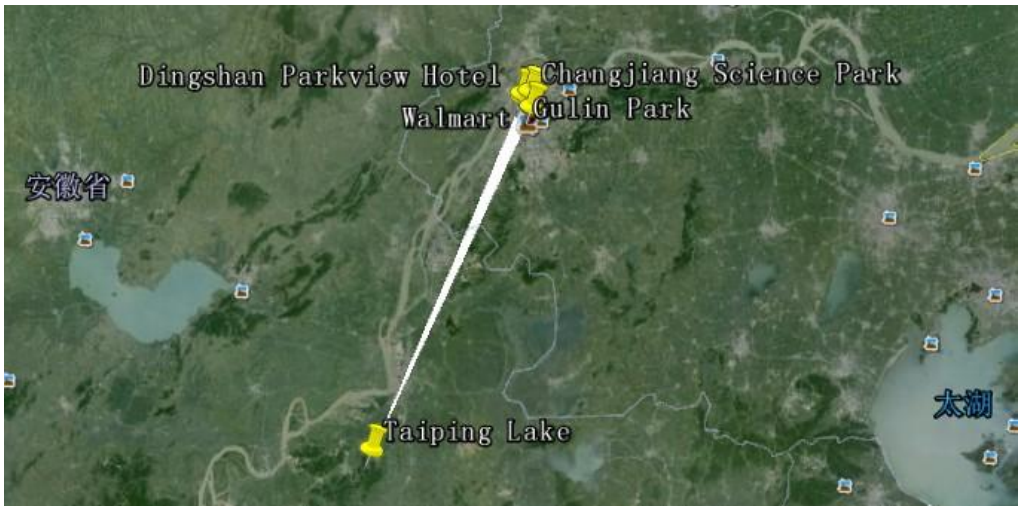

Figure 6. A cluster discovered by EM algorithm for Gaussian Mixture Model

4.4 Discussions

The general results of all the algorithms indicate that DBSCAN performs the best among the four algorithms, which is not surprising. Since DBSCAN is a density based clustering algorithm, it is the density and distance measure determine the clusters. In this experiment, the density corresponds to the number of check-ins in an area within

the predefined distance threshold. As the number of check-ins reflects the frequency of visits, we supposed the dense areas to be the areas that the person visits often in reality. Thus, they tend to be actually meaningful places. The results of the experiment proved such expectation.

However, this conclusion is different with the ones from other researches using GPS data to identify meaningful places (Nurmi 2009, Bhattacharya 2009). In their studies, DBSCAN didn't perform very well. That is because unlike the check-ins on Foursquare, which were made by people actively, the GPS trajectories were logged by the machine automatically. In this situation, they preprocessed the data by pruning it according to a time threshold. However, the assumption that time consuming events are meaningful is not always true in reality, for example, the traffic jams. Thus, when they attempted DBSCAN on their data, a lot of non-meaningful places were detected. Whereas in our experiment, there was a low possibility that users kept checking in non-meaningful places.

From the experiment, we also find that the algorithms' performances depend on the distribution of the data. For the data sets with highly concentrated data, all the algorithms returns good results, because with many check-ins in a same venue, the data is somehow self grouped.

For the second user type, which is the ones with scattered check-ins concentrated in several areas, K-means is most suitable algorithm, because with a suitable "K" value, an optimal partition of the space is easier to reach in this kind of data. However, when using K-means, defining the parameter "K" is always the toughest part. In this experiment, we set it according to the suggested cluster numbers by the EM algorithm, which leaded to the best precision among all the algorithms, but with it

alone, it is unrealistic to determine “K” especially when users’ behavior varies. Another significant drawback is that it takes all the points into account without considering noise, which makes it not robust. Additionally, it is non-deterministic in nature (Zhou, 2004). If we want to take advantage of K-means, we may combine it with other clustering algorithms to help define “K”.

As to the scattered check-ins, which is the most common case in reality whilst the most difficult one to be properly clustered, DBSCAN is a better algorithm than the others.

Although the outcome from DBSCAN was outstanding, it was not perfect. The two parameters of DBSCAN, i.e., Eps and MinPts, are not easy to define. Because of the requirement of parameter setting, it suffers the granularity problem. This problem occurs when user defined granularity is different from the algorithm’s. It can happen when several places are so close to each other that the algorithms fail to distinguish them or the other way around, that the algorithms divided one meaningful place into sub-clusters. The later situation is addressed by the “tolerance factor”.

Not only in DBSCAN, but also in SLINK, this problem is leaded by the distance parameter. It is not easy to find a suitable parameter that balances the two situations. If the distance is set larger, more places will be grouped together, whilst smaller ones is set, there will be lots of small clusters and some more points will be considered as noise as they are not able to be grouped with other points. In some circumstance, one fixed distance parameter is not enough for a single data set. One participant went travelling to another city and checked in at several locations there. On his meaningful places list, the whole city was one place to him, while in his city of residence, his office and a restaurant nearby were considered as two places as they

were associated with different meanings. One possible solution to this kind of users, who have a strong variation in the density of their check-ins, could be a pre-calculation of the user's center mass, where most check-ins exist. The distance parameter for clustering will then depends on the distance between the current check-in location and the center mass. A threshold can be defined to determine where to the change the distance parameter. For doing so, further experiments on human spatial cognition are needed.

4.5 Summary

In order to choose a suitable algorithm to identify individual's meaningful places, an experiment was conducted. In the experiment, we first asked 12 participants for their Foursquare check-in histories and performed four existing algorithms (SLINK, K-means, DBSCAN and EM Algorithm for Gaussian Mixture Model) on each dataset to discover meaningful places. Then, the participants were required to provide their personal meaningful places list based on their check-ins, which served as the ground truth to evaluate the performances of the algorithms. For the evaluation, metrics like precision, recall, tolerance factor and F1-score were calculated for each algorithm on each dataset.

Each algorithm had its own pros and cons. SLINK tended to discover many insignificant places. It worked well on the type of users with a highly concentrated check-in data, but not on the ones with scattered distributed check-ins. The "K" value of K-means algorithm was difficult to define. However, if combined with other algorithms that could suggest a suitable "K" value, it would return a good result on the second type of users, which were the ones with scattered check-ins concentrated in several areas. DBSCAN had the similar user types with SLINK that it was suitable to.

An advantage of it was that almost all the places it detected were actually meaningful. EM Algorithm for Gaussian Mixture Model yielded good results on first and second type of user. However, with an assumption that the data fitted Gaussian distribution, it could cause serious problems.

In general, all the algorithms performed well on the type of users with highly concentrated check-ins, which was not surprising. Both SLINK and DBSCAN balanced very well in all indexes on the data from this type of users. However, they all produced poor results on the type of users with scattered check-ins, with DBSCAN slightly better than the others. For the users with scattered check-ins concentrated in several areas, K-means demonstrated its advantage. EM algorithm for Gaussian mixture model had the poorest overall performance in the experiment, while DBSCAN balanced the best among the four algorithms.

In conclusion, DBSCAN is the most suitable algorithm for our data from Foursquare check-ins. Seeing that the data distribution affects the algorithms performance severely, if we could have a priori knowledge of the data distribution, it is suggested to choose the algorithms accordingly.

Chapter 5: Modeling Individual's Familiarity of Places

With meaningful places been identified from the check-in history, how familiar is the person with these places become the next thing worth considering. In this chapter, I only focus on ranking discovered meaningful places according to an individual's familiarity with them.

5.1 Methodology

As been addressed in Chapter 2, the familiarity of a place can be inferred as the frequency of visits as well as the extensity and intensity of the experiences there. The frequency of visits, which corresponds to the check-ins on Foursquare, has already been employed to discover the meaningful places. If we want to measure the extensity and intensity of the experiences in a place, we should look into the details of each check-in.

The extensity can be represented by the duration of stay in a place. However, from Lindqvist's investigation on Foursquare users (2011), different people have different check in habit, some people check in when they first arrive in a venue, some check in when they are going to leave, while the majority don't have a clear habit, instead, they check in whenever they were reminded to do so. As a consequence, it is very complicated to estimate the duration that an individual stays in a place though there is a timestamp within each check-in.

With the extensity of experience in a place nearly impossible to be calculated, is the intensity able to measure? Fortunately, the answer is positive. The intensity can be represented by the affected responses aroused by that place. When checking in, users can add a short text description named "shout" and also photos. Suggested by

Bower (1970), linguistic encoding is preferred to stand for experience. Hence, the "shout" along with the check-in is treated as an expression of the experience there. When talking about photographing, people take photos for various reasons (Flickr, 2013), but a same reason shared by most people is the thing been photographed caught the photographer's eyes. Therefore people's impression on the place where the photo is taken is generally stronger compared with places without any photo taken activities. As a certain Foursquare user's behavior is consistent, the checked in places with a "shout" or photos are considered to be more familiar to the user than the ones without them. After checking in a place, user can also mark it "like", which is a coarse representation of emotional attachment to it and results in a higher familiarity. In summary, the intensity of experience in a place can be depicted by "shout", "photos" and "like" inside a check-in item.

We treat "shout", "photos" and "like" each as the result of an activity happened along with the check-in activity in the checked in place, i.e., micro blog writing activity, photo taking activity and liking activity. All these activities increase the intensity of experience in that place. Thus, we consider that they have the same influences. To measure the familiarity of a place based on the check-ins, a weighting is introduced. The weighting of each check-in is calculated as follows,

1. Each check-in has an initial weighting of 1;
2. If there is a "shout" in a check-in, the weighting will plus 1;
3. If there are photos in a check-in, the weighting will plus 1;
4. If the check-in is marked "like" by the user himself, the weighting will plus 1.

Therefore, the weighting of a check-in is at least 1 and can sum up to at most 4.

After weighting each check-in, a clustering algorithm will be performed on the check-in data to discover meaningful places. Concerning about the overall performances of the clustering algorithms compared in Chapter 4, with the good completion on discovering meaningful places, DBSCAN is selected. The weighting of each discovered place is then the sum of the weightings of all the check-ins belong to this place. The familiarity of each place is represented by the weighting. As the numerical values of the weightings do not have actual meanings in reality, we then rank the discovered places according to their weightings. The ranking of the discovered meaningful places stands for the relative familiarity of the individual with them. Thus, our model is able to provide the ordinal measure of individual's familiarity of places.

In order to find out whether the proposed method is able to model individual's familiarity of places, we will compare it with a random ranking of the discovered places. To answer if considering affected responses aroused by the place (expressed as "shout", "photos", and "likes" in Foursquare) helps to improve the ranking results or not, it will also be compared with the method only considers the frequency of visits (i.e., the number of check-ins). In the later method, all the check-ins are weighted as 1. When the meaningful places are discovered, the weighting of each place is calculated in the same way as the proposed method. Then the discovered places will be ranked in accordance with these weightings.

5.2 Implementation of the Method

To implement the proposed method and the comparison, a website was made in the form of an online survey[13].

5.2.1 Introduction of the Website

This study was shortly introduced on the first page (Figure 7).

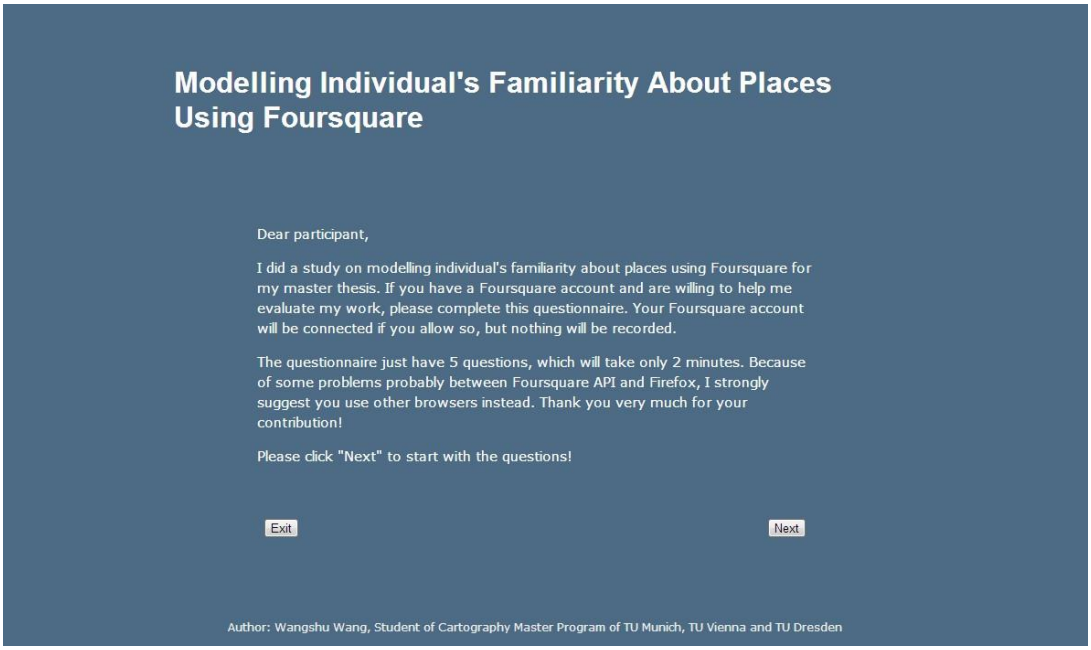

Figure 7. The first page of the online survey with an introduction of this study

Starting with the survey, we asked for some general information from the participants first, which were their age, gender and occupation. Then their Foursquare account would be connected, where a redirect page would occur to request their permission for it (Figure 8).

[13] The survey is at the link of http://web.student.tuwien.ac.at/~e1129622/foursquare/index.html .

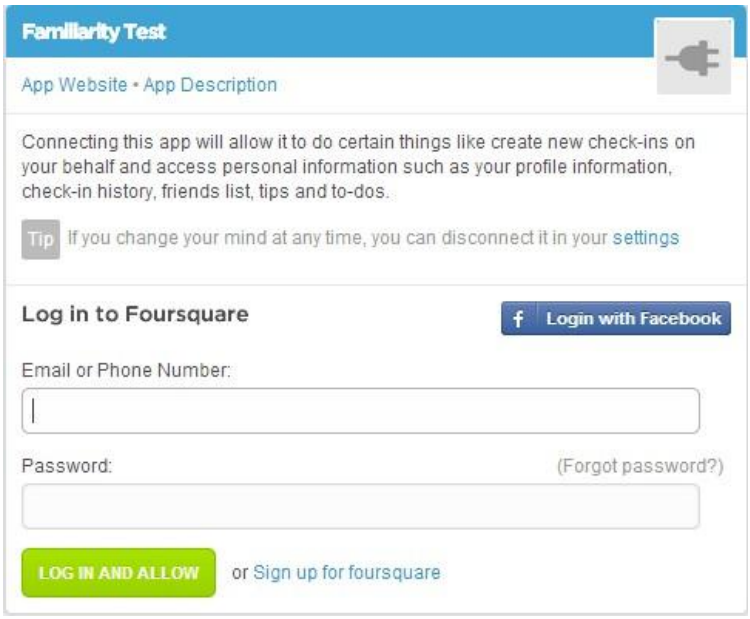

Figure 8. Redirect page of Foursquare

When connection was successful, a page showing a list of the participant's discovered meaningful places would turn up, which required the participant to rank the places according to his familiarity to them (Figure 9).

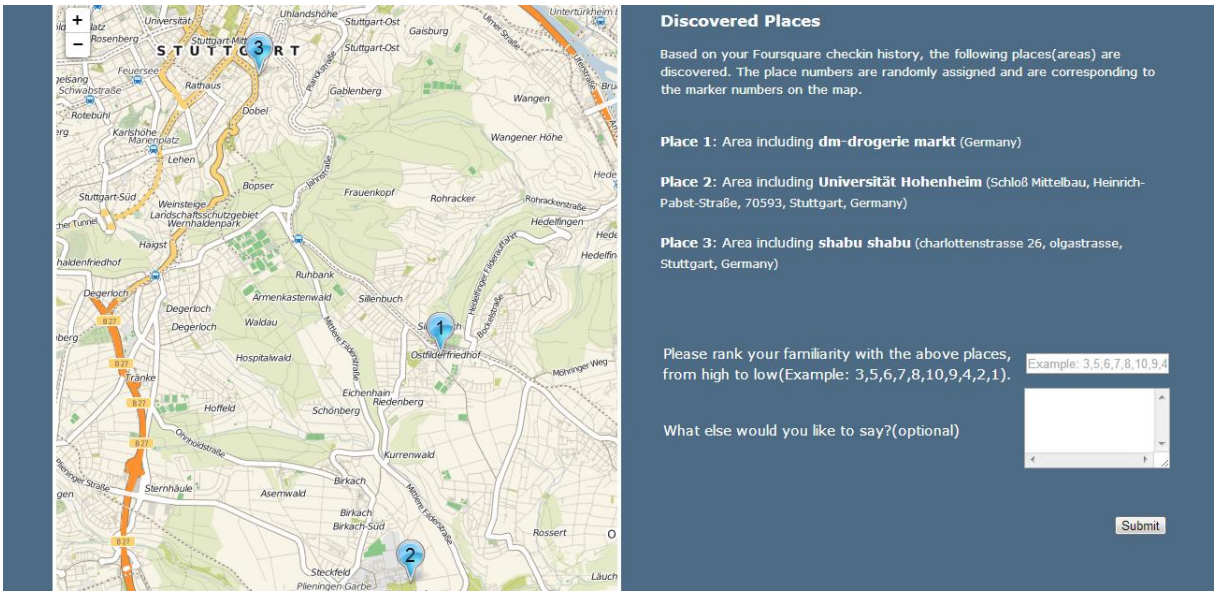

Figure 9. A screenshot of the webpage where users can input their ranking of the places

A participant could have lots of places discovered, for simplicity, if more than 10 places were discovered, only the top 10 places would be presented to him. These places were randomly displayed and were also located on a map, with the corresponding numbers in the markers to give the participant a clearer impression on the places. A short explanation of it and a declaration stating that no personal information of the participant would be recorded was also available on the page. After submitting the ranking, the survey was going to the end with a page that thanks them for their contribution.

5.2.2 Implementation of the Website

Foursquare API was leveraged to connect our website with participant's Foursquare account. The detailed information on the connecting process was covered in Chapter 3. Once connected, the participant's check-in history was retrieved via the "checkins" endpoint.

All the manipulating on the check-in history was written in JavaScript. The history was read and stored in an array. Each item in the array was a check-in with its location name, address, latitude and longitude pairs, together with its "shout", photos and "like". Two weightings were then added to the items in the array. "weighting1" was calculated under the criteria for the proposed method stated earlier. "weighting2" was simply set to 1. The parameters of DBSCAN, Eps and MinPts, were set to 100 meters and 3 respectively. For calculating the great-circle distance, that is, the shortest distance over the earth's surface between two check-in points, the "haversine" formula [14] was used. After performing the clustering method,

[14] http://en.wikipedia.org/wiki/Haversine_formula

"weighting1" and "weighting2" of each cluster were then calculated, followed by performing a sorting algorithm based on "weighting1" to the weighted clusters. The ranking resulted from here was noted in each place item. A shuffle function was written to randomize the order the clusters. The original ranking of each place was stored in an array in this order and denoted as "rankingWithaff", which represented the ranking of the proposed method. The random ranking was the order now, which was stored in another array and denoted as "rankingRandom". The sorting algorithm was performed again based on "weighting2", so that the ranking without considering the affected responses was achieved. It was stored in a third array in the order of the random ranking and denoted as "rankingWithoutaff". This randomized order was the order to display the meaningful places on the screen and to locate them on the map. The map was implemented using the open source Leaflet library[15].

A database was connected via PHP code to receive and store participants' information, including the required age, gender, occupation and ranking of the places, with optional comments. In addition, "rankingWithaff", "rankingWithoutaff" and "rankingRandom" were also uploaded to the database. Taking the privacy issues into account, the detailed places information remained hidden.

5.3 Evaluation Framework

5.3.1 The Experiment

The survey was then sent to possible users of Foursquare and allowed two weeks for the experiment.

[15] http://leafletjs.com/

We received 23 effective responses from 14 females and 9 males. Their ages ranged from 23 to 50 with an average age of 27. Similar to the experiment in Chapter 4, counting to 9, students were also the major group of this experiment. There was no retiree took part in. Among the employees, financial field took the biggest part with a number of 3. Other fields covered by the participants are management, life and physical sciences, engineering, health, social service, education and transportation.

5.3.2 Evaluation Metrics

Spearman's rank correlation coefficient (Lehman, 2005) was employed to evaluate the rankings between participants' and ours.

Spearman's rank correlation coefficient, denoted by ρ, measures the strength of association between two ranked variables. The value of ρ is inside $[-1,1]$. If ρ is equals to:

1. 1, then the association between the two rankings is positively perfect. This means the two rankings are the same.
2. 0, then the rankings are completely independent.
3. −1, then the association between the two rankings is negatively perfect. This means one ranking is the reverse of the other.

Depending on if there are tied ranks in the data, two formulas are available to calculate ρ. Since tied ranks didn't exist in our data, the one we used was:

Formula 12. $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

Where, d_i is the difference in paired ranks and n is the total number of cases to be compared.

An online software written in R was employed to fulfill the calculation task (Wessa, 2013). For each participant, we calculated the correlation between the ranking provided by the participants and the ranking generated by the method concerning the affected responses, the one engendered by the method only considering the number of check-ins and a random ranking, respectively.

In order to properly interpret the results of ρ to determine whether our method concerning the affected responses is significantly different with the other two rankings or not, paired t-tests were then performed on them pairwise. A paired t-test measures how different two groups are, where the subjects for these two groups are the same or matched. For the testing, a set of hypotheses should be given first and a p-value will be calculated to decide if the null hypothesis should be rejected or not. If the p-value is small enough, then we can reject the null hypothesis and prove the significance of the difference between the two groups. The level of significance is most commonly set as 0.05, that is, any test with a p-value under 0.05 would be significant (Texasoft, 2008).

In our experiment, the hypotheses are:

H_o: $\mu_1 = \mu_2$ (Means of the two groups of correlations are equal)

H_a: $\mu_1 \neq \mu_2$ (Means of the two groups of correlations are not equal)

5.4 Results

Table 10 shows the statistical description of the ρ value for each ranking. "ρ_with affected responses" stands for the ρ value calculated between the ranking of our proposed method concerning the affected responses and the participants' ranking. "ρ_with affected responses" represents the correlation between the ranking only

considering the number of the check-ins and the participants' ranking. "ρ_random" is the ρ value calculated between the random ranking and the participants' ranking.

Table 10. Statistics of ρ for each ranking

	Mean	Standard Error of the Mean	Median	Sample Variance	Maximum	Minimum
ρ_with affected responses	0.3329	0.1192	0.5	0.3269	1	-1
ρ_without affected responses	0.3097	0.1232	0.5	0.3488	1	-1
ρ_random	0.0459	0.1256	0.0476	0.3631	1	-1

Among all the 23 participants, the minimum values of all the rankings are -1, which means the participant's ranking and our results are exactly the opposite way around. While the maximum values were 1, which means the rankings were the same with the participant's ranking. When comparing the mean value of ρ, the ranking by the method taking the affected responses into account has the strongest association with the participants' ranking. It is slightly better than the ranking generated by the method without considering the affected responses, while they are both much better than the random ranking. Indicated from Figure 10, "ρ_with affected responses" and "ρ_without affected responses" are the same in many times. The overall results implied that there is positive association between the ranking of our proposed method and participants' ranking, but it is not strong.

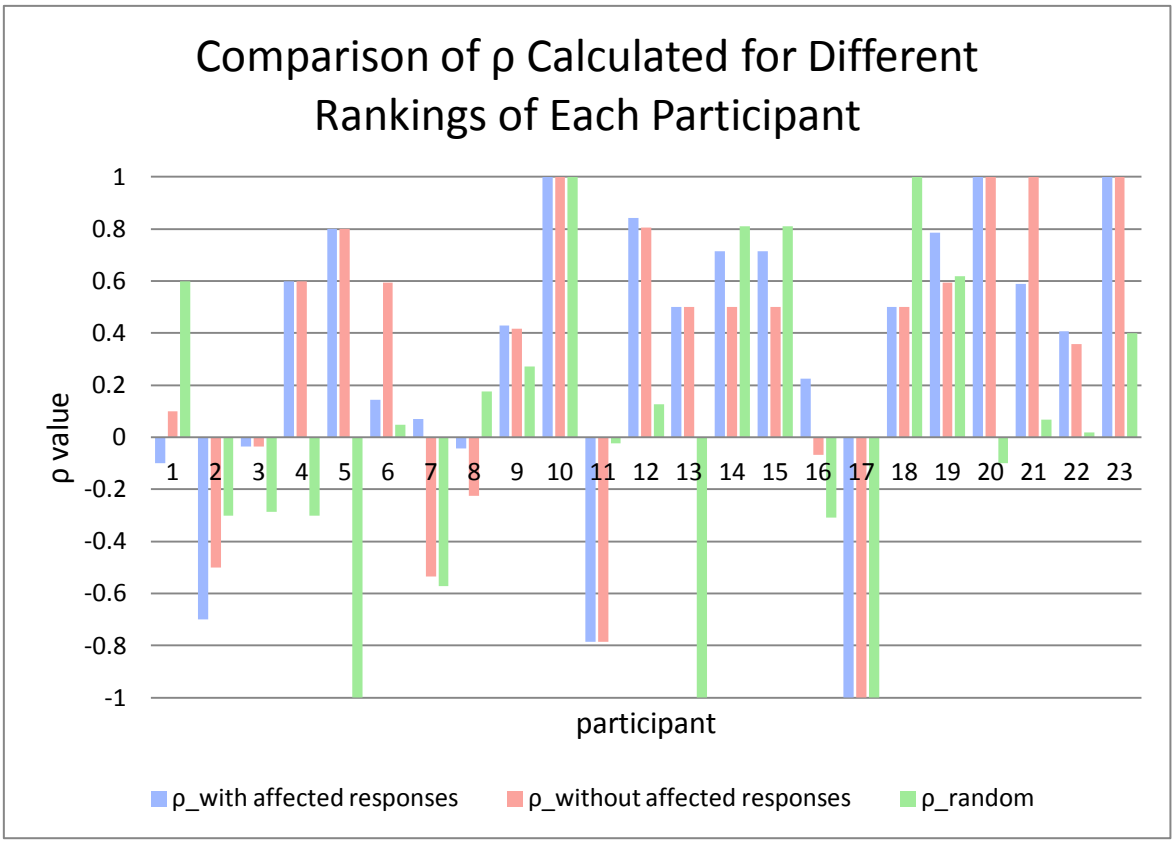

Figure 10. Comparison of ρ calculated for different rankings of each participant

As the values of ρ were not persuading enough, paired t-tests were then conducted to measure if the differences between these methods were significant. With a two-tail p-value smaller than 0.05 (Table 11), it is evidenced that the mean value of ρ_with affected responses is significantly larger than that of ρ_random. In other words, the correlation between the ranking considering the affected responses and the participants' ranking is significantly stronger than that between the random ranking and the participants' ranking. Thus, we can conclude that the method proposed by us is able to model individual's familiarity of places.

Table 11. Results of the t-test between ρ_with affected responses and ρ_random

	ρ_with affected responses	ρ_random
Mean	0.3329	0.0459
Variance	0.3269	0.3631
Observations	23	23
Degree of Freedom	22	
T Stat	2.1189	
P(T<=t) one-tail	0.0228	
T Critical one-tail	1.7171	
P(T<=t) two-tail	0.0456	
T Critical two-tail	2.0738	

However, to our disappointment, the evidence against the null hypothesis is quite weak implied from Table 12. That is to say, taking the affected responses into account may not significantly improve the performance of the model.

Table 12. Results of the t-test between ρ_with affected responses and ρ_without affected responses

	ρ_with affected responses	ρ_without affected responses
Mean	0.3329	0.3097
Variance	0.3269	0.3488
Observations	23	23
Degree of Freedom	22	
T Stat	0.5047	
P(T<=t) one-tail	0.3094	

T Critical one-tail	1.7171
P(T<=t) two-tail	0.6188
T Critical two-tail	2.0738

5.5 Discussions

In section 5.1, we brought forward a method that takes the affected responses aroused by a place (expressed as "shout", “photos”, and “likes” in Foursquare) into consideration to model individual’s familiarity of the place. So as to ascertain if this method can model individual’s familiarity of places and to see whether considering the affected responses is necessary, we also introduced two other rankings to compare with the ranking generated by our proposed method. The first one is a random ranking. The average values of ρ for the random ranking are close to 0, which indicates that this ranking and the participants’ ranking are nearly independent with each other. This demonstrates its randomness. When comparing the results of our method with the random one, it shows its superiority. Further, proven by the paired t-test, our method is significantly better than the random one. Thus, it is confirmed that our method is able to model individual’s familiarity of places. Therefore, our second objective, to model individual’s familiarity of places, is attained.

Another ranking for comparison is the one engendered by a method that only considers the frequency of visits, i.e., the number of check-ins. Unexpectedly, the result of the paired t-test indicates that the difference between these two methods is not significant. That is, considering the affected responses may not necessarily help to model individual’s familiarity of places. Because of this surprising result, we then

performed another paired t-test to uncover the possibility to model individual's familiarity of places with the check-ins alone. As it was not in our original intention, we did not state the result in section 5.4. This paired t-test was between "ρ_without affected responses" and "ρ_random". The result shows that the method only considers the check-ins might not be significantly better than the random one (two sided t-test, $t(22) = 1.9122$, $p = 0.0690$). An interesting pattern is revealed, that is, if only concerning the frequency of visits to a place to model the familiarity with it, the significance of the model cannot be demonstrated. Since we failed to certify the superiority of the method considering the affected responses to the one without it, we could not reach the conclusion that these features in Foursquare representing the affected responses helps to model individual's familiarity of places better neither. Especially when look into the details of the comparison in figure 10, we find sometimes the ranking without these features shows a higher correlation with the participants' ranking. Rather than deducing them as unnecessary, we prefer to leave this question to further studies.

The improper weighting approach may be responsible for the insignificant result. Because we failed to find any proof of qualitative measure of the influencing factors, we simply took them as individual activities with equal influences. This way of weighting is too rough. If we want to model familiarity better, a more precise approach to weight them should be introduced. To increase the accuracy of the weighting on the text descriptions, natural language processing can be implemented. If the text is written in English, each word in a sentence can be analyzed and its affective rating can be looked up in the Affective Norms for English Words (ANEW) (Bradley and Lang, 1999). The value of valence and arousal in it provides a qualitative

measure of the emotional rating. For other languages, similar indexes are also available.

Facing the fact that the mean value of ρ is quite low, the association between the ranking of our method and the participants' ranking is rather weak. One reason for it is the limitation of Foursquare API. The "checkins" endpoint has a limitation of 250 items. When set as default, it will return the newest 250 check-ins. If the participant has a check-in number larger than this, the rest will be automatically ignored. A part of an individual's check-in history can hardly reflect his overall familiarity to places. Against Foursquare API's limitation, some researchers collected Twitter messages which contain Foursquare check-ins instead (Noulas *et al.*, 2011). Although users may not connect to their Twitter account whenever they check in at some places, it is a possible solution to compensate the limitation. In addition, starting from the connection between Twitter and Foursquare, more social media may be involved to discover an individual's meaningful places, e.g., Facebook, Flickr and Instagram.

Another reason could be knowledge fading over time. The temporal influence on knowledge fading was not considered in this study, which could cause a decrease on familiarity of places too. For instance, a user checked in at a place very often one year ago, which would result in a high familiarity from our measurements. He didn't go there for a long time and the familiarity of that place faded over time. He felt the place not as familiar as a place he just checked in several times newly. In consequence, the familiarity rankings of these two places are reversed between his ranking and ours. To solve this problem, the temporal influences can be added at the weighting phase, e.g., set a coefficient which has a negative correlation with the time interval between now and the check-in time. The overall weighting will be the

calculated weighting of all the factors multiplies this coefficient, resulting in higher weighting for recent check-ins and lower weighting for faraway ones.

The last possible reason is the changing of users' check in behavior. In Lindqvist's research on Foursquare users, she pointed out that users' check in behavior may changed because of safety reasons. One of her participant had stopped checking in at home after knowing this could be potentially dangerous. In this case, because she checked in at home before, her home could appear on the place list we provided but with a low familiarity as there were not so many check-ins there. When she is provided the choices, home is usually at a high rank over the other places, thus the association between the two rankings become low. Unfortunately, this kind of problems is very difficult to predict and overcome.

5.6 Summary

To achieve the second objective, we put forward a method that takes the "shout", "photos" and "like" of a check-in item into consideration to measure the individual's familiarity of places. An experiment was designed to rank the discovered meaningful places according to an individual's familiarity with them and evaluate the ranking results. The experiment was implemented by an online survey. By accessing the participant's Foursquare account, his familiarity of places were modeled and quantified as place ranking. Together with the ranking provided by the participant himself and the rankings from two other methods, a method generating random ranking and a method that only considers the frequency of visits (i.e., the number of check-ins), the place rankings were uploaded to our database. The three computer generated place rankings were then evaluated by calculating the rank correlation coefficients with the ranking provided by the participant. Paired t-tests were then

performed between our proposed method and the other two methods to determine the ability of our method to model individual's familiarity of places and the necessity of concerning the affected responses.

The significant difference between our method and the random method evinced that our method is able to model individual's familiarity of places. However, the weak significance of the ascendancy of the method considering the affected responses over the one without concerning it was unexpected. Starting from this point, several possible explanations and future improvements were put forward.

Chapter 6: Conclusions and Outlook

6.1 Conclusions

This thesis aimed at modeling individual's familiarity of places, for which, two objectives was set in the beginning.

- The first objective is to identify individual's meaningful places.

  To derive meaningful places from the check-ins on Foursquare, a suitable clustering algorithm is required. Four existing clustering methods were introduced and one algorithm from each category was chosen for comparison. An experiment was therefore conducted. Based on users' check-in history, different algorithms demonstrated their superiorities on data from different types of users. All the algorithms performed well on the type of user with a highly concentrated check-in data, which was not surprising. Both SLINK and DBSCAN yielded very good balance on the data from this type of user. On the contrary, for the type of user with scattered check-ins, the algorithms' performances dropped down dramatically, with DBSCAN slightly better than the others. K-means did a good job on another type of user, which were the ones with scattered check-ins concentrated in several areas. EM algorithm for Gaussian mixture model had the poorest overall performance in the experiment. In general, DBSCAN balanced the best among the four algorithms, therefore, it was chosen to help fulfill the second objective.
- The second objective is to model individual's familiarity of places.

  The text descriptions, photos and likes accompanying a check-in reflect the affected responses of an individual. They were taken into account to weight each check-in and eventually to measure individual's familiarity of places. Another

experiment was made to evaluate this modeling framework, in which the participants were asked to rank the discovered places according to their familiarity with them. The evaluating result showed there were positive association between the method's ranking and the participants', but it was not strong. Compared with random ranking and the ranking only considering the number of check-ins, this method demonstrates its ability to model individual's familiarity of places, but for better performances, our model needs further improvements.

In summary, we have developed an approach to model individual's familiarity of places using Foursquare data. As the first study that attempts to build this model, it attests the possibility of modeling individual's familiarity of places using social media.

6.2 Outlook

From this study, quite a lot of problems were uncovered. In the place identification part, the algorithms' performances vary according to different data distributions. Hence, preprocessing of the data sets to detect their distribution could be helpful. Cheng (2011) calculated the radius of gyration in his research on human mobility patterns. The gyration measure can be brought in to future study to help recognize the data distribution. Then, the most appropriate clustering algorithm can be implemented to the corresponding data.

In the familiarity modeling phase, more works can be complemented. To weight each check-in more accurately, a preprocessing of the text descriptions can be supplemented. The affective rating of the words can be found in ANEW (Bradley and Lang, 1999), which can then be utilized to calculate the affective rating of the whole sentence user wrote while checking in.

To acquire suitable weightings of the influencing factors, supervised method for knowledge discovery can also be considered. If we have more participants, we can divide them into a training set and a testing set. In this way, we could first train our model on the training set to gain an optimal. Then we use the testing set for evaluation.

As discussed in Chapter 5, the knowledge fading effect ought not to be ignored. The time interval between the check-in created time and current time should be calculated first. Then, the exponential nature of forgetting (Ebbinghaus, 1885) can be taken into account to define a coefficient. The overall weighting will be the calculated weighting of all the factors multiplies this coefficient, resulting in higher weightings for recent check-ins and lower weightings for faraway ones. A more complex approach could consider the checked in venues but not the check-ins themselves. This is because not all the check-ins are new encounters to a venue, the memory fading over time and strengthening when visiting again are to the venues.

For a better representing of individual's familiarity of a place, more social media can be replenished in future study. With millions of geotagged photos available on Flickr, it serves as a very good data source for understanding human spatial knowledge. In common with Foursquare, together with each photo, there are text descriptions, which could be analyzed in the same way as the ones on Foursquare. Twitter has also gained tremendous popularity in the past few years. With a large number of geotagged tweets and the unlimited public access to the data through the Twitter API, this micro-blogging network can also play an important role in modeling individual's knowledge of a place. Similarly, other social media like Facebook and Instagram can also contribute to it.

In general, using social media to model individual's familiarity of places is a novel study. Therefore, more investigation could be made in environmental psychology, geographic information science and computer science to further complement it.

Bibliography

Aitken, S. C., Cutter, S. L., Foote, K. E., and Sell, J. L., 1989. Environmental perception and behavioral geography. In *G.L. Gaile and C.J. Wilmott (Eds.), Geography in America.* Columbus: Merrill Publishing Company, 218-238.

Ankerst, M., et al., 1999, OPTICS: Ordering Points To Identify the Clustering Structure. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, 49-60.

Appleyard, D., 1970. Styles and methods of structuring a city. *Environment and Behavior*, 2 (1), 100–117.

Ashbrook, D., and Starner, T., 2002. Learning Significant Locations and Predicting User Movement with GPS. In *Proceedings of the 6th IEEE International Symposium on Wearable Computers*, 101–108.

Äyrämö, S., and Kärkkäinen, T., 2006. Introduction to partitioning-based clustering methods with a robust example, Reports of the Dept. of Math. Inf. Tech. (Series C. Software and Computational Engineering), 1/2006, University of Jyväskylä.

Bhattacharya, S., 2009. Place Identification: A Comparative Study. (Master's Thesis). University of Helsinki, Helsinki, Finland.

Bower, G. H., 1970. Analysis of a mnemonic device. American Scientist, 58, 496-510.

Bradley, M.M., and Lang, P.J.,1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Chen, W. C., Battestini, A., Gelfand, N., and Setlur, V., 2009. Visual summaries of popular landmarks from community photo collections. In *Proceedings of the 17th ACM international conference on Multimedia.* New York, NY, USA: ACM, 789–792.

Cheng, Z., Caverlee, J., Lee, K. and Sui, D., 2011. Exploring Millions of Footprints in Location Sharing Services. ICWSM, The AAAI Press.

Collins, M., 1997. The EM Algorithm. URL http://faculty.washington.edu/fxia/courses/LING572/EM_collins97.pdf.

Cranshaw, J., et al., 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City, In *Proceeding of 6th Int'l AAAI Conf. Weblogs and Social Media*, AAAI Press, 81–88.

Dempster, A., Laird, N., and Rubin, D., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society,* 39(1), 1–38.

Downs, R. M. and Stea, D., 2009. *Image and environment: cognitive mapping and spatial behavior.* Chicago: Aldine.

Ebbinghaus, H., 1885. *Memory: a contribution to experimental psychology.* Translated by H. A. Ruger and C .E. Bussenius, 1913. New York: Teachers College, Columbia University.

Edelsbrunner, H., Kirkpatrick, D., and Seidel, 1983. R.: On the shape of a set of points in the plane. Information Theory, IEEE Transactions on 29(4), 551-559.

Ester, M., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of Second International Conference on Knowledge Discovery and Data Mining,* 226-231.

Flickr, 2008. http://code.flickr.net/2008/10/30/the-shape-of-alpha/.

Flickr, 2013.
http://www.flickr.com/groups/573182@N24/discuss/72157606445761692/

Foursquare, 2013a. http://foursquare.com/about. Referenced November 21, 2013.

Foursquare, 2013b. https://developer.foursquare.com/docs/.

Foursquare, 2013c. https://developer.foursquare.com/overview/auth .

Foursquare, 2013d. https://developer.foursquare.com/docs/users/venuehistory .

Fraley, C., and Raftery, A.E., 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97, 611-631.

Fraley, C., and Raftery, A.E., 2006. MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. Technical Report no. 504, Department of Statistics, University of Washington.

Golledge, R.G., Stimson R. J., 1997. *Spatial behavior: a geographic perspective.* London: The Guilford Press, 161-166.

Goodchild, M.F., 2007. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. 2.

Hart, R.A., and Conn, M. K.,1991. Developmental perspectives on decision making and and action in environments. In *T. Gärling and G. W. Evans (Eds.) Environment, Cognition and Action: An Integrated Approach.* New York: Plenum Press, 277-294.

Hart, R.A., and Moore, G.T., 1973. The development of spatial cognition: A review. In *Downs, R. M. and Stea, D., Image and environment: cognitive mapping and spatial behavior.* Chicago: Aldine, 246-288.

Heller, K. A., and Ghahramani, Z., 2005. Bayesian Hierarchical Clustering, In *Proceedings of the 22nd international conference on Machine learning*, 297-304.

Hollenstein, L. and Purves, R.S., 2010. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1, 21–48.

Kang, J. H., Welbourne, W., Stewart, B., and Borriello, G. 2005. Extracting places from traces of locations. SIGMOBILE Mob. Comput. Commun. Rev. 9, 3, 58-68.

Keßler, C., Maué, P., Heuer, J.T., and Bartoschek, T., 2009. Bottom-Up Gazetteers: Learning from the Implicit Semantics of Geotags. In *K. Janowicz, M. Raubal, and S. Levashkin, eds. GeoSpatial Semantics.* Springer Berlin Heidelberg, 83-102.

Komninos, A., Stefanis, V., Plessas, A., Besharat, J., 2013. IEEE Pervasive Computing, 12(4), 20-28.

Krämer, B., 1995. Classifications of generic places: Explorations with implications for evaluation. *Journal of Environmental Psychology*, 15:3-22.

Lange, K., 1995. A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica* 5, 1-18.

Lehman, A., et al., 2005. *Jmp For Basic Univariate And Multivariate Statistics: A Step-by-step Guide*. Cary, NC: SAS Press, 123.

Lindqvist, J. et al., 2011. I'm the Mayor of My House: Examining Why People Use Foursquare. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2409-2418.

MacQueen, J. B., 1967. Some Methods for classification and Analysis of Multivariate Observations, In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, 1:281-297.

Marmasse, N. and Schmandt, C., 2000. Location-aware information delivery with ComMotion. In *Proceedings of the 2nd International Symposium on Handheld and Ubiquitous Computing (HUC),* volume 1927. Springer, 361–370.

Mehrabian, A., and Russell, J. A., 1974. *An Approach to Environmental Psychology.* Cambridge, MA: MIT Press.

Montello, D. R., 1998. A new framework for understanding the acquisition of spatial knowledge in large-scale environments. In M. J. Egenhofer and R. G. Golledge (Eds.), *Spatial and temporal reasoning in geographic information systems*pp. New York: Oxford University Press, 143-154.

Naaman, M., 2011. Geographic Information from Georeferenced Social Media Data. *SIGSPATIAL Special*, vol. 3, no. 2, 54–61.

Noulas, A., Scellato S., Mascolo C., and Pontil M., 2011. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. *The Social Mobile Web*, volume WS-11-02 of AAAI Workshops, AAAI.

Nurmi, P., 2009. Identifying Meaningful Places (PhD Thesis). University of Helsinki, Helsinki, Finland.

Piaget, J. and Inhelder, B., 1967. *The Child's Conception of Space.* New York: Norton.

Rattenbury, T., and Naaman, M., 2009. Methods for extracting place semantics from flickr tags. ACM Transactions on the Web 3, 1.

Relph, E., 1976. *Place and Placelessness.* London: Pion Books.

Reynolds, D. A., 2009. Gaussian Mixture Models. Encyclopedia of Biometrics, 659-663.

Rokach, L., and Maimon, O., 2005. Clustering Methods. In L. Rokach and O. Maimon (Eds.), *Data Mining and Knowledge Discovery Handbook*, 321-352.

Shemyakin, F. N., 1962. General problems of orientation in space and space representations. In B.G. Anan'yev et al. (Eds.), *Psychological Science in the USSR*: (Vol.1), NTIS Report No. TT6211083 (pp. 184-255). Washington, DC: Office of Technical Services.

Sibson, R., 1973. SLINK: An Optimally Efficient Algorithm for the Single-Link Cluster Method, *The Computer Journal* ,16 (1): 30-34.

Siegel, A. and White, S., 1975. The Development of Spatial Representations of Large-Scale Environments. In W. H. Reese (Eds.), *Advances in Child Development and Behaviour*. New York: Academic Press, 9–55.

Texasoft, 2008. Understanding Statistical Hypothesis Testing, URL http://www.stattutorials.com/understanding-hypothesis-testing.html

Tuan, Y.-F., 1977. *Space And Place: The Perspective of Experience*. Minneapolis: University of Minnesota Press, 138-184.

Ulrich, R.S., 1983. Aesthetic and affective response to nature environments. In I. Altman and J. F. Wohlwill (Eds.), *Human behavior and Environment.* New York: Plenum Press, 85-125.

Wessa, P., 2013. Free Statistics Software, Office for Research Development and Education, version 1.1.23-r7, URL http:// www.wessa.net/.

Winkielman, P., Schwarz, N., Reber, R., and Fazendeiro, T.A., 2003. Cognitive and affective consequences of visual fluency: When seeing is easy on the mind. In L.M. Scott and R. Batra(Eds.), *Persuasive imagery: A consumer response perspective*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 75–89

Zajonc, R.B., 1968. Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9 (2, Pt.2), 1–27.

Zhou, C. et al. 2004. Discovering personal gazetteers: An interactive clustering approach. In Proc. ACMGIS

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., and Terveen, L., 2007. An Experiment in Discovering personally meaningful places: An interactive clustering approach. ACM Trans. Inf. Syst., 25 (3).