Recommending Attractive Thematic Regions by Semantic Community Detection with Multi-sourced VGI Data

Abstract: Attractive regions can be detected and recommended by investigating users' online footprints. However, social media data suffers from short noisy text and lack of *a priori* knowledge, impeding the usefulness of traditional semantic modelling methods. Another challenge is the need for an effective strategy for the selection/recommendation of candidate regions. To address these challenges, we propose a comprehensive workflow which combines semantic and location information of social media data to recommend thematic urban regions to users with specific interests. This workflow is novel in: (1) developing a data-driven geographic topic modelling method which utilizes the co-occurrence patterns of self-explanatory semantic information to detect semantic communities; (2) proposing a new recommendation strategy with the consideration of region's spatial scale. The workflow was implemented using a real-world dataset and evaluation conducted at three different levels: semantic representativeness, topic identification and recommendation desirability. The evaluation showed that the semantic communities detected were internally consistent and externally differentiable and that the recommended regions had a high degree of desirability. The work has demonstrated the effectiveness of self-explanatory semantic information for geographic topic modelling and highlighted the importance of including region spatial scale into the model for an effective region recommending strategy.

Keywords: location recommendation; geographic topic modelling; community detection; multi-sourced VGI

1. Introduction

Location-based social networks (LBSNs) and volunteered geographic information (VGI) (Goodchild 2007) provide considerable opportunities for studying traditional issues such as human mobility and urban structure from a new perspective (Mislove et al. 2011, Stefanidis et al. 2013, Tsou et al. 2014, Huang and Wong 2015, Longley and Adnan 2016). Besides spatial information, semantic information such as text and hashtags are also attached in LBSNs to indicate the activities related to the social media content. By combining the online footprint and

semantic information, an attractive region detection and recommender system can be developed, and is of great value if desirable regions can be recommended to users.

Previous work on location recommendation focussed on recommending venues and POIs (points of interest) (Ye et al. 2011, Liu et al. 2013, Yuan et al. 2013, Feng et al. 2015, Ying et al. 2017). These studies were limited in their range of application scenarios. For example, a user may prefer to visit a region with many shops in order to compare options, rather than be simply recommended a single venue with no alternatives. Under such circumstances, POI recommender systems are of limited value as they fail to offer guiding information when users' demands are to explore thematic regions with multiple desirable venue options.

Consequently, a major motivation for our research is to develop an effective workflow to detect attractive thematic regions from VGI data and provide useful information for users intending to explore regions of specific themes. To achieve this, there are two theoretical challenges to be considered.

The first challenge is how to develop an effective method to investigate the geographic pattern of social media topic and identify those regions relating to different themes. Statistical methods, such as Latent Dirichlet Allocation (LDA), are traditionally used for geographic topic modelling (Li et al. 2007, Pan and Mitra 2011, Long et al. 2012, Wang et al. 2012, Ghosh and Guha 2013, Cheng and Wicks 2014, Gao et al. 2017), which, nevertheless, have limited effectiveness in handling social media data. A major reason is that statistical methods commonly require large amounts of well-organized documents as training data, which is inconsistent with the social media environment where short and noisy texts predominate (Prateek and Vasudeva 2016); another reason is that some statistical methods require predefined parameters, such as counts of topics, which are subjective due to a lack of *a priori* knowledge.

The second challenge is to develop a recommending strategy which takes account of varying spatial scales. An example is that regions with large areas may have more visits than those with small areas because they have more venues. This scale issue is specific to region recommendation, as venue recommendations typically treat the data as points and thus do not consider their sizes.

To address the above challenges, a comprehensive region recommendation workflow is proposed. The workflow consists of three modules: (1) interest inference, (2) candidate prediction and (3) query response. In the interest inference module, we developed a new data-

driven geographic topic modelling method as the semantic information attached to social media post is well self-explanatory. The candidate prediction module includes a new recommending strategy which models mutual reinforcement between region attractiveness and user expertise with the consideration of region's spatial scale and user redundant visits. In the query response module, a list of recommended regions is returned to respond user queries. This workflow was implemented with a real-world dataset and evaluated from different aspects. The results demonstrated the effectiveness of our methods.

The remainder of the paper is organized as follows: Section 2 summarizes the related work. Section 3 gives an overview of the proposed workflow and elaborates each module. Section 4 provides the experimental results and demonstrates the effectiveness of our methods. Section 5 gives a discussion about the performance of our recommendation strategy and presents the limitation and future work. Finally, Section 6 concludes the whole paper.

2. Related Work

Our work aims to recommend attractive urban regions by considering several interrelated challenges: how to detect thematic regions using VGI data; and how to develop an effective rating and ranking strategy for region recommendation. In this section, some previous studies on geographic topic discovery and location recommendation are presented and discussed.

2.1. Geographical Topic Discovery

The aim of geographical topic discovery (Yin et al. 2011) is to discover topics from spatial big data (e.g., GPS-associated document, geo-tagged social media data) in a geographical context. Among the geographical topic modelling methods, statistical models such as Latent Dirichlet Allocation (Blei et al. 2003) are commonly used for geographic topic modelling（Zhang et al. 2013, Lansley and Longley 2016, Steiger et al. 2015）. However, the LDA method has limitations in relation to social media data because short noisy texts are predominant and *a priori* knowledge such as a predefined count of topics is unavailable (Prateek and Vasudeva 2016). In other research studies, other forms of information have been used, such as images (Rykov et al. 2016) and heterogeneous unstructured articles (Adams and Janowicz 2012), to facilitate the process of topic discovery. While in our work, to detect thematic regions, we proposed a data-driven geographic topic modelling method based on the assumption that the semantic information attached to social media post, such as hashtags, is sufficiently self-explanatory and the potential topics can be thus indicated. The method is data-driven and

requires no well-organized training data or *a priori* knowledge such as topic counts, thus reducing potential perceptual biases.

2.2. Location Recommendation in LBSNs

The increasing availabilities of location-based social networks (LBSNs) provide researchers with new tools and data sources to study and recommend attractive locations.

According to Bao et al. (2015), there are two main types of stand-alone recommended locations: POIs and regions. POI recommender systems aim to provide users with individual locations, such as museums or restaurants that match user preferences (Bao et al. 2012, Yuan et al. 2013, Griesner et al. 2015, Ayala et al. 2017, Ying et al. 2017). However, POI recommender systems are limited in effectiveness, when users prefer to explore regions where there are multiple desirable venues. This gap can be bridged by region detection and recommendation (Kurashima et al. 2010, Sun et al. 2015). From the level of the individual, Huang (2016) introduced a methodology to cluster and predict a user's next location based on the sparse online footprints accumulated over a long period of time. In addition to facilitating user travel, region detection and region recommendation also aid in the study of the social dynamics of a city at a large scale (Cranshaw et al. 2012, Lee et al. 2014). However, among existing methods, few efforts have been made to integrate the related parameters of region spatial scale and users' redundant user visits into the model, which may undermine the effectiveness of current region recommending strategies.

3. Proposed Methods

In this section, our proposed workflow is described, followed by a detailed description of each data handling process.

3.1. Analytic Workflow

Our analytic workflow consists of three main modules (Figure 1): (1) interest inference module; (2) candidate prediction module; and (3) query response module. Several data stores including user information, check-ins and venue information, are incorporated into the workflow.

Figure 1. Analytic Workflow

Interest Inference Module: The semantic content attached to the LBSNs is used to investigate the user activity pattern associated with the geo-tagged content (e.g. geo-tagged photos). This step starts by retrieving geo-tagged media data (check-in dataset) from Instagram

API and venues information (venue dataset) from Foursquare API. The hashtags in the titles attached to the geo-tagged photos are extracted, based on which an undirected 'hashtag network' model is built where each hashtag is denoted as a network node and the co-occurrence frequency between hashtags is assigned as a weighting to the edge connecting the corresponding nodes in the 'hashtag network' (Tag Network Construction). A greedy optimization method is then implemented to explore the communities from the hashtag network, and a common topic is assigned to the hashtags of the same community (Semantic Exploration). The topics of geo-tagged photos are further detected by introducing the topics of attached hashtags as an indicator, which enables multiple topics to be assigned to one photo, as shown in Figure 2 (Interest Identifying).

Figure 2. Instagram photo sample. The photo has a title that contains hashtags (e.g., #foodporn, #seafood, #delicious, #friendships, #saturdaynight) to annotate it, indicating multiple topics for one single photo.

Candidate Prediction Module: In this module, candidate attractive regions and experienced users are predicted. Firstly, the geo-tagged photos (check-ins), whose topics have been identified, are further processed by a density-based clustering algorithm to acquire the attractive regions. Secondly, derived from Hyper-Induced Topic Search (HITS) (Zheng and Xie 2011), a new method is used to predict region attractiveness and user expertise with the input of attractive regions, topic-identified check-ins and venues in terms of the varying geographic range (Interest-aware Candidate Selection).

Query Response Module: This phase starts when users initiate a query asking for a certain count of regions matching an interest (e.g., food) within a spatial range (e.g., 5 km). Regions meeting the criteria are selected from the candidate region set. Finally, a list of regions is ranked in descending order by attractiveness score and recommended to users.

3.2. Interest Inference Module

As noted above, LDA methods have several limitations when dealing with geo-tagged social media data, so we develop a new method of modelling the geographic pattern of topics of social media contents. The assumption is that the semantic information, such as hashtags, is highly relevant to the topics of a social media post, and that hashtags of similar semantic meaning have a high probability of co-occurrence (Figure 2). Consequently, in a rich hashtag environment, a

'hashtag network' model is built to investigate the geographic patterns of topics. Relevant data models are defined below:

Let P be a collection $P = [p_1, \dots, p_n]$ of social media posts $p_i = (l_i, t_i, tgs_i)$, where l_i is the location of post p_i, t_i is the time-stamp and tgs_i is the union of the hashtags attached to p_i. Let $HT = \bigcup_{i=1}^{n} tgs_i = [tg_1, \dots, tg_m]$ be the hashtag union of tgs_i, i.e., if and only if $item \in tgs_i (i = 1, \dots, n)$, $item \in HT$.

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

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

Where p_i is a post item from collection P, tg_j and tg_k are hashtag items from HT.

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

$$A_{jk} = \sum_{i=1}^{n} CRC(p_i, tg_j, tg_k) \quad (2)$$

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

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

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (3)$$

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

To detect communities, the Louvain algorithm is employed (Blondel et al. 2008). As a greedy method inspired by the optimization of modularity, the Louvain algorithm is data-driven and does not require an *apriori* selection of the community count, so that network communities can be detected with less potential perceptual biases. The algorithm is applied to the 'hashtag

network' *HTN* and several hashtag communities can be detected. Based on the semantic meanings of hashtags, a topic can be assigned to each community (Figure 3). The topics of social media post $p_i = (l_i, t_i, tgs_i)$ can be identified by referring to the topics of $p_i.tgs_i$, making it possible to provide multiple topics for one post.

Figure 3. Hashtag community detection based on co-occurrence pattern and modularity, with node and edge size proportional to occurrence frequency and node colour indicating community category.

3.3. Candidate Prediction Module

In this module, candidate attractive regions and experienced users are predicted.

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

$$MinPts = \frac{1}{n}\sum_{i=1}^{n} p_i \quad (4)$$

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

Candidate Prediction: For discovered regions, an HITS-based model is used to predict user expertise and region attractiveness. The idea of Hyper-Induced Topic Search (Zheng and Xie 2011) model is illustrated in Figure 4. Users and regions are all perceived as nodes and a user check-in to one region is regarded as a directed link from user node to region node. The mutual reinforcement relationship is that a user who visits many attractive regions is more likely to be an experienced user and a region that is visited by many experienced users is more likely to be an

attractive region. A hub score is assigned to a user and an authority score to a region respectively to indicate user expertise and region attractiveness.

Figure 4. HITS-based candidate prediction model

One problem is that a region visited frequently does not necessarily have high quality and attractiveness, as the chances that a region is visited are probably proportional to its spatial scale (more specifically, venue count). Another problem is that a user with a specific interest may affect the calculation of region attractiveness and user expertise if a user visits one specific region overmuch due to personal preferences. Consequently, we build data models as below: given *U* is a user set with records of region visiting history, and *R* is a region set with records of users who have visited. For each user $u_i \in U$ and each region $r_i \in R$, a hub score $h(u_i)$ and an authority score $a(r_i)$ are calculated respectively using the following equations:

$$h(u_i) = \sum_{r_j \in u_i.r}(1 + \ln(\sigma)) * a(r_j) \quad (5)$$

$$a(r_i) = \sum_{u_j \in r_i.\mathrm{u}}(1 + \ln(\lambda)) * h(u_j) / n(r_i) \quad (6)$$

Where $u_i.r$ are the regions that user u_i has visited, $r_i.u$ are the users who have visited region r_i, σ is the number of visits of user u_i to region r_j, λ is the number of visits of user u_j to region r_i, and $n(r_i)$ is the count of venues of a specific category (e.g., food venues, shopping venues, etc.) in region r_i. By using these equations, the increase of one single user's contribution to the region authority score decelerates with the increment of the count of user's visits. Moreover, by introducing venue count into the model, the calculated authority value can reflect the average attractiveness of the venues within the region so that, taking account of the impact of region spatial scale, the desirability of the recommended region is predicted.

3.4. Query Response Module

In this module, queries from users for attractive regions are responded to. A user query is represented as below:

$$q = [interest, location, spatial\ range, count\ ] \quad (7)$$

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

Algorithm 1: Query Response Method

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

Output: A ranked list of region recommendations L

Begin

Retrieve regions R' that match user interest i

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

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

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

end

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

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

End

4.Experiment and Evaluation

In this section, we describe an application prototype implementation of our proposed workflow and evaluate effectiveness of our method with a real-world dataset.

4.1 Application Scenarios and Architecture

We implemented the proposed recommending workflow with an android mobile application. A snapshot of the user interface of the application is shown in Figure 5.

A typical usage scenario starts when a user logs into the application to obtain personalized recommendations. The user inputs query location, interest (e.g. food, coffee), spatial range and item count in the system. After clicking the "Go" button, a list of regions matching the request is returned on the screen (Figure 5b). For the application architecture, Microsoft Access database was used to store the region attributes and spatial information and Google Maps API used for map viewing and geocoding.

(a) (b)

Figure 5. Application user interface. (a) home page;(b) recommendation list

4.2. Experimental Settings

There were two kinds of data source needed in the proposed workflow: geo-tagged social media photo and Points of Interest (POIs). A set of geo-tagged photos from Hong Kong were retrieved using the Instagram API between 11-Nov-2014 and 15-Nov-2015. The dataset had 1,774,596 geo-tagged photos generated by 57,662 users. A social media post included the post ID, user ID, the attached hashtags, the time-stamp of the online post and the location (longitude and latitude) indicating the place of posting (Table 1). All user IDs were anonymized for privacy protection.

Table 1. Description of social media post field

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

The POIs data was retrieved via Foursquare API and 32,485 venues were collected in Hong Kong in total. A POI item included the venue ID, venue category, venue location (longitude and latitude), and venue rating (Table 2). The venue rating is a numerical score (0 through 10) calculated from a wide variety of signals derived from user explicit feedback, such as: liking or disliking a venue or leaving a positive or negative tip, as well as user implicit signals, such as: whether the venue tends to have many loyal customers, the credibility and expertise of the users and so on. This rating algorithm had been validated in metropolitan areas and trusted by users for accuracy and reliability in indicating venue desirability (Yang and Sklar 2016, Yang and Sklar 2018).

Table 2. Description of POIs field

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

An anomaly detection method was applied to remove commercial advertising accounts. The count of photos posted by each user was first investigated. Figure 6 shows the proportions of users, by number of posts, with its long tail distribution. The calculations showed that the average count of posted photos per user, μ, was 30.1 and the standard deviation σ, 66.4. The three-sigma rule (Pukelsheim 1994) was introduced, which stated that, for both normally distributed and non-normally distributed variables, most cases should fall within the three-sigma intervals. Therefore, those users with photo counts outside the three-sigma intervals (i.e., $\mu + 3\sigma$, 229) were recognized as outliers and their posted photos removed from the photo dataset. After data cleaning, 1,432,733 photos, generated by 56,878 users, remained.

Figure 6. Proportion of users by number of posts

4.3. Evaluation Approaches

In this section, we present our evaluation framework and then give a detailed description of the evaluation measurements.

4.3.1. Framework of Evaluation

The evaluation framework is illustrated in the Figure 7. Regarding our analytical workflow, evaluations were made at the three levels: semantic representativeness, topic identification, recommendation desirability.

Figure 7. Evaluation Framework

Each aspect needs to answer the following questions:

- **Semantic Representativeness**: How well can the semantic similarity between hashtags be represented by the connectivity between vertices in a 'hashtag network'? In other words, from the semantic perspective, are the semantic communities detected internally consistent and externally different?
- **Topic Identification**: To what extent can the topics of the social media content (i.e. photos) be accurately identified by the proposed method?
- **Recommendation Desirability**: How effective is our recommending strategy? Do the recommended regions match user expectations?

By introducing quantitative measurements from these three levels, the effectiveness of the proposed geographic topic modelling method and personalized recommending strategy were evaluated.

4.3.2. Quantitative Measurements

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

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

$$CSS(C_m, C_n) = \frac{\sum_i \sum_j sim(tag_i^m, tag_j^n)}{\text{Tg\_Cnt}(C_m) * \text{Tg\_Cnt}(C_n)} \quad (8)$$

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

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

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

$$MAE = \sum_i^n \frac{|SR_i - UR_i|}{n} \quad (9)$$

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

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

$$nDCG_p = \sum_{i=1}^{p} \frac{2^{rel_i - 1}}{log2(i+1)} / IDCG \quad (10)$$

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

4.4. Experimental Results

In this section, the evaluation results of semantic representativeness, topic identification and recommendation desirability are presented respectively.

4.4.1. Evaluation of Semantic Representativeness

The hashtag network *HTN* was built first with the experimental dataset. To strengthen the robustness of the division of *HTN*, hashtags appearing only once in the dataset were removed from the network as such hashtags were highly probably generated by user typing mistakes and contained little semantic information. The Louvain algorithm was implemented into the *HTN* and several hashtag communities were detected. The divisions of *HTN* were visualized using a comparison word cloud of the representative hashtags in each community (Figure 8).

Figure 8. A comparison word cloud of 20 hashtag communities

Table 3. shows the semantic similarities between hashtag communities. For a hashtag community C_n, intra community semantic similarity (*intra-CSS*) calculated by $CSS(C_n, C_n)$ was introduced (see Equation 8) to measure the internal semantic consistency within the hashtag community C_n and inter community semantic similarity (*inter-CSS*) calculated by $CSS(C_n, C_m)$ ($n \neq m$) to measure the semantic differences between communities C_n and C_m. Semantically, high *intra-CSS* can indicate that the detected community is internally consistent; significant differences between *intra-CSS* and *inter-CSSs* can indicate the detected community is externally differentiable.

Table 3. Comparison of the semantic similarities between communities

		inter-CSS		
Hashtag Communities	**intra-CSS**	**maximum inter-CSS**	**minimum inter-CSS**	**average inter-CSS**
Drinking&Bar	0.85	0.59	0.32	0.33
Travel&Wander	0.90	0.65	0.27	0.37
LifeStyle&Happiness	0.89	0.68	0.37	0.50
Sports&Fitness	0.86	0.56	0.32	0.37
Food	0.87	0.66	0.15	0.43
DisneyLand&Toys	0.86	0.57	0.23	0.47

Tattoo	0.91	0.34	0.22	0.24
Coffee	0.84	0.49	0.22	0.32
Pets&Animals	0.89	0.56	0.21	0.39
Shop&Luxury	0.85	0.59	0.19	0.21
Watch&SportsCar	0.91	0.37	0.26	0.29
Concert&LiveMusic	0.89	0.58	0.23	0.39
Bike&Cycling	0.92	0.45	0.23	0.25
Vegetarian	0.86	0.49	0.13	0.33
Makeup	0.86	0.54	0.26	0.36
Kids&Baby	0.91	0.49	0.26	0.30
Movie&Video	0.88	0.53	0.29	0.35
Muslism	0.85	0.48	0.19	0.30
Big Bang Group	0.88	0.51	0.23	0.30
Art&Design	0.85	0.48	0.22	0.30

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

'LifeStyle&Happiness' were mainly emotion-related (e.g., 'amazing', 'happy', 'enjoy'). These emotion-related hashtags tended to co-occur with other activity-oriented hashtags to express user emotional reactions and evaluations upon the activities, leading to communities with more external connectivity and higher *inter-CSS* values. Among all communities, the *intra-CSSs* mostly achieved very high values (over 0.85) and the differences between *intra-CSSs* and average *inter-CSSs* were mostly over 0.4 (the exception is 'LifeStyle&Happiness', where the difference is 0.39). The results verify, semantically, that the divisions of semantic communities were internally consistent and externally different.

4.4.2. Evaluation of Topic Identification

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

As shown in Table 4, photo samples identified as 'food' topic were drawn in sizes *N*, where *N* = 1000, 2000, 3000, 4000, to assess the topic identification results.

Table 4. 'Food' photo identification

Sample Number	True Positive, TP	False Negative, FN	False Positive, FP	True Negative, TN	Total Size
1	251	19	31	699	1000
2	520	40	59	1381	2000
3	734	68	96	2102	3000
4	1124	83	152	2641	4000

Figure 9 shows the *precision*, *recall* and *F-measure* calculated from Table 4. The fluctuations of *precision*, *recall* and *F-measure* was investigated with a varying sampling scale. As the size of a data sample grew, the values of the three measurements fluctuated and gradually stabilized around 0.9. When *N* =4000, the *precision* was 0.881, *recall* was 0.931and the *F-measure,* 0.905. This means that on average, for every 10 photos identified as 'food' topic by our method, there were at least 9 items that were actually food-relevant photos, and on the other hand, for every 10 food-relevant photos posted by users, there were 9 items accurately identified and assigned to 'food' topic by the method. Such a high ratio of true positive items demonstrates

the good performance of the method for photograph topic modelling and identification. As the spatial cluster algorithm was implemented based on the spatial distribution of geo-tagged photos of specific topics, effective exploration of the thematic regions can also be demonstrated by good topic identification performance.

Figure 9. *Precision*, *Recall* & *F-measure* of 'food' photos

4.4.3. Evaluation of Recommendation Desirability

By using the *k-dist* plot and Equation 4, the DBSCAN method parameter was set as *eps* =71 and *minPts* =1820, and 21 'food' theme regions were detected. By using Foursquare API, 1960 venues of food category were retrieved in those 21 regions.

Figure 10. Top 5 recommended regions of 4 different themes. Green regions for 'food' theme; Yellow for 'Drinking&Bar' theme, Blue for 'Shop&Luxury' theme; Red for 'Art&Design' theme.

Baseline methods: The regions recommended by our methods were compared with those recommended by three baseline methods: rank-by-users, rank-by-visits and rank-by-attractiveness. With the rank-by-users method, the more visitors a region receives, the more attractive the region is. With the rank-by-visits method, similarly, the more check-ins a region receives, the more attractive the region is. For the third baseline method, rank-by-attractiveness, the attractiveness of a region was calculated with the following mutual reinforcement equations:

$$h(u_i) = \sum_{r_j \in u_i.r} \sigma * a(r_j) \quad (11)$$

$$a(r_i) = \sum_{u_j \in r_i.\mathrm{u}} \lambda * h(u_j) \quad (12)$$

Where $u_i.r$ is the region set visited by users u_i , $r_i.u$ is the user set who have visited region r_i, σ is the number of visits of user u_i to region r_j, λ is the number of visits of user u_j to region r_i, and the attractiveness of each region is decided by the authority score. In the rank-by-attractiveness method, the influences of user specific preferences and spatial scale of region size are not taken into consideration compared with our method (Equations 5 and 6).

Ground truth: A survey of users' ratings on the recommended regions is needed to evaluate the effectiveness of recommendation. As a region spans across certain spatial ranges and covers

many venues, an indicator is needed to evaluate users' general attitude towards the recommended regions as a whole. Therefore, the evaluation indicator was calculated using:

$$UR_i = \sum_{j=1}^{n} \frac{VR_i^j}{n} \quad (13)$$

Where UR_i is the user rating score for the i^{th} recommended region, VR_i^j is the user rating score for the j^{th} venue in the i^{th} region. The user venue rating scores can be retrieved from Foursquare API (*vrating* in Table 2). As mentioned, this rating score is calculated based on a variety of users' real-world explicit feedback and has been validated for metropolitan regions for accuracy and reliability in indicating venue desirability. By using Equation 13, the user satisfaction level for a whole region can be derived. The higher the UR_i is, the more desirable the region is. The effectiveness of our recommendation was evaluated by comparing the ranking and scoring results with the results generated based on *UR*. Table 5 lists the top 10 regions recommended by each method. The *rel* in $nDCG_p$ (Equation 10) was set based on *UR*, which means the region with higher *UR* had higher *rel*. The regions were sorted in descending order by *UR*. As there were 21 detected regions in total, the *rel* of the first region in the list was set as 21. The *rel* was decreased by 1 to the next region in the sorted list and finally the *rel* of the last region was set as 1.

Table 5. Top 10 regions recommended by each method

Rank	Our method	Rank-by-users	Rank-by-visits	Rank-by-attractiveness	Rank-by-UR (Ground Truth)
1	Star Ferry Pier	Lyndhurst Terrace	Lyndhurst Terrace	Lyndhurst Terrace	Star Ferry Pier
2	Canton Road(Tsim Sha Tsui)	Causeway Bay	Causeway Bay	Causeway Bay	Hong Kong Station-Man Cheung Street
3	Tai Hang	Cameron Road(Tsim Sha Tsui)	Cameron Road(Tsim Sha Tsui)	Cameron Road(Tsim Sha Tsui)	Statue Square
4	Hong Kong Station-Man Cheung Street	Sunning Road	Sunning Road	Sunning Road	Canton Road(Tsim Sha Tsui)

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

To evaluate the performance of recommendations in various spatial ranges, different values were assigned to p in $nDCG_p$ (Equation 10), as shown in Table 6. It shows that for all ranking methods, the values of *nDCGs* increased with the increment of p and the best *nDCGs* were achieved when p =21. On average, our recommending strategy can achieve *nDCGs* with an average value as much as around 0.91, which were much greater than those of the three baseline methods (0.2-0.4), indicating that regions with higher user ratings were given more recommendation priority by our strategy. Such results show that items recommended by our method could well match user expectations, demonstrating the advantages of our method over the several baseline methods in terms of effectively ranking and recommending desirable regions.

Table 6. Normalized Discounted Cumulative Gain (*nDCG*) for each ranking method

$nDCG_p$	**Our method**	**Rank-by-users**	**Rank-by-visits**	**Rank-by-attractiveness**
$nDCG_{10}$	0.905	0.225	0.220	0.221
$nDCG_{15}$	0.905	0.228	0.222	0.224
$nDCG_{21}$	0.914	0.389	0.384	0.386

The agreement between our scoring strategy and user rating was further investigated. The results are shown in Table 7. For rank-by-users and rank-by-visits strategies, the user count and visit count were used respectively as the attractiveness score for each region. The *MAE* (mean absolute error, Equation 9) indicates that the attractiveness scores predicted by our method achieved a better agreement with ground truth than the baseline methods.

Table 7. *MAE* (mean absolute error) based on ground truth

	Our method	**Rank-by-users**	**Rank-by-visits**	**Rank-by-attractiveness**
MAE	0.353	0.424	0.458	0.444

5. Discussion

This section includes a comprehensive discussion of the performance of the recommending strategy and then presents the limitations and potential future work.

5.1 Recommendation Performance Analysis

In calculating *nDCG* (Equation 10), the highly desirable regions were given more relevance than marginally desirable regions. Highly desirable regions appearing lower in recommendation list would be penalized as the graded relevance value was reduced logarithmically proportional to the position of the result. The proposed method clearly outperformed the other baseline methods in terms of *nDCG* (Table 6) because it provided a more consistent recommendations order, especially among the highly desirable regions at the top of the list, in relation to ground truth, than baseline methods (Table 5). This, in turn, demonstrated the method's superiority over baseline methods in effectively meeting user needs in finding regions with high attractiveness. The top several regions recommended by our method were all located in the downtown, mainly in Tsim Sha Tsui, Central, Mong Kok, and Causeway Bay, which are the major shopping and recreation areas in Hong Kong. For example, among all 21 recommended regions, 5 were located

in Tsim Sha Tsui, 4 in Central, 2 in Mong Kok and 3 regions in Causeway Bay. Such highly concentrated spatial distribution of attractive regions, in some ways, indicated the high-density urban living environment of Hong Kong. In addition, the proposed method was effective and useful in finding regions with small area yet high desirability, which meets the need of users to find regions where venues of superior quality are located within a reasonably accessible distance. The top 3 recommendations (Star Ferry Pier Station, Canton Road, Tai Hang) by our method were all small-area regions with good quality venues attractive to users. In contrast, for the baseline methods, region size was given much weight in deciding recommendation priorities. This is because, for the baseline methods, region attractiveness was strongly positively correlated with numbers of visits, which inevitably related to region size. For example, the experiments revealed that regions such as Lyndhurst Terrace, Causeway Bay and Cameron Road were highly recommended by the baseline methods, as these were large regions with large numbers of venues, which added to the amounts of visits and visitors and eventually, priority for recommendation. However, such priority ranking failed to effectively reflect user satisfaction levels with the regions, as the service quality of the venues varied from item to item and the user average rating to the whole region might be compromised by unsatisfying venues. The proposed strategy, however, took account of both region size and user redundant visits so that region attractiveness was more accurately modelled.

Basically, to select and rank recommendations from a set of candidate regions, three particular aspects were considered by the proposed method. Firstly, our consideration was that larger regions tend to include more venues and consequently attract more users and visits, and such influence of region spatial area on the effectiveness of region recommendation needs to be modelled. Consequently, the proposed method included a new measurement (Equation 6) to quantify region attractiveness. Secondly, the influence of redundant visits by specific users was taken into account. The method assumption was that specific groups of users may have particular tastes and certain venues may be frequently visited by them, however, such personal preferences can hardly represent the general opinion of the total user pool. A decay effect was thus introduced into the model so that the single user influence decreased as the number of their visits increased. Thirdly, the interactive reinforcement effects were recognised in predicting the attractiveness of regions and the expertise of users. This is derived from the intuitive perception that the more a region is visited by experienced users, the more attractive the region is, and

equally, the more a user visits attractive regions, the more experienced the user is likely to be. The proposed recommending strategy aimed to distinguish between local experts and common users so that the attractive regions were more accurately identified.

5.2 Limitations and Future work

Detecting and recommending attractive regions with VGI data in this study has certain limitations. Firstly, the boundaries of the regions were manually defined to match the approximate geographical distribution of the check-ins, causing inaccurate prediction of region attractiveness and user expertise. Research to develop an improved region boundary outlining method would be of future benefit. Secondly, there can be a mismatch between check-ins and POIs. Theoretically, each check-in can be allocated to a POI. In practice, however, the check-in dataset and POIs dataset came from different platforms, and there was no accurate way of mapping check-ins to POIs. Since a mutual reinforcement process between users and POIs was used for inferring attractiveness and expertise, such mismatches might negatively affect the effectiveness in predicting the region attractiveness. Thirdly, using social media data to detect urban regions inevitably suffers from data bias, as social media users are a skewed sample from the whole population, mainly consisting of specific groups of people and young generation (Aslam 2018). How to quantify and alleviate the influence of data bias and improve the recommendation results could be a potential future research direction. Fourthly, as a result of the data availability, this research related to only one city and the generalization is limited. More cities, of different sizes can be analysed as data becomes available. Fifthly, a potential alternative to Algorithm 1 is to swap the first two lines: i.e. apply the spatial filter before the interest filter. Datasets with various scales can be used to evaluate which way is more efficient.

6. Conclusion

The widespread availabilities of location-awareness technologies and location-based social networks (LBSNs) enable researchers to study the urban structure from new perspectives. A primary motivation for this research was to develop an effective workflow to detect and recommend attractive regions with the awareness of answering the following questions: (1) how to investigate the geographic patterns associated with a specific topic embedded in social media content, under the conditions that short noisy texts are predominant and *a priori* knowledge is absent; (2) how to develop an appropriate scoring and ranking strategy which accurately recommends region, taking particular account of the effects of differing region sizes.

To do so, an unsupervised data-driven geographic topic modelling method was developed, making use of the self-explanatory information (i.e., hashtags) attached to social media content. Secondly, by taking the effects of spatial scale (region size) into account, a mutual reinforcement model was then developed to predict region attractiveness and user expertise. The new workflow was implemented with a real-world dataset and a framework introduced to evaluate the proposed methods at three different levels. The evaluation demonstrated the effectiveness of the newly proposed geographic topic modelling method and recommending strategy.

Our research examined the feasibility and effectiveness of taking advantage of the self-explanatory text information to investigate the issues of geographic topic modelling. Further, for region recommendation part, this study has demonstrated the need to take the effects of spatial scale into account so that the rating model can more accurately reflect the general satisfaction level of users over the whole region.

Acknowledgements
We would like to thank the editor Dr. Shawn Laffan and the anonymous reviewers for their insightful comments and substantial help on improving this article. We also thank Jittin Chaitamart for providing the valuable sample data and Dr. Zhang Xiaokang for helpful suggestions for improvement. This study is funded by The Hong Kong Polytechnic University [1-ZVF2, 1-ZEAB]

Funding
This work was supported by The Hong Kong Polytechnic University [1-ZVF2, 1-ZEAB]

References

Adams, B. and Janowicz, K., 2012. On the Geo-Indicativeness of Non-Georeferenced Text. *In*: J.G. Breslin, ed. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 4–7 June 2012 Dublin. Palo Alto, California: The AAAI Press, 375-378.

Aslam, S. , 2018*. Instagram by the Numbers: Stats, Demographics & Fun Facts* [online]. Available from: https://www.omnicoreagency.com/instagram-statistics/ [Accessed 21 July 2018]

Ayala, V. A. A.*, et al.*, 2017. A Delay-Robust Touristic Plan Recommendation Using Real-World Public Transportation Information. *In*: J. Neidhardt et al., eds. *2nd Workshop on Recommenders in Tourism*, 27 August 2017 Como, Italy. New York, NY, USA: ACM.

Bao, J., Zheng, Y. and Mokbel, M. F., 2012. Location-based and preference-aware recommendation using sparse geo-social networking data. *In*: I. Cruz and C. Knoblock,

eds. *Proceedings of the 20th international conference on advances in geographic information systems*, 07 – 09 November 2012 Redondo Beach, CA, USA. New York, NY, USA: ACM, 199-208.

Bao, J.*, et al.* 2015. Recommendations in location-based social networks: a survey. *GeoInformatica,* 19(3), 525-565.

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

Blondel, V. D.*, et al.* 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment,* 2008(10), P10008.

Cheng, T. and Wicks, T. 2014. Event detection using Twitter: a spatio-temporal approach. *PloS one,* 9(6), e97807.

Cranshaw, J.*, et al.*, 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. *In*: J.G. Breslin, ed. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 4–7 June 2012 Dublin. Palo Alto, California: The AAAI Press, 58-65.

Ester, M.*, et al.*, 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *In*: E. Simoudis et al., eds. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 2-4 August 1996 Portland, Oregon. Palo Alto, California: The AAAI Press, 226-231.

Feng, S., *et al.*, 2015. Personalized Ranking Metric Embedding for Next New POI Recommendation. *In*: Q. Yang and M. Wooldridge, eds*. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 25–31 July 2015 Buenos Aires, Argentina. Palo Alto, California: The AAAI Press, 2069-2075.

Gao, S.*, et al.* 2017. A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science,* 31(6), 1245-1271.

Ghosh, D. and Guha, R. 2013. What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science,* 40(2), 90-102.

Goodchild, M. F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal,* 69(4), 211-221.

Griesner, J.-B., Abdessalem, T. and Naacke, H., 2015. POI recommendation: towards fused matrix factorization with geographical and temporal influences. *In*: H. Werthner and M. Zanker, eds. *Proceedings of the 9th ACM Conference on Recommender Systems*, 16-20 September 2015 Vienna, Austria. New York, NY, USA: ACM , 301-304.

Huang, Q. 2016. Mining online footprints to predict user's next location. *International Journal of Geographical Information Science*, 31(3), 523-541.

Huang, Q. and Wong, D. W. 2015. Modeling and visualizing regular human mobility patterns with uncertainty: An example using Twitter data. *Annals of the Association of American Geographers,* 105(6), 1179-1197.

Kurashima, T.*, et al.*, 2010. Travel route recommendation using geotags in photo sharing sites. *In*: J. Huang, ed. *Proceedings of the 19th ACM international conference on Information*

and knowledge management, 26-30 October 2010 Toronto, ON, Canada. New York, NY, USA: ACM , 579-588.

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

Lee, I., Cai, G. and Lee, K. 2014. Exploration of geo-tagged photos through data mining approaches. *Expert Systems with Applications,* 41(2), 397-405.

Li, Z*., et al.*, 2007. Exploring LDA-Based Document Model for Geographic Information Retrieval. *In*: C. Peters et al., eds. *8th Workshop of the Cross-Language Evaluation Forum*, 19-21 September 2007 Budapest, Hungary. Berlin: Springer, 842-849.

Liu, B*., et al.*, 2013. Learning geographical preferences for point-of-interest recommendation. *In*: R. Chani et al., eds. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 11-14 August 2013 Chicago, IL, USA. New York, NY, USA: ACM, 1043-1051.

Long, X., Jin, L. and Joshi, J., 2012. Exploring trajectory-driven local geographic topics in foursquare. *In*: A. K. Dey ed. *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 5-8 September 2012 Pittsurgh, PA, USA. New York, NY, USA: ACM, 927-934.

Longley, P. A. and Adnan, M. 2016. Geo-temporal Twitter demographics. *International Journal of Geographical Information Science,* 30(2), 369-389.

Mikolov, T*., et al.* 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mislove, A*., et al.* 2011. Understanding the Demographics of Twitter Users. *In*: N. Nicolov and J. G. Shanahan eds. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 17-21 July 2011 Barcelona, Catalonia, Spain. Palo Alto, California: The AAAI Press, 554-557.

Newman, M. E. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences,* 103(23), 8577-8582.

Pan, C.-C. and Mitra, P., 2011. Event detection with spatial latent Dirichlet allocation. *In*: G. Newton ed. *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, 13-17 June 2011 Ottawa, ON, Canada. New York, NY, USA: ACM, 349-358.

Prateek, M. and Vasudeva, V., 2016. Improved topic models for social media via community detection using user interaction and content similarity. *In*: B. Novikov ed*., 2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, 28 August-4 September 2016 St. Petersburg, Russia. Washington, D.C., USA: IEEE , 1-7.

Pukelsheim, F. 1994. The three sigma rule. *The American Statistician,* 48(2), 88-91.

Rykov, Y., Nagornyy, O. and Koltsova, O. 2016. Semantic and Geospatial Mapping of Instagram Images in Saint-Petersburg. *In*: L. Pivovarova and T. Lando eds. *Proceeding of the AINL FRUCT 2016 conference,* 10-12 November 2016 St. Petersburg, Russia. Washington, D.C., USA: IEEE, 110 - 113

Stefanidis, A., Crooks, A. and Radzikowski, J. 2013. Harvesting ambient geospatial information from social media feeds. *GeoJournal,* 78(2), 319-338.

Steiger, E., Resch, B. and Zipf, A. 2015. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science,* 30(9), 1694-1716.

Sun, Y.*, et al.* 2015. Road-based travel recommendation using geo-tagged images. *Computers, Environment and Urban Systems,* 53, 110-122.

Tsou, M.-H.*, et al.* 2014. Mapping ideas from cyberspace to realspace: visualizing the spatial context of keywords from web page search results. *International Journal of Digital Earth,* 7(4), 316-335.

Wang, X., Gerber, M. S. and Brown, D. E., 2012. Automatic crime prediction using events extracted from twitter posts. *In*: S.J. Yang et al., eds. *Proceedings of the 5th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction*. 3-5 April 2012 College Part, MD, USA. Berlin: Springer, 231-238.

Yang, S. and Sklar, M., 2016. Detecting Trending Venues Using Foursquare's Data. *In*: I. Guy and A. Sharma eds. *Proceedings of the Poster Track of the 10th ACM Conference on Recommender Systems*, 17 September 2016 Boston, USA. New York, NY, USA: ACM.

Yang, S. and Sklar, M., 2018. *Finding the Perfect 10: How We Developed the Foursquare Venue Rating System* [online]. Acailable from: https://engineering.foursquare.com/finding-the-perfect-10-how-we-developed-the-foursquare-venue-rating-system-c76b08f7b9b3 [Accessed 9 June 2018].

Ye, M.*, et al.*, 2011. Exploiting geographical influence for collaborative point-of-interest recommendation. *In*: W. Ma and J. Nie eds. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 24-28 July 2011 Beijing, China. New York, NY, USA: ACM, 325-334.

Yin, Z.*, et al.*, 2011. Geographical topic discovery and comparison. *In*: S. Sadagopan et al., eds. *Proceedings of the 20th international conference on World wide web*, 28 March – 01 April 2011 Hyderabad, India. New York, NY, USA: ACM, 247-256.

Ying, Y., Chen, L. and Chen, G. 2017. A temporal-aware POI recommendation system using context-aware tensor decomposition and weighted HITS. *Neurocomputing,* 242, 195-205.

Yuan, Q.*, et al.*, 2013. Time-aware point-of-interest recommendation. *In*: G. Jones and P. Sheridan eds. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 28 July-01 August 2013 Dublin, Ireland. New York, NY, USA: ACM, 363-372.

Zhang, L., Sun, X. and Zhuge, H., 2013. Location-driven geographical topic discovery. *In*: Y. Pan et al., eds.*, 2013 Ninth International Conference on Semantics, Knowledge and Grids*, 3-4 October 2013 Beijing, China. Washington, D.C., USA: IEEE, 210-213.

Zheng, Y. and Xie, X. 2011. Learning travel recommendations from user-generated GPS traces. *ACM Transactions on Intelligent Systems and Technology (TIST),* 2(1), 2.

Zhou, H., Wang, P. and Li, H. 2012. Research on adaptive parameters determination in DBSCAN algorithm. *JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE,* 9(7), 1967-1973.