MACHINE-READING PLACES & SPACES

Generative Probabilistic Modelling of Urban Thematic Zones & Contexts

NUOZHI LIU¹ and IMMANUEL KOH²
¹,²Singapore University of Technology and Design.
¹nuozhi_liu@sutd.edu.sg, 0000-0002-3709-6822
²immanuel_koh@sutd.edu.sg, 0000-0002-1181-1082

Abstract. In this paper, a "place" is conceptualised as a composition of dynamic socioeconomic activities and collective perceptions. We apply generative probabilistic modelling to explore urban contextual semantics. By analogy to sorting documents into different topics, this research retrieves data embedding for each urban regions and classify them with thematic zones. Using Singapore as a case study, topic modelling is applied to retrieve perceptual and functional thematic zones from Instagram and TripAdvisor respectively. A subsequent analysis shows strong correlations among certain regions with functional and perceptual consistency. In addition, with our proposed uniqueness and diversity indices, a strong negative correlation at 0.82 is found, suggesting that a region could be more unique if the functions tend to be dominated by certain types of functional and perceptual thematic zones.

Keywords. Machine Learning; Natural Language Processing; Generative Probabilistic Models; Urban Data Modelling; Thematic Zones; Topic Modelling; SDG 11.

1. Introduction

In the city, a place is associated with a space that provides facilities in supporting socioeconomic activities. The sense of places in turn plays a critical role in affecting how people feel about these spaces. With rapid urbanization and globalization, urban planners need to acknowledge and be attuned to the diversity of places. More urgently, in view of the United Nations' Sustainable Development Goals such as SDG 11 (Sustainable cities and communities), urbanization approaches that recognize the diversity of places could significantly make cities more inclusive and participatory. Previous studies on place functions and human behaviours have provided insights on the features of places, and consequently, facilitate in the proper classification of urban regions. Discovering the functions of places enables planners to monitor change of land use (Gao et al., 2017) as well as to allocate resources accordingly (Yuan et al., 2012). However, despite prior observations of these functional characteristics, it remains unclear how places correlate to users' perceptions. For example, the perception of any
given restaurant in the city may differ according to its urban context (e.g., Central Business District (CBD) as compared to a school) despite it playing the same function of food and beverages. Hence, the purpose of this research is to study the relationship between space functions and human perceptions of urban regions. It aims to answer whether and how similar function of spaces can have an impact on people’s feeling about places. It also investigates how patterns of urban regions might be inferred from their urban contextual semantics. It demonstrates both functional and perceptual thematic zones beyond their existing homogeneously allocated administrative districts, in order to illustrate how a spatial region could be re-composed by socio-economic activities as well as human perceptions. The datasets used for this research are taken from both TripAdvisor and Instagram. By analogy to reading documents, this research employed topic modelling techniques to read a city. Both Latent Dirichlet allocation (LDA) and supervised LDA have been applied to the datasets in transforming the sparse urban data to dense informative data embedding. This paper contributes to the formulation of a proposed framework in extracting the latent structure of features from the regions of interests. The latent structure defines two types of thematic zones – a functional thematic zone for describing the physical characteristics and a perceptual thematic zone for exploring the human perceptions of the space. To measure the similarity among regions and the correlation between these two aspects of data, the research uses a diversity index and uniqueness index (as derived from Jensen-Shannon divergence). Both indices are meant for evaluating the urban-scale contexts by comparing regions of interests.

2. Background

The sense of place relating to place attachment, identity, and place making has been studied previously by others. Most frequently adopted methodologies by researchers are questionnaires, interviewing, and gathering feedbacks, and it could take a long period of time in collecting data (Lewicka, 2008; Savage et al., 2004; Shamai & Ilatov, 2005; Yuen, 2005). However, it is difficult to make use of the fully collected dataset, because the raw data typically consists of unstructured texts such as captions, comments, feedbacks, and open-ended questions. Furthermore, due to the diverse backgrounds of participants, summarising human perceptions from texts could also be challenging. Instead, this research introduces generative probabilistic modelling to synthesize such data in an unsupervised manner. Moreover, as the discussion of sense of places hitherto remains inconclusive (Cross, 2001; Hashemnezhad et al., 2013), this research aims to shed light on the field of urban studies via a statistical approach.

Another important task for urban studies is to detect and estimate the land use in Geographic Information System (GIS). It has been achieved by using census surveys, municipal records, and satellite imagery, etc. Meanwhile, with an unprecedented scale of information generated by users on the internet, emerging machine learning algorithms have brought new opportunities for understanding the built environment. For example, “social sensing” is proposed to discover socioeconomic environments through social network data by considering individuals as spatial detectors (Ali et al., 2011; Liu et al., 2015; Wang et al., 2019). One of the most popular techniques being LDA. It is a probabilistic topic model for uncovering the hidden thematic structures in large volume of documents, such as, books, news, articles, etc (Blei et al., 2003). While
others have extended it to urban studies by revealing spatial functions, incorporating geospatial data and location-based social media like Twitter and Foursquare (Gao et al., 2017; Ghahramani et al., 2021; Yuan et al., 2012), these studies have mainly focused on urban land use with limited attention paid to human crowd perception and without any synthesis of both. Therefore, our research attempts to evaluate and synthesize both urban functions and human perceptions via a quantitative approach that could complement further qualitative studies by applying variational inference of unstructured data at scale.

3. Methodology

3.1. RESEARCH DESIGN

The collected raw data are points of interests (POIs) which are geo-locations with attributes in texts. To aggregate attributes based on analysis unit, regions of interests are introduced. Similar to ‘Discovers Regions of different Functions’ or DRoF (Yuan et al., 2012) but simplified, a region of interest is a geographic region with elements where the elements are aggregated attributes of POIs inside. By defining the geographic regions, Singapore territory is subdivided into hexagons of 500 meters radius. As shown in Figure 1, each hexagon is regarded as a region of interest, aggregating POIs inside. If no POI is found, the region will be removed. LDA is then applied to these regions of interest in order to extract the latent structure of features. Each region will be categorised with a thematic zone based on its latent features. This procedure is applied to two datasets, namely Instagram for perceptual thematic zones and TripAdvisor for functional thematic zones. The correlation between function and perception in the same region can thus be analysed and measured.

3.2. DATA COLLECTION AND PRE-PROCESSING

Instagram is one of the major social media platforms in Singapore, as people use Instagram to document their life and interact with their family members and friends. Although Instagram is an image-based platform, the captions used constitute a directly
suggestive form of how people express themselves. A crawler was run to retrieve captions of Instagram posts from all locations in Singapore during the year of 2019. Text pre-processing, such as tokenization and lemmatisation, was then applied to these captions, splitting texts into words while removing tenses, plurality, etc. However, the “#” tags were kept during the process as they are representative of meaning rather than of punctuation in the context of Instagram.

For datasets that are representative of urban functions, we use TripAdvisor's place metadata. Compared to general Map service providers such Open Street Map, TripAdvisor captures more detailed features about a place's characteristics, such as types of cuisines, price ranges, etc. Furthermore, urban functions should also consider dynamic demands, rather than a static usage. For example, when considering how people visits functional places, this research collected passenger volume data of the most popular transport modes in Singapore, including Mass Rapid Transit (MRT) and Light Rail Transit (LRT). These volumes were grouped using transition cuboid into 24 time bins (12 for weekdays and 12 for weekends) (Yuan et al., 2012). Both TripAdvisor and traffic data are inputs to our supervised LDA to retrieve functional thematic zones.

3.3. TOPIC MODELLING

Probabilistic topic models are based on word frequencies. For instance, given two papers, one with high occurrence of the words "urban" and "plan", and the other with the words "probability" and "Markov chain", intuition will suggest that the former may be related to topics in urban planning, while the latter to topics in statistics. However, in practice, a paper could be a mixture of various topics in different proportions. For example, a paper related to generative urban design may mention the words "urban" and "probability", thus suggesting a mix of topics from urban planning and statistics. The definition of a topic of a paper, in such a context, thus refers to a distribution of topics instead of a single 'topic', and the goal of topic models is to capture this kind of intuition.

One of the most popular topic models is LDA. Given the words contained in a set of documents, LDA performs a generative process that infers the latent thematic structure through a posterior distribution. It considers each document as a distribution over topics, and each topic is a distribution over words. The generative process computes a joint probability distribution over both the observed and hidden random variables, where the documents and words are observed, and the topics are hidden (Blei, 2011; Blei et al., 2003). In extending this algorithm to urban studies, the methodology from Yuan at al. (2012) has been adapted for this research. Each region is considered a document, while the elements within the region liken to the words within the document. The elements are either keywords extracted from Instagram or functions from TripAdvisor's metadata as explained in Section 3.2. The hidden random variables are thematic zones. In addition, the vocabulary of the documents refers to processed attributes gathered from all POIs, or in other words, our metadata.

Thematic zones provide another perspective to see the relation between regions and zoning. Unlike how standard zoning in urban planning sees residential and commercial as disparate zones, a thematic zone mixes them in different proportions. For example, one thematic zone might have more residential areas with little commercial activities, while another might have more commercial functions but with few residential
buildings. These two thematic zones should not be simply regarded as either residential or commercial zones. Therefore, LDA is applied to uncover these thematic zones. As shown in Figure 2, the region belongs to certain thematic zone which is defined by mixing 3 themes such as foods, happy family, and fitness. Each theme refers to a specific subset of elements retrieved from the metadata. Such a definition of thematic zone will also have a greater capability in abstracting senses collectively, such as human perception which tends to be more complex as compared to functions.

The notations are defined as follows: \( K \) number of thematic zones are denoted as \( \beta_{1:K} \). Each \( \beta_k \) is a distribution over metadata. Proportion of theme \( k \) in the \( i \)th region is denoted as \( \theta_{i,k} \), while the theme of this region \( \theta_i \). Inferred theme of element of the region is denoted as \( z_{i,n} \), while \( w_{i,n} \) denotes the observed element. The value of \( n \) is the index of the element. The joint posterior distribution over the inferred and observed variables is described as:

\[
p(\beta_{1:K}, \theta_{1:K}, z_{1:N}, w_{1:N}) = \prod_{j=1}^{K} p(\beta_j) \prod_{i=1}^{P} p(\theta_i) (\prod_{n=1}^{N} p(z_{i,n} \mid \theta_i) p(w_{i,n} \mid \beta_{1:K}, z_{i,n}))
\]

This research employed supervised topic modelling to include the factor of traffic behaviours as part of consideration of functional thematic zones. A typical approach of supervised LDA is using authors as additional references to enhance the topic modelling of articles (Roberts et al., 2013). In contrast, this research uses transition cuboid as additional references to complement TripAdvisor dataset.
The number $K$ of thematic zones was chosen based on the recommended indicators, including residuals, semantic coherence and exclusivity. The goal is to find a $K$ number with a lower residual but with a relatively higher semantic coherence and exclusivity.

### 3.4. DIVERSITY INDEX AND UNIQUENESS INDEX

For the region of $i$, the output values of $\theta_{Ti}^{TA}$ and $\theta_{Ti}^{IG}$ respectively indicate the functional theme distribution and perceptual theme distribution. The research concatenates both $\theta$ values into one feature vector for each region. The similarity between regions of interest could be measured using the distance matrix. Figure 3. shows the selection of distance measures. It first considers the regions of three areas – (1) Marina Bay area in CBD, (2) Punggol neighbourhood area and (3) Changi airport area. These three areas are then considered as three distinctive areas from one another in terms of both functions and people activities. Both cosine distance and Jensen-Shannon Divergence (JSD) are capable in separating these areas, although JSD performed best in showing the nuances between regions within the same area.

Based on the $\theta$ values and JSD, this research introduces the **Uniqueness** and **Diversity** indices. Diversity index compares the difference within the regions, while uniqueness index compares the difference crossing regions with aggregation of JSD.

\[
\text{Uniqueness } U_i = \sum_{j=1}^{I} D_{ij}, j \in \{1, 2, \ldots , I\}, \forall i \text{, where, } D_{ij} = JSD(P \parallel Q)
\]

\[
P = (\theta_{i}^{TA_1}, \theta_{i}^{TA_2}, \ldots , \theta_{i}^{TA_K}, \theta_{i}^{IG_1}, \theta_{i}^{IG_2}, \ldots , \theta_{i}^{IG_K})
\]

\[
Q = (\theta_{j}^{TA_1}, \theta_{j}^{TA_2}, \ldots , \theta_{j}^{TA_K}, \theta_{j}^{IG_1}, \theta_{j}^{IG_2}, \ldots , \theta_{j}^{IG_K})
\]

\[
\text{Diversity } H_i = -\sum p_i \log p_i
\]

### 4. Results and Discussions

The model retrieved 30 thematic zones (prefix “TA”) for TripAdvisor, and 30 thematic zones (prefix “IG”) for Instagram. For the functional thematic zones, more differences have been observed. For example, as shown in Table 1, Zone TA-27 contains functions...
that are mostly bars and clubs, while TA-7 contains more varied activities, events, and leisure parks. Mid-range restaurant takes up the largest proportion for both Zone TA-2 and Zone TA-9. However, the former consists of more hotel types, while the latter more cuisine types.

Table 1. Top 10 ranked elements of functional thematic zones

<table>
<thead>
<tr>
<th>Theme</th>
<th>To 10 Ranked elements</th>
<th>Interpreted tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA-2</td>
<td>Brew Pub, Central Asian cuisine restaurant, Gastropub, Hotel (&gt; $300), Marina View Hotel, Mid-range restaurant, Other Outdoor Activities, Scandinavian cuisine restaurant, Swedish cuisine restaurant, Wine Bar</td>
<td>Mid-range restaurant</td>
</tr>
<tr>
<td>TA-7</td>
<td>Equipment Hire, Fun &amp; Games, Game &amp; Entertainment Centres, Gardens, Nature &amp; Parks, Outdoor Activities, Playgrounds, Room Escape Games, Sports Camps &amp; Clinics, Sports Complexes</td>
<td>Activities</td>
</tr>
<tr>
<td>TA-9</td>
<td>Cafe, Cheap restaurant, Chinese cuisine restaurant, Healthy, Italian cuisine restaurant, Medicinal foods, Mid-range restaurant, Singaporean cuisine restaurant, Soups, Trendy Hotel</td>
<td>Cafe &amp; Restaurant</td>
</tr>
<tr>
<td>TA-27</td>
<td>Bars &amp; Clubs, Dance Clubs &amp; Discos, Family Hotel, Fine-dining restaurant, French cuisine restaurant, Great View Hotel, Hotel (&gt; $300), Karaoke Bars, Nightlife, Wine Bars</td>
<td>Nightlife</td>
</tr>
</tbody>
</table>

As for perceptual thematic zones, many words are found frequently co-occurring in different thematic zones, but in varied contexts, such as, "time", "happy", "love", etc. For example, as shown in Figure 4, all 10 thematic zones mention the word "happy", with most zones mentioning the words "love" and "time" in parallel, while the nuances of contextual words serve to differentiate those zones. In addition, some thematic zones are found to be more ethnicity-related. In particular, IG-1 is dominated with Malay words, such as, "selamat" (congratulations), "kasih" (love), "hidup" (life), and "hati" (heart).
Pearson correlation was computed between functional and perceptual thematic zones (Figure 5). By way of cross comparing, only few correlations can be found between functional and perceptual zones. Relatively high positive correlation can be found for TA-6 (Indian cuisine) / IG-13 (Hawker food), and TA-4 (Sightseeing & Landmarks) / IG-30 (Sea). For the former pair, it shows hawker food tends to be relatively consistent in space functioning and human perception, especially in Indian Cuisine, and those thematic zones are near Little India, Chinatown and Toa Payoh. The latter pair are instead found near Sentosa, East Coast Park, and Marina Bay Sands, which are typically regarded as tourist destinations in Singapore. There are also some negative correlations. For instance, TA-11 (Japanese & French cuisine) and IG-10 (Family events) could less likely co-exist in functions and perceptions.

In terms of the uniqueness index, we found that the most unique two regions are those near the National and Hougang stadiums. Neighbourhoods are supposed to be generic in Singapore. Surprisingly, some of the neighbourhoods are the unique regions themselves, such as Telok Blangah. Yet, some regions are not as unique as people expected. For example, Changi airport shows relatively lower uniqueness, especially the area near the Jewel, an entertainment and shopping complex inside the Changi airport. The reason could be due to the presence of similar shops, department stores, landscape, playground. Figure 5. shows the distribution of normalised uniqueness index. By showing the relation between diversity index and uniqueness index for all
regions, a strongly negative correlation can be found at -0.82. It implies that if a region becomes more unique in both functions and perceptions, the region tends to be dominant by certain types of functional and perceptual thematic zones.

5. Conclusion

We have explored how one might 'read' the city in the form of urban thematic zones with probabilistic topic modelling. By appropriating natural language processing and topic modelling for urban studies, we propose that an urban thematic zone should be regarded as a mixture of different scenarios of certain proportions. By using Singapore as a case study, we have used TripAdvisor and Instagram datasets to retrieve functional and perceptual thematic zones respectively. A further correlation analysis found that some thematic zones are consistent in functions and perceptions. Our proposed uniqueness and diversity indices have also found a strongly negative correlation. We believe our contribution lies in a methodology that could be complementary to research on sense of places via a statistical perspective on urban functions and perceptions. It could enable urban planners and designers to evaluate urban regions located in any cities around the world with a vast amount of unstructured data using machine learning in "reading" places at scale. In addition, such a statistical view of "thematic zone" also reframes questions like "what makes a place a place", by seeing a place as a proportional mixture of tangible and intangible elements. It challenges the notion of a place beyond that of a static function, but reckons it as that which is constituted by its dynamic socioeconomic activities and collective perceptions. To address the limitation of our current research such as dataset biasness, future works would be further verified in a qualitative manner, alongside a more comprehensive selection of machine learning models to compare factors like probability model perplexity.

References


