Geo-spatial Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgsi20

Learning the spatial co-occurrence for browsing interests extraction of domain users on public map service platforms

Guangsheng Dong, Rui Li, Huayi Wu, Wei Huang, Hongping Zhang, Vincent Tao & Quan Liu

To cite this article: Guangsheng Dong, Rui Li, Huayi Wu, Wei Huang, Hongping Zhang, Vincent Tao & Quan Liu (2022): Learning the spatial co-occurrence for browsing interests extraction of domain users on public map service platforms, Geo-spatial Information Science, DOI: 10.1080/10095020.2022.2140078

To link to this article: https://doi.org/10.1080/10095020.2022.2140078

Published online: 15 Nov 2022.

GEO-SPATIAL INFORMATION SCIENCE
https://doi.org/10.1080/10095020.2022.2140078

Learning the spatial co-occurrence for browsing interests extraction of domain users on public map service platforms

Guangsheng Dong [a,b,c], Rui Li [a,b,c], Huayi Wu [a,b,c], Wei Huang [d], Hongping Zhang [d,e], Vincent Tao [f] and Quan Liu [f]

[a]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China; [b]Hubei Luojia Laboratory, Wuhan, China; [c]Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China; [d]National Geomatics Center of China, Beijing, China; [e]School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, China; [f]Shanghai Weizhizhuoxin Information Technology Co. Ltd, Shanghai, China

ABSTRACT
Public Map Service Platforms (PMSPs) provide embedded map services in domains such as forests and rivers. Users from different domains (Domain Users) prefer specific spatial features, and extracting the Browsing Interests of Domain Users (BIDUs) can help elucidate users' access intentions and provide suitable recommendations. Previous research has found that access frequency of spatial features is an indicator of users' browsing interests; however, high-frequency spatial features are sparsely distributed, resulting in inaccurate extraction of browsing interests. Our objective is to model the spatial co-occurrence of spatial features and employ BIDUs extraction to address this limitation. First, to extract spatial features in tiles, we proposed a k-nearest neighbor method for Point-of-Interest (POI) extraction and a template-based method for Land Uses/Land Covers extraction. Then, we developed the word2vec model to construct a POI semantic space to quantify spatial co-occurrence and employed multi-domain user classification to verify its effectiveness. Finally, a combined word2vec and singular value decomposition model is proposed to perform topic extraction as a representation of BIDUs. Compared with the baseline models, the proposed model integrates spatial co-occurrence from massive POIs to achieve high-accuracy BIDU extraction. Our findings can help construct domain user profiles and support the development of intelligent PMSPs.

ARTICLE HISTORY
Received 5 May 2022
Accepted 20 October 2022

KEYWORDS
Browsing interest extraction; spatial co-occurrence; Point-of-Interest (POI) semantic space; word2vec; Public Map Service Platform (PMSP)

1. Introduction

Public Map Service Platforms (PMSPs), such as Google Maps, OpenStreetMap, Baidu Map, Amap, and Tianditu, provide access to geographic information and significantly influence the daily lives of people globally (Dong et al. 2020; Li et al. 2019). Traditional PMSPs are dominated by portal websites. PMSPs have entered a new phase of development and are now embedded in applications in various domains such as ocean, forest, and river domains to support spatial analysis, natural resource management, and urban planning. The China Geospatial Information Industry Report stated that PMSPs reached one billion daily active users, and Tianditu, Baidu Map, and AMap supported nearly one million domain applications in 2021. Domain applications have become the fundamental service mode of PMSPs, and users visit specific domain applications (Domain Users) and thus have similar objectives as other users in the same domain.

The Browsing Interests of Domain Users (BIDUs) include explicit and implicit interests. The users' search content in a PMSP portal is an indicator of explicit interest, such as typing "airport" into the Google Maps portal to search for the location of an airport. However, it is not possible to input words for domain applications. PMSPs provide an embedded map for spatial visualization in which zooming in, zooming out, and panning are allowed map operations. In this case, BIDUs are implicit and hidden in the access content, including Points-of-Interest (POIs) and Land Use/Land Cover features (LULCs). For example, LULCs include built-up, agricultural, river, and forest regions while POIs include hospitals, banks, schools, and museums. In the popular COVID-19 map designed by Johns Hopkins Coronavirus Resource Center, cases are presented in the embedded map, and the purpose of users' visits is to understand the localization features of the epidemic at a given point in time. The Internet hosts many domain applications; however, there is no research on BIDU extraction, resulting in PMSP providers being ignorant of user requirements. This study aimed to address this lack of research and extract implicit BIDUs to help PMSP providers understand domain user requirements and develop personalized and intelligent PMSPs (Dong et al. 2022).

BIDUs have common access content and spatial features. Current research on browsing interest

CONTACT Rui Li ruili@whu.edu.cn

extraction is summarized as either individual- or group-oriented. In individual-oriented studies, scholars collect users' vision and touch information using eye-tracking technology to determine browsing interests. However, such experiments are costly, and most have been conducted on small samples. In group-oriented studies, scholars have considered access frequency as an indicator of BIDUs from the perspective of time, space, and spatial features. The higher the access frequency, the greater the browsing interest. However, popular spatial features are sparsely distributed, and it is impossible to accurately extract BIDUs by considering only access frequency (Wang et al. 2014).

Learning the spatial co-occurrence of spatial features is key to addressing the limitation of spatial sparsity (Cheng et al. 2014). For example, restaurants tend to appear around shopping malls while bookstores tend to appear around schools. This study was the first attempt to extract BIDUs based on spatial co-occurrence. The access content for common web services, such as social media (Zheng, Ge, and Wang 2019), news (Li et al. 2011), and shopping (Yu et al. 2021), is text. Scholars have used the topic model to extract topics as users' browsing interests based on word co-occurrence, thereby providing a reference for our research. However, the dimension of spatial features is higher than that of text sequences, and quantifying the spatial co-occurrence of spatial features is the key to extracting BIDUs accurately.

The objective of this study is to extract the BIDU topics based on the spatial co-occurrence of visited spatial features. The research questions are summarized as follows:

(1) How can we accurately extract spatial features from visited tiles?
(2) How can spatial features be quantified by considering spatial co-occurrence? How can the accuracy of spatial feature quantification be evaluated?
(3) How can BIDU topics be extracted based on spatial co-occurrence? How can the accuracy of topic extraction be evaluated?

To solve the above research problems, we propose the following solutions, which are the primary contributions of this study.

(1) Spatial feature extraction for domain users. BIDUs are reflected by the spatial features of the tiles, which are composed of LULCs in the vector tiles and POIs in the annotation tiles. We proposed a K-Nearest Neighbor (KNN) method for POI extraction and a template-based method for LULC extraction.
(2) POI semantic space construction to model the spatial co-occurrence of POIs. We used the word2vec model to quantify the spatial co-occurrence distribution of 65.27 million POIs in China and construct the POI semantic space based on Tobler's First Law of Geography. The POI semantic space was considered suitable for quantifying visited POIs. We also evaluated the quantization accuracy by multi-domain user classification experiments. The results indicate that the POI semantic space can achieve a higher classification accuracy for domain users compared with the baseline models.
(3) BIDU extraction by topic. We proposed the W2 V-SVD model, a combined model of the Word2vec and Singular Value Decomposition model (SVD), to perform BIDU extraction of topics based on the user – POI matrix, which was constructed using the results of (1) and (2). Compared with the traditional Latent Dirichlet Allocation model (LDA), the proposed model integrates the spatial co-occurrence of massive POIs to achieve a higher consistency of BIDUs extraction.

The remainder of this paper is organized as follows. In Section 2, we present related work. In Section 3, we describe the proposed methods for BIDU extraction. In Section 4, we discuss experiments that validate the effectiveness of the proposed model. Section 5 concludes the study and suggests directions for future research.

2. Related work

2.1. Browsing interest extraction on PMSPs

(1) Individual-oriented browsing interest extraction

Researchers have determined indicators of individual-oriented browsing interests from vision- (Unrau and Kray 2019) and touch-based interactions (Manson et al. 2012). Eye trackers and data acquisition systems allow us to determine where, when, and how long an individual's visual attention is directed toward an object, which could elucidate vision-based browsing interests (Krafka et al. 2016). When a user interacts with a PMSP using a mouse or their fingers, touch-based operations, such as the number of clicks and duration of cursor placement, can identify browsing interests (Dong et al. 2019). However, these experiments require volunteer recruitment and a professional system for data collection. Thus, their costs are very high, and only small samples can be used for data collection. Such methods cannot be applied to PMSPs with a large number of users.

(2) Group-oriented browsing interest extraction

Many group-oriented studies are based on spatio-temporal modeling (Chen et al. 2020; Dlamini et al. 2021). Scholars have modeled the access frequency of content by considering temporal and spatial features using massive access logs. Higher access frequency is correlated with greater browsing interest. (Li et al. 2018) found that the time series of user access frequency is periodic and conforms to the rhythms of work and rest. (Fisher 2007) developed a hotmap based on the access frequency of tiles. (Quinn and Gahegan 2010) indicated that frequently accessed tiles cover popular POIs such as highways, coastlines, parks, and perennial tourist attractions. García quantitatively modeled the relationship between spatial features and their access frequency using OLS (García Martín et al. 2013) and ANN models (García et al. 2013).

At present, research on BIDU extraction from PMSPs is still in its infancy. The access frequency of tiles or spatial features is an indicator for extracting BIDUs, that is, the higher the access frequency, the greater the browsing interest. This approach is both concise and intuitive. However, popular spatial features or tiles are sparsely distributed, making it difficult to extract BIDUs accurately. Users in different domains prefer specific spatial features. Currently, the domain user types interested in specific spatial feature types remain unknown.

2.2. Browsing interests extraction on common web services

Text is the main access content on common web services, such as social media (Zheng, Ge, and Wang 2019), news (Li et al. 2011), and shopping (Yu et al. 2021). Scholars have employed text mining to extract topics of browsing interest (Sharma, Kumar, and Chand 2017). Topic models include bag-of-words and word-embedding models. Bag-of-words models include Latent Semantic Analysis (LSA) (Dumais 2004), probabilistic LSA (pLSA) (Hofmann 2001), and LDA models (Blei, Ng, and Jordan 2003). The LDA model has remained the mainstream topic model for browsing interest extraction in recent years (Tontodimamma et al. 2021; Sutherland and Kiatkawsin 2020; Seo and Cho 2021; Jung and Yoon 2020). With the development of deep learning (Chen et al. 2022), a word-embedding model has also been employed. The most popular word-embedding model is Google's word2vec (Mikolov et al. 2013), which exhibits advantages in measuring the semantic similarity between words and discovering potential relationships between concepts.

Topic interests comprise an important form of users' browsing interests; however, we did not find any research on topic extraction on PMSPs. The access contents on a PMSP are tiles as well as spatial features (POIs and LULCs) covered in the tiles. Its two-dimensional spatial distribution is more complex than that of a one-dimensional word sequence. Therefore, quantifying the spatial features of tiles is a key problem in current research. In recent years, using the word embedding model as a foundation, researchers have proposed Poi2vec (Feng et al. 2017), Plcace2vec (Zhai et al. 2019; Yan et al. 2017), and Location2vec (Zhu et al. 2019) to quantify spatial features, which provide references for our research. We defined the problem of BIDUs extraction as extracting commonly visited spatial features in domain applications. We determined that topic extraction could address this problem and provide a new perspective and method for research on browsing interest extraction from PMSPs.

3. Methodology

3.1. Framework

PMSPs, which are public-oriented platforms, provide embedded map services and integrate with applications from different domains. For example, an embedded PMSP supports maritime weather publishing in the maritime domain and river monitoring in river management. Users from different domains access the PMSP for different purposes, while users in the same domain have similar browsing interests.

The users' access process is driven by their interests through map operations such as zooming and panning (Dong et al. 2022). For example, when staff needs to query the location and length of specific rivers, they zoom into the map to find River-A, pan along the river, and then zoom out to evaluate River-B by a similar process. In this process, we cannot directly capture the user's potential interest from access behavior. The commonly visited spatial features represent the preferences of domain users, which are embodied in visited LULCs and POIs. Therefore, our framework is proposed to achieve BIDUs extraction in Figure 1, which is detailed in Sections 3.2, 3.3, and 3.4.

3.2. POI semantic space construction to model the spatial co-occurrence of POIs

3.2.1 POI spatial corpus

In linguistics, a corpus refers to a language resource that consists of a large and structured set of texts. A corpus contains many documents, each of which is composed of sentences and words. The sequence of sentences and words in the documents indicates context. A dictionary can be generated based on all words in the corpus. Analogous to the organization of the text, scholars have converted the spatial distribution of POIs into the text to generate a POI corpus in which the POI types as words in the dictionary. The key

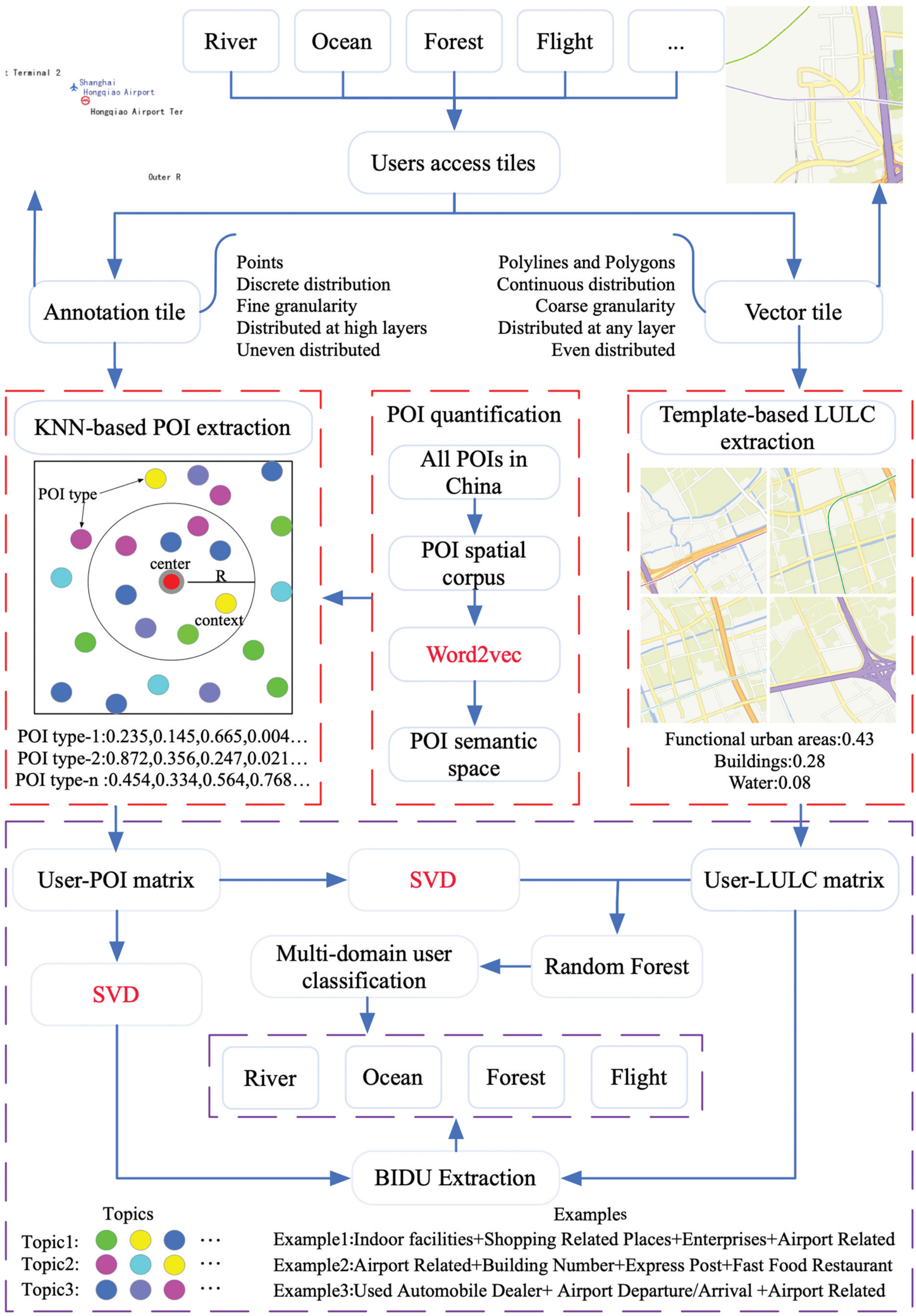

Figure 1. BIDU extraction framework.

process underlying this approach is the transformation of two-dimensional distributed POIs into a one-dimensional sequence considering spatial co-occurrence. Tobler's First Law of Geography states that "everything is related to everything else, but near things are more related than distant things" (Tobler 1970). Therefore, the spatial context of a POI can be expressed using the nearest POIs, which is the theoretical basis for constructing the POI corpus.

Some scholars have built a POI corpus in one city or multiple cities to study their structures. However, a larger POI corpus, such as that at the national scale, has not yet been constructed. Because the spatial range of users accessing PMSPs spans the entire country, we constructed a nation-scale POI corpus using all POIs in China.

Considering the large spatial range and amount of POI data, we adopted a simple and fast method to construct the POI corpus. Each POI document comprises $(POI_{center}, POI_{context})$, where POI_{center} indicates the coordinates of the center POI. We constructed a buffer with a radius of R to retrieve the covered k nearest POIs as $POI_{context}=\{POI_1, I_2, \cdots, POI_k\}$, where $\mathrm{dis}(POI_{center}, POI_{\lambda-1}) < \mathrm{dis}(POI_{center}, POI_{\lambda}), 1 < \lambda \leq k$, and $\mathrm{dis}(POI_{center}, POI_{\lambda})$ indicates the Euclidean distance between POI_{center} and POI_{λ}. We built $POI^{\theta}_{context}$ in three levels, where $\theta=1, 2, 3$. The POI type of each level was used as the dictionary, and their sizes in the $POI^1_{context}$, $POI^2_{context}$, $POI^3_{context}$ were 24, 268, and 899, respectively.

3.2.2 Word2vec model for POI semantic space construction

Word2vec, a word-embedding model proposed by Google in 2013 (Mikolov et al. 2013), uses a neural network to learn word associations from a large corpus. The model maps each word to a high-dimensional vector based on its contextual content. The word vectors retain the semantic information of the words, such that the cosine similarity between the vectors can measure the semantic similarity of words. Compared to other embedding models, word2vec is among the most widely used in spatial modeling because of its stability and efficiency; thus, we used this model to construct the POI semantic space. The word2vec model provides two training approaches: skip-gram and Continuous Bag-of-Words (CBOW). Compared with the skip-gram model, the continuous input and training process of CBOW can better reflect the context relationships that characterize words (Yao et al. 2017). Therefore, in this study, a CBOW-based word2vec model was adopted to extract POI vectors.

We assumed that the size of the POI spatial corpus at level θ is H, the sampling window of the context of POI_h is c, and the maximum likelihood estimation of the word2vec model can be expressed as $\frac{1}{H}\sum_{h=1}^{H}\log\rho(POI_h|POI^{h+c}_{h-c})$, where POI^{h+c}_{h-c} represents using POI_h as the center and c as the sampling window to construct the POI context. $\rho(POI_h|POI^{h+c}_{h-c})$ is defined as

$$\rho(POI_h|POI^{h+c}_{h-c}) = \frac{\exp(-\mathrm{E}(POI_h|POI^{h+c}_{h-c}))}{\sum_{i=1}^{H}\exp(-\mathrm{E}(POI_i|POI^{h+c}_{h-c}))}, \quad (1)$$

where E is an energy function, and $\mathrm{E}(POI_i, POI_j) = -(POI_i POI_j)$ (Yao et al. 2017). equation (1) indicates the occurrence probability of POI_h when the current context is c. The POI spatial corpus was composed of three-level corpora, and a word2vec model was built for each level to construct the POI semantic space.

3.3 Spatial feature extraction for individuals

To distinguish the access behavior in different sessions, we divided the access process by the time threshold. If the interval between two records in a user's access process exceeds a threshold, the process is divided into two sessions.

The user access process is characterized by changes in a user's browsing interest. Zooming-in indicates increasing interest, whereas zooming-out indicates decreasing interest; the maximum interest is assigned to the browsing target. To understand the user's browsing interests, we extracted browsing targets using the HGMM-RF model (Dong et al. 2022) and then retrieved the spatial features around the targets. The spatial information provided by the Web Map Tile Service (WMTS) is presented in tiles, including LULCs (polyline/polygon) in vector tiles and POIs (points) in annotation tiles, as shown in Figure 2. The differences between the vector and annotation tiles are listed in Table 1. The information on LULCs is evenly distributed in each tile; however, there are few information types, resulting in low information density. The POI distribution was spatially uneven. For example, they are dense in urban areas and sparse in rural areas. POI types are classified into fine granularity with a high information density. The POIs in the annotation tiles and LULCs in the vector tiles are complementary, and their combination can express users' browsing interests. We separately extracted POIs and LULCs around the browsing targets.

3.3.1 KNN – based POI extraction

A session contains more than one browsing target, and a single target cannot reflect the user's browsing interest. Therefore, we used all browsing targets in a session to express users' browsing interests. KNN is a classic classification method (Djenouri et al. 2019); however, in this study, KNN indicates a method for spatial retrieval. Similar to the method for constructing a POI corpus, we traversed all the targets in each session to establish a buffer with a radius R, retrieved

Figure 2. Annotation and vector tiles on a PMSP.

Table 1. Differences between a vector tile and an annotation tile.

annotation tile	vector tile
POIs	LULCs
Points	Polylines and Polygons
Discrete distribution	Continuous distribution
Fine granularity	Coarse granularity
Distributed at high layers	Distributed at any layer
Uneven distributed in space	Even distributed in space

the k nearest POIs, and used the POI types as the spatial semantic context of each target. We constructed browsing targets with three levels of POI types respectively, $\boldsymbol{POI}_q^{\theta}=\{\boldsymbol{POI_1}, \boldsymbol{POI_2}, \cdots, \boldsymbol{POI_s}\}$, where s is the number of POIs in each session. We converted all POIs in the session into a matrix $\boldsymbol{POI}_q^{\theta}$ based on the POI semantic space as follows

$$\boldsymbol{POI}_q^{\theta}=\begin{bmatrix} P_{11} & \cdots & P_{1\beta} \\ \vdots & \ddots & \vdots \\ P_{\alpha 1} & \cdots & P_{\alpha\beta} \end{bmatrix}, \tag{2}$$

where β is the hyperparameter of the word2vec model and represents the dimension of the output vector.

3.3.2 Template-based LULC extraction

We extracted LULCs from the vector tiles. Vector tiles are images generated from vector data, and different LULCs are distinguished by color. Referring to "web map symbols and annotations in Tianditu," we established a color template for LULC extraction in vector tiles with manual identification, as shown in Table 2.

The template includes LULC types such as Road, Water, Green land, Land, and Residential area. Each LULC type has a layer range for display. For example, green land appears from layer 11 to 18, while buildings appear from layer 16 to 18. Each LULC type is represented by one or multiple colors, and the colors of the same LULC type in different layers may be different. For example, the County road has three displayed colors from layer 11 to 18.

LULC colors are composed of RGB values. To eliminate the impact of subtle color differences in the tiles, an interval of ± 2 was adopted for R, G, and B when performing LULC recognition. The results extracted by the template-based method are expressed as

$$\boldsymbol{LULC}_q^{\theta}=\begin{bmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mn} \end{bmatrix}, \sum_{e=1}^{n} f_{mn}(e) \leq 1, \tag{3}$$

where $\boldsymbol{LULC}_q^{\theta}$ is a matrix composed of m tiles in a session, and each tile is an n-dimensional LULC vector. $f_{mn}(e)=\frac{pixel(e)}{256*256}$, where $pixel(e)$ represents the

Table 2. The color-based template for LULCs extraction.

LULC types		layer	color R±2	G±2	B±2
Road	Freeways	11-18	186	160	241
	National road	11-14	254	205	120
		15-18	254	205	110
	Province road	11-14	254	240	158
		15	254	240	130
		16-18	254	235	130
	County road	9-14	255	244	175
		15	255	244	140
		16-18	254	240	145
	Other road	12	223	223	215
		13-18	253	254	255
Water		1-18	171	198	239
Green land		11-18	187	215	141
Land		1-18	245	244	238
Residential area	Functional urban areas (universities, shopping malls, hospitals, industrial areas, parking lots)	14-18	236	238	203
	Buildings	16	249	250	243
		17-18	249	250	254

pixel number of a specific LULC in a tile and the total pixel number in a tile is 256×256. Therefore, $f_{mn}(e)$ represents the proportion of pixels in a specific LULC. Because some pixels were not in the color range of the template, $\sum_{e=1}^{n} f_{mn}(e) \leq 1$.

3.3.3 User – interest matrix construction

We extracted the POIs and LULCs around the browsing targets at the session-scale. Because the number of browsing targets in each session is different, the POIs and LULCs in each session are also different. To convert POIs into a fixed-size matrix that can be fed to the model, we averaged the POI vectors in the session and obtained the ω_q^θ as

$$\omega_q^\theta = \frac{1}{s}\left[\sum_{d=1}^{s} P_1(d), \sum_{d=1}^{s} P_2(d), \cdots \sum_{d=1}^{s} P_\beta(d)\right]. \quad (4)$$

We also calculated the average value of each LULC to obtain the vector representation σ_q^θ in a session, as follows

$$\sigma_q^\theta = \frac{1}{s}\left[\sum_{d=1}^{s} f_1(d), \sum_{d=1}^{s} f_2(d), \cdots \sum_{d=1}^{s} f_n(d)\right]. \quad (5)$$

Then, we constructed the user – POI matrix $\boldsymbol{A}$,

$$\boldsymbol{A} = \left[\omega_1^\theta, \omega_2^\theta, \cdots \omega_q^\theta, \cdots \omega_u^\theta\right]^{\mathrm{T}}, \quad (6)$$

and user – LULC matrix $\boldsymbol{B}$ of the session,

$$\boldsymbol{B} = \left[\sigma_1, \sigma_2, \cdots \sigma_q, \cdots \sigma_u\right]^{\mathrm{T}}, \quad (7)$$

where $1 < q < u$, u indicates the number of browsing sessions in the data.

3.4. BIDU extraction using the proposed W2 V-SVD model

Users within a given domain have similar browsing interests. We constructed a classification model to classify multi-domain users based on POI dimension for validating the accuracy of spatial feature quantification. Random Forest (RF) is a classic classification model with high accuracy, robustness, and interpretability in many applications (Breiman 2001). Therefore, we used an RF model for multi-domain user classification. Then, we proposed the W2 V-SVD model to extract BIDUs from the user dimension. The flowchart of this process is shown in Figure 3.

3.4.1 Multi-domain user classification to validate spatial feature quantification

The representation of browsing interests includes the user – POI matrix $\boldsymbol{A}$ and user – LULC matrix $\boldsymbol{B}$. The dimension of the POI vector is a hyperparameter of the word2vec model. The POI spatial corpus was large in this study; thus, we needed a high-dimensional vector to train the model. Multi-domain user classification experiments require the fusion of POIs and LULCs. The POI vector dimension is 400, whereas the LULC dimension is only 10. When the difference in their dimensions is large, the weight of the LULC is small. Therefore, we applied the SVD model to achieve dimensionality reduction of the POI vectors in $\boldsymbol{A}$ and facilitate the fusion of the POI and LULCs. We rewrote $\boldsymbol{A}$ as

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & \cdots & A_{1\beta} \\ \vdots & \ddots & \vdots \\ A_{u1} & \cdots & A_{u\beta} \end{bmatrix}. \quad (8)$$

For the nonzero real matrix $\boldsymbol{A}$, $\boldsymbol{A} \in R^{u \times \beta}$, the SVD decomposes $\boldsymbol{A}$ into three real matrices

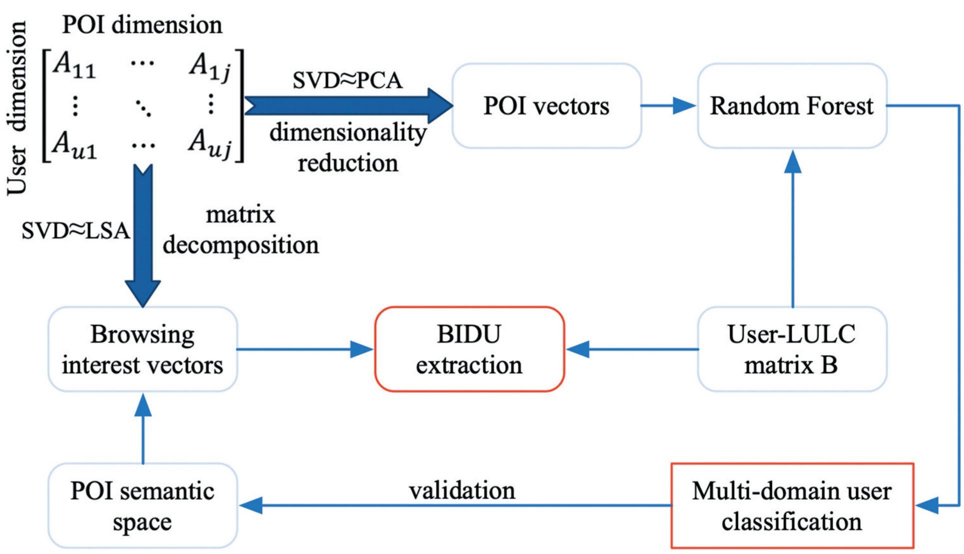

Figure 3. W2V-SVD model used for multi-domain user classification and BIDU extraction.

$$A = U\Sigma V^{\mathrm{T}}, \tag{9}$$

where U is an orthogonal matrix of order u, V is an orthogonal matrix of order β, Σ is a diagonal matrix composed of non-negative elements arranged in descending order, $UU^{\mathrm{T}} = I$, $VV^{\mathrm{T}} = I$, $\Sigma = \mathrm{diag}(\varphi_1, \varphi_2, \cdots, \varphi_{\mathrm{r}})$, $\varphi_1 \geq \varphi_2 \geq \cdots \geq \varphi_r \geq 0$, $r = \min(\mathrm{q}, \beta)$. φ_r is the singular value of A, the column of U is the left singular vector, and the column vector of V is the right singular vector. Truncated SVD is commonly used to improve SVD efficiency. The singular values are sorted in descending order. Only the t column vectors of U_t and t row vectors of V_t corresponding to t singular values Σ_t are retained. The rest of the matrix is discarded to obtain the approximate decomposition $\widehat{A}$, as follows

$$A \approx U_t \Sigma_t {V_t}^{\mathrm{T}} = \widehat{A}. \tag{10}$$

In this study, the SVD refers to a truncated SVD. SVD transforms high-dimensional A into low-dimensional vectors, which is equivalent to Principal Component Analysis (PCA) (Wall, Rechtsteiner, and Rocha 2003), where t represents the number of principal components in PCA and ${V_t}^{\mathrm{T}}$ represents the principal components. In our experiments, t=30. We used ${V_t}^{\mathrm{T}}$ and B as independent variables to construct an RF model to achieve multi-domain user classification. When the accuracy of the multi-domain user classification is higher, the quantification of the spatial features is more accurate, and the representation of the BIDUs is more reasonable.

3.4.2 BIDU extraction by topic

BIDUs can be represented as topics to yield a general understanding of user interests. We propose a W2 V-SVD model to extract BIDUs. The essence of the POI semantic space constructed based on the word2-vec model is the spatial co-occurrence. When the distance between any two POI vectors is small, their co-occurrence probability is high. We decomposed user – POI matrix A based on the SVD model from the user dimension. In the decomposition results, U_t represents the user – topic matrix, ${V_t}^{\mathrm{T}}$ represents the topic – POI matrix, and t represents the number of topics.

The W2 V-SVD model is similar to the LSA model. In the LSA model (Schütze, Manning, and Raghavan 2008), A is generated by the one-hot model or the Term Frequency – Inverse Document Frequency (TFIDF) model (Guo and Yang 2016), where the POI vectors represent the frequency of visited POIs, and the corresponding ${V_t}^{\mathrm{T}}$ represents the topic-POI matrix. In the W2 V-SVD model, A consists of POI vectors calculated using the W2 V model (abbreviation of the word2vec model). The POI vectors constructed using the word2vec model were semantically computable. A well-known example is the vector calculation, "King – Man + Woman = Queen" (Drozd, Gladkova, and Matsuoka 2016). Thus, the vectors in the ${V_t}^{\mathrm{T}}$ indicate the topics of browsing interest in the POI semantic space. However, the browsing interests generated by the W2 V-SVD model were incomprehensible. We needed to explain their semantics approximately using existing POIs. We calculated the distance $\mathrm{D}(v, P)$ between each topic vector v in ${V_t}^{\mathrm{T}}$ and all POI types P using cosine similarity, where $\mathrm{D}(v, P) \in [-1, 1]$. When $|\mathrm{D}(v, P)|$ is larger, P is closer to the topic vector v. We selected the k-nearest POIs as the composition of $\mathrm{Topic}(v)$ as

$$\begin{gathered} \mathrm{Topic}(v) = \{P_1, P_2, \cdots, P_k\}, \\ \mathrm{D}(v, P_1) > \mathrm{D}(v, P_2) > \cdots > \mathrm{D}(v, P_k), \end{gathered} \tag{11}$$

and the Figure 4 shows an example of topic generation.

The POI vectors represent the spatial co-occurrence of POIs in China. User browsing interests were mapped into the POI semantic space to enhance semantic information. Using the k-nearest POIs as $\mathrm{Topic}(v)$ expanded the semantic expression of BIDUs. We found similar POIs with spatial co-occurrence from all potential POIs, which overcame the limitation of extracting BIDUs solely from sparsely distributed visited POIs.

3.5. Pseudo-code for BIDU extraction

```
Input: domain, user, POISet
Output: topics;
Main Function (){
   POISemanticSpace= CreatePOISemanticSpace (POISet);
   For each domain:
        For each user:
             For each session:
                  annotation_tiles, vector_tiles=CollectTiles (user);
                  visited_POIs= KNN (annotation_tiles);
                  visited_POI_vectors=ConvertPOI2Vectors (visited_POIs,
                   POISemanticSpace);
                  visited_LULCs=Template (vector_tiles);
             ω=average (visited_POI_vectors);
             σ= average (visited_LULCs);
        A=ConvertUserMatrix (ω);
        B=ConvertUserMatrix (σ);
        topics=ExtractBIDU (A, POISemanticSpace);
   Random Forest (×=[(A+B)/A/B],Y=domain);
   }
Sub Function CreatePOISemanticSpace (POISet){
   POI_spatial_corpus=KNN (POISet);
   POI_semantic_space=Word2vec (POI_spatial_corpus);
   return POI_semantic_space;
}
Sub Function EctractBIDU (A, POISemanticSpace){
   TopicVectors=SVD (A);
   For each TopicVectors:
        topics=Topic (TopicVectors, POISemanticSpace);
   return topics;
   }
```

$$Topic(v) = Shopping\ Related\ Places + Photo\ Finishing + bakery + \text{Personal Care Items Shop} + \text{Clothing Store}$$

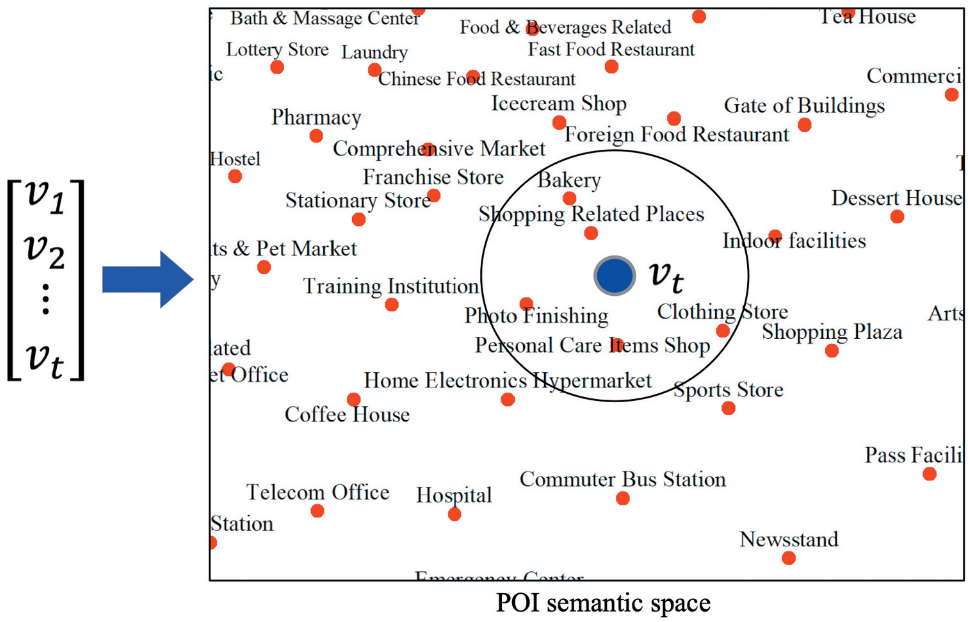

Figure 4. An example of topic generation.

4. Experiments and analysis

4.1. Data source and experimental settings

4.1.1 Nationwide POI data in China from Amap

The POI is an important data source that represents human spatial behavior (Sari Aslam et al. 2021; Mei et al. 2022). In total, 65.27 million POIs in China were collected by Amap in 2018. The POI types included three levels subdivided from the first to the third level, with 24, 268, and 899 types for the first, second, and third levels, respectively (https://lbs.amap.com/api/webservice/download). Table 3 presents some examples of the POI types used in this research.

4.1.2 User access logs from Tianditu

The data used in this research were collected from the user access logs of Tianditu, which is a networked geographic information-sharing and service portal. Tianditu was built by the National Geomatics Center of China to provide integrated geographic information services (https://www.tianditu.gov.cn/). WMTS is one of the main services provided by Tianditu, with nearly

Table 3. Examples of the POI types used in this study.

First level	Second level	Third level
Auto Dealers	Toyota Franchised Sales	Toyota Sales
Auto Repair	Chrysler Franchised Repair	Jeep Repair
Motorcycle Service	Motorcycle Sales	BMW Motorcycle Sales
Food & Beverages	Chinese Food Restaurant	Shanghai Food
Shopping	Supermarket	Wal-Mart
Daily Life Service	Information Centre	Enquire of Hotel
Sports & Recreation	Sports Stadium	Gym Center
Medical Service	Special Hospital	Special Hospital
Accommodation Service	Hotel	Five-star Hotel
Tourist Attraction	Park & Square	Park
Commercial House	Building	Commercial-residential Building
Governmental Organization & Social Group	Governmental Organization	State Level Organization & Institution
Science/Culture & Education Service	Media Organization	TV Station
Culture & Education	School	Facilities within the School
Transportation Service	Parking Lot	Public Parking Lot
Finance & Insurance Service	ATM	Bank of China ATM
Enterprises	Company	Network Science and Technology
Road Furniture	Warning Sign	Camera
Place Name & Address	Normal Place Name	Country Name
Public Facility	Public Toilet	Public Toilet
Incidents and Events	Public Event	Conference
Indoor facilities	Indoor facilities	Indoor facilities
Pass Facilities	Gate of Buildings	Main Gate of Buildings

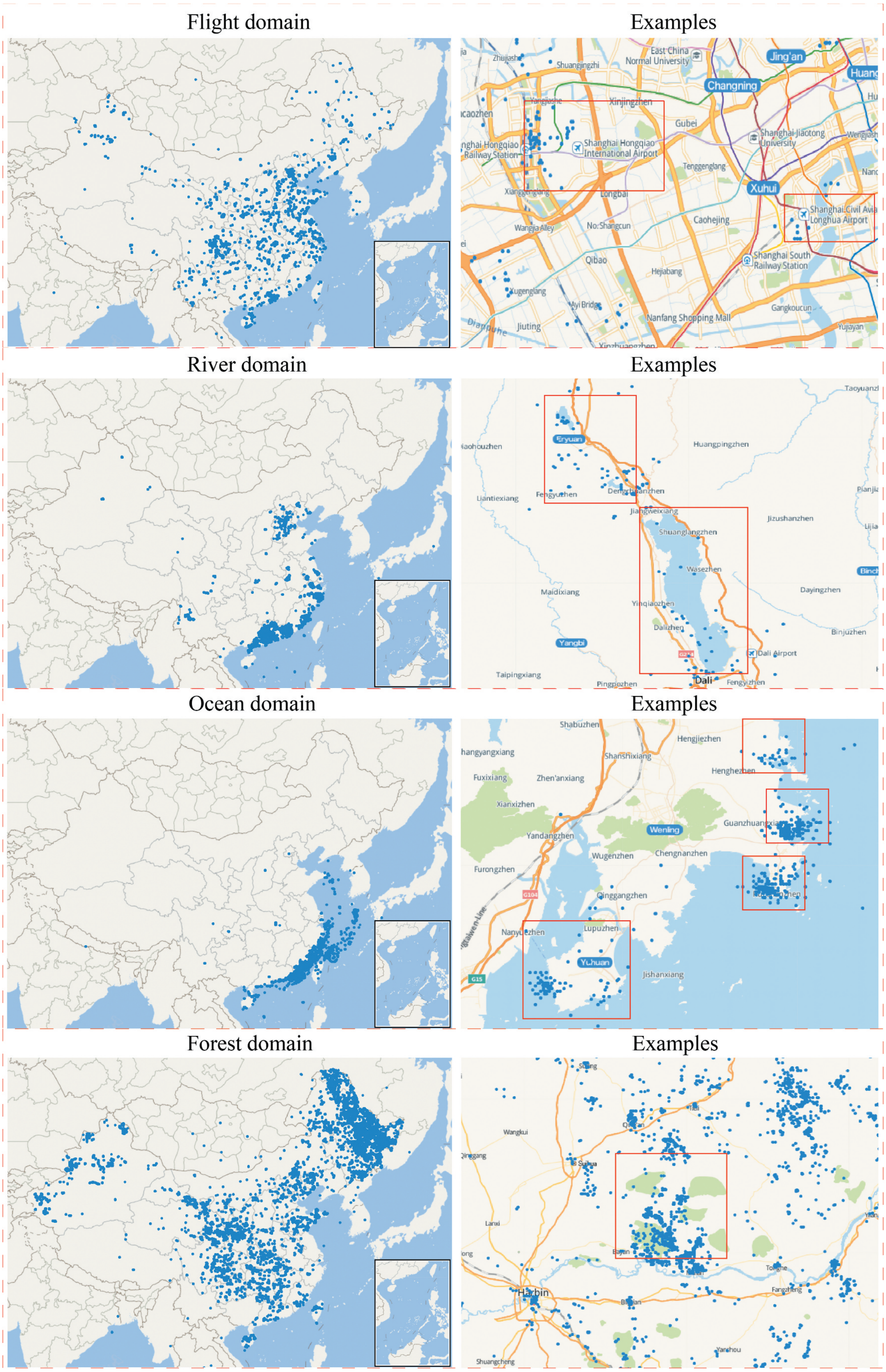

Figure 5. Experimental data for flight, river, ocean, and forest domains.

500 million tile accesses every day in recent years, and the numbers are growing every year. Therefore, it is the primary PMSP in China. The data used in this study comprised logs from October 1 to 19, 2020.

We selected four domains with a large number of users as the study cases: flight, river, ocean, and forest. Figure 5 shows the experimental data. Users browsed airports in the flight domain, rivers and lakes in the river domain, islands in the ocean domain, and forests and mountains in the forest domain. Referring to the source of the user access data, we manually labeled the topics using the access POIs, as shown in Table 4.

The topics in each domain were manually labeled using two approaches. Taking the river domain as an example, we tracked the source website from the access logs and found that the website was used for river and lake management. Thus, the topics of this domain were manually determined to be river related. In addition, we verified the reliability of manually determined topics by visualizing user browsing locations. For example, in Figure 5, the user's browsing targets were distributed around the river, which is consistent with the topic. These two approaches can ensure the accuracy of topic extraction, which is then used to verify the accuracy of the BIDU extraction.

Some examples of fields in the logs are shown in Table 5, including IP, time, layer, row, and column. To facilitate our research, we converted the latitude and longitude coordinates of the tiles according to the layer, row, and column.

4.1.3 Experimental settings

We constructed a POI spatial corpus by imitating a text corpus. The two-dimensional spatial distribution of POIs was converted into one-dimensional sequences wherein the POI types are words and POI sequences are sentences. Referring to the parameter settings in the related literature (Liu et al. 2020), when the maximum retrieval radius is 1000 m in the KNN method, the surrounding nearest neighbor POIs can be covered. However, in urban areas with dense POIs, the sentence length in the POI spatial corpus is longer than 100, which is larger than the average sentence length in the text corpus of 30. Thus, we set the maximum number of neighbors to 30 (Chen et al. 2018). The word2vec model was used to train the POI semantic space. This process was implemented using the Gensim package; the training method was CBOW with a window size of 5. In natural language processing, the dimension of the vectors in the semantic space trained by word2vec is 50–1000, which is determined by the size of the corpus (Liu et al. 2020). The training of our POI spatial corpus included 65.27 million POIs, which is the largest POI spatial corpus to our knowledge; thus, we set the POI vector dimension to 400. Both SVD and *t*-distributed Stochastic Neighbor Embedding (t-SNE) models were implemented using the scikit-learn package.

Table 4. Dataset of the domain users accessing Tianditu.

domain	records of visited tiles	manually labeled topics
Flight	7,124,961	Airport, Airport Departure, Enquire of Baggage, Departure Lounge
River	2,711,485	River, Lake, Bridge
Ocean	2,006,608	Ocean, Island, Port, Beach, Gulf, Strait
Forest	7,444,252	Forest, Mountain, River, Lake, Resort, Tourist Attraction

4.2 Validation of spatial feature quantification

4.2.1 POI semantic space visualization

The POI vector dimension was 400 resulting in the POI semantic space could not be visualized. The *t*-SNE model was used to map the POI semantic space onto a 2-dimensional semantic space to develop dimension-reduction visualization (Van der Maaten and Hinton 2008). We employed dimension reduction of POI vectors in the POI semantic space at the second level, as shown in Figure 6. POI types in the first, second, and third levels numbered 24, 268, and 899, respectively. Considering the clarity and informativeness of the visualization, we utilized only second-level POI semantic space for visualization. Adjacent POIs indicate a high probability of spatial co-occurrence and similar spatial semantics.

For example, the "Hospital" is adjacent to the "Emergency Center;" the "Bank" is adjacent to the "ATM;" "Shopping Related Places" appear near the "Coffee House," "Clothing Store" and various restaurants. "Taxi," "Parking Lot," "Service Area," "Charging Station" and "Airport Related" are adjacent; the sales and repair of automobiles related POIs are concentrated. Thus, the POI semantic space can quantify the spatial co-occurrence of POIs.

4.2.2 Validation of spatial feature quantification by multi-domain user classification

We adopted the popular indicators of accuracy, precision, recall, and F1 for the classification evaluation. The Area Under the Receiver Operator Characteristic (AUROC) was also used as the metric, which is suitable for imbalanced datasets.

Table 5. Examples of logs from the Tianditu.

IP	Time	Layer	COL	ROW
User1 IP	9 October 2020:14:14:32	17	113288	47020
User2 IP	9 October 2020:14:15:10	14	13566	7045
User2 IP	9 October 2020:14:15:10	14	13567	7045

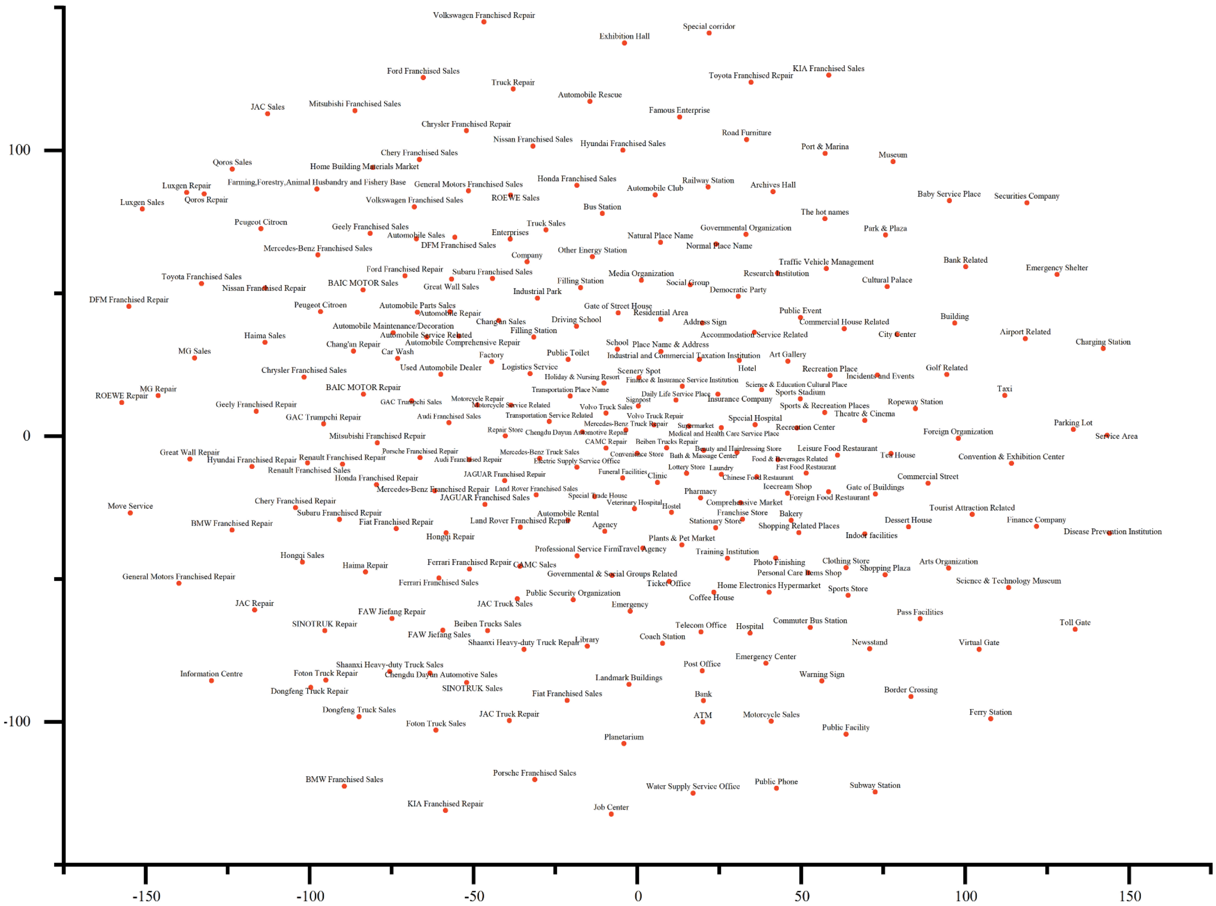

Figure 6. Second-level POI semantic space visualization.

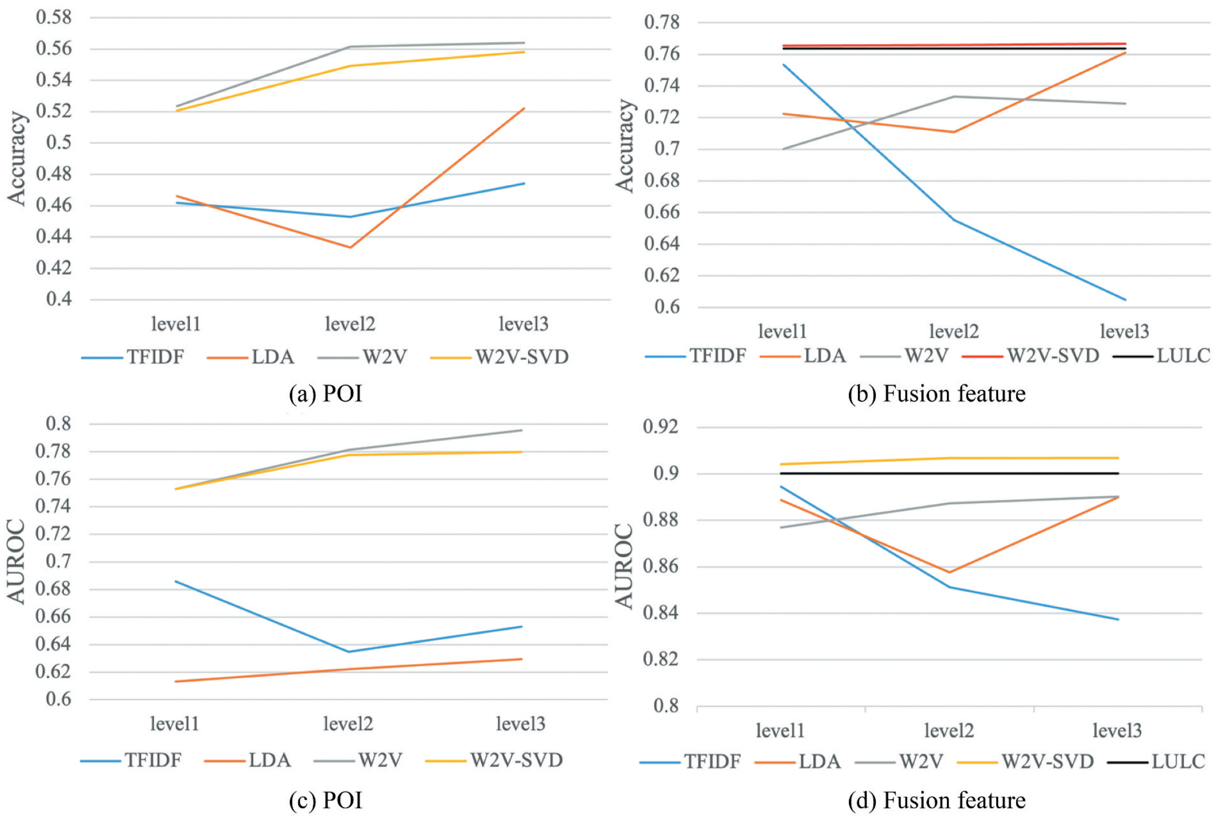

Figure 7. Overall accuracy and AUROC of multi-domain user classification with POI based on POI (a)(c), LULC (b)(d), and fusion features (b)(d).

4.2.2.1 Overall classification accuracy. Through multi-domain user classification, we evaluated the accuracy of spatial feature quantification. When the accuracy and AUROC curves of the multi-domain user classification were high, the POI semantic space effectively quantified spatial co-occurrence. We used POIs, LULCs, and their fusion features as the input and explored the preferences of different domain users based on the classification accuracy of the RF model. We also discuss the impact of the POI levels. The POI types include three levels that are subdivided from the first to the third level. In this study, we compared the accuracy of the proposed model and baselines, including the TFIDF, LDA, W2 V (Yao et al. 2017), and W2 V-SVD models. The results are shown in Figure 7.

Figure 7 shows the change in the overall classification accuracy and AUROC for domain users by POI level based on POI, LULC, and fusion features; the horizontal axis represents the POI level, such as level1, level2, and level3, and the vertical axis represents the classification accuracy/AUROC. As shown in Figure 7(a), the overall accuracy of all models increased as the POI level increased. When the POI level is higher, POIs can provide increasingly fine granular information. The LDA model achieved higher accuracy than the TFIDF model, especially in level3, and the accuracy of LDA was increased by 4.8% compared with that of TFIDF. As a classic probabilistic topic model approach, LDA has proven to be superior to TFIDF based on word frequency statistics. The accuracy of W2 V-based models (including the W2 V and W2 V-SVD models) was 4% higher than that of the TFIDF and LDA models. The advantage of the W2 V-based model is that it uses the POI semantic space to quantify the spatial co-occurrence of the POIs. The accuracy of the W2 V model was slightly higher than that of the W2 V-SVD model because the W2 V-SVD model reduced the dimensions of the POI vectors, resulting in information loss. The POI dimension of the W2 V-SVD model (30) was significantly lower than that of the W2 V model (400), which is conducive to fusion with LULCs.

Figure 7(b) shows the classification of domain users based on LULC and fusion features, where TFIDF, LDA, W2 V, and W2 V-SVD represent the classification models based on the fusion features, and LULC indicates the classification models based only on LULCs. Although there are only several types of LULC, the accuracy of the model based on LULCs is increased by 20% compared with the model based on POIs. Visited POIs were unevenly distributed across the different domains. Especially for the ocean domain, POIs are only distributed on the coast or islands; however, many visited tiles are in the ocean, where POIs do not exist. There was no POI in these sessions, which made accurate classification difficult. All visited tiles are composed of different LULCs, where water is the majority in marine areas and land is the main body in the forest. The uneven distribution of LULCs in the visited tiles in different domains is the user preference and basis of accurate classification. By fusing POIs with LULCs, we found that the accuracy of the TFTDF model decreased rapidly as the POI level increased, whereas the accuracy of the LDA model increased. The accuracy of the W2 V model was lower than that of the LDA model.

As shown in Figure 7(c,d), the accuracy levels are consistent with the AUROC. Although the dataset was imbalanced, this did not affect overall accuracy. The W2 V-based model achieved the highest AUROC value. When only using POIs in Figure 7(c), the W2 V model reaches the largest AUROC, whereas when using the fusion feature in Figure 7(d), the W2 V-SVD model achieves the highest AUROC. Thus, the SVD model facilitated the feature fusion of POIs and LULCs.

The above experimental results show that the POI semantic space constructed based on the W2 V-SVD model can accurately quantify spatial features, which provides a solid foundation for BIDU extraction. In the next section, we describe the classification accuracy for the four domains in detail.

4.2.2.2 Classification accuracy for each domain. We performed a fine-grained comparative analysis of users in distinct domains. Figure 8(a–f) show the precision, recall, and F1 of models based on POI and fusion features at three levels in flight, river, ocean, and forest domains. Figure 8(g) shows the classification accuracy based on LULCs. Figure 8(h) shows the importance of LULCs in the RF model.

The classification accuracies in different domains were unbalanced. Comparing POI (b), LULC (g), and fusion features (a) in Figure 8, the precision, recall, and F1 in ocean and forest domains reached 0.8; precision in the flight domain was 0.8 while recall and F1 were approximately 0.4; and all indicators in the river domain were balanced, reaching approximately 0.6.

The classification accuracies of the models were also unbalanced. Comparing (a) and (b) in Figure 8, we found that the F1 of the TFIDF and LDA models in the four domains were extremely imbalanced. The F1 in the forest was significantly higher than those in the other three domains. The accuracies of the TFIDF and LDA models were lower than those of the W2 V-based models, particularly in Figure 8(b). The classification accuracy of the W2 V-SVD model was more balanced in the four domains than that of the other models.

For all models, using the fusion features is better than using only POIs as the model input, indicating that the fusion features were effective. As the POI level increased, the classification accuracies of the TFIDF and LDA models improved significantly, while the performance of the W2 V-based models was relatively

Figure 8. Precision, recall, and F1s obtained by the four models based on POI and fusion features at three levels in flight, river, ocean, and forest domains (a–f). Classification accuracy based on LULCs (g). Importance of LULCs in the RF model (h).

stable, indicating that the POI semantic space was more effective than the bag-of-words model (Figure 8(b,d and f)).

As stated previously, users in different domains have different feature preferences. In Figure 8(b), the precision of the models in the flight domain reached 0.8, indicating that users in this domain prefer specific POI types that do not exist in other domains. In Figure 8(h), the Gini importance of the RF model based on LULCs represents the importance of each LULC in the multi-domain user classification. Water and land were the most important LULCs, and their corresponding domains were ocean and forest, respectively, yielding the highest classification accuracy for these domains.

Analyzing model performance for user classification in multiple domains, the proposed W2 V-SVD model achieved higher accuracy than the baseline

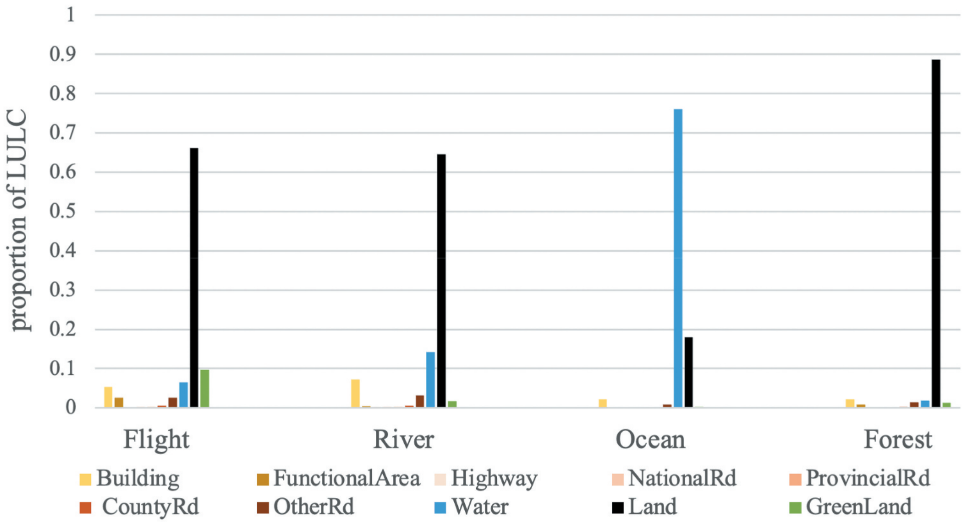

Figure 9. Proportion of LULCs in each domain to indicate domains users' preferences.

models, indicating its effectiveness for quantizing users' browsing interests. In the next section, we present the obtained BIDUs for the four domains.

4.3 Results and evaluation of BIDU extraction

We explored the distribution of user interests in different domains considering LULCs and POIs.

4.3.1 Analysis of BIDUs based on LULCs

We presented the proportion of LULCs in each domain to indicate the domain user preferences, as shown in Figure 9. The proportion of water was the highest in the ocean domain, indicating that most users visited the ocean. Land accounted for the highest proportion in the forest domain; however, the proportion of green land was small. Only forests in cities are represented as green land in Tianditu, such as scenic areas and parks, while forests in remote areas are represented as land. In the river domain, the proportion of land was also the highest; however, the proportion of water was higher than that in the flight and forest domains. In the flight domain, the land also comprised the main body and contained a small proportion of water, forests, and buildings. The proportion of roads was the smallest in all four domains.

LULCs are divided into coarse-grained types, and only 10 LULCs can be detected from the tiles. Thus, the ability of this indicator to express users' browsing interests is limited. This highlights the advantages of our proposed W2 V-SVD model, a POI-based method, for extracting BIDUs.

4.3.2 Analysis of BIDUs based on POIs

The key to evaluating the topic model was to validate the extraction semantically consistent topics, indicated by the coherence score. In this study, we used the coherence score to evaluate semantically consistent interests of the proposed model. In multi-domain user classification, the third-level POI achieved the highest accuracy. Thus, we used the third-level POI features for experimentation.

4.3.2.1 Coherence evaluation of BIDUs. We used the word-vector-based indicator WESim, proposed by (Fang et al. 2016), to evaluate the coherence score for the extracted BIDUs. WESim measures semantic consistency by calculating the average semantic similarity between the Top-T POI pairs in the BIDUs. Top-T refers to the number of key POIs in a given topic. The fewer the key POIs, the more focused the topics. Figure 10 compares the performance of the W2 V-SVD and LDA models in the four domains based on WESim.

In Figure 10, we compared the semantically consistent interests of the W2 V-SVD and LDA models in the four domains. The horizontal axis represents the Top-T POIs for calculating WESim, and the vertical axis represents WESim. From the WESim curve of the W2 V-SVD and LDA models in the four domains, we found that, as Top-T increases, all WESim exhibit a downward trend, indicating that the semantic consistency decreases. In the flight, river, and ocean domains, the WESim of the W2 V-SVD model was higher than that of the LDA model for all TOP-T, indicating that the W2 V-SVD model achieved better semantic consistency than the LDA model. In the forest domain, when Top-T = 2, the WESim of the LDA model was higher than that of the W2 V-SVD model, but as Top-T increased, the WESim of W2 V-SVD was higher than that of the LDA model. The browsing interests of the W2 V-SVD model are composed of k-nearest POIs, and the generation process demonstrated its superior semantic consistency.

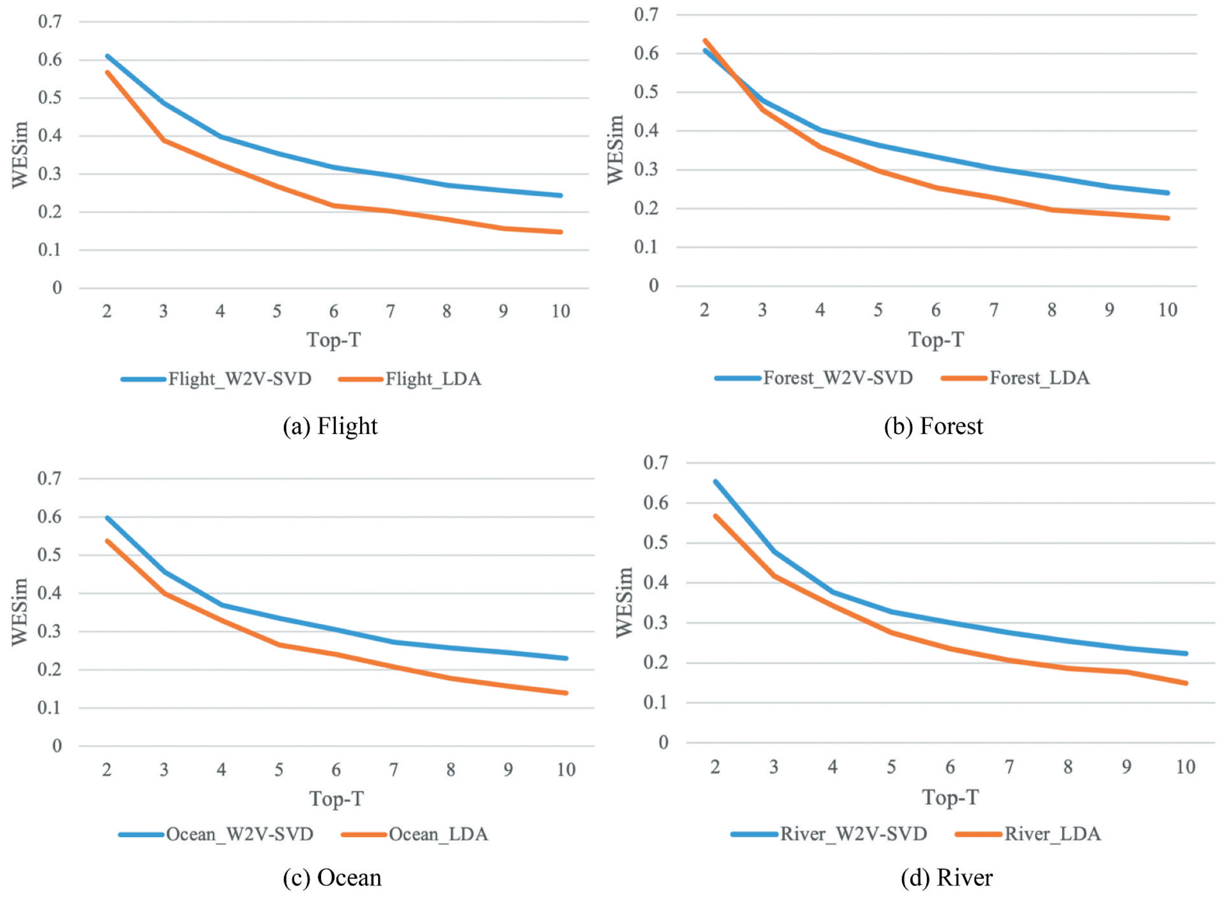

Figure 10. Comparison of semantically consistent interests in the W2V-SVD and LDA models across domains.

Table 6. Comparing the topics of the W2 V-SVD model and LDA model in the flight domain.

topic	W2V-SVD	LDA
1	Village-level Place Name+Filling Station+PetroChina+Mountain +Below Town level Government and Institution	Bus Station Related+Company+Enterprises+Automobile Repair +Hardware Store
2	**Airport**+**Airport Departure/Arrival**+Enquire of Baggage +Departure Lounge+Public Parking Lot	Training Institution+Public Parking Lot+Gate of Street House+Science & Education Cultural Place+Parking Lot Entrance & Exit
3	Mountain+River+Lake+Expressway Exit+Resort	Village-level Place Name+Name of Intersection+Road Name+Company +Bus Station Related
4	Mountain+Personal Care Items Shop+Brand Clothing Store +Clothing Store+Children's Store	**Airport Departure/Arrival** +**Airport**+Public Parking Lot+Bus Station Related+Company
5	Company+Enterprises+Gate of Street House+Building Number +Automobile Sales	Company+Shopping Related Places+Gate of Street House+Enterprises +Convenience Store
6	Convenience Store+Road Name+Name of Intersection+**Airport** +Tobacco & Wine Franchise Store	Indoor facilities+Public Toilet+Public Phone+Name of Intersection +Company
7	Indoor facilities+Shopping Related Places+Enterprises+**Airport Related**+Franchise Store	Tourist Attraction Related+Name of Intersection+Public Parking Lot+Road Name+Public Toilet
8	**Airport Related**+Building Number+Express Post+Fast Food Restaurant+Supermarket	Accommodation Service Related+Hotel+Bus Station Related+Name of Intersection+Residential Quarter
9	Used Automobile Dealer+ **Airport Departure/Arrival** +**Airport Related**+Repair Store+Automobile Parts Sales	Building Number+Gate of Street House+Name of Intersection+Residential Quarter+Main Entrance of Street House Gate
10	River+Water Sports Centre+Lake+Resort+Import Volkswagen Sales	Road Name+Main Entrance of Street House Gate+Public Toilet+Public Parking Lot+Parking Lot Entrance & Exit

4.3.2.2 Qualitative evaluation of BIDU representation. BIDUs can be represented as POI types. For example, in the river domain, rivers, bridges, lakes, and other domain-related POIs appear frequently, whereas other unrelated POIs such as islands, beaches, and airports do not appear. Qualitative analysis of topic extraction is an important approach for evaluating the performance and applicability of the method (Kim, Park, and Lee 2020). The qualitative analysis indicator in this study was whether the extracted topics contain the manually labeled topics in Table 4.

We applied the W2 V-SVD and LDA models to extract the TOP-5 POIs as the topic, which are also the BIDUs, in the four domains (Tables 6–9). We manually labeled key POIs and emphasized them in bold font. All key POIs are ground-truth labels that

Table 7. Comparing the topics of the W2 V-SVD model and LDA model in the river domain.

topic	W2V-SVD	LDA
1	Village-level Place Name+Filling Station+Below Town level Government and Institution+Mountain+Township-level Place Name	Village-level Place Name+Name of Intersection+Below Town level Government and Institution+Bus Station Related+Mountain
2	Mountain+**Lake**+Fishing Spot+**River**+Resort	Parking Lot Entrance & Exit+Public Parking Lot+Logistics Service +Residential Quarter+Training Institution
3	Name of Intersection+Road Name+Public Toilet+Company+Factory	Main Entrance of Street House Gate+Bus Station Related+Furniture Store+Factory+Company
4	**Lake+River**+Natural Place Name+Four-star Hotel+Ropeway Station	Machinery and Electronics+Industrial Park+Factory+Company +Mountain
5	Name of Intersection+Road Name+Mobile Handsets Sales +Transportation Service Related+Children's Store	Building Number+Village-level Place Name+Bus Station Related +Tourist Attraction Related+Below Town level Government and Institution
6	Building Number+Residential Quarter+Northeastern Chinese Food +Below Town level Government and Institution+Card & Chess Room	Company+Gate of Street House+Factory+Enterprises+Name of Intersection
7	**River**+**Bridge**+Park Box Office+Vocational Technical School+Indoor facilities	Name of Intersection+Bus Station Related+Gate of Street House +Road Name+Residential Quarter
8	**Bridge**+Accommodation Service Related+Chinese Food Restaurant +Volvo Truck Sales+Shopping Plaza	Accommodation Service Related+Hostel+Hotel+Village-level Place Name+Recreation Place
9	Buddhist & Taoist Temple+Bus Station Related+Public Parking Lot +Island+Tourist Attraction Related	Village-level Place Name+Company+Bus Station Related +Enterprises+Name of Intersection
10	Other Farming, Forestry, Animal Husbandry and Fishery Base+Gate of Street House+Bus Station Related+Buddhist & Taoist Temple +Elementary School	Name of Intersection+Shopping Related Places+Road Name +Hardware Store+Chinese Food Restaurant

Table 8. Comparing the topics of the W2 V-SVD model and LDA model in the ocean domain.

topic	W2V-SVD	LDA
1	**Island**+Mountain+**Gulf and Strait**+Village-level Place Name +**River**	Building Number+Gate of Street House+Chinese Food Restaurant+Park +Accommodation Service Related
2	**Island**+Beauty and Hairdressing Store+Medical Supplies+Bakery +Food & Beverages Related	Accommodation Service Related+Company+Enterprises+Hostel+Bus Station Related
3	**Gulf and Strait**+**Island**+Accommodation Service Related+**Ferry Terminal**+Special Trade House	Building Number+Name of Intersection+Company+Road Name+Bus Station Related
4	Mountain+**River**+Fishing Spot+Resort+Lake	Company+Factory+Village-level Place Name+Gate of Street House +Town Level Government and Institution
5	Company+Building Number+Enterprises+Residential Quarter +Industrial Park	Village-level Place Name+Below Town level Government and Institution +Accommodation Service Related+Mountain+Tunnel
6	**River**+**Port & Marina**+Building Number+Tourist Attraction Related +Basketball Stadium	**Gulf and Strait**+Main Entrance of Street House Gate+Bus Station Related+Company+Governmental & Social Groups Related
7	Enterprises+Company+Home Building Materials Market+Finance & Insurance Institution+Automobile Sales	Village-level Place Name+Shopping Related Places+Convenience Store +Bus Station Related+Science & Education Cultural Place
8	Building Number+**Port & Marina**+Accommodation Service Related +Below Town level Government and Institution+Residential Quarter	Company+Name of Intersection+Road Name+Factory+Tourist Attraction Related
9	**Port & Marina**+Name of Intersection+Hotel+**Beach** +Transportation Service Related	Village-level Place Name+**Island**+Tourist Attraction+Mountain+Tourist Attraction Related
10	Tourist Attraction+Buddhist & Taoist Temple+Beach+Brand Clothing Store+Indoor facilities	Name of Intersection+Village-level Place Name+Road Name+Below Town level Government and Institution+Governmental Organization Related

represent the topics in this domain, as shown in Table 4. The more key POIs occur, the more accurate the extracted topics. In contrast, if key POIs do not appear in all topics, then we cannot determine the real BIDUs of the domain.

In Tables 6–9, the topics of the W2 V-SVD model are more accurate than those of the LDA model. In the flight domain, key POIs such as "Airport," "Airport Departure/Arrival," "Enquire of Baggage," "Departure Lounge," and "Airport Related," appeared many times in the W2 V-SVD model results, whereas only "Airport Departure/Arrival" and "Airport" appeared once in the LDA model results. In the river domain, the key POIs of the W2 V-SVD model include "River," "Lake," and "Bridge," however, there are no key POIs in the LDA results. In the ocean domain, the key POIs of the W2 V-SVD model are composed of "Island," "Gulf and Strait," "Ferry Terminal," "River," "Port & Marina," and "Beach," whereas the LDA results only include "Island" and "Gulf and Strait". In the forest domain, the key POIs of the W2 V-SVD model contain "Mountain," "Resort," "River," "Lake," "Other Farming, Forestry, Animal Husbandry, and Fishery Base," and "Tourist Attraction Related". In the LDA results, "Mountain," "River," "Tourist Attraction," "Other Farming, Forestry, Animal Husbandry, and Fishery Base," and "Tourist Attraction Related" appear. These results are consistent with the POI distribution shown in Figure 5. In summary, the BIDUs of the W2 V-SVD model were more reasonable than those of the LDA model.

The topics of the W2 V-SVD model are similar to those of the LDA model. In addition to the representative topics in the four domains, the two models extracted some common POIs in topics that could not be distinguished as significantly related to the given domain. For example: "Name of Intersection," "Road Name," "Company," "Convenience Store,"

Table 9. Comparing the topics of the W2 V-SVD model and LDA model in the forest domain.

topic	W2V-SVD	LDA
1	Village-level Place Name+Township-level Place Name+Filling Station +Mountain+Below Town level Government and Institution	Building Number+Gate of Street House+Residential Quarter+Road Name+Below Town level Government and Institution
2	**Mountain**+Resort+River+Lake+Fishing Spot	Town Level Government and Institution+Township-level Place Name +Police Station+Governmental Organization Related +Accommodation Service Related
3	**Other Farming, Forestry, Animal Husbandry, and Fishery Base** +Name of Intersection+Road Name+Public Toilet+Toll Gate	**Tourist Attraction Related**+Public Toilet+Prefecture Level Government and Institution+Public Parking Lot+**Tourist Attraction**
4	Name of Intersection+Road Name+Public Toilet+Expressway Entrance+Bus Station Related	Residential Quarter+Gate of Street House+Company+Training Institution+Commercial House Related
5	**River**+Township-level Place Name+**Tourist Attraction** Related +Expressway Parking Area+Expressway Filling Station Service Area	Village-level Place Name+Below Town level Government and Institution+**Mountain**+**River**+Bridge
6	Township-level Place Name+Town Level Government and Institution +Police Station+Governmental & Social Groups Related+District & County Level Government and Institution	Shopping Related Places+Chinese Food Restaurant+Daily Life Service Place+Hardware Store+Convenience Store
7	Company+Building Number+Governmental & Social Groups Related +Gate of Street House+Residential Quarter	Name of Intersection+Road Name+District & County Level Government and Institution+Governmental & Social Groups Related +Governmental Organization Related
8	**Tourist Attraction Related**+Building Number+Residential Quarter +Gate of Street House+Northeastern Chinese Food	Bus Station Related+Main Entrance of Street House Gate+Company +Automobile Repair+Gate of Street House
9	Enterprises+Township-level Place Name+Company+Hardware Store +Home Building Materials Market	Public Parking Lot+Finance & Insurance Institution+ Construction Company +Business Office Building + Advertisement and Decoration
10	Village-level Place Name+Township-level Place Name+Filling Station +Mountain+Below Town level Government and Institution	Building Number+Gate of Street House+Residential Quarter+Road Name+Below Town level Government and Institution

"Village-level Place Name," "Township-level Place Name," "Gate of Street House," "Residential Quarter," and "Bus Station Related,". The POIs appearing in both models indicated that the results of the W2 V-SVD model were similar to those of the LDA model.

5. Conclusion

Users in the same domain on a given PMSP have similar browsing interests. We proposed spatial feature extraction approaches for POIs and LULCs in tiles. The word2vec model can be employed to construct a POI semantic space, which can then be used to model spatial POI co-occurrence and perform multi-domain user classification for validation. We proposed the W2 V-SVD model to achieve BIDU extraction by topic. This study can help PMSP providers understand the requirements of domain users and promote the optimization of intelligent PMSPs.

The W2 V-SVD model searches for the k-nearest POIs as BIDUs. Thus, the semantic consistency of the W2 V-SVD model was better than that of the LDA model. However, some POIs that have never been visited will appear in BIDUs and produce confusing results. In this study, the template-based method for LULC extraction is simple. We plan to use deep learning models, such as convolutional neural networks, to extract high-level semantic features including the shape and relationship of spatial features.

Acknowledgments

The authors thank the National Geomatics Center of China and Tianditu for supporting this work.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Natural Science Foundation of China [grant numbers: U20A2091 41771426], Zhizhuo Research Fund on Spatial-Temporal Artificial Intelligence [grant number ZZJJ202204], and LIESMARS Special Research Funding.

Notes on contributors

Guangsheng Dong received BS degree from Central South University, China, in 2015 and PhD degree from Wuhan University, China, in 2021. Currently, he is a post-doc and focuses on spatial-temporal data mining and intelligent geographic information service.

Rui Li is currently a full professor in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. Her scientific interests include networking communication, spatial-temporal computing, and networks GIS.

Huayi Wu is a full professor in GeoInformatics and the Vice Director of the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His scientific interests include high-performance geospatial computing and intelligent geospatial web services.

Wei Huang is the Director of the public platform department in the National Geomatics Center of China. His scientific interests are the development of the public service platform of geographical information and spatial cloud computing.

Hongping Zhang is a senior engineer in the department of the public platform in the National Geomatics Center of China, majoring in GIS web service and integrated application.

Vincent Tao, male, born in November 1967, is a Canadian. In 1998, he received his doctor's degree in geomatics from the University of Calgary, Canada, the founder and CEO of Wayz.ai. and a high-level overseas talent in Shanghai. Mainly engaged in the research of spatio-temporal artificial intelligence technology. An entrepreneurial entrepreneur recognized by the industry as having both business operation and in-depth technical skills, and an authoritative expert in the international map industry, he has made a number of technological inventions, published more than 200 papers, and was awarded as a tenured professor of York University and the national chief professor of space information in Canada.

Quan Liu received his Master's degree of Engineering from University of California, Los Angeles in 2018. Currently he is Vice President of Engineering at Wayz.ai. His research interest includes knowledge graph, spatio-temporal AI and big data.

ORCID

Guangsheng Dong http://orcid.org/0000-0001-7676-497X
Rui Li http://orcid.org/0000-0001-5167-2956
Huayi Wu http://orcid.org/0000-0003-3971-0512
Wei Huang http://orcid.org/0000-0002-9382-7212
Hongping Zhang http://orcid.org/0000-0002-2618-3533
Vincent Tao http://orcid.org/0000-0001-5382-2364
Quan Liu http://orcid.org/0000-0002-7515-0834

Data availability statement

The data used in this paper was collected by the National Geomatics Center of China and Tianditu. Due to the nature of this research, participants of this study did not agree for their data to be shared publicly for protecting the users' privacy. The logs from OpenStreetMap could become the alternative (https://planet.openstreetmap.org/tile_logs/).

References

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022. [Online]. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. doi:10.1023/A:1010933404324.

Chen, J., Q. Chen, X. Liu, H. Yang, D. Lu, and B. Tang. 2018. "The Bq Corpus: A Large-Scale Domain-Specific Chinese Corpus for Sentence Semantic Equivalence Identification." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 4946–4951. doi:10.18653/v1/D18-1536.

Cheng, X., X. Yan, Y. Lan, and J. Guo. 2014. "Btm: Topic Modeling Over Short Texts." *IEEE Transactions on Knowledge and Data Engineering* 26 (12): 2928–2941. doi:10.1109/TKDE.2014.2313872.

Chen, H. C., K. T. Putra, S. S. Tseng, C. L. Chen, and J. C. W. Lin. 2020. "A Spatiotemporal Data Compression Approach with Low Transmission Cost and High Data Fidelity for an Air Quality Monitoring System." *Future Generation Computer Systems* 108: 488–500. doi:10.1016/j.future.2020.02.032.

Chen, H. C., K. T. Putra, C. E. Weng, and J. C. W. Lin. 2022. "A Novel Predictor for Exploring PM2. 5 Spatiotemporal Propagation by Using Convolutional Recursive Neural Networks." *Journal of Internet Technology* 23 (1): 165–176.

Djenouri, Y., A. Belhadi, J. C. W. Lin, and A. Cano. 2019. "Adapted K-Nearest Neighbors for Detecting Anomalies on Spatio–Temporal Traffic Flow." *IEEE Access* 7: 10015–10027. doi:10.1109/ACCESS.2019.2891933.

Dlamini, S., S. G. Tesfamichael, G. D. Breetzke, and T. Mokhele. 2021. "Spatio-Temporal Patterns and Changes in Environmental Attitudes and Place Attachment in Gauteng, South Africa." *Geo-Spatial Information Science* 24 (4): 666–677. doi:10.1080/10095020.2021.1976599.

Dong, W., H. Liao, Z. Zhan, B. Liu, S. Wang, and T. Yang. 2019. "New Research Progress of Eye Tracking-Based Map Cognition in Cartography Since 2008." *Acta Geographica Sinica* 74 (3): 599–614. doi:10.11821/dlxb201903015.

Dong, G., R. Li, J. Jiang, H. Wu, and S. C. McClure. 2020. "Multigranular Wavelet Decomposition-Based Support Vector Regression and Moving Average Method for Service-Time Prediction on Web Map Service Platforms." *IEEE Systems Journal* 14 (3): 3653–3664. doi:10.1109/jsyst.2019.2944527.

Dong, G., R. Li, H. Wu, W. Chen, W. Huang, and H. Zhang. 2022. "Browsing Behavior Modeling and Browsing Interest Extraction in the Trajectories on Web Map Service Platforms." *Expert Systems with Applications* 195: 116590. doi:10.1016/j.eswa.2022.116590.

Drozd, A., A. Gladkova, and S. Matsuoka. 2016. "Word Embeddings, Analogies, and Machine Learning: Beyond King-Man+woman=queen." In Proceedings of Coling 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 3519–3530. [Online] https://aclanthology.org/C16-1332.pdf

Dumais, S. T. 2004. "Latent Semantic Analysis." *Annual Review of Information Science and Technology* 38 (1): 188–230. doi:10.1002/aris.1440380105.

Fang, A., C. Macdonald, I. Ounis, and P. Habel. 2016. "Using Word Embedding to Evaluate the Coherence of Topics from Twitter Data." In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1057–1060. doi:10.1145/2911451.2914729.

Feng, S., G. Cong, B. An, and Y. M. Chee. 2017. "Poi2vec: Geographical Latent Representation for Predicting Future Visitors." In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 102–108. doi:10.1609/aaai.v31i1.10500.

Fisher, D. 2007. "Hotmap: Looking at Geographic Attention." *IEEE Transactions on Visualization and Computer Graphics* 13 (6): 1184–1191. doi:10.1109/Tvcg.2007.70561.

García Martín, R., J. P. de Castro Fernández, E. Verdú Pérez, M. J. Verdú Pérez, and L. M. Regueras Santos. 2013. "An OLS Regression Model for Context-Aware Tile Prefetching in a Web Map Cache." *International Journal of Geographical Information Science* 27 (3): 614–632. doi:10.1080/13658816.2012.721555.

García, R., E. Verdú, L. M. Regueras, J. P. de Castro, and M. J. Verdú. 2013. "A Neural Network Based Intelligent System for Tile Prefetching in Web Map Services." *Expert Systems with Applications* 40 (10): 4096–4105. doi:10.1016/j.eswa.2013.01.037.

Guo, A., and T. Yang. 2016. "Research and Improvement of Feature Words Weight Based on TFIDF Algorithm." In 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference, 415–419. doi: 10.1109/ITNEC.2016.7560393.

Hofmann, T. 2001. “Unsupervised Learning by Probabilistic Latent Semantic Analysis.” *Machine Learning* 42 (1): 177–196. doi:10.1023/A:1007617005950.

Jung, S., and W. C. Yoon. 2020. “An Alternative Topic Model Based on Common Interest Authors for Topic Evolution Analysis.” *Journal of Informetrics* 14 (3): 101040. doi:10.1016/j.joi.2020.101040.

Kim, S., H. Park, and J. Lee. 2020. “Word2vec-Based Latent Semantic Analysis (W2V-LSA) for Topic Modeling: A Study on Blockchain Technology Trend Analysis.” *Expert Systems with Applications* 152: 113401. doi:10.1016/j.eswa.2020.113401.

Krafka, K., A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. 2016. “Eye Tracking for Everyone.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2176–2184, Las Vegas, NV, USA. doi: 10.1109/CVPR.2016.239.

Li, R., G. Dong, J. Jiang, H. Wu, N. Yang, and W. Chen. 2019. “Self-Adaptive Load-Balancing Strategy Based on a Time Series Pattern for Concurrent User Access on Web Map Service.” *Computers & Geosciences* 131: 60–69. doi:10.1016/j.cageo.2019.06.015.

Li, R., J. Fan, H. Wu, J. Jiang, and G. Dong. 2018. “Group-User Access Patterns and Tile Prefetching Based on a Time-Sequence Distribution in Cloud-Based GIS.” *Computers, Environment and Urban Systems* 69: 17–27. doi:10.1016/j.compenvurbsys.2017.12.002.

Liu, K., L. Yin, F. Lu, and N. Mou. 2020. “Visualizing and Exploring POI Configurations of Urban Regions on POI-Type Semantic Space.” *Cities* 99: 102610. doi:10.1016/j.cities.2020.102610.

Li, W., X. Wang, R. Hu, and J. Tian. 2011. “User Interest Modeling by Labeled LDA with Topic Features.” In 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, 6–11. doi: 10.1109/CCIS.2011.6045022.

Manson, S. M., L. Kne, K. R. Dyke, J. Shannon, and S. Eria. 2012. “Using Eye-Tracking and Mouse Metrics to Test Usability of Web Mapping Navigation.” *Cartography and Geographic Information Science* 39 (1): 48–60. doi:10.1559/1523040639148.

Mei, Y., Z. Gui, J. Wu, D. Peng, R. Li, H. Wu, and Z. Wei. 2022. “Population Spatialization with Pixel-Level Attribute Grading by Considering Scale Mismatch Issue in Regression Modeling.” *Geo-Spatial Information Science* 25 (3): 365–382. doi:10.1080/10095020.2021.2021785.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *arXiv preprint arXiv:1301.3781*. doi:10.48550/arXiv.1301.3781.

Quinn, S., and M. Gahegan. 2010. “A Predictive Model for Frequently Viewed Tiles in a Web Map.” *Transactions in GIS* 14 (2): 193–216. doi:10.1111/j.1467-9671.2010.01191.x.

Sari Aslam, N., M. R. Ibrahim, T. Cheng, H. Chen, and Y. Zhang. 2021. “ActivityNet: Neural Networks to Predict Public Transport Trip Purposes from Individual Smart Card Data and POIs.” *Geo-Spatial Information Science* 24 (4): 711–721. doi:10.1080/10095020.2021.1985943.

Schütze, H., C. D. Manning, and P. Raghavan. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Seo, Y. D., and Y. S. Cho. 2021. “Point of Interest Recommendations Based on the Anchoring Effect in Location-Based Social Network Services.” *Expert Systems with Applications* 164: 114018. doi:10.1016/j.eswa.2020.114018.

Sharma, D., B. Kumar, and S. Chand. 2017. “A Survey on Journey of Topic Modeling Techniques from SVD to Deep Learning.” *International Journal of Modern Education and Computer Science* 9 (7): 50–62. doi:10.5815/ijmecs.2017.07.06.

Sutherland, I., and K. Kiatkawsin. 2020. “Determinants of Guest Experience in Airbnb: A Topic Modeling Approach Using LDA.” *Sustainability* 12 (8): 3402. doi:10.3390/su12083402.

Tobler, W. R. 1970. “A Computer Movie Simulating Urban Growth in the Detroit Region.” *Economic Geography* 46 (sup1): 234–240. doi:10.2307/143141.

Tontodimamma, A., E. Nissi, A. Sarra, and L. Fontanella. 2021. “Thirty Years of Research into Hate Speech: Topics of Interest and Their Evolution.” *Scientometrics* 126 (1): 157–179. doi:10.1007/s11192-020-03737-6.

Unrau, R., and C. Kray. 2019. “Usability Evaluation for Geographic Information Systems: A Systematic Literature Review.” *International Journal of Geographical Information Science* 33 (4): 645–665. doi:10.1080/13658816.2018.1554813.

Van der Maaten, L., and G. Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research* 9 (86): 2579–2605.

Wall, M. E., A. Rechtsteiner, and L. M. Rocha. 2003. *“Singular Value Decomposition and Principal Component Analysis.” a Practical Approach to Microarray Data Analysis, 91-109*. doi:10.1007/0-306-47815-3_5.

Wang, X., D. Chen, G. Lu, Y. Peng, and C. Hu. 2014. “Web Map Service Log Analysis.” In International Conference on Wireless Algorithms, Systems, and Applications, 22–33. doi:10.1007/978-3-319-07782-6_3.

Yan, B., K. Janowicz, G. Mai, and S. Gao. 2017. “From Itdl to Place2vec: Reasoning About Place Type Similarity and Relatedness by Learning Embeddings from Augmented Spatial Contexts.” In Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems, 1–10. doi: 10.1145/3139958.3140054

Yao, Y., X. Li, X. Liu, P. Liu, Z. Liang, J. Zhang, and K. Mai. 2017. “Sensing Spatial Distribution of Urban Land Use by Integrating Points-Of-Interest and Google Word2vec Model.” *International Journal of Geographical Information Science* 31 (4): 825–848. doi:10.1080/13658816.2016.1244608.

Yu, Y., B. H. Nguyen, F. Yu, and V. N. Huynh. 2021. “Discovering Topics of Interest on Steam Community Using an LDA Approach.” In International Conference on Applied Human Factors and Ergonomics, 510–517. doi: 10.1007/978-3-030-80840-2_59.

Zhai, W., X. Bai, Y. Shi, Y. Han, Z. R. Peng, and C. Gu. 2019. “Beyond Word2vec: An Approach for Urban Functional Region Extraction and Identification by Combining Place2vec and POIs.” *Computers, Environment and Urban Systems* 74: 1–12. doi:10.1016/j.compenvurbsys.2018.11.008.

Zheng, W., B. Ge, and C. Wang. 2019. “Building a TIN-LDA Model for Mining Microblog Users’ Interest.” *IEEE Access* 7: 21795–21806. doi:10.1109/ACCESS.2019.2897910.

Zhu, M., W. Chen, J. Xia, Y. Ma, Y. Zhang, Y. Luo, Z. Huang, and L. Liu. 2019. “Location2vec: A Situation-Aware Representation for Visual Exploration of Urban Locations.” *IEEE Transactions on Intelligent Transportation Systems* 20 (10): 3981–3990. doi:10.1109/TITS.2019.2901117.