Computers, Environment and Urban Systems 80 (2020) 101442

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A framework for extracting urban functional regions based on multiprototype word embeddings using points-of-interest data

Sheng Hu[a], Zhanjun He[a,b], Liang Wu[a,b,*], Li Yin[c], Yongyang Xu[a], Haifu Cui[a]

[a] Faculty of Information Engineering, China University of Geosciences, Wuhan 430074, China
[b] National Engineering Research Center of Geographic Information System, Wuhan 430074, China
[c] Department of Urban and Regional Planning, State University of New York, Buffalo, NY 14214, USA

ARTICLE INFO

Keywords:
Urban functional regions
Word embeddings
Points-of-interest
Spatial clusters

ABSTRACT

Many studies are in an effort to explore urban spatial structure, and urban functional regions have become the subject of increasing attention among planners, engineers and public officials. Attempts have been made to identify urban functional regions using high spatial resolution (HSR) remote sensing images and extensive geo-data. However, the research scale and throughput have also been limited by the accessibility of HSR remote sensing data. Recently, big geo-data are becoming increasingly popular for urban studies since research is still accessible and objective with regard to the use of these data. This study aims to build a novel framework to provide an alternative solution for sensing urban spatial structure and discovering urban functional regions based on emerging geo-data – points of interest (POIs) data and an embedding learning method in the natural language processing (NLP) field. We started by constructing the intraurban functional corpus using a center-context pairs-based approach. A word embeddings representation model for training that corpus was used to extract multiprototype vectors in the second step, and the last step aggregated the functional parcels based on an introduced spatial clustering method, hierarchical density-based spatial clustering of applications with noise (HDBSCAN). The clustering results suggested that our proposed framework used in this study is capable of discovering the utilization of urban space with a reasonable level of accuracy. The limitation and potential improvement of the proposed framework are also discussed.

1. Introduction

Urban functional regions, which are closely related to the spatial and social structure of urban environments, are important and essential content of urban planning (Antikainen, 2005; Bracken, 2014). There has been continued and sustained interest in developing classification and integrating approaches for accurately inferring and extracting regions of different functions in cities (Hu et al., 2015; Karlsson, 2007; Pei et al., 2014; Zhi et al., 2016), as this problem is applied in a wide variety of areas, including city transportation, public security, management and sustainable development (Dear & Flusty, 1998; Janowicz, Scheider, Pehle, & Hart, 2012; Regan et al., 2015). Sensing the spatial and social structure of urban environments facilitates not only the daily living of citizens but also designing better urbanization strategies for the future.

Urban functional (or urban functions) regions are an important geospatial attribute of urban land, which is generally determined by two perspectives: land use types (Jokar Arsanjani, Helbich, Bakillah, Hagenauer, & Zipf, 2013) and human activities (Zhong, Huang, Arisona, Schmitt, & Batty, 2014). For the former, research efforts have been made to classify urban land use and identify urban functional regions at a fine-grained spatial and temporal resolution using HSR remote sensing images (Song, Lin, Li, & Prishchepov, 2018; Vatsavai et al., 2011; Wen, Huang, Zhang, & Benediktsson, 2016; Zhang et al., 2017; Zhang, Du, Wang, & Zhou, 2018), such as scene-based (Zhang & Du, 2015; Zhang, Du, & Wang, 2015) and object-oriented classification (OOC) (Blaschke et al., 2014; Hu & Wang, 2013) models. However, HSR remote sensing data are good at extracting physical characteristics (for example, spectral, shape and texture features) of ground objects but fail to extract socioeconomic information relating to urban functional regions (Heiden et al., 2012; Liu et al., 2015; Pei et al., 2014; Van de Voorde, Jacquet, & Canters, 2011). Meanwhile, the research scale and throughput have also been limited by the accessibility of HSR remote sensing data compared with the open source data and geo-data (Liu et al., 2015). For the latter, enabled by new and emerging big geo-data such as points-of-interest (POIs), social media, mobile and trajectory

* Corresponding author at: Faculty of Information Engineering, China University of Geosciences, Wuhan 430074, China.
E-mail address: wuliang@cug.edu.cn (L. Wu).

https://doi.org/10.1016/j.compenvurbsys.2019.101442
Received 9 May 2019; Received in revised form 5 November 2019; Accepted 9 November 2019

data, there has been a recent surge in inferring urban functional regions to study the spatial and social structure of urban environments (Barbosa et al., 2018; Zheng, Capra, Wolfson, & Yang, 2014). Previous studies have demonstrated that emerging big geo-data and their associated methods have been employed for sensing the spatial and social structure of urban environments with different degrees of success (Chen et al., 2017; Hu et al., 2016; Liu et al., 2017; Wu et al., 2018).

POIs are of practical importance in studying the spatial and social structure of urban environments because they can effectively infer land parcels with diverse functions (McKenzie & Janowicz, 2017) and high access from Map services (i.e., Google Map) (Yao, Li, Liu, Liu, Liang, Zhang, and Mai, 2017, Yao, Liu, Li, Zhang, Liang, Mai, and Zhang, 2017). Attempts have also been made to extract urban functional regions using POIs data, and several status quo works can be identified (Gao, Janowicz, & Couclelis, 2017; Jiang, Alves, Rodrigues, Ferreira Jr, & Pereira, 2015; Long & Shen, 2015; Yao, Li, et al., 2017; Yuan, Zheng, & Xie, 2012; Zhong et al., 2014). From a probabilistic topic model-based perspective, Yuan et al. (Yuan et al., 2012) first developed a DRoF framework to infer the territory of urban functional regions and identify the functional intensity in the regions by a clustering method and kernel density estimation using POIs data and GPS trajectory datasets. An analogy strategy from urban elements to NLP textual materials was introduced, where a region was regarded as a document, a function was deemed as a topic, human mobility related to the region was used as words, and POIs were treated as metadata. Thus, the urban regions and their functionality could be represented by a joint probability distribution of a latent Dirichlet allocation (LDA) topic model (Blei, Ng, & Jordan, 2003). More recently, based on Yuan's work, Gao et al. (Gao et al., 2017) improved the conventional topic model and proposed a popularity-based LDA model to study urban functional regions. Supported by POIs data, this novel framework incorporates location-based social network user check-ins into the LDA topic model to replace POIs frequencies as the determination of a region.

However, the LDA-based method is a great algorithm for topic modeling, but still has a theoretical limitation because LDA utilizes the bag of words (BoW) method in which each document is regarded as a vector of word frequencies to transform textual information to vector characteristics (Blei et al., 2003; Yang, Chua, & Sun, 2015). However, the BoW method does not consider the mutual position of the words and ignores the context relationships in documents, which might be potential and important information (Wallach, 2006; Zhang, Du, & Wang, 2017). For urban functional issues, LDA topic models take only POIs frequencies as the determination of a region's functionality and ignore inner spatial correlations. As a result, most of the POIs spatial information is not transformed as potential features in the unsupervised learning process. To address this problem, Yao et al. (Yao, Li, et al., 2017) first introduced the word embedding-based model Word2Vec into urban functional zoning issues. As another unsupervised method in the NLP field, Word2Vec is an open source language learning model that projects words to high-dimensional vector spaces based on context relationships in documents (Goldberg & Levy, 2014; Mikolov, Chen, Corrado, & Dean, 2013). In their work, a traffic analysis zone (TAZ)-based corpus was constructed using POIs. Moreover, POIs categories and TAZs were represented by high-dimensional embedding vectors. More recently, a novel embedding-based model called Place2Vec, that extends the Word2Vec model, has been proposed to learn POIs category representations (Yan, Janowicz, Mai, & Gao, 2017). Based on the Place2Vec model and POIs data, Zhai et al. (Zhai et al., 2019) constructed a neighborhood area–based corpus and trained high-dimensional characteristic vectors of POIs categories to identify urban functional regions.

Although the Word2Vec model overcomes the deficiencies of the LDA topic models and considers the inner spatial correlations of urban functional regions, this model ignores the ubiquitous homonymy and polysemy issues of words and embeds each word using only a single vector (Yan et al., 2017; Yang et al., 2015; Zhai et al., 2019). Similarly, homonymy and polysemy issues of urban regions should be considered when extracting urban functional regions based on the NLP model. For example, restaurants are found both in residential areas and in business districts, but the semantic signatures (i.e., the functional intensity and service groups) might be distinctive (Gao et al., 2017).

In this study, we aim to infer urban functional regions by integrating a topic word embedding (TWE) model and POIs data to address these issues:

- HSR remote sensing data are good at extracting physical characteristics but difficult to be obtained.
- We consider ubiquitous homonymy and polysemy of urban regions based on the NLP model.

We assume that POIs categories in land parcels with diverse functions, such as residential, commercial, industrial, transportation, business and infrastructure, can be represented by high-dimensional vectors. Then, a spatial clustering method is introduced to aggregate the parcel functions. Finally, a case study is conducted to evaluate our method. Comparing with existing studies that infer urban functional regions with emerging geo-data, our method has the following advantages. First, based on an open source language learning model Word2Vec, this study employs a widely used POIs data with high spatiotemporal resolution. POIs data is available for almost all major map services (such as Google Map Service, OpenStreetMap Service and Gaode Map Services), while the source code and implementation of Word2Vec are provided by Google in an open way. As a result of their high accessibility, the research scale and throughput can be not limited by data and method. Second, our study employs a training-then-cluster strategy. Comparing with existing studies, this strategy can take advantage of machine learning in dealing with big data and feature representation. Third, comparing with the previous LDA-based method, our method addresses the spatial relationships between POIs and geographical context information in study area, instead of simple information of POIs' frequencies. Furthermore, our study considers ubiquitous homonymy and polysemy of urban regions in different topics, instead of only embedding each word in a single way.

The remainder of this paper is organized as follows. In Section 2, the study area and POIs database are briefly reviewed. Section 3 introduces the proposed representation framework, including the intraurban functional corpus and a series of analyses methods. Section 4 depicts the results of our analyses. In Section 5, we draw conclusions and discuss future works.

2. Study area and data

The study area is the main urban area of Wuhan city, the capital of Hubei Province, which is located in central China. Wuhan city has a favored geographical location in the inland of China (depicted as Fig. 1a) and is the center city of urban agglomeration in the middle reaches of the Yangtze River. Wuhan has gradually become a major comprehensive economic hub in recent years, and the urban functions in the intraurban area are highly heterogeneous and mixed. Considering the majority of contributions to the citizens' activities in the city, the main urban area within the third ring road is investigated as the intraurban area in this study.

In this study, the POIs dataset is obtained by an open source data platform, Peking University Research Data (State Information Center, 2018). This dataset is derived from the Gaode Map Services (https://ditu.amap.com/), which is a famous map service provider in China, with a time span to September 30, 2018. We obtained all the POIs records with a total of > 65.2 million data and extracted a total of 467,325 POIs in the study area. In geographic information systems (GISs), a POIs can be a house, a shop, a bus stop, even a mailbox, and so on. After preprocessing, the data in our study contain 6 core fields: POIs name, multilevel categories, address, coordinate location (latitude and

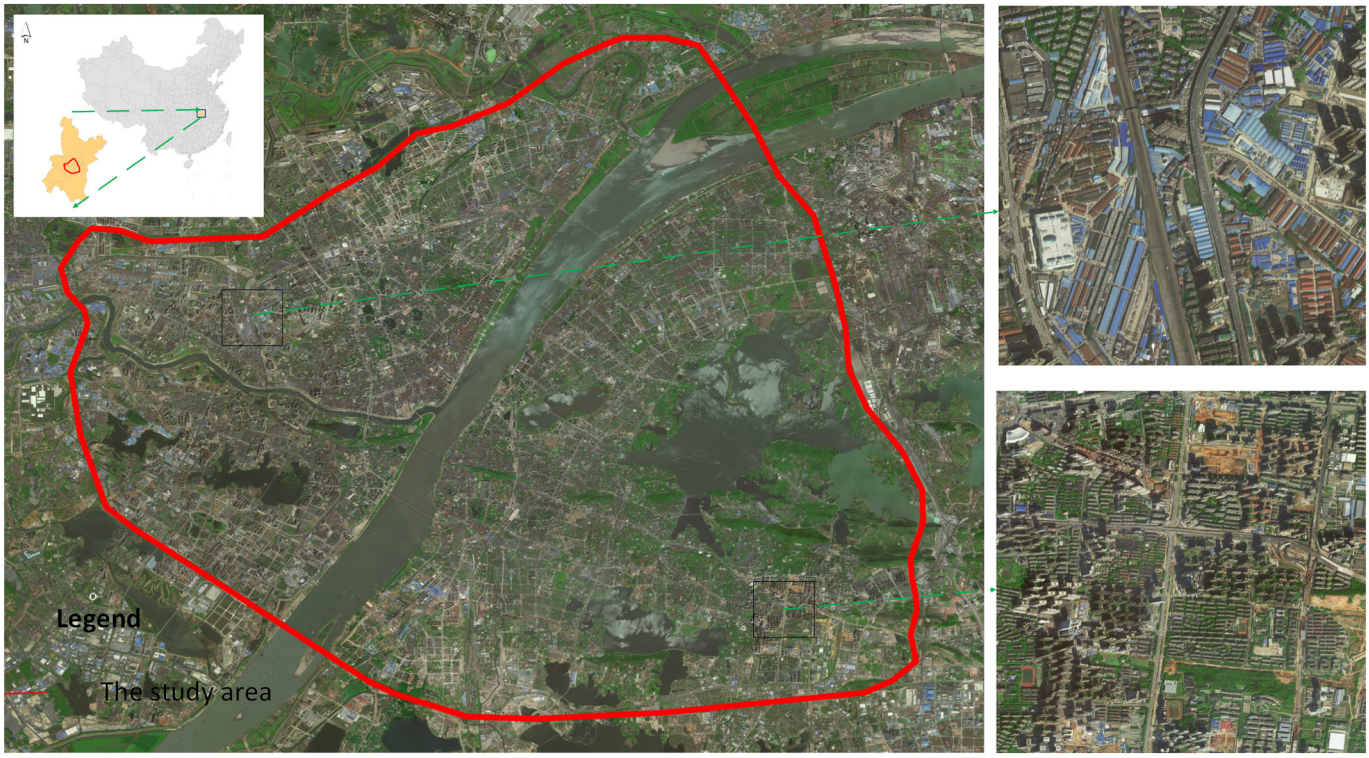

Fig. 1. The research area - the main urban area of Wuhan city. The red line denotes the third ring road, which depicts the main urban area in this study. The upper-left picture illustrates the location of Wuhan city in mainland China and depicts the study area in Wuhan city. The right pictures denote specific remote sensing (RS) images of the study area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

longitude), district name and publication time (Table 1). Among them, multilevel categories are mainly divided into three categories: top-level, second-level and third-level, where the descriptions of the POIs are provided in greater detail along with the category-level upgrades. In this POIs dataset, there are 16 labels in the top-level category (Table 2), 155 labels in the second-level category and > 869 labels in the third-level category. In our study, we use two level categories, the top and second-level category, are employed and ignore the third-level category because the third-level category is relatively complex and superfluous. For example, Chinese food restaurants in the second-level category can be divided into > 22 third-level labels, such as Sichuan food restaurants and Cantonese food restaurants, which seem to be unnecessary and could be ignored.

3. Methodology

In this work, a novel POIs-based framework is built for discovering the utilization of urban space, which consists of three basic parts: intraurban functional corpus, urban functional representation and functional parcel clustering. The flowchart of the proposed framework is illustrated in Fig. 2. The aim of our study is to analogize urban functional elements (e.g., POIs category) to NLP materials (e.g., word) to determine the utilization of urban space. First, an intraurban functional corpus is constructed via a center-context pair-based approach. Then, we implement the word embeddings representation model to train the above corpus to transform the second-level POIs category into multi-prototype vectors. Finally, we introduced a spatial clustering method based on a cosine distance similarity measurement to aggregate the functional parcel and employ the top-level POIs category to identify the clusters.

3.1. Building of intraurban functional corpus

In the field of NLP or linguistics, a corpus often refers to a large collection of well-sampled and structured sets of texts where language studies, theoretical or applied, can be conducted with the aid of computer tools (Ng and Zelle, 1997). A corpus may contain many documents, while documents may consist of many words. Moreover, the frequency and the sequence arrangements of words in a document usually reflect the particular contextual relationships and semantics of this document. In our study, we analogize the study area as a corpus and POIs categories as words. Thus, we assume that the spatial distributions and interactions of different POIs categories reflect the context of geographical space and the utilization of urban space. To construct urban functional documents in the context of the corpus of the study area, a center-context pairs-based approach is introduced. This approach was also adopted by Gao et al., 2017; Yan et al., 2017; Zhai et al., 2019 in recent GIS studies. First, 1500 searching points were randomly selected within the study area as the searching center of the urban functional documents based on the unduplicated and evenly spatially distributed selection criterion of the searching center (Fig. 3a). Then, the surrounding POIs around each searching center were identified using a nearest neighbor approach (Fig. 3b). To identify enough contextual information and control nonoverlap of POIs between searching pairs considering the study area, the number of surrounding POIs around a searching center was set to 150. Thus, 1500 center-context searching pairs were constructed and regarded as urban functional documents with urban spatial contextual information.

Considering the importance of the POIs spatial distributions in a searching pair, to assign each document to realistic meanings and associate the words in an organized form, we introduce a shortest path method, proposed by Yao et al., to reorganize the arrangement of words in each document. Based on the greedy algorithm and the shortest path algorithm, this method associates POIs by geographical relationships (e.g., distance and direction) and illustrates the spatial distribution characteristics and positional relationships between POIs to some extent. This method first constructs the shortest path that passes all the contextual POIs in each searching pair and then records these POIs in

Table 1
Examples of POIs dataset in the study area.

Name	Categories			Address	Coordinate location		District name	Publish time
	Top-level	Second-level	Third-level		Latitude	Longitude		
Baodao Park	Tourism Attraction	Park & Square	Park	Taibei 2nd. Road	114.284	30.598	Jiang'an District	2018-07-07 15:16:52
Champagne International Hotel	Residence	Hotel	Five-star Hotel	Hanxi 3rd. Road	114.223	30.590	Hankou District	2018-07-07 15:18:28
China University of Geosciences	Science and Education	School	University & College	388, Lumo Road	114.399	30.529	Hongshan District	2018-07-07 15:50:51

sequential order. Using the new ordered sequences, the final urban functional documents for training are constructed. More details about this method can be found in Yao's work (Yao, Li, et al., 2017).

3.2. Word embeddings representation model

Word embedding refers to a kind of language feature learning technique that has been shown to boost the performance in NLP tasks such as sentiment analysis (Socher et al., 2013). In this technique, the vocabulary words can be mapped to vectors of real numbers in a continuous space via neural networks or dimensionality reduction methods. Word2Vec, proposed by Mikolov et al. (Mikolov, Chen, Corrado, & Dean, 2015), is an open-source, state-of-the-art model that produces word embedding. Taking a large corpus as input, by training two-layer neural networks (NNs), linguistic contexts of words are reconstructed, and each word is represented as a unique vector. Additionally, the Word2Vec model provides two mathematical models, skip-gram and continuous BoW, for training NNs. Due to the ease of operation and high scalability of this model, Word2Vec is widely used, and there are also many variants. Specifically, in the Word2Vec model (or most word embedding models), each word, including polysemous or homophonic words, is only embedded as a single vector, causing the problem that part of the word discrimination is insufficient. For example, the word *apple* means a kind of fruit in a food-topic document while it might mean Apple Inc. in the topic of information technology. However, this word (*apple*) was typically represented using only a single embedding vector in the Word2Vec model (Yang et al., 2015).

To solve this problem, Yang et al. proposed a flexible and high-performance framework for multiprototype word representation called TWE (Yang et al., 2015). This framework assumes that under different topics derived from LDA (Blei et al., 2003), each word can be represented as different embeddings. Given a sequence of words $S = \{w_1, w_2, \ldots, w_n\}$, each word token w_i is discriminated into a specific topic t_i after Gibbs sampling (Griffiths & Steyvers, 2004) in LDA, and finally, a word-topic pair $< w_i,\ t_i >$ is formed, which is used to learn TWE. To learn topical word vectors, three TWE models, TWE-1, TWE-2 and TWE-3, were designed in terms of different training methods. (Yang et al., 2015) provided detailed information on the TWE framework and the toolkit. According to the requirements of analysis and consideration of model complexities, the TWE-1 and TWE-2 models are introduced and used for clustering tasks and correlation analysis, respectively, in this paper. For a better understanding and comparison, here we give a brief introduction of skip-gram, TWE-1 and TWE-2.

Skip-gram is a powerful and common mathematical model for learning word representations in which the goal is to predict context words for a given target word in a sliding window (Mikolov et al., 2013). By embedding each word as a unique word vector, the vectors of the target word are used as features to predict the context words. The maximum likelihood function of the skip-gram can be estimated as:

$$I(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p(w_i \mid w_{i-s}^{i+s}) \tag{1}$$

where s denotes the window size, n denotes the number of words vocabulary, and $w_{i-s}{}^{i+s}$ denotes context words of target word w_i. The softmax function is used in skip-gram for computing the probability $p(w_i | w_{i-s}{}^{i+s})$ as:

$$p(w_i \mid w_{i-s}^{i+s}) = \frac{\exp(w_i, w_{i-s}^{i+s})}{\frac{1}{n}\sum_{i=1}^{n} \exp(w_i, w_{i-s}^{i+s})} \tag{2}$$

Based on the basic skip-gram model, TWE-1 and TWE-2 have been expanded to different degrees and aspects. Typically, using the word-topic pair $< w_i,\ t_i >$, TWE-1 attempts to learn vector embeddings for words and topics separately and simultaneously as:

Table 2
The top-level category taxonomy of POIs.

No.	Top-level category	Abbreviation
1	Transportation facilities	TrabsFac
2	Residence	Residen
3	Sports/Recreation	Spr/Rec
4	Public facility	PubFac
5	Corporate business/Factory	Cop/Fact
6	Medical service	MedServ
7	Business building	BusBuil
8	Governmental and Public organizations	Gov/Pub
9	Daily life service place	LifeServ
10	Science and Education	Sci/Edu
11	Shopping mall	ShopMal
12	Car service	CarServ
13	Road facility	RoadFac
14	Bank/Financial	Bank/Fina
15	Tourism attraction	TourAtr
16	Food and Beverage place	FooBeve

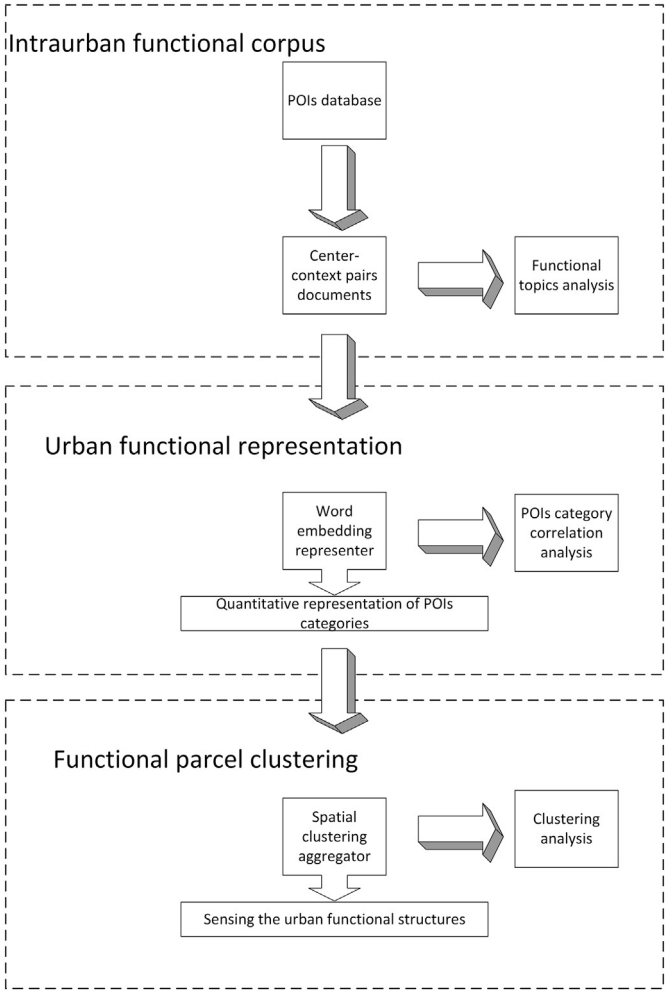

Fig. 2. The proposed framework flowchart.

$$I(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p(w_i \mid w_{i-s}^{i+s}) + \log p(w_i \mid t_i) \tag{3}$$

In TWE-1, the topic probability of topic t_i, denoted by $p(w_i|t_i)$ obtained from LDA, helps to predict context words. Based on a word w and the topic t, the topical word embedding w^t can be denoted as $w^t = w \oplus t$. Meanwhile, by aggregating over all TWE of each word in a specific document, the document embedding d can be denoted as $d = \Sigma_{w \in d} w' w^t$, where w' represents the weight of the words in d, which can be weighted with TF-IDF scores (Aizawa, 2003).

In contrast to TWE-1, TWE-2 regards a word-topic pair $< w_i, t_i >$ as a pseudoword and learns a unique embedding w^t. The maximum likelihood function of the TWE-2 can be estimated as:

$$I(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p(<w_{i+s}, t_{i+s}> | <w_i, t_i>) \tag{4}$$

3.3. Similarity measurement and clustering method

In high-dimensional vector space, given two high-dimensional vectors w_i and w_j, the measurement of word similarity is usually calculated by the cosine distance, denoted as:

$$S(w_i, w_j) = 1 - \cos(\theta) = 1 - \frac{w_i \cdot w_j}{|w_i||w_j|} \tag{5}$$

The change interval of $S(w_i, w_j)$ is $[-1, 1]$. In our study, a cosine distance-based similarity measurement was introduced to implement the correlation analysis of POIs categories and the preprocessing of the clustering method.

In the embedding space, vectors of the word have features such as high dimensionality, denseness, and unclear semantics. A single word vector does not explicitly express the semantics of a word. Therefore, it is particularly important to find an efficient clustering method, which can divide the high-dimensional embedding space based on the cosine-based distance and accurately cluster the vectors of words, thus effectively mining the internal features of the word vectors and clustering attributes. In response to this issue, this paper introduces a robust, state-of-the-art clustering algorithm, HDBSCAN. HDBSCAN is a hierarchical density-based clustering algorithm. While inheriting all the benefits of DBSCAN (Campello, Moulavi, Zimek, & Sander, 2015; McInnes, Healy, & Astels, 2017), a classic density-based cluster algorithm, this algorithm allows varying density clusters by condensing the dendrogram into a smaller tree that is used to select the most stable clusters. More details about the HDBSCAN algorithm can be found in McInnes's work (McInnes et al., 2017).

Compared with the popular clustering algorithm applied in existing studies (i.e., *k*-means) that discovers urban functional parcels using big geo-data, the HDBSCAN algorithm has the following advantages. First, HDBSCAN is a density-based spatial clustering algorithm that can find clusters of arbitrary shapes and is insensitive to noise data. HDBSCAN is stable over runs and subsampling and has good stability over parameter choices. Second, the HDBSCAN algorithm is friendly for parameter setting. As is known, the number of clusters is a hard parameter to obtain for the *k*-means method. In DBSCAN, the values of *min samples* and *eps* are also difficult to select. There are two crucial parameters: *min samples* and *min cluster size* in HDBSCAN. *min samples* is the parameter inherited from DBSCAN for the density-based space transformation, while *min cluster size* is an unintuitive parameter for one that is not hard to choose. Third, importantly, HDBSCAN is noise aware; it has an assumption of data samples that are not assigned to any cluster by calculating the cluster membership score ranging from 0.0 to 1.0. A score of 0.0 represents a sample that is not in the cluster at all, while a score of 1.0 represents a sample that is at the heart of the cluster. In our study, noise data represent mixed functional parcels with many functions. However, it seems difficult to distinguish the mixed functional parcels in *k*-means.

4. Implementation and results

4.1. Functional topic analysis

In the preprocessing LDA topic model, we set 100 as the total number of topics and ran 1000 iterations of the Gibbs sampling process to derive the posterior distribution over topic assignments. Fig. 4

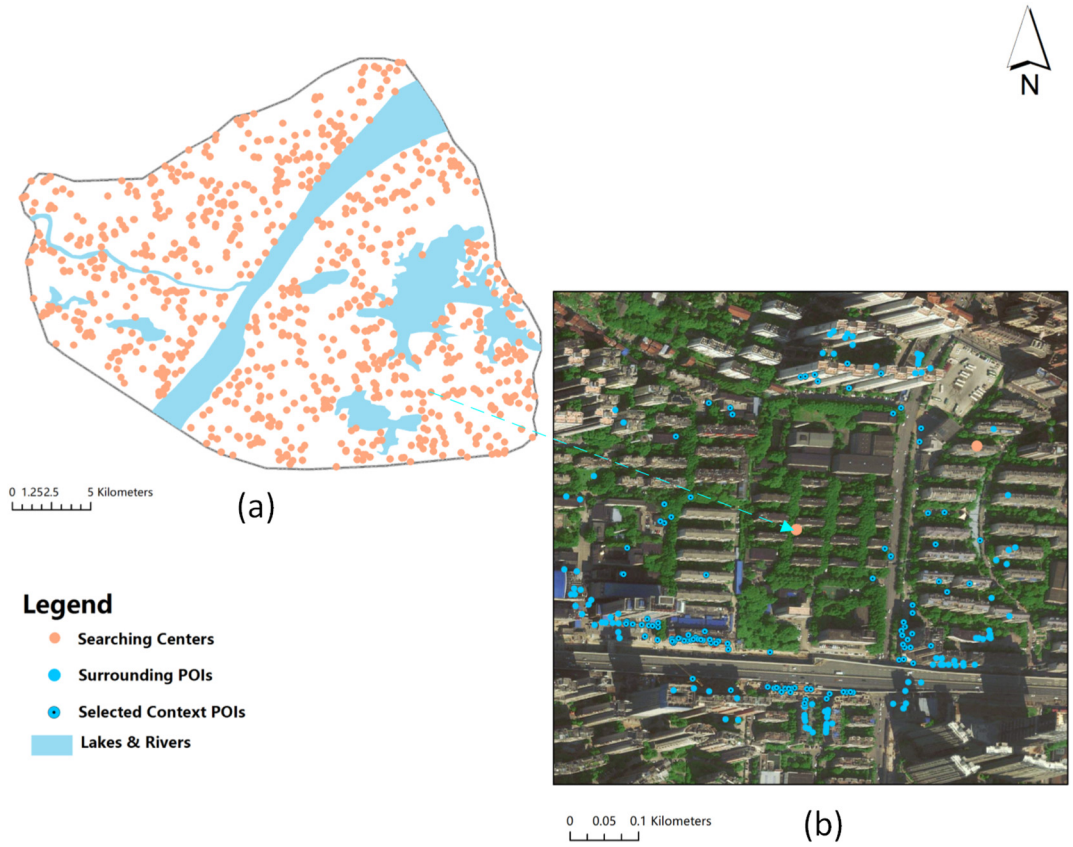

Fig. 3. The spatial locations of the center-context pairs. (a) illustrates the locations of 1500 unduplicated searching points. (b) illustrates the locations of contextual POIs around the sample searching center.

displays twelve interesting topics related to urban functions.

As illustrated in Fig. 4, we find that different zones are represented as interesting topics that output POIs categories with certain probabilities. Specifically, topic 4 represents a financial district that consists of various frequently occurring POIs categories, including *financial institutions, commercial premises* and some *business companies*, such as *securities companies* and *insurance companies*. Topic 12 is a recreation-related topic that includes *recreation centers, Chinese restaurants and bath and massage centers* with high probabilities. Topic 13 represents a healthcare district that contains *hospital, special hospital, healthcare location and emergency center*. Topic 83 represents an educational district that consists of *research institution, convention & exhibition center, school, and library*. Topic 91 is a scenic-related topic that contains *scenic spot, ticket office, park & plaza, museum, cultural palace and holiday location*. Similarly, topic 21, topic 90, topic 32, topic 41, topic 74, topic 73 and topic 96 are fundamental topics across all cities, which may suggest residential district, convenience district, automobile service district, cultural district, industrial district, sports district and beverages district, respectively.

4.2. POIs category correlation analysis

To demonstrate multiprototype features of POIs categories, several example POIs categories were selected and used to find the most similar of example POIs categories in different topics by the TWE model. For comparison purposes, the skip-gram-based Word2Vec model was used to find similar POIs categories. Note that the similarity of POIs categories, measured by cosine distance, demonstrates the spatial correlation between different categories of POIs. A higher value of cosine-distance-based similarity value suggests a more pronounced correlation between POIs categories.

In Table 3, we show the most similar POIs categories of three typical example words, *ATM, parking lot and Chinese restaurant*, which are typical polysemous POIs categories in urban areas. For each example POIs category *p*, we first illustrate the result obtained from the skip-gram-based Word2Vec model; then, we list the results under two representative topics of the example POIs categories obtained from TWE-2, denoted as *p#topic-1* and *p#topic-2*.

As illustrated in Table 3, we find different similarities between POIs categories by different models in two aspects. From the homogeneity perspective, there are common items between similar categories returned by the TWE-2 model and the Word2Vec model. For example, the *recreation center and convenience store* exhibit significant correlations with *Chinese restaurants in* both the Word2Vec and TWE-2 models under topic 12 (a recreation-related topic in Fig. 4). *ATM* is highly related to *banks* in both models, which means that the correlation between different POIs categories can be explored by both the TWE model and the Word2Vec model to some extent.

From the heterogeneity perspective, the TWE model can demonstrate multiprototype features of POIs categories. Taking *ATM* as an example, we find that bank is the most similar category in the Word2Vec model. However, in different parcels with different topics (e.g., sports-related district and educational district, denoted by topic 73 and topic 83, respectively), the most similar categories change to *sports store or coffee shop*. However, *bank* is still one of the similar categories if a *bank* exists in this parcel with a specific topic. Similar phenomena occur over *parking lots and Chinese restaurants*. The result indicates that the TWE model can discriminate multiprototype features of different topics related to urban functional regions.

4.3. Clustering results and analysis

A few trials were performed to determine appropriate values for the main parameters of the HDBSCAN cluster method, such as *min cluster*

Topic 4	
Category	Prob.
Finance Institution	0.7459
Commercial Premises	0.0578
Securities Company	0.0476
Insurance Company	0.0449
Travel Agency	0.0249
Palace	0.0240
Education Location	0.0169
Coffee Shop	0.0103
Barbershop	0.0045
Commercial Street	0.0040

Topic 12	
Category	Prob.
Recreation Center	0.5509
Chinese Restaurant	0.3094
Bath & Massage Center	0.0679
Pharmacy	0.0299
Convenience Store	0.0055
Pet Hospital	0.0050
Dessert Shop	0.0050
Travel Agency	0.0041
Stationary Store	0.0037
Training Institution	0.0028

Topic 13	
Category	Prob.
Hospital	0.4310
Parking Lot	0.2573
Special Hospital	0.0906
Healthcare Location	0.0900
Pharmacy	0.0499
Bus Station	0.0195
Emergency Center	0.0150
Palace	0.0127
Enterprises	0.0064
Governmental Organization	0.0058

Topic 21	
Category	Prob.
Residential Area	0.9601
School	0.0180
Supermarket	0.0036
Logistics	0.0027
Recreation Location	0.0018
Laundry	0.0014
Leisure Food Restaurant	0.0014
Education Location	0.0014
Theatre	0.0009
Shopping Locations	0.0005

Topic 90	
Category	Prob.
Life Service Location	0.4125
Convenience Store	0.3501
Logistics	0.1183
Shopping Locations	0.0282
Clinic	0.0231
Public Security Organization	0.0149
Telecom Office	0.0093
Electronics Market	0.0088
Lottery Store	0.0088
Supermarket	0.0031

Topic 32	
Category	Prob.
Auto Dealers	0.4286
Automobile Service Related	0.1574
Auto Repair	0.1533
Enterprises	0.1481
Used Car Market	0.0354
Finance Institution	0.0198
Filling Station	0.0172
Governmental & Socialgroups	0.0094
Energy Station	0.0078
Lottery Store	0.0068

Topic 41	
Category	Prob.
Parking Lot	0.5310
Park & Plaza	0.2495
Museum	0.0361
Ticket Office	0.0330
Coffee Shop	0.0284
Art Gallery	0.0264
Sports Stadium	0.0193
Theatre	0.0153
Transportation Related	0.0142
Education Location	0.0097

Topic 74	
Category	Prob.
Commercial Premises	0.6213
Enterprises	0.2348
Finance Institution	0.0337
Parking Lot	0.0239
Hotel	0.0192
Life Service Location	0.0109
Professional Service Firm	0.0083
Franchise Store	0.0063
Agency	0.0063
Science & Technology Museum	0.0042

Topic 73	
Category	Prob.
Sports Stadium	0.6844
Parking Lot	0.1416
Supermarket	0.0400
Education Location	0.0277
Sports Location	0.0251
Arts Organization	0.0155
Franchise Store	0.0107
Cultural Palace	0.0086
Sports Store	0.0075
Life Service Location	0.0054

Topic 83	
Category	Prob.
Research Institution	0.6375
School	0.3461
Convention & Exhibition Center	0.0052
Library	0.0026
Coffee Shop	0.0012
Franchise Store	0.0006
Arts Organization	0.0006
Recreation Location	0.0006
Public Security Organization	0.0006
Foreign Organization	0.0003

Topic 91	
Category	Prob.
Scenic Spot	0.9540
Ticket Office	0.0198
Park & Plaza	0.0117
Public Phone	0.0059
Museum	0.0020
Cultural Palace	0.0012
Holiday Location	0.0008
Sports Store	0.0004
Telecom Office	0.0004
Laundry	0.0002

Topic 96	
Category	Prob.
Icecream Shop	0.1953
Foreign Restaurant	0.1671
Food & Beverages Related	0.1269
Fast Food Restaurant	0.1218
Coffee Shop	0.1019
Bakery	0.0967
Dessert Shop	0.0633
Sports Location	0.0267
Photo Finishing	0.0267
Commercial Street	0.0152

Fig. 4. Twelve interesting topics with their top-10 ranked POIs types related to urban functional regions.

Table 3
Similar POIs categories of example categories by TWE-2 and word2vec.

Example category	Similar categories
ATM	Bank, sports stadium, training institution
ATM#topic-73	Sports store, clinic, bank
ATM#topic-83	Coffee shop, bank, convention & exhibition center
Parking lot	Sports stadium, education location, scenic spot
Parking lot#topic-41	Park & plaza, ticket office, museum
Parking lot#topic-13	Hospital, special hospital, healthcare location
Chinese restaurant	Recreation center, barbershop, convenience store
Chinese restaurant#topic-12	Recreation center, bath & massage center, convenience store
Chinese restaurant#topic-14	Foreign restaurant, lottery store, commercial location

size (denoted by c) and *min samples* (denoted by s) in our case study, which is achieved by running the cosine-distance measurement-based HDBSCAN cluster method with different values of c (ranging from 2 to 126) and s (ranging from 1 to 126) and evaluating the results based on the silhouette metric and the number of clusters. Moreover, to evaluate the interpretation and consistency of the HDBSCAN results, a silhouette metric is introduced to determine the main parameters of HDBSCAN (Rousseeuw, 1987). Note that the silhouette metric value ranges from 0 to 1, and a great silhouette metric value with stable contextual trends suggests a better assignment for the cluster in this study.

As illustrated in Fig. 5, we find that as c and s increase separately; from a global perspective, the number of clusters decreases in an elbow shape (depicted as an orange line), and the silhouette metric values decrease in a stepwise manner (depicted as a blue line). From the local perspective, the highest silhouette metric value is obtained when both $c=7$ and $s=7$, but at this point, the contextual trends of the silhouette metric values and the number of clusters tend to be extremely unstable. However, as the value of c changes, the contextual trends of silhouette metric values and the number of clusters show stability for the values of $c=50$ and $c=90$ (Fig. 5a). The same results occur when $s=50$, $s=78$ and $s=97$ (Fig. 5b). The fine-tuned result means that when the main parameters are set to above-specified values, the evaluation indicators (such as the silhouette metric and the number of clusters) can exhibit local and contextual invariance and stability to some extent; that is, the cosine-distance based HDBSCAN clustering results tend to be contextually stable and reliable to some extent. Therefore, in subsequent analyses, six clustering schemes are implemented using $c=50$ and $c=90$, and $s=50$, $s=78$ and $s=97$, as these specific clustering parameters correspond to reliable clusters.

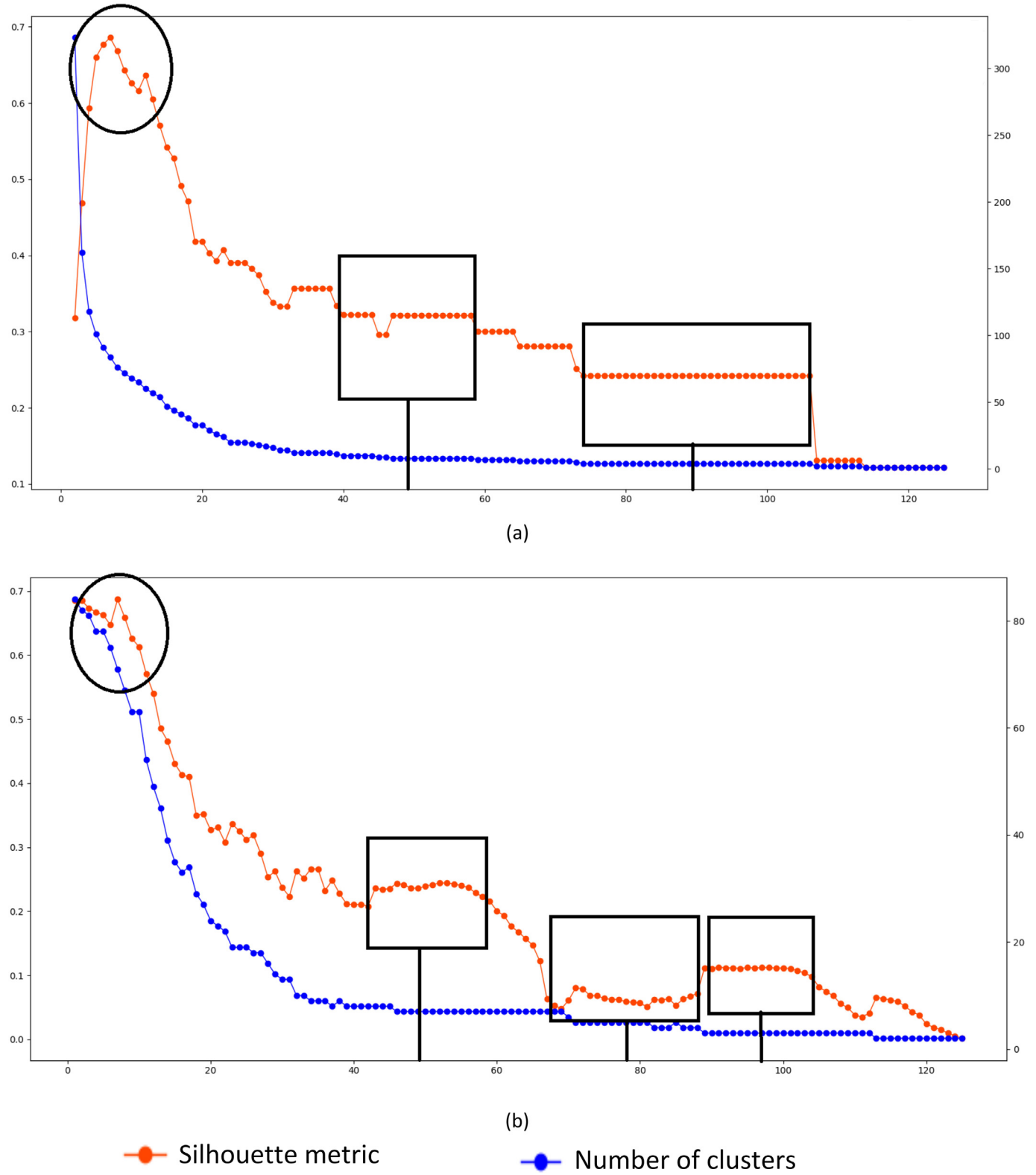

Fig. 5. The silhouette metric values (left side of y-axis) and the number of clusters (right side of y-axis) with different experimental results. (a) by varying the value of *c*. (b) by varying the value of *s*.

For a better understanding of the clustering results and to understand cluster patterns visually, two clustering schemes ($c=50$, $s=50$ and $c=90$, $s=50$) were selected to map the cluster results because the silhouette metric values of these two schemes are greater than those of others (Table 4).

Fig. 6 depicts the clustering results for the clustering scheme with $c=50$, $s=50$ (scheme A) and that with $c=90$, $s=50$ (scheme B). Fig. 5.a shows 4 diverse clusters in the study area, while Fig. 6.b shows 7 detailed clusters. Note that the POIs database is utilized to interpret the clustering results, and POIs enrichment factors are calculated for

Table 4
The silhouette metric value and the number of clusters of six clustering schemes.

Cluster parameters	Silhouette metric value	Number of clusters
$c=50$, $s=50$	0.226	7
$c=50$, $s=78$	0.103	4
$c=50$, $s=97$	0.112	4
$c=90$, $s=50$	0.174	4
$c=90$, $s=78$	0.103	4
$c=90$, $s=97$	0.112	4

individual clusters (Tables 5 and 6). The final types of clusters are determined through enrichment factors of POIs categories and analysis based on governmental land use planning and remote sensing images of the study area.

When $c=90$ and $s=50$, cluster 1 mainly represents a corporate business area or factory; cluster 2 can be considered a shopping mall; cluster 3 mostly contains hospitals, daily life service place and food and beverage locations; cluster 4 largely includes tourism attractions, science and education areas and transportation facilities. When $c=50$ and $s=50$, clusters tend to be more specific. Cluster 1 basically represents corporate business areas or factories; cluster 2 mainly represents shopping malls; cluster 3 mostly contains tourism attractions and public facilities; cluster 4 largely represents transportation facilities; cluster 5 represents science and education locations; cluster 6 mainly contains hospitals, food and beverage locations and daily life service place; cluster 7 largely represents governmental and public organizations.

Therefore, we found that cluster 1 and cluster 2 from scheme A and cluster 1 and cluster 2 from scheme B are identical, both in terms of geography (Fig. 6) and category characteristics (Table 7). Therefore, scheme B can be considered a further decomposition of clusters identified in scheme A. Moreover, via analysis of the region clustering, we can visually and clearly understand the distribution characteristics of urban land use. The region clustering results with clustering schemes B are annotated as follows (Fig. 7):

Cluster 1 (Corporate business areas and factories): In this cluster, the land use type mainly contains corporate business areas and factories, which are highly related to daily work and commuting. This land use type mainly relates to the POIs categories, including all kinds of companies and factories. In Table 6, we found that the value of the enrichment factor of this land use type in this cluster is as high as 3.524,

Table 5
Enrichment factors of POIs categories (EP) grouped by clusters for scheme A.

Land use types	Cluster1	Cluster2	Cluster3	Cluster4
TrabsFac	0.757	0.750	0.737	1.432
Residen	0.683	0.563	1.002	0.711
Spr/Rec	0.593	0.631	1.084	1.187
PubFac	1.110	0.625	0.655	1.561
Cop/Fact	3.524	1.017	0.592	0.517
MedServ	0.617	0.686	1.635	0.596
BusBuil	0.973	0.823	0.888	1.152
Gov/Pub	0.950	0.824	1.270	0.814
LifeServ	0.846	0.789	1.412	0.683
Sci/Edu	0.633	0.499	0.802	1.987
ShopMal	0.771	2.337	1.073	0.489
CarServ	0.964	0.812	0.669	0.392
RoadFac	0.554	0.887	1.226	0.799
Bank/Fina	1.087	0.461	0.901	0.797
TourAtr	0.131	0.127	0.190	4.060
FooBeve	0.611	0.705	1.469	0.742

far more than other land use types, which indicates the dominant position in this cluster. Moreover, from the factory's perspective, this land use type mainly contains industrial parks and high-tech parks, located in the surrounding areas of the third ring road where the land price is relatively low. From the corporate business area's perspective, this land use type largely includes all kinds of companies located in the central business areas.

Cluster 2 (Shopping malls): In this cluster, the land use type is single and largely contains the POIs categories such as supermarket, convenience store and all kinds of comprehensive markets. As a place to satisfy people's most basic shopping services, this land use type is usually closely related to the distribution of residential areas. Therefore, the distribution in the interior of the city is very uniform and widespread.

Cluster 3 (Tourism attractions and public facilities): This land use type mainly contains tourism attractions and public facilities, which correspond to the tourism-related POIs categories (*park, zoo*) and some public POIs categories (*public toilet and emergency shelter*). There are many tourist attractions in Wuhan, attracting many tourists every year. Moreover, we found that this type of land use presents a very distinct feature of clustering distribution and mainly clusters around East Lake and the Yangtze River, which are famous tourism places for sightseeing.

Cluster 4 (Transportation facilities): In this cluster, the land use

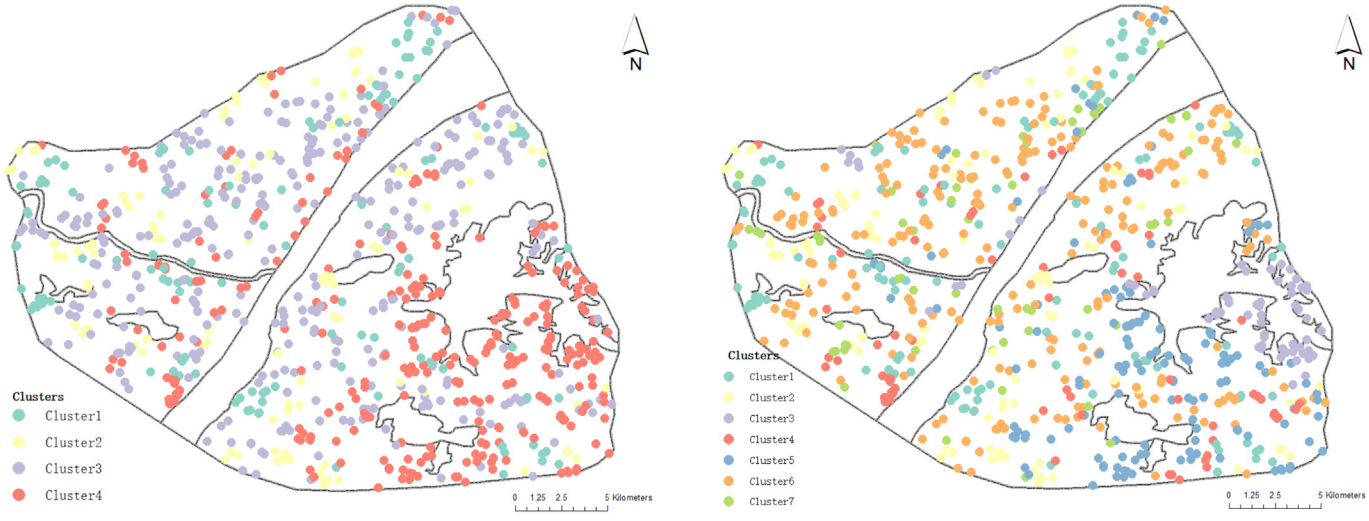

Fig. 6. Cosined-distance measurement-based HDBSCAN clustering results for (a) scheme A with $c=90$, $s=50$ and (b) scheme B with $c=50$, $s=50$.

Table 6
Enrichment factors of POIs categories grouped by clusters for scheme B.

Land use types	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
TrabsFac	0.757	0.750	1.279	2.413	1.048	0.719	0.797
Residen	0.683	0.563	0.560	0.811	0.741	1.056	0.760
Spr/Rec	0.593	0.631	1.075	1.165	1.168	1.170	0.769
PubFac	1.110	0.625	4.048	1.044	0.602	0.579	0.741
Cop/Fact	3.524	1.017	0.426	0.813	0.410	0.506	0.839
Hospital	0.617	0.686	0.242	0.739	0.674	1.748	1.162
BusBuil	0.973	0.823	0.475	1.055	1.510	0.852	1.002
Gov/Pub	0.950	0.824	0.862	0.656	0.874	0.764	3.258
LifeServ	0.846	0.789	0.440	0.916	0.644	1.513	1.120
Sci/Edu	0.633	0.499	0.492	0.981	3.309	0.795	0.754
ShopMal	0.771	2.337	0.206	0.718	0.471	1.145	0.840
CarServ	0.964	0.812	0.251	0.636	0.293	0.649	0.762
RoadFac	0.554	0.887	0.453	0.868	0.990	1.205	1.294
Bank/Fina	1.087	0.461	0.248	1.052	0.919	0.977	0.650
TourAtr	0.131	0.127	15.429	0.215	0.755	0.119	0.191
FooBeve	0.611	0.705	0.355	0.828	0.861	1.661	0.817

Table 7
The correspondence between land use types and clusters.

Land use types	Scheme A	Scheme B
Corporate business area or factory	Cluster 1	Cluster 1
Shopping mall	Cluster 2	Cluster 2
Tourism attraction place	Cluster 3	Cluster 4
Public facility	Cluster 3	Cluster 4
Transportation facility	Cluster 4	Cluster 4
Science and education place	Cluster 5	Cluster 4
Medical services place, food and beverage place and daily life service place	Cluster 6	Cluster 3
Governmental and public organization	Cluster 7	Cluster 3

type basically represents transportation facilities, which consist of transportation-related POIs categories such as railway stations, ports, bus stations, parking lots and other stations. The distributions of these POIs categories are widespread and anthropogenic. Moreover, this distribution partly reveals the underdeveloped regions of transportation in the study area.

Cluster 5 (Science and education locations): In this cluster, this land use type mainly relates to schools, institutions, libraries, museums and other education-related places. Wuhan has many colleges and universities so that the distribution of this land use type is widespread and clustered.

Cluster 6 (Medical services, food and beverage and daily life service place): Compared with Cluster 2 and Cluster 4, this cluster tends to be more complex and mixed. In this cluster, the land use type contains medical services locations (such as *hospitals, clinics and health care locations*), food and beverage locations (such as all kinds of *restaurants, coffee houses and dessert stores*), daily life service place (such as *post office, telecom office and express distribution center*). These POIs categories are closely related to citizens' daily lives and are widely spread around any corner of a city. However, the cluster is mixed, and details could not be devised in our study area.

Cluster 7 (Governmental and public organizations): In this cluster, the land use type mainly means governmental and public organizations, including the POIs categories such as *social groups, governmental organizations, security organizations and other public institutions*. This land use type is largely located in the central area of the city area to facilitate better services for urban residents.

5. Discussion

In this study, we attempt to infer urban functional regions using POIs data by integrating the TWE model and HDBSCAN clustering method. In this section, we firstly discuss the reason word embedding model is implemented in two aspects. On the one hand, the proliferation of crowdsourcing technology and the emergence of large-scale datasets bring unprecedented opportunities for the development of embedding learning and unsupervised word embeddings in the NLP field. Based on an analogy strategy (from urban elements to NLP textual materials), we could make full use of the advantages of training neural networks and embedding representation method to tackle problems in urban studies, and recent studies also prove that this analogy strategy has achieved great success in many fields. On the other hand, word embedding can project words to high-dimensional vector spaces based on context relationships in documents. Therefore, via word embedding model and analogy framework, high-dimension real valued vectors of POIs categories can be obtained to quantify the characteristics of POIs and urban space. Using these quantified vectors, the researchers can then use it for urban clustering, classification or visualization for urban studies.

Moreover, in this study, we try to explore an NLP-based plan for inferring urban functional regions mostly considering ubiquitous homonymy and polysemy of urban regions. Simultaneously, as an improvement of existing NLP-based method, some comparisons between proposed method and existing baseline method have been implemented. Three types of embedding vectors, specifically LDA-based, Word2Vec-based and TWE-based, are obtained. Before clustering, we introduce Hopkins statistic test to determine whether given data (above vectors) contains significant clusters or not. Hopkins statistic metric (denoted by H) tests the spatial randomness of the embedding vector by measuring the probability that this vector is generated by a uniform data distribution. Note that A value for this metric higher than 0.75 indicates a clustering tendency at the 90% confidence level and a higher value suggests a better result. Table 8 shows that of H values of three types of embedding vectors are both higher than 0.75 and TWE-based vectors has the highest H value.

Furthermore, cosined distance-based k-means clustering is conducted using these three types of embedding vectors and silhouette metric value is used. We run this method with different values of k (range from 2 to 10) 10 rounds and evaluate the clustering result using the average values. Fig. 8 shows that the silhouette metric values obtained by the k-means clustering based on TWE representation are much higher than that based on Word2Vec and LDA at mostly k values. The highest silhouette metric value is obtained when $k = 2$ and as the value of k increases, the silhouette metric value increases significantly. Therefore, based on the results of both Hopkins statistic test and silhouette metric, our proposed embedding plan is more robust to clustering and obtains a better cluster effect.

Second, we discuss the choices of the clustering method. Clustering algorithms have deep and extensive applications in GIS studies and urban planning, especially in exploring the utilization of urban space

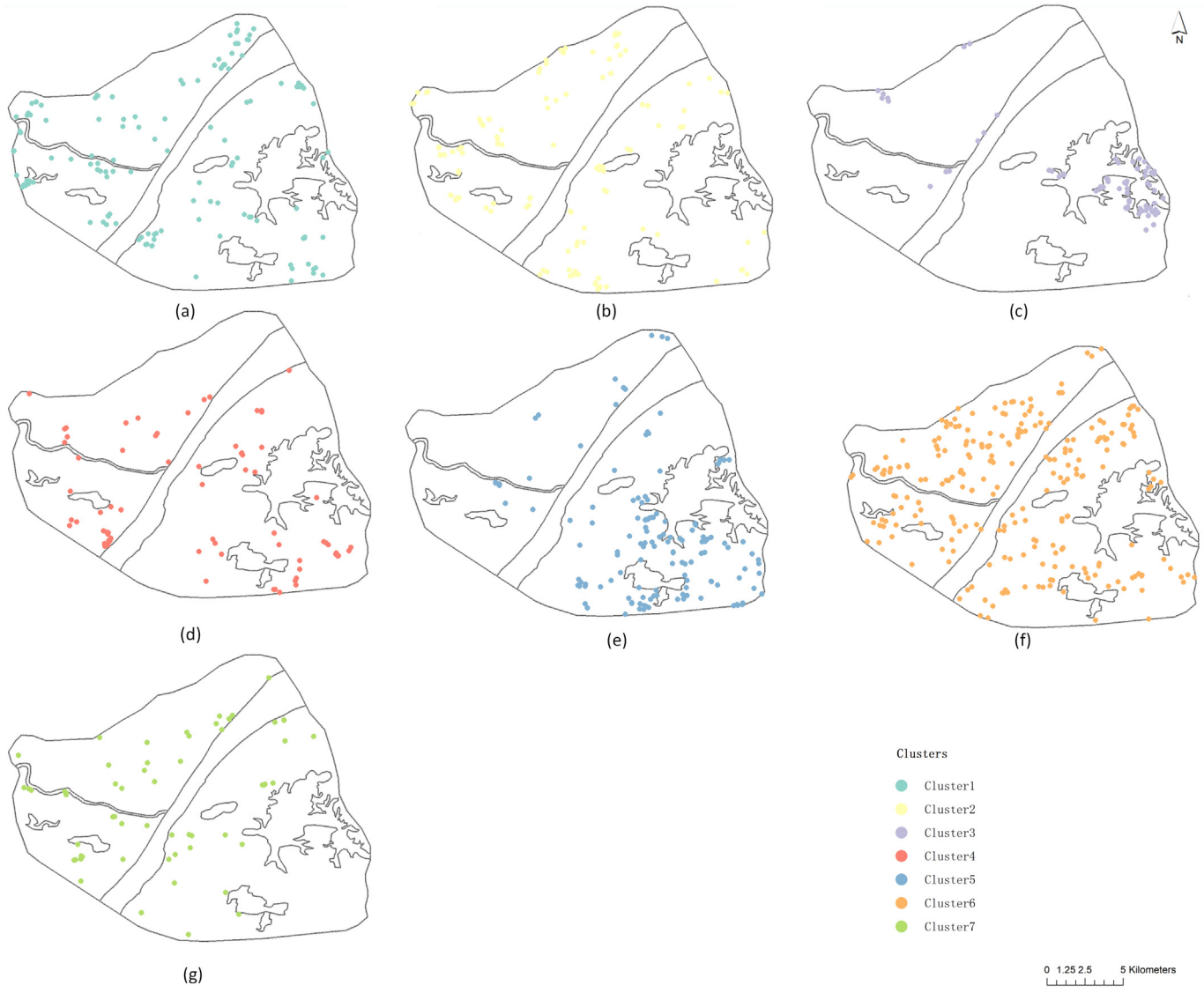

Fig. 7. The separated geographical distribution of (a) clusters 1 to (g) 7.

Table 8
The Hopkins statistic test of three types of embedding vector.

Embedding vector	Hopkins statistic metric (*H*)
LDA-based	0.925
Word2Vec-based	0.965
TWE-based	0.977

and land use. Based on mobile phone call data, Pei et al. (Pei et al., 2014) first constructed a vector of aggregated mobile phone data composed of a time factor and total volume factor to character the activities of residents and then applied a semisupervised fuzzy *c*-means clustering method to determine the land use in Singapore. Yao et al. (Yao, Li, et al., 2017) implemented a *k*-means-based clustering model to aggregate the urban functions region based on the POIs database. In addition, Chen et al. (Chen et al., 2017) introduced a dynamic time warping distance-based *k*-medoid clustering method, a modified *k*-means algorithm, to delineate urban functional areas. The application of the *k*-means method and its modified clustering method has achieved great success in GIS studies because these clustering methods are easy to implement and expand. However, these clustering methods ignore the data features (such as the special characteristics), and it is difficult

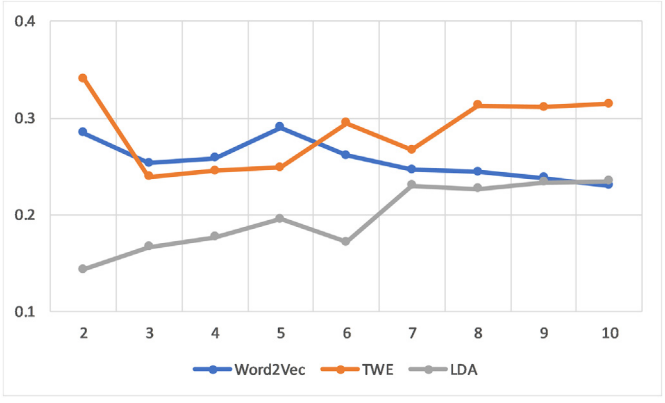

Fig. 8. The silhouette metric values of cosined distance-based *k*-means clustering method based on different embedding vectors.

to initialize and adjust the parameters (such as the number of clusters). In our case study, we also implement the *k*-means-based clustering model with the cosine distance similarity measurement-based on the high-dimensional word embedding vectors. Fig. 9 shows the value of

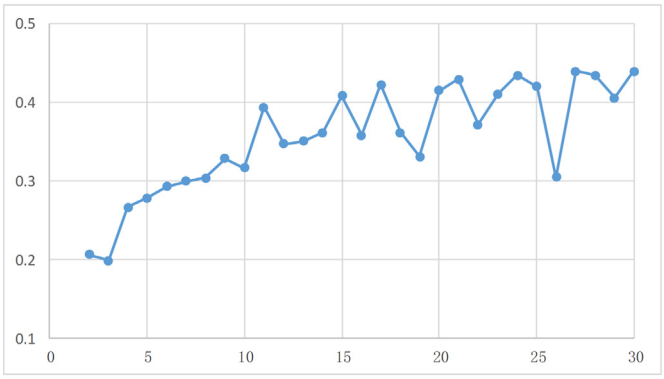

Fig. 9. The silhouette metric values of different experimental results by varying the number of clusters *k*.

evaluation with different *k* values. The result reveals that as the value of *k* changes, the value of the evaluation criteria remains extremely volatile, resulting in difficulty in finding the elbow points, which are usually considered to be points with better clustering results (Ketchen & Shook, 1996). Therefore, the *k* value, the number of clusters for subsequent analysis, is difficult to define when the *k*-means-based clustering method is applied.

In addition, we discuss the regional division method of urban space in the section of building an urban functional corpus. There are three popular regional division methods in related urban studies and urban planning: grid-based division, traffic area zone-based division and center-context pair-based division. In our study, a center-context pair-based approach is utilized to divide urban space into functional parcels as basic research units for two reasons: grid-based division and traffic area zone-based division generally suffer from controversial issues in response to the scale of division. For example, in a grid-based division approach, the length of each grid, which has a great influence on subsequent clustering results and analysis, is difficult to define. In contrast, the center-context pairs-based approach does not require any specific boundary (such as grid net and traffic lines) to divide the area but rather a method for randomly generating many searching centers and constructing the surrounding context by spatial POIs distributions. This approach can construct more flexible training samples by altering the number of searching centers and be free from the above issue.

The approach and the results of our study can be used to help urban researchers understand, represent, and reason the spatial and social structure of urban environments and help urban planners design better urbanization strategies for the future to better serve residents and their cities and improve urban vitality. For example, we further examine the urban functional diversity at two different levels using our clustering results in the study area. We construct a regular grid net of 1 km × 1 km and the main roads to divide our study area. Based on the calculation of the cluster enrichment factors in each grid and TAZ, we identify the diversity of urban functional areas and their geographical characteristics and contextual relations. As shown in Fig. 10, our study area is divided into 544 grids, and each grid is denoted by an attribute of clusters. The *other* item in the clusters represents the mixed area that is difficult to assign a single urban function. Fig. 11 displays the TAZ-level urban functional diversity in the study area based on traffic analysis zones.

To examine the applicability of our proposed framework, based on POI database of OSM we performed our method on a European city - Munich, one of financial, cultural and technological central cities of Germany with high density areas and mix-used urban functions. Fig. 12 shows the clustering results with clustering scheme of 7 clusters. We found a significant geographical clustering pattern in Munich. Therefore, the clustering results in Munich reveal that our proposed framework can effectively sense the spatial structures of urban space with high density areas.

An understanding of the spatial and social structure of urban environments, especially intraurban functions, is one of the central themes in urbanology and geography and is of considerable conceptual and practical interest because the question is intimately related to the problem of urban planning, government management and sustainable development. This study seeks to contribute to the literature in three ways. First, via a series of basic statistic and correlation analyses, this study can provide new perspectives for understanding the context of geographical space of intraurban areas and an alternative method for

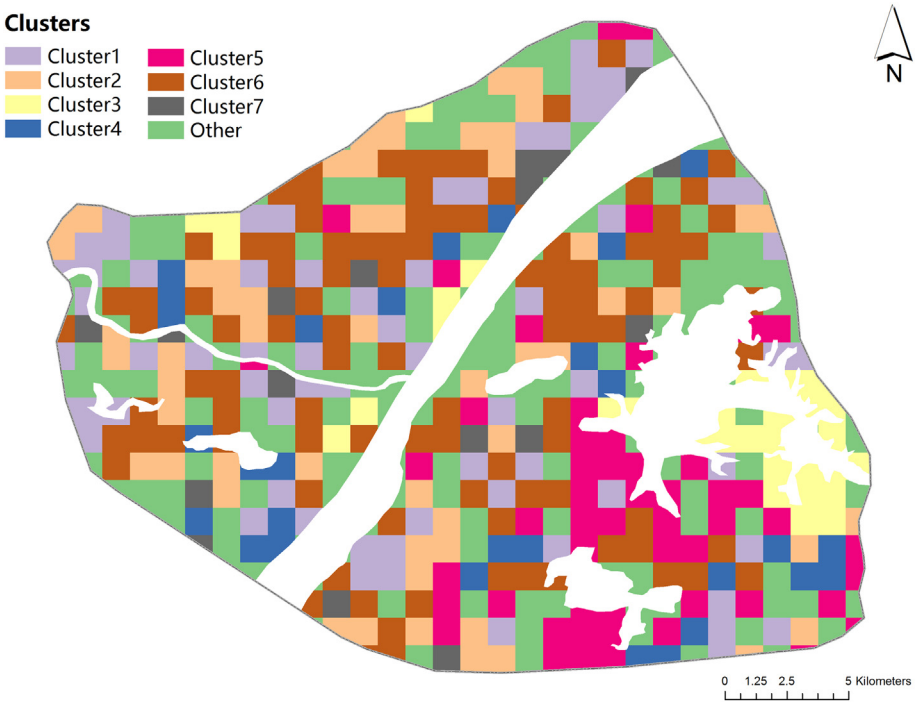

Fig. 10. The grid-level urban functional diversity in the study area based on a 1 km × 1 km grid net. The blank area indicates rivers and lakes.

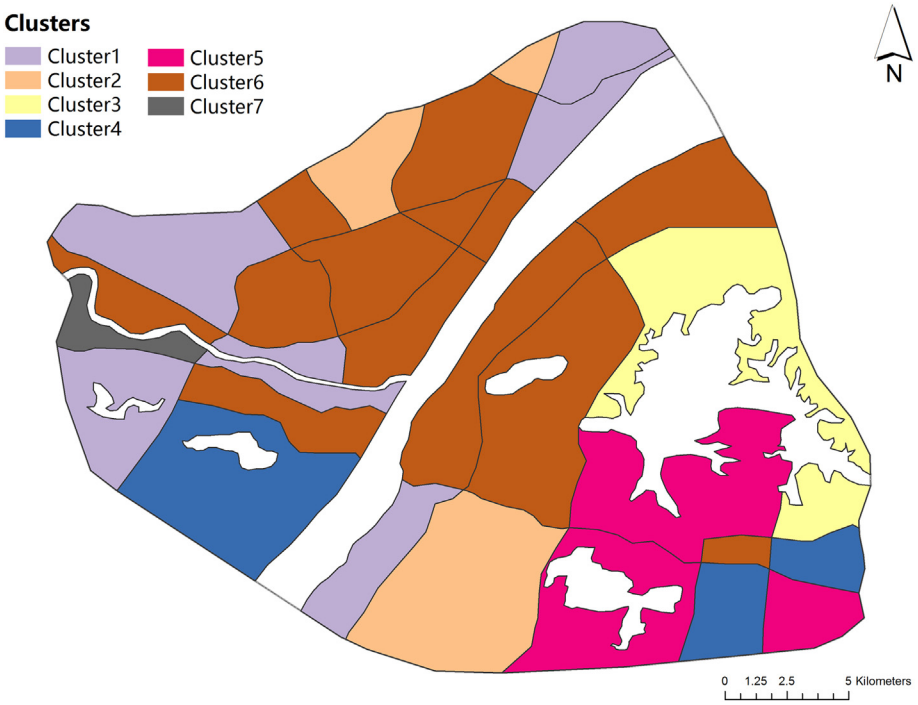

Fig. 11. The TAZ-level urban functional diversity in the study area based on traffic analysis zones. The blank area indicates rivers and lakes.

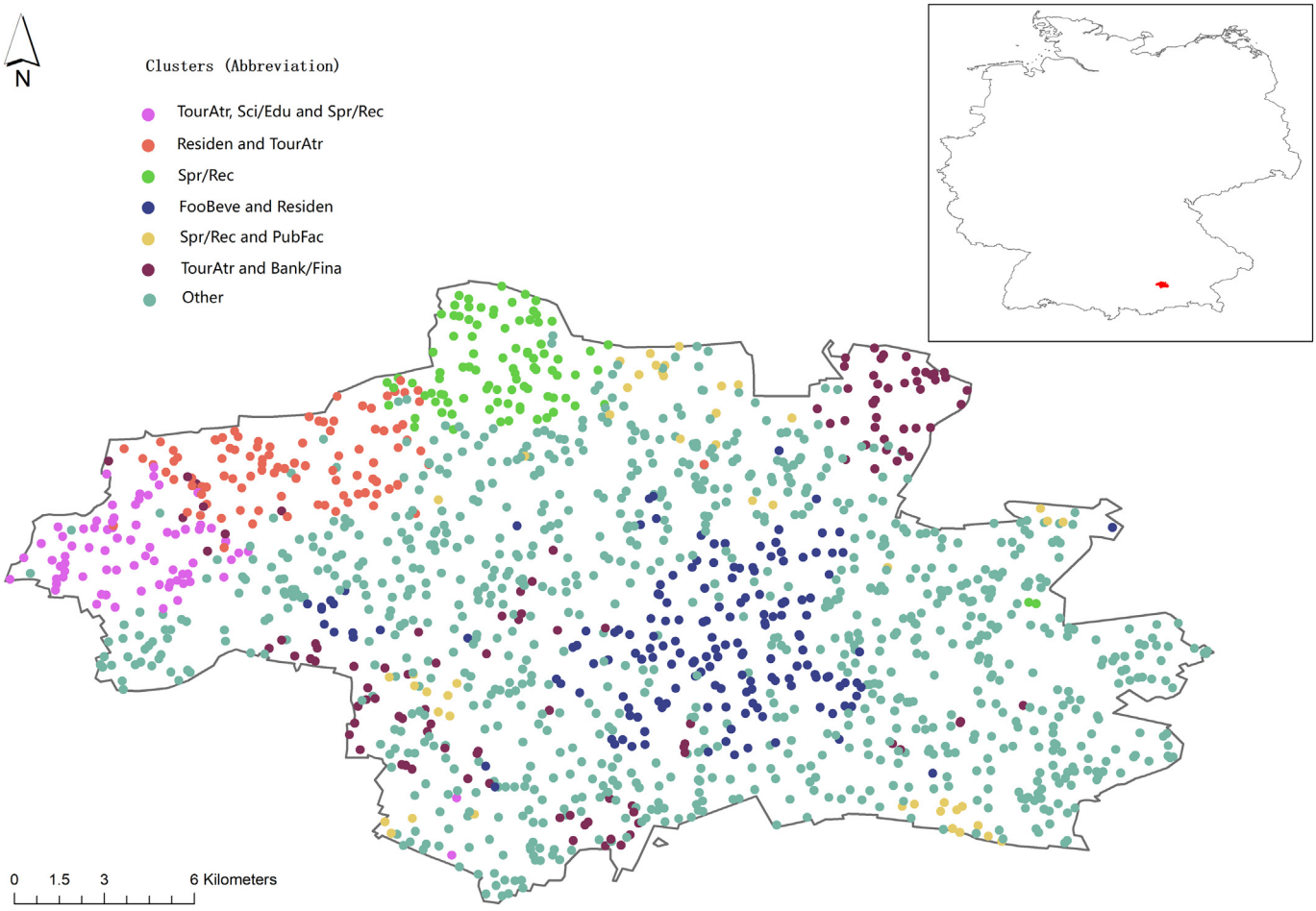

Fig. 12. Clustering results using our proposed method in Munich, Germany. The meaning of clusters abbreviation shows in Table 2.

conducting large-scale emerging geo-data collection (such as POIs data) to explore urban functional regions. Second, the study develops a systematical and integral framework using POIs data and an embedding learning method. The framework has two distinctive features. (1) In the framework, we analogize the urban elements to textual materials in the NLP field and provide multiprototype high-dimensional vectors. (2) We introduce an open-sourced and state-of-the-art clustering method and improve solving the spatial issue in this study. Third, the model and methods developed in the study are also applicable to other cities that can be performed reproducibly and universally applicable with readily available POIs data to help discover the utilization of urban space.

However, our research has the following shortcomings. First, because of the lack of land use status data or land use planning data, our study is unable to infer the utilization of urban space based on a more precise supervised classification framework. Second, little consideration was made about the mixed-use issue of urban functional regions using single dataset. In future research, we plan to fuse different kinds of emerging big geo-data and construct joint geospatial features of urban parcels to accurately identify urban functional structures and explore mixed-use of urban functions.

6. Conclusion

Discovering the utilization of urban space is one of the long-lasting questions in urban studies and planning. Using the analogy framework between urban elements (i.e., urban parcel, POIs) and NLP terms (i.e., document, word) has been a popular and effective method for discovering the utilization of urban parcels and has become a significant measure to solve the problems in urban studies. Based on this analogy framework, many mature models, such as topic models and word embedding models, have been introduced from the NLP domain to urban functional studies. However, these models still have some inherent defects when they are applied to urban function studies. For example, the problem of ignoring the order of the words in the document is not considered in the topic models, and the problem of word polysemy and ambiguity is not considered in the word embedding models. In response to the defects of these two types of models, we built a novel framework to solve these problems. In this framework, based on the open-source and massive POIs data, the TWE model, an improved word embedding model, was introduced to generate multiprototype word embedding representation. The TWE model can integrate the topic characteristics of the document into the word embedding model, resulting in a kind of multiprototype word embedding vector. The experimental results show that the word embedding vectors generated by the introduced model can express the meaning of POIs categories more accurately and distinctly than that of word2vec, a popular word embedding model. Then, by aggregating the embedding vectors of POIs categories into the high-dimension urban functional parcel vectors, we further introduced the HDBSCAN cluster model to cluster the high-dimension vectors of urban functional parcels in the study area. The clustering results reveal that the HDBSCAN clustering method can effectively sense the spatial structures of urban space and accurately identify land parcels with diverse functions.

Acknowledgment

We thank the anonymous reviewers for their valuable comments and suggestions. This study was supported by the National Natural Science Foundation of China [grant number 41871311], the National Key R & D Program of China [No. 2017YFB0503600] and the National Natural Science Foundation of China (grant number 41671400).

References

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management, 39*, 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3.

Antikainen, J. (2005). The concept of functional urban area. *Informationen zur Raumentwicklung, 7*, 447–452.

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., ... Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports, Human mobility: Models and applications, 734*, 1–74. https://doi.org/10.1016/j.physrep.2018.01.001.

Blaschke, T., Hay, G. J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R. Q., Van der Meer, F., Van der Werff, H., Van Coillie, F., et al. (2014). Geographic object-based image analysis–towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing, 87*, 180–191.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bracken, I. (2014). *Urban planning methods: Research and policy analysis.* Routledge.

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data, 10*, 5:1–5:51. https://doi.org/10.1145/2733381.

Chen, Y., Liu, X., Li, X., Liu, X., Yao, Y., Hu, G., ... Pei, F. (2017). Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method. *Landscape and Urban Planning, 160*, 48–60. https://doi.org/10.1016/j.landurbplan.2016.12.001.

Dear, M., & Flusty, S. (1998). Postmodern Urbanism. *Annals of the Association of American Geographers, 88*, 50–72. https://doi.org/10.1111/1467-8306.00084.

Gao, S., Janowicz, K., & Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS, 21*, 446–467. https://doi.org/10.1111/tgis.12289.

Goldberg, Y., & Levy, O. (2014). *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint.* arXiv:1402.3722.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*, 5228–5235. https://doi.org/10.1073/pnas.0307752101.

Heiden, U., Heldens, W., Roessner, S., Segl, K., Esch, T., & Mueller, A. (2012). Urban structure type characterization using hyperspectral remote sensing and height information. *Landscape and Urban Planning, 105*, 361–375.

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems, 54*, 240–254.

Hu, S., & Wang, L. (2013). Automated urban land-use classification with remote sensing. *International Journal of Remote Sensing, 34*, 790–803. https://doi.org/10.1080/01431161.2012.714510.

Hu, T., Yang, J., Li, X., Gong, P., Hu, T., Yang, J., ... Gong, P. (2016). Mapping urban land use by using landsat images and open social data. *Remote Sensing, 8*, 151. https://doi.org/10.3390/rs8020151.

Janowicz, K., Scheider, S., Pehle, T., & Hart, G. (2012). Geospatial semantics and linked spatiotemporal data –past, present, and future. *Semant Web, 3*, 321–332.

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., Jr., & Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems, 53*, 36–46.

Jokar Arsanjani, J., Helbich, M., Bakillah, M., Hagenauer, J., & Zipf, A. (2013). Toward mapping land-use patterns from volunteered geographic information. *International Journal of Geographical Information Science, 27*, 2264–2278.

Karlsson, C. (2007). Clusters, functional regions and cluster policies. *JIBS and CESIS Electronic Working Paper Series, 84*, 1010–1018.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal, 17*, 441–458. https://doi.org/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G.

Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., & Hong, Y. (2017). Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science, 31*, 1675–1696. https://doi.org/10.1080/13658816.2017.1324976.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., ... Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers, 105*, 512–530. https://doi.org/10.1080/00045608.2015.1018773.

Long, Y., & Shen, Z. (2015). Discovering functional zones using bus smart card data and points of interest in Beijing. *Geospatial Analysis to Support Urban Planning in Beijing* (pp. 193–217). Springer.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software, 2*, 205.

McKenzie, G., & Janowicz, K. (2017). The effect of regional variation and resolution on geosocial thematic signatures for points of interest. *The annual international conference on geographic information science* (pp. 237–256). Springer.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space. arXiv preprint.* arXiv:1301.3781.

Mikolov, T., Chen, K., Corrado, G.S., Dean, J.A., 2015. Computing numeric representations of words in a high-dimensional space. Google Patents.

Ng, H. T., & Zelle, J. (1997). Corpus-Based Approaches to Semantic Interpretation in NLP. *AI Magazine, 18*(4), 45. https://doi.org/10.1609/aimag.v18i4.1321.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science, 28*, 1988–2007. https://doi.org/10.1080/13658816.2014.913794.

Regan, C. M., Bryan, B. A., Connor, J. D., Meyer, W. S., Ostendorf, B., Zhu, Z., & Bao, C. (2015). Real options analysis for land use management: Methods, application, and implications for policy. *Journal of Environmental Management, 161*, 144–152.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation

of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642). .

Song, J., Lin, T., Li, X., & Prishchepov, A. V. (2018). Mapping urban functional zones by integrating very high spatial resolution remote sensing imagery and points of interest: A case study of Xiamen, China. *Remote Sensing, 10*, 1737. https://doi.org/10.3390/rs10111737.

State Information Center (2018). *Map POIs (point of interest) data.* https://doi.org/10.18170/DVN/WSXCNM.

Van de Voorde, T., Jacquet, W., & Canters, F. (2011). Mapping form and function in urban areas: An approach based on urban metrics and continuous impervious surface data. *Landscape and Urban Planning, 102*, 143–155.

Vatsavai, R. R., Bright, E., Varun, C., Budhendra, B., Cheriyadat, A., & Grasser, J. (2011). Machine learning approaches for high-resolution urban land cover classification: A comparative study. *Proceedings of the 2nd international conference on computing for geospatial research & applications* (pp. 11). ACM.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. *Proceedings of the 23rd international conference on machine learning* (pp. 977–984). ACM.

Wen, D., Huang, X., Zhang, L., & Benediktsson, J. A. (2016). A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation. *IEEE Transactions on Geoscience and Remote Sensing, 54*, 609–625.

Wu, L., Cheng, X., Kang, C., Zhu, D., Huang, Z., & Liu, Y. (2018). A framework for mixed-use decomposition based on temporal activity signatures extracted from big geo-data. *International Journal of Digital Earth, 0*, 1–19. https://doi.org/10.1080/17538947.2018.1556353.

Yan, B., Janowicz, K., Mai, G., & Gao, S. (2017). From ITDL to Place2Vec: Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts. *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '17. ACM, New York, NY, USA* (pp. 35:1–35:10). . https://doi.org/10.1145/3139958.3140054.

Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015). Topical word embeddings. *Twenty-ninth AAAI conference on artificial intelligence. Presented at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., & Mai, K. (2017). Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science, 31*, 825–848. https://doi.org/10.1080/13658816.2016.1244608.

Yao, Y., Liu, X., Li, X., Zhang, J., Liang, Z., Mai, K., & Zhang, Y. (2017). Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *International Journal of Geographical Information Science, 31*, 1220–1244. https://doi.org/10.1080/13658816.2017.1290252.

Yuan, J., Zheng, Y., & Xie, X. (2012). Discovering regions of different functions in a City using human mobility and POIs. *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '12* (pp. 186–194). New York, NY, USA: ACM. https://doi.org/10.1145/2339530.2339561.

Zhai, W., Bai, X., Shi, Y., Han, Y., Peng, Z.-R., & Gu, C. (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. https://doi.org/10.1016/j.compenvurbsys.2018.11.008.

Zhang, X., & Du, S. (2015). A linear Dirichlet mixture model for decomposing scenes: Application to analyzing urban functional zonings. *Remote Sensing of Environment, 169*, 37–49. https://doi.org/10.1016/j.rse.2015.07.017.

Zhang, X., Du, S., & Wang, Y.-C. (2015). Semantic classification of heterogeneous urban scenes using intrascene feature similarity and interscene semantic dependency. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8*, 2005–2014.

Zhang, X., Du, S., & Wang, Q. (2017). Hierarchical semantic cognition for urban functional zones with VHR satellite images and POIs data. *ISPRS Journal of Photogrammetry and Remote Sensing, 132*, 170–184. https://doi.org/10.1016/j.isprsjprs.2017.09.007.

Zhang, X., Du, S., Wang, Q., & Zhou, W. (2018). Multiscale geoscene segmentation for extracting urban functional zones from VHR satellite images. *Remote Sensing, 10*, 281. https://doi.org/10.3390/rs10020281.

Zhang, Y., Li, Q., Huang, H., Wu, W., Du, X., & Wang, H. (2017). The combined use of remote sensing and social sensing data in fine-grained urban land use mapping: A case study in Beijing. *China. Remote Sensing, 9*, 865. https://doi.org/10.3390/rs9090865.

Zheng, Y., Capra, L., Wolfson, O., & Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology, 5*(38), 1–38:55. https://doi.org/10.1145/2629592.

Zhi, Y., Li, H., Wang, D., Deng, M., Wang, S., Gao, J., ... Liu, Y. (2016). Latent spatio-temporal activity structures: A new approach to inferring intra-urban functional regions via social media check-in data. *Geo-spatial Information Science, 19*, 94–105.

Zhong, C., Huang, X., Arisona, S. M., Schmitt, G., & Batty, M. (2014). Inferring building functions from a probabilistic model using public transportation data. *Computers, Environment and Urban Systems, 48*, 124–137.