

doi: 103969/j. issn. 0490-6756. 2019. 02. 011

基于时空语义挖掘的城市功能区识别研究

于璐¹, 何祥¹, 刘嘉勇²

(1. 四川大学电子信息学院, 成都 610065; 2. 四川大学网络空间安全学院, 成都 610065)

摘要: 针对目前城市功能区划分大多依靠人工完成, 且未充分使用城市中时空数据的问题, 提出一种基于时空语义挖掘的城市功能区识别方案. 首先, 选取某城市矩形区域为研究样本, 并以建筑物为划分依据将研究样本划分为有效的基础区域; 然后, 对各基础区域内的新浪微博位置签到数据及 POI(Points of Interest) 数据进行时空语义挖掘, 采用狄利克雷多项式回归(DMR)主题模型生成区域的功能性向量; 最后, 通过向量聚类, 依据 POI 类别比例完成区域的功能性识别. 实验结果表明, 本方案相比基于 POI 密度的 *k-means* 聚类方案和基于潜在狄利克雷分布(LDA)主题模型的城市功能区识别方法具有更高的准确性, 位置签到数据所表征出的人们活动模式可以揭示城市功能区之间的差异, 在城市地理空间分析上具有良好的效果.

关键词: 时空数据挖掘; 城市功能分区; 主题模型; 签到数据

中图分类号: TP181 **文献标识码:** A **文章编号:** 0490-6756(2019)02-0246-07

Discovering urban functional regions based on sematic mining from spatiotemporal data

YU Lu¹, HE Xiang¹, LIU Jia-Yong²

(1. College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China;
2. College of Cybersecurity, Sichuan University, Chengdu 610065, China)

Abstract: To tackle the problem that the current urban functional regions division are manual completed and do not fully use the spatiotemporal data in urban regions, an approach for detecting urban functional regions is proposed based on sematic mining from spatiotemporal data. In which, a rectangular area of the city is first selected as a research sample and divided it into some valid basis region units according to its buildings. Dirichlet multinomial regression (DMR) topic model is then implemented for the check-in and POI(points of interest) data from Sina weibo in these basis region units and the functional vectors of the basis region units are obtained. Finally, the functional regions are discovered with vector clustering algorithm and POI's category proportion. The experimental results show that this approach has higher accuracy compared with the *k-means* clustering method based on POI density and urban functional area detecting approach based on latent Dirichlet allocation (LDA) topic model. Therefore, The activity patterns of people identified by location check-in data can reveal the differences between urban functional areas and have a good effect on urban geospatial analysis.

Keywords: Spatiotemporal data mining; Urban district; Topic Model; Check-in data

收稿日期: 2018-06-06

基金项目: 科技部国家重点研发项目(2017YFB0802904)

作者简介: 于璐(1994-), 女, 山西晋城人, 硕士生, 研究方向为数据挖掘. E-mail: im.izzie.yu@gmail.com

通讯作者: 刘嘉勇. E-mail: ljy@scu.edu.cn

1 引言

传统的城市功能分区研究大多以卫星遥感、问卷调查、实地走访等得来的数据为基础,再辅以指标体系进行城市功能区识别^[1-3]。但这些方法耗费的人力成本过高,并且分析过程中带有调查人员的主观因素,难以准确地对城市功能区进行长期动态监测。

随着互联网的高速发展和感知技术的逐渐成熟,城市中的时空数据迅速积累,主要包括交通工具的移动数据^[4,5]、移动网络数据^[6]和基于位置的服务(Location Based Service, LBS)数据^[7]等。这些数据获取成本低、内容丰富而且现势性好,可以反应出人们的活动模式,成为解决城市功能区识别问题的另一思路。

目前,基于城市时空数据的功能区识别主要有三个研究方向。一些学者分析区域内人们活动的时间特征和 POI 分布特征等,再通过聚类分析得到区域的功能^[8]。这种方法通常挖掘单一类型时空数据特征,缺乏数据之间的关联与融合,导致得到的结果也缺乏可靠性。另一部分学者把城市功能区划分问题转换为分类问题,借助于外部人员区域标注,使用不同分类器得出区域的功能属性^[9],这种方法更好地挖掘多类型数据间的潜在结构,并得到了更高的准确率,但是需要大量人工标注,实际使用时存在困难。最新的研究方向使用主题模型进行时空语义挖掘^[10],发现区域的隐含功能,不仅可以融合多源地理数据,而且用非监督的方式完成功能区识别。

主题模型在自然语言处理领域应用广泛,它在文本隐含语义挖掘中表现良好,将文档从词项空间映射到主题空间,被表示为多个主题的概率分布。而这个思想也可以映射到区域功能性发现上,将一个区域视为一篇文档,区域中活跃的时空数据是文档中的单词,区域的功能则是文档的主题,那么使用主题模型便可以得到每个区域的功能分布。LDA(Latent Dirichlet Allocation)主题模型^[11]是文本语义挖掘的经典模型,它是一个包含隐藏量的贝叶斯生成模型,尝试用主题的概率分布表示文档,用单词的概率分布刻画主题,但对于主题表达,它只考虑了文本中的单词,缺少与文档关联的其他数据对主题的贡献,故出现了许多基于 LDA 的扩展模型。DMR(Dirichlet Multinomial Regression)模型^[12]是 LDA 派生出的主题模型,与其他主题模型相

比,此模型中文档的狄利克雷主题分布先验参数考虑了文档相关特征的影响,引入了更复杂有效的辅助特征,增强了模型的主题提取效果。

基于以上研究,本文以位置签到数据和 POI 数据为基础,提出一种基于时空语义挖掘的城市功能区识别方法。以建筑物作为研究区域中的基础功能单元,由位置签到数据提取功能单元中的用户行为模式、文本特征,再结合 POI 密度一同输入 DMR 主题模型,得到功能单元的功能向量,然后对向量进一步聚类,聚合出具有相似功能的区域,并且对其标注进行语义层面解释,实现区域的功能识别。

2 城市功能区识别思路

基于时空语义挖掘的城市功能区识别方法,尝试通过主题模型发现区域隐含的功能。与文本主题挖掘类比,基础功能单元相当于语料中的文档,功能单元中的时空数据类似文档中的单词,通过主题模型后,得到的单元功能分布相当于文档主题分布,如图 1 所示。

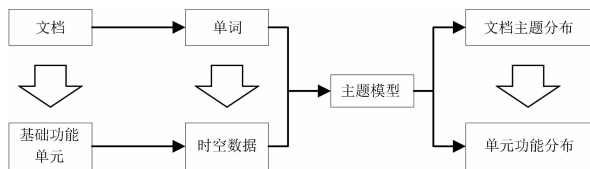


图 1 区域-功能和文档-主题的类比

Fig. 1 Analogy between region-function and document-topic

本文使用的城市时空数据为具有代表性的新浪微博位置签到数据,每条签到数据中包含用户信息、签到位置的空间坐标、发布时间和发布文本等,能够从不同角度反应人们的动态活动模式。同时从百度地图获取研究区域中的 POI,每条记录包含物理实体的名称、空间坐标、地址和类型等,由其类型计算出不同基础功能单元的 POI 类别密度特征。

城市功能区识别的主要流程如图 2 所示,首先将研究区域以建筑物为划分依据分为空间独立的基础功能单元,并根据空间坐标将离散的微博签到数据转换为签到事件集合分配到各个单元;然后,以基础功能单元为对象,提取其签到事件集合的行为模式和文本特征,并计算 POI 类别密度,代入 DMR 主题模型,得到基础功能单元的功能向量;由于获得的功能向量并没有明确的功能语义,故对功能向量聚类分析,得到具有相似功能的单元簇,

最后,根据功能单元中 POI 结构标注出各区域簇的功能属性,给予语义性解释,完成功能区识别.

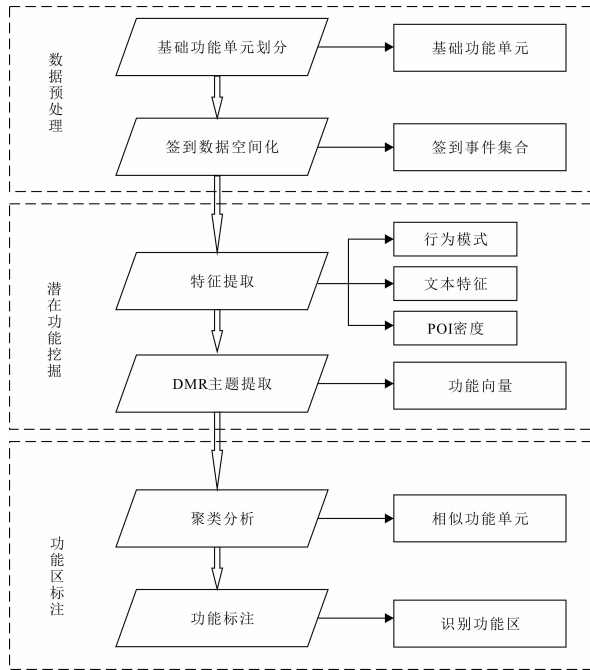


图 2 城市功能区识别基本流程

Fig. 2 Flow diagram of discovering urban functional regions

3 基于主题模型的城市功能区识别

为了获得城市功能区,需要先进行基础功能单元划分,识别出研究区域中建筑物的轮廓^[13],再使用连通区域标记算法^[14]划分出基础功能单元.然后将研究区域中的签到数据空间化,映射到各个基础单元.接下来详细介绍基础单元特征提取、潜在功能挖掘以及功能区标注的过程.

3.1 特征提取

3.1.1 行为模式 位置签到数据中,用户每次的签到行为可以表示为

$$C = \{user, latitude, longitude, time, text\}$$

其中, *user* 为用户标识; *latitude* 为签到位置的纬度; *longitude* 为签到位置的经度; *time* 为签到的时间; *text* 为签到时发布的文本,它们的集合构成了用户的一次移动行为,表示用户在某个时间出现在了某个地点.基础功能单元中的用户行为模式定义如下.

定义 1 用户行为模式矩阵 *P*. 行为模式矩阵表示某时间段内用户出现在某基础功能单元的平均次数.将一天划分为 12 个时间段,每段 2 个小时,并且区分工作日和周末,得到 24 个时间区间.

统计每个区域每个时间段用户签到行为 *C* 的平均个数,形成行为模式矩阵 *P*,如图 3 所示.

| | | | | | | | | |
|-------|---|-------|----|----------|--|----|----|-------|
| | 3 | | | | | | | R_n |
| | | | 59 | | | | 19 | ⋮ |
| | | 4 | | 15 | | | | ⋮ |
| | | | 6 | | | | | R_j |
| | 8 | | | | | 50 | | ⋮ |
| | | | 9 | | | | 10 | ⋮ |
| 23 | | | | 60 | | | | ⋮ |
| | | 5 | | | | | | R_1 |
| t_1 | ⋯ | t_i | ⋯ | t_{24} | | | | |

图 3 用户行为模式矩阵

Fig. 3 User behavior pattern matrix

在图 3 的用户移动模式矩阵中,横轴为时间区间 t_1, \dots, t_{24} ,纵轴为区域 R_1, \dots, R_n , *n* 为基础单元的个数,矩阵中的数字表示某种行为模式平均出现的次数.如阴影数字 6 表示,在区域 R_j 中时间区间 t_i 内用户平均每天到此区域 6 次.如此,得到每个区域 24 维的行为模式向量.

3.1.2 文本特征扩充 位置签到的文本数据以短文本居多,特征提取存在困难,本文采用基于 Word2vec 词向量模型的特征扩充方法扩展区域内的文本特征,以缓解特征稀疏问题. Word2vec 将词语投影到一个向量空间,属于 distributed representation 词向量^[15,16].它基于词语语义的分布假设理论,提出一种基于神经网络的词向量训练模型,通过目标词与上下文的关系,得到目标词的低维词向量,不仅训练效率高,而且在大规模语料上训练得到的词向量在句法和语义上也具有很强的相关性.由于 Word2vec 词向量模型可以发现词之间的语义关系,故使用它寻找关键词的相近词,用于扩充短文本的特征,同时在一定程度上增强主题,更好地体现功能性.

文本特征扩充的具体步骤如下. (1)数据预处理:将采集到的大量微博语料分词并去除停用词和干扰词; (2)训练词向量模型:配置 Word2vec 模型参数,代入数据训练; (3)抽取关键词:统计分析得出已有语料文本的平均长度为 17 个词语,故计算待扩充文本中词语的 TF-IDF 值,并选取前 10 个作为关键词; (4)文本扩充:遍历这些关键词,根据之前得到的 Word2vec 模型扩充 5 个最相近词作为扩充文本特征.

3.1.3 POI 密度 对于每一个基础功能单元, 构建一个 POI 密度特征向量. 对于每个区域 r , 第 i 种类型 POI 的数量为 n_{ri} , 区域 r 中所有 POI 的数量为 s_r , 那么区域内第 i 种 POI 的密度 v_{ri} 为

$$v_{ri} = \frac{n_{ri}}{s_r} \quad (1)$$

区域 r 的 POI 密度特征向量为 $x_r = (v_{r1}, v_{r2}, \dots, v_{rF}, 1)$. 其中 F 为 POI 种类的数量, 最后一位的 1 是一个默认值, 为了之后描述每个主题的均值^[12]. POI 的类型如表 1 所示.

表 1 POI 类型

Tab. 1 POI's category

| 种类 ID | POI 种类 | 种类 ID | POI 种类 |
|-------|--------|-------|--------|
| 1 | 休闲娱乐 | 7 | 金融 |
| 2 | 地产小区 | 8 | 生活服务 |
| 3 | 政府机构 | 9 | 宾馆 |
| 4 | 餐饮 | 10 | 公司企业 |
| 5 | 教育培训 | 11 | 购物 |
| 6 | 交通设施 | 12 | 医疗 |

3.2 潜在功能挖掘

将 3.1 得到的行为模式和文本特征作为基础功能单元的“文本”, 集合各个基础单元的“文本”形成“文档集”输入 DMR 主题模型, 同时每个基础功能单元都有 POI 密度特征作为先验数据. 根据可观察的“文本”中的特征, 推断隐藏变量, 即基础单元的功能向量, 最终每个基础单元会被表示为各个功能下的概率分布.

DMR 模型的生成过程如图 4 所示, 具体为:

- (1) 给定研究区域中第 r 个基础功能单元, 对于每个隐含功能 k , 由超参数 σ 生成服从高斯分布的引入向量 λ_k , 它与 POI 密度特征的长度相同, 并由先验参数 η 生成基础单元的功能-特征狄利克雷分布 β_k ;
- (2) 令 $\alpha_{r,k} = \exp(x_r^T \lambda_k)$, 其中 x_r 是基础单元 r 的 POI 密度特征, θ_r 是服从先验参数为 α_r 的狄利克雷功能分布;
- (3) 对于基础单元 r 中的第 n 个特征 $f_{r,n}$, 它的功能分布 $z_{r,n}$ 是服从步骤(2)得到的 θ_r 的多项分布, 可以由此确定这个单词的功能分布 $\beta_{z_{r,n}}$;
- (4) 根据特征 $f_{r,n}$ 服从 $\beta_{z_{r,n}}$ 的多项分布, 生成特征 $f_{r,n}$;
- (5) 遍历基础单元 r 中的 N 个特征, 重复执行步骤(3)~(4), 生成基础单元 r ;
- (6) 遍历研究区域中的 R 个基础单元, 重复执行步骤(1)~(5), 生成整个研究区域.

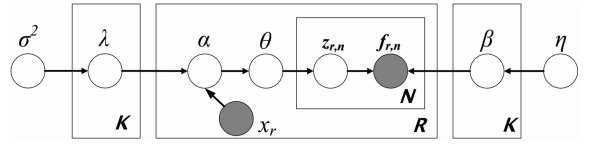


图 4 DMR 模型生成过程

Fig. 4 Generation process of DMR model

每个基础单元功能分布的先验参数 α 考虑了 POI 密度特征对功能的影响, 不同的 POI 密度具有不同的 α 取值, 可见, 基础单元的“语义主题”由人们动态的活动模式和静态的地理实体共同决定. DMR 主题模型的训练中, 采用吉布斯采样估计模型参数, 直到收敛, 输出 θ 和 β 的值, 其中 θ 的大小为 $R \times K$, 也就是 R 个区域的功能向量, 向量长度为 K .

3.3 功能区标注

功能向量没有明确的语义表达, 不能对功能区定性判断. 故先将功能向量使用 k-means 算法^[17,18]聚类分析, 使得具有相似功能的区域聚合在一起, 并用平均轮廓系数确定聚类中心的个数. 然后, 对聚合后的区域簇由其内部的 POI 分布结构进行功能标注, 科教区、居民区、商业区、工作区、生活设施区和混合区.

POI 分布结构包括区域簇内不同类别 POI 的频率密度 (Frequency Density, FD) 和该区域簇中不同类别 POI 所占比例 (Category Proportion, CP), 对于区域簇 c , 计算方式如下.

$$FD_{c,i} = \frac{n_{c,i}}{N_i}, i=1, 2, \dots, m \quad (2)$$

$$CP_{c,i} = \frac{FD_{c,i}}{\sum_{i=1}^m FD_{c,i}}, i=1, 2, \dots, m \quad (3)$$

其中, i 为 POI 类别; m 为 POI 类别总数; $n_{c,i}$ 为 c 中类别 i 的 POI 数量; N_i 为类别 i 的 POI 总数; $FD_{c,i}$ 表示区域簇 c 中第 i 种 POI 的频率密度; $CP_{c,i}$ 表示区域簇 c 中第 i 种 POI 的频率密度所占该簇所有类别 POI 频率密度的比例.

类别比例代表着一种类型 POI 在某一区域中的重要程度, 而不同类型的 POI 与功能的相关度是不同的, 研究确定, 在当前区域簇中, 若属于某种功能的 POI 类别比例之和超过 50%, 则该区域簇以该功能为主, 为单一城市功能区; 若属于每种功能的 POI 类别比例之和均未超过 50%, 则该区域簇为混合区.

4 实验结果分析

4.1 实验数据及评价指标

本文选取成都市内一环路附近的矩形区域为研究对象,采集该范围内的签到微博,又以现有用户为基础,增量获取微博文本用于基础功能单元的文本特征扩充.数据集的统计信息如表 2 所示,微博总数共计 32 万余条,其中位置签到数据为 16 万余条,涵盖的用户有 10527 个,签到的地点数为 418 个,研究范围内的 POI 总数为 3382 条.

为了衡量及评测区域功能性发现的结果,对研究范围内基础功能单元的功能手工标注,各个功能区的标注情况如表 3 所示.

表 2 实验数据集

Tab. 2 Experimental dataset

| 标识 | 大小 |
|-------|---------|
| 微博总数 | 328346 |
| 签到数 | 162099 |
| 用户数 | 10527 |
| 地点数 | 418 |
| 签到/用户 | 15.398 |
| 签到/位置 | 387.796 |
| POI 数 | 3382 |

评价指标采用区域功能识别准确率^[19]衡量模型识别的可靠性,即正确标注的区域占识别总区域的比例.并辅以区域识别率衡量模型功能识别的可用性,即识别出的区域占总区域的比例.

4.2 实验过程

根据城市功能区识别的步骤对研究对象内基础单元的功能进行识别.首先通过基础功能单元划分得到 417 个基础单元,为了避免面积过小区域的干扰,只选择总面积为研究范围内所有单元面积之和 85% 的基础单元,得到基础功能单元 349 个.由基础单元内签到数据及 POI 数据获得其行为模式、文本特征和 POI 密度,通过 DMR 主题模型挖掘隐含功能向量,在训练主题模型时,通过模型困惑度选择最优的主题数,此处为 10 个.在功能区标注中,使用 *k-means* 聚类算法,并由轮廓系数确定为 6 个聚类中心.聚合后区域簇的 POI 结构如表 4 和表 5 所示,表 5 中的加粗黑体部分为该区域簇中所属功能类别比例之和超过 50% 的 POI 类型,则区域簇为该功能为主的单一功能区.根据类别比例判定标准,依次将区域簇 1~6 标注为居民区、科教区、工作区、商业区、生活设施区和混合区.

表 3 标注样本

Tab. 3 Labeled data

| 功能性标注类别 | 数量 |
|---------|-----|
| 科教区 | 60 |
| 居民区 | 184 |
| 商业区 | 68 |
| 工作区 | 19 |
| 生活设施区 | 10 |
| 混合区 | 8 |

表 4 区域簇内不同类别 POI 的频率密度

Tab. 4 Frequency density of different POI's categories in regional clusters

| POI 种类 | 区域簇 1 | 区域簇 2 | 区域簇 3 | 区域簇 4 | 区域簇 5 | 区域簇 6 |
|--------|----------|----------|----------|----------|----------|----------|
| 休闲娱乐 | 0.041732 | 0.007874 | 0.023622 | 0.690550 | 0.000000 | 0.236220 |
| 地产小区 | 0.801711 | 0.004310 | 0.047413 | 0.099135 | 0.004310 | 0.043103 |
| 政府机构 | 0.042858 | 0.071429 | 0.514286 | 0.085714 | 0.000000 | 0.285715 |
| 餐饮 | 0.094719 | 0.003300 | 0.039603 | 0.567656 | 0.103300 | 0.198021 |
| 教育 | 0.017648 | 0.611768 | 0.125490 | 0.127452 | 0.000000 | 0.117647 |
| 交通设施 | 0.150000 | 0.000000 | 0.200000 | 0.450000 | 0.100000 | 0.100000 |
| 金融 | 0.070832 | 0.000000 | 0.129167 | 0.512499 | 0.062500 | 0.225000 |
| 生活服务 | 0.142429 | 0.075759 | 0.121213 | 0.363644 | 0.100020 | 0.196970 |
| 宾馆 | 0.118133 | 0.000000 | 0.016949 | 0.557624 | 0.067797 | 0.239495 |
| 公司企业 | 0.028979 | 0.009553 | 0.477388 | 0.171974 | 0.001592 | 0.310510 |
| 购物 | 0.082210 | 0.018405 | 0.077915 | 0.580370 | 0.006135 | 0.234970 |
| 医疗 | 0.000000 | 0.000000 | 0.133334 | 0.266668 | 0.133333 | 0.466667 |

表 5 区域簇内不同 POI 的类别比例

Tab. 5 Category proportion of different POI in regional clusters

| POI 种类 | 区域簇 1 | 区域簇 2 | 区域簇 3 | 区域簇 4 | 区域簇 5 | 区域簇 6 |
|--------|-----------------|-----------------|-----------------|-----------------|-----------------|----------|
| 休闲娱乐 | 0.026226 | 0.008726 | 0.012391 | 0.154372 | 0.000000 | 0.088995 |
| 地产小区 | 0.503824 | 0.004776 | 0.024871 | 0.022162 | 0.007444 | 0.016239 |
| 政府机构 | 0.026934 | 0.079155 | 0.269771 | 0.019161 | 0.000000 | 0.107642 |
| 餐饮 | 0.059525 | 0.003657 | 0.020774 | 0.126899 | 0.178421 | 0.074603 |
| 教育 | 0.011091 | 0.677936 | 0.065826 | 0.028492 | 0.000000 | 0.044323 |
| 交通设施 | 0.094265 | 0.000000 | 0.104911 | 0.100597 | 0.172721 | 0.037675 |
| 金融 | 0.044513 | 0.000000 | 0.067755 | 0.114569 | 0.107951 | 0.084768 |
| 生活服务 | 0.089508 | 0.083953 | 0.063583 | 0.081292 | 0.172756 | 0.074207 |
| 宾馆 | 0.074239 | 0.000000 | 0.008891 | 0.124656 | 0.117100 | 0.090227 |
| 公司企业 | 0.018211 | 0.010586 | 0.250416 | 0.038445 | 0.002750 | 0.116983 |
| 购物 | 0.051664 | 0.020396 | 0.040871 | 0.129741 | 0.010596 | 0.088524 |
| 医疗 | 0.000000 | 0.000000 | 0.069941 | 0.059613 | 0.230295 | 0.175815 |

为了验证提出模型的有效性,将本文提出的基于 DMR 主题模型的城市功能区识别方案与现有的区域功能性分析方法进行比较,分别为基于 POI 密度的 *k-means* 聚类方案^[8](本文简称为 POI+k-means)和基于 LDA 主题模型的功能区识别方案^[6].

4.3 实验结果分析

实验结果如表 6 所示,POI+k-means 方案功能识别准确率最低,原因是 POI 数据不够准确,如若区域内 POI 数据没有及时更新,则会对区域功能识别结果产生影响.而基于 LDA 的功能区识别方案,由于使用了行为模式数据和区域内的文本数据,所以效果相比 POI+k-means 方案要好,体现出人们动态的活动模式对区域功能性的影响.本文提出的城市功能区识别方案表现最好,有 90.31% 的准确率.DMR 主题模型相比 LDA,引入 POI 密度特征对区域功能的影响,同时使用 Word2vec 扩充文本特征后,文本具有更好的语义表达,提高了识别的准确率.

表 6 本课题模型与对比模型的指标数据

Tab. 6 Evaluation of different algorithms

| | POI+k-means | LDA | 本文方案 |
|-------|-------------|--------|--------|
| 区域识别率 | 0.7106 | 0.8195 | 0.8195 |
| 识别准确率 | 0.7096 | 0.8564 | 0.9031 |

5 结 论

本文通过大量位置签到数据和 POI 数据实现了区域功能识别,挖掘海量签到数据中蕴含的区域

时空特性和文本特征,用 DMR 主题模型发现区域功能性,并用聚类算法使得区域聚合以更好地对其语义标注,最终可以对区域的功能性做合理性的解释.实验结果表明,本方法在城市功能区识别上有较高的准确率,在城市规划、商业投放和资源配置等领域具有应用前景.今后的研究工作将进一步融合多源数据进行分析,提高区域功能识别的效果,并且在有少许标注的情况下,使用半监督学习的方式进一步深入探索.

参考文献:

- [1] 郭理桥,林剑远,王文英.基于高分遥感数据的城市精细化管理应用[J].城市发展研究,2012,19:57.
- [2] Tian G, Wu J, Yang Z. Spatial pattern of urban functions in the Beijing metropolitan region [J]. Habitat International, 2010, 34: 249.
- [3] 史玉峰,王艳.基于自组织神经网络的城市功能分区研究[J].计算机工程,2006,32:206.
- [4] Pu Y, Song X, Ge Y. An improved framework to discover functional urban regions using pedestrian trajectories [C]// Proceedings of the International Conference on Geoinformatics. [s. l.]:IEEE, 2016.
- [5] 王永坤,王海洋,潘平峻,等.面向公共安全的时空数据挖掘综述[J].重庆邮电大学学报:自然科学版,2018,30:40.
- [6] 向峰.基于移动网络数据的用户行为与城市感知研究[D].武汉:华中科技大学,2014.
- [7] Brillhante I, Macedo J A, Nardini F M, et al. Where shall we go today?: planning touristic tours with tripbuilder [C]// Proceedings of the Acm In-

- ternational Conference on Information & Knowledge Management. [s. l.]: ACM, 2013.
- [8] Cranshaw J, Schwartz R, Hong J I, *et al.* The livehoods project: utilizing social media to understand the dynamics of a city [J]. Soc Sci Electron Publish, 2012; 1.
- [9] Pan G, Qi G, Zhang D, *et al.* Land-use classification using taxi GPS traces [J]. IEEE Trans Intell Transp Syst, 2013, 14: 113.
- [10] 陈世莉, 陶海燕, 李旭亮, 等. 基于潜在语义信息的城市功能区识别——广州市浮动车 GPS 时空数据挖掘 [J]. 地理学报, 2016, 71: 471.
- [11] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. J Mach Learn Res, 2003, 3: 993.
- [12] Mimno D, Mccallum A. Topic models conditioned on arbitrary features with dirichlet multinomial regression [J]. U Mass Amherst, 2012, 2008: 411.
- [13] Xie Z, Ye Z, Wu L. Research on building polygon map generalization algorithm [C]// Proceedings of the Eighth Acis International Conference on Software Engineering. [s. l.]: IEEE, 2007.
- [14] Shapiro L G, Stockman G. Computer vision [M]. 2ed edit. Upper Saddle River, NJ, USA: Prentice Hall, 2001.
- [15] Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space [J]. Comput Sci, 2013, 2013: 1.
- [16] 阳馨, 蒋伟, 刘晓玲. 基于多种特征池化的中文文本分类算法 [J]. 四川大学学报: 自然科学版, 2017, 54: 287.
- [17] 杨杰明, 吴启龙, 曲朝阳, 等. MapReduce 框架下基于抽样的分布式 k-means 聚类算法 [J]. 吉林大学学报: 理学版, 2017, 55: 109.
- [18] 徐健锐, 詹永照. 基于 Spark 的改进 k-means 快速聚类算法 [J]. 江苏大学学报: 自然科学版, 2018, 39: 316.
- [19] Zhan X, Ukkusuri S V, Zhu F. Inferring urban land use using large-scale social media check-in data [J]. Networks & Spatial Economics, 2014, 14: 647.

引用本文格式:

- 中文: 于璐, 何祥, 刘嘉勇. 基于时空语义挖掘的城市功能区识别研究 [J]. 四川大学学报: 自然科学版, 2019, 56: 246.
- 英文: Yu L, He X, Liu J Y. Discovering urban functional regions based on sematic mining from spatiotemporal data [J]. J Sichuan Univ; Nat Sci Ed, 2019, 56: 246.