杨桂松, 郭东升, 何杏宇,等. 基于 LBSN 数据聚类分析的城市 POI 感知方法[J]. 智能计算机与应用,2024,14(7):43-49.
DOI:10.20169/j.issn.2095-2163.240706

基于 LBSN 数据聚类分析的城市 POI 感知方法

杨桂松[1], 郭东升[1], 何杏宇[1,2], 卢海军[3]

(1 上海理工大学 光电信息与计算机工程学院, 上海 200093; 2 上海理工大学 出版印刷与艺术设计学院, 上海 200093;
3 上海诺基亚贝尔股份有限公司, 上海 201201)

摘 要: 城市 POI 的分布情况客观反映了一个城市各行各业的发展情况,传统获取 POI 的测绘手段成本高、更新周期长、时效性差,而基于位置的社交网络(Location-Based Social Network, LBSN)平台的发展为实现城市 POI 的感知提供了一种新思路。本文提出一种基于 LBSN 数据聚类分析的城市 POI 感知方法,首先,对 LBSN 数据进行预处理,包括清洗重复数据、删除无效数据、数据预分类等,以提高数据的有效性;其次,提出一种改进的 DBSCAN 算法,对处理后的数据进行聚类分析,从而得到准确度较高的城市各类 POI 分布情况。实验结果表明,与传统的 DBSCAN 算法以及 K-means 算法相比,本文提出的算法有更好的聚类效果,且在聚类指标上有更大的 CH 指数值和更小的 DBI 指数值。

关键词: 城市 POI 感知; 基于位置的社交网络; 数据预处理; 改进的 DBSCAN 算法
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)07-0043-07

Urban POI perception method based on LBSN data cluster analysis

YANG Guisong[1], GUO Dongsheng[1], HE Xingyu[1,2], LU Haijun[3]

(1 School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2 College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai 200093, China; 3 NOKIA ShanghaiI Bell CO., LTD, Shanghai 201201, China)

Abstract: The distribution of urban POI objectively reflects the development of all walks of life in a city. Traditional surveying and mapping methods to obtain POI have high cost, long update cycle and poor timeliness. The development of location-based social network (Location-based Social Network, LBSN) platforms provide a new way to realize the perception of urban POI. For that, this paper proposed an urban POI perception method based on LBSN data cluster analysis. Firstly, this method preprocessed the LBSN data, including cleaning duplicate and invalid data, data pre-classification, etc., to improve the validity of the data. Then, this method proposed an improved DBSCAN algorithm to cluster the processed data, so as to obtain the distribution of various urban POIs with high accuracy. Finally, compared with the traditional DBSCAN algorithm and K-means algorithm, the experimental results show that the proposed method performs better in clustering result, and has a larger CH index value and a smaller DBI index value in clustering indicators.

Key words: urban POI perception; Location-based Social Network; data preprocessing; improved DBSCAN algorithm

0 引 言

在地理信息系统中,兴趣点(Point of Interest, POI)可以指一栋房子,一个公交站等[1]。具体而言,POI 指能够反映一个区域内具有某类特征的地点,这些地点往往是有趣或有用的。POI 作为一个基本概念被广泛应用于地图学、导航等领域[2]。一条 POI 数据主要包括名称、地址、类别等信息。

城市的 POI 数据可以细致而真实地反映城市的发展与变迁。例如一个城市旅游景点类的 POI

基金项目: 国家自然科学基金(61802257,61602305); 上海市自然科学基金(18ZR1426000, 19ZR1477600); 浦东新区科技发展基金产学研专项资助项目(PKX2021-D10)。

作者简介: 杨桂松(1982-),男,博士,副教授,硕士生导师,主要研究方向:物联网与普适计算; 郭东升(1997-),男,硕士研究生,主要研究方向:移动群智计算,数据挖掘分析; 卢海军(1979-),男,硕士,主要研究方向:负责诺基亚 OLT 全球产品研发和交付,以及中国区固网事业部管理。

通讯作者: 何杏宇(1984-),女,博士,副教授,硕士生导师,主要研究方向:物联网和移动群智计算。Email:xy_he@ usst.edu.cn

收稿日期: 2023-04-12

的分布以及发展情况能够客观、准确地反映该城市旅游行业的发展状况。POI 广泛地应用于商业选址,城市计算,旅游推荐等领域。文献[3]通过研究 POI 数据,从行为空间互动理论视角探究建设用地多功能混合规律,从而丰富现有土地混合利用理论与方法体系;文献[4]基于不同类型的生活服务业 POI 数据,对城市生活服务业设施的空间分布特征和空间配套进行研究,实现对城市服务业布局规划的指导;文献[5]通过收集用户的多维偏好,设计了一个基于 POI 的位置推荐系统,推荐用户更喜欢的旅游景点。因此对城市里各类 POI 以及其发展规律的感知,能够准确地了解城市内各行各业的发展状况及发展规律,不仅有利于从业者掌握其所在行业的发展情况还有利于政策决策者制定合理的城市发展规划。然而,传统的 POI 感知方法主要由专业人员通过测绘手段获取,获取到的 POI 数据精度高、属性完整,但更新成本更高、更新周期长、时效性差,急需找到一种成本低且快捷的 POI 感知方式。

近年来,很多研究者对 POI 的感知进行了研究。文献[6]利用多媒体用户上传的照片数据,提出一种无监督的方法来提取每个地标的代表性视图和图像,从而实现对城市 POI 的感知;文献[7]提出了一种新型的基于密度的 DBSCAN 的聚类算法 P-DBSCAN,对大量带有地理位置的照片进行聚类分析,以发现人们关注的兴趣点;文献[8]和文献[9]通过对包含位置的图片进行异构特征提取,提出一种改进的自适应光谱聚类方法,实现对 POI 的感知;文献[10]提出一种概率访问 POI 识别方法,使用一种新型的层次贝叶斯模型对用户的移动轨迹进行分析,从而得出用户偏爱的 POI;文献[11]通过对微博签到数据进行空间以及属性的匹配,实现对城市 POI 的感知更新。为了挖掘具有潜在宝贵价值的不受欢迎的景点,文献[12]提出一种层次多线索融合(HMCF)的方法,使用来自多源用户生成的内容来全面描述 POI,最后通过对 POI 进行分层建模,找到潜在的 POI。

上述工作中,部分研究是从图片数据中获取位置信息并对其进行分析,实现对 POI 的感知,这些方式需要大量包含位置信息的图片数据,而这些数据往往较难获得且对图片数据处理的过程较为复杂,难以从中获取到有效的信息。部分研究虽然使用了一些较易获取的数据,但是使用的模型算法较为复杂,需要消耗大量的计算资源与时间成本。随着互联网技术的进步和移动设备的广泛使用,基于位置的社交网络(Location-based Social Network,LBSN)也大量出现。目前国外著名的 LBSN 包括 Yelp、Foursquare、Gowalla 和 Facebook。国内流行的 LBSN 平台包括美团、大众点评、微博等。这些社交网络平台允许用户在平台上即时地交流自己的想法,并与彼此建立联系。因为 LBSN 与地理信息紧密相连,LBSN 用户可以记录其所在的位置,为之前访问过的地方写评论,并通过分享或评论朋友的签到信息、评论和评分与朋友进行互动,LBSN 用户分享的数量庞大、内容丰富、成本低的位置信息,评论信息给城市 POI 感知带来了新的思路。为了有效并且快速地感知城市的 POI 分布情况,本文提出了一种基于 LBSN 数据聚类分析的城市 POI 感知方法,针对 LBSN 数据存在的重复率高,数据精度低等问题,对数据进行预处理,包括清洗重复数据、删除缺失值数据、删除虚假数据、数据分类等,以保证数据的有效性;利用一种改进的 DBSCAN 密度空间聚类算法(DBSCAN-H),通过计算两个位置的 Haversine 空间距离,以准确地反映位置数据之间的关系,从而准确地感知城市各类 POI 的分布情况。

1 问题模型

本文的研究目标是准确地获取城市各类 POI 的分布情况。由于 LBSN 用户上传的签到数据往往存在大量重复,精度低等问题,甚至存在一些虚假恶意的数据。因此,首先需要对 LBSN 数据进行预处理,包括删除重复、缺失值和虚假的数据,对数据进行特征分析;得到有效性较高的 LBSN 数据后,对数据进行聚类分析,得出较为准确的城市各类 POI 的分布情况。具体的城市 POI 感知研究步骤如图 1 所示。

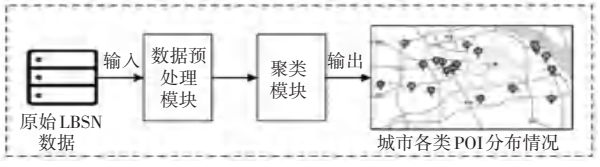

图 1 城市 POI 感知研究步骤

Fig. 1 Research steps of urban POI perception

2 LBSN 数据预处理与 DBSCAN-H 聚类算法

2.1 LBSN 数据预处理

原始的 LBSN 数据集存在数据不完整以及准确性较低的问题,甚至存在一些虚假恶意的数据,如果直接使用原始的数据集进行聚类分析,效率较低,且

得到的城市 POI 的分布情况也不准确。为了获取准确的城市 POI 分布情况，数据预处理是必不可少的一个步骤。数据预处理的方法主要包括数据清理、数据集成、数据变换和数据规约[13]。数据清理指通过填写缺失的值、光滑噪声数据、识别或删除离群点并解决不一致性来清理数据，主要是对重复，异常，错误的数据进行处理；数据集成指将不同来源、不同格式的数据集成在一起，实现统一存储，从而保证数据集的完整性；数据变换指利用规范化、归一化等方法将原本不规范的数据转化成适用于数据挖掘的形式；数据规约指只利用原始数据集上的部分优质数据，通过分析这些数据，在保证了挖掘质量的同时缩短挖掘的时间。上述的方法，往往根据不同数据集的要求，按需使用。本文对 LBSN 数据进行处理的流程如图 2 所示。

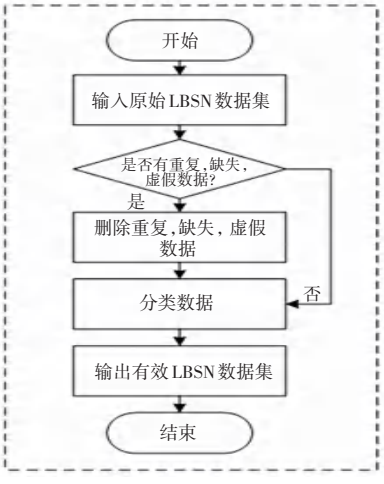

图 2　LBSN 数据预处理流程

Fig. 2　LBSN data preprocessing process

（1）数据重复处理：本文对 LBSN 数据中位置经纬度相同的数据进行删除，防止重复的数据在后续的聚类分析中被重复的计算，以保证聚类结果的准确性。

（2）数据缺失处理：在城市 POI 的感知过程中，数据集中最重要的一个特征就是数据包含的位置。虽然 LBSN 数据是基于位置的社交用户上传的数据，但有些用户为了保护其自身的隐私，会隐藏其位置信息，且这些隐藏的位置信息一般难以通过其他手段填补，因此本文针对缺失位置信息的数据进行删除操作。

（3）虚假数据处理：由于 LBSN 平台是一个开放的平台，所有人都可以注册。在这些用户中不乏一些恶意的用户，上传数据中的位置信息往往是虚假的，这些虚假的位置信息会严重影响城市 POI 感知的准确性。本文采用确定城市的经纬度范围，将不属于这个范围内的数据进行删除，排除这些虚假数据，以保证使用的数据都是属于该城市的。

（4）数据分类处理：LBSN 数据包含的丰富的位置信息往往属于不同的 POI 类别。为了更准确地研究城市各类 POI 的分布情况以及其布局特征规律，本文将 LBSN 数据按照不同的行业类别进行分类操作。

2.2　DBSCAN-H 聚类算法

聚类算法是研究分类问题的一种统计分析方法，在数据挖掘的过程中起到十分重要的作用[14]。聚类算法的目标是将一组数据集按照特定的标准分为不同的簇，使同一个簇中的数据的特征往往是一致的。随着机器学习、深度学习技术的快速发展，聚类算法也成为研究热点。常见的聚类算法有：K-means、谱聚类、DBSCAN 等，这些聚类算法都有不同的适用范围。K-means 算法是一种简单快速的聚类算法，使用场景包括数值型聚类问题，文本聚类等[15]；谱聚类算法是建立在图谱理论基础上的一种聚类算法，能够在任意形状的样本空间聚类且收敛于全局最优解，被广泛应用于计算机视觉、图形聚类等应用场景[16]；DBSCAN 算法是一个具有代表性的基于密度空间的聚类算法，对于空间数据的聚类具有很好的效果[17]。在本文的基于 LBSN 数据的城市POI 感知方法中，LBSN 用户上传的数据中包含着大量的位置信息以及对该位置的描述，考虑到各种聚类算法的应用场景以及优缺点，本文采用一种改进的 DBSCAN 聚类算法实现对基于 LBSN 数据的聚类分析，最终得到城市中各类 POI 的分布情况。

DBSCAN 算法基本定义：

（1）邻域：给定任意样本点 a，该点半径为 Eps 内的区域称为样本点 a 的邻域，Eps 称为邻域半径。

（2）核心点：如果给定样本点 a 的邻域内至少包含 MinPts 个样本点，那么称点 a 为核心点，MinPts 称为簇最小点数。

（3）密度直达：对于样本集合 D，如果样本点 b 在 a 的邻域内，并且 a 为核心点，称样本点 a 到 b 关于 Eps 和 MinPts 密度直达。

（4）密度可达：对于样本集合 D，给定一串样本点 $a_1, a_2, \ldots, a_i, \cdots, a_n, a = a_1, b = a_n$，若样本点 a_i 从 a_{i-1} 密度直达，那么样本点 a 到样本点 b 关于 Eps 和 MinPts 密度可达。

（5）密度相连：存在样本集合 D 中的一点 o，如

果样本点 o 到点 a 和点 b 都是密度可达的,那么 a 和 b 密度相连。

一个对城市位置坐标点进行 DBSCAN 聚类的简单示例如图 3 所示,在该示例中设定簇最小点数 MinPts = 3,邻域半径 *Eps* 如图 3 中双向箭头所示,星形点为核心 POI 坐标点,方块、三角、椭圆的位置坐标点称为边界 POI 坐标点,圆形点称为噪声 POI 坐标点,方块点与核心 POI 坐标点 1 的关系为密度直达,与其他两个核心 POI 坐标点的关系为密度可达,与三角和椭圆点的关系为密度相连,密度相连具有对称性,因此三角点与方块和椭圆点的关系也为密度相连,椭圆点与方块和三角点的关系也为密度相连。在城市 POI 感知的过程中,DBSCAN 算法的目的是找到密度相连位置坐标点的最大集合。

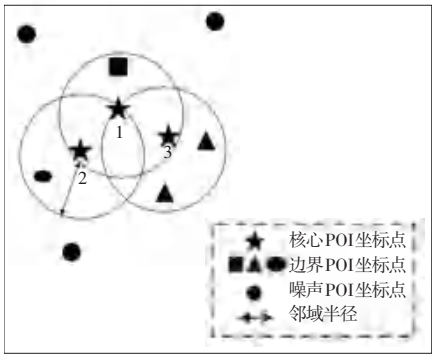

图 3 DBSCAN 聚类示例图

Fig. 3 DBSCAN cluster example

DBSCAN 算法的聚类效果取决于邻域半径 *Eps* 以及簇最小点数 MinPts 的取值,而针对不同的数据集中的数据维度以及数据量的不同,这两个参数往往难以确定。在目前的研究中,最常见的参数确定方式是使用 K-dist 图[18]。使用 K-dist 图确定这两个参数的过程:首先,计算数据集中每个样本点与其第 K 个邻点间的距离 k - distance;其次,对每一个样本点的 k - distance 从大到小排序并绘制 K-dist 图;最后,把 K-dist 图中斜率出现明显拐点的 k - distance 值作为邻域半径 *Eps*,K 的取值作为簇最小点数 MinPts。

在计算两个样本点之间 k - distance 的过程中,传统的计算方式为计算两个样本点之间的欧氏距离(Euclidean Distance),欧氏距离代表着在欧几里得空间中两点的直线距离。假设两个样本点的经纬度坐标为 $A(x_1,y_1)$,$B(x_2,y_2)$。那么 A、B 之间的欧氏距离 d 可由公式(1)计算:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1}$$

但欧氏距离往往并不能准确的反映出两个位置坐标的真实距离,因此本文使用 Haversine 距离来表示两个位置之间的距离。Haversine 距离公式是一种基于球面模型的地理空间坐标点之间的距离计算公式,通过该方法来计算两个地球上的位置坐标之间的距离能够比较准确的反映两个位置之间的真实距离,因此本文使用 Haversine 距离公式来计算两个样本点之间的 k - distance,针对 $A(x_1,y_1)$,$B(x_2,y_2)$ 这两个经纬度坐标,两点之间的 Haversine 距离 D 可由公式(2)计算:

$$D = 2r \times \arcsin\sqrt{\sin^2\left(\frac{y_1 - y_2}{2}\right) + \cos y_1 \cos y_2 \times \sin^2\left(\frac{x_1 - x_2}{2}\right)} \tag{2}$$

其中,r 为地球的半径 6 371 km,x_1、x_2、y_1、y_2 为弧度值。

本文提出的 DBSCAN-H 算法的主要流程为:首先,计算 LBSN 数据集中的所有位置坐标点与其第 K 个临近点的 Haversine 距离作为该位置坐标点的 k - distance;其次,对 k - distance 排序并绘制 K-dist 图,从而确定 DBSCAN-H 算法的参数,即邻域半径 *Eps* 以及簇最小点数 MinPts;初始化簇的个数为 0,并从 LBSN 数据集中选取一个未处理的位置坐标点 a,若该点的 *Eps* 邻域内包含的样本点数不少于 MinPts,则样本点 a 作为核心 POI 坐标点,以 a 点为核心创建一个簇,并将 *Eps* 邻域内与 a 点密度直达的样本点加入该簇中;将与该簇中所有核心 POI 坐标点密度可达的位置坐标点加入到该簇,直到所有与 a 点密度相连的位置坐标点都加入到该簇;最后,选择一个未被加入任何簇的位置坐标点,重复上述过程,直到没有新的位置坐标点可以加到任何一个簇中,聚类算法结束,未被加入任何簇的位置坐标点为噪声 POI 坐标点。使用 DBSCAN-H 算法对 LBSN 数据集进行聚类的算法伪代码如算法 1 所示。

算法 1 DBSCAN-H 聚类算法

输入 LBSN 位置坐标点集合 D。

输出 聚类结果 C

1. 使用 Haversine 距离公式计算所有样本点的 k - distance,确定邻域半径 *Eps* 以及确定簇最小点数 MinPts。

2. Number_C=0 //初始化簇的个数为 0

3. for 样本集合 D 中每一个未访问点 a

4. 将 a 标记为已访问
5. 计算样本点 a 邻域半径 Eps 内的样本点个数 N
6. if N < MinPts
7. 将 a 标记为噪声点
8. else
9. 建立新簇 C
10. 将 a 加入到簇 C 中
11. for 样本点 a 邻域半径内的未被访问的样本点 a'
12. 将 a' 标记为已访问
13. 计算样本点 a' 邻域半径 Eps 内的样本点个数 N'
14. if N' > MinPts
15. $N = N + N$'
16. end if
17. if a' 不属于任何其他簇
18. 将 a' 加入到簇 C 中
19. end if
20. end for
21. end if
22. end for

3 实验分析

3.1 数据集介绍

为了验证本文所提算法的有效性，使用真实的开源 Foursquare 数据集来进行模拟实验。Foursquare 数据集为纽约市的移动用户 2012 年 4 月到 2013 年 2 月间在各个地点的打卡数据，包含 1 083个用户，227 428 条签到数据，涉及到的地点 42 981个。

对该数据集进行预处理后，本文获取到该数据集中分类为餐饮类的数据 4 998 条，在此基础上进行相关实验。

3.2 对比算法与评价指标

3.2.1 对比算法

(1)传统 DBSCAN 算法(DBSCAN-E)[19]：使用 DBSCAN 算法对城市餐饮购物类 POI 数据进行分析，已获得城市的商业结构特征，该算法确定邻域半径以及簇最小点数的方法，公式(3)：

$$Eps = \sum_{k=1}^{20} mean_i/k/n \tag{3}$$

其中，$mean_i$ 是围绕第 i 个核心点的最近 POI 距离之和的平均值；k 是邻近核心点的最近 POI 的数量；n 为 POI 总数。

确定好参数 Eps 后，通过计算每个点的邻域半径内点的数量的期望值作为聚类最小点数，公式(4)：

$$\text{Minpts} = \frac{1}{n}\sum_{i=1}^{n} count_i i \tag{4}$$

其中，$count_i i$ 是样本点 i 的邻域内样本点的数量。

(2) K-means 聚类算法[20]：是一种基于无监督学习和划分的聚类算法。K-means 聚类将样本集合划分成 k 个子集，构成 k 个类。在该对比文献中，K-means 算法被用来对地铁车站相关的 POI 数据进行聚类分析，以实现地铁车站的精细化分类。

3.2.2 评价指标

(1) *CH* (Calinski-Harabaz index)指数：通过计算类中各点与类中心的距离平方和来度量类内的紧密度，通过计算各类中心点与数据集中心点距离平方和来度量数据集的分离度，*CH* 的值由分离度与紧密度的比值得到。*CH* 值越大代表类自身越紧密，类与类之间越分散，聚类结果更优。*CH* 指数的具体计算公式(5)：

$$CH = \frac{tr(\boldsymbol{B}_k)}{tr(\boldsymbol{W}_k)} \times \frac{N_E - k}{k - 1} \tag{5}$$

其中，N_E 表示数据集训练样本数；k 表示类别数；$\boldsymbol{B}_k$ 表示类别之间的协方差矩阵；$\boldsymbol{W}_k$ 表示类内样本点之间的协方差矩阵；tr 表示矩阵的迹。

(2) *DBI* (Davies-Bouldin Index)指数：又称为分类适确性指标，是用来评估聚类算法优劣的一个重要指标。*DBI* 指数是计算任意两类别的类内距离平均之和除以这两类的中心距离，并求最大值。*DBI* 值越小代表着类内距离越小，类间距离越大，聚类结果越优。*DBI* 指数的具体计算公式(6)：

$$DBI = \frac{1}{k}\sum_{1}^{k} \max \frac{s_i + s_j}{d_{i,j}}, i \neq j \tag{6}$$

其中，s_i 表示类 i 中每个点与该类质心的平均距离，$d_{i,j}$ 表示类 i 和类 j 质心之间的距离。

3.3 聚类结果分析

本文使用 K-dist 图来确定 DBSCAN 算法在 Foursquare 数据集下的邻域半径以及簇最小点数。在 $K = 4$ 的情况下，Foursquare 数据集基于 Haversine 距离的 K-dist 图如图 4 所示，明显看出在 K-dist 图斜率出现明显拐点时，Foursquare 数据集的邻域半径 Eps 的范围在[0.2,0.4]。在具体的实验中，本文设定该数据集的邻域半径 Eps 为 0.4。

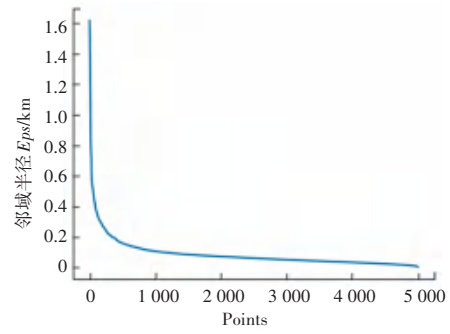

图 4 Foursquare 数据集基于 Haversine 距离的 K-dist 图

Fig. 4 K-dist diagram based on Haversine distance in Foursquare dataset

通过本文提出的 DBSCAN-H 算法,Foursquare 数据集上的所有的样本点被识别为 14 个不同的类,41 个噪声点。Foursquare 数据集 DBSCAN-H 算法聚类结果如图 5 所示。

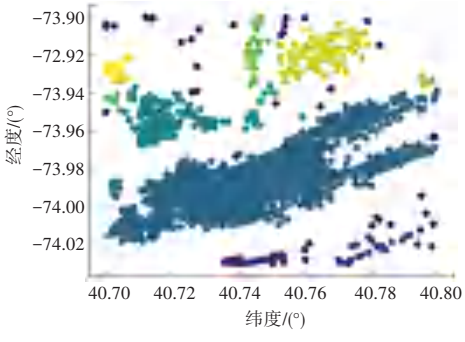

图 5 Foursquare 数据集 DBSCAN-H 算法聚类结果

Fig. 5 Clustering result of DBSCAN-H algorithm in Foursquare dataset

纽约真实地图聚类结果如图 6 所示,可以发现纽约市的餐饮行业十分繁荣,基本覆盖了整个城市,且大部分的餐饮店都有集聚的现象。

图 6 纽约真实地图聚类结果

Fig. 6 Clustering result in real New York map

Foursquare 数据集 DBSCAN-E 算法聚类结果如图 7 所示,将 DBSCAN-E 算法中的邻域半径 *Eps* 设置为 0.6,簇最小点数 MinPts 设置为 16;Foursquare 数据集 K-means 算法聚类结果如图 8 所示,K-means 算法中簇的个数 *K* 与 DBSCAN-H 算法聚类结果中簇的个数都设置为 14。如图 7 中 DBSCAN-E 算法聚类生成的簇的数量为 6,明显少于图 5 中使用 DBSCAN-H 算法生成的簇的数量 14。有一些能够可以分成不同簇的样本点被分到了同一个簇,这是因为使用 Haversine 距离来计算两个地理位置之间的距离能够更加准确的反映各个样本点之间的真实距离,因此聚类效果也更好。对比图 5 和图 8,虽然两种算法聚类的簇的个数一样,但是可以明显的发现在 Foursquare 数据集上 K-means 聚类算法的效果比较差,有大量明显属于同一个类的样本点被聚集到不同的簇中。

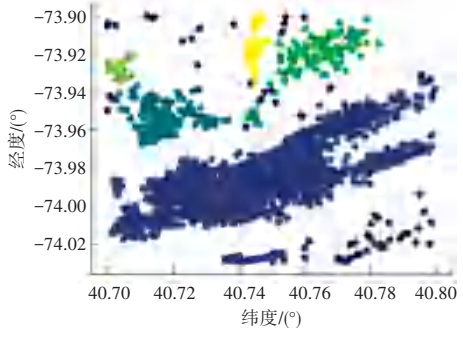

图 7 Foursquare 数据集 DBSCAN-E 算法聚类结果

Fig. 7 Clustering result of DBSCAN-E algorithm in Foursquare dataset

本文提出的 DBSCAN-H 算法与两种基准算法的 *CH* 值与 *DBI* 值见表 1,可以看出 DBSCAN-H 算法的 *CH* 值大于其他两个基准算法的 *CH* 值,说明在使用 DBSCAN-H 算法得到的聚类结果中类之间的关系更加紧密,类与类之间的关系更分散,聚类的结果更优。本文提出的 DBSCAN-H 算法的 *DBI* 值比其他两个基准算法的 *DBI* 值小,说明类内距离越小,类间距离越大,也同样表明本文提出的算法的聚类结果更优。

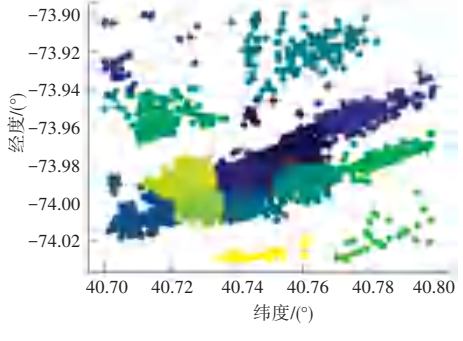

图 8 Foursquare 数据集 K-means 算法聚类结果

Fig. 8 Clustering result of K-means algorithm in Foursquare dataset

表 1　不同算法 *CH* 值与 *DBI* 值对比表

Table 1　Comparison of *CH* value and *DBI* value of different algorithms

	CH	*DBI*
DBSCAN-H	2 644 883.5	0.020 2
DBSCAN-E	937 006.6	0.029 4
K-means	5 894.1	0.792 9

4　结束语

本文提出了一种改进的 DBSCAN 算法对基于位置社交网络的数据进行聚类分析，从而快捷、准确地感知城市中各类 POI 的分布情况。针对 LBSN 数据中存在的数据质量较低的问题，本文首先对这些数据进行了相应的预处理，为聚类算法提供更加精准、有效的数据；其次，通过计算两个位置坐标的 Haversine 距离来确定 DBSCAN-H 算法的参数，使该算法更适合本文中的地理空间数据的聚类分析。实验结果表明，与两种基准算法对比，本文提出的 DBSCAN-H 算法在聚类的效果以及聚类评价指标方面具有更好的表现。在今后的工作中，在得到城市各类 POI 准确的分布后，将通过对 LBSN 用户上传的数据进行时间序列分析，挖掘城市各类 POI 的生命周期变化，从而实现对城市 POI 的全生命周期感知。

参考文献

[1] ZHANG Wei, GAO Xinyuan, LI Ruishan. Multi-source POI data fusion based on the spatial location information[J]. Periodical of Ocean University of China, 2014, 44(7): 111-116.

[2] 路新江. 基于移动感知数据的城市兴趣点生命周期预测研究[D]. 西安：西北工业大学，2018.

[3] 李建春，起晓星，袁文华. 基于 POI 数据的建设用地多功能混合利用空间分异研究[J]. 地理科学进展，2022，41(2)：239-250.

[4] 张家旗，刘晏男，宋斌玢. 基于 POI 数据的郑州市主城区生活服务业空间分布特征研究[J]. 世界地理研究，2022，31(2)：399-409.

[5] WAHURWAGH R A, CHOURAGADE P M. Personalized POI travel recommendation with multiple tourist information [C]// Proceedings of IEEE International Conference on Electronics, Communication and Computing Technologies. Coimbatore: IEEE, 2019: 1-3.

[6] LYNDON S K, MOR N. Generating diverse and representative image search results for landmarks[C]// Proceedings of the 17th International Conference on World Wide Web. Beijing: IEEE, 2008: 297-306.

[7] SLAVA K, FLORIAN M, DANIEL K.P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geotagged photos[C]// Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application. Washington, D. C: IEEE, 2010: 38.

[8] YANG Yiyang, GONG Zhiguo. Identifying points of interest by self-tuning clustering[C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing: IEEE, 2011: 883-892.

[9] YANG Yiyang, GONG Zhiguo. Identifying points of interest using heterogeneous features [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2015, 5(4):68.

[10] KYOSUKE N, HIROYUKI T, TAKESHIK, et al. Probabilistic identification of visited point - of - interest for personalized automatic check - in [C]// Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Seattle: IEEE, 2014:631 - 642.

[11] ZHOU M, WANG M, HU Q. A POI data update approach based on Weibo check-in data[C]// Proceedings of the 21st International Conference on Geoinformatics. Kaifeng:IEEE, 2013:1-4.

[12] YANG Y, DUAN Y, WANG X, et al. Hierarchical multi-clue modelling for POI popularity prediction with heterogeneous tourist information [J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(4): 757-768.

[13] VOLKAN C, OKTAY A. A comprehensive review on data preprocessing techniques in data analysis [J]. Journal of Engineering Sciences, 2022, 28(2):299-312.

[14] ZHANG Yonglai, ZHOU Yaojian. Review of clustering algorithms [J]. Journal of Computer Applications, 2019, 39(7): 1869-1882.

[15] 李珺，刘鹤，朱良宽. 基于改进的 K-means 算法的关联规则数据挖掘研究[J]. 小型微型计算机系统，2021，42(1)：15-19.

[16] 王少将，刘佳，郑锋. 机器学习层谱聚类综述[J]. 计算机科学，2023，50(1)：9-17.

[17] DENG D. DBSCAN clustering algorithm based on density[C]// Proceedings of the 7th International Forum on Electrical Engineering and Automation (IFEEA). Hefei:IEEE, 2020: 949-953.

[18] 杨帆，徐建刚，周亮. 基于 DBSCAN 空间聚类的广州市区餐饮集群识别及空间特征分析[J]. 经济地理，2016，36(10)：110-116.

[19] 翟青，高玉洁，魏宗财. 基于 DBSCAN 的南京商业空间聚类研究[J]. 南京邮电大学学报(社会科学版)，2022，24(3)：82-92.

[20] 赵源，王越，胡华. 基于 POI-K-means 地铁车站聚类方法研究[J]. 智能计算机与应用，2022，12(5)：114-118.