中文图书分类号：TP391
密　　级：公开
UDC：004
学　校　代　码：10005

硕士学位论文

MASTERAL DISSERTATION

论　文　题　目：基于用地特征的城市轨道交通站点分析方法研究

论　文　作　者：李彤

学　　科：计算机科学与技术

指　导　教　师：才智　副教授

论文提交日期：2019 年 5 月

UDC：004　　学校代码：10005
中文图书分类号：TP391　　学　　号：S201607054
密　　级：公开

北京工业大学工学硕士学位论文

题　　目： 基于用地特征的城市轨道交通站点分析方法研究

英文题目： RESEARCH ON ANALYSIS METHOD OF OF URBAN RAIL TRANSIT BASED ON LAND USE CHARACTERISTICS

论 文 作 者： 李彤

学 科 专 业： 计算机科学与技术

研 究 方 向： 大数据技术与应用

申 请 学 位： 工学硕士

指 导 教 师： 才智 副教授

所 在 单 位： 信息学部（原计算机学院）

答 辩 日 期： 2019 年 5 月

授予学位单位： 北京工业大学

独 创 性 声 明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京工业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签　　名：_____

日　　期：　　年　　月　　日

关于论文使用授权的说明

本人完全了解北京工业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签　　名：　　日　　期：　　年　　月　　日

导师签名：　　日　　期：　　年　　月　　日

摘 要

随着现代科技和经济的迅速发展，城市轨道交通已成为城市客运交通系统的重要组成部分。站点客流影响未来轨道交通站点的规划建设，同时也是轨道调度运营的可靠依据。国内外学者根据站点历史客流数据提出了诸多预测不同条件下站点客流的方法，但这并不能预测规划建设和正在建设的站点客流量，与此同时，可靠的站点规划方法少之又少。为解决上述问题，本文提出了基于用地特征计算站点相似性方法，证明了土地使用相似的站点具有相似的客流量，为城市轨道交通站点选址和土地开发提供思路。本文在城市轨道交通站点周边用地特征和客流方面的主要研究如下：

(1) 提出了划分有界区域的 RC-*tree*（Colored R-*tree*）算法。为了分析站点周边用地情况，本文利用 POI(Point of Interest)的经度和纬度特征生成 R-*tree*，利用提出的算法将 R-*tree* 着色，从而划分区域，为提取 *k* 个 POIs 做准备。

(2) 在 RC-*tree* 算法生成的边界区域内，基于空间和语义，利用多样性或比例性方法提取 *top-k* POIs。每个站点周边有部分的 POIs 是相同的，这些 POIs 在计算站点之间的相似度时会产生很大误差，故提取 *k* 个有代表性的 POIs 从空间和语义两方面计算站点相似度。为实现在有界区域内获取 *top-k* POIs，本文在原有的多样性和比例性方法进行了修改。

(3) 利用提取的 *k* 个 POIs，计算站点间的相似性。POI 属性包含名字和类别，故在计算站点间相似度时，不仅利用 *k* 个 POIs 的文本相似性，而且更注重 POIs 的关系相似度。

(4) 基于各站点进出分时客流，计算站点间的相似度。为验证基于用地特征的站点相似性方法的准确性，分别统计站点进出双方向分时 AFC 刷卡记录，计算站点间的相似度。

(5) 基于天气特征和土地利用特征的站点分类。为再次验证多样性和比例性方法计算站点相似的可靠性，利用机器学习方法基于用地特征站点相似性将站点分类，将天气特征（即雨天和非雨天）对客流影响的分类结果作为验证集。

比较基于进出分时客流和用地特征的站点相似性，以及基于天气特征和用地特征的站点分类，实验结果表明本文提出基于用地特征的站点相似性方法有效。利用基于用地特征计算站点相似性，将站点分类，分析和验证站点用地特征和客流的关系以及天气对客流的影响，这对规划城市轨道交通线网、站点选址、调整

站点周边土地的使用和缓解运营压力起到指导性作用。

关键词：轨道交通；RC-*tree* 算法；有界区域的多样性和比例方法；客流量；相似度

Abstract

With the rapid development of modern technology and economy, urban rail transit has become one of important components of the urban transportation system. Passenger flow of urban rail transit not only affects the planning and construction of the future rail transit station, but also is the basis of orbital operation scheduling. Although domestic and foreign scholars have proposed a number of methods to predict passenger flow based on historical data of stations, these methods can't preditc the passenger flow of planning and construction and constructing stations in the absence of historical data. In view of this limitation, this paper proposed a method to calculate the similarity of stations based on the land use characteristic around rail transit stations, which proves stations with similar land use have similar passenger flow. The main researches of this paper on land use and passenger flow around urban rail transit are as follows:

(1) Proposed the RC-*tree*(Colored R-*tree*) algorithm for dividing bounded regoins. In order to analyze the land use around stations, this paper used the longitude and latitude characteristics of POI (Point of Interest) to generate R-*tree*, and used the proposed algorithm to color R-*tree*, so as to divide regions and prepare for extracting *k* POIs.

(2) Extracted *top-k* POIs with diversity or proportion methods based on spatial and semantic within the boundary region generated by the RC-*tree* algorithm. It produces large errors when calculating the similarity between the stations with same POIs. Therefore, this paper extracted *k* representative POIs to calculate the similarity between stations based on both spatial and semantic aspects. In order to achieve *top-k* POIs in bounded areas, the original diversity and proportion methods were modified.

(3) Calculated the similarity between stations by using the most representative *k* POIs extracted. It includes name and category with attributes of POI. So when calculate the similarity between stations, not only use the text similarity of *k* POIs, but also pay more attention to the relationship similarity of POIs.

(4) Calculated the similarity between stations based on the time interval passenger flow of each station's entry and exit. In order to validate the accuracy of similarity method of stations based on land use characteristics, This paper proposes a similarity calculation method based on, and to verify the accuracy of the method,

Calculated the similarity between entry and exit of station with records of AFC.

(5) Classified stations based on weather characteristics and land use characteristics. In order to verify the reliability of diversity and proportion methods in calculating similarity of stations again, used the machine learning method to classify stations based on land use characteristics and used as verification sets with the classification results of weather characteristics (i.e. rainy and non-rainy).

Comparing the similarity of stations based on the time interval passenger flow of each station's entry and exit and land use characteristics, as well as the classification of stations based on weather and land use characteristics, the experimental result showed that it is effective of the proposed similarity method of stations based on land use characteristics. It plays a guiding role in planning urban rail transit, selecting location, adjusting the use of land around the station, and alleviating the pressure on passengers by using land use characteristics to calculate similarity of stations, classify stations, analyze and verify the relationship between land use characteristics and passenger flow, and the impact of weather on passenger flow.

Keywords: Urban Rail Transit, RC-*tree* Algorithm, Diversity and Proportion methods in the Boundary Region, Passenger Volume, Similarity

目 录

摘 要.....I
第 1 章 绪论.....1
1.1 研究背景及意义.....1
1.1.1 我国轨道交通发展现状.....1
1.1.2 我国轨道交通存在的主要问题.....3
1.1.3 研究意义.....4
1.2 国内外研究现状.....4
1.2.1 轨道交通站点客流研究.....4
1.2.2 轨道交通站点周边用地研究.....6
1.2.3 天气特征对轨道交通站点客流影响研究.....6
1.3 本文主要研究内容.....7
1.4 本文组织结构.....9
第 2 章 数据处理.....11
2.1 基础数据来源.....11
2.2 OD 分时客流量的相似度计算方法.....12
2.3 北京市 POI 数据预处理.....16
2.4 本章小结.....18
第 3 章 划分有界区域的 RC-*tree* 算法.....19
3.1 RC-*tree* 定义.....19
3.1.1 R-*tree* 结构.....19
3.1.2 RC-*tree* 思想.....20
3.2 算法实现.....20
3.3 本章小结.....24
第 4 章 基于用地特征的站点相似性计算方法.....25
4.1 多样性方法.....25
4.1.1 基于空间的多样性方法.....25
4.1.2 基于语义的多样性方法.....27
4.2 比例性方法.....27
4.3 基于有界区域的多样性和比例性方法.....28

4.4 基于用地特征的站点相似度.....32
4.4.1 文本相似性.....32
4.4.2 类别相似性.....32
4.5 本章小结.....33
第 5 章 OD 分时客流量相似度与基于用地特征的站点相似度比较.....35
5.1 近似质量，参数 k 和 β 的取值.....35
5.1.1 近似质量.....35
5.1.2 参数 k.....35
5.1.3 参数 β.....35
5.2 实验结果与分析.....36
5.2.1 k 值变化.....36
5.2.2 β 值变化.....39
5.2.3 *NUM* 值变化.....40
5.3 本章小结.....41
第 6 章 基于天气特征和用地特征的站点分类.....43
6.1 基于用地特征的站点分类.....43
6.2 不同天气特征客流量的站点分类.....44
6.3 实验结果与分析.....46
6.4 本章小结.....50
结 论.....51
参 考 文 献.....53
攻读学位期间发表的学术成果.....57
致 谢.....59

第 1 章　绪论

1.1　研究背景及意义

预计到 2050 年，城市居民将占全球人口的 75%，快速的城市化发展已使基础设施捉襟见肘，完善城市轨道线网，与其他公共交通连通，能帮助城市无缝工作，更好发展。

城市轨道交通已成为大部分城市通勤的最重要方式。完善城市轨道交通建设，不仅提高市民的出行效率、改善生活质量，更能缓解城市中心人口密集、住房紧张的压力和改善空气质量、绿化面积小等城市通病。城市轨道交通具有客运量大、正点率高、交通事故发生率低等特点。与此同时，随着轨道交通发展的不断完善，站点周边用地也倍受关注，土地开发强度和性质等都有了明显的变化，这也给轨道交通带来更多效益。

城市轨道交通的迅速发展，促使客流日益增长，吸引越来越多的学者研究站点客流分布、预测、换乘等问题，并且研究站点客流对周边用地的影响也颇多。但目前所提出的方法仅对已有历史数据的站点适用，本文提出基于用地特征的城市轨道站点分析方法，证明了相似的站点具有相似的客流，对站点选址提供思路，以及对正在建设和规划的站点客流预测具有指导意义。

本章阐述了我国轨道交通的发展现状及存在的问题，以北京市轨道交通为例，介绍了北京市轨道交通发展现状，提出了基于用地特征的轨道交通站点分类方法的研究意义。

1.1.1　我国轨道交通发展现状

我国的城市轨道交通是城市公共交通系统的核心，主要有地铁、轻轨、市郊铁路、有轨电车以及悬浮列车等多种类型。国外城市轨道交通建设起步较早，发达国家早已形成完善的城市轨道交通网络。由于经济实力和技术水平的局限性，我国城市轨道交通建设起步较晚，在 2000 年之前，仅北京、上海、广州三个城市拥有轨道交通线网。随着我国经济的飞速发展和城市化进程的加快，城市轨道交通进入大发展时期，越来越多的城市已经制定了城市轨道交通系统发展计划，特别是以安全和速度著称的地铁，其建设不仅缓解地面交通拥堵，更能促进就业和推动城市化进程。截至 2018 年底，我国内陆地区已有 24 个省份、35 个城市、

178 条城市轨道交通线路投入运营，总里程长达 5398 公里，客运量超过 200 亿人次[1]。

如今，我国拥有世界上最长的地铁系统，北京作为我国的首都和轨道交通发展最为迅速的城市，机动车保有量增长迅速，地面交通体系和管理面临严峻挑战。为有效缓解地面交通压力，北京市高度重视轨道交通建设，城市轨道交通建设和发展令人瞩目。

北京是全国最早建设轨道交通的城市，地铁一号线是中国第一条轨道交通线路，始建于 1965 年 7 月，1969 年 10 月通车。2000 年之前，北京的轨道交通运行里程仅有 114 公里。进入 21 世纪以来，为迎接 2008 年北京奥运会，北京市进一步加快轨道交通建设，逐步完善城市轨道交通网，着力建立一个以公共运输网络为主体、以快速轨道交通为骨干、与城市发展进程相适应的现代化城市综合交通体系。

截至 2018 年 12 月，北京地铁运营线路共有 22 条，均采用地铁系统，覆盖北京市 11 个辖区，运营里程 637 公里，共设车站 391 座。如图 1-1 所示，其中换乘站 59 座（两线换乘站 56 座、三线换乘站 3 座），换乘站所占比例为 15.09%，年客流量已超过 45 亿人次。

图 1-1 北京市轨道交通线路图

Fig. 1-1 Beijing rail transit route map

1.1.2　我国轨道交通存在的主要问题

随着轨道交通的快速发展，站点客流也大幅增加，出现了一些急需解决的问题。例如：高峰时段乘客量大，列车超载，乘车难；站点客流差异大，站点密度和客流不均衡；站外限流，等待时间较长；换乘量大，换乘难；站点规划依据方法少等问题。

高峰时段乘车难，列车运营超载现象严重。根据各线高峰每小时断面流量和满载率数据分析，大部分线路的高峰满载率大于 1，轨道交通设施供给能力与客流需求间的矛盾较为突出[2]。

部分站点周边用地特征类型较为综合，站点流量较大，且站点密度较低，不仅给乘客带来困扰，而且给运营管理增添压力。以北京市为例，西二旗站和上地站影响用地范围大，周围公司和住宅较多，客流很大，目前周边只有两个站点，造成乘车拥挤。

高峰时段，满载、超载情况较为严重，运营部门在尽最大所能调整运行图，缩小发车间隔、提高运力的同时，对一些进站流量高的车站采取了站外临时限流的措施，以保证运营安全，但严重影响出行效率。例如北京市轨道交通网络共有 59 个换乘车站，各换乘站最大换乘客流多出现在工作日早高峰时段，其中西直门站换乘客流最大；东直门、西二旗、国贸等站换乘客流量紧随其后。

针对站点规划问题，目前主要有解析法和系统分析法两种规划方法。解析法就是根据城市人口和土地分布，运用数学和运筹学建立、求解目标函数，找到最优路径；所谓系统分析法，是利用客流和土地发展构建初始线路，之后再使用交通规划方法预测线路的合理性，选择最好的线路实施，该方法结合了历史数据和经验。而实际中，土地使用情况不断发生改变，现有的方法无法利用土地使用特征预测规划站点的规模。

除了上述问题外，我国轨道交通运行不足还表现在轨道网断线多、半径线多，乘客被动换乘，严重影响出行效率；随着疏解人口和房价门槛的双重作用，超过 50 公里圈层的通勤客流需求也正在逐步增长，由于目前缺乏应有的市域快线系统，导致轨道交通长距离出行速度慢、效率低，直接影响了轨道交通对长距离出行的吸引力。

轨道交通客流量特征实质上是站点乘客时间累加、空间分布和交换，因此对站点客流特征的研究十分必要。同时，站点周边土地的使用情况是导致站点客流量不均衡的重要原因之一。对站点周边土地使用研究亦十分重要，这对站点的规划建设，运营管理和站点周边土地开发起指导作用。

1.1.3 研究意义

依据城市轨道交通站点周边的土地利用特征，计算站点相似性和站点客流的关系，掌握站点客流的特征，及周边用地特征对客流的规模、时间分布等特征对站点的影响，可以作为预测相似站点客流是否合理的参考依据。

城市轨道交通作为城市总体规划的一部分，以周边的土地利用为基础进行规划，通过对站点客流规模和周边土地利用关系的研究，掌握各类站点用地特征对站点客流的影响，在规划城市轨道交通时，结合各类站点周边用地交通需求的特征，使轨道交通建设更加合理。

基于用地特征相似的站点分类，可根据正在建设和将要建设的站点周边的土地利用情况，找到其站点属于的类别，即使没有历史客流量数据也可预测该站点的客流，为设计站点规模提供重要依据，使建设站点更具有合理性。

根据不同天气特征对客流的影响，和对站点进站客流时间分布特征和站点周边土地利用关系的研究，分析出同种类型站点的客流时间分布的特征，为运营组织提供建议，建立多样化的运营组织体系。

通过站点周边土地的利用对轨道交通站点客流的影响分析，掌握轨道交通站点与周边用地的互动关系，对站点周边土地的合理利用与开发提供指导作用，使土地利用和轨道交通相互促进，共同发展。

1.2 国内外研究现状

随着国内外城市轨道交通的不断发展，国内外学者对轨道交通的研究也不断深入。本文的研究涉及城市轨道交通系统客流、站点周边土地利用情况和天气对站点客流的影响。国内外的相关研究将在以下三个小节中详细介绍。

1.2.1 轨道交通站点客流研究

城市轨道交通系统的管理在很大程度上取决于对客流的分析。为了实现科学合理的列车调度，满足乘客的出行需求，分析客流具有重要意义。客流是一个定期动态变化过程，研究空间和时间的客流分布有助于及时合理地安排城市轨道交通系统，节省成本，为乘客提供更好的服务。

国内外针对轨道交通客流的研究主要包括客流分布、客流预测和接驳问题。付博峰等[3]主要采用空间句法轴线图模型预测轨道车站出入口客流分布，将预测问题转化为对站点周边路网行人活动分布比例关系，建立预测轨道站出入口客流

分布系数的方法。袁江等人[4]针对目前广州地铁运营出现的问题，在不同时间和空间维度上分析实际运营客流分布特征和规律，提出设计阶段客流预测数据的选用原则、车厢站立密度对站台乘客上下车时间的影响及车站布局形式与站台客流分布的关系。周玮腾[5]通过构建时刻表扩展网络，设计时刻表扩展网络下的 k 短路径搜索算法，考虑列车容量限制下乘客的拥挤因素和留乘延误因素，构建随机用户均衡条件下的客流动态分配模型。有三种实用策略可缓解地铁系统的压力，提高列车班次，提高列车容量和应用快车。徐晓明等[6]提供了模糊乘客到达率和下车率的地铁系统客流变化模型，用于调查上述策略中的客流。

客流分布的预测是城市轨道交通系统管理的基础，尤其是在城市交通网络发生改变，和轨道交通重新规划时。随着机器学习的广泛应用，基于机器学习的方法已经成为预测客流的主流，例如卡尔曼滤波器和基于支持向量机（SVM）的方法。Xu 等[7]使用聚类算法找出高峰时段的客流间隔，从列车调度的角度研究客流时间聚合规则。Yujuan[8]提出了一种预测转移客流的方法，以避免乘客拥堵，该方法基于分解模型方法，并通过代表性的个体方法来合并历史客流数据以分解数据。Shuzhi[9]和 Yang[10]分别使用神经网络和多尺度径向基函数网络来预测地铁中不同客流的不规则变化。目前，关于城市轨道交通系统中不同类型客流预测的研究很少。在 Yuxing[11]中提出了一种新颖的混合模型（即 Wavelet-SVM），它结合了小波和 SVM 模型，以实现准确的客流预测。在分析客流量及其影响因素的基础上，B. Valentina[12]分析了目前的客流预测方法。根据城市之间的客流量，找到了预测城市间公交系统客流量的方法。同时，在客流预测中，模拟方法也是一种流行有效的方法。G. Shengguo[13]根据乘车时间分布计算每条路径上的乘客比例。根据进站和时间，可以获得一段时间内所有乘客的出行时间分布，并且还可以估计该时段内每条路径上的乘客比例。Z. Ling[14]分析了客流的基本因素，包括空间布局，服务对象，乘客规模以及世界大城市轨道交通系统的历史演变。王静等[15,16]从乘客的时空分布和车站客流的变化分析了城市轨道交通系统的客流量。Bizhuang[17]和李金海等[18]从网络的角度找到客流特征，并对城市轨道交通系统的规划，建设和管理提出建议。

在城市轨道交通系统中，许多站点，特别是换乘站，必须按不同的情况处理问题，如乘客拥挤等。然而，大多数现有的研究都是基于历史交通数据，因此，若不利用历史交通数据集很难继续研究。本文通过站点的用地特征对站点分类，站点的客流只作为验证集，这样根据新建和规划的站点其周边土地利用情况也可以预测其客流量。

1.2.2 轨道交通站点周边用地研究

目前，关于客流量与站点周围土地利用之间关系的研究主要集中在土地类型的分类及其对客流的影响。

Jun[19]评估了首尔大都市区的土地使用特征对地铁乘客量的影响。Ma[20]提出了一种多目标规划模型，基于免疫遗传算法的改进算法，将交通和土地的利用集成到公交导向发展（TOD）中为找到土地使用的最优方案。Calvo[21]描述并深入分析了地铁对人口和土地利用的影响。Binglei[22]建立了 TOD 模式下城市轨道交通系统与土地利用之间协调关系的评估体系。Yena[23]描述了地铁系统的演变，揭示了地铁站周围土地成本与人口分布之间的关系。Seungil[24]发现车站周围的可持续土地利用对城市轨道交通系统很重要，因为它对客流有长期影响。由于车站周围土地的开发速度很快，因此对车站周围土地利用特征的探索对于指导新建站和现有站的土地利用规划和调整非常重要[25,26,27]。Kuby[28]利用多元回归模型分析了影响美国轨道交通的客流的因素，结果表明最重要的因素是土地利用等。长期以来，城市土地和城市交通规划的理论和实践只关注城市轨道交通对土地的需求，但没有充分考虑对实时交通的影响。近年来，城市土地利用与城市交通的关系越来越受到学者们的关注，马航[29]就中国城市轨道交通土地利用的发展与问题做了详细的介绍。

诸多学者对城市轨道交通和周边土地利用情况的研究，均是结合站点客流进行分析，并没有就站点的土地利用情况进行分析，本文是基于站点周边 POIs 本身特征进行分析计算站点的相似性和分类，并利用真实的乘客流量数据进行了验证。

1.2.3 天气特征对轨道交通站点客流影响研究

在众多影响轨道交通客流的外部因素中，天气特征变化备受关注。天气变化影响交通出行方式、出行路线等的改变，从而间接影响轨道交通客流的变化。Arana[30]利用多元线性回归方程的方法分析了气象条件对西班牙 Gipuzkoa 城市的公交数据的影响。Zhou[31]通过收集深圳市的公共交通和气象记录的智能卡数据，考虑天气和公共交通乘客的日内变化以及调查天气对个别公共交通用户的乘车行为的影响分析天气条件对公共交通总量的影响。Koetse[32]介绍了关于气候变化和天气条件对交通运输部门影响的实证文献的调查，证明了天气对交通确实有影响。杜恒[33]分析天气因素和轨道交通客流分布的关联关系，通过不同天气下对客流的预测，利用客流预测模型对乘客出行目的地进行预测。Meyer[34]探讨了当

前轨道交通系统适应气候变化的实践，并制定了交通基础设施不同组成部分的概念框架，这些组成部分将受到一系列气候变化的不同影响。

不同天气特征对轨道交通客流的影响，大多数学者是根据已有的客流数据对站点的客流做预测，这些均基于原有客流数据之上，但对于新建或将要建设的站点来说这类方法并不可行，本文是根据站点周边土地的使用情况将站点分类，分析了不同天气特征对各类别站点客流的影响，这使正在规划的站点在不同天气特征下的客流也可预测，对规划轨道交通具有指导意义。

1.3　本文主要研究内容

在城市轨道交通系统中，进出站的客流对站点的建设和管理有很大的影响。此外，根据站点的客流还可以有效的管理和利用站点周边的土地。尽管有很多研究成果[35,36,37]提出了通过客流的历史数据预测站点客流，但对于正在建设和规划建设的站点来说，它们是没有历史数据的，因此利用这些方法很难预测其客流。本文提出了基于用地特征的站点相似度计算方法将站点分类，数据主要以北京市轨道交通为例，证明了相似用地特征的站点具有相似客流，该方法的提出对城市轨道交通的规划、新站点的客流预测和选址具有指导意义。本文的主要研究内容包括：

1. 提出基于进出站分时客流（OD，Origin 出发地点，Destination 目的地）计算站点间相似度的方法。统计每个站点每小时进站客流量，每天出站客流量，计算统计分时客流量的站点相似度。

2. 提出划分 POIs 的 RC-*tree* 算法，找到站点覆盖范围，生成站点的有界区域。

3. 提出基于有界区域内提取 *top-k* POIs 方法。本文提出了有界区域的多样性和比例性方法，用于提取有代表性的 POIs，每个 POI 的得分通过其语义和空间两个维度计算。

4. 提出基于用地特征计算站点相似性方法。利用每个站点具有代表性的 k 个 POIs 基于文本相似性和类型相似性计算站点间的相似性。

5. 基于进出站分时客流的站点相似度结果与基于用地特征的站点相似度结果实验比较。调整影响基于用地特征的站点相似度结果的系数值，分析两个实验结果的近似质量，验证提出方法的有效性。

6. 利用聚类算法，基于用地特征站点相似性将站点分类，并加入天气特征对客流量的影响将站点分类，分析两种分类结果，进一步验证提出用地特征站点相

似度计算方法的可靠性。

本文主要贡献是提出划分站点影响的有界区域的 RC-*tree* 算法和基于空间和语义有界区域内提取 *top-k* POIs 的多样性和比例性方法。

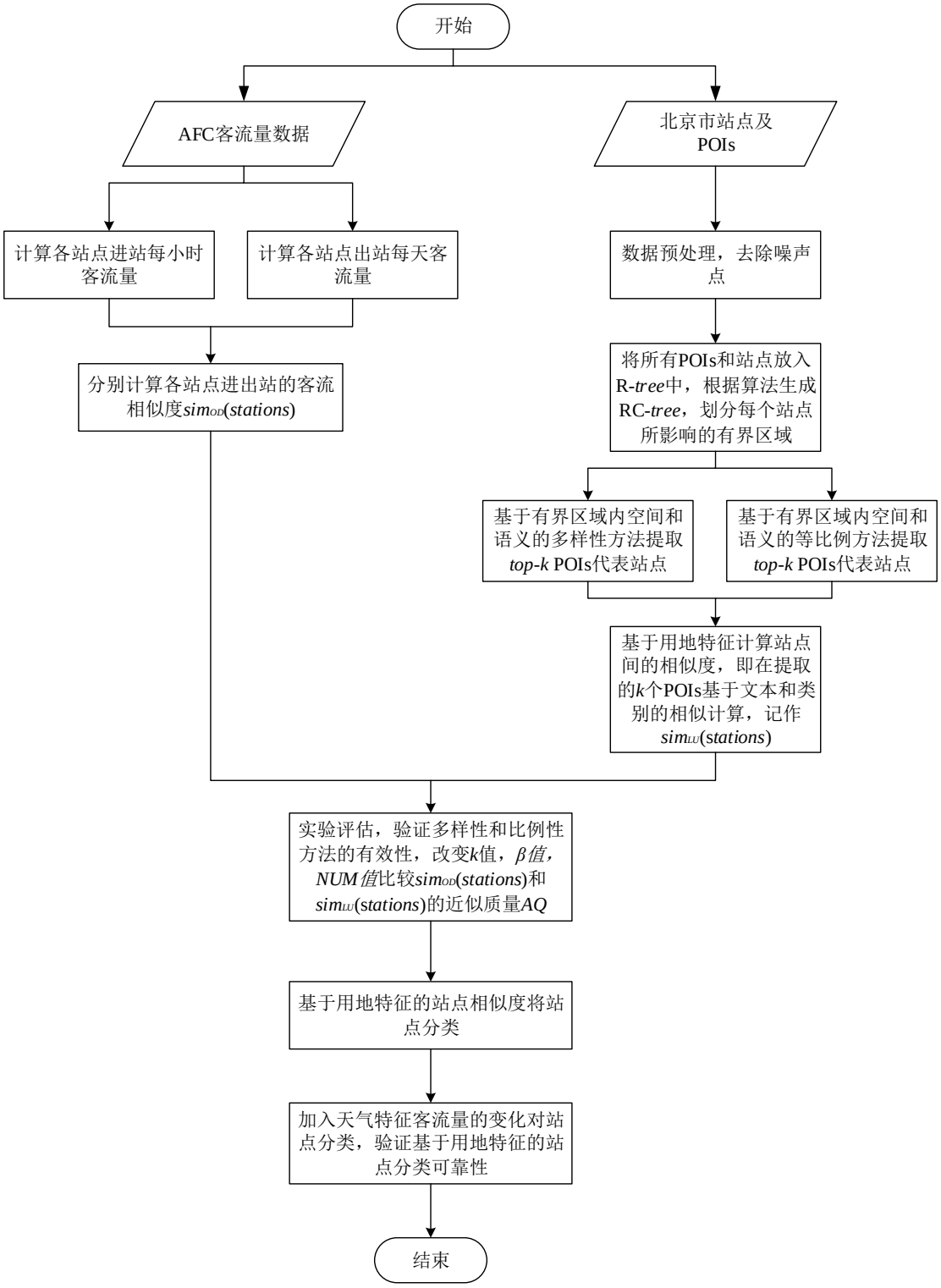


图 1-2 研究技术路线图

Fig. 1-2 The research technical route

1.4　本文组织结构

本文具体章节安排如下：

第 1 章：绪论。本章阐述了本文的研究背景和意义，介绍了轨道交通站点客流预测方法、站点周边用地特征、天气特征对轨道交通影响的研究现状，最后简要说明本文的研究内容和组织结构。

第 2 章：数据处理。本章首先说明了本文研究所用数据的来源，然后介绍了基于 AFC 系统，对进站和出站客流量的不同时间维度的统计，提出基于分时客流的站点相似性计算方法，最后简要介绍了基于用地特征的站点相似度的数据预处理方法。

第 3 章：划分有界区域的 RC-*tree* 算法。本章首先提出了 RC-*tree* 的定义和结构。然后，主要介绍了划分各站点影响 POIs 范围的具体算法实现。本章主要输出各站点所覆盖的有界区域。

第 4 章：基于用地特征的站点相似性计算方法。本章首先介绍了基于空间和语义的多样性和比例性方法提取 POIs，第 3 章输出的是各个站点有界区域的 POIs，且站点并未均在有界区域的中心，故提出了基于有界区域空间和语义的多样性和比例性方法提取 *top-k* POIs，之后根据提出的方法选取的 *top-k* POIs 基于文本相似性和类别相似性计算站点间的相似度。

第 5 章：基于 OD 分时客流的站点相似度与基于用地特征的站点相似度比较。本章重点比较站点分时客流的相似度和用地特征的相似度，并将相关实验进行详细介绍，改变影响提出方法结果的 k 和 β 值，对实验数据进行了充分整理、比较、分析，从而证明提出方法的有效性和优越性。

第 6 章：基于天气特征和用地特征的站点分类。本章研究分析加入天气特征后各站点客流的变化将站点分类；利用机器学习方法将站点周边的土地使用情况将站点分类，将两种分类结果交叉验证，进一步证明提出的方法的有效性。

第 2 章　数据处理

本章首先对本文研究的北京市相关数据来源做了必要的说明，其次对北京市轨道交通进出站客流的相似度计算方法和用地数据预处理做基本的介绍和分析。

2.1　基础数据来源

（1）进出站客流数据

从北京市交通委获悉，2008 年 6 月 9 日，北京市轨道交通在全网开通使用 AFC 系统（Automatic Fare Collection System），即城市轨道交通自动售检票系统。AFC 系统的应用，通过乘客的刷卡可以准确的记录乘客进站，出站信息，从而分析和掌握客流时、空分布规律，统计各条线路和各站点不同时间维度的客流量，为城市轨道交通运营组织和规划提供数据支持，对站点限流，站点接驳以及突发情况做合理规划方案和应急措施。本文所用的轨道交通站点进出站客流数据均来源于 AFC 系统。

本文使用的是 2013 年 6 月至 7 月北京市轨道交通 AFC 刷卡记录统计站点进出客流量记录。数据存放于 Oracle 数据库中，共有 483,614,919 条进出站记录，15,081,258 张交通卡，用以统计 227 个站点的 OD 客流量。

（2）北京市用地信息数据

POI 名为兴趣点，也可称为位置点，具有数据来源可靠、更新速度快、获取成本低等优势[38]。本研究借助 JavaScript 和 PHP 语言，爬取了高德地图 520,000 条数据，POI 包含的主要信息有名称，经纬度，类别等，具体信息如表 2-1 所示。

表 2-1 POI 属性表

Tab. 2-1 Attribute of POI

名称	经度	纬度	类别	地址	区域
国贸(地铁站)	116.461841	39.909104	交通设施服务	建外街道南郎家园西北方向	朝阳区
乐成国际 2 期	116.47333	39.89615	商务住宅	西大望路 SOHO 现代城南	朝阳区

POI 中的类别属性共有 20 种，包含汽车服务，汽车销售，汽车维修，摩托车服务，餐饮服务，购物服务，生活服务，体育休闲服务，医疗保健服务，住宿服务，风景名胜，商务住宅，政府机构及社会团体，科教文化服务，交通设施服

务，金融保险服务，公司企业，道路附属设施，地名地址信息和公共设施。若该 POI 有多个类别，用“|”分割，如“餐饮|酒店|电影院”。将数据以 JSON 的格式存储，这样以便实验中通过经纬度和类别查找数据。

2.2 OD 分时客流量的相似度计算方法

根据站点进出乘客的数量，可以直观地比较两个站点间的客流差异。但本文所比较的站点间客流量相似性，不仅仅是客流数量上的相似，重点是比较站点间客流分布趋势的相似性。因此，为了准确计算站点间的客流相似性，我们定义如下机制：利用日流量和每小时流量准确描述各站点的客流量，采用相似度计算方法计算站点间的客流相似度。

我们使用等效的时间描述了一段时间内站点的客流量情况。具体而言，对于进站客流量，以一小时为时间间隔，即将 5:00-24:00，划分为 19 段时间。对大多数站点而言，工作日的客流量和休息日的客流量会有很大区别，由于进站是按小时计算客流量，所以本文在进站统计平均客流量时没有包含周末和重大节日的客流量。站点 s 进站的客流量，即 $f_{s\bullet H}$，表示站点 s 在 H 至 H+1 一小时内的进站乘客数，整数 $H\in[5, 23]$，表示轨道交通系统的运营时间，例如 H=5 表示 5:00~6:00。采用欧氏距离的思想，两个站点 s_1 和 s_2 间的进站客流差异可以通过如下公式计算：

$$dif(s_1, s_2, Enter)=\sqrt{\sum_{H=5}^{23}(f_{s_1\bullet H}-f_{s_2\bullet H})^2} \tag{2-1}$$

为了多维度的验证本文提出的基于用地特征的站点相似度方法的可靠性，以天为时间间隔统计一周内出站客流量，即 $f_{s\bullet D}$，表示站点 s 在第 D 天的出站乘客数，整数 $D\in[1, 7]$表示北京市轨道交通系统的运营日，例如 D=5 表示星期五。则两个站点 s_1 和 s_2 间的出站乘客差异可计算为：

$$dif(s_1, s_2, Exit)=\sqrt{\sum_{D=1}^{7}(f_{s_1\bullet D}-f_{s_2\bullet D})^2} \tag{2-2}$$

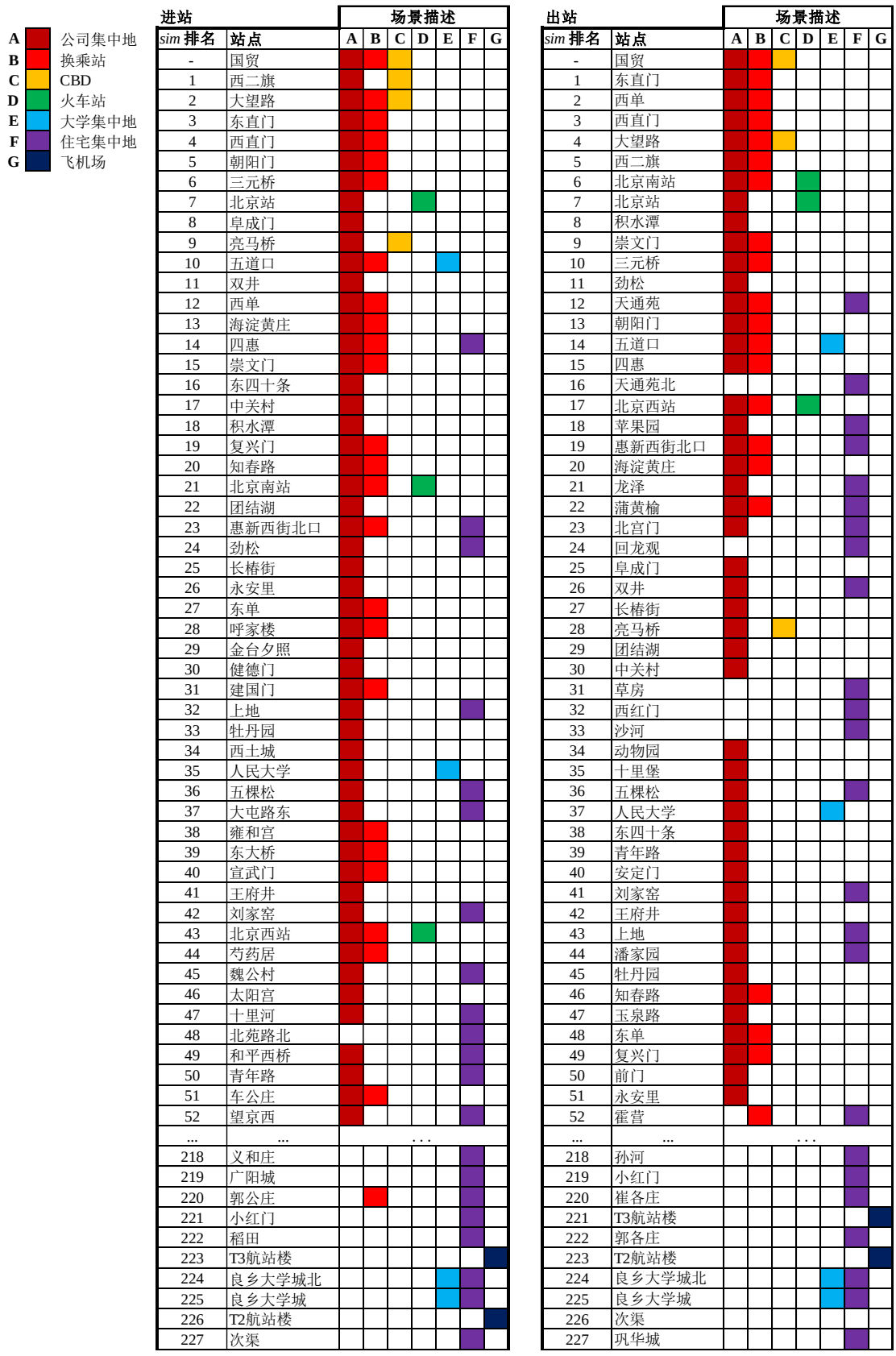
两个站点 s_1 和 s_2 的时间感知客流量的相似度可以根据公式(2-1)和(2-2)，即 $dif(s_1, s_2, E\bullet)$ 计算， $E\bullet$ 表示进站或者出站，其描述如下：

$$sim(s_1, s_2, E\bullet)=\frac{1}{dif(s_1, s_2, E\bullet)+1} \tag{2-3}$$

基于上述公式，可以计算出站点进出站客流量的相似度。本文以国贸站（见

图 1-1，标签 A）作为示例分析。根据北京市轨道交通系统的数据，图 2-1 列出了与国贸站进出站最相似和不同的站点。从图中可以看出客流量与国贸站具有较高相似度的进出站有西二旗站（见图 1-1，标签 C），大望路站（见图 1-1，标签 F）、东直门站（见图 1-1，标签 G）和西单站（见图 1-1，标签 H）。根据相应的场景描述，实验结果是合理的。国贸站位于北京中心商业区，周边分布众多商场和公司。西二旗站和大望路站所影响的区域内也有很多公司，因此它们在进站每小时的客流量分布与国贸站相似。

图 2-2 (a)中，国贸站位于北京市的 CBD，晚上进站客流量要远远高于早上的客流量，呈现晚高峰单峰状。根据图 2-1，与国贸站进站相似度最高的是西二旗站，从图 2-2 (a)和(b)可以看出这两个站点的进站客流量的曲线走势完全一样。劲松站（见图 1-1，标签为 B）周边用地不仅仅是住宅区，还有学校，商场，远郊公交站始发站，这使该站点进站客流量不仅集中于早高峰。而天通苑站（见图 1-1，标签为 E）周边均为住宅区，所以其进站客流量均集中于早上，客流量趋势随时间形成早高峰单峰状。图 2-3 介绍了一周内以天为时间间隔的站点出站客流量趋势。在西单站和国贸站周边有许多商场，这使这两站的周末出站客流量大于工作日；而西直门站（见图 1-1，标签 I）周边公司较多，次渠南站（见图 1-1，标签 J）是住宅集中地，所以这两站工作日的出站客流量较多，与国贸站和西单站不同。

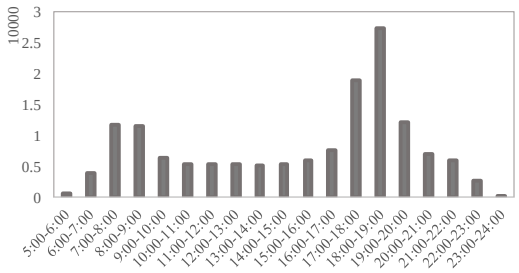


进站		场景描述						
sim 排名	站点	A	B	C	D	E	F	G
-	国贸	■	■	■				
1	西二旗	■		■				
2	大望路	■	■	■				
3	东直门	■	■					
4	西直门	■	■					
5	朝阳门	■	■					
6	三元桥	■	■					
7	北京站	■			■			
8	阜成门	■						
9	亮马桥	■		■				
10	五道口	■	■			■		
11	双井	■						
12	西单	■	■					
13	海淀黄庄	■	■					
14	四惠	■	■				■	
15	崇文门	■	■					
16	东四十条	■						
17	中关村	■						
18	积水潭	■						
19	复兴门	■	■					
20	知春路	■	■					
21	北京南站	■	■		■			
22	团结湖	■						
23	惠新西街北口	■	■				■	
24	劲松	■					■	
25	长椿街	■						
26	永安里	■						
27	东单	■	■					
28	呼家楼	■	■					
29	金台夕照	■						
30	健德门	■						
31	建国门	■	■					
32	上地	■					■	
33	牡丹园	■						
34	西土城	■						
35	人民大学	■				■		
36	五棵松	■					■	
37	大屯路东	■					■	
38	雍和宫	■	■					
39	东大桥	■	■					
40	宣武门	■	■					
41	王府井	■						
42	刘家窑	■					■	
43	北京西站	■	■		■			
44	芍药居	■	■					
45	魏公村	■					■	
46	太阳宫	■						
47	十里河	■					■	
48	北苑路北						■	
49	和平西桥	■					■	
50	青年路	■					■	
51	车公庄	■	■					
52	望京西	■					■	
...	...	...						
218	义和庄						■	
219	广阳城						■	
220	郭公庄		■				■	
221	小红门						■	
222	稻田						■	
223	T3航站楼							■
224	良乡大学城北					■	■	
225	良乡大学城					■	■	
226	T2航站楼							■
227	次渠						■	

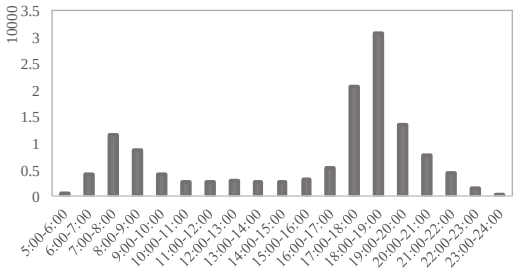
出站		场景描述						
sim 排名	站点	A	B	C	D	E	F	G
-	国贸	■	■	■				
1	东直门	■	■					
2	西单	■	■					
3	西直门	■	■					
4	大望路	■	■	■				
5	西二旗	■	■					
6	北京南站	■	■		■			
7	北京站	■			■			
8	积水潭	■						
9	崇文门	■	■					
10	三元桥	■	■					
11	劲松	■						
12	天通苑	■	■				■	
13	朝阳门	■	■					
14	五道口	■	■			■		
15	四惠	■	■					
16	天通苑北						■	
17	北京西站	■	■		■			
18	苹果园	■					■	
19	惠新西街北口	■	■				■	
20	海淀黄庄	■	■					
21	龙泽	■					■	
22	蒲黄榆	■	■				■	
23	北宫门	■					■	
24	回龙观						■	
25	阜成门	■						
26	双井	■					■	
27	长椿街	■						
28	亮马桥	■		■				
29	团结湖	■						
30	中关村	■						
31	草房						■	
32	西红门						■	
33	沙河						■	
34	动物园	■						
35	十里堡	■						
36	五棵松	■					■	
37	人民大学	■				■		
38	东四十条	■						
39	青年路	■						
40	安定门	■						
41	刘家窑	■					■	
42	王府井	■						
43	上地	■					■	
44	潘家园	■					■	
45	牡丹园	■						
46	知春路	■	■					
47	玉泉路	■						
48	东单	■	■					
49	复兴门	■	■					
50	前门	■						
51	永安里	■						
52	霍营		■				■	
...	...	...						
218	孙河						■	
219	小红门						■	
220	崔各庄						■	
221	T3航站楼							■
222	郭各庄						■	
223	T2航站楼							■
224	良乡大学城北					■	■	
225	良乡大学城					■	■	
226	次渠							
227	巩华城						■	

图 2-1 国贸站在等量时间的客流量与北京市轨道站点的相似度

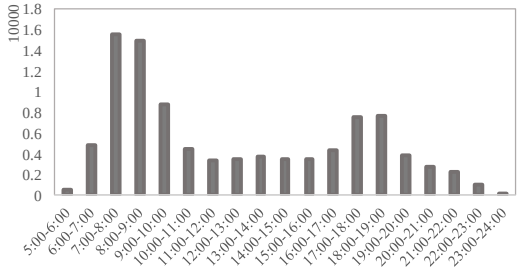
Fig. 2-1 The transit stations in Beijing most similar and dissimilar to *guomao* station calculated from passenger flow over equivalent time



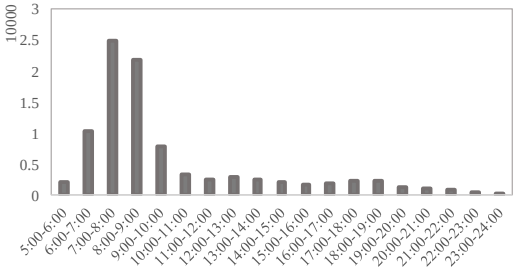
(a) 国贸站



(b) 西二旗站



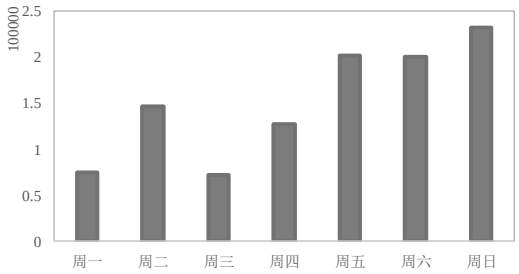
(c) 劲松站



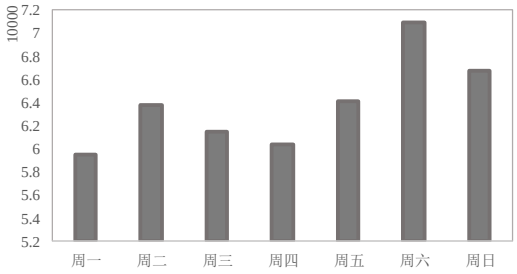
(d) 天通苑站

图 2-2 进站每小时客流量

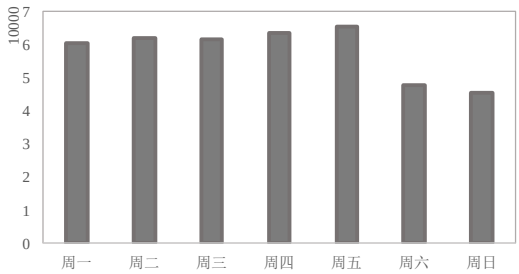
Fig. 2-2 Passenger flow over equivalent time (hourly) of Entering



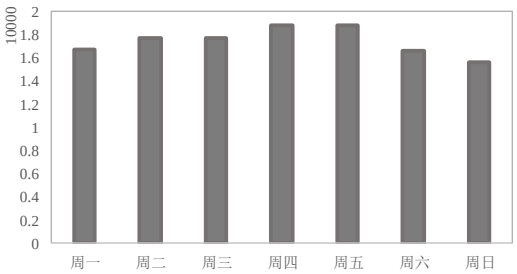
(a) 国贸站



(b) 西单站



(c) 西直门站



(d) 次渠南站

图 2-3 出站每日客流量

Fig. 2-3 Passenger flow over equivalent time (daily) of Exiting

2.3 北京市 POI 数据预处理

本文使用的北京市 POIs 数据是通过抓取高德地图所获取的，直接保存成 JSON 数据格式，但这些数据是不完整、不一致和有噪声的，存在一些“脏”数据，本小节主要清洗 POIs 数据。

数据清洗的主要方法包括：缺失数据删除法、缺失值填补法、3σ 检测法、基于分布的异常值检测法和基于密度聚类法。下面简单介绍各方法具体实现步骤。

1） 缺失值删除方法

删除数据是处理缺失数据最基本和最传统的技术，主要包括列表/样本删除和成对删除两种方法：

列表/样本删除。直接删除存在缺失值的列表/样本数据，即所有的缺失样本无论在一个变量上还是在多个变量上均被排除。该方法的优点是保证了剩余列表/样本的数据集的完整性，但会导致完整数据集样本量少。虽然这是处理缺失数据的默认方法，但在大多数情况下，列表/样本删除的缺点远大于其优点。

成对删除。仅在缺少测试特定假设所需的特定数据点时才会删除，如果数据集中的其他位置缺少数据，则在统计测试中使用现有值。由于成对删除使用的是观察到的所有信息，因此与列表/样本删除相比保留了更多的信息。但这种方法的缺点是模型参数的模糊定义会导致估计的标准误差和测试统计数据出现偏差，且产生一个非正定的互相关矩阵，这可能会影响进一步分析。

2） 缺失值填补方法

为了保证数据的完整性，用填补值替换缺失数据。在不删除具有任何缺失值的样本情况下，利用其他可利用信息预测缺失值，替换丢失的数据来保留所有样本。

①均值/中位数/众数填补法。该方法用于数值型变量，缺失值可利用均值、中位数和众数填补，变量服从正态分布，利用均值填补；如果变量存在很大异常值，则用中位数填补。

②热卡填补法。最简单的单值填补法之一，其思想是：在数据集中找到与缺失值最相似的样本，用该样本的数据值填充缺失值。不同问题，判定相似的标准也不同。最常见的方法是利用相关系数矩阵判定缺失值最相关的样本。

③回归填补法。回归是衡量一个变量和其他变量的关系，将缺失值作为因变量，相关变量作为自变量，利用回归的方法将自变量和因变量拟合，再将预测值

填补缺失处。

④类似响应模式填补法（SRPI）。该方法的思想是：在一组匹配数据量中利用具有相似分数用户估算缺失值。该方法与热卡填补法相似，但需要用户指定一组匹配数据量 Z 和一组不完整变量 Y，标准化 Z 中的变量，计算每个变量 Y 与每个变量 Z 的距离，最小距离值的变量 Y 替代缺失值。

⑤多重填补法（M-试探法）。利用一组合理值（或范围）替代缺失值，例如参数的有效置信区间。多重填补算法流程分为三个步骤：第一步填补阶段，丢失的数据被填充 M 次，生成 M 个完整的数据集；第二步分析阶段，使用标准程序分析 M 个完整数据集；第三步汇集阶段，组合 M 个完整数据集的结果推断最后结果。

⑥基于贝叶斯填补法。分别将缺失的属性作为预测项，再根据最简单的贝叶斯方法，对这个预测项进行预测[39]。

3） 3σ 检测方法

3σ 原理思想来源于切比雪夫不等式，可简单描述为：若数据服从正态分布，则异常值被定义为结果值中与平均值偏差超过三倍标准差的值。在统计学中，平均值小于一个标准差、二个标准差、三个标准差以内的百分比分别为 68.27%、95.45%及 99.73%[40]。

4） 基于分布的异常值检测方法

Grubbs 检验。Grubbs 检验基于正态假设，应用该方法前首先要验证数据是否满足正态分布。该方法一次检测一个异常值，计算数据集的平均值 avg 和标准差 s，计算任一点 x_i 统计量 $g_i=|x_i - avg|/s$ 与临界表的大小，若 g_i 大于临界表的大小，表明 x_i 是异常点，从数据集中删除该异常点。

5） 基于密度聚类方法

DBSCAN 是最常见的聚类算法之一，在一组数据点中，基于距离（通常用 Euclidean distance）和形成密集区域的最小点数将接近点分组，将低密度区域中的点标记为异常值。针对难以手动查找关联和结构的数据点，可以使用 DBSCAN 查找相关模式和预测趋势。DBSCAN 算法需要两个参数：

eps：两点之间的最小距离。只有当两个数据点之间的距离小于或者等于 *eps* 时，才会将这些点视为相邻。

*min*Points：形成密集区域的最小点数。 例如，如果将 *min*Points 参数设置为 5，那么需要至少 5 个点才能形成密集区域。若区域内的点数小于 5，则这些点为噪音。

本文将 POIs 数据去噪步骤如下：

1）对于经纬、维度缺失的数据，根据该数据的名称在 GPSspg 网站（http://www.gpsspg.com/）查询高德地图经纬度，填充缺失值；

2）对于数据类别缺失的，利用热卡填补法，在数据中找到一个与其最相似的对象，用这个相似对象的类别来进行填充；

3）20 种类别中，个别类别，如道路附属设施，地名地址信息和公共设施等，在提取 *top-k* 有代表的 POIs 时，是无用的，并且是噪声，将属于这些类别的数据删除。

2.4 本章小结

本章首先对本文进行后续研究所使用的基本数据来源做了说明，并结合下文进行进出分时客流量的相似性分析，基于用地特征的站点相似度研究的数据需求提出计算不同时间段的进出站客流量的站点间的相似度的方法，以及对北京市 POIs 数据进行预处理。

第 3 章　划分有界区域的 RC-*tree* 算法

POI 包含空间（经纬度）和语义（名称，类别）等信息，为划分站点所覆盖范围，本文采用 R-*tree* 结构保存 POIs，本章主要介绍划分每个站点所影响区域的 RC-*tree* 算法。

3.1　RC-*tree* 定义

3.1.1　R-*tree* 结构

随着空间数据的不断增加，对于空间数据的存储和处理的需求也急剧增加，但空间数据不仅包含文本、图像等信息，还有空间位置信息。迄今为止，专家和学者们提出了许多空间索引结构，例如网络索引、四叉树，R-*tree* 等。本文主要通过 R-*tree* 结构保存 POIs。

Guttman 提出了 R-*tree* 的概念[41]，R-*tree* 是一种平衡树结构，广泛用于索引多维数据，例如地理数据，坐标等。R-*tree* 由根节点、中间节点和叶节点组成，所有的空间数据，例如经纬度，坐标等，保存在叶节点上，生成一个个最小的矩形；根节点和中间节点存储空间范围，包含多个叶节点生成较大的矩形，并且 R-*tree* 能够保证对空间数据的搜索只需要访问很小一部分的节点。如图 3-1 所示，给出了 R-*tree* 的空间索引结构。由 R-*tree* 的结构可知，其有以下特性：

1）R-*tree* 中所有的叶节点均在树的高度的最后一层；

2）如果根节点不是叶节点，则根节点至少有两个子节点；

3）叶节点存储的数据数量 N，满足 m<=N<=M，其中存储的数据最小数量是 m，2<=m<=M/2；

4）中间节点包含子节点数 W，满足 m<=W<=M。

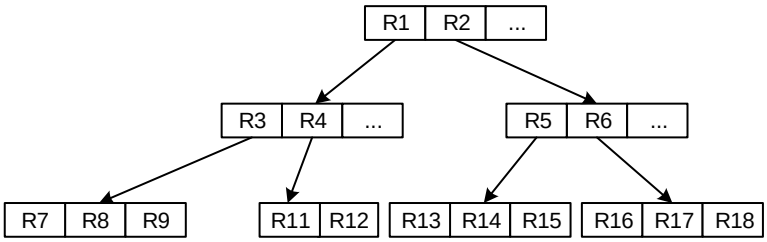


图 3-1 R-*tree* 的结构图

Fig. 3-1 The structure of R-*tree*

3.1.2 RC-*tree* 思想

定义 1：RC-*tree*。给定一组 POIs 和站点集合，RC-*tree* 是一种 R-*tree* 形式的数据结构，存储城市中的所有 POIs 和站点，所有节点均被涂有颜色。RC-*tree* 中每种颜色（以区分彼此）代表一个非重叠区域，并且每个区域内最多包含一个站点。

为了进一步研究各站点的 POIs 对客流的影响，提出了一种新的方法划分每个站点所影响覆盖的 POIs。分别将站点记作 $station=\{s_1, s_2, \ldots, s_l\}$，POIs 记作 $pois=\{p_1, p_2, \ldots, p_m\}$。更具体地内容为：首先将所有站点和 POIs 储存在 R-*tree* 的叶节点中；然后根据叶节点中 POIs 的距离和节点关系，提出了 POIs 划分算法（详见 3.2 章）为 R-*tree* 中的所有节点着色。着色的 R-*tree* 被记作 RC-*tree*=$\{R_1(c_1), R_2(c_2), \ldots, R_n(c_n)\}$，其中 c_i 是节点 R_i 对应的颜色。

3.2 算法实现

由于 R-*tree* 将对象空间按范围划分，当进行一个高维空间查询时，只需要遍历少数几个叶子节点所包含的指针，查看这些指针指向的数据是否满足要求即可。这正是满足储存空间 POIs 的要求，所以将所有 POIs 和站点生成一棵 R-*tree*。但生成的 R-*tree* 要满足每个叶子节点中至多包含一个站点。为提取每个站点所影响、覆盖的区域的 POIs，生成的 R-*tree* 的叶节点维度应为不能包含两个站点的最小密度。即每个站点在一个叶节点中，有的叶节点可以不包含站点。

已经得到了一棵包含所有站点及站点周边 POIs 的 R-*tree* 索引，接下来要为每个站点划分其影响范围的 POIs。生成的 R-*tree* 中的叶节点可能包含一个或零个站点，我们将所有包含站点的叶节点涂为不同颜色。

在 R-*tree* 生成之后，可以获得其叶节点存储的所有站点和 POIs，然后通过对 POIs 着色划分为某个站点影响的范围，因此 RC-*tree* 包含各站点的颜色。

首先，包含站点的叶节点被涂为不同颜色（以此区分），然后，对于仅包含 POIs 的叶节点（即 R_i）使用如下三个规则来处理不同的情况，以便将 POIs 划分到某个站点区域内：

1）如果 R_i 只有一兄弟节点包含站点，则 R_i 被涂为该兄弟节点的颜色。

2）如果 R_i 有多个兄弟节点包含站点，计算 R_i 中的 POIs 与其兄弟节点中站点的距离，R_i 被涂为距离最近的站点的颜色。

3）如果 R_i 所有的兄弟节点均不包含站点，查询 R_i 的祖先的兄弟分支，计算分支中包含站点距 R_i 的距离，R_i 被涂为距离最近的包含站点的叶节点的颜色。

最后父节点被涂为其孩子节点的颜色，将整个 R-*tree* 填满颜色。

算法 3.1 RC-*tree* 算法划分站点覆盖范围

Input: *stations*: all stations in Beijing; *pois*: all POIs around stations
Output: RC-*tree*:a full-color R-*tree*

```
1    R-tree rt <- generateRTree(pois, stations)
2    while rt.leaf.contain(stations) > 1 do
3        rt <- splitNode(rt.leaf);
4    end
5    RC_tree rct <- rt
6    while rct.leaf.contain(stations) == 1 do
7        paint(rct.leaf, color(random(c_i)));
8    end
9    while rct.leaf.contain(stations) == 0 do
10       if rct.leaf.silbings.contain(stations) == 1 then
11           paint(rct.leaf, color(rct.leaf.silbing));
12       end
13       else if rct.leaf.silbings.contain(stations) > 1 then
14           s_t <- ∀ s_j ∈ rct.leaf.silbings;
15           s_min <- min(distance(∀ s_j ∈ s_t; ∀ p_i ∈ rct.leaf);
16           paint(rct.leaf; color(s_min ∈ rct.leaf.silbing);
17       end
18       else if rct.leaf.silbings.contain(stations) == 0 then
19           leaves <- search(rct.leaf.ancestors.branches);
20           le <- ∀ leaf_i ∈ leaves.contain(stations) = 1;
21           leaf_closest <- closest(relationship(rct.leaf, leaf_i ∈ le));
22           paint(rct.leaf; color(leaf_closest));
23       end
24   end
25   while hasColor(rct.node) == 0 do
26       paint(rct.node; color(rct.node.children));
27   end
28   return rct;
```

算法 3-1 介绍了划分有界区域的 RC-*tree* 算法的整个过程，图 3-2 是算法的流程图。算法的输入是北京市的所有站点和 POIs，输出是生成 RC-*tree*。在算法的开始，生成 R-*tree* 存储所有站点和 POIs，如若 R-*tree* 的叶节点中包含站点个数大于 1 时要将该节点分裂为 2 个节点（Lines 1-3）；然后将生成的 R-*tree* 转化成 RC-*tree*，但此时的 RC-*tree* 没有被着色（Line 5）；所有包含站点的叶节点均

被涂为不同的颜色（Lines 6-7）；对于不包含站点的叶节点（即 R_i），如果 R_i 只有一个兄弟节点包含站点，则将 R_i 涂为该兄弟节点的颜色（Lines 10-11）；如果 R_i 有多个兄弟节点包含站点，计算 R_i 中 POIs 到站点的距离，则 R_i 的颜色为距离最近站点的颜色（Lines 13-16）；如果 R_i 的兄弟节点都不包含站点，R_i 的颜色为距离最近的父节点包含站点的兄弟节点的子节点颜色（Lines 18-22）；最后，父节点的颜色为其子节点的所有颜色（Lines 25-26）；返回涂满颜色的 RC-*tree*（Line 28）。

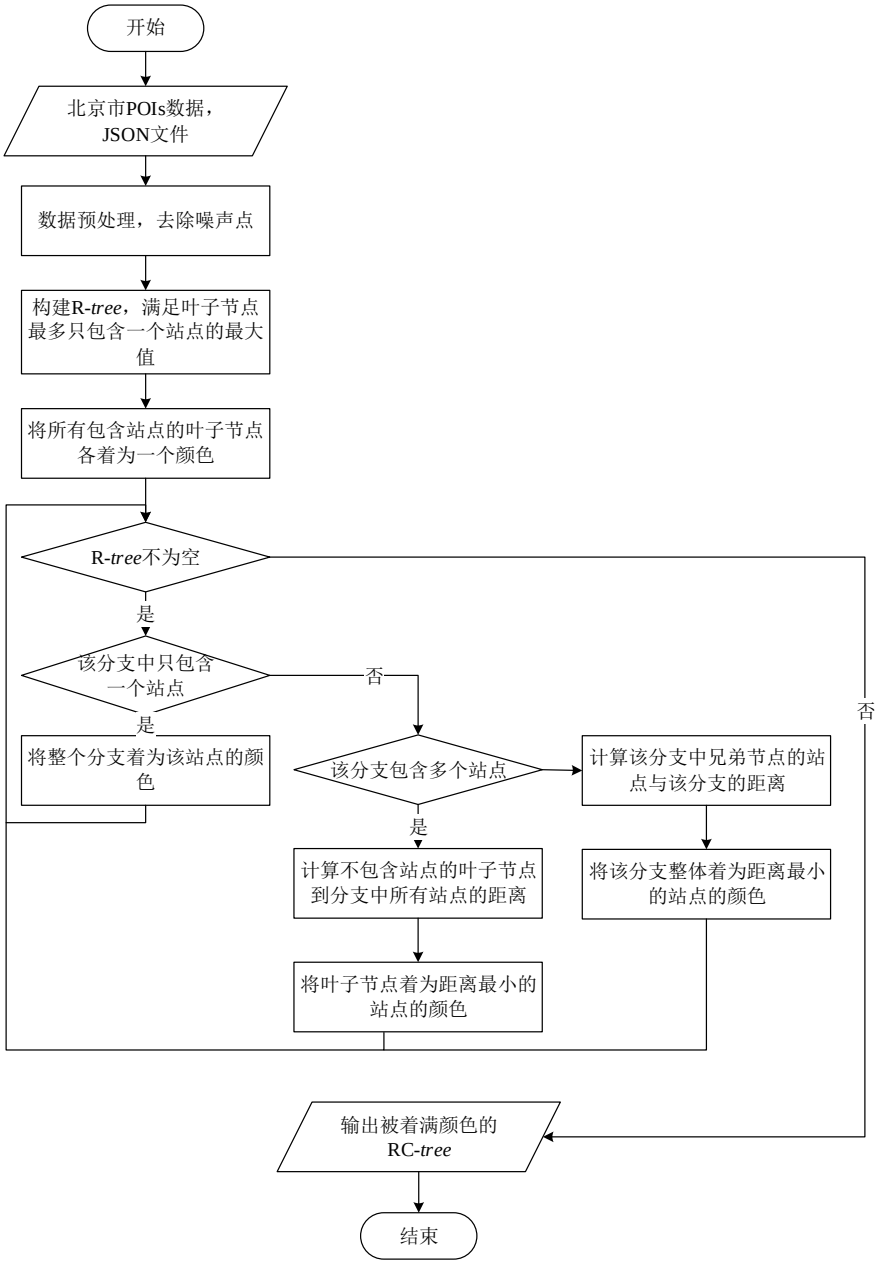


图 3-2 算法 3-1 流程图

Fig. 3-2 The flowchart of Algorithm 3-1

为了更加清晰的介绍提出的算法，在此通过例子来详细介绍。如图 3-3 所示，生成 R-*tree*，保证 R-*tree* 的任一叶节点中最多包含一个站点。

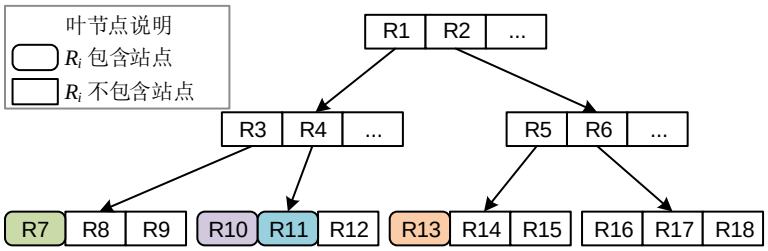


图 3-3 R-tree 例子

Fig. 3-3 An R-tree example

将包含站点节点的 R_7, R_{10}, R_{11}, R_{13} 分别涂为绿色，紫色，蓝色和橙色。如果 R_i 只有一个兄弟节点包含站点，则 R_i 被涂为该兄弟节点的颜色。图 3-3 中因只有 R_7 包含站点，则 R_8 和 R_9 被涂为绿色，同理 R_{14} 和 R_{15} 被涂为 R_{13} 的颜色（橙色）。如果 R_i 有多个兄弟节点包含站点，则计算 R_i 的 POIs 与兄弟节点中的站点之间的距离，并且 R_i 被涂为包含最近站点的兄弟节点的颜色。R_{10} 和 R_{11} 均包含站点，则要计算 R_{12} 与 R_{10} 的站点和 R_{11} 的站点的距离，$dis(R_{10}, R_{11})$和 $dis(R_{10}, R_{12})$，假设 $dis(R_{10}, R_{11}) < dis(R_{10}, R_{12})$，则 R_{12} 被涂为蓝色。如果 R_i 的所有节点都不包含站点，则通过 R_i 的祖先搜索其他分支的叶节点，将 R_i 涂为与此最密切关系的叶节点的颜色。R_{16}、R_{17} 和 R_{18} 这一分支均不包含站点，则搜索 R_6 兄弟的分支，R_5 中只有 R_{13} 包含站点，则 R_6 整个分支均被涂为橙色。根据步骤 3 的方法，逐层涂色，将整个 R-*tree* 涂满颜色，生成 RC-*tree*。得到的 RC-*tree* 如图 3-4 所示。

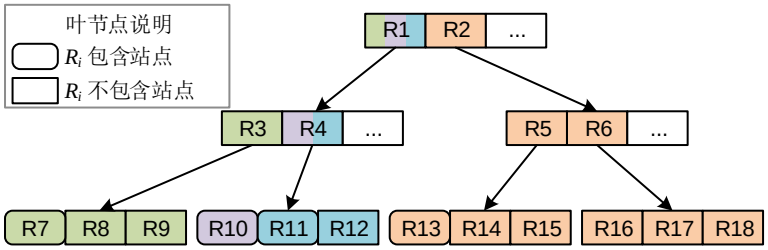
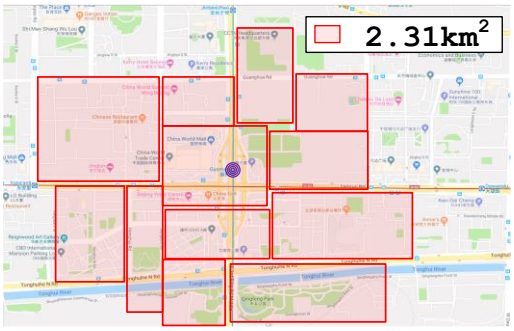


图 3-4 根据图 3-2 生成的 RC-tree 例子

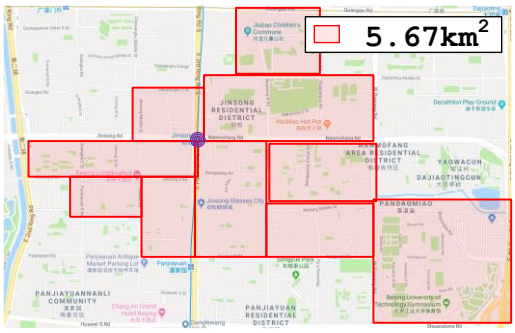
Fig. 3-4 The RC-tree example according to Fig. 3-3

根据算法 3-1、图 3-3 和 3-4，可以看出提出的划分 POIs 的 RC-*tree* 算法可以将所有 POIs 划分给不同站点，然后圈出相同颜色的 POIs 的边界，即站点的有界区域，$\{R_7, R_8, R_9\}$，$\{R_{10}\}$，$\{R_{11}, R_{12}\}$，$\{R_{13}, \ldots, R_{18}\}$。同时，在图 3-5 中提供了 4 个站点的有界区域内，图中红色方块代表站点覆盖的区域，紫色圆圈代表站点。

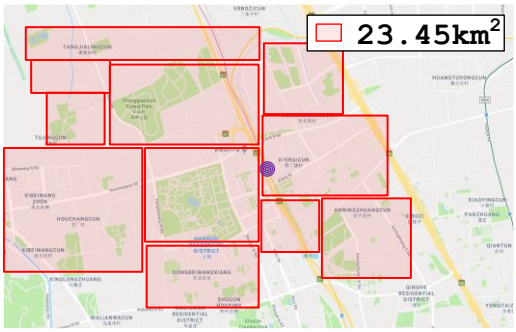
可以看出，每个站点具有不同的区域分布和 POIs 的稀疏性，每个站点的覆盖区域也十分明显。



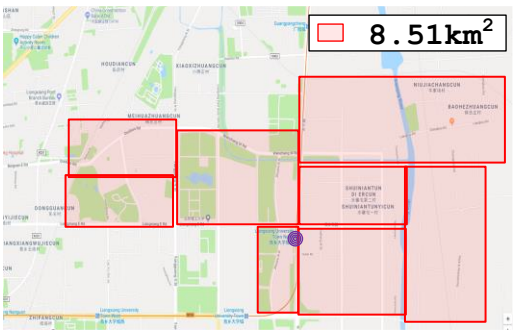
(a) 国贸站的有界区域



(b) 劲松站的有界区域



(c) 西二旗站的有界区域



(d) 良乡大学城北站的有界区域

图 3-5 图 1-1 中标签 A-D 的站点的有界区域

Fig. 3-5 the bounded areas of selected stations are the station labeled A-D respectively in Fig. 1-1

3.3 本章小结

本章介绍了 R-*tree* 的结构及特点，基于 R-*tree* 提出了新的结构 RC-*tree*。为了找到各站点影响的有界区域，提出了划分有界区域的 RC-*tree* 算法。该算法通过节点被涂为的颜色差异确定边界，为第 4 章提取 *top-k* 有代表的 POIs 提供数据支持。

第4章　基于用地特征的站点相似性计算方法

定义 2：*top-k* POIs。根据定义 1，给定 RC-*tree* 中任意独立的 R_i 和一整数 k，*top-k* POIs 表示 R_i 对应站点的有界区域内最具有代表性的 k 个 POIs。

RC-*tree* 中的每个节点都包含一组站点和 POIs，可以描述为 $R_i \subset \{s_i \cup \text{POIs}\}$。*top-k* POIs 的提取方法是基于空间和语义计算 POIs 的动态分数，站点 s_i 提取的 *top-k* POIs 保存在数组 kPOI(s_i)中。

划分有界区域的 RC-*tree* 算法，所有的 POIs 被划分成不同的站点区域（即被涂为站点颜色）。通过定位 POIs 的位置可以找到每个站点的有界区域。然而，如果一个站点边界区域内的所有 POIs 直接用于计算站点的相似度，那么所有站点具有很高的相似度。这是因为在站点的有界区域中，存在数百个 POIs，并且大多数 POIs 是相同或无意义的，例如公共厕所、快餐店等。为了提取站点有代表的 POIs，提出基于空间和语义有界区域的提取 *top-k* POIs 多样性方法和比例性方法。

4.1　多样性方法

多样性，一常见生态学名词，近期，查询结果的多样化引起了广泛的关注，多样性是通过平衡与查询点的不相似性来提高返回结果质量的方法。本文认为提取 k 个 POIs 时应该多样化，避免一些重要的节点被忽略。故提出基于空间和语义的多样性方法选取分值最高的 k 个 POIs。

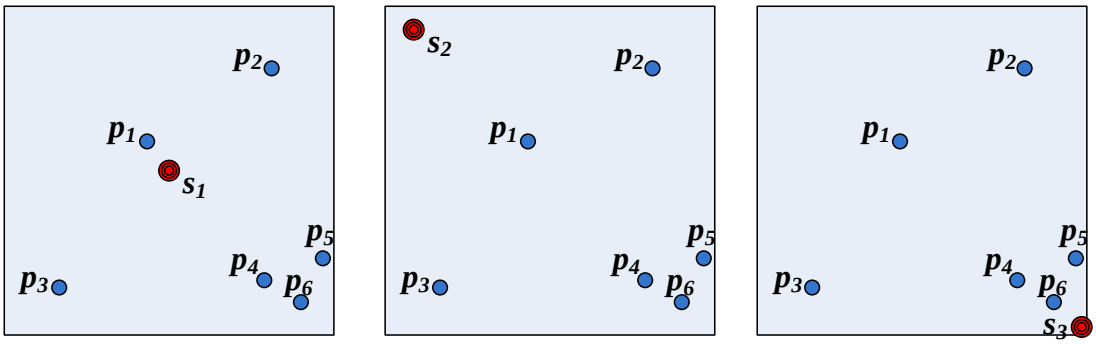
4.1.1　基于空间的多样性方法

空间多样性是指通过已经选择 POIs 的方向对候选 POIs 中同一方向的点进行削弱，即空间多样性为分布多样性，多个方位，故根据三角形特性计算空间分布多样性的公式如下：

$$Div_{spa}(p_i) = \frac{1}{l} \sum_{p_j \in eQue} \frac{dis(p_j, p_i)}{dis(p_i, s) + dis(p_j, s)} \tag{4-1}$$

其中，$l = |eQue|$ 表示已经提取了 POIs 的个数，p_i 是候选 POIs 队列中任意的 POI，p_j 表示已提取队列中的任意 POI；s 表示边界区域内的站点，$dis(\cdot,\cdot)$ 表示两个位置点的欧氏距离。为了计算简单和方便结合空间和语义的比例，其值域为(0, 1]。

为了清晰地解释空间 POIs 分布多样性的计算过程，图 4-1 列举了在同一边界范围内站点与周边 POIs 位置情况。其中，矩形框表示边界区域，三个圆圈表示站点的位置（s_1, s_2, s_3），点（$p_1, \ldots, p_6$）表示 POIs 的位置。



(a)站点位于中心　　(b) 站点远离 POIs　　(c) 站点靠近 POIs

位置坐标<x, y>

$s_1 :< 10, 10 >$　$s_2 :< 1, 19 >$　$s_3 :< 20, 0 >$

$p_1 :< 9, 12 >$　$p_2 :< 16, 17 >$　$p_3 :< 4, 3 >$

$p_4 :< 16, 4 >$　$p_5 :< 19, 5 >$　$p_6 :< 18, 2 >$

图 4-1 站点周围 POIs 位置分布的示例

Fig. 4-1 Example of POIs position distributions all around stations

如图 4-1 (a)所示，s_1 与点 p_i（i=1, 2, 3, 4, 5, 6）的距离分别为 $dis(s_1, p_1) = 2.236$，$dis(s_1, p_2) = 9.2195$，$dis(s_1, p_3) = 9.2195$，$dis(s_1, p_4) = 8.485$，$dis(s_1, p_5) = 10.296$ 和 $dis(s_1, p_6) = 11.314$，首先选取距离最近的点 p_1 作为 *top*-1 POI，且将 p_1 加入队列 *eQue* 中，此时 $l = 1$；其次，根据公式(4-1)（即 $\frac{dis(p_1, p_i)}{dis(p_i, s_1) + dis(p_1, s_1)}$，其中 i= 2, 3, 4, 5, 6）计算，从 p_2， p_3， p_4， p_5 和 p_6 中选取得分最高者作为第二个 POI，则将 p_6 加入队列 *eQue* 中，$l = 2$（即 $eQue = \{p_1, p_6\}$）；然后，从 p_2， p_3， p_4，和 p_5 中选取得分最高的点作为第三个 POI，此时 i= 2, 3, 4, 5，计算公式为 $\frac{1}{2} * (\frac{dis(p_1, p_i)}{dis(p_i, s_1) + dis(p_1, s_1)} + \frac{dis(p_6, p_i)}{dis(p_i, s_1) + dis(p_6, s_1)})$，则将 p_3 加入队列 *eQue* 中。基于上述计算过程，可得到第四-六顺序的 POI，则图 4-1 (a)中提取 POIs 的排列顺序为 $p_1 -> p_6 -> p_3 -> p_2 -> p_4 -> p_5$。若令站点 s_1 为向量起点，p_1 和 p_6 分别为两个向量的终点（即 $v_1 = [s_1, p_1]$ 和 $v_2 = [s_1, p_6]$），我们发现向量 v_1 和 v_2 的方向完全相反，且 s_1， p_1 和 p_6 趋于在同一条直线上。根据公式(4-1)计算图 4-1 (a)的空

间多样性排序结果符合我们的预期。

4.1.2　基于语义的多样性方法

最大相关边界法[42]的提出奠定了多样性的发展，其核心思想是通过迭代的方式将候选集合加到结果集合中，在选择候选集中元素时，基于候选集中与已选结果集中的内容相似度较低的元素的原则加入结果集，该方法降低了结果集的数据冗余。衡量多样性的标准目标是最大化节点之间的差异总和。而相似性和多样性本身正交，则 POIs 的语义多样性（记作 Div_{sem}）可根据其语义相似度（记作 Sim_{sem}）计算，公式为：

$$Div_{sem}(p_i)=\frac{1}{l}\sum_{p_j\in eQue}(1-Sim_{sem}(p_j,p_i)) \tag{4-2}$$

其中，$Sim_{sem}(\cdot,\cdot)$ 是 p_j和p_i 的语义相似性。则两个 POIs 的相似度可根据 *Jaccard* 公式计算，定义为：

$$Sim_{sem}=\frac{p_j\bigcap p_i}{p_j\bigcup p_i} \tag{4-3}$$

在计算 POIs 的语义和空间多样性之后，通过添加系数 β 来组合两个多样性值。系数 β 可以调整空间和语义所占比例，以此影响 POI 的最终多样性值。POI 的最终多样性值计算如下，基于 POIs 的最终多样性的值，可以提取各站点 *top-k* 有代表的 POIs。

$$Div(p_i)=\beta*Div_{sem}(p_i)+(1-\beta)*Div_{spa}(p_i) \tag{4-4}$$

4.2　比例性方法

比例性是根据各类别的数量所占总数的比例，选取各类别对应比例的数量。比例特性和多样性正好是相反的，用比例性方法提取 *top-k* POIs 也是基于地理空间和语义文本两方面。

在语义文本方面，本文根据 POIs 的类别标签数量所占总数的比例，进行权值削弱。由于 *k* 的值可能较小，为了比例性覆盖不同的集合，本文使用增量选择策略，即在每次迭代中，可以追加一个不同的节点（有利于某种程度的循环选择）。为此，引用 G.Fakas[43]使用比例性公式(4-5)如下：

$$Pro_{sem}(p_i)=\frac{sum(type(p_i))}{\alpha*ext(type(p_i)+1)} \tag{4-5}$$

其中，p_i 表示有界区域内的任意候选 POIs，$type(p_i)$表示点 p_i 的类别，$sum(type(p_i))$表示任意类别 p_i 的数量，$ext(type(p_i))$是点 p_i 的类别在结果集中的数量（即，该类别已经被添加 *ext*-1 次），α 是一个常数，可以调整比例（一般 α=2，经验值[44,45]）。如果给定 *ext* 值，等价于给定了 $ext(type(p_i))$，将相应的值表示为 $Pro_{sem}(p_i)[ext]$。这个公式的基本原理是满足任意一个 POI 的类别被添加到结果集队列时，它的 $Pro_{sem}(p_i)[ext]$的结果会明显削弱，从而依次选择其他 POIs 的类别。

对于空间比例性，基于到站点不同方向的 POIs 的比例来提取 POIs，在每个方向上提取相应数量的 POIs。为了划分不同的方向，相同方向的概念描述如下：以站点为中心，将有界区域划分为多个扇区（或大多数情况下为象限），并且同一扇区中的 POIs 被认为是在同一方向，例如图 4-1 (a) 中，p_4，p_5 和 p_6 是同一方向。POIs 的空间比例性计算公式如下：

$$Pro_{spa}(p_i)=\frac{sum(dir(p_i))}{\alpha * ext(dir(p_i)+1)} \tag{4-6}$$

类似多样性方法提取 POIs 时的计算（即公式(4-2)），最终，比例性方法提取 POIs 的空间和语义的比例值也是通过添加系数 β 来计算，其最终公式如(4-7)所示。

$$Pro(p_i)=\beta * Pro_{sem}(p_i)+(1-\beta) * Pro_{spa}(p_i) \tag{4-7}$$

4.3 基于有界区域的多样性和比例性方法

然而，在实际应用中，站点的位置是不确定的，大多数情况下不是边界区域的中心，例如图 4-1 (b)和(c)中的 s_2 和 s_3。在计算 POIs 的分布多样性时，这种情形会导致计算结果偏向距离站点较远的 POIs。此外，公式(4-1)仅考虑了方向分布的多样性。通常情况下，距离站点较远的 POIs 对该站点的客流影响少于距离较近的 POIs。因此，本文提出归一化调整率 $\mathbb{B}$，该值值域为(0, 1)，考虑站点与 POIs 的距离以削弱离站点 *s* 较远的 POIs，则 $\mathbb{B}$ 的定义公式如下：

$$\mathbb{B}=1-\frac{dis(p_i,s)}{\max(dis(p_{(.)},s))+1} \tag{4-8}$$

其中，$\max(dis(p_{(.)},s))$ 表示在相同的边界内所有的 POIs 到站点 *s* 的最大距离。

因此，基于空间有界区域的多样性公式可更新定义为：

$$Div_{spa\_border}(p_i) = \mathbb{B} * \frac{1}{l} \sum_{p_j \in eQue} \frac{dis(p_j, p_i)}{dis(p_i, s) + dis(p_j, s)} \tag{4-9}$$

基于公式(4-1)和公式(4-9)，图 4-1 (a)，(b)，(c)提取 *top*-6 的 POIs 顺序如表 4-1 所示。从表 4-1 可以看出，基于公式(4-9)的计算结果，除了考虑方向分布的多样性，也要考虑 POIs 与站点的距离，距离站点较远的 POIs 对客流的贡献低于较近的 POIs，在排序中应位于后位。

表 4-1 根据图 4-1POIs 多样性的序列表

Tab. 4-1 The diversification sequences of POIs according to Fig. 4-1

图	公式	排序结果
4-1 a)	4-1	p_1 -> p_6 -> p_3 -> p_2 -> p_4 -> p_5
4-1 a)	4-9	p_1 -> p_4 -> p_3 -> p_2 -> p_5 -> p_6
4-1 b)	4-1	p_1 -> p_6 -> p_3 -> p_2 -> p_5 -> p_4
4-1 b)	4-9	p_1 -> p_3 -> p_2 -> p_4 -> p_5 -> p_6
4-1 c)	4-1	p_6 -> p_2 -> p_3 -> p_5 -> p_1 -> p_4
4-1 c)	4-9	p_6 -> p_5 -> p_4 -> p_3 -> p_1 -> p_2

基于有界区域的多样性方法提取 *top-k* POIs 公式更改为：

$$Div_{border}(p_i) = \beta * Div_{sem}(p_i) + (1-\beta) * Div_{spa\_border}(p_i) \tag{4-10}$$

算法 4-1 描述了基于多样性方法的有界区域内 *top-k* POIs 的提取方法的整个过程，图 4-2 是算法的流程图。算法的输入是边界区域的 POIs 和站点，系数 β 和 k 的值，其输出是 k 个 POIs。首先，在边界区域内，计算 POIs 和站点的欧氏距离（Lines 1-2）。将距离站点最近的 POI 加入结果集（即 *eQue* 队列），作为被提取的第一个 POI，同时从候选集（即 *pois*）中删除（Lines 4-5）。当前提取的 POI 数量等于 1，即 $l = 1$（Line 6）。如果 l 小于 k（Line 7），则继续提取 POIs，操作如下：首先根据公式(4-2)和(4-3)计算每个候选 POI 的语义多样性（Line 9）；然后基于公式(4-9)计算每个候选 POI 的空间分布多样性（Line 10）；之后利用公式(4-10)计算每个 POI 的最终分数（Line 11）；最后选择分数最高的 POI 加入 *eQue* 队列中（Line 13），同时 l 值自增 1，候选队列中删除此 POI（Lines 14-15）。Line 17 返回提取的 *top-k* POIs 集合，*eQue* 队列。

与多样性方法类似，POIs 的空间比例性计算公式如下：

$$Pro_{spa\_border}(p_i) = \mathbb{B} * \frac{sum(dir(p_i))}{\alpha * ext(dir(p_i) + 1)} \tag{4-11}$$

比例性方法在有界区域内提取 POIs 的最终公式为：

$$Pro_{border}(p_i) = \beta * Pro_{sem}(p_i) + (1-\beta) * Pro_{spa\_border}(p_i) \quad (4\text{-}12)$$

算法 4-1 基于多样性方法提取 *top-k* POIs

Input: *pois*: POIs in the bounded area; *s*: the station; β: the coefficient; *k*: the number of extracted

Output: *eQue*: the *top-k* extracted POIs

```
1   for each p_i ∈ pois do
2     p_i.dis = dis(s, p_i);
3   end
4   eQue.append(p_i), where p_i has min(p_i, dis);
5   pois.delete(p_i);
6   l = 1;
7   while l < k do
8     for each p_i ∈ pois do
9       p_i.Div_sem = calculateSemDiv(p_i)    根据公式(4-2)和(4-3);
10      p_i.Div_spa = calculateSpaDiv(p_i)    根据公式(4-9);
11      p_i.Div = ß* p_i.Div_sem + (1-ß) * p_i.Div_spa;
12    end
13    eQue.append(p_i), where p_i has max(p_i, Div);
14    l = l+ 1;
15    pois.delete(p_i);
16  end;
17  return eQue;
```

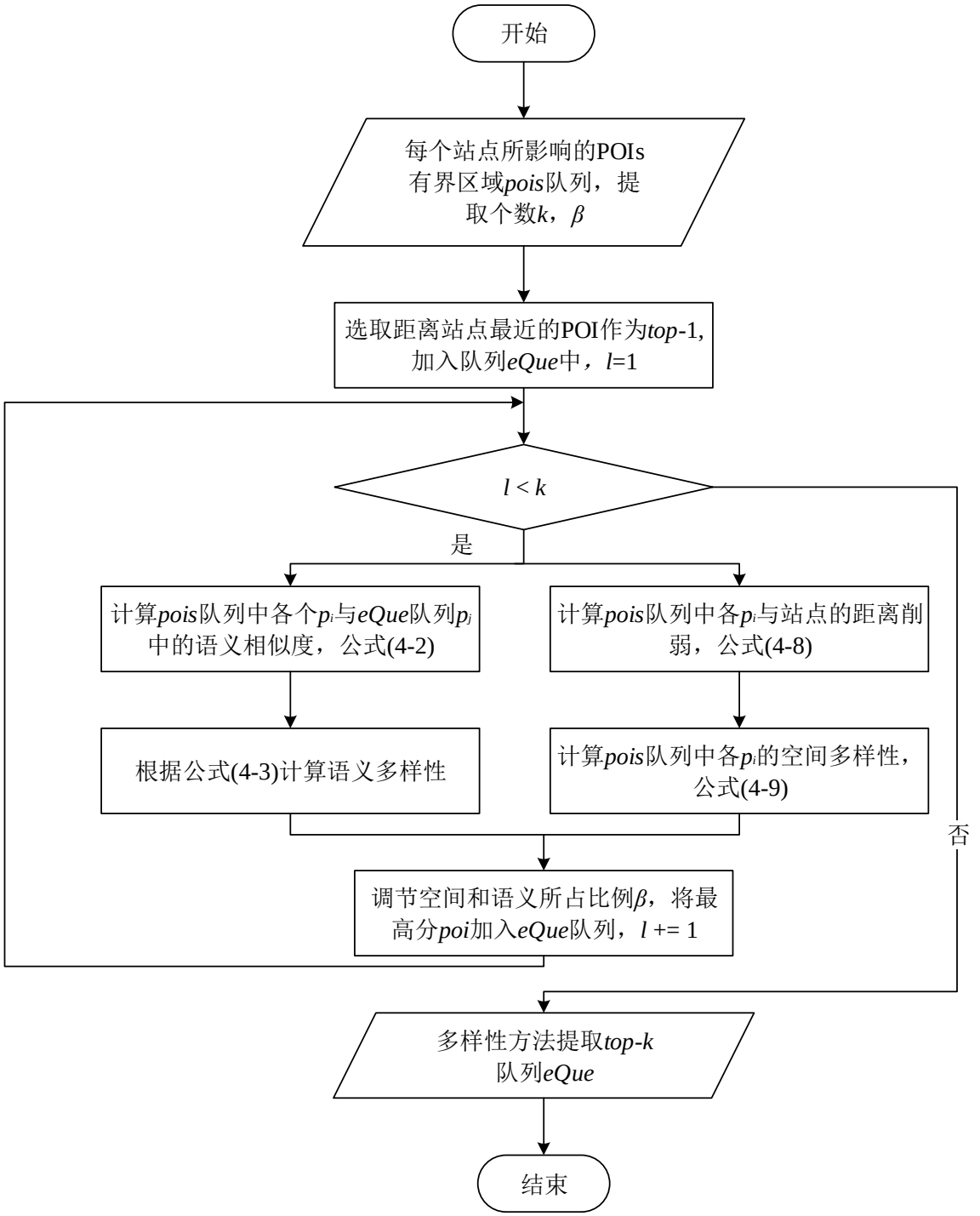


图 4-2 算法 4-2 流程图

Fig. 4-2 The flowchart of Algorithm 4-2

本文选取 POI 到站点的距离最小值作为 *top*-1，然后根据提出的算法提取 *top k*-1 个 POIs。在实际中，亦可以利用国内外对 POIs 的评分作为初始值，选取初始分数最高的作为 *top-1*。如大众点评(http://www.dianping.com/)，Google 地图(http://www.google.cn/maps/)等。

4.4 基于用地特征的站点相似度

通过前三节的介绍，本文提出了基于多样性和比例性的方法提取 *top-k* POIs。由于 POIs 本身没有详细的属性信息，因此每个站点只有三个属性信息：站点的有界区域，有界区域中 POIs 的语义信息和类型信息。 通过计算 POIs 语义和类型的相似性，可以获得站点之间的相似度。在本节中，详细介绍站点语义相似度的计算过程。

4.4.1 文本相似性

由于站点的 top-*k* POIs 可以被认为是 POIs 的矢量 *k*POIs(s_i)，因此可以采用 *Jaccard* 相似度来计算两个站点的 POIs 的文本相似性，如下所述。

$$sim_{text}(kPOIs(s_1), kPOIs(s_2)) = \frac{text(s_1) \bigcap text(s_2)}{text(s_1) \bigcup text(s_2)} \tag{4-13}$$

其中，*k*POIs(s_1)和 *k*POIs(s_2)表示站点 s_1 和 s_2 的 top-*k* POIs；*text*(s_1)={ t_{11}, t_{12}, …, t_{1k} }， *text*(s_2)={ t_{21}, t_{22}, …, t_{2k} }， t_{ij} 表示第 *i* 个站点中第 *j* 个 POI 的名称文本（例如万达广场等）。

在站点的有界区域内，基于语义和空间特征提取 *top-k* POIs。例如 *k*=5，大望路站，记作 s_1，望京站，记作 s_2，s_1 和 s_2 提取的 *top*-5 POIs 的文本名称分别为 *text*(s_1)={金地中心，SOHO 现代城，华贸广场，宏鑫写字楼，万达广场}；*text*(s_2)={望京通信大厦，望京 SOHO，中福百货，望京港旅大厦，望京街道综合商城}。根据公式(4-13)，站点 s_1 和 s_2 的文本相似性可计算为：

$$sim_{text}(kPOI(s_1), kPOI(s_2)) = \frac{text(s_1) \bigcap text(s_2)}{text(s_1) \bigcup text(s_2)} = \frac{0}{10} = 0\% \tag{4-14}$$

4.4.2 类别相似性

通过公式(4-14)的计算可以看出，仅考虑 POIs 的文本相似度并不能准确的反映两个站点的相似性。在 POIs 的相似度计算中，本节利用 POIs 的类别相似度来计算站点间的相似度。虽然文本中没有类似的 POIs，但如果两个 POIs 属于同一类别，则可以将它们视为相似的 POIs。对于上文的同一示例，如果大望路站 s_1 和望京站 s_2 的 POIs 类别分别为 type(s_1)={写字楼，写字楼，广场，写字楼，商场}；type(s_2)={写字楼，写字楼，商场，写字楼，农贸市场}，则 s_1 和 s_2 的类别相似度可计算如下：

$$sim_{type}(kPOI(s_1), kPOI(s_2)) = \frac{type(s_1) \bigcap type(s_2)}{type(s_1) \bigcup type(s_2)} = \frac{4}{10} = 40\% \quad (4\text{-}15)$$

则基于用地特征的站点相似度为提取的 *k* 个 POIs 的文本相似度和类别相似度的和。即：

$$\begin{aligned} sim(kPOI(s_1), kPOI(s_2)) = sim_{text}(kPOI(s_1), kPOI(s_2)) \\ + sim_{type}(kPOI(s_1), kPOI(s_2)) \end{aligned} \quad (4\text{-}16)$$

但通过对公式(4-14)和(4-15)的比较可以看出，在计算两个站点相似度时，利用 POIs 的类别计算，得到的站点相似度更加准确。在本文的实验中，在提取了站点的 *top-k* POIs 后，也只利用了 POIs 间的关系相似度即类别相似计算站点间的相似性。

4.5　本章小结

本章主要介绍了在站点的有界区域内基于空间和语义提取 *top-k* POIs 和基于用地特征的站点相似度计算的方法。本文提出了多样性和比例性两种方法提取 *top-k* POIs。这两种方法都是从空间和语义两个维度介绍的。

本章首先介绍了多样性和比例性方法，但由于 POIs 与站点间距离对站点的影响有差异，距离站点越近的 POIs 对站点影响越大，因此提出了新的基于有界区域的多样性和比例性方法。该新方法的提出，是通过举例表明，实验结果更加符合实际和预期，为计算站点间的相似性提供了输入数据。最后计算站点的相似性，即 POIs 的文本相似性值和类别相似性值的和。通过基于用地特征的站点相似性与前文的客流相似性的比较，验证本章提出方法是否可靠。

第 5 章　OD 分时客流量相似度与基于用地特征的站点相似度比较

本章主要是根据前一章内容，在 k（提取 POIs 的数量）和系数 β（见公式(4-10)和(4-12)）的不同取值下，基于有界区域内空间和语义特征提取 *top-k* POIs 的多样性和比例性方法，以用地特征计算站点间的相似度，通过北京轨道交通系统站点分时客流量的相似性验证提出方法的可靠性，详细给出实验的设置和基准，演示和评估 k、β 和 *NUM*（比较相似站点的个数）的变化对实验结果的影响。

5.1　近似质量，参数 k 和 β 的取值

5.1.1　近似质量

通过比较基于用地特征的站点相似性（即 $sim_{LU}(stations)$ 集合）与北京市轨道交通 OD 分时客流量的站点相似性（即 $sim_{OD}(stations)$ 集合），近似质量（*Approximation Quality*，记作 *AQ*）可通过如下公式计算：

$$AQ = \frac{\{top_{NUM}(sim_{LU}(stations))\} \cap \{top_{NUM}(sim_{OD}(stations))\}}{NUM} \tag{5-1}$$

其中，*NUM* 是比较相似站点的数量，并且 *AQ* 值域为[0, 1]。

5.1.2　参数 k

在提取 *top-k* 方法中，k 值对相似度的近似质量（即 *AQ*）计算具有显著影响。如前所述，站点周围有许多相同且毫无意义的 POIs，例如公共设施，快餐店等。如果在相似度计算中提取并使用这些 POIs，则计算出的各站点间相似性极高且不准确。在实验评估中，我们利用一组不同 k 值，即{5, 15, 25, 35, 50}，分析不同 k 值下提出方法计算站点间相似度的 *AQ* 值。实验结果表明，不同 k 值下站点相似度的趋势遵循正态分布，当 k 在 25 和 35 之间时，获得了最好的近似质量。k 值的设置足以反映相似性变化，其中 50 代表更大值。

5.1.3　参数 β

在所提出的提取 *top-k* POIs 多样性和比例性方法中，系数 β 也起着重要作用，

其可以调整 POIs 的语义和空间特征比例值对相应的多样性和比例性方法的影响。在实验中，通过比较不同空间特征的比例值 β（公式(4-10)和(4-12)），令其取值为{0.1, 0.5, 0.9}，则 1-β（即语义特征的比例，公式(4-10)和(4-12)）的取值为{0.9, 0.5, 0.1}，分析近似质量的变化趋势。β 的三种取值表示三种 POIs 提取情况，偏好语义、平衡和偏好空间。

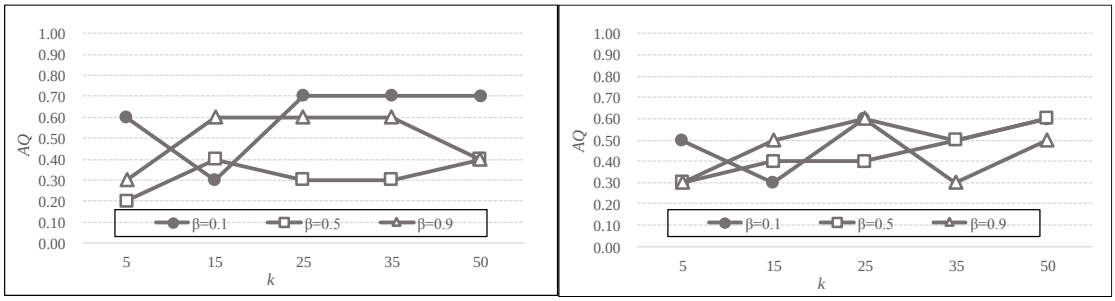
5.2 实验结果与分析

从公式(5-1)可以看出，k，β 和 NUM 的取值影响 AQ。如果 AQ 越接近 1，这不仅表明设定的变量 k，β 的组合好，而且还表明提出的多样性或比例性方法是有效的。

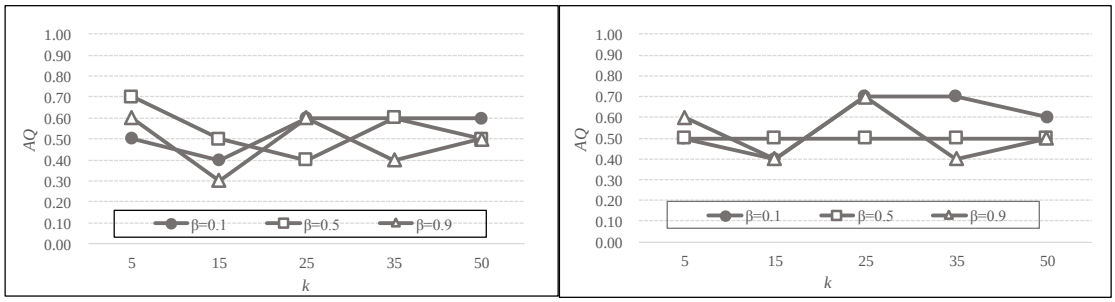
5.2.1 k 值变化

根据提出的多样性和比例性方法，当比较与国贸站相似的前 10 个站点（即 $NUM = 10$）时，不同 k 值下的 AQ 也有所不同，如图 5-1 所示。从图 5-1 中可以发现，所提出方法的最高近似质量仅为 0.70。由于本实验中只利用 10 个相似站点比较，不同值的 k 和 β 的近似质量的规律变化不明显。但仍然可以看出，在大多数情况下，当 $k = 25$ 时可以达到最高的 AQ。

在 $NUM = 20$ 的情况下，利用不同 k 值和 β 值所提的多样性和比例性方法计算的 AQ 如图 5-2 所示，可以看出，随着 k 的增加，AQ 的变化趋势是先上升后下降。在提出的比例性方法中，AQ 达到的最高值是 0.80，这明显高于 NUM=10 时的 AQ。当 $k = 35$ 时，基于多样性方法的计算，AQ 达到最高值。比例性方法是 $k = 25$ 时，AQ 达到最高，且高于基于多样性计算的 AQ。这表明比例性方法优于多样性方法。这是因为基于多样性的方法趋向于提取不同类型的 POIs，但是这只会提取小部分的特殊 POIs，并不能代表站点。此外，当系数 β 为 0.5 时，所提出方法的近似质量高于 0.1 和 0.9，表明 POIs 的语义和空间特征对提取 POIs 是同等重要的。



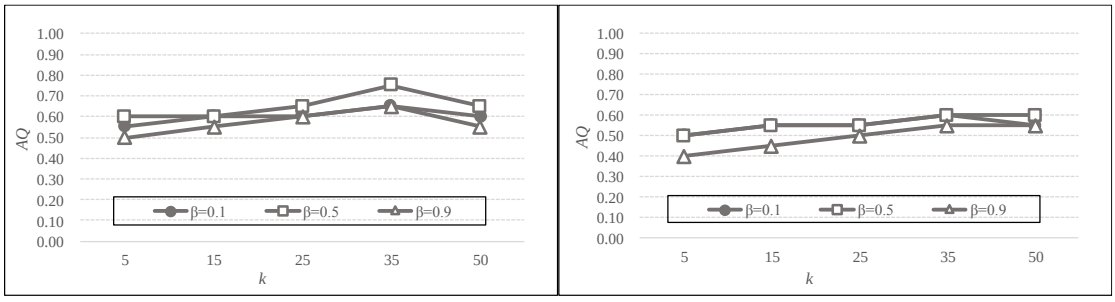
(a) 基于进站的多样性方法　　(b)基于出站的多样性方法



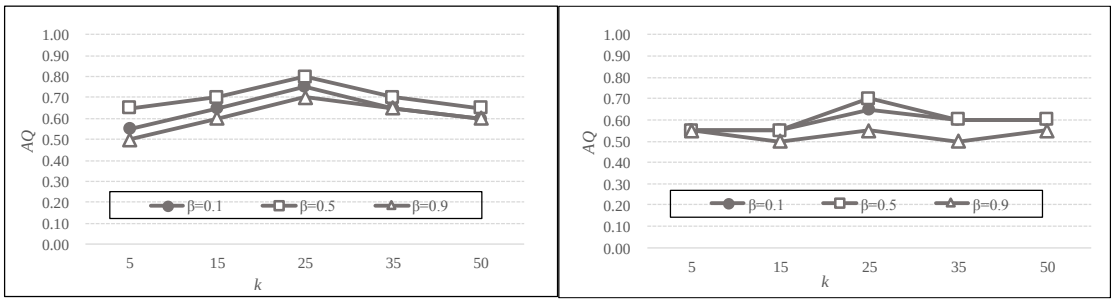
(c) 基于进站的比例性方法　　(d)基于出站的比例性方法

图 5-1 *NUM*=10，国贸站 *k* 值变化的 *AQ* 值

Fig. 5-1 *AQ* varying *k* for guomao station with *NUM*=10



(a) 基于进站的多样性方法　　(b)基于出站的多样性方法



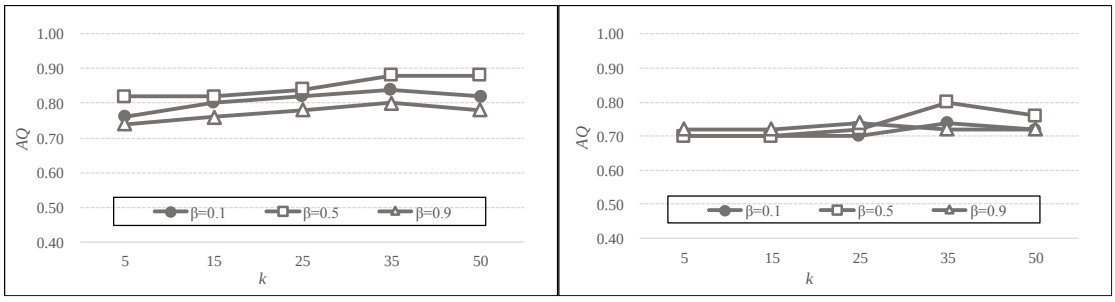
(c) 基于进站的比例性方法　　(d)基于出站的比例性方法

图 5-2 *NUM*=20，国贸站 *k* 值变化的 *AQ* 值

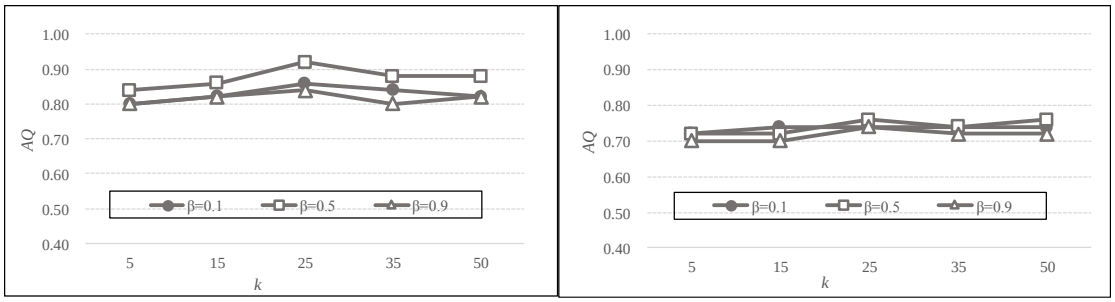
Fig. 5-2 *AQ* varying *k* for guomao station with *NUM*=20

当 *NUM* = 20 时，*AQ* 较 *NUM* = 10 时是有所提升的，但其准确性并没有达到预期效果，所以将 *NUM* 进一步扩大。当扩展到 50 时，国贸站在不同方法、*k*

值和 β 值下的 *AQ* 如图 5-3 所示。从图中可以看出，在 k=25，β=0.5 时，基于比例性方法 *AQ* 取得最高值 0.92，此时的 *AQ* 较高于 *NUM* = 20，并且此时最低 *AQ* 也为 0.7。实验表明随着 *NUM* 的增加，*AQ* 逐步升高。



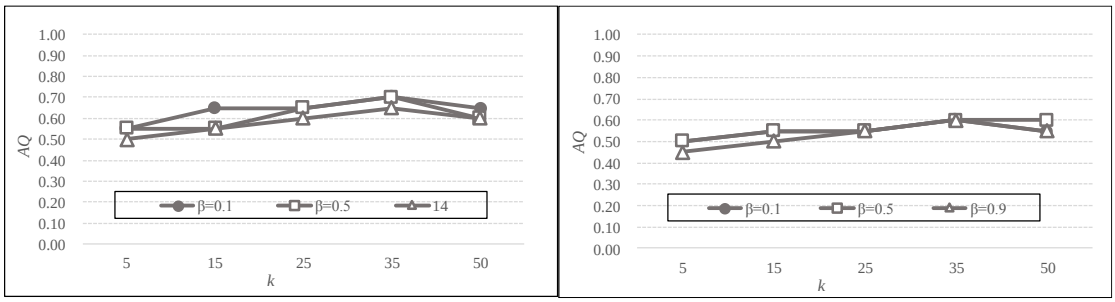
(a) 基于进站的多样性方法　　(b)基于出站的多样性方法



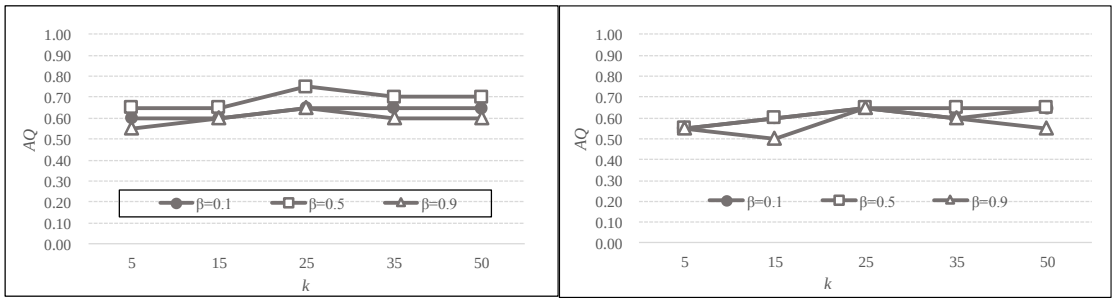
(c) 基于进站的比例性方法　　(d)基于出站的比例性方法

图 5-3 *NUM*=50，国贸站 *k* 值变化的 *AQ* 值

Fig. 5-3 *AQ* varying *k* for guomao station with *NUM*=50



(a) 基于进站的多样性方法　　(b)基于出站的多样性方法



(c) 基于进站的比例性方法　　(d)基于出站的比例性方法

图 5-4 *NUM*=20，天通苑站 *k* 值变化的 *AQ* 值

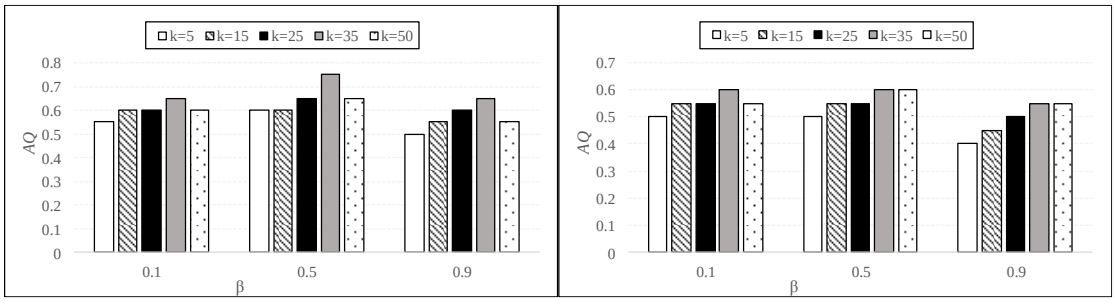
Fig. 5-4 *AQ* varying *k* for tiantongyuan station with *NUM*=20

除此之外，无论是基于多样性方法还是比例性方法，进站的 *AQ* 均高于出站 *AQ*，这是因为进站时采用以一小时为时间间隔，细粒度统计客流量，与站点周边的用地情况的关联度有很大关系。

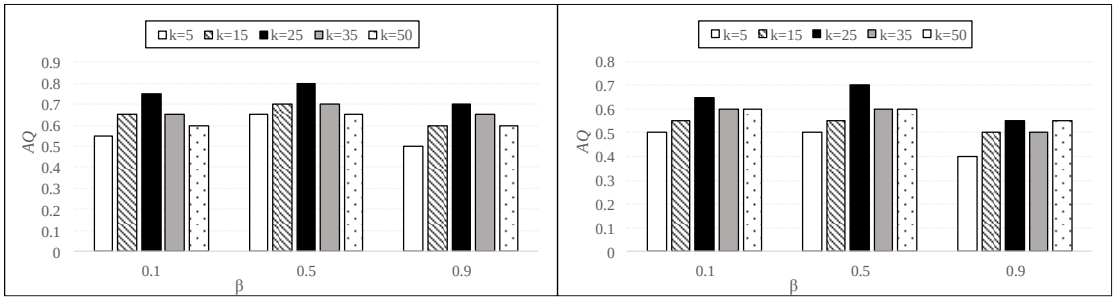
同时，为了全面比较所提出的方法，使用不同站点和 *k* 值，分析 *AQ* 的变化情况。即 *NUM*=20 时，天通苑站的 *AQ* 如图 5-4 所示。如前所述，天通苑站周边是住宅区，与国贸站有很大不同，从图 5-4 可以看出，随着 *k* 的增加，基于多样性方法的 *AQ* 并没有发生很大变化。这是因为天通苑站位于住宅区，POIs 类型相对简单，提取的 POIs 数量对其代表性影响不大。与前面的实验相同，基于比例性方法得到的 *AQ* 高于多样性方法。当 β 设置为 0.5 且 k = 25 时，基于比例性方法取得最高 *AQ* 值。

5.2.2　β 值变化

本节在比较不同 β 的设置值（0.1，0.5 和 0.9）时，基于多样性和比例性方法得到 *AQ* 性能如图 5-5 所示。显然，当 β = 0.5 时，*AQ* 取得最高值，这证明了站点周边的土地使用方法正是 POIs 语义和空间的平衡。另一个有趣的观察结果是当设定 k = 25 或 35 时，产生了最高 *AQ*。根据高斯分布，当设定 β 和 k 取值范围内为中间值时，*AQ* 最大，此时相似度最接近，而且进站和出站都证明了此结论。



(a) 基于进站的多样性方法　　(b)基于出站的多样性方法



(c) 基于进站的比例性方法　　(d)基于出站的比例性方法

图 5-5 *NUM*=20，国贸站 β 值变化的 *AQ* 值

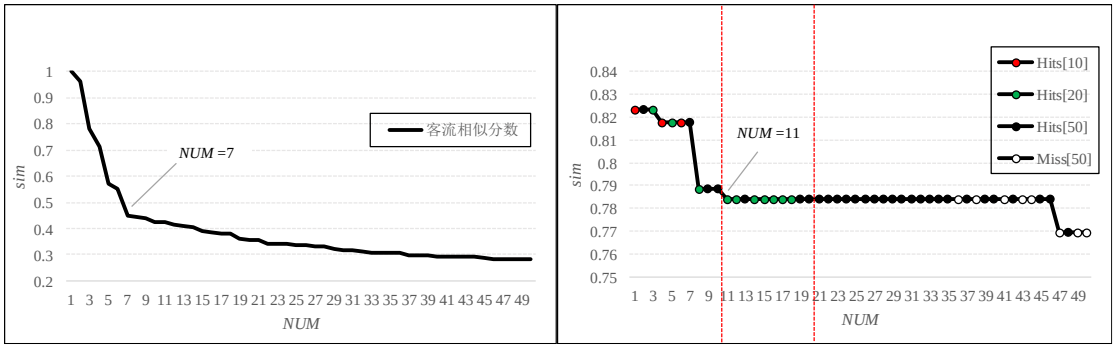
Fig. 5-5 *AQ* varying β for guoamo station with *NUM*=20

5.2.3 *NUM* 值变化

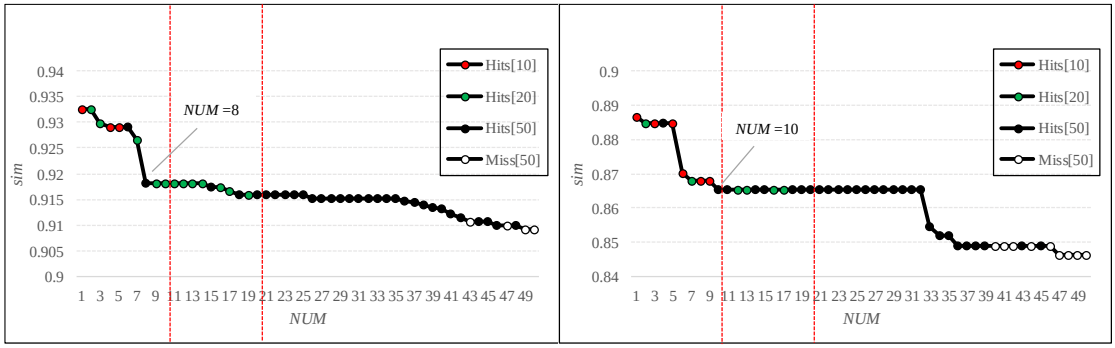
在本小节中，我们用不同的 *NUM* 值（即 10，20 和 50，且设定 β = 0.5）分析了国贸站相似站点的趋势及其相应的 *AQ*，从中进一步研究了多样性、比例性方法和 *k* 对相似性结果的影响，得出以下结论：

（1）首先，根据公式(2-1)和(2-2)，统计进站客流量与国贸站相似的前 50 个站点，相似度 *sim* 得分如图 5-6 (a)所示，从中可以很容易地看出，开始时 *sim* 急剧下降，直到 *NUM* = 7 缓慢下降。图 5-6 (c)基于比例性方法，*k*=25 时的曲线下降走势最符合。

（2）其次，基于比例性方法的 *AQs*（*Hits*）通常高于多样性方法，并且随着 *NUM* 的增加，结果更明显。即当 *NUM* = 10，20，50 时，图 5-6 (c)中 *Miss*(*Hits*)分别是{4} / {10}，{4} / {20}，{4} / {50}，而图 5-6 b)的 *Miss*(*Hits*)是{7} / {10}，{7} / {20}，{8} / {50}。



(a) 国贸站进站客流量相似度　　(b) k=25，多样性方法



(c) k=25，比例性方法　　(d) k=50，比例性方法

图 5-6 β=0.5，国贸站 *NUM* 值变化的相似度

Fig. 5-6 *Sim* varying *NUM* for guomao station with β=0.5

（3）最后，根据 *k* 值的变化分析，较大的 *k* 值（例如图 5-6 (d)中 *k*=50）导致较低 *AQ*，但此时的 *sim* 和下降趋势几乎无变化（即 *NUM* 从 10 到 30 具有相同的分数），这和图 5-6 (b)中的多样性方法结果相类似。

图 5-6 的结果进一步证明了先前的结论，即比例性方法更适合于站点周围的土地利用分析，并且 k 设置为中间值时获得更好的结果。

5.3　本章小结

本章主要是比较基于用地特征的站点相似度和基于静态分时客流量的站点相似度的 AQ 实验。分析所提出的多样性和比例性方法、k 值、β 值、NUM 值的变化对 AQ 的影响。实验结果表明基于比例性方法的 AQ 优于多样性方法，同时 k 值的变化对 AQ 的影响服从高斯分布，当 k 值取值为[25, 35]时，且 POIs 的语义和空间值平衡时（即 β=0.5），AQ 达到最高值。

同时，实验结果中 AQ 的取值范围在 0.6 - 0.9 之间，这表明本文提出的多样性方法和比例性方法基于站点的用地特征计算站点间的相似度是可靠地。并且发现站点周围土地利用的相似性对公共交通环境影响很大，可以为规划或新建站点交通状况提供有意义的指导或预测。该方法对新站点的位置、站点周围土地利用的调整和交通压力的缓解起着指导作用。

第 6 章　基于天气特征和用地特征的站点分类

天气的变化影响人们的出行方式，从而影响轨道交通客流。据有关资料显示，2012 年 7 月 23 日，受暴雨影响，北京市房山线、9 号线客流量增加近一倍。为应对突如其来的降雨对轨道客流影响，北京地铁公司调整了列车运行计划，延长早、晚高峰时段，并缩短高峰和平峰时段的列车运行间隔。由此可见，天气特征不同的情况下，轨道交通客流会呈现不同的规律，尤其降雨、大风和温差较大等情况，均会使轨道交通客流发生波动，这给轨道交通公司和车站客运组织及站点的工作人员将会带来巨大的冲击。

本章首先利用传统的机器学习算法将本文提出的站点用地相似性方法将站点分类，然后利用不同天气特征对客流量的影响对站点分类，进一步验证本文提出的基于用地特征的站点分类效果的可靠性。

6.1　基于用地特征的站点分类

根据第 5 章的介绍，使用多样性或比例性方法提取的 *top-k* POIs 代表站点，以此计算站点间的相似度，利用站点间的相似度对站点分类，主要使用 K-means 聚类算法。

K-means 算法是由 Steinhaus、Lloyd、Ball&Hall 和 McQueen 分别在各自的科学研究领域独立提出的[46]。K-means 算法是一个迭代的过程，聚类为 *k* 个簇，每一个簇中所有数据的均值作为该簇中心。K-means 算法流程[47]为：

1） 随机选择 *k* 个对象作为初始聚类中心，每个对象代表一个类别的中心；

2） 计算样本中其余对象到聚类中心的欧氏距离，将对象分配到距离最近的簇；

3） 计算每个类别中所有对象的均值作为该类别的新聚类中心，并计算所有样本到其所在类别聚类中心的距离平方和；

4） 若聚类中心和距离平方和的值发生改变，重复 2）- 3）步，否则聚类结束。

在利用 K-means 算法聚类时，本文采取了两种数据格式进行分类，一种数据是直接利用站点间的相似性矩阵做输入，另一种是利用站点的坐标做输入。但目前只知道站点与站点间的相似度，各站点的坐标是没有办法知晓的，为了聚类效果的可视化，本文利用站点间的相似度当做站点间的距离，令某一个站点当做原

点，根据站点间的距离，确定其他站点的坐标后再利用 K-means 算法将站点分类。具体实现方法如算法 6-1 所示。

算法 6-1 基于用地特征的站点相似度分类

Input: *similarity*: similarity between stations; *k*: number of clusters
Output: *centList*: clustering result

```
1   Matrix mat <- np.ones(similarity.shape()) * 1;
2   Matrix distance <- mat – similarity;
3   Point point[0][0] <- (0, 0);
4   point <- calculatePoint(distance);
5   centList <- K-means(k, point);
6    return centList;
```

利用站点客流量相似度矩阵作为 K-means 算法的输入是直接且直观的，但其结果难于可视化，从而利用站点的相似性计算站点的坐标再分类，通过相似度矩阵的分类结果验证计算站点坐标分类结果的正确性。

6.2 不同天气特征客流量的站点分类

本实验 AFC 系统客流数据包含 6 月和 7 月的乘客进出站记录，此时天气特征分为雨天和非雨天。统计雨天和非雨天各站点每天进站的客流量，得出有些站点的客流量很小，如次渠南站等，可视为噪声，在分类时将其去掉。

基于天气特征对站点分类，大致分为五类：①上午雨天，下午非雨天对客流量影响较大的站点；②上午非雨天，下午雨天对客流量影响较大的站点；③全天雨天对客流量影响较大且增大客流量的站点；④全天雨天减少客流量的站点；⑤全天雨天对客流量影响较小即不受天气影响的站点。为分析天气特征对站点客流量的影响，具体操作如下：

（1）计算各站点分时段雨天和非雨天的平均客流量，分别记作 *avg_rain_morning*，表示上午雨天，下午非雨天；*avg_rain_afernoon*，表示上午非雨天，下午雨天；*avg_rain*，表示全天雨天；*avg_not_rain*，表示非雨天；

（2）利用各站点的 *avg_rain_morning*，*avg_rain_afernoon* ，*avg_rain* 分别与 *avg_not_rain* 的差值作为将站点分类的依据。第①类站点满足 $|avg\_rain\_morning - avg\_not\_rain| > \frac{avg\_not\_rain + avg\_rain\_morning}{2} \times 0.05$，表示雨天对早高峰客流影响较大，其他时段影响较小的站点；同理，满足

$$|avg\_rain\_afernoon - avg\_not\_rain| > \frac{avg\_not\_rain + avg\_rain\_afernoon}{2} \times 0.05$$

记作第②类，雨天对晚高峰影响较大的站点；

表 6-1 不同天气特征下站点的分类结果

Tab. 6-1 Classification results of stations in different weather characteristic

类别	站点
第一类	八里桥，北苑，次渠，大瓦窑，稻田，分钟寺，丰台东大街，丰台站，俸伯，高米店北，高米店南，管庄，广阳城，果园，黄村火车站，黄村西大街，回龙观东大街，霍营，九棵树，旧宫，科怡路，梨园，篱笆房，立水桥，立水桥南，良乡大学城，良乡大学城北，良乡大学城西，良乡南关，六里桥，马家堡，南邵，苹果园，清源路，沙河，石门，首经贸，顺义，苏庄，天宫院，天通苑，天通苑北，天通苑南，通州北苑，西红门，西苑，小红门，新宫，义和庄，亦庄桥，育新，枣园，长阳，朱辛庄
第二类	北海北，北京站，北苑路北，灯市口，东大桥，东单，动物园，丰台科技园，高碑店，国家图书馆，海淀黄庄，建国门，金台夕照，灵境胡同，南锣鼓巷，荣京东街，王府井，西单，西四，新街口，永安里，园博园，张自忠路，中关村
第三类	安河桥北，奥体中心，八宝山，八角游乐园，北宫门，草房，草桥，常营，成寿寺，传媒大学，慈寿寺，褡裢坡，大红门，大井，丰台南路，公益西桥，巩华城，古城，光熙门，郭庄子，海淀五路居，和平门，黄渠，回龙观，火器营，纪家庙，角门东，角门西，劲松，林萃桥，临河里，刘家窑，六里桥东，龙泽，马泉营，南法信，泥洼，潘家园，蒲黄榆，七里庄，青年路，沙河高教园，芍药居，生命科学园，十里堡，十里河，石榴庄，四惠东，宋家庄，孙河，太阳宫，天坛东门，同济南路，土桥，万寿路，五棵松，西局，亦庄文化园，永泰庄，玉泉路，圆明园，张郭庄
第四类	安华桥，白堆子，白石桥南，北京大学东门，北京西站，朝阳门，车公庄，车公庄西，磁器口，大屯路东，大望路，大钟寺，东四，东四十条，阜成门，复兴门，国贸，国展，和平西桥，呼家楼，花园桥，惠新西街北口，健德门，经海路，军事博物馆，亮马桥，柳芳，木樨地，南礼士路，人民大学，荣昌东街，三元桥，天安门东，天安门西，团结湖，万源街，望京，望京西，魏公村，五道口，西钓鱼台，西二旗，西土城，西直门，宣武门，雍和宫，长春桥，知春里，知春路
第五类	T2 航站楼，T3 航站楼，安定门，安贞门，北京南站，北土城，北新桥，菜市口，崇文门，崔各庄，大葆台，东直门，鼓楼大街，郭公庄，和平里北街，后沙峪，花梨坎，惠新西街南口，积水潭，金台路，莲花桥，牡丹园，农业展览馆，平安里，前门，森林公园南门，上地，生物医药基地，双井，双桥，四惠，陶然亭，西小口，肖村，长椿街

（3）计算 $level = avg\_rain - avg\_not\_rain$ ，当 $level > 0$ 时，记作第③类站点，该类站点受雨天影响客流量增加；当 $level < 0$ 时，表明站点非雨天的客流量多于雨天客流量，继续比较 *level* 与雨天和非雨天客流量平均值的 5%的大小，即 *level* 与 $value = \frac{avg\_rain + avg\_not\_rain}{2} \times 0.05$ 的大小，若 $level < value$ 记作第④类站点，表示该站点的客流量几乎不受天气的影响；否则记作最后一类站点，该类站点受天气影响且雨天客流量减少。通过上述步骤的计算，站点分成五类的结果为表 6-1 所示。

6.3 实验结果与分析

通过 6.1 节的介绍，提出了通过站点相似度和换化站点坐标点两种方法利用 K-means 算法将站点聚类。通过第 5 章的分析和总结，在此选择 $k = 35$、$\beta = 0.5$ 多样性方法和 $k = 25$、$\beta = 0.5$ 比例性方法的站点相似性进行聚类。直接将站点的相似度作为聚类算法的输入，和利用算法 6-1 在站点的相似度上计算站点的坐标作为聚类算法的输入，$k = 35$、$\beta = 0.5$ 多样性方法的聚类结果分别为表 6-2 和 6-3 所示。通过两个表的分类结果对比发现，无论是基于站点的相似度直接聚类还是利用相似度计算站点的相对坐标聚类，这两种聚类算法的分类结果完全一致，这也与我们的预期一样，利用坐标聚类只是为了能更加清晰的展示站点的分类结果。

基于多样性方法用地特征的分类结果，我们发现第一类站点周边的用地主要是住宅小区，该类用地特征较为单一，因此此类为典型的居住型站点；第二类站点周边有大量的写字楼和办公区，该类为就业型；第三类站点周边用地大部分是小区住宅，还有少部分的商业公司，该类为偏居住型；第四类与第三类恰好相反，该类站点周边以办公用地为主，同时沿线分布了娱乐中心、大型商场和交通枢纽，但也有不少居住用地，该类偏就业型；第五类站点周边土地利用较为综合，住宅和办公各半，该类为职住均衡型。

图 6-1 展示了 $k = 35$、$\beta = 0.5$ 多样性和 $k = 25$、$\beta = 0.5$ 比例性方法的站点用地特征相似性的站点分类结果，更加清晰的展示了基于用地特征的站点分类。图 6-1 (b)是使用比例性方法将站点分类，分别为居住型、就业型、偏居住型、偏就业型和职住均衡性，与多样性方法得出的结果一致。

表 6-2 基于多样性方法站点相似性的分类结果

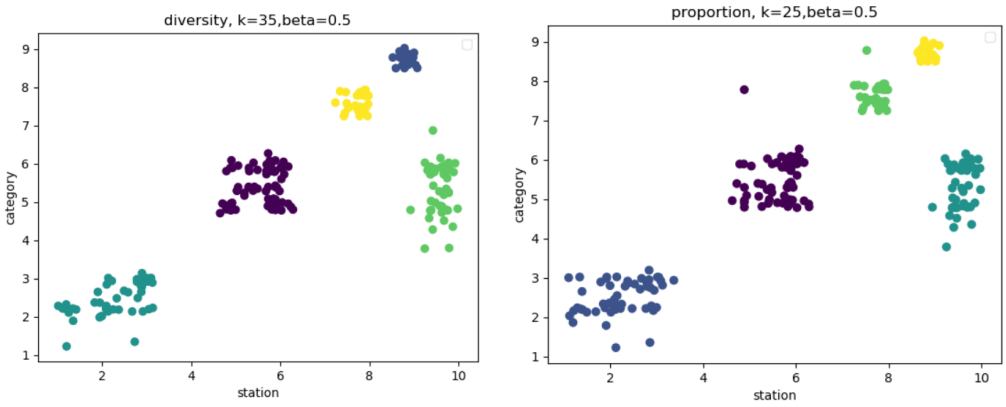
Tab. 6-2 Classification results based on station similarity of diversity method

类别	站点
第一类	霍营，立水桥，北苑，俸伯，新宫，高米店南，枣园，清源路，黄村西大街，黄村火车站，义和庄，天宫院，沙河，稻田，长阳，广阳城，良乡大学城，良乡大学城西，苏庄，旧宫，次渠，管庄，八里桥，篱笆房，通州北苑，石门，马家堡，回龙观东大街，天通苑南，果园，朱辛庄，育新，次渠南，小红门，南邵，丰台东大街，良乡南关，顺义，天通苑，西红门，六里桥，西苑，亦庄桥，梨园，苹果园，分钟寺，首经贸，丰台站，良乡大学城北，高米店北，九棵树，科怡路，立水桥南，天通苑北
第二类	金台夕照，东单，荣京东街，西四，灵境胡同，西单，高碑店，国家图书馆，永安里，王府井，建国门，中关村，动物园，灯市口，海淀黄庄，南锣鼓巷，北海北，东大桥，丰台科技园，园博园，天安门东，天安门西，北京站
第三类	七里庄，大井，西局，惠新西街南口，芍药居，太阳宫，海淀五路居，慈寿寺，马泉营，南法信，沙河高教园，巩华城，生命科学园，四惠东，传媒大学，土桥，宋家庄，龙泽，玉泉路，同济南路，光熙门，公益西桥，亦庄文化园，和平门，临河里，八宝山，奥体中心，天坛东门，劲松，角门西，孙河，古城，八角游乐园，五棵松，万寿路，长椿街，永泰庄，林萃桥，安河桥北，北宫门，圆明园，蒲黄榆，刘家窑，回龙观，十里堡，青年路，褡裢坡，黄渠，常营，草房，潘家园，十里河，成寿寺，石榴庄，大红门，角门东，草桥，纪家庙，泥洼，六里桥东，丰台南路，火器营，张郭庄，大瓦窑，郭庄子
第四类	国贸，复兴门，三元桥，亮马桥，团结湖，呼家楼，花园桥，西直门，大钟寺，西二旗，望京西，望京，阜成门，国展，柳芳，荣昌东街，宣武门，万源街，和平西桥，北苑路北，西土城，知春路，五道口，北京大学东门，知春里，魏公村，经海路，白堆子，雍和宫，军事博物馆，车公庄，白石桥南，木樨地，南礼士路，大望路，朝阳门，东四十条，人民大学，新街口，大屯路东，惠新西街北口，张自忠路，东四，磁器口，健德门，安华桥，西钓鱼台，车公庄西，北京西站，长春桥
第五类	北土城，安贞门，农业展览馆，上地，东直门，崔各庄，花梨坎，郭公庄，平安里，大葆台，肖村，四惠，T2 航站楼，T3 航站楼，菜市口，生物医药基地，双井，前门，崇文门，西小口，森林公园南门，安定门，鼓楼大街，积水潭，北京南站，和平里北街，北新桥，牡丹园，金台路，莲花桥，后沙峪，双桥，陶然亭

表 6-3 基于多样性方法站点坐标的分类结果

Tab. 6-3 Classification results based on station coordinates of diversity method

类别	站点
第一类	南邵，丰台东大街，良乡南关，顺义，天通苑，西红门，清源路，黄村西大街，立水桥南，黄村火车站，义和庄，天宫院，沙河，稻田，长阳，朱辛庄，广阳城，良乡大学城，良乡大学城西，良乡大学城北，苏庄，旧宫，次渠，管庄，八里桥，篱笆房，通州北苑，石门，马家堡，回龙观东大街，天通苑南，果园，九棵树，育新，次渠南，小红门，六里桥，西苑，亦庄桥，梨园，苹果园，天通苑北，分钟寺，首经贸，丰台站，科怡路，霍营，北苑，俸伯，新宫，高米店北，高米店南，枣园
第二类	金台夕照，东单，荣京东街，动物园，建国门，西单，高碑店，国家图书馆，永安里，王府井，天安门东，天安门西，北京站，西四，灵境胡同，中关村，北海北，东大桥，丰台科技园，园博园，灯市口，海淀黄庄，南锣鼓巷
第三类	光熙门，公益西桥，丰台南路，和平门，临河里，八宝山，奥体中心，长椿街，劲松，角门西，七里庄，孙河，古城，八角游乐园，五棵松，宋家庄，龙泽，玉泉路，北宫门，同济南路，郭庄子，圆明园，蒲黄榆，大红门，刘家窑，回龙观，十里堡，青年路，褡裢坡，黄渠，常营，张郭庄，潘家园，十里河，成寿寺，石榴庄，泥洼，角门东，草桥，纪家庙，大瓦窑，六里桥东，火器营，万寿路，草房，永泰庄，林萃桥，大井，西局，亦庄文化园，惠新西街南口，芍药居，太阳宫，天坛东门，慈寿寺，马泉营，南法信，沙河高教园，巩华城，生命科学园，四惠东，传媒大学，海淀五路居，土桥
第四类	知春里，长春桥，安华桥，柳芳，木樨地，亮马桥，团结湖，呼家楼，花园桥，惠新西街北口，大钟寺，西二旗，望京西，国贸，阜成门，国展，北京西站，望京，荣昌东街，宣武门，万源街，和平西桥，三元桥，北苑路北，西土城，知春路，五道口，北京大学东门，西钓鱼台，魏公村，经海路，白堆子，雍和宫，军事博物馆，西直门，车公庄，白石桥南，车公庄西，南礼士路，大望路，朝阳门，东四十条，人民大学，新街口，大屯路东，健德门，张自忠路，复兴门，东四，磁器口
第五类	东直门，崔各庄，花梨坎，郭公庄，平安里，大葆台，肖村，四惠，T2 航站楼，T3 航站楼，菜市口，安贞门，生物医药基地，北京南站，前门，崇文门，北土城，西小口，森林公园南门，安定门，鼓楼大街，积水潭，上地，和平里北街，北新桥，牡丹园，金台路，莲花桥，双井，后沙峪，双桥，陶然亭，农业展览馆



(a) 基于多样性方法的分类结果　　(b) 基于比例性方法的分类结果

图 6-1 基于站点用地特征的 K-means 分类结果

Fig. 6-1 *K-means* classification results based on land use characteristics of station

根据上一节基于不同天气特征客流量的变化分类结果，我们发现第一类站点受雨天影响，早高峰客流量变化明显，而且进站客流量日分布规律表现为早高峰客流量较大。早高峰进站乘客多，必然是周边居住小区较多才导致，这与基于用地特征的第一类结果一样。同样，第二类站点受雨天影响晚高峰客流量变化明显，且进站客流量日分布规律表现为晚高峰客流量大，因此为就业型。第三类站点受雨天影响，分时客流量形态主要表现为早晚高峰变化明显，但早高峰变化量明显高于晚高峰，进站客流量出现一高一低双峰，此区域包含较大比例居住人群和较小比例的工作人群，与基于用地特征的第三类一致。第四类站点受雨天影响，客流量早晚高峰变化呈减少趋势，且晚高峰变化更为明显，对进站客流量而言，出现一低一高双峰，晚高峰时段乘客返程比例高于早高峰离家乘客比例，属于偏就业型。第五类站点受天气影响变化不大，此类站点早高峰客流量与晚高峰客流量大致相等，居住人群的出行比例和工作人群的出行比例均衡，表明站点用地职住均衡。

通过分析天气特征对客流量变化影响，进一步验证了我们所提出的基于用地特征的多样性和比例性方法对站点分类的有效性，可为无历史客流量数据的站点根据其周边的用地情况在天气变化时，对其同一类站点的客流量形态规律进行分析，总结不同天气特征对该站点客流量的影响规律，且为轨道交通运营部门预测极端天气对客流量的影响变化，提供制定合理的行车组织计划和管理方案等作为重要依据。

比较基于用地特征的站点分类，和基于天气对客流量影响的分类结果，图 6-2 展示了提出的两种方法的分类与天气特征对客流量影响的分类的不同，从图中可以看出两种方法的结果与天气分类结果的差异性很小，且比例性方法的分类

结果与天气的分类结果更接近，这进一步验证了前文提到的比例性方法优于多样性方法。

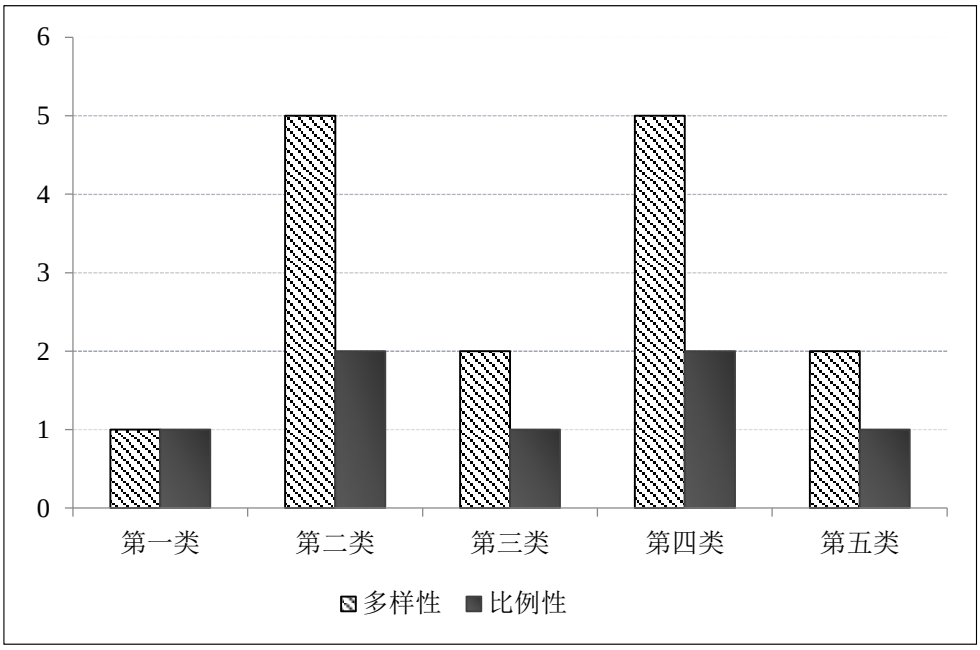


图 6-2 多样性和比例性方法的分类与天气客流量变化的分类的差异性

Fig. 6-2 Differences between classification of diversity and proportional methods and changes passenger volume of weather

6.4 本章小结

本章主要是通过不同天气特征对客流量变化的站点分类，验证基于用地特征的站点分类的效果。首先，主要介绍了利用机器学习的方法将站点基于用地特征分类，主要展示了当 $k = 35$、$\beta = 0.5$ 多样性方法和 $k = 25$、$\beta = 0.5$ 比例性方法的分类结果；然后介绍了基于不同天气客流量变化的站点分类方法。两个分类结果均分为五类，通过实验对比和分析，证明了基于用地特征的站点分类方法的可靠性，这可以实现新建站点和规划站点的客流量预测，并针对极端天气对无历史数据的站点客流量影响提供数据支撑和运营方案。

结 论

本文旨在基于用地特征的站点分类方法研究。利用城市轨道交通周边的土地使用，分析土地利用和客流量关系，将站点分类，可预测规划中的站点客流量，对相似站点周边的土地使用情况进行二次开发和利用具有指导作用，并对城市交通站点规划提供思路。

现将本文主要内容与贡献总结如下：

1. 提出了基于进出分时客流量计算站点间的相似度方法。计算进出站分时段和全天的客流量，根据其客流量提出计算方法，计算站点基于分时客流量的相似性，作为基于用地特征的站点相似性比较的验证集。

2. 提出了划分有界区域的 RC-*tree* 算法，找到站点的覆盖范围，生成站点的有界区域。每个站点所影响的 POIs 有界区域作为提取 *top-k* POIs 方法的数据输入。

3. 提出了基于有界区域内提取 *top-k* POIs 方法。提出了多样性和比例性方法用于提取有代表性的 POIs，每个 POI 的得分通过其语义和空间两个维度计算。根据提取 POIs 的个数 k 值和空间特征所占比例 β 值的不同取值，得到不同的结果。

4. 提出了基于用地特征计算站点的相似性计算方法。

5. 实验比较了基于进出分时客流量的站点相似度结果与基于用地特征的站点相似度结果，验证了提出方法的有效性。

6. 基于天气和用地特征的站点分类。进一步验证了提出的基于用地特征的站点相似度计算方法的可靠性。

虽然本文在城市轨道交通客流量、土地利用、以及不同天气对客流量影响的分类取得了一些阶段性的研究结果，但目前还有一些问题和更复杂的实验没有研究，仍然需要进一步地深入探索。

1. 通过对实验结果的分析总结可以看出，本文提出的方法是可行和可靠的，但使用的客流数据和土地利用数据的时间不一致，方法正确性还是有待进一步提高。

2. 提出的划分区域 RC-tree 算法中，仅考虑了站点和 POIs 的直线距离，与实际距离偏差有一定的差异。

参 考 文 献

[1] 周晓勤.中国城市轨道交通的发展现状及机遇[J].城市轨道交通,2018(10):23.

[2] 吴爽王波，李世民等.北京轨道交通络运营问题分析及规划建设改进建议[C].2014 两岸四地城市轨道交通学术研讨会,北京,2014.1-13.

[3] 傅搏峰,吴娇蓉,华陈睿.轨道站出入口客流分布系数估计方法[J].同济大学学报(自然科学版),2012,40(11):1660-1665.

[4] 袁江,彭磊.广州地铁运营客流分布特征研究与应用[J].都市快轨交通,2018,31(04):63-68.

[5] 周玮腾,韩宝明.考虑列车容量限制的地铁网络客流分配模型[J].华南理工大学学报(自然科学版),2015,43(08):126-134+143.

[6] Xu X , Li K , Li X . Research on passenger flow and energy consumption in a subway system with fuzzy passenger arrival rates[J]. Proceedings of the Institution of Mechanical Engineers Part F Journal of Rail & Rapid Transit, 2015, 93(12):228-254.

[7] D. Xiaobing, H. Hua, L. Zhigang, Z. Haiyan, and S. Sivasundaram. The aggregation mechanism mining of passengers flow with period distribution based on suburban rail lines. Thirtieth International Conference on Very Large Data Bases, 2017.

[8] S. Yujuan, Z. Guanghou, Y. Huanhuan, and N. Huimin. Passenger flow prediction of subway transfer stations based on nonparametric regression model. Discrete Dynamics in Nature and Society, 2014.

[9] Z. Shuzhi, N. Tonghe, W. Yang, and G. Xiangtao. A new approach to the prediction of passenger flow in a transit system. Computers and Mathematics with Applications, 2010.

[10] L. Yang, W. Xudong, S. Shuo, M. Xiaolei, and L. Guangquan. Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks. Transportation Research Part C, pages 306–328, 2017.

[11] S. Yuxing, L. Biao, and G. Wei. A novel wavelet-svm short-time passenger flow prediction in beijing subway system. Neurocomputing, pages 109–121, 2015.

[12] B. Valentina, Emilia, C. J. Lakhmi, Z. Xiangmo, S. Fuqian, Z. Yiming, and P. Qingge. Study on Passenger Flow Analysis and Prediction Method of the Public Transport Operation Passenger Line of the Adjacent City. 2017.

[13] G. Shengguo and W. Zhong. Modeling passenger flow distribution based on travel time of urban rail transit. Journal of Transportation Systems Engineering and Information Technology, pages 124–130,2011.

[14] Z. Ling, L. Yuejun, L. Yu, and L. Ying. Research on the spatialsystem-based rail transit systems of the world cities. Procedia Engineering, pages 699–708, 2016.

[15] 王静，刘剑锋，孙福亮. 北京市轨道交通线网客流分布及成长规律[J]. 城市交通，2012，10(2)：26–32.

[16] 王静，刘剑锋，马毅林等. 北京市轨道交通车站客流时空分布特征城市交通[J]. 城市交通，2013(6):18-27.

[17] C. Bizhuang, W. Zhongqiang, and W. Yilin, Yang. Analysis on passenger flow characteristics of shanghai rail transit network and its enlightenment. Urban Transport of China, pages 28–34, 2013.

[18] 李金海，李明高，杨冠华，郭印. 北京轨道交通网络化客流特征及成长趋势分析[J]. 交通工程，2017(3):53-57.

[19] Jun M , Choi K , Jeong J , et al. Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul[J]. Journal of Transport Geography, 2015.

[20] Ma X , Chen X , Li X , et al. Sustainable station-level planning: An integrated transport and land use design model for transit-oriented development[J]. Journal of Cleaner Production, 2018, 170:1052-1063.

[21] Calvo F , OAJD , Fernando Ar án. Impact of the Madrid subway on population settlement and land use[J]. Land Use Policy, 2013, 31(none):627-639.

[22] D. Binglei, Xie, Chuan. An evaluation on coordinated relationship between urban rail transit and land-use under tod mode. Journal of Transportation Systems Engineering and Information Technology,pages 9–13, 2013.

[23] K. Yena, Song;Hyun. Evolution of subway network systems; subway accessibility; and change of urban landscape: A longitudinal approach to seoul metropolitan area. International Journal of Applied Geospatial Research (IJAGR), 2015.

[24] Y. H. Seungil, Lee;Changhyo. Urban structural hierarchy and the relationship between the ridership of the seoul metropolitan subway and the land-use pattern of the station areas. Cities, pages 69–77, 2013.

[25] S. H. Doina, Olaru;Brett. Residential location and transit-oriented development in a new rail corridor. Transportation Research Part A, pages 172–173, 2017.

[26] 谢明隆. 轨道客流特征与土地利用的互动关系研究[A]. 中国城市规划学会城市交通规划学术委员会、福州市人民政府.公交优先与缓堵对策——中国城市交通规划 2012 年年会暨第 26 次学术研讨会论文集[C].中国城市规划学会城市交通规划学术委员会、福州市人民政府:,2012:5.

[27] 李世民,安栓庄,贺腊妮.轨道交通与城市用地开发互动影响研究[J].都市快轨交通,2013,26(05):25-29.

[28] Kuby M , Barranda A , Upchurch C . Factors influencing light-rail station boardings in the United States[J]. Transportation Research, Part A (Policy and Practice), 2004, 38(3):0-247.

[29] Hang M , Ningning H , Mingyang C . The Development and Problems of Land Use Along Urban Rail Transit in China[C]. International Conference on Optoelectronics & Image Processing. IEEE, 2010.

[30] Arana P , Cabezudo S , M. Peñalba. Influence of weather conditions on transit ridership: A statistical study using data from Smartcards[J]. Transportation Research Part A, 2014,

59(1):1-12.

[31] Zhou M , Wang D , Li Q , et al. Impacts of weather on public transport ridership: Results from mining data from different sources[J]. Transportation Research Part C: Emerging Technologies, 2017, 75:17-29.

[32] Koetse M J , Rietveld P . The impact of climate change and weather on transport: An overview of empirical findings[J]. Transportation Research Part D: Transport and Environment, 2009, 14(3):205-221.

[33] 杜恒. 天气因素对轨道交通客流的影响[A]. 中国城市规划学会城市交通规划学术委员会.2017 年中国城市交通规划年会论文集[C].中国城市规划学会城市交通规划学术委员会:中国城市规划设计研究院城市交通专业研究院,2017:14.

[34] Meyer M D , Weigel B . Climate Change and Transportation Engineering: Preparing for a Sustainable Future[J]. Journal of Transportation Engineering, 2011, 137(6):393-403.

[35] 赵珍祥.基于 PWNN 模型的轨道交通客流预测分析[J]. 黑龙江交通科技,2018,41(11): 162 - 164.

[36] 郇宁,谢俏,叶红霞,姚恩建.基于改进 KNN 算法的城轨进站客流实时预测[J].交通运输系统工程与信息,2018,18(05):121-128.

[37] 吕慎,过秀成.轨道线网客流预测方法研究[J].系统工程理论与实践,2001(08):106-110.

[38] 李国强. 基于 AFC 和 POI 数据的轨道交通站点客流影响因素挖掘[A]. 中国城市规划学会城市交通规划学术委员会.创新驱动与智慧发展——2018 年中国城市交通规划年会论文集[C].中国城市规划学会城市交通规划学术委员会:中国城市规划设计研究院城市交通专业研究院,2018:11.

[39] AZARKHAIL M, PETER W. Uncertainty management in model-based imputation for missing data[C]. Annual Reliability abd Maintainability Symposium(RAMS), 2013-01-28:1-7.

[40] XiaofengZhu, Shichao Zhang. Missing value estimation for mixed attribute data set[J]. IEEE Trans on Knowledge and Data Engineering, 2011,23(1):110-121.

[41] Guttman A. R-trees: a dynamic index structure for spatial searching [M]. ACM, 1984.

[42] J. G. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In SIGIR, pages 335–336, 1998.

[43] G. Fakas, Z. Cai, and N. Mamoulis. Diverse and proportional size-l object summaries for keyword search. ACM SIGMOD International Conference, pages 363–375, 2015.

[44] V. Dang and W. B. Croft. Diversity by proportionality: an electionbased approach to search result diversification. In International Acm Sigir Conference on Research and Development in Information Retrieval, pages 65–74, 2012.

[45] G. J. Fakas, Z. Cai, and N. Mamoulis. Diverse and proportional size-l object summaries using pairwise relevance. The VLDB Journal, 25(6):791–816, 2016.

[46] 王千,王成,冯振元,叶金凤. K-means 聚类算法研究综述[J]. 电子设计工程, 2012, 20(07):21-24.

[47] Anil K J. Data clustering: 50 years beyond K-Means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

攻读学位期间发表的学术成果

[1] Zhi Cai, Tong Li, Xing Su, Limin Guo, Zhiming Ding. Research on Analysis Method of Characteristics Generation of Urban Rail Transit. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, 二次修改后审阅中.

[2] Kun Lang, Wei Han, Tong Li, Zhi Cai. Authority Based Ranking in Propaganda Summary Generation Considering Values. Advances in Intelligent Systems and Computing, ISSN: 2194-5357.

[3] 才智, 李彤, 兰许, 曹阳, 丁治明. 基于多样性的地理空间兴趣点检索方法, 发明, 201611254804.X, 实审, 2017.01.26.

[4] 才智, 李彤, 兰许, 丁治明. 狭隘范围内文献的多样性查询方法, 发明, 201710163193.6, 实审, 2017.04.20.

[5] 才智, 李彤, 郎琨, 才博远, 苏醒. 一种基于有界区域多样性和比例性的站点 *top-k* 个 POI 的方法, 发明, 201811156725.4, 实审, 2018.09.30.

致 谢

白驹过隙，三年的时间恍如昨日，这三年的研究生活，有太多的感慨，我的导师、同学和家人都给予了我很大的指导、帮助和鼓励，在此我要衷心的感谢大家。

首先，我要感谢才智副教授在研究生期间对我学业上的指导和生活上的照顾。才老师兢兢业业、勤勤恳恳，他的钻研态度深深地影响了我，并且有着严谨的学术态度和尽职尽责的工作精神。同时，才老师在毕业课题的方向把握、算法创新和实验设计等方面的指导，锻炼了我思考问题的思维和解决问题的方式方法。在此，我很感谢才老师对我的细心栽培，会一直谨记于心。

此外，我还要感谢苏醒老师，苏老师治学严谨，学术上无半点马虎，在我完成小论文初稿时，苏老师百忙之中抽出空来对我的论文认真批改，严格把关论文内容，提出许多中肯的指导意见，使我顺利完成论文，万分感谢苏老师的帮助。

其次，我要感谢实验室的全体师生，感谢师兄、师姐们对我的帮助，还有阚海鹏、闫琦、郭耀光等同学在学业和生活中的关怀和帮助。

再次，我要感谢我的朋友和家人。感谢所有朋友们在北京工业大学的细心照顾，让我度过了愉快而难忘的时光。感谢家人们对我的鼓励和支持，陪我走过这十多年的求学之路，感谢他们的一直信任。

最后，感谢评阅、评议和答辩委员会的各位老师在百忙之中给予我的指导。