摘要

普遍存在的定位设备衍生了海量的轨迹数据。大规模轨迹数据的分析应用成为轨迹数据领域的重要问题。轨迹聚类不仅能帮助了解城市交通流量、拥堵状况和路网负载情况，还可以作为交通规划和管理的重要依据。然而，各场景下大规模轨迹数据的分析性能仍是巨大的挑战。快速大规模轨迹聚类可以在短时间内识别出乘客出行的热点路段，为数据驱动的公交网络优化提供依据。本文针对大规模轨迹聚类以及公交网络优化展开了深入研究，主要工作如下：

（1）轨迹预处理阶段，将原始轨迹数据映射到路网结构，并且分析路网的边在轨迹中的连续共现概率，提出了一种基于路网结构的距离度量方式 LORE（the Longest Overlapped Repeating Edges）。LORE 仅使用顺序交叉遍历即可返回相似值，其最坏情况下时间复杂度为$O(m+n)$。相较于主流的距离度量方式，极大地缩短了距离计算的时间。对比实验表明，LORE 距离度量能有效提高大规模轨迹聚类的效率。

（2）轨迹聚类阶段，提出了快速大规模轨迹聚类算法 KMTC（K-Mediod-based Trajectory Clustering）和基于中心表的质心路径细化算法 CPRP（Centroid Path Refine based on Pivot-table），有效处理大规模轨迹数据，优化了轨迹聚类的性能。KMTC 算法是针对轨迹数据的特性改进了 k-medoids 算法，通过在轨迹分配与质心路径细化过程中应用三角不等式与边界剪枝策略，提升了算法的聚类效果与执行速度。CPRP 算法首先利用中心表将相似的轨迹绑定为一组，减少轨迹分配阶段的分配次数；然后对路网各边建立倒排索引，进一步减少距离计算成本；最后通过图遍历算法避免轨迹数据重复访问，得到质心路径。通过实验表明，KMTC 和 CPRP 的结合能够进一步的提升轨迹聚类效果以及算法性能，比传统轨迹聚类方法在时间方面提升了一个数量级。

（3）提出了一个公共交通路线规划的解决方案 RDDC（Route Design based on Demand and Connectivity），覆盖通勤需求热点路段且保障公交网络连通性的公交线路规划。通过扩展 KMTC 算法并结合网络连通性因素，设计一种启发式方法 ERD（Extension based Route Design）来实现 RDDC。使用 Lanczos 方法估计有界误差的公交网络的自然连通性，避免矩阵运算带来的计算膨胀，加速算法的收敛。基于边界思想，提出 IUB（Incremental Update on Bound）算法，利用边界增量进行贪婪策略扩展选择边，减少迭代过程中的重复矩阵计算。在两个真实的数据集对本文提出的方法进行实验验证，结果表明本文提出的方法是有效可行的，并且相对于仅依赖通勤需求的优化方案，通勤最短距离降低 20~30%。

关键词： 距离度量；轨迹聚类；边界剪枝；公交网络优化；Lanczos 矩阵估计

Abstract

Ubiquitous positioning devices generate massive amounts of trajectory data. The analysis and application of large-scale trajectory data has become an important issue in the field of trajectory data. Trajectory clustering can not only help understand urban traffic flow, congestion and road network load, but also serve as an important basis for traffic planning and management. However, the analysis performance of large-scale trajectory data in various scenarios is still a huge challenge. Rapid large-scale trajectory clustering can identify hot spots for passengers to travel in a short time, providing a basis for data-driven bus network optimization. This paper conducts in-depth research on large-scale trajectory clustering and bus network optimization. The main work is as follows:

(1) In the trajectory preprocessing stage, the original trajectory data is mapped to the road network structure, and the continuous co-occurrence probability of the edges of the road network in the trajectory is analyzed, and a distance measurement method based on the road network structure LORE (the Longest Overlapped Repeating Edges) is proposed. LORE can return similar values only using sequential cross traversal, and its worst-case time complexity is O(m+n). Compared with the mainstream distance measurement method, it greatly reduces the time of distance calculation. Finally, the comparative experiments show that the LORE distance metric can effectively improve the efficiency of large-scale trajectory clustering.

(2) In the trajectory clustering stage, a fast large-scale trajectory clustering algorithm KMTC (K-Mediod-based Trajectory Clustering) and a centroid path refine algorithm CPRP (Centroid Path Refine based on Pivot-table) are proposed, which can effectively deal with large-scale trajectories data, optimizing the performance of trajectory clustering. The KMTC algorithm improves the k-medoids algorithm according to the characteristics of the trajectory data. By applying the triangle inequality and the boundary pruning strategy in the process of trajectory assignment and centroid path refinement, the clustering effect and execution speed of the algorithm are improved. The CPRP algorithm first uses the pivot-table to bind similar trajectories into a group to reduce the number of allocations in the trajectory allocation stage; then establishes an inverted index for each side of the road network to further reduce the cost of distance calculation; finally uses the graph traversal algorithm to avoid trajectory data duplication Access to get the centroid path. Experiments results show that the combination of KMTC and CPRP can further improve the trajectory clustering effect and algorithm

performance, which is an order of magnitude faster than the traditional trajectory clustering method in terms of time.

(3) A public transportation route planning solution RDDC (Route Design based on Demand and Connectivity) is proposed, which covers the hot spots of commuting demand and ensures the connectivity of the bus network. By extending the KMTC algorithm and combining network connectivity factors, a heuristic method ERD (Extension based Route Design) is designed to realize RDDC. The Lanczos method is used to estimate the natural connectivity of the bus network with bounded errors, avoiding the calculation expansion caused by matrix operations, and accelerating the convergence of the algorithm. Based on the boundary idea, the IUB (Incremental Update on Bound) algorithm is proposed, which uses the boundary increment to expand the selection edge with a greedy strategy, and reduces repeated matrix calculations in the iterative process. The method proposed in this paper is verified experimentally on two real data sets. The results show that the method proposed in this paper is effective and feasible, and for the optimization scheme that only depends on commuting needs, the shortest commuting distance is reduced by 20-30%.

Keywords: Distance metric；Trajectory clustering；Boundary based pruning；Bus network optimization； Lanczos matrix estimation

目录

第一章 绪论

§1.1 研究背景与意义

随着科学技术不断发展以及定位设备的普及使用，轨迹数据[1][2]正在以指数形式增长。它可以由各种各样的设备产生，如全球定位系统（GPS）设备、摄像头和射频识别（RFID）读取器[3]。一辆带定位设备的汽车一小时内可以产生 15G 的庞大数据量[4]。这些具有时空特性的轨迹点蕴含着大量的研究价值，可以为交通流量的预测和优化以及城市规划管理提供诸多的可能。

轨迹数据的挖掘与分析有诸多应用价值，如在 Foursquare[5]中，将用户在社交平台打卡信息的顺序作为轨迹，用于推荐系统；在像 GeoLife[6][7]这样的 GPS 轨迹分享网站上，人们上传自己在生活、工作、学习等场景的生活轨迹数据；许多出租车都安装了 GPS 传感器，它们的实时位置以数据流的形式报告给交通管理系统[8]；生物学家们为了了解动物如何随着时间的推移移动和与环境互动，记录麋鹿等动物的运动轨迹来验证[9]；战场中军事卫星监视指定区域，收集可能的入侵者的行动，他们的轨迹被军事网络收集[10]。轨迹聚类是指将相似轨迹样本划分到同一类别中的聚类算法，其通常基于轨迹样本的空间和时间信息，以及轨迹样本之间的相似度来确定聚类结果。

轨迹聚类的主要作用是具有在高维度、非线性、噪声和不完整性等特点的轨迹数据中发现相似的轨迹集合，通过分析集合内、集合间的关系，从而可以实现数据压缩、轨迹分类、相关分析、异常检测等。因此轨迹聚类算法的应用非常广泛，例如在交通流量监测、智能交通系统、旅游规划、自然灾害监测等领域都有应用。更具体的来说，轨迹聚类可以快速的反应出用户需求量大的道路，为决策者提供帮助。

此外，经济高速增长和城市化进程的推进为我们提供现代化生活条件的同时，也带来了交通拥堵、规划落后、环境污染等巨大的压力和挑战[11]，这些问题已成为制约城市现代化发展与经济发展的瓶颈[12]。据公安部统计，截至 2020 年底，我国机动车保有量达 3.72 亿辆，其中汽车 2.81 亿辆。世界银行的一项研究表明，伦敦的公共交通系统在环境治理方面每年可减少大约 4.3 万吨的碳排放。美国公共交通协会（American Public Transportation Association）统计，居住在公共交通服务区的人每年仅在减少拥堵方面就节省了 8.65 亿小时的出行时间和 4.5 亿加仑的燃料。一个经常使用公共交通工具的家庭每年可节省超过 9700 美元[13]。道路交通是人类社会活动的必要环节，然而城市道路建设的速度赶不上机动车数量的增长速度，交通供给与需求严重不平衡，道路拥堵仍是当前社会亟待解决的重要问题[14]。公共交通（如公共汽车、地铁）出行作为一种安全、实惠、方便的出行方式不仅降低了碳排放，还缓解了交通

拥堵[15]。虽然使用公共交通可以带来诸多好处，但我国公共交通系统还存在着不足之处，还有很大的改进空间。良好的公交系统对于城市可持续发展以及人民生活质量有诸多益处。

以往的公共交通规划主要有两种方法，一是调查了解人们在出行时选择不同的交通方式[16][17],这种方法不但花费了大量时间和成本，而且调查过于静态性，无法反映出城市地区的发展；二是通过需求感知路线规划出新的线路[13] [18][19]，这种方式虽可以动态的反应城市的发展变化，且成本较低，但大多需要建造新的公交车站。然而对于像北京市这样覆盖面很广的城市来说，修建新的公交车站不仅是非刚需的而且还会增加额外成本。相比之下，利用道路网络的连通性，可以在不产生任何额外的成本情况下实现规划出新的线路。因此，对于公交网络的优化显得格外的必要。其次，除了满足乘客的通勤需求外，理想的公交网络还应尽可能地连通以方便乘客换乘，即新的公交路线应与现有路线连接良好，以便提供更多换乘选择。

随着云计算、数据挖掘、人工智能等新技术的发展，从低密度价值的轨迹数据中挖掘知识的新方法不断产生，但现有的方案在面对大规模数据集时表现平平，主要原因及挑战如下：

数据质量：公交网络优化需要大量的实时数据和历史数据，包括车辆位置、乘客出行、客流量、车速等。但是，这些数据往往存在精度低、采样不均等问题，影响了优化方案的准确性和可靠性。

算法效率：公交网络优化需要考虑很多因素，如客流量、路网、班次等，因此模型非常复杂，且涉及到大量的复杂计算和优化问题，如路径搜索等。这会增加研究成本和实施难度。但是，现有的算法往往存在效率低、收敛速度慢等问题，导致优化方案难以实现实时性和高效性。

用户体验：公交网络优化需要综合考虑乘客出行需求、舒适度等因素，以提高用户体验。但是，现有的优化方案往往过于注重效率和成本控制，忽视了乘客出行的个性化和舒适度，影响了用户体验。

由于现有的公交网络方法无法有效应对上述挑战，因此基于轨迹数据的公交网络优化研究具有非常重要的价值。

§1.2 国内外研究现状

§1.2.1 轨迹聚类模型

轨迹聚类是指将一组轨迹划分为多个类别的过程，目的是发现轨迹间的相似性和差异性。轨迹聚类模型主要分为无监督聚类模型、半监督聚类模型、有监督聚类模型。

（1）无监督聚类模型

无监督聚类模型是一种基于数据本身的结构特点，将数据点自动分组的方法，不需要任何预定义的类别标签或指导信息。无监督聚类模型的目标是发现数据集中的固有结构，例如密度区域、聚类中心等。以下总结几种常用的无监督聚类方法。

将一个簇定义为一个密度较高的区域，该区域与其他区域相对密度较高，这就是密度聚类。密度聚类可以自动发现任意形状的簇，并且不需要预先指定聚类的数量。基于此思想，DBSCAN[20]被提出来之后，被广泛的应用在轨迹聚类研究中。但DBSCAN 对参数却十分的敏感，Mean Shift[21]算法则正好可以解决 DBSCAN 对邻域半径敏感问题，只要确定核函数的宽度，即使邻域半径的轻微变化也并不会影响聚类结果。K-means[22]通过不断地迭代优化聚类中心点和数据点分配的过程来达到目标。K-means 聚类算法在每一次迭代时，需要计算所有数据点与所有聚类中心的距离，这样的计算量非常大，尤其是在处理大规模数据集时。因此 Kumar 等人提出 k-means++[23]来对聚类中心优化，通过计算每个数据点到已有聚类中心的最小距离，再随机选取距离大于这个最小距离的一个点作为新的中心，从而达到更优的效果。

将数据点组织成一个树形结构，其中每个叶节点表示一个单独的数据点，而每个内部节点表示由其子节点组成的数据点集合的聚类，这就是层次聚类。层次聚类一般分为：凝聚型和分裂型。这两种方法不同在于，前者是自底向上从单个数据点开始，不断将距离最近的数据点合并成一个聚类，直到达到聚类数目为止；后者是自顶向下从所有数据点开始，将其划分为两个或多个子聚类，然后对每个子聚类递归地执行同样的过程，直到达到聚类数目为止。在凝聚型聚类思想上，为更准确地捕捉聚类结构，提出了基于重心的凝聚型聚类算法 AGCG（Agglomerative Clustering with Centers of Gravity）[24]；在分裂型的框架下，提出了一种应对大规模轨迹聚类的方法，首先将数据划分为不同的子集，并将分裂操作应用于每个子集中的数据，从而加速聚类过程[25]。

（2）半监督聚类模型

半监督聚类是一种聚类方法，它结合了有标签和无标签的数据，并且使用小部分已经被标记的数据就可以工作。半监督聚类的目标是通过利用已经标记的数据，提高聚类的准确性。该方法的基本思想是将已标记数据用于约束聚类的结果，并尽可能地使未标记数据与已标记数据相似。

通过有标记数据来指定数据点的标签，并将其嵌入到低秩矩阵中，然后，使用平滑性约束来对低秩矩阵进行优化，以便更好地利用有标记和无标记数据的信息[26]。使用两个神经网络模型来学习特征表示和特征相似度，利用相似度来进行聚类，不但提高了聚类的效果，还可以学习到更好的特征表示和特征相似度。

（3）有监督聚类模型

有监督聚类模型需要同时考虑无标签数据和标签数据，通过将标签数据的信息纳

入到聚类过程中，来改进无标签数据的聚类结果。在对监督模型理解上，有监督模型更可能地使得聚类结果优于半监督模型，同样具备较小计算资源需求的优势。以下是对常见几种有监督聚类算法的总结。

顶点划分算法[27]（Vertex Partitioning Algorithm，VPA），它将聚类问题转化为二分图划分问题，并利用已有类别信息指导聚类过程。它通常用于解决具有多个属性或特征的数据的聚类问题。顶点划分算法将数据集中的每个数据点表示为一个顶点，并将数据点之间的相似性表示为边。然后，算法将问题转化为将这个二分图划分为两个独立的子图的问题，其中一个子图包含所有属于同一类别的顶点，另一个子图则包含属于不同类别的顶点，通常使用最大权独立集合算法求解，即找到一个最大权的独立集合，该集合包含尽可能多的同类别的顶点，并且两个集合之间的权重最大。VPA 算法的优点是计算复杂度相对较低，因为它只需要找到一个最大权独立集合即可。但其对噪声和异常值比较敏感，且难以处理高维稠密数据集。

神经网络是一种模拟人类大脑中神经网络的人工系统。Zhang 等人提出 DeepMove[28]将轨迹数据转化为二维地图上的图像，然后使用卷积神经网络进行特征提取，最后使用 K-Means 算法对特征进行聚类。Lv 等人采用卷积神经网络对轨迹进行建模，使用对比损失函数来训练网络，并使用聚类算法对轨迹进行聚类[29]。将轨迹矢量化表示后，一种基于自注意力机制的轨迹聚类将每个轨迹点表示为一个向量，并使用向量之间的点积来计算它们之间的相似度，使用这些相似度计算一个加权平均值，用于更新每个轨迹点的向量表示[30]。

§1.2.2 公交网络优化

公交网络优化是城市规划和交通领域的一个研究领域，主要解决如何设计新的公交线路网络或如何在现有网络中重新设计公交线路的问题[31][32]。该问题是一个复杂的、非线性的、非凸的、多目标的 NP-难问题[33]。其重点在于在运营和资源约束下，优化代表公交网络效率的若干目标的结果，以满足用户需求。公交网络优化的一般过程包括用户出行需求的生成、公交站点及路线网络的建设、制定有目标和约束的优化模型、建设候选公交线路，以及最终解的计算。

（1）传统公交网络优化

主要考虑从人口普查和家庭出行调查中收集的客流和用户需求[16][17]。在一般调查中，会获得多种类型的信息，例如起点-目的地、运输方式和距离、出行目的、出行选择的路线、已付车费、支付类型、按时间划分的使用频率，以及社会经济和态度因素[32]。由于大城市地区的相关调查通常每次花费大量的人力和财力，因此通常的做法是每几年进行一次此类调查。

传统公交网络优化的目标和约束条件大致为：最短距离、最短出行时间、最大客

流和最大区域覆盖等[17]。基于此目标，人们提出了多种方法来制定和解决公交线路网络优化问题[17][34]。Fan 和 Machemehl[35]将公交线路网络设计问题表述为一个多目标非线性混合整数模型。然后结合 Dijkstra 的最短路径算法和 Yen 的 k-最短路径算法[36]，生成所有候选路径。最后，利用遗传算法进行路径选择。Ceder 和 Wilson[37]将出行时间视为一个约束条件，并构建了最小化需求差异和最短路径的路线。Chakroborty 和 Dwivedi[38]使用了需求驱动的节点修正方法，他们估计每个网络节点的入境和出境乘客，并用需求来指导路线的建设。

（2）需求感知路线规划

与传统通过调查的路线规划相比，需求感知路线规划旨在有效发现通勤者的及时需求，其方法可分为两组：网络（重新）设计；特定路线优化或添加新路线。

具体而言，Chen 等人[39]利用夜间出租车轨迹，首先检测包含频繁接送的区域，然后将其划分为多个集群，并将每个集群中的一个位置确定为候选公交站点，最后通过两种启发式方法从所有候选公交路线中确定热门路线。Pinelli 等人[40]通过推导频繁的移动模式并在现有站点规划新路线，在不考虑换乘和连通性的情况下，基于手机塔的移动轨迹重新设计了整个公交网络。Liu 等人[13]建议发现运营不好的路线，并根据从出租车和公共汽车行程记录中提取的热门起始目的地对，通过优化现有交通网络中的路线，而不是摧毁现有的交通系统。轨迹上的反向 k 最近邻（RkNNT）[19]是一种基于现有公交网络中的轨迹数据估计公交路线需求的工具，用于规划两个给定站点之间的最大通行能力的路线。

（3）基于连通性路线优化

作为衡量公交网络换乘便利性的关键指标之一[17][41]，连通性在交通领域得到了广泛的应用和研究。通过将公交网络建模为由顶点和边组成的无向图，可以用多种方式测量连通性，例如顶点和边连通性[42]、代数连通性[43]和自然连通性[44]（也称为 Estrada 指数的扩展），这些更适合于复杂网络。在这些措施中，自然连通性可以说是公交网络最合适的一种，因为它不会因较小或较大的图形变化而出现剧烈变化。相反，它可以随着更多的修改而单调变化。

由以上分析可知，当前的公交网络优化主要是基于静态的调查分析或者是通勤需求感知规划，都没有考虑网络的连通性。尽管 Peng 等人[43]尝试过考虑网络连通性来优化路线，但是旨在添加多条离散的边来满足需求，且并没有考虑通勤需求，无法切合我们的需求。因此，将网络连通性考虑进来是一项非常有意义的事。

§1.3 论文研究内容

本文研究的主要内容是基于大规模轨迹数据聚类下的公交网络优化，主要为了解决现有的公交网络优化方案中无法适应复杂网络以及过分偏重传统指标等问题。本文