面向大规模图数据的关键词覆盖最优路径规划方法

李梓杨[1]，陈鹏程[1]，于炯[1,2]，蒲勇霖[3]，何贞贞[2]，李雪[2]，郑世杰[1]

（1. 新疆大学软件学院，新疆 乌鲁木齐 830002；2. 新疆大学信息科学与工程学院 新疆 乌鲁木齐 830017；
3. 南京信息工程大学软件学院，江苏 南京 210044）

摘 要：针对个性化自驾游路径规划中存在规划路径无法满足不同用户个性化需求的问题，提出了基于不同用户兴趣点的关键词覆盖最优路径规划方法。首先，建立路网信息预处理模型并通过路网信息预处理算法绘制路网信息查询图；其次，使用倒排索引算法根据用户设定的个性化需求对路网信息查询图进行剪枝，在减小大规模数据处理内存开销的同时提升了关键词覆盖最优路径规划方法的执行效率；最后，通过双向并行拓展方式的关键词覆盖最优路径拓展算法实现满足用户兴趣点的个性化旅游路径推荐。实验结果表明，关键词覆盖最优路径规划方法不仅实现了满足用户个性化需求的路径规划，而且通过剪枝和双向并行拓展的方式提高了方法的执行效率。

关键词：图数据；路径规划；动态规划；倒排索引算法；双向并行拓展

中图分类号：TP391

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023171

Keyword-aware optimal route planning method for large-scale graph data

LI Ziyang[1], CHEN Pengcheng[1], YU Jiong[1,2], PU Yonglin[3], HE Zhenzhen[2], LI Xue[2], ZHENG Shijie[1]

1. School of Software, Xinjiang University, Urumqi 830002, China
2. School of Information Science and Engineering, Xinjiang University, Urumqi 830017, China
3. School of Software, Nanjing University of Information Science and Technology, Nanjing 210044, China

Abstract: Focused on the problem that the planned routes cannot meet the personalized demand of different users in route planning of personalized self-driving tour, a keyword-aware optimal route planning method based on different user interests was proposed. Firstly, the road network information preprocessing model was set up and the road network information query graph was built by the road network information preprocessing algorithm. Secondly, the inverted index algorithm was proposed to prune the road network information query graph according to the personalized requirements from users, which improved the execution efficiency of keyword-aware optimal route planning method and reduced the memory cost of large-scale data processing effectively. Finally, the keyword-aware optimal route planning algorithm was proposed to realize personalized recommendation according to user interest by bidirectional parallel extension. The experimental results show that the method not only realizes the route planning to meet the individual needs of users but also improves the execution efficiency of the method through pruning and bidirectional parallel extension.

Keywords: graph data, route planning, dynamic programming, inverted index algorithm, bidirectional parallel extension

---

收稿日期：2023−04−12；**修回日期：**2023−07−04

通信作者：陈鹏程，cpc@stu.xju.edu.cn

基金项目：国家自然科学基金资助项目（No.62262064, No.62266043, No.61966035）；新疆维吾尔自治区重点研发计划基金资助项目（No.2022295358）；新疆维吾尔自治区自然科学基金资助项目（No.2022D01C56）；新疆大学博士研究生创新基金资助项目（No.XJU2022BS072）

Foundation Items: The National Natural Science Foundation of China (No.62262064, No.62266043, No.61966035), The Key Research and Development Project in Xinjiang Uygur Autonomous Region (No.2022295358), The Natural Science Foundation of Xinjiang Uygur Autonomous Region (No.2022D01C56), Xinjiang University Doctor Postgraduate Innovation Project (No.XJU2022BS072)

0 引言

随着旅游行业的不断发展，高质量个性化旅游成为人们关注的焦点。为了提高自驾游的旅行质量和用户体验，在出行前做好旅游路径的规划是必要的。旅游路径规划问题是针对用户设定多个访问兴趣点（PoI, point of interest）的旅行计划问题[1]，其中，精确算法（如 Dijkstra 算法等）虽然能得到精确解，但其搜索能力较差；启发式算法能够在短时间内得到可行解，但解的质量并不高，只适用于小规模的 PoI 规划[2-3]。因此，智能化元启发算法（如遗传算法[4]等）成为解决该问题的主流方法，但未应用于大规模图数据。图数据指由顶点和边组成的图结构，同时包含顶点和边的坐标等矢量信息。旅游路径规划还包括一类特殊路径规划，能够为地图服务旅游行程中的路径规划提供有效支持，这类路径查询问题被定义为基于关键词覆盖的最优路径查询（KORS, keyword-aware optimal route search），综合考虑路径关键词覆盖条件、路径行程代价约束以及路径流行度三类因素间的组合优化性[5]。现阶段的算法策略是分段拓展策略，即将路径划分为多段后分别进行拓展操作，但各个分段路径的关联度较高，拓展未能并行化处理[6]。

针对上述问题，本文提出了面向大规模图数据的关键词覆盖最优路径规划方法，解决游客从某点出发，途经游客兴趣点，最终到达目的地的最优路径规划问题，且该路径应当满足游客的行程代价，从而提高路径规划效率。

本文的主要贡献如下：提出基于动态规划思想的关键词覆盖最优路径规划方法，实现满足用户个性化需求的旅游路径规划；同时，以高效的数据处理方法减小计算资源的消耗，并为动态推荐旅游路径奠定基础。

1 相关研究

根据对环境信息理解的不同，路径规划可分为环境信息完全已知的全局路径规划、环境信息部分未知或完全未知的局部路径规划[7]。针对此类问题，现阶段的研究主要有传统算法路径规划问题、智能化元启发算法路径规划问题和基于关键词路径规划问题。

传统算法策略路径规划问题的研究侧重于环境信息完全已知的全局路径规划，其研究的约束条件为路径代价，通过拓展找寻最短路径，是研究旅游路径规划问题的重要理论基础，如 Dijkstra 算法[2]、Floyd 算法[8]、A*算法[9]和 Prim 树算法[10]。其中，Floyd 算法是利用动态规划的思想寻找给定的加权图中多源点之间最短路径算法；Dijkstra 算法、A*算法和 Prim 树算法都基于贪心策略。Dijkstra 算法解决有权图中最短路径问题，但其搜索能力较差，双向搜索有利于提升搜索效率和算法性能；A*算法是一种静态路网中求解最短路径最有效的直接搜索方法；Prim 树算法用于解决无向图中最小生成树问题。

智能化元启发算法路径规划问题的研究侧重于环境信息部分未知的局部路径规划，其研究的约束条件为路径代价、兴趣点。解决思路主要有以下 2 种。1) 采用单个算法[11-12]的智能化元启发算法，典型算法有粒子群优化（PSO, particle swarm optimization）算法[11]、蚁群优化（ACO, ant colony optimization）算法[12]。以上智能算法总体来说在解决环境信息部分未知时有效，但普遍存在算法搜索时间长、容易陷入局部最优等问题。2) 采用 2 种或 2 种以上算法相结合的智能化元启发算法。文献[13]提出了一种融合寻路算法，极大地避免了陷入局部最优，并得到了一条平滑的路径，其时间效率比传统算法提升 60%，比单个智能算法提升 70%。研究表明，2 种或 2 种以上算法相结合的智能化元启发算法有效解决了路径规划时遇到的陷入局部最优解问题。

基于关键词路径规划问题的研究成果实现了覆盖用户兴趣点的路径规划。首先，面向关键词路网查询，文献[14]开发了一种高效的索引结构，通过检索与查询相关的数据对象来有效地精简搜索空间，实验证明了该方法的高效性；此类查询为后续的路径规划奠定了基础。此类研究偏向环境信息完全已知的全局路径规划，研究的约束条件为路径代价、兴趣点、路径流行度。文献[15]为了解决处理大规模路网信息时内存开销急剧上升，导致路网不能处理的问题，提出一种大规模路网图下关键词覆盖最优路径查询（KORL, keyword-aware optimal route query on large-scale road network）算法，但查询效率仍需提高，预处理方法也可被替代。文献[16]为了解决搜索规模会随着搜索深度的增加而呈指数级增长的问题，提出了一种针对关键词覆盖最优路径查询的分段并行展开算法，在几乎不损失精度的情况下显著缩短了执行时间。本文研究属于此类基于关键词路径规划研究。

目前，现有的研究成果大多面临以下 3 个问题：1) 面对大规模图数据，直接进行路径规划会占用过多的内存资源，需要减小内存的消耗；2) 现有路径规划算法的执行效率仍有待进一步提升；3) 现有研究大多只适用于普通自驾的路径规划，无法实现满足更多用户个性化需求的旅游路径推荐。

针对上述问题，本文提出一种面向大规模图数据的关键词覆盖最优路径规划方法。本文与现有研究成果的不同之处总结如下。

1) 现有研究成果大都适用于自驾游路径规划，但传统的路径规划无法满足不同用户的个性化需求。本文通过引入倒排索引算法，实现了满足用户兴趣点的个性化旅游路径推荐；同时本文提出了关键词倒排索引信息树，为动态推荐提供了实现方法。

2) 现有研究成果大多采用智能化元启发算法路径规划，但面向大规模图数据路网时存在内存消耗过大的问题。本文构建了路网信息查询图以及剪枝方法，提出了路网信息预处理算法，有效减少了计算资源的消耗，提升了方法的性能。

3) 现有研究成果大多采用智能化元启发算法路径规划，但面向大规模图数据路网时存在执行时间过长的问题。本文通过双向并行拓展的方式以及倒排索引结构，有效降低了时间开销。

2　问题建模与分析

为了解决自驾游过程中途径兴趣点的个性化路径规划问题，本节建立了路网信息预处理模型、关键词倒排索引信息树模型以及关键词覆盖最优路径拓展模型。其中，路网信息预处理模型描述了路网信息图中节点与边之间的关系，为后续的路网信息查询图奠定了基础；关键词倒排索引信息树模型通过顶点上的关键词索引集合，将相对应的边存储至叶子节点，以此降低路径关联度，并根据游客兴趣点的选择进行相对应的剪枝，为后续的路径拓展任务奠定了基础；关键词覆盖最优路径拓展模型将剪枝后的路网信息查询图进行相对应的拓展，实现路径规划。

2.1　路网信息预处理模型

全国道路信息矢量图描述如下，将全国道路信息矢量数据集导入地理信息系统，可以查阅数据集中的节点坐标、道路类型和道路名称、道路坐标、路径长度等，以及通过对道路的拓扑得到相应的节点信息。

定义 1　路网信息图。由实际全国道路信息矢量图经拓扑处理后得到的。路网信息图 $G(V,E)$由一个无向图的顶点集 V 和边集 E 构成，其中，顶点 $v \in V$ 表示路网信息图的一个顶点，顶点上的兴趣本质是一个空间文本对象，既包含空间经纬度信息 $v.\mathrm{geo}$；又包含对该点兴趣的文本描述关键词集合 $v.\psi=\{v.K_1,v.K_2,\cdots\}$。边 $e \in E$ 表示 2 个邻接顶点 v_i,v_j 之间的边，记为(v_i,v_j)。每条边包含一个属性值表示对应边的代价，记为 $\mathrm{BS}(v_i,v_j)$。对于任意一条边 e，代价值表示其路径长度。

为了方便筛选包含特定关键词的顶点，根据路网信息图中的关键词与顶点的对应关系构建关键词倒排索引，如表 1 所示。

表 1　　**关键词倒排索引**

关键词	包含关键词顶点集
K_1	v_5
K_2	v_3
K_3	v_4,v_1
K_4	v_2,v_6,v_7

由定义 1 可知，路网信息图阐述了图中顶点和边的关系，描述了每个顶点所包含的关键词集合，以及每条边的权重的含义，为路网信息查询图的建立以及关键词倒排索引信息树的构建奠定了基础。

定义 2　路网信息查询图。分别选取距离出发点和目的地较近的 N 个顶点，得到 N^2 条路径。路网信息查询图 $M(V,E)$由顶点集 V 和边集 E 构成，通过路网信息预处理得到对应的路网信息查询图。其中，当顶点 v_i 向另一个顶点 v_j 拓展时，路径的代价值会进行更新，即 $\mathrm{BS}=v_i.\mathrm{BS}+\mathrm{BS}(v_i,v_j)$，对应的关键词集合也会更新，即 $v.\psi=v_i.\psi \cup v_j.\psi$。路径 $P=(v_s,v_1,\cdots,v_t)$则表示由 v_s 出发，途径 $v_1,v_2,v_3,\cdots$，直到 v_t 的一条路径，该路径的代价值为 $\mathrm{BS}(P)=\sum_{i=1}^{n}\mathrm{BS}(v_{i-1},v_i)$，路径对关键词的覆盖 $P.\psi=\bigcup_{v\in P}v.\psi$。因此，通过对路网信息查询图顶点和边进行抽象，可得到如图 1 所示的路网信息查询图。

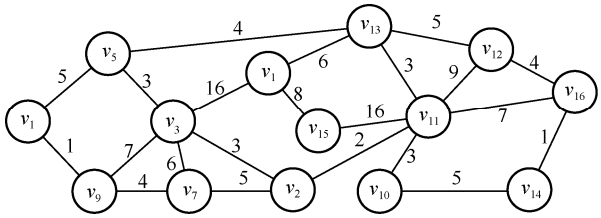


图 1　路网信息查询图

定义 2 描述了路网信息查询图中顶点和路径之间的关系，以及路径代价值和路径关键词覆盖。

根据定义 1 和定义 2 的描述可知，路网信息查询图是通过路网信息预处理模型获取的。首先，将代价值 BS 初始化为无穷大，正向搜寻为从 v_s 开始，扫描弧为 $L(n,m)$，正向搜索的标签为 d_{f}，此段的代价总和为 $d_{\mathrm{f}}.\mathrm{BS}=\sum_{s}^{n}\mathrm{BS}(v_i,v_j)$；反向搜寻为从 v_t 开始，扫描弧为 $L(m,n)$，反向搜索的标签为 d_{r}，此段的代价总和为 $d_{\mathrm{r}}.\mathrm{BS}=\sum_{t}^{m}\mathrm{BS}(v_i,v_j)$，其中 $n,m\in V$。正向搜寻和反向搜寻分别在正向表和反向表中进行，同时，使用小根堆存储路径，顶堆的路径代价值分别为 $\mathrm{top_f}$，$\mathrm{top_r}$。向前搜索扫描时，当扫描到弧 $L(n,m)$并且 m 在反向扫描中时，更新代价值 BS，即

$$d_{\mathrm{f}}.\mathrm{BS}+L(n,m).\mathrm{BS}+d_{\mathrm{r}}.\mathrm{BS}<\mathrm{BS}_{\max} \tag{1}$$

反向搜索同理。搜索相遇如图 2 所示。

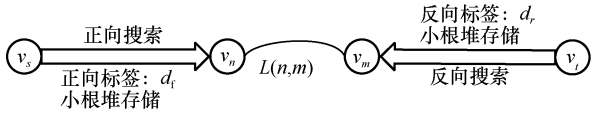


图 2 搜索相遇

定理 1 路径搜索定理。设有路网信息图 G，其中顶点 $v\in V$。在路网信息图 G 中，当开始搜寻时，$\mathrm{top_f}$ 和 $\mathrm{top_r}$ 分别表示正向表和反向表中堆顶元素的路径代价值，若顶点 v 在路径上且满足

$$\mathrm{top_f}+\mathrm{top_r}<\mathrm{BS}_{\max} \tag{2}$$

则双向搜索一定相遇；反之，若顶点 v 不在路径上且满足

$$\mathrm{top_f}+\mathrm{top_r}\geqslant\mathrm{BS}_{\max} \tag{3}$$

则双向搜索一定无法相遇。

证明 详见附录 1。

定理 1 描述了构建路网信息查询图时双向并行搜寻相遇时的判定规则，为后续关键词路径拓展奠定基础。此外，路网信息预处理算法使用了此判定规则。

2.2 关键词倒排索引信息树模型

通过建立路网信息预处理模型，可以根据出发点、目的地的不同得到不同的路网信息查询图，为后续的个性化路径提供了一定的基础，同时，也减少了一部分的内存开销。因此，在关键词倒排索引信息树模型中，借助预处理模型得到的路网信息查询图，进行下一步的存储与剪枝。

定义 3 关键词倒排索引信息树。根据关键词集合 ψ（如表 1 所示），将路网信息查询图 M（如图 1 所示）中的路径按照关键词划分，根据划分结果构建关键词倒排索引信息树。以路网信息查询图为根节点，子节点存储关键词，叶子节点存储包含该关键词的 N 条路径。关键词倒排索引信息树如图 3 所示。

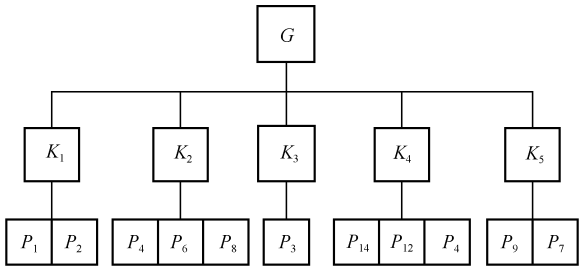


图 3 关键词倒排索引信息树

由定义 3 可知，关键词倒排索引信息树的基础模型阐述了各节点之间的关系以及节点中存储的内容，为后续的个性化路径以及路径剪枝提供了支持，同时为关键词覆盖最优路径奠定了基础。

定义 4 关键词顶点。若存在一个顶点 v_m 和关键词 k_i，满足 $k_i\in v_m.\psi$ 且 $k_i\in Q.\psi$，则称顶点 v_m 是查询 Q 的一个关键词顶点。给定查询 $Q(v_s,v_t,\psi)$，关键词倒排索引信息树按照查询中的起点与终点构建成一棵完整的查询树，叶子节点中保存的路径按照查询 Q 中关键词的选择进行路径的初步剪枝。

由定义 3 和定义 4 可知，根据给定的查询 Q，可以实现路网信息查询图的初步剪枝，这个查询为动态路径推荐提供了基础。当路网信息查询图存储至关键词倒排索引信息树中时，其叶子节点须按以下要求存储：通过关键词倒排索引信息表，将每个顶点兴趣 v_m 按照关键词分类，将其包含的路径 P 存储在相应的叶子节点中，在每个叶子节点中需要选取合适的存储方案，即满足

$$\mathrm{BS}(P_{i+1}(v.k_i))<\mathrm{BS}(P_i(v.k_i)) \tag{4}$$

由式(11)可知，每个叶子节点中存储的路径 P 是按路径代价的升序方式排列的，其每个关键词下的节点可以选取合适的存储方案。

2.3　关键词覆盖最优路径拓展模型

关键词倒排索引信息树模型为个性化路径规划提供了媒介。在路网信息查询图中按照需求进行选择，未选择的兴趣所包含的路径被初步剪枝，剪枝后的路网信息查询图进行下一步的关键词覆盖最优路径拓展。

定义 5　关键词路径。在满足定义 4 的前提下，路径 P 不能被其他路径所替代。通过关键词倒排索引信息树剪枝后的路径满足关键词路径。关键词路径如图 4 所示，其中，三角形表示查询 Q 中的关键词，圆形表示未在查询 Q 中的关键词。

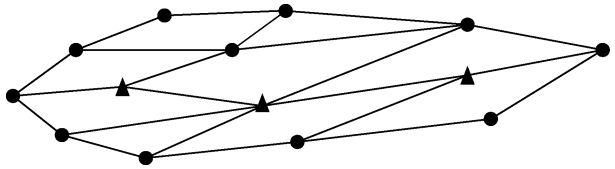


图 4　关键词路径

定义 6　关键词路径拓展剪枝。将满足关键词路径的路网信息查询图通过关键词覆盖最优路径拓展模型进行处理，最终会得到一条自驾游推荐路径。关键词路径拓展剪枝如图 5 所示，其中虚线表示已被剪枝。

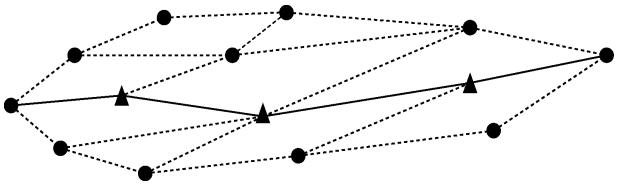


图 5　关键词路径拓展剪枝

由定义 5 和定义 6 可知关键词路径的基本信息。结合定义，在路径拓展时需要进行如下操作。定义以下 2 个功能函数，向前搜索的函数为 $\pi_{\mathrm{f}}(v)$，表示对 v_s 开始搜寻到 v_n 的估价；反向搜索的函数为 $\pi_{\mathrm{r}}(v)$，表示对 v_t 开始搜寻到 v_m 的估价。同时保证这 2 个函数在搜索任意弧 $L(n,m)$ 满足 $\pi_{\mathrm{f}}+\pi_{\mathrm{r}}=\mathrm{constant}$。满足这个要求需要定义一致的势函数，表示为

$$P_{\mathrm{f}}(v)=\frac{1}{2}(\pi_{\mathrm{f}}(v)-\pi_{\mathrm{r}}(v))+\frac{\pi_{\mathrm{r}}(t)}{2} \tag{5}$$

$$P_{\mathrm{r}}(v)=\frac{1}{2}(\pi_{\mathrm{r}}(v)-\pi_{\mathrm{f}}(v))+\frac{\pi_{\mathrm{f}}(s)}{2} \tag{6}$$

当 π_{f} 和 π_{r} 为下限时，$P_{\mathrm{f}}(t)=0$，$P_{\mathrm{r}}(s)=0$，正向搜索和反向搜索分别在各自的表中进行，同时，使用小根堆存储，设 v_{f} 和 v_{r} 为 2 个小根堆的堆顶元素。正向搜索的路径长度为

$$d_{\mathrm{f}}(v_{\mathrm{f}})+P_{\mathrm{f}}(v_{\mathrm{f}})-P_{\mathrm{f}}(s)=\mathrm{top}_{\mathrm{f}}-P_{\mathrm{f}}(s) \tag{7}$$

反向搜索路径长度为

$$d_{\mathrm{r}}(v_{\mathrm{r}})+P_{\mathrm{r}}(v_{\mathrm{r}})-P_{\mathrm{r}}(t)=\mathrm{top}_{\mathrm{r}}-P_{\mathrm{r}}(t) \tag{8}$$

当正向搜索与反向搜索相遇，即满足式(9)时，搜索即可停止。

$$\left[\mathrm{top}_{\mathrm{f}}-P_{\mathrm{f}}(s)\right]+\left[\mathrm{top}_{\mathrm{r}}-P_{\mathrm{r}}(t)\right]\geqslant\left[\mathrm{BS}-P_{\mathrm{f}}(s)+P_{\mathrm{f}}(t)\right] \tag{9}$$

式(9)可以简化为

$$\mathrm{top}_{\mathrm{f}}+\mathrm{top}_{\mathrm{r}}\geqslant\mathrm{BS}+P_{\mathrm{r}}(t) \tag{10}$$

为了解决个性化自驾游中路径规划的问题，本文提出了路网信息预处理模型、关键词倒排索引信息树模型以及关键词覆盖最优路径拓展模型。模型与模型之间相辅相成，每个模型的输出结果为下一个模型的输入条件。下面，基于以上模型提出关键词覆盖最优路径规划方法，以实现满足不同用户个性化需求的关键词覆盖旅游路径规划。

3　关键词覆盖最优路径规划方法

本节在上述模型的基础上提出了关键词覆盖最优路径规划方法，该方法包含了路网信息预处理算法、倒排索引算法和关键词覆盖最优路径拓展算法，通过对大规模路网信息图的预处理得到路网信息查询图，降低内存的占用，然后对路网信息查询图进行进一步的处理，最终得到最优路径，该方法主要分为以下 3 个步骤，具体流程如图 6 所示。

步骤 1　将路网信息矢量图导入数据库中，进行拓扑处理，确保任意两点间的可达性；然后运用

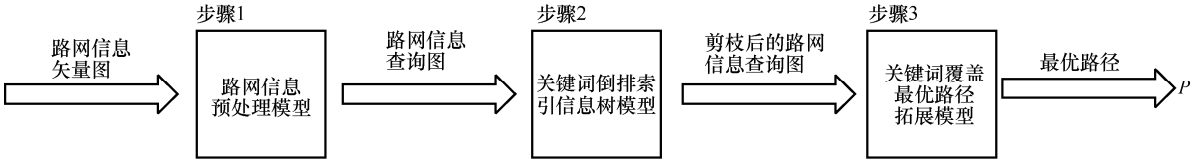


图 6　关键词覆盖最优路径规划方法流程

路网信息预处理算法得到路网信息查询图。

步骤 2 根据关键词信息将得到的路网信息查询图存储于关键词倒排索引信息树，再根据相应查询 Q 进行剪枝，得到符合查询 Q 的路网信息查询图。

步骤 3 依据关键词倒排索引信息树剪枝后得到的路网信息查询图，进行后续的关键词路径拓展，最终得到最优路径。

3.1 路网信息预处理算法

为了解决路网信息图规模庞大导致内存资源消耗过大，进而产生大规模数据无法处理的问题，本文设计了路网信息预处理算法。根据路网信息预处理模型设计关键词覆盖的双向 Dijkstra（Kbd_stra, keyword-aware the bidirectional Dijkstra）算法。首先，根据路网信息图的可达性，进行双向拓扑查询；然后，找到起点、终点各自邻近的 5 个顶点，并使用 Kbd_stra 算法进行路网信息查询图的绘制。其选取依据为人们日常出行时所能考虑周边最近出发点的个数，具体步骤如算法 1 所示。

算法 1 路网信息预处理算法

输入 路网信息图 $G(V,E)$，用户设定的查询 $Q(v_s,v_t,\psi)$

输出 路网信息查询图 $M(V,E)$

1) $M.V \leftarrow \varnothing$
2) $M.E \leftarrow G.E$ /*G 为路网信息图，E 为边集*/
3) for each $e \in G.E$ /*执行拓扑过程*/
4) if ! reachable (e)
5) $G.E \leftarrow G.E-e$
6) else
7) e.BS = st_length(geom) /*BS 为路径代价*/
8) end if
9) end for
10) γ = nearest($Q.v_s$,5)
11) τ = nearest($Q.v_t$,5)
12) for each $v_0 \in \gamma$, $v_n \in \tau$ /*根据定理 1，执行 Kbd_stra 算法进行路网信息查询图路径绘制*/
13) if $\text{top}_\text{f} + \text{top}_\text{r} \geqslant \text{BS}_{\max}$
14) (v_i,v_j).BS= ∞
15) end if
16) end for
17) for each v_i, $v_j \in G.V$ /*路径绘制时的顶点信息*/
18) if (v_i,v_j).BS $< \infty$
19) $M.V \leftarrow M.V \bigcup \{v_i, v_j\}$
20) end if
21) end for
22) return $M(V,E)$

算法 1 的输入为 $G(V,E)$，用户设定的查询为 $Q(v_s,v_t,\psi)$，输出为 $M(V,E)$。第 1)～2)行表示输入的路网信息图；第 3)～9)行表示可达性路径拓扑的过程，这个过程的条件为当 2 条或多条道路形成一个交叉路口时，它必须是该交叉路口处的一个节点，并计算该路段的路径长度；第 10)～11)行表示根据用户给定的查询，找到起点、终点各自邻近的 5 个点进行路网信息查询图绘制，并根据定理 1 判断双向查询过程是否能够相遇，即在路网信息查询图中判断起点和终点的连通性；第 12)～21)行表示绘制路网信息查询图时的顶点信息加入路网信息查询图中，最终得到路网信息查询图。

算法 1 的核心目标是根据用户给定的查询条件，由路网信息图生成对应的路网信息查询图，以满足用户给定的查询请求。面向大规模图数据的关键词覆盖最优路径规划方法，通过 Kbd_stra 算法快速构建路网信息查询图，减小计算资源，其相遇条件满足式(1)。

在时间复杂度方面，路径拓扑的时间复杂度为 $O(n)$，在绘制路网信息查询图时，使用 Kbd_stra 算法进行路径搜索，最坏情况下的时间复杂度为 $O((V\log V+E))$，因此算法 1 的时间复杂度为 $T(n)=O(n+(V\log V+E))$。

3.2 倒排索引算法

本文采用路网信息预处理算法绘制路网信息查询图，构建关键词倒排索引信息树模型解决推荐个性化自驾游路径问题，根据关键词倒排索引信息树模型提出倒排索引算法。该算法根据兴趣点构建倒排索引结构，得到一棵索引信息树，树的叶子节点保存路网信息查询图中的路径信息，保留前 3 个旅游方案。因此，每个叶子节点至多保留 3 个合适的方案信息，确保后续关键词覆盖最优路径拓展的高效进行。具体步骤如算法 2 所示。

算法 2 倒排索引算法

输入 路网信息查询图 $M(V,E)$，用户设定的查询 $Q(v_s,v_t,\psi)$

输出　剪枝后的路网信息查询图 $M'(V,E)$

1) T=initTree(M)
2) for each $v \in M.V$　/*倒排索引信息树第二层关键词存储*/　/*M 为路网信息查询图，V 为顶点集*/
3)　$T.\text{sub} = T.\text{sub} \cup v.\psi$　/*T.sub 表示关键词倒排索引信息树第二层关键词存储*/
4) end for
5) for each $k \in T.\text{sub}$　/*倒排索引信息树第三层路径存储*/
6)　if $k \notin Q.\psi$
7)　$T.\text{sub} = T.\text{sub} - k$
8)　end if
9)　k.sub = createInvertIndex(k, p.BS, ASC)
10) end for
11) M' = buildGraph(T)
12) return M'

算法 2 的输入为路网信息查询图 $M(V,E)$和用户设定的查询 $Q(v_s,v_t,\psi)$，输出为剪枝后的路网信息查询图 $M'(V,E)$。第 1)行表示初始化树的结构；第 2)～4)行表示在关键词倒排索引信息树第二层中的关键词按照路网信息查询图中的关键词类进行归并，得到关键词种类；第 5)～10)行表示在关键词倒排索引信息树中叶子节点上，按照给定的查询 Q 对关键词倒排索引信息树进行剪枝，并存储对应关键词下的路径，其中第 9)行的升序排序满足式(4)；第 11)行表示得到剪枝后的路网信息查询图 $M'(V,E)$。

算法 2 的核心目标是根据用户给定的查询 Q，进行路网信息图的初步剪枝，由路网信息查询图剪枝后生成满足查询 Q 的路网信息查询图。倒排索引算法的输入为算法 1 的输出结果。倒排索引算法首先根据路网信息查询图构建了倒排索引结构，其次根据查询 Q 进行了进一步的剪枝，为后续的算法奠定了基础，进一步为关键词覆盖的路径规划提供支持，以及为推荐个性化自驾游路径提供了实现方法，提升了方法的性能。

在时间复杂度方面，在关键词倒排索引信息树构建与查询的过程中时间复杂度为 $O(n\log n)$，因此算法 2 的时间复杂度为 $T(n)=O(n\log n + n\log n)$。

3.3　关键词覆盖最优路径拓展算法

为了解决分治法中各分段路径关联度高、拓展时不能并行化处理的问题，提出关键词覆盖最优路径拓展算法。根据关键词覆盖最优路径拓展模型，设计关键词覆盖双向 A*（Kbd_a*, keyword-aware the bidirectional A*）算法。该算法在剪枝后的路网信息查询图 $M'(V,E)$上，从终点和起点进行路径拓展，依据 2 点间的路径距离消耗为路径代价，进行每一段路径的对比，最终得到一条最优规划路径。具体步骤如算法 3 所示。

算法 3　关键词路径拓展算法

输入　剪枝后的路网信息查询图 $M'(V,E)$，用户设定的查询 $Q(v_s,v_t,\psi)$

输出　最优规划路径 P

1)$v_n=v_s$
2)$v_m=v_t$
3)while($\text{top}_f+\text{top}_r$>=BS+$p_r(t)$)　/*根据定义 5 和定义 6，执行 Kbd_a*算法进行路径搜寻*/ /*$p_r(t)$为势函数*/
4)　openList1.clear()
5)　openList2.clear()
6)　if　$v_n \notin$ closeList1　/*正向搜索*/
7)　β=getAllNear (v_n) /*获取 v_n 的所有邻近节点*/
8)　openList1.insert(v_i, P.BS)
9)　openList1.heapfy(P.BS) /*BS 为路径代价*/
10)　u = openList1.removeTop()
11)　closeList1.insert(u)
12)　end if
13)　if　$v_m \notin$ closeList2　/*反向搜索*/
14)　γ=getAllNear (v_m) /*获取 v_m 的所有邻近节点*/
15)　openList2.insert(v_i, P.BS)
16)　openList2.heapfy(P.BS)
17)　r = openList2.removeTop()
18)　closeList2.insert(r)
19)　end if
20)　$v_n=u$　/*正向搜索下一节点*/
21)　$v_m=r$　/*反向搜索下一节点*/
22) end while
23) V = clostList1 $\cup$ clostList2
24)P = buildPath(V)
25)Return P

算法 3 的输入为剪枝后的路网信息查询图

$M'(V,E)$，用户设定的查询为 $Q(v_s,v_t,\psi)$，输出为最优规划路径 P。第 1)～2)行表示输入起点、终点，准备开始路径拓展；第 3)～22)行表示搜索路径的并行化处理的过程，每个过程都需要对路径进行比较、存储，其相遇条件满足式(10)，路径长度满足式(7)和式(8)；第 23)～24)行表示相遇之后，获取最优规划路径。

算法 3 的核心目标是根据剪枝后的路网信息查询图快速规划覆盖关键词路径的路径。关键词覆盖最优路径拓展算法的输入为算法 2 的结果，关键词覆盖最优路径拓展算法使用双向并行化的思想，提高了路径规划的效率，实现更高效的路径规划。

在时间复杂度方面，路径并行化处理的时间复杂度为 $O((E+V)\log V)$，因此算法 3 的时间复杂度为 $T(n)= O((E+V)\log V)$。

时间复杂度分析如下。面向大规模图数据的关键词覆盖最优路径规划方法由 3 个算法构成，其时间复杂度为 3 个算法的时间复杂度总和，总体时间复杂度为 $T(n)=O(n+(V\log V+E)+n\log n+n\log n+(E+V)\log V)$。可以化简为 $T(n)= O(n\log n)$。

4　实验与分析

为了验证关键词覆盖最优路径规划方法的可行性与有效性，本文通过设置对比实验、消融实验，将关键词覆盖最优路径规划方法与传统算法策略及相关研究成果进行对比。通过对处于不同规模下的图数据预处理阶段和关键词个数选择以及关键词覆盖最优路径阶段进行实验设计，验证了路网信息预处理的可行性以及关键词覆盖最优路径拓展算法的有效性，讨论了关键词覆盖最优路径规划方法的优化效果和执行开销。

4.1　实验设置

本文所有实验环境运行在 Intel(R) Xeon(R) Gold 5117 CPU @ 2.00 GHz (2 处理器)的服务器上，采用的编程语言为 Java。具体实验环境参数如表 2 所示。

表 2　　实验环境参数

环境	类别	描述
硬件	CPU	Intel(R) Xeon(R) Gold 5117 CPU @ 2.00 GHz
	内存	256 GB
软件	数据库	PostgreSQL
	编译环境	IntelliJ IDEA2020
	辅助软件	QGIS 3.26.1

实验采用中国全国道路网数据，路网图边集具有路径长度属性。本文通过辅助软件从全国道路网中截取不同规模的路网信息图 G_1～G_4，将路径长度视为方法中的路径代价，并采用 10 000 个关键词随机为路网信息图中各顶点生成关键词描述[17]。数据集具体信息如表 3 所示。

表 3　　数据集具体信息

路网信息图	顶点数/10^3 个	边数/10^3 条	顶点平均关键词数
G_1	230	230	6.81
G_2	2 142	2 670	7.16
G_3	2 670	3 544	7.59
G_4	3 242	4 218	7.4

为了证明关键词覆盖最优路径规划方法的有效性，本文设计对比了主流智能化元启发算法，包括 ACO[18]、PSO[19]。其中，ACO 算法是一种用来寻找优化路径的概率型算法；PSO 算法是一种基于群体协作的随机搜索算法。

4.2　关键词覆盖率对比

本文实验设计与文献[4-5,15]中关键词查询实验相关工作相似；同时，设计了 2 个参数，即查询图规模、图规模，以节点为单位。为了验证本文算法的可行性和有效性，本文分别从关键词覆盖率、预处理开销、查询因素方面设计 6 组实验，验证了关键词覆盖最优路径规划方法的执行效率与性能。

实验 1　关键词覆盖率对比。为验证路径是满足用户查询 Q 的个性化路径，需要设定一个评价指标关键词覆盖率 M，该评价指标为路径所含用户查询关键词个数 $P.\psi$ 与用户查询 Q 中关键词 $Q.\psi$ 个数之比，即 $M=\dfrac{|P.\psi|}{|Q.\psi|}$。为了观察关键词覆盖情况，需要在控制图规模和关键词个数不变的前提下改变查询图规模，评价标准为关键词覆盖率。实验 1 采用 G_3，关键词个数分别为 2、4、6、8。结合消融实验的消融关键词倒排索引信息树观察，比较不同关键词个数下的关键词覆盖率 M，如图 7 所示。

在消融关键词倒排索引信息树后，可知在不同关键词个数下，覆盖率具有不稳定性，说明所提方法满足了覆盖兴趣点的路径规划；同时，在未消融关键词倒排索引信息树时，Kbd_a*算法要求覆盖到查询 Q 下的所有关键词，因此，覆盖率可以达到 100%，说明关键词倒排索引信息树在所提方法中占

据核心地位，在满足覆盖兴趣点的同时实现了个性化路径规划。通过消融关键词倒排索引信息树实验得到的关键词覆盖率可以观察到，消融时 Floyd 算法平均覆盖率为 58%；Dijkstra 算法平均覆盖率为 55%；Prim 算法平均覆盖率为 58%，关键词覆盖双向 A*算法平均覆盖率为 54%；未消融时所有算法的关键词覆盖率均达到 100%，对于其每个算法的平均提升率为 42%、45%、42%、46%。分析可知本文方法中的关键词倒排索引信息树提升了关键词的覆盖率，可以满足用户的个性化自驾游路径；同时，证明该树具有一定的剪枝能力。

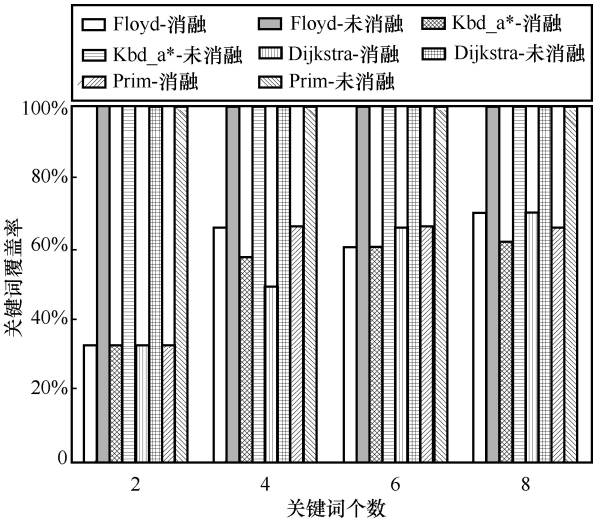
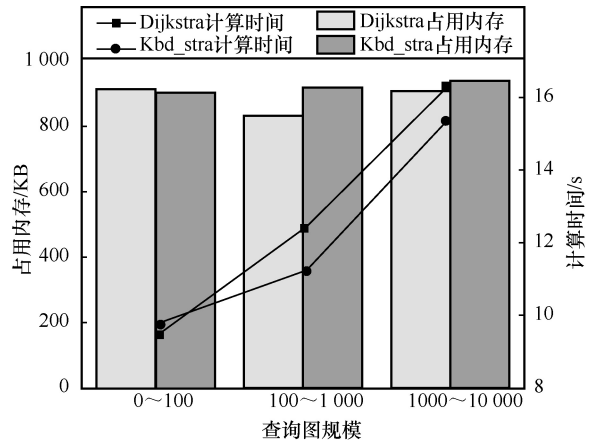


图 7　关键词覆盖率对比

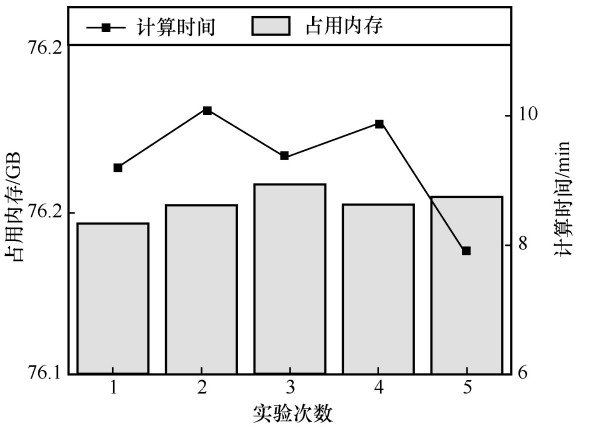
4.3　预处理开销对比

实验 2　预处理时空开销实验。为了验证 Kbd_stra 算法在预处理时空开销上的优越性，在不同的查询图规模下进行对比，采用 G_1～G_4，求取不同查询图规模的占用内存和预处理计算时间的平均值。预处理时空开销对比如图 8 所示。

如图 8(a)所示，Dijkstra[7]算法平均占用内存为 885.074 KB，平均时间为 12.71 s；Kbd_stra 算法平均占用内存为 920.90 KB，平均时间为 12.12 s。通过对比分析可知，在面向不同节点数的查询图规模时，与 Dijkstra 算法相比，本文的 Kbd_stra 算法内存占用更稳定，且时间消耗更低。如图 8(b)所示，Floyd 算法平均占用内存为 76 GB，平均时间为 9.29 min。Floyd 算法在实验中服务器内存资源充足时，内存开销远超 Dijkstra 算法，且耗时更长；服务器内存资源相对紧张时，Floyd 算法会因内存分配不足而终止；在一定的程度上体现了 Kbd_stra 算法的性能优势。



(a)　Dijkstra算法与Kbd_stra算法对比



(b)　Flody算法补充对比

图 8　预处理时空开销对比

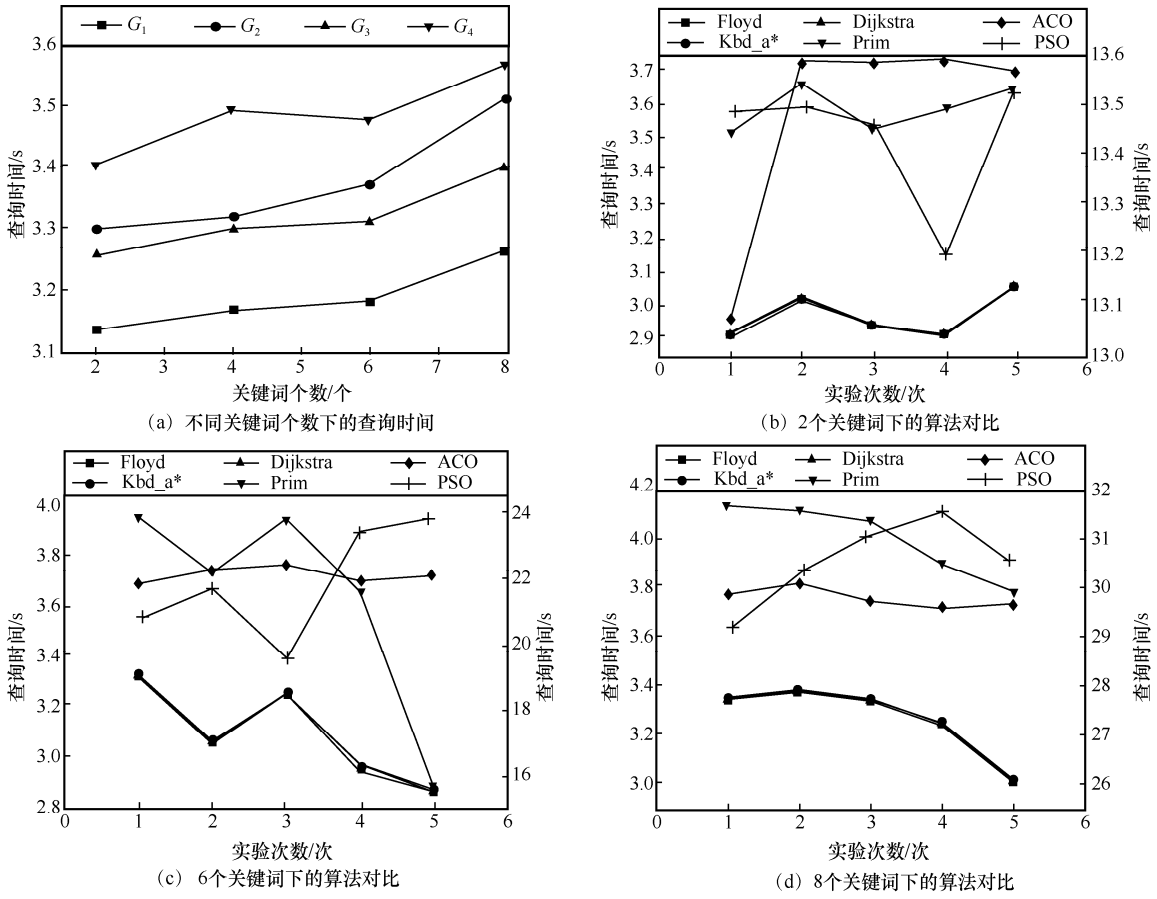
从图 8 可知，随着查询图规模的增加，Kbd_stra 算法以内存开销为代价降低了算法执行的时间开销，其内存消耗上略高于 Dijkstra 算法但低于 Floyd 算法；其原因为 Kbd_stra 算法中用到 2 个表进行存储，需要多消耗一部分内存，但其内存开销在合理可接受的范围内。

4.4　不同查询因素对算法效率的影响

本节针对 3 个影响因素，即关键词个数、查询图规模、图规模，设置了实验 3～实验 5，并对结果进行了分析。

实验 3　关键词个数对算法效率的影响实验。为了判断关键词个数对算法效率的影响，采用 G_1～G_4，对比不同关键词个数下的查询时间；同时为了证明关键词覆盖双向 A*算法的优势，在不同关键词个数下对比了算法效率，结果如图 9 所示。

从图 9(a)可知，当关键词为 6 个以上时，查询时间会显著上升，这是由于关键词增多时，遍历的深度也会大幅增加，即使采用关键词倒排索引算法也会产生大量中间结果，导致查询时间显著上升。当图规模增大时，查询时间随关键词个数增加而增加。



（a）不同关键词个数下的查询时间

（b）2个关键词下的算法对比

（c）6个关键词下的算法对比

（d）8个关键词下的算法对比

图 9　不同关键词个数下的算法效率

这是因为在更大规模的图中，关键词个数增加会导致最优路径更难被拓展出来。

图 9(b)～图 9(d)中，右纵坐标表示 PSO 算法的查询时间，左纵坐标表示其他算法的查询时间，从图 9(b)～图 9(d)可知，通过预处理、剪枝过程的查询图大小已经可以使不同算法达到最好状态，智能化元启发算法本身含有初始种群大小，以及代数的轮转，如 ACO 算法[18]和 PSO 算法[19]；因此，在不同关键词个数下的效果较差，而 Kbd_a*算法相对处于中间阶段，但结合实验 6 来看，其响应时间平稳，在关键词覆盖最优路径规划方法中起到了作用。

实验 4　查询图规模对算法效率的影响实验。设定关键词个数为 6，采用 G_3 比较不同查询图规模下的响应时间，结果如图 10 所示。

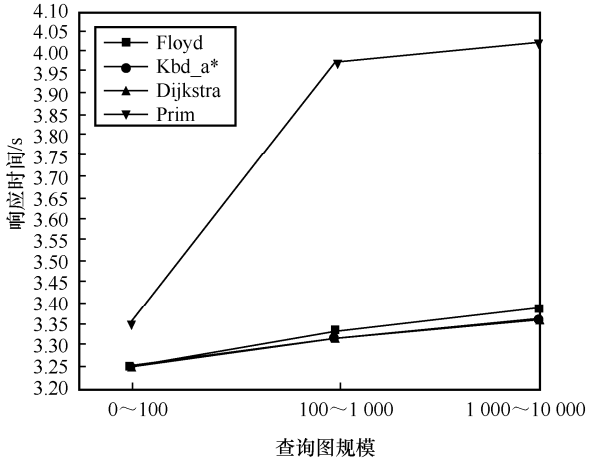


图 10　不同查询图规模下的时间对比

从图 10 可以看出，本文的 Kbd_a*算法具有一定的优势，但优势并不明显。这是因为通过算法 1 和算法 2 的修剪，查询图已剪枝到所有实验算法都可达到最好状态的规模。因此，在这种情况下，规划路径所用的时间相对接近。

实验 5　图规模对算法效率的影响实验。为了判断图规模对算法效率的影响，本节在相同关键字个数下进行实验，对比不同图规模下的计算时间，实验结果如图 11 所示。从图 11 可知，相较于对比算法，本文的 Kbd_a*算法较稳定，而且时间较短。

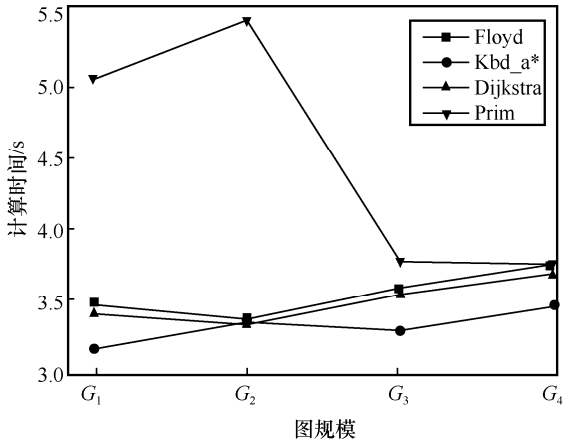


图 11　不同图规模下的计算时间

4.5　关键词覆盖最优路径规划方法消融实验

实验 6　消融实验。为了测试本文所提预处理阶段的可行性和关键词倒排索引信息树的作用，本节进行了消融实验，实验结果如图 12 所示。从图 12(a)可知，包含预处理阶段的响应时间具有一定的优势，验证了预处理阶段的可行性。

图 12(b)中，右纵坐标表示 Prim 算法查询时间，左纵坐标表示其他算法查询时间。从图 12(b)可知，在不同查询图规模下，本文的 Kbd_a*算法查询时间较长，这是因为在路径搜索过程中 Kbd_a*算法有估价函数，其估计的距离必须小于或等于真实距离。当估价距离小于或等于真实距离时，搜寻时剪掉的路径空间较多，最坏的情况下需要搜寻整个路径空间，查询时间接近 Dijkstra 算法。在个别情况下，Kbd_a*算法查询时间接近 Dijkstra 算法。图 12(c)中，右纵坐标表示 PSO 算法查询时间，左纵坐标表示其他算法查询时间。从图 12(c)可知，不同关键词个数下本文的 Kbd_a*算法查询时间较短且曲线较平缓，具有一定的优势，因此结合实验 3 和实验 1 可知，倒排索引信息树具有一定的深度遍历能力和剪枝能力，具有一定的可扩展性。

上述实验结构表明，本文所提关键词覆盖最优路径规划方法是有效的，实现了个性化自驾游路径规划，提升了关键词覆盖最优路径规划方法的性能。

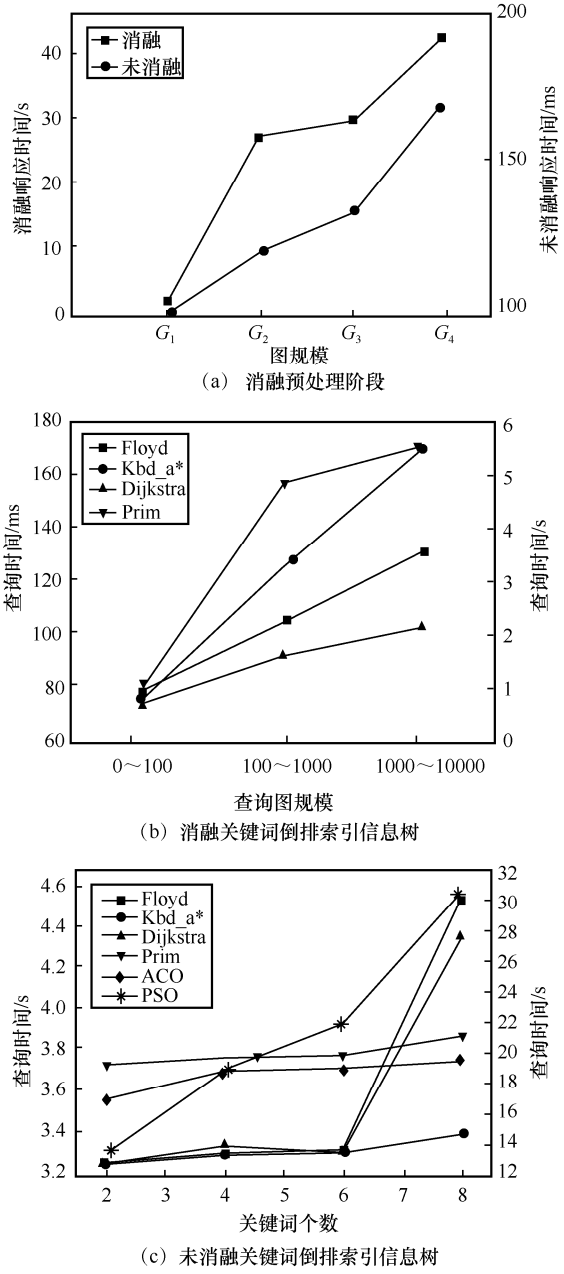


图 12　消融实验

5　结束语

本文提出的面向大规模图数据的关键词覆盖最优路径规划方法通过构建算法结合的方法，使用路网信息预处理模型、关键词倒排索引信息树模型以及关键词覆盖最优路径拓展模型，实现了个性化自驾游路径的推荐，同时方法性能得到提升。

但本文算法也存在一定的局限性。首先，在预处理阶段，本文的 Kbd_stra 算法在一定程度上依赖内存资源，需要进一步降低内存的消耗问题；其次，在关键词倒排索引信息树中，树形结构较简单，面对长路径的遍历能力还有待提高。因此，未来的研究工作将主要集中于以下 3 个方面。1) 通过研究新 Kbd_stra 算法，进一步降低预处理的内存消耗，为后续算法提供支撑，或使用机器学习与传统算法相结合的方式提升路径的精确度。2) 进一步优化关键词倒排索引信息树，增强树的可扩展性，使其更好地适应其他应

用场景。3) 通过进一步优化路径拓展方法，使其在没有关键词倒排索引信息树的结构下，也具有一定的优势；同时也可以考虑加入智能化元启发式算法进行解决。

附录 1 路径搜索定理证明

证明 根据反证法可知，假设存在一条从 v_s 到 v_t 的路径，此时 $\mathrm{BS}(v_s,v_n)<\mathrm{top_f}$ 并且 $\mathrm{BS}(v_t,v_m)<\mathrm{top_r}$。由式(1)可知，正反向搜索可以扫描到弧 $L(n,m)$，此时满足

$$\begin{aligned}&\mathrm{BS}(v_s,v_n)+\mathrm{BS}(v_n,v_m)+\mathrm{BS}(v_t,v_m)<\\&\mathrm{top_f}+\mathrm{BS}(v_n,v_m)+\mathrm{top_r}\end{aligned}\tag{11}$$

由式(1)和式(11)可知，路径代价值小于初始路径代价值，即

$$\mathrm{top_f}+\mathrm{BS}(v_n,v_m)+\mathrm{top_r}<\mathrm{BS_{max}}+\mathrm{BS}(v_n,v_m)\tag{12}$$

式(12)与式(3)矛盾。由此可知，若 v 不在路径上，此时小根堆中的路径代价值无法相加；反之，若 v 在路径上，此时小根堆中的路径代价值可以相加。定理 1 成立。

证毕。

参考文献：

[1] ZHANG H F, GE H W, YANG J L, et al. Review of vehicle routing problems: models, classification and solving algorithms[J]. Archives of Computational Methods in Engineering, 2022, 29(1): 195-221.

[2] FENG Y, WANG H, LU H, et al. A novel faster all-pair shortest path algorithm based on the matrix mul-tiplication for GPUs[J]. arXiv Preprint, arXiv: 2208.04514, 2022.

[3] HUANG T, GONG Y J, ZHANG Y H, et al. Automatic planning of multiple itineraries: a niching genetic evolution approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(10): 4225-4240.

[4] 丛玉良, 孙闻晞, 薛科, 等. 基于改进的混合遗传算法的车联网任务卸载策略研究[J]. 通信学报, 2022, 43(10): 77-85.
CONG Y L, SUN W X, XUE K, et al. Research on task offloading strategy of Internet of vehicles based on improved hybrid genetic algorithm[J]. Journal on Communications, 2022, 43(10): 77-85.

[5] 金鹏飞, 牛保宁, 张兴忠. 高效的多关键词匹配最优路径查询算法KSRG[J]. 计算机应用, 2017, 37(2): 352-359.
JIN P F, NIU B N, ZHANG X Z. KSRG: an efficient optimal route query algorithm for multi-keyword coverage[J]. Journal of Computer Applications, 2017, 37(2): 352-359.

[6] 刘蒙蒙, 牛保宁, 杨茸. 关键词最优路径查询的分段拓展算法[J]. 计算机工程, 2022, 48(6): 79-88.
LIU M M, NIU B N, YANG R. Segmentation expansion algorithm for keyword-aware optimal route query[J]. Computer Engineering, 2022, 48(6): 79-88.

[7] TANG Z Z, MA H Z. An overview of path planning algorithms[J]. IOP Conference Series: Earth and Environmental Science, 2021, 804(2): 022024.

[8] WANG S, LIU B, LIU W P, et al. Research on the shortest path for crossing desert based on Floyd algorithm[C]//Proceedings of 2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC). Piscataway: IEEE Press, 2021: 1-4.

[9] JU C Y, LUO Q H, YAN X Z. Path planning using an improved A-star algorithm[C]//Proceedings of 2020 11th International Conference on Prognostics and System Health Management (PHM-2020 Jinan). Piscataway: IEEE Press, 2020: 23-26.

[10] LIU R, FENG F. Optimization model of pipeline laying based on immune genetics and improved second-order prim algorithm[C]//Proceedings of 2021 International Conference on Communications, Information System and Computer Engineering (CISCE). Piscataway: IEEE Press, 2021: 620-626.

[11] THAMMACHANTUEK I, KETCHAM M. Path planning for autonomous mobile robots using multi-objective evolutionary particle swarm optimization[J]. PLoS One, 2022, 17(8): e0271924.

[12] 刘丽珏, 罗舒宁, 高琰, 等. 基于回溯蚁群-粒子群混合算法的多点路径规划[J]. 通信学报, 2019, 40(2): 102-110.
LIU L J, LUO S N, GAO Y, et al. Multi-point path planning based on the algorithm of colony-particle swarm optimization[J]. Journal on Communications, 2019, 40(2): 102-110.

[13] LIU L S, WANG B, XU H. Research on path-planning algorithm integrating optimization A-star algorithm and artificial potential field method[J]. Electronics, 2022, 11(22): 3660.

[14] ATTIQUE M, AFZAL M, ALI F, et al. Geo-social top-k and skyline keyword queries on road networks[J]. Sensors, 2020, 20(3): 798.

[15] 郝晋瑶, 牛保宁, 康家兴. 大规模路网图下关键词覆盖最优路径查询优化[J]. 软件学报, 2020, 31(8): 2543-2556.
HAO J Y, NIU B N, KANG J X. Optimization of keyword-aware optimal route query on large-scale road networks[J]. Journal of Software, 2020, 31(8): 2543-2556.

[16] LIU M M, NIU B N, YANG R. A segmented parallel expansion algorithm for keyword-aware optimal route query[J]. GeoInformatica, 2022: 1-27.

[17] JIN P F. Research on efficient keyword-aware optimal route query processing method[D]. Jinzhong: Taiyuan University of Technology, 2017.

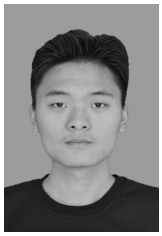
[18] LI X T, HUANG T P, CHEN H H, et al. Path planning of mobile robot based on dynamic chaotic ant colony optimization algorithm[C]//Proceedings of 2022 IEEE 10th International Conference on Information, Commu-

nication and Networks (ICICN). Piscataway: IEEE Press, 2023: 515-519.

[19] YUAN D C. Research on path-planning of particle swarm optimization based on distance penalty[C]//Proceedings of 2021 2nd International Conference on Computing and Data Science (CDS). Piscataway: IEEE Press, 2021: 149-153.

[作者简介]

李梓杨（1993− ），男，新疆乌鲁木齐人，博士，新疆大学副教授、硕士生导师，主要研究方向为大数据分析、机器学习。



陈鹏程（1998− ），男，河南正阳人，新疆大学硕士生，主要研究方向为图计算、路径规划。

于炯（1964− ），男，北京人，博士，新疆大学教授、博士生导师，主要研究方向为网格计算、并行计算、分布式系统。

蒲勇霖（1991− ），男，山东淄博人，博士，南京信息工程大学讲师，主要研究方向为边缘计算、绿色计算等。

何贞贞（1994− ），女，新疆乌鲁木齐人，新疆大学博士生，主要研究方向为图查询优化、深度学习等。

李雪（1996− ），女，江苏新沂人，新疆大学博士生，主要研究方向为深度学习、图像处理等。

郑世杰（1998− ），男，四川广安人，新疆大学硕士生，主要研究方向为增量学习、异常检测。