Joinable Search over Multi-source Spatial Datasets: Overlap, Coverage, and Efficiency

Efficient and Privacy-Preserving Spatial-Feature-Based Reverse kNN Query

Yandong Zheng , Member, IEEE, Rongxing Lu , Fellow, IEEE, Yunguo Guan , Songnian Zhang , Jun Shao , Senior Member, IEEE, Fengwei Wang , Member, IEEE, and Hui Zhu , Senior Member, IEEE

IEEE TRANSACTIONS ON SERVICES COMPUTING, VOL. 16, NO. 4, JULY/AUGUST 2023

汇报人： 张政楷

学习内容

Reverses kNN问题：给定空间数据q，找寻其他将q视为k最近邻的top-k个空间数据tk

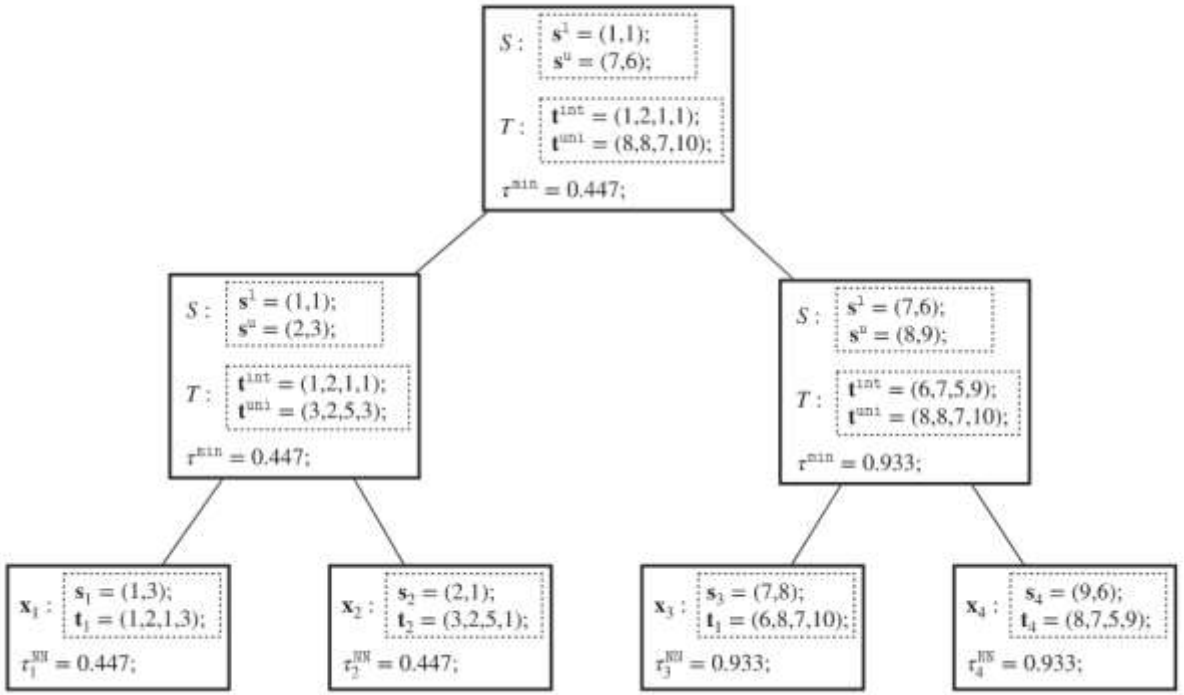
本文场景：在线约会系统中，用户A希望找到将自己视为前k个选择的用户，且性格合拍
本文提出改进：基于空间特征的RkNN

空间特征由两个值决定：

1. 由经纬度决定的空间位置$s_i \rightarrow dict$
2. 由人物性格决定的特征向量$t_i \rightarrow Jaccard$

$$Sim(\mathbf{x}_i, \mathbf{x}_j) = \alpha * \left(1 - \frac{D(\mathbf{s}_i, \mathbf{s}_j) - \phi_s}{\psi_s - \phi_s}\right) + (1 - \alpha) * \frac{J(\mathbf{t}_i, \mathbf{t}_j) - \phi_t}{\psi_t - \phi_t}$$

索引方式：
提前计算出一个每个空间数据的第k个最近邻的dict和Jaccard计算出相似度Sim
利用R-Tree将其索引
1，空间位置：S
2，特征：T
3，其相似度：τ
最后使用SHE加密



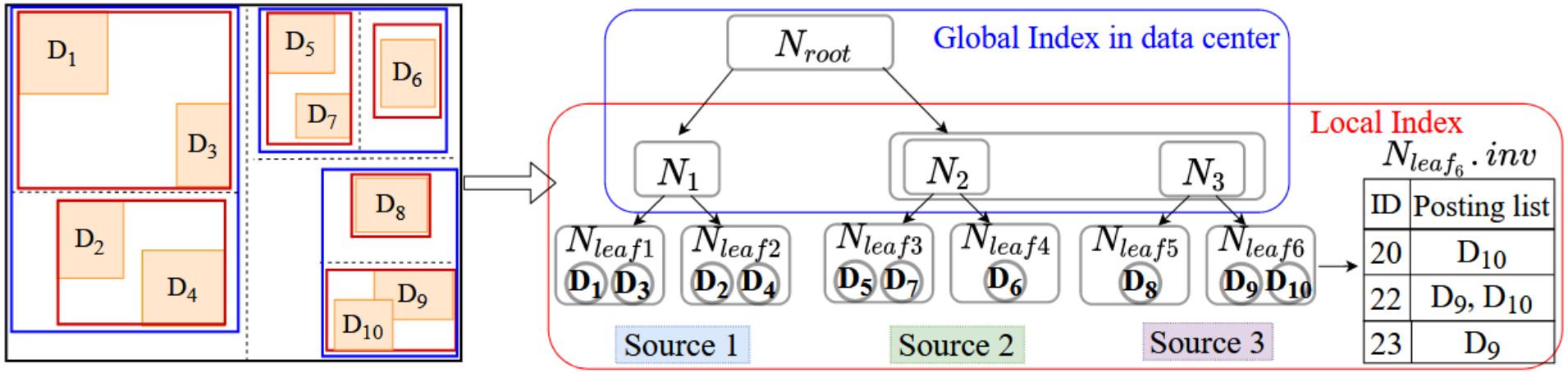
学习内容

数据集joinable query问题场景：数据增强
给定一个空间数据集，计算其可连接的其他空间数据集数据集，并做如下两个操作

可连接：$dict\ （S_1, S_2）\ \leq \delta \quad \delta = 2^{\theta}\sqrt{2}$

1.Overlap Joinable Search：找出尽可能覆盖给定数据集的Top-k个数据集
2.Coverage Joinable Search：找出尽可能覆盖整片区域的Top-K个数据集

索引方式：
构建$2^{\theta} \times 2^{\theta}$的网格，将其网格进行编号后构建倒排索引： 网格号：$\{D_1, D_2, \ldots D_i\}$
利用R-Tree构建MBR将所有数据集包含进索引，并在叶子节点存放MBR中包含网格号所对应的倒排索引



学习内容

提到概念：
MCP（Maximum Coverage Problem）：用于Overlap Joinable Search的理论

即存在总集合$U = \{u_1, u_2, ..., u_n\}$以及存在多个子集$S_i = \{s_{i1}, s_{i2}, ..., s_{i3}\}$使得找到k个子集$S_1, ...S_k$使得其能够尽可能覆盖$U$中的所有元素。

考虑将概念MCP引入关键词集合：引入关键词集合的相似度问题
既考虑数据集的Joinable Search，又考虑两集合之间的关键词集合Keywords

问题实际场景1：某城市政府需要将原有的商业旅游景区进行扩建，修建2期工程，需要找寻k个数据集满足扩建需求

A：原有空间数据集
D=$\{D_1, D_2, ..., D_n\}$

Query →

找出与A连接，且关键词高度重合的数据集
Joinable Union Search

问题实际场景2：我们需要知道某一地区的全貌，由于各空间数据集数据并不全面，我们需要查询某一区域的所有数据集以尽可能还原地区全貌

B：原有空间数据集
D=$\{D_1, D_2, ..., D_n\}$

Query →

找出与B连接并尽可能覆盖B包含的区域，且关键词不重合的数据集Joinable Union Search

学习内容

$$spatial = （x, y） \rightarrow D_{vector} \qquad dict(d_i, q_d)$$

$$Keywords = k_1, k_2, ..., k_n \rightarrow K_{vector} \qquad Jaccard(K_i, q_k)$$

两阶段：
1，找出可连接的m（m>k）个空间数据集
2，再在m个数据集中计算Jaccard算出关键词相似度，取前K个数据集返回结果R
问题：欧氏距离d和关键词相似度Jaccard的权重问题

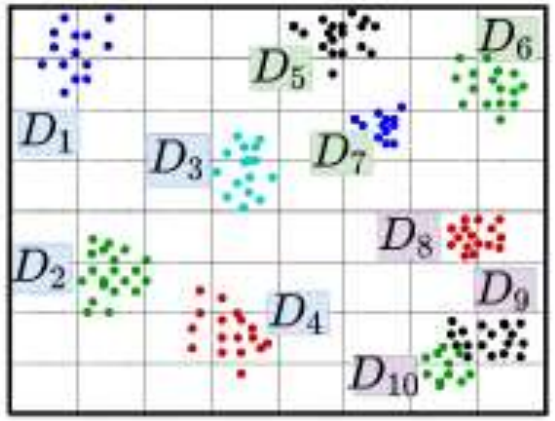
改为单阶段，计算两者的相似度：

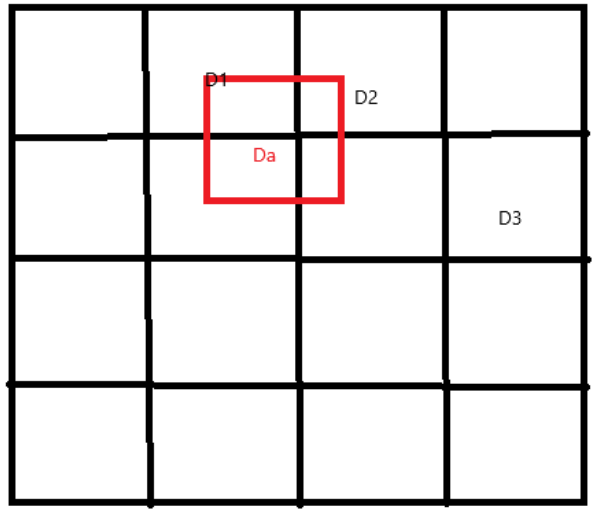
Joinable Union Search：$Sim(d_i, q) = \lambda_1 * Dict(d_i, q_d) + \lambda_2 * Jaccard(d_i, q_k)$

Joinable Overlap Search：$Sim(d_i, q) = \lambda_1 * Cov(d_i, q_d) + \lambda_2 * (1 - Jaccard(d_i, q_k))$

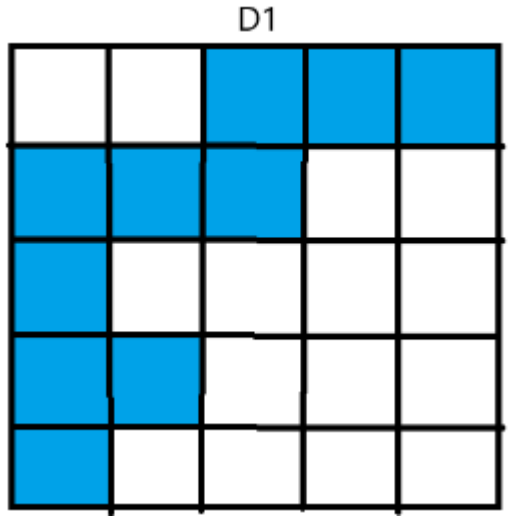
学习内容

将整个空间进行网格划分





D_1, D_2



每个数据集存储三项内容：
数据集中对应空间点所在空间的（a，b）左下右上的边界
数据集内点所在的网格转为hash
Keywords集合转为向量

构建R-Tree：
每个最小外接矩形进行交集判断，
当判断到叶子节点时将数据集放入候选集和内

将搜索出来的数据集对应的哈希运算，求出两数据集重复的小网格，
将小网格内的空间点计算欧氏距离dict
并计算出总的keywords相似度Jaccard

最后计算出总的Sim

THANKS