



哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

大数据分析

大作业系统设计报告

(2019 年度春季学期)

组	员	<u>1160300312 靳贺霖</u>
组	员	<u>1160300314 朱明彦</u>
学	院	<u>计算机学院</u>
教	师	<u>杨东华、王金宝</u>

计算机科学与技术学院

目录

第 1 章 问题描述	3
1.1 数据	3
1.2 范围查询	3
1.3 k NN 查询	3
1.4 Reverse k NN 查询	3
第 2 章 系统设计	4
第 3 章 系统工作流程	4

分布式空间近似关键字查询系统

第 1 章 问题描述

1.1 数据

空间对象集合 $D = o_1, o_2, \dots, o_n$, 对于 D 中任意一个对象 $o_i = (loc_i, kw_{i,1}, \dots, kw_{i,m})$, 即包含 \mathbb{R}^d 维欧式空间中一个点 loc_i 和一组关键字 $kw_{i,1}, \dots, kw_{i,m}$, 记为 $o_i.loc = loc_i$ 和 $o_i.kw = \{kw_{i,1}, \dots, kw_{i,m}\}$ 。在本项目中主要关注 $d = 2$ 的情况, \mathbb{R}^2 对于现实中的应用有着很大的价值, 但项目中提出的方法可以扩充到任意有限维欧式空间。

1.2 范围查询

输入: $Q = (Q_{rs}, Q_{rt})$, 其中 Q_{rs} 是一个空间范围 (\mathbb{R}^d 维欧式空间中的超立方体); Q_{rt} 为关键字近似条件, $Q_{rt} = \{(kw_1, \theta_1), \dots, (kw_K, \theta_K)\}$, 其中 θ_i 为阈值。

输出: $O = \{o | o \in D, o.loc \in Q_{rs}, \forall (kw_i, \theta_i) \in Q_{rt}, \exists o.kw_j, ED(kw_j, kw_i) \leq \theta_i\}$, 其中 $ED(kw_j, kw_i)$ 表示两个关键字 kw_i 和 kw_j 之间的编辑距离。

1.3 kNN 查询

输入: $Q = (Q_s, Q_t, k)$, 其中 $Q_s = loc$ 是 \mathbb{R}^d 维欧式空间中一个点, 即查询发出的位置; $Q_t = \{(kw_1, \theta_1), \dots, (kw_K, \theta_K)\}$; k 为表示最近邻居的数量。

输出: 对 $O_t = \{o | o \in D, \forall (kw_i, \theta_i) \in Q_t, \exists o.kw_j, ED(kw_j, kw_i) \leq \theta_i\}$, 根据 $|O_t|$ 的大小进行定义,

- 如果 $|O_t| \leq k$, 则 $O_{kNN} = O_t$ 即为最终结果。
- 如果 $|O_t| > k$, $O_{kNN} = \{o | o \in O_t, \forall o_i \in O_t - O, Dis(loc, o_i) \geq Dis(loc, o_j) \text{ 对 } \forall o_j \in O \text{ 成立}\}$ 并且 $|O_{kNN}| = k$ 。

1.4 Reverse kNN 查询

输入: 与1.3节输入相同, 不再赘述。

输出: $O_{RkNN} = \{o_{R_1}, \dots, o_{R_M}\}$, 对于 O_{RkNN} 中的任一元素 o_{R_i} 均有 $o_{kNN} \in O_{R_i-kNN}$ 且 $o_{R_i} \in D$, 其中 $o_{kNN}.loc = Q_s, o_{kNN}.kw = Q_t$; O_{R_i-kNN} 是以 $(o_{R_i}.loc, o_{R_i}.kw, k)$ 为输入的 kNN 查询结果。

第 2 章 系统设计

主要思路

- 存储部分，类似 Spark、HDFS 进行处理
- 索引和算法部分，两层索引结构，组织不同节点间的索引使用 RT-CAN [3]，在本地使用以 R 树为核心，结合 MHR-Tree [1] 进行范围查询，结合 Voronoi Diagrams [2] 进行 k NN 查询和 Reverse k NN 查询。

第 3 章 系统工作流程

参考文献

- [1] Li F, Yao B, Tang M, et al. Spatial approximate string search[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 25(6): 1394-1409.
- [2] Sharifzadeh M, Shahabi C. Vor-tree: R-trees with voronoi diagrams for efficient processing of spatial nearest neighbor queries[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 1231-1242.
- [3] Wang J, Wu S, Gao H, et al. Indexing multi-dimensional data in a cloud system[C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010: 591-602.