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

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 kNN 查询 3

1.4 Reverse kNN 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 $D = o_1, o_2, \ldots, o_n$，对于 D 中任意一个对象 $o_i = (loc_i, kw_{i,1}, \ldots, kw_{i,m})$，即包含 $\mathbb{N}$ 维欧式空间中一个点 loc_i 和一组关键字 $kw_{i,1}, \ldots, kw_{i,m}$，记为 $o_i.loc = loc_i$ 和 $o_i.kw = \{kw_{i,1}, \ldots, kw_{i,m}\}$。

1.2 范围查询

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

输出： $O = \{o | o \in D, o.loc \in Q.Q_s, \forall (kw_i, \theta_i) \in Q.Q_t, \exists o.kw_j, \mathrm{ED}(kw_j, kw_i) \leq \theta_i\}$，其中 $\mathrm{ED}(kw_j, kw_i)$ 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 $\mathbb{N}$ 维欧式空间中一个点，即查询发出的位置；$Q_t = \{(kw_1, \theta_1), \ldots, (kw_K, \theta_K)\}$；$k$ 为表示最近邻居的数量。

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

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

1.4 Reverse kNN 查询

输入： 与1.3节输入相同，不再赘述。

输出： $O_{\mathrm{R}k\mathrm{NN}} = \{o_{R_1}, \ldots, o_{R_M}\}$，对于 $O_{\mathrm{R}k\mathrm{NN}}$ 中的任一元素 o_{R_i} 均有 $o_{k\mathrm{NN}} \in O_{R_i - k\mathrm{NN}}$ 且 $o_{R_i} \in D$，其中 $o_{k\mathrm{NN}}.loc = Q_s, o_{k\mathrm{NN}}.kw = Q_t$；$O_{R_i - k\mathrm{NN}}$ 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 kNN 查询和 Reverse kNN 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.