第 38 卷 第 4 期 计 算 机 工 程 2012 年 2 月
Vol.38 No.4 Computer Engineering February 2012

·软件技术与数据库· 文章编号：1000—3428(2012)04—0082—02 文献标识码：A 中图分类号：TP311.12

数据集中单纯型连续近邻链查询方法

李 松[1a]，张丽平[1a]，蔡志涛[2]，郝晓红[1b]，王 淼[1a]

(1. 哈尔滨理工大学 a. 计算机科学与技术学院；b. 计算中心，哈尔滨 150080；
2. 盐城生物工程高等职业技术学校汽车电子工程系，江苏 盐城 224051)

摘 要：为解决数据集中的单纯型连续近邻链查询问题，提出一种基于 Voronoi 图的查询方法。给出单纯型连续近邻链查询的定义，利用 Voronoi 图的性质对大量数据点进行精减，设计可准确查询出数据集中单纯型连续近邻链的查询算法。实验结果表明，随着待查连续近邻链所含数据点规模的增大，该方法的效率比传统基于 R 树方法更高。

关键词：空间数据库；数据集；最近邻查询；连续近邻链；R 树；Voronoi 图

Query Method of Simple Continus Near Neighbor Chain in Dataset

LI Song[1a], ZHANG Li-ping[1a], CAI Zhi-tao[2], HAO Xiao-hong[1b], WANG Miao[1a]

(1a. School of Computer Science and Technology; 1b. Computation Center, Harbin University of Science and Technology, Harbin 150080, China;
2. Department of Automotive Electronic Engineering, Yancheng Biology Engineering Higher Vocational School, Yancheng 224051, China)

【Abstract】To handle the Simple Continues Near Neighbor Chain(SCNNC) query in dataset, the method based on the Voronoi diagram is proposed. The definition of SCNNC query is given and many data points are deleted based on the properties of the Voronoi diagram. The algorithm which can accurately query the simple continues near neighbor chain in dataset is put forward. Experimental results show that with the increasing numbers of the points in the simple continue near neighbor chain, the method has more advantages than the method based on R tree.

【Key words】 spatial database; dataset; Nearest Neighbor(NN) query; continues near neighbor chain; R tree; Voronoi diagram

DOI: 10.3969/j.issn.1000-3428.2012.04.027

1 概述

对海量空间数据对象的最近邻(Nearest Neighbor, NN)查询及其变种查询问题的研究[1-9]在空间定位技术、机器人技术、物联网技术、计算机游戏、地理信息系统、空间数据库和智能查询技术等领域具有重要意义。

最近邻查询的一个变种问题是连续近邻链(Continues Near Neighbor Chain, CNNC)查询问题，所谓连续近邻链查询，就是在给定空间的数据集中查找满足某些空间关系特征的一组有序数据点集，该集合中后一数据点均是前一数据点的近邻(最近邻或第 k 最近邻)。根据限制条件不同，连续近邻链查询还可细分为单纯型连续近邻链(Simple Continues Near Neighbor Chain, SCNNC)查询和多态型连续近邻链查询等。已有的近邻查询方法[1-9]主要集中在数据集中特定点的最近邻和反向最近邻等方面，无法处理近邻链查询问题。为了弥补已有方法的空白，本文基于计算几何中的 Voronoi 图对数据集中单纯型连续近邻链问题进行了研究，提出了一种有效的查询数据集中单纯型连续近邻链的方法。

2 基本定义

单纯型连续近邻链查询问题是最近邻查询问题的一个重要的变种问题，本节给出最近邻查询的形式化定义如下所示：

定义 1(最近邻查询)[1] 假设有一 d 维空间的点集 S 和一个查询点 q，最近邻查询就是找出 S 的子集 NN(q)：NN(q)=$\{s \in S | \forall p \in S: D(q,s) \leqslant D(q,p)\}$。若要找出 k 个最近邻，该定义则可扩展成 k 个最近邻的查询，即 kNN(q)=$\{p_1,p_2,\cdots,p_k\}$。其中，$\forall p \in S-k\text{NN}(q)$，$s \in k\text{NN}(q)$，且 $D(q,s) \leqslant D(q,p)$；$D(p_i, q) \leqslant D(p_j,q)$，$1 \leqslant i \leqslant j \leqslant k$。

在定义 1 中，$D(q,s)$和 $D(q,p)$等表示 2 个数据点之间的距离。该距离可以是欧氏空间中的直线距离，也可以是曲面距离和网络路径距离。基于定义 1，下文给出单纯型连续近邻链的形式化定义：

定义 2(连续近邻链) 设有一 d 维空间中的数据点集 P，P 中一组有序数据点的集合记为 L，$L=\{p_m, p_{m+1},\cdots, p_n\}$。其中，$p_{i+1} \in \text{NN}(p_i)$(或 $p_{i+1} \notin \text{NN}(p_i)$且 $p_{i+1} \in k\text{NN}(p_i)$)，$i=m,m+1,\cdots, n-1$，称 L 为 P 集中的一条连续近邻链；p_m 称为链首点；p_n 称为链尾点。

定义 3(单纯型连续近邻链查询) 设 L 为数据集 P 中的一条连续近邻链，若 L 满足以下条件，则称 L 为单纯型连续近邻链：

(1) $\forall p_i,p_j \in L$，若 $i \neq j$，则有 $p_i \neq p_j$；

(2) $\forall p_i,p_j \in L$，若 $p_{i+1} \neq p_{j+1}$，则有 $p_i \neq p_j$；

(3) $\forall p_i \in L$，$p_{i+1} \notin \{p_m, p_{m+1},\cdots,p_i\}$；

(4) $\forall p_i \in L$，若 $p_{i+1} \notin \text{NN}(p_i)$且 $p_{i+1} \in k\text{NN}(p_i)$，则有$(k-1)$ $\text{NN}(p_i) \subseteq \{p_m, p_{m+1},\cdots,p_i\}$。

在数据集 P 中，查找一条 p_m 到 p_n 的单纯连续近邻链 L 的查询，简称为单纯型连续近邻链查询。

3 单纯型连续近邻链的查询方法

查询单纯型连续近邻链的一种可行方法是利用 R 树进行查询，但由于利用 R 树索引结构进行查询往往需要大量冗余距离的计算，在连续近邻链中的数据点较多的情况下效率并

基金项目：黑龙江省教育厅科学技术研究基金资助项目(11551084)

作者简介：李 松(1977－)，男，副教授、博士，主研方向：数据库技术；张丽平，讲师、硕士；蔡志涛，讲师；郝晓红，高级实验师；王 淼，博士研究生

收稿日期：2011-08-03 **E-mail：**lisongbeifen@163.com

不十分理想。为了弥补该方法的不足，本节着重引进计算几何中的 Voronoi 图[10]对该问题进行处理，以下给出 Voronoi 图的定义和性质。

定义 4(Voronoi 图)[10] 给定一组数据点 $D=\{d_1,d_2,\cdots,d_n\}\subset R^2$，其中，$2<n<\infty$，当$(i\neq j)$时，$d_i\neq d_j$。Voronoi 区域由下式给出：$VP(d_i)=\{d|D(d,d_i)\leqslant D(d,d_j)\}$。$D(d,d_i)$为 d 与 d_i 之间的最小距离。d_i 称为 Voronoi 生成点，由 d_i 所决定的 Voronoi 区域 $VP(d_i)$称为 Voronoi 多边形，Voronoi 多边形的棱记为 $VL(d_i)$。由 $V(D)=\{VP(d_1), VP(d_2),\cdots,VP(d_n)\}$所定义的图形被称为 Voronoi 图(如图 1 所示)。共享相同棱的 Voronoi 多边形被称为邻接多边形，它们的生成点被称为邻接生成点。

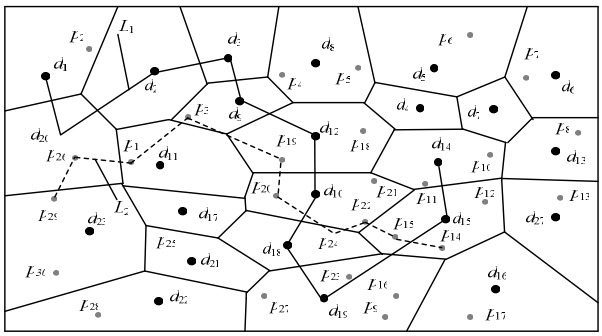


图 1　部分 Voronoi 图及单纯连续近邻链示例

在图 1 中，数据点集$\{d_1,d_{20},d_2,d_3,d_9,d_{12},d_{10},d_{18},d_{19},d_{15},d_{14}\}$即为由部分生成点所构成的一条 d_1 到 d_{14} 的单纯型连续近邻链。

性质 1[10] 离 Voronoi 生成点 d_i 最近的生成点 d_j 必是 d_i 的邻接生成点。

性质 2[10] 一个 Voronoi 多边形 $V(p_i)$内任何其他的点 $s(s\notin P)$到 p_i 的距离必小于该点到其他 Voronoi 生成点的距离。

基于定义 4 和性质 1、性质 2，本节给出基于 Voronoi 图的单纯型连续近邻链查询算法如下所示：

算法 SCNNC_V_ SEARCH(p_m,p_n,p)

输入 数据点集 P，数据对象点 p_m,p_n

输出 p_m 到 p_n 的单纯型连续近邻链 $L=\{p_m,p_{m+1},\cdots,p_{i-1},p_i,\cdots,p_n\}$

```
begin:
    L← ∅ ;
    if   p 集的 Voronoi 图没生成 then
        GeVoronoi(P); //生成 P 集的 Voronoi 图
    else
    p←pm;
(1) L←NNV(p,P);
//利用 Voronoi 图在 p 的邻接生成点中计算 p 的最近邻
if NNV(p,P)≠pn then
        P←P-p;
        p←NNV(p,P);
        转(1);
    else    // 计算到链尾点
        L←pm;
    return L;
end
```

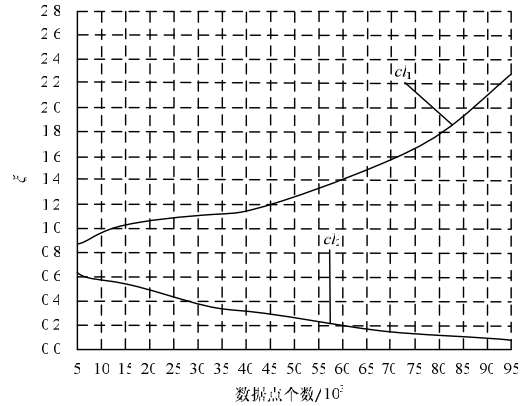
利用 SCNNC_V_SEARCH 算法可准确地查询出数据集中的单纯型连续近邻链。例如，若在图 1 中的空间数据集中查询 d_1 到 d_{14} 的单纯型连续近邻链，利用 SCNNC_V_SEARCH 算法即可得出单纯型连续近邻链为$\{d_1,d_2,d_3,d_9,d_{12},d_{10},d_{18},d_{19},d_{15},d_{14}\}$。

分析 SCNNC_V_SEARCH 算法可知，该算法生成 Vornoi 图的时间复杂度为 $O(n\log n)$，利用 Voronoi 图计算 p 的最近邻的复杂度为 $O(s)$，由于一个 Voronoi 多边形最多有 6 个邻接生成点，因此 $s\leqslant 6$，设所求单纯型连续近邻链 L 共有 m 个点，则该算法核心部分的时间复杂度为 $O(n\log n+(m-1)s)$。由 SCNNC_V_SEARCH 算法的时间复杂度可知，该算法的查询效率主要由数据集的大小、所求连续近邻链的数据点数目和利用 Voronoi 图计算 p 的最近邻的计算量所决定。

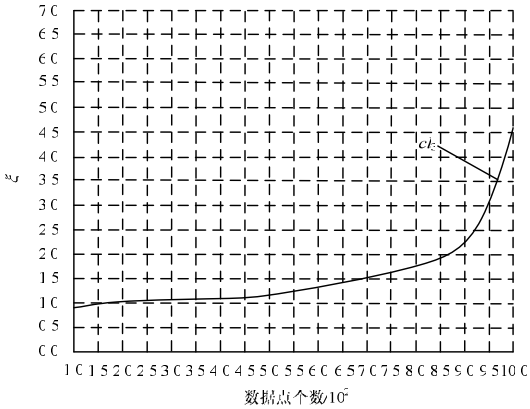
由于 Voronoi 生成点的邻接生成点数量有限，因此 Voronoi 图构造完后，生成点的最近邻的计算量较小；又由于 Voronoi 图可在曲面上进行构建，因此该算法可扩展到曲面上查询数据集中的单纯型连续近邻链。

4　实验结果与分析

本文第 2 节提出了基于 Voronoi 图的查询方法来处理单纯型连续近邻链问题。本节着重对传统的基于 R 树的查询方法和基于 Voronoi 图的查询方法进行性能分析与实验比较。为了分析所提方法的查询效率，在 Pentium4、CPU 1.8 GHz、内存 2 GB、Windows XP 环境下，利用 C++ builder 6.0 软件工具对所提出的方法进行实验分析。所用实验数据是随机生成的模拟数据对象。图 2 给出了实验结果。



(a)SCNNC_V/SCNNC_R(t=200)



(b)SCNNC_V/SCNNC_R(n=42 000)

图 2　SCNL_V 与 SCNL_R 的效率比较

图 2(a)和图 2(b)表示不同情况下，利用 Voronoi 图进行查询的方法(SCNNC_V)和传统的基于 R 树索引结构进行查询的方法(SCNNC_R)的查询效率比较情况。在图 2(a)中，横坐标表示数据集中数据点的个数，纵坐标 ξ 表示 SCNNC_V 与 SCNNC_R 的查询效率的比率。曲线 cl_1 和 cl_2 分别表示在不同大小的静态和动态数据集中查询包含 200 个数据点的单

(下转第 87 页)