第 37 卷 第 2 期
Vol.37 No.2
计 算 机 工 程
Computer Engineering
2011 年 1 月
January 2011
• 软件技术与数据库 •
文章编号：1000—3428(2011)02—0052—02 文献标识码：A 中图分类号：TP391

球面上的 K 最近邻查询算法

张丽平[a]，李 松[a]，郝晓红[b]
(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)

摘 要：针对球面上数据对象点集的特征和 K 最近邻查询的需求，提出 2 种处理球面上 K 最近邻查询的算法：基于查询轴的 K 最近邻查询算法(PAM 方法)和基于查询圆面的 K 最近邻查询算法(PCM 方法)。对 2 种算法进行实验比较，结果表明，PAM 方法和 PCM 方法都适合处理球面上的最近邻查询问题，PAM 方法在存储量和查询复杂度方面相对于 PCM 方法具有一定优势，但 PAM 方法的可扩展性远低于 PCM 方法，尤其不适合处理受限查询和带方向的查询。

关键词：最近邻；球面；查询轴；查询圆面；索引结构

Algorithms for K-Nearest Neighbor Query on Sphere

ZHANG Li-ping[a], LI Song[a], HAO Xiao-hong[b]
(a. School of Computer Science and Technology; b. Computation Center, Harbin University of Science and Technology, Harbin 150080, China)

【Abstract】 According to the characteristics of the datasets on the sphere, the algorithm of the K-Nearest Neighbor query based on the query axis (PAM) and the algorithm of the K-Nearest Neighbor query based on the query circular planar(PCM) are presented. Theoretical research and experimental results show that both the two methods can handle the problem of the K-Nearest Neighbor query on the sphere, compared with the PCM, PAM has advantages on the memory capacitance and the query efficiency, but the expansibility of PAM is poor and PCM has high scalability.

【Key words】 nearest neighbor; sphere; query axis; query circular planar; index structure
DOI: 10.3969/j.issn.1000-3428.2011.02.018

1 概述

随着空间定位技术、地理信息系统和智能查询技术的发展，对空间对象的近邻查询及其变种的研究成为空间数据库领域研究的热点和难点。近年来，国内外对空间对象的近邻关系查询问题进行了大量的工作，取得了一定的研究成果[1-5]，但其主要是对二维平面中的近邻查询问题进行分析，没有进一步给出球面上的数据对象集的最近邻查询的算法，研究成果在具体应用中具有一定的局限性。本文着重对球面上数据对象点的 K 最近邻查询算法进行研究。

2 球面上的 K 最近邻查询算法

根据球面上数据对象点的特征和 K 最近邻查询的要求，本节给出基于查询轴的 K 最近邻查询算法(PAM 方法)和基于查询圆面的 K 最近邻查询算法(PCM 方法)。

2.1 基于查询轴的 K 最近邻查询算法(PAM 方法)

定义 1 设 $P=\{p_1, p_2,\cdots, p_n\}(2\leq n\leq\infty)$为球面 S^2 上的对象点集，$\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 分别为点 $p_i\in S$ 和 $p_j\in S$ 的位置矢量，点 p_i 和 p_j 之间的最短距离定义为通过点 p_i 和 p_j 的大圆(其中心点即为球的中心)中较小弧段的长度。这个距离用公式表达为：

$$d(p_i, p_j)=\arccos(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j)\leq\pi$$

称此距离为点 p_i 和 p_j 之间的球面距离。

定义 2 过查询点 q 和球心 o 的直线称之为 q 的查询轴，q 的查询轴具有唯一性。q 的查询轴与球面相交的另一点 q' 称为 q 的球面对称点。以查询轴作为一维刻度轴，查询轴上的数据点到查询点 q 的距离称为轴查询距离。球面上的数据点在查询轴上的投影称之为轴投影点。

查询轴及查询圆面如图 1 所示，直线 qq'是查询轴，查询轴上的点 o_3 是球面上的点 p_{12} 的轴投影点。由球的性质可知，判断球面上点集之间的弧的长短可以转化为判断欧式空间内的直线段的大小。且球面上的数据对象点到查询点 q 之间的球面距离大小关系在 q 的查询轴上投影后保持不变。若查询点 q 的位置固定，球面上其他数据点在球面上移动，移动点到查询点 q 的距离关系在查询轴上因数据点的移动而做相应变化，其变化情况与球面上的一致。球面上数据点到 q 的距离大小关系及其动态距离关系的变化在 q 的查询轴上可得到较好的保持。由此，可将查询点 q 在球面数据集中的 K 最近邻问题降维到 q 的查询轴上进行处理，从而降低了查询的难度。基于查询轴的方法主要适用于球面上的数据对象点是静态或动态、查询点 q 的更新频率较低的情况。

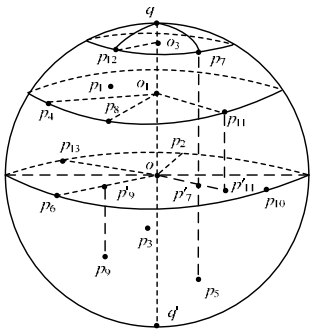


图 1 查询轴及查询圆面

若球面 S^2 上数据集中的数据点是静态的，数据集的动态变化主要限于增加或删除数据点，此时可用二叉树或 B 树来处理一维查询轴空间内的查询点 q 的 K 最近邻查询问题。当球面数据集中增加点或删除点时，相应的树索引结构可进行局部的插入或删除更新。具体算法如算法 1 所示。

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：张丽平(1976－)，女，讲师、硕士，主研方向：数据结构，数据库理论；李 松，讲师、博士；郝晓红，高级实验师
收稿日期：2010-07-02 **E-mail：**zhanglptg@163.com

算法 1 *SPA_SNN*()
输入 球面上的静态数据对象点集 P，查询点 q
输出 q 的 k 个最近邻

```
begin
    S[]←∅;
    if 没有构造 q 的查询轴 then
       Con_PSA(q)；//构造 q 的查询轴
    ProjA(P);   //投影操作
    for pi∈P do
       S[]←D(pi,q);
    //计算各数据点到 q 的轴查询距离
    Construct(BST_TREE,S[]); //建立二叉排序树
    利用二叉排序树查询 q 的 k 个最近邻；
end
```

算法将数据集 P 中的数据对象点向查询轴上投影的时间复杂度为 $O(n)$，建立二叉排序树的时间复杂度为 $O(n\log n)$，利用二叉排序树查询 q 的K最近邻的时间复杂度为 $O(k\log n)$，故算法的总时间复杂度为 $O(n\log n)$。

若球面上查询点 q 是静态的，数据集中其他点在球面上是沿圆弧轨迹进行连续移动，则可在 q 的查询轴上分步进行查询 q 的 k 个最近邻。具体的算法如算法 2 所示。

算法 2 *DPA_SNN*()
输入 球面上的动态态数据对象点集 P，查询点 q
输出 q 的 k 个最近邻

```
begin
    if 没有构造 q 的查询轴 then
        Con_PSA(q)；      //构造 q 的查询轴
    ProjA(P);
     M←PO(P,d);
    //在限距 d 内找到所有投影对象集 M
     S←NN(q,k,M);
     while 到下一时刻 do
        for pi∈P do
            if 数据点 pi 离开限距范围
                  或进入限距范围 then
               更新 M；
               S←NN(q,k,M);
        for pj，pk∈S do
           if  j<k and |pjq|>|pkq| then
              S'←pj, pj←pk, pk←S';   //次序更新
              更新最近邻集 S，
    endwhile
    return S；
end
```

2.2 基于查询圆面的 K 最近邻查询算法(PCM 方法)

上文提出的基于查询轴的 K 最近邻查询算法对具有受限最近邻查询和带有方向关系的最近邻查询无效。为了弥补该方法的不足，本节给出基于查询圆面的 K 近邻查询算法。

定义 3 如图 1 所示，过球心 o 和查询点 q 的查询轴垂直的平面与球面相交而成的圆面称为查询圆面。查询圆面有 2 面，正对 q 的一面称为 q 的近查询圆面；背对 q 的一面称为 q 的远查询圆面。

定义 4 查询圆面将球面分成 2 个半球面，以查询点 q 为顶点的半球面称为近半球面，以 q 的球面对称点 q'为顶点的半球面称为远半球面。近半球面在查询圆面的投影即为近查询圆面，远半球面在查询圆面的投影即为远查询圆面。

定义 5 查询点 q 在近查询圆面上的投影 q_o 和球心 o 重合，称为 q 的近查询投影点；q 的球面对称点 q'的投影 q_o'亦和球心 o 重合，称为 q 的远查询投影点。

由球的性质可知，近半球面上的数据点在 q 的近查询圆面上投影后，投影点到近查询投影点 q_o 之间的球面距离大小关系和球面上的原始距离对应关系一致。远半球面上的数据点在 q 的远查询圆面上投影后，投影点 p_i'到远查询投影点 q_o'之间的球面距离大小关系和球面上的原始点 p_i 到 q 的原始距离大小对应关系相反。

基于查询圆面的 K 近邻查询算法如算法 3 所示。

算法 3 *CPP_SNN*()
输入 球面上的数据对象点集 P，查询点 q
输出 q 的 k 个最近邻

```
begin
    S←∅; S'←∅;S''←∅;
    if 没有构造 q 的远近查询圆面 then
       Con_FNC(q); //构造 q 的远近查询圆面
    ProjP(P);ProjP(q);
    if 在查询圆面上没有构造查询树 then
       Con_tree(NST_tree);   //构建近查询树
       Con_tree(FST_tree);   //构建远查询树
    if n'>k then
       S←NNS(NST_tree,k);
     //用 NST 树查询 q 的 k 个最近邻
       return S;
     elseif n'<k then
       S'←NNS(NST_tree,n');
     //用 NST 树查询 q 的 n'个最近邻
      S''←NNS(FST_tree,(k-n''));
    //用 FST 树查询 q 的第 n'+1 到第 k'个最近邻
     S←S'∪ S''; //将 S'和 S''的有序数据合并入 S；
end
```

3 算法比较与实验分析

本文2种方法的性能特征如表1所示。根据球面上数据集的特点和查询要求，可科学选用最适合的查询方法。其中，“√”表示该方法适用于对应的查询类型；“×”表示该方法不适用于对应的查询类型。

表 1 2 种算法的性能特征

方法	球面查询	存储量	查询复杂性	适用数据结构	受限查询	带方向查询	可扩展性
PAM	√	低	低	二叉树、B 树等	×	×	较高
PCM	√	较低	较低	R 树、TPR 树等	√	√	高

在Pentium 4, 3.0 GHz CPU、1 GB内存、Windows XP 环境下，利用C++ builder 6.0开发工具对本文提出的2种方法进行了实验比较分析。所用实验数据是在球面上随机生成的模拟动态数据对象。实验结果如图2所示。

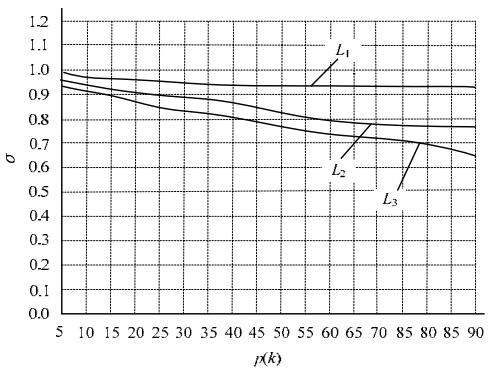


图 2 PCM 方法与 PAM 方法的查询效率比较

(下转第 56 页)