文章编号:1671 - 2021(2009)03 - 0579 - 06

基于网格索引的 Top - k偏好查询算法

孙焕良[1],姜 超[1],孙丽梅[1],廖廷悟[2]

(1.沈阳建筑大学信息与控制工程学院,辽宁 沈阳 110168; 2.中国人民银行海口中心支行,海南 海口 570105)

摘 要:**目的** 设计基于网格索引的 Top - *k*偏好查询算法,提高 Top - *k*偏好查询问题的解决效率. **方法** 利用网格索引,采用概念划分的方法,实现基于范围查询和 NN查询两种方式的 Top - *k*偏好查询算法. **结果** 通过真实数据集测试结果表明算法能够结合网格索引的优点,与基于 *R*树索引的传统算法相比,在 *k*值不断增加的情况下,查询效率提高了 50%,能适应多种空间特征数据对象集合. **结论** 网格索引可以有效处理 Top - *k*偏好查询.

关键词:数据挖掘; Top - k偏好查询;网格索引;概念划分

中图分类号: TP311.131 **文献标志码**: A

0 引 言

空间数据库管理是数据挖掘研究中的一个重点领域,空间数据不但包含大量的地理空间实体信息,而且包含了许多非空间信息如大小、类型、价格等属性. 利用空间数据库中的空间数据挖掘技术研究工具来支持和改善人类活动和生活的质量,如全球变化、社会经济可持续发展、精细农业和智能交通等[1-3]. Top - k偏好查询解决了在空间邻近区域内选择具有最好设施空间位置的问题[4].

空间索引是指依据地理对象的位置、形状或地理对象之间的某种空间关系,按一定的顺序排列的一种数据结构[5]. 现有的 Top - k偏好查询算法中,大多采用经典的空间数据结构 R树[6]来对空间数据进行索引,如文献[4]中的 SP、GP、BB、FJ等算法,其中一些算法通过在 R树上进行剪枝[7-8]来提高算法效率. 在基于 R树 Top - k偏好查询算法中,由于对 R树任意结点的访问都必须从树的根结点开始,采用深度优先的方式访问[9] R树,树的层数越高,访问的中间结点就越多,而真实的数据信息全部存放在叶子结点上,每次查询需要对整个树结构进行一次深度遍历,这样的访问方式影响了 R树的访问速度. 空间索引的性能优劣直接影响空间数据库和地理信息系统的整体性能[10]. 故笔者提出采用网格为索引结构,结合概念划分算法[11]来实现 Top - k偏好查询,并且分别采用范围查询和 NN查询进行实现. 实验测试显示了笔者提出的网格索引上的范围查询 TopRAN - G(Top - k range query based on Grid index)算法与 NN查询 TopNN - G(Top - k NN query based on Grid index)算法能够结合网格索引的优点,提高了 Top - k偏好查询的效率.

1 Top - k偏好查询

给定兴趣对象集合 D, Top - k偏好查询返回对象集合 D中具有最高评价值的 k个对象. 对象的评价值是通过空间邻近区域特征的质量进行定义,邻近区域特征包括设施或服务等. Top - k空间偏好查询实例如图 1所示.

图中表示了范围查询和 NN查询,在空间[0, 1]范围内白色点表示兴趣对象点,灰色和黑色分别表示不同种类的特征对象点. 在购房问题中,购房者不但看重商品房的质量,同时也关注商品房周围的生活环境和相应的配套设施. 图中的兴趣对象 p_1, p_2, p_3 相当于可供选择的商品房,灰色和

收稿日期: 2008 - 05 - 21
基金项目:辽宁省自然科学基金项目(20052006);辽宁省教育厅攻关计划项目(05L354)
作者简介:孙焕良(1969—),男,副教授,博士,主要从事数据库、数据仓库和数据挖掘等方面研究.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

黑色点相当于房屋周围的配套设施(幼儿园、医院、商场等).当购房者要求在一定范围内,具有较高配套设施的房屋时,位于点 p_1 的房屋成为理想的购房选择.当购房者将房屋与配套设施的距离和最小放在首位进行考虑时,位于点 p_2 的房屋则成为理想的购房选择.

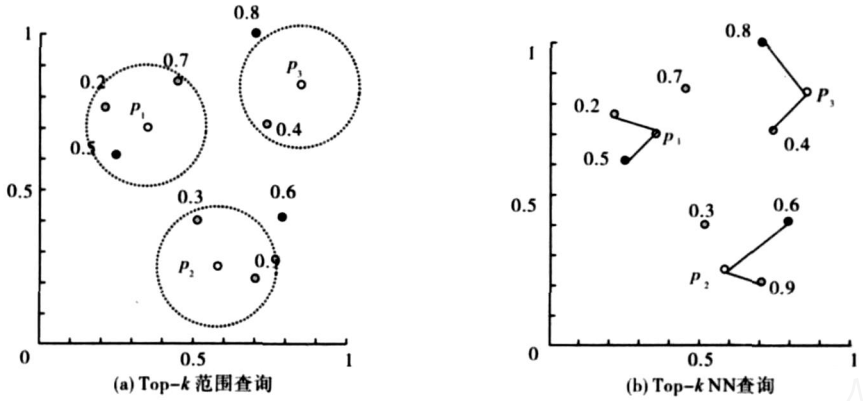

图 1 Top - k空间偏好查询实例

由于在地理实体中所包含的信息更为繁多和复杂,查询结果可能受到几个特征对象集合的共同影响,所以对于给定的兴趣对象点 $p \in D$ 和 m 种特征集合 $F_1, \cdots, F_m$ 而言,p 点所在的地理位置的评价值定义为:$\gamma^{\theta}(p) = agg\{\gamma_c^{\theta}(p) \mid c \in [1, m]\}$,其中 agg 为单调聚集运算符号,典型的聚集运算符号有:求和、求最大值和求最小值,$\gamma_c^{\theta}(p)$ 为兴趣对象点 p 在特征数据集合 Fc 和邻近条件 θ 下的评价值.两种最直观的邻近条件分别为最近邻居邻近(NN)和范围邻近,因此笔者实现了这两种度量函数下的 Top - k偏好查询.

2 基于网格索引的 Top - k偏好查询算法

Man Lung Yiu等人在提出的算法[4]中使用R树为空间数据建立索引,鉴于 R树的种种缺点,笔者考虑使用适于主存的网格结构为数据集合建立索引,并采用具有较高查询效率的 CPM 概念划分[11]算法来提高查询效率.算法基于网格索引,采用概念划分的剪枝方法,实现了基于范围查询和 NN 查询两种方式的 Top - k偏好查询算法.

2.1 概念划分算法

概念划分是网格索引中一种高效的剪枝方法,若给定一个单元格 c 和查询 q,$mindist(c, q)$ 为对象 p,$p \in c$,与 q 之间的最短距离,best_NN 为已找到的 q 的最近邻居链表,best_dist 为 k^{th} 最近邻居与查询点之间的距离.如果 $mindist(c, q) \geq$ best_dist,单元格中不可能存在比当前 NNs 更靠近查询点 q 的对象,因此可以不必查找单元格 c.基于这种思想,可将所有的单元格 c 根据 $mindist(c, q)$ 进行排序,并按照 $mindist(c, q)$ 的升序方式访问单元格.对于每个当前访问的单元格,计算其内部每个对象 p 与查询点 q 的距离 $dist(p, q)$,并根据 $dist(p, q)$ 判断是否更新 best_NN 链表.Kyriakos Mouratidis等人根据此模型得到概念划分算法.

概念划分算法将查询点周围的单元格(cell)合并成概念上的矩形(rectangle),并为每个矩形定义了层数和方向.如图 2所示.图中,U、D、L、R 分别为查询点的上、下、左、右 4个方向;层数为查询点和当前矩形之间相隔的矩形个数.概念划分算法初始化一个空堆:(1)将查询点所在单元格的 keydistance 置为 0并插入到堆中;(2)将第 0层上各个方向的矩形 DIR 根据其与查询点的最短距离 $mindist(DIR_0, q)$ 进堆.然后,反复进行出堆操作,当返回的对象为概念矩形时,将概念矩形所包含的单元格根据其与查询点之间的距离插入到堆中;当返回的对象为单元格时,则在此单元格

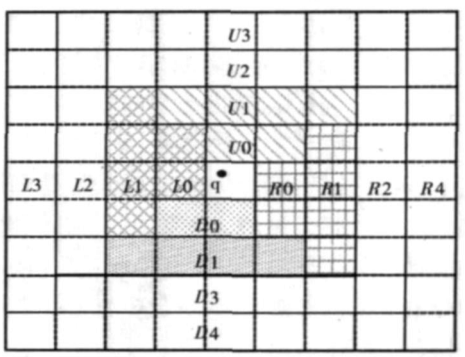

图 2 概念划分原理

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

内进行相应类型的查找.

2.2　基于网格索引 Top - k偏好查询算法

笔者结合概念划分算法的核心思想,鉴于网格索引可高效构建和访问空间数据以及适用于二维空间数据这些特点,使用网格索引进行 Top - k偏好查询.与使用 R树索引进行 Top - k偏好查询的算法不同.

(1)在初始化阶段,网格索引将单元格以及单元格内所包含的空间数据全部存储在主存中;而使用 R树为空间数据建立索引时,主存中只存储 R树的头指针; (2)当发起一个范围或 NN查询时,网格索引所查找的数据空间随着 k值的不断增大而不断复杂化;而 R树索引由于已为每一类型的空间数据建立一棵 R树,所以查找的数据空间较为单一; (3)当发起多重查询时,网格索引为每个查询点周围的单元格进行概念划分,并可能出现不同查询点的概念矩形相互交叠的情况;而 R树索引中每个查询点的查询操作是相互独立的.笔者提出的 Top - k偏好查询算法 CPTK算法很好地克服了这些问题,采用 CPM剪枝方法来减少搜索空间.为了便于理解,给出算法中符号与意义的对应关系,见表 1.

表 1　符号所代表的意义

符号	意义
$mindist(x, y)$	x与 y之间的最短距离
DIR_{lvl}	在 lvl层, DIR方向上的概念矩形
c_q	查询点 q所在的单元格
δ	单元格边的长度
influence_list	存放受查询影响区域的单元格
maxdist	存放当前查询点 q与得到各种特征集合的 NN中的最大距离
best_NN	存放每个特征集合的 NN结果

CPTK算法结合了概念划分的思想,并根据不同的查询请求进行不同的初始化处理.以下给出了在网格索引的条件下,解决 Top - k偏好查询问题中的范围查询算法与 NN查询算法. Top-RAN - G算法首先将查询点所在的单元格放入最小堆中,并将查询点周围的单元格划分成概念矩形放入最小堆中,其次反复进行出堆操作直到将所有的受影响的单元格全部遍历一遍.如下给出了网格索引下 Top - k偏好范围查询算法 Top-RAN - G的伪码.

算法 1. TopRAN - G (G, q, r)

输入:初始查询半径 r,查询点 q, Top - k参数 k

输出:最优的 k个查询结果　// G:网格索引;空间　q:查询点; r:范围查询的半径

将堆 H初始化为空

1)将以查询点为圆心, r为半径的圆内所包含的单元格存入 influence_list中

2)将 $<c_q, 0>$插入到 H中

3)将每个方向 DIR的 $<DIR_0, mindist(DIR_0, q)>$插入堆 H中

4)得到堆顶元素 X

5) if堆顶元素 X为单元格 $<c, mindist(c, q)>$

6)对于单元格中的每个特征对象 pf,如果需要,更新范围查询结果

7)将单元格从 influence_list中移除

8) else //堆顶元素 X为概念矩形 $<DIR_{lvl}, mindist(DIR_{lvl}, q)>$

9)对于 DIR_{lvl}中的每个单元格

10)将单元格 $<c, mindist(c, q)>$插入堆 H中

11)将概念矩形 $<DIR_{lvl+1}, mindist(DIR_{lvl+1}, q) + \delta>$插入堆 H中

12) while influence_list中没有单元格

当进行范围查询时

(1)根据给定的范围 r半径确定所能影响到的单元格区域,利用概念划分思想将查询点周围的单元格合并成概念矩形 rect,并根据 mindist (rect, q)插入到最小堆 H中去 (步骤 1～3); (2)反复进行出堆操作,当返回对象 X为概念矩形时,将其包含的单元格 c根据 mindist(c, q)插入到堆中;当返回对象 X为单元格时,在单元格所包含的对象链表中查找结果 (步骤 4～11); (3)当所有受影响的单元格全部搜索一遍时,如图 3所示,算法结束.该算法的主要目的在于在给定的

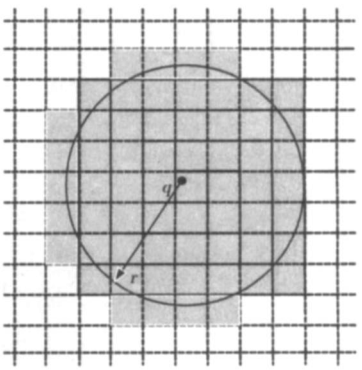

图 3　范围查询的影响区域

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

范围空间里搜索 k个具有最高评价值之和的查询点(步骤 12).

TopRAN - G算法首先将查询点所在的单元格放入最小堆中,并将查询点周围的单元格划分成概念矩形放入最小堆中,其次反复进行出堆操作直到将所有的受影响的单元格全部遍历一遍并在每个特征数据集上找到查询点的 NN. 如下给出了网格索引下采用 NN 查询的方法进行 Top - k偏好算法 TopNN - G的伪码.

算法 2 TopNN - G (G, q)

输入:查询点 q, Top - k参数 k

输出:最优的 k个查询结果 // G:网格索引空间; q:查询点

将堆 H初始化为空,全局变量 maxdist = 0

1)将 $<c_q, 0>$插入堆 H中

2)将每个方向 DIR 的 $<DIR_0, minidist(DIR_0, q)>$插入堆 H中

3)得到堆顶元素 X

4) if堆顶元素 X为单元格 $<c, minidist(c, q)>$

5)对于单元格中的每个特征对象 pf,如果需要,更新 NN 查询结果与 maxdist,并标记此单元格

6) else //堆顶元素 X为概念矩形 $<DIR_{lvl}, minidist(DIR_{lvl}, q)>$

7)对于 DIR_{lvl}中的每个单元格

8)将单元格 $<c, mindist(c, q)>$插入堆 H中

9)将概念矩形 $<DIR_{lvl+1}, minidist(DIR_{lvl+1}, q) + \delta>$插入堆 H中

10) if best_NN的数量 =k值

11) if以查询点 q为圆心, maxdist为半径的圆包含的单元格未被标记

12)将单元格插入到 influence_list中

13) while influence_list中没有单元格

当进行 NN 查询时,算法需要从每个特征数据集合中找到相应的 NN 作为 Top - k偏好查询的结果. 首先初始化全局变量 maxdist = 0, maxdist用来存放已找到 NN 中与查询点的最远距离,以下的处理思想与范围查询大致相同. (1)利用概念划分思想将查询点周围的单元格合并成概念矩形 rect,并根据 mindist(rect, q)插入到最小堆 H中去(步骤 1～2); (2)反复进行出堆操作,当返回对象 X为概念矩形时,将其包含的单元格 c根据 minidist(c, q)插入到堆中;当返回对象 X为单元格时,在单元格所包含的对象链表中查找结果,并适时更新 maxdist(步骤 3～10). (3)若每个特征数据集合都找到了一个 NN 结果,并且以查询点为圆心, maxdist为半径的圆覆盖到的单元格全部搜索一遍,如图 4所示,则算法结束(步骤 11～14). 该算法的主要目的在于在空间中搜索一个具有最高评价值之和的查询点,评价值之和由查询点在每个特征集合上的 NN 构成.

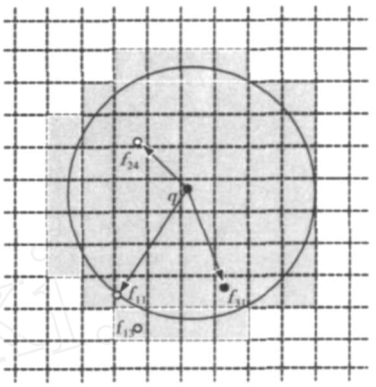

图 4 NN 查询的影响区域

3 算法性能测试

算法采用 VC ++实现,实验在处理器为 P4 2.4G Hz,主存为 1GB,操作系统为 WindowsXP 的微机上进行,实验测试的数据来自德国城市 Oldenburg的交通网络,因为现实生活中的特征对象也是分布在交通网络上. 实验测试了使用网格索引和使用 R树索引,在不同条件下,其效率发生的变化情况.

笔者研究了不同的变量参数对算法性能的影响,变量参数由表 2给出(其中默认值由粗体标出). 在每次的性能测试中,只有一个变量参数发生变化,其余的变量参数全部使用固定默认值. 用一些符号来表示各算法 TopNN - R、TopRAN - R、TopNN - G、TopRAN - G,替换图中的名称.

表 2 变量参数的范围

变量名称	变量取值范围					
查询对象数量	15	30	**45**	60	75	90
特征数据的数量	500	1 000	**5 000**	10 000	15 000	
特征集合数量, m值	1	2	3	**4**	5	6
结果数量, k值	2	4	**8**	16	32	
查询半径, r	0.002	0.005	**0.02**	0.05	0.2	0.5

图 5给出了在默认值的条件下,特征数据数量对范围算法效率的影响. 可以看出在网格索引上特征数据数量对 Top - k偏好范围查询的效率变化幅度不大;而在 R树索引上 Top - k偏好范

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

围查询的效率随着特征数据数量的增多而降低. 图 6给出了在默认变量参数值的条件下,特征数据数量对 NN 查询算法效率的影响. 可以看出在网格索引上特征数据数量对 Top - k偏好 NN 查询的效率没有很大影响;而在 R树索引上 Top - k偏好 NN 查询的效率随着特征数据数量的增多而降低.

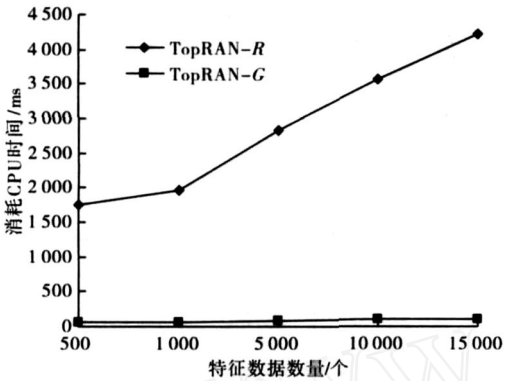

图 5　特征数据量对范围查询的影响

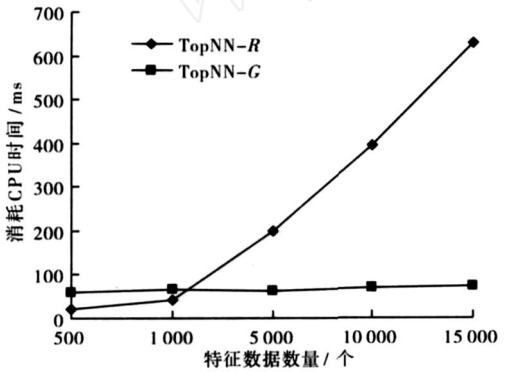

图 6　特征数据量对 NN 查询的影响

图 7给出了在默认变量参数值的条件下,特征数量对范围查询算法效率的影响. 可以看出在网格索引上特征集合数量对 Top - k偏好范围查询的效率变化幅度不大;而在 R树索引上 Top - k偏好范围查询的效率随着特征集合数量的增多而降低. 图 8给出了在默认变量参数值的条件下,特征数量对 NN 查询算法效率的影响. 可以看出在网格索引和 R树索引上特征集合数量对 Top - k偏好 NN 查询的效率随着特征集合数量的增多而降低;但在 R树索引上 Top - k偏好 NN 查询的效率变化幅度更为明显.

图 9给出了在默认变量参数值的条件下,查询数据数量对范围查询算法效率的影响. 可以看出在网格索引和在 R树索引上查询点数量对 Top - k偏好范围查询的效率随着查询点数量的增多而降低;但在 R树索引上 Top - k偏好范围查询的效率变化幅度更为明显. 图 10给出了在默认变

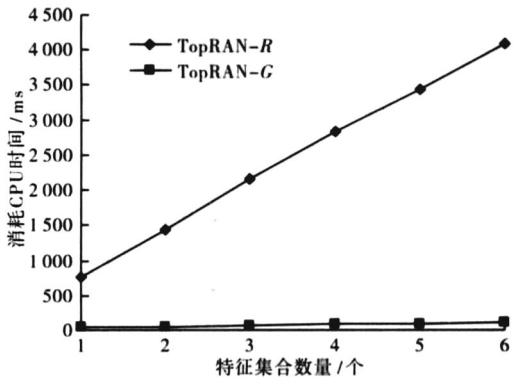

图 7　特征数量对范围查询算法效率的影响

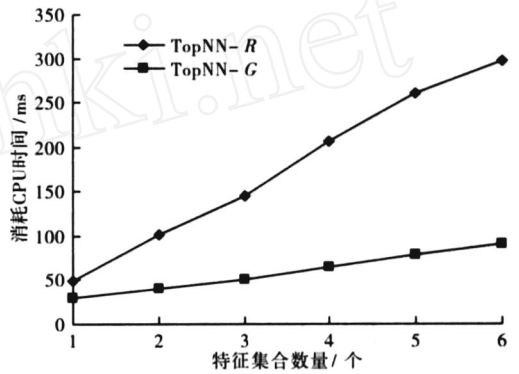

图 8　特征数量对 NN 查询算法效率的影响

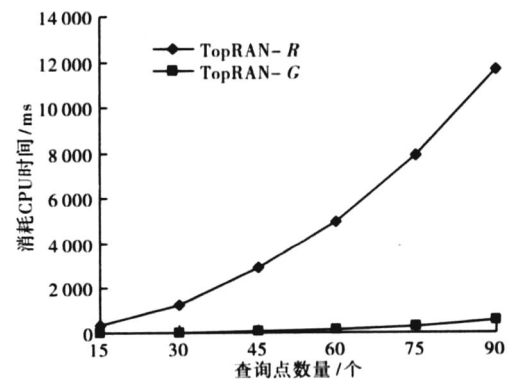

图 9　查询数据数量对范围查询算法效率的影响

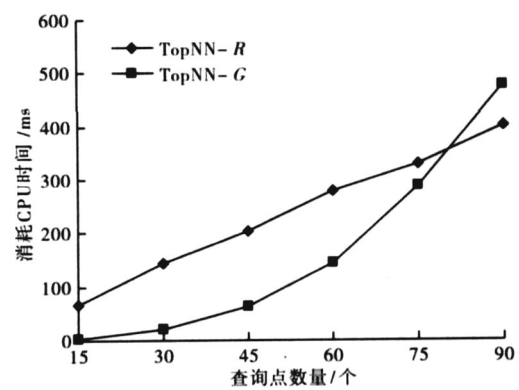

图 10　查询数据数量对 NN 查询算法效率的影响

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

量参数值的条件下,查询数据数量对 NN 查询算法效率的影响.可以看出在查询点数量比较少的时候,网格索引的 NN 查询效率要高于 R树索引的 NN 查询效率;但随着查询点数量的继续增多网格索引的 NN 查询效率低于 R树索引的 NN 查询效率.

实验测试还表明在默认变量参数值的条件下, Top - k的 k值变化,对网格索引和 R树索引上的范围查询和 NN 查询没有明显影响,但网格索引的查询效率要高于 R树索引的查询效率.

4 结 论

笔者设计了基于网格索引的 Top - k偏好查询算法,采用范围查询和 NN 查询两种方式实现了 Top - k偏好查询算法: TopRAN - G算法和 TopNN - G算法,并利用概念划分的方法来提高查询效率.实验测试显示了所提出的网格索引上的范围查询 TopRAN - G算法与 NN 查询 TopNN - G算法能够结合网格索引的优点,进行快速准确的 Top - k偏好查询.

参考文献:

[1] 孙焕良,朱叶丽,姜超.交通网络中移动对象 RNN 查询算法[J].小型微型计算机系统, 2007(8): 73 - 76.

[2] 孙焕良,朱叶丽,姜超,等.交通网络中移动对象定点 CRNN 查询算法[J].沈阳建筑大学学报:自然科学版, 2007, 23(4): 688 - 691.

[3] Man L Y, Dai X Y, Mamoulis N, et al. Top - k spatial preference queries[C]//Proceedings of the 23rd International Conference on Data Engineering. Istanbul: IEEE Computer Society Press, 2007: 1076 - 1085.

[4] 王青山.面向对象地理数据模型的研究与实践[D].郑州:信息工程大学测绘学院, 2000.

[5] Antonin G. R - trees: A dynamic index structure for spatial searching[C]// Proceedings of Annual Meeting. Boston: ACM Press, 1984: 47 - 57.

[6] Hjaltason G R, Hanan S. Distance browsing in spatial databases[J]. ACM Trans. Database Syst, 1999, 24(2): 265 - 318.

[7] Stefan B, Christian B, Daniel A K, et al. A cost model for nearest neighbor search in high - dimensional data space[C]// Proceedings of the 16th ACM SIGACT - SIGMOD - SIGART Symposium on Principles of Database Systems. Tucson: ACM Press, 1997: 78 - 86.

[8] Nick R, Stephen K, Fréédéric V. Nearest neighbor queries[C]// Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data. San Jose: ACM Press, 1995: 71 - 79.

[9] 陈述彭,周成虎.地理信息系统导论[M].北京:科学出版社, 2000.

[10] Kyriakos M, Marios H, Dimitris P. Conceptual partitioning: an efficient method for continuous nearest neighbor monitoring[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data 2005. Baltimore: ACM Press, 2005: 634 - 645.

Top-k Preference Query Algorithms Based on Grid Index

SUN Huanliang[1], JIANG Chao[1], SUN Limei[1], LIAO Tingwu[2]

(1. School of Information and Control Engineering, Shenyang Jianzhu University, Shenyang China, 110168; 2. Haikou Central SVB - Branch of the People Bank of China, Haikou China, 570105)

Abstract: To improve the problem-solving efficiency, a Top-k preference query algorithm based on grid index was designed. By indexing objects on grid structure, the conceptual partitioning method was adopted to carry out Top-k preference query algorithm which was based on the range query and the NN query. The query algorithm which was experimentally evaluated with real dataset shows that the algorithm has the advantages of the grid index, and it processes Top-k preference queries accurately and can be applied to handling multiple feature datasets in spatial database. Compared with the traditional algorithm based on R tree, the query efficiency is highly improved and the grid index is suitable for the Top-k preference query.

Key words: data mining; Top-k preference query; grid index; conceptual partitioning

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net