

空间资源索引与 top-k 查询研究

于世龙, 黄宏斌, 邓 苏[†]

(国防科学技术大学 信息系统工程重点实验室, 长沙 410073)

摘要: 以空间资源索引结构 R-tree 为基础, 考虑人们对空间资源能力指标的要求, 利用道路网络模型进行空间距离的计算, 提出了一种包含能力维度信息的空间资源索引结构和 top-k 查询算法, 并与传统遍历算法进行对比实验, 验证了查询算法的有效性。

关键词: 空间资源索引; 能力指标; 道路网络; top-k 算法

中图分类号: TP301.6 **文献标志码:** A **文章编号:** 1001-3695(2014)01-0134-03

doi:10.3969/j.issn.1001-3695.2014.01.032

Spatial resource index and top-k query research

YU Shi-long, HUANG Hong-bin, DENG Su[†]

(Science & Technology on Information System Engineering Laboratory, National University of Defense Technology, Changsha 410073, China)

Abstract: Based on the spatial resource index of R-tree, concerning about the request of resource ability guide line, this paper used the road networks model for computing the spatial distance, and proposed a spatial resource index that contained the information of ability dimensionality and the top-k query algorithm. In the end, it implemented the algorithm in the contrast of traditional traversing algorithm in the experiment, and validated the impact of the algorithm.

Key words: spatial resource index; ability guide line; road networks; top-k algorithm

随着物联网技术的发展, 人们对空间资源的搜索需求越来越大。而且在搜索的过程中, 人们往往会关注某一资源对象是否有能力来满足用户的需求。考虑到这一问题, 本文利用道路网络模型, 提出包含资源对象能力维度信息的空间资源索引结构以及 top-k 查询算法, 并通过实验与传统遍历查询算法进行对比分析, 验证了该算法的可行性和有效性。

1 相关研究工作

R-tree 通常被作为一种查询和管理空间资源的索引结构。文献[1]基于 R-tree 索引, 从用户偏好的角度出发, 提出一种可以反映用户偏好的空间资源 top-k 查询算法, 并在此基础上提出了许多改进的算法, 主要利用查询点的近邻资源特征, 将那些用户可能感兴趣的近邻给予较高的排序权重。文献[2]在 R-tree 的基础上, 提出了 IR²-tree 空间资源索引结构。该结构将文本计算与空间的欧式距离计算比较融合在一起, 在 R-tree 中每一个节点加入特殊的文本标记语言, 避免了空间资源的查询在距离与文本相似度之间频繁地导入/导出, 减少了空间资源 top-k 算法的复杂度。

用户在查询时, 往往对资源的内容匹配有着强烈的需求, 传统的 R-tree 已经不能满足人们对查询效率的要求。文献[3, 4]提出了一种将倒排索引表与传统 R-tree 结合起来的索引结构 IR-tree, 使得人们在处理用户对资源的索引请求时, 能够同时利用资源的位置信息和文本内容来进行匹配筛选。文献[5]提出一种带方向约束的资源查询算法, 解决了用户需要查询某一特定方向区域资源的问题。

以上这些文献在计算空间资源的距离时都采用的是欧式距离, 而忽略了实际的道路情况。文献[6]将道路网络模型引入到空间资源 top-k 查询中, 使得查询结果更符合实际情况。该文献在进行查询时考虑两个因素, 一个是到查询点的最短路径, 另一个是查询关键词的文本相关度。但是一个值得考虑的问题是, 用户查询某一空间资源通常是为了利用这一资源, 那么提供的查询算法所查找出来的结果是否能够令用户满意意味着这一查询是否有效。因此, 在文献[6]的基础上, 引入资源的能力信息, 这一信息表示的是资源能够满足用户某种需求的指标水平。例如, 某一用户要查找距离他最近的两家饭店, 如果仅仅是按照文本信息和位置信息进行匹配查找可能找到的饭店的口感不如距离稍远一些的饭店, 因此本文综合以下这三方面来考虑: a) 资源到查询点的距离; b) 资源的文本匹配度; c) 资源的能力指标。

2 准备工作

2.1 道路网络模型

以一个无向图 $G = (V, E, W)$ 来定义道路网络模型, 其中 V 表示顶点的集合, E 表示图的边, W 表示一个权重的集合, 即与边有关的网络距离。对于任意 $v \in V$, v 表示网络中道路的交叉点或者网络图的边界点; 边 $(v, v') \in E$ 表示网络中的一段道路; 而权重 $w \in W$ 表示一段路途的距离长度, 即 $|v, v'|$, 由于是无向图, 因此这种距离是双向的, 即 $|v, v'| = |v', v|$ 。

2.2 资源对象描述建模

定义空间对象集 P , 对于任意的 $p \in P$, 它都在道路图 G 的

收稿日期: 2013-04-26; 修回日期: 2013-05-24

作者简介: 于世龙(1988-), 男, 辽宁沈阳人, 硕士研究生, 主要研究方向为信息管理、辅助决策; 黄宏斌(1975-), 男, 教授, 博士, 主要研究方向为信息管理、人工智能; 邓苏(1963-), 男(通信作者), 教授, 博导, 主要研究方向为智能决策支持技术、人工智能(xmgzysl@126.com)。

边 E 上。本文将 p 定义为一个三元组 $(p.l, p.d, p.a)$, 分别表示资源的位置、描述文本和能力信息。空间对象集 P 的大小用 $|P|$ 表示。对于一个对象 p 与所在边的两个端点的距离分别用 $|p, v|$ 和 $|p, v'|$ 表示, 其中 v 和 v' 分别为对象 p 所在边 (v, v') 所在边的两个端点。在道路网络图 G 中, 对象 p 和 p' 的距离用 $|p, p'|$ 来表示。在边的集合 E 中, (v, v') 与 (v', v) 是相同的, 距离 $|p, v| = |v, v'| - |p, v'|$ 。因此, 知道点 p 的到一个端点的距离就可以知道点 p 到另一个端点的距离。一个道路网络模型的形象化表示如图1所示。

2.3 查询问题的描述

在图1中的 $q.l$ 为查询点所在的位置, 例如要查询距离查询点最近的宾馆, 如果以欧式距离来计算, 则查询的结果应该是点 p_4 , 而以道路网络模型来计算, 则查询的结果应该是点 p_6 。但这也是只考虑了距离和文本(即宾馆)的匹配要求, 而没有考虑到实际资源的能力情况, 在这个例子中它应该是这一宾馆的质量, 其中可能包括住宿条件、服务质量等。因此, 定义空间道路网络的 top- k 查询为 $Q_N = \langle q.l, q.d, q.a, q.k \rangle$, 其中: $q.l$ 表示查询点的位置; $q.d$ 表示查询的文本描述信息; $q.a$ 表示待查询资源的能力匹配要求; $q.k$ 表示查询结果的数量。一般认为查询点 $q.l$ 位于道路网络的边上或者可以就近找到一个边。在文献[7,8]中, 用 $\tau(p) = \alpha \times \delta(p.l, q.l) + (1 - \alpha) \times \theta(p.d, q.d)$ 来表示 top- k 查询, 其中: $\delta(p.l, q.l)$ 表示查询点与空间资源点之间距离的因素, $\theta(p.d, q.d)$ 表示两者之间文本相似度这一因素。文献[6]利用公式:

$$\tau^*(p) = \frac{\theta(p.d, q.d)}{1 + \alpha \times \delta(p.l, q.l)}$$

来表示 top- k 查询, 并且通过实验验证了其在道路网络模型中的良好衡量标准。本文定义 $\tau(p)$ 为

$$\tau(p) = \frac{\theta(p.d, q.d) + \lambda(p.a, q.a)}{1 + \alpha \times \delta(p.l, q.l)} \quad (1)$$

其中: $\lambda(p.a, q.a)$ 表示查询的能力要求与空间资源能力信息的匹配指标。 $\alpha \in R^+$, 分别表示三者的权重。从式(1)中可以看出, 如果 $\alpha = 0$, 则位置要求将被弱化为0; 如果 $\alpha > 1$, 则位置的重要性在文本和能力匹配之上。

网络距离接近度 δ 表示空间对象与查询点距离的重要性, 因此其定义如下:

$$\delta(p.l, q.l) = \|p.l, q.l\| \quad (2)$$

文本相似度 θ 表示空间对象文本描述与查询的文本匹配程度^[9]。文本相似度定义为

$$\theta(p.d, q.d) = \frac{\sum_{t \in q.d} w_{t,p.d} \times w_{t,q.d}}{\sqrt{\sum_{t \in p.d} (w_{t,p.d})^2 \times \sum_{t \in q.d} (w_{t,q.d})^2}} \quad (3)$$

权重 $w_{t,p.d} = 1 + \ln(f_{t,p.d})$, $f_{t,p.d}$ 表示词条 t 在 $p.d$ 中出现的频率, 并且权重 $w_{t,q.d} = \ln(1 + \frac{|P|}{df_t})$ 。其中: $|P|$ 表示所抽取的集合中资源对象的数目, df_t 表示在其描述文本中包含 t 的数目。能力匹配度 λ 表示空间对象的能力指标与查询要求的匹配程度, 其定义为

$$\lambda(p.a, q.a) = \ln\left(\left|\frac{p.a}{q.a}\right|\right) \quad (4)$$

式(4)表示当空间对象的能力指标不能满足查询要求时, 其能力匹配度 $\lambda < 0$; 相反, 如果能够满足查询要求, 则 $\lambda > 0$ 。

3 研究方法

从索引结构和 top- k 查询算法两个方面来进行道路网络模型中空间资源的查询研究。

3.1 索引结构

一般情况下, 包括文本描述信息的空间资源索引结构有文献[3,4]中提到的利用倒排文档的 IR-tree, 以及文献[10]中所提到的道路网络框架。空间资源索引的示例如图2所示。

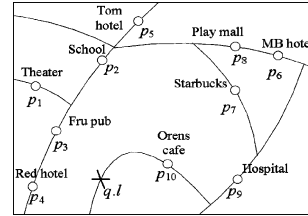


图1 道路网络

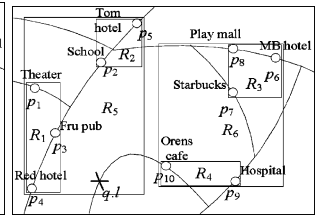


图2 空间资源索引示例

在 IR 树索引结构的基础上, 本文提出 AIR-tree (ability inv-File R-tree), 它不仅包含了空间位置信息和文本描述信息, 同时增加了空间对象的能力维度信息, 如图3所示。

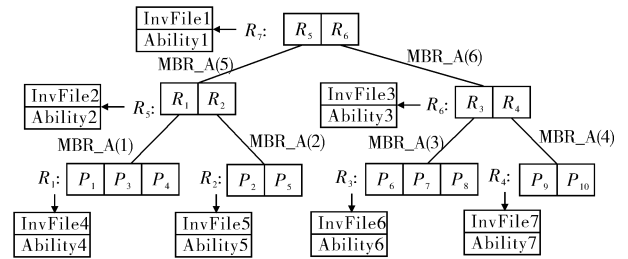


图3 AIR-tree 索引结构

在计算空间对象之间的距离时要用到道路网络模型中节点的邻接关系 (v, v') 以及节点间的道路信息, 其中道路信息包括了道路所包含的空间对象信息。图3中的 MBR 是由若干条相邻接的空间对象所在的边 (v, v') 组成的, 其中 $MBR_A(i)$ 表示的是该 MBR 的能力指标, 在这里 $MBR_A(i) = \max(A(p_j), p_j \in MBR(i))$, 取最大值是为了避免遗漏某个空间对象。

3.2 道路网络遍历查询

基本的查询算法将 Dijkstra's 算法在位置查询的邻接上进行了一些扩展。 K 个最优解是以一个综合指标值降序的方式排列在一个堆中的。当执行过程中不能找到一个比第 k 个对象的分数值高的对象或者整个网络被遍历结束则算法停止。

算法1是道路网络遍历查询处理算法。该算法在文献[6]的基础上进行修改, 其输入是一个查询 Q_N , 返回值是前 k 个 $\tau(p)$ 值最大的空间对象。第4行访问道路网络, 找出 $q.l$ 所在的边 (v, v') 。第5行是计算 $q.l$ 与节点 v 和 v' 的距离, 第6行节点 v 和 v' 以到 $q.l$ 距离增加的顺序加入到 N 中, 并且标记已经访问。第7行 findCandidates ($ID_{(v,v')}, q.d, q.a, \varepsilon$) 过程用来找到文本描述和能力匹配在边 (v, v') 上的候选空间对象 C , 这个对象的 $\tau(p)$ 分数高于 ε ($H^{q,k}$ 中的第 k 个分数), 然后添加到 $H^{q,k}$ 。当 $H^{q,k}$ 中插入一个新对象时, $H^{q,k}$ 中的第 k 个对象被移除, 并且 ε 更新为新的第 k 个对象的分数。该算法持续访问距离 $q.l$ 最近的节点 v (第9行), 处理还没有访问到的节点 v 。当整个网络被遍历结束或者距离其他未访问的空间对象的最小网络距离的综合分数小于或等于第 k 个对象的分数。一个

空间对象的综合分数 $\tau^-(v)$ 是通过查询点与对象的距离,最大文本匹配度 ($\theta=1$),以及能够满足用户的能力匹配 $q.a=p.a$,

从而 $\lambda=0$,即 $\tau^-(v) = \frac{1}{1 + \alpha \times \delta(v.l, q.l)}$ 。

算法 1: QueryProcessingAlgorithm(Query Q_N)

输入:top- k spatial keyword query on road networks based on ability

$Q_N = \langle q.l, q.d, q.a, q.k \rangle$ 。

输出:Report the top- k objects found。

1 MaxHeap $H^{q.k} \leftarrow \text{space}$ //q.k best objects in decreasing order of τ

2 $\varepsilon \leftarrow 0$ //k-th score in $H^{q.k}$, $|H^{q.k}| < q.k, \varepsilon = 0$ 。

3 MinHeap $N \leftarrow \text{space}$ //vertices v in increasing order of $|v, q.l|$ 。

4 $(v, v') \leftarrow \text{network edge in which } q.l \text{ lies}$ 。

5 compute $|v, q.l|$ and $|v', q.l|$ 。

6 insert v and v' into N , mark (v, v') as visited。

7 $C \leftarrow \text{findCandidates}(\text{ID}_{(v, v')}, q.d, q.a, \varepsilon)$ 。

8 update $H^{q.k}$ (and ε) with $p \in C$ 。

9 $v \leftarrow N.\text{pop}()$ //Vertice v in N with minimum $|v, q.l|$ 。

10 while $v \neq \text{space}$ and $\left(\frac{1}{1 + \alpha \times \delta(v.l, q.l)} \leq \varepsilon \right)$ do

11 for each non-visited adjacent edge (v, v') of v do

12 $C \leftarrow \text{findCandidates}(\text{ID}_{(v, v')}, q.d, q.a, \varepsilon)$

13 update $H^{q.k}$ (and ε) with $p \in C$ 。

14 insert v' into N , mark (v, v') as visited

15 end for

16 end while

17 return $H^{q.k}$ 。

这里之所以是以 $\tau^-(v)$ 作为终止条件,是因为在 (v, v') 这条边上,即使有空间对象能够满足能力匹配的要求,但是由于起点 v 到查询点的距离是最小的,并且已经假设文本匹配度 $\theta=1$,如果 v 到查询点的分数 $\tau^-(v)$ 都不大于 ε ,则这条边上的其他对象的分数更不能大于 ε ,因此继续寻找下一个未被访问到的端点 v 。

4 实验评估

在本章利用本文提出的查询算法与传统的网络图遍历算法进行对比分析。数据集采用真实数据 California place names Sequoia 和 Streets 数据(<http://www.rtreportal.org>),主要是描述美国加利福尼亚州的某个地方的真实空间数据以及街道数据集。用计算机随机生成 0~1 之间的数来表示每一个空间资源对象的能力指标。查询点的能力要求取 $q.a=0.5$ 。采用电脑主要配置为:Pentium® Dual-core CPU 3.06 GHz, 2 GB RAM 内存。对参数的设定如表 1 所示。

表 1 参数设置

参数	取值
所求结果数 k	10, 20, 30, 40, 50
关键字个数	1, 2, 3, 4, 5
参数 α	0.01, 0.1, 1, 10, 100

实验结果如图 4~6 所示。

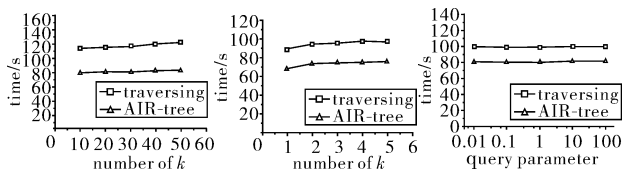


图 4 选取不同的

图 5 选取不同的

图 6 选取不同的

k 值对比分析

关键字个数对比分析

参数 α 对比分析

从上面的实验结果中可以看出,本文提出的 AIR-tree 索引结构的查询时间是优于传统的道路网络遍历算法的,在图 4 中随着 k 值的增大,查询时间也会相应地有所增加,这是因为所需要的结果集增大了;图 5 中随着关键字个数的增加,进行比较匹配的次数增多,所以查询时间也会增大;图 6 中随着参数 α 的变化,查询时间的变化不是十分明显。

5 结束语

空间对象作为一种资源接入到物联网中,成为网络中的主体,人们对空间资源的查询需求越来越广泛。本文以空间资源索引为基础,考虑到人们对空间资源能力指标的要求,利用道路网络模型,提出了一种包含能力维度信息的空间资源索引结构和查询算法,并与传统遍历算法进行对比实验,验证了其可以应用于具体环境中。

本文还有许多没有研究到的地方。例如,如何对空间资源的能力指标进行综合评估,才能更加准确地表示资源的真实情况;而且本文考虑的是静态资源,如果是动态资源,将如何进行索引的构建,这些都是在未来的工作中需要不断研究的。

参考文献:

- [1] YIU M L, DAI Xiang-yuan, MAMOULIS N, *et al.* Top- k spatial preference queries [C]//Proc of the 23rd IEEE International Conference on Data Engineering. 2007:1076-1085.
- [2] FELIPE I D, HRISTIDIS V, RISHE N. Keyword search on spatial databases [C]//Proc of the 24th IEEE International Conference on Data Engineering. 2008:656-665.
- [3] CONG Gao, JENSEN C S, WU Ding-ming. Efficient retrieval of the topk most relevant spatial Web objects [C]//Proc of VLDB Endowment. 2009:337-348.
- [4] MOHAN A K, YELISALA H, PRASAD M K. IR Tree-an adept index for handling geographic document searching[J]. International Journal of Engineering Research and Applications, 2012,2(3): 1433-1438.
- [5] LI Guo-liang, FENG Jian-hua, XU Jing. DESKS: Direction-aware spatial keyword search [C]//Proc of the 28th IEEE International Conference on ICDE. 2012:474-485.
- [6] ROCHA-JUNIOR J B, NØRVÅG K. Top- k spatial keyword queries on road networks [C]//Proc of the 15th IEEE International Conference on EDBT. New York:ACM Press, 2012:168-179.
- [7] LI Zhi-sheng, LEE K C, ZHENG Bai-hua, *et al.* IR-tree: an efficient index for geographic document search [C]//Proc of IEEE Transactions on Knowledge and Data Engineering. 2010:585-598.
- [8] ROCHA-JUNIOR J B, GKORGKAS O, JONASSEN S, *et al.* Efficient processing of top- k spatial keyword queries [C]//Proc of the 12th International Symposium on SST. 2011:205-222.
- [9] ZOBEL J, MOFFAT A. Inverted files for text search engines [J]. ACM Computing Surveys, 2006,38(2):6.
- [10] PAPADIAS D, ZHANG Jun, MAMOULIS N, *et al.* Query processing in spatial network databases [C]//Proc of the 29th International Conference on Very Large Data Base. 2003:802-813.