文章编号:1001 - 2486(2005)03 - 0081 - 05

基于尺度选择性的空间数据源选择与预取策略*

陈 荦,吴秋云,景 宁,唐 宇
(国防科技大学 电子科学与工程学院,湖南 长沙 410073)

摘 要:合理高效的数据源选择策略是提高空间信息检索系统效率的重要因素之一。针对以往研究中对空间尺度语义关注程度的不足,提出一种结合空间尺度语义选择检索数据源的方法。该方法综合考虑人眼视觉特性,计算出参与分布式查询计划生成的数据源集合,过滤在尺度意义上对最终查询结果无影响的数据源,并通过将数据源的简化表达进行预取,进一步减少了查询的整体执行代价。实验结果表明,提出的方法在大尺度查询和小尺度查询时均具有良好的性能。

关键词:空间信息检索;数据源选择;空间信息系统
中图分类号:TP391 文献标识码:A

Geospatial Data Sources Selection and Prefetching Strategy Based on Scale Selectivity

CHEN Luo, WU Qiu-yun, JING Ning, TANG Yu
(College of Electronic Science and Engineering, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Effective data source selection strategy is one of the important issues of the spatial information retrieval. This paper proposes a method to select data source based on spatial scale semantics. The method determines the data source set for the distributed query plan fully considering the human vision features, filters those data sources which don't contribute to the final result. Through prefetching the simplified representation of data sources, the method can further reduce the overall cost of a query. The experiment shows that the proposed method has favorable performance both in small and large size query.

Key words: spatial information retrieval; data source selection; spatial information system

如何面向空间信息的特点,提高空间信息应用系统的互操作性和可扩展性,是当前该领域的一个研究热点。空间信息栅格 SIG 是我国在该领域的一个重要研究项目[1],其中一个重要的技术问题是如何在大量分布的空间信息资源中选取那些对于分布式空间信息查询有效的数据源,并剪除无关的数据源。

目前国际上对于该问题的主要研究方法是基于数据集成领域的成果进行扩展。TSIMMIS[3,4]等系统对分布数据资源的内容的查询能力进行建模,在较细的粒度上实现对不同数据资源的区分,在生成查询计划时进行有效剪枝。Versatile[5]从信息检索的角度出发,通过相似度的排序来过滤不相关或弱相关的数据源。上述方法能够在一定程度上解决空间信息集成问题,但是,由于空间信息具有丰富的语义,因此,在结合空间特征(如拓扑和尺度)进行信息集成时,仍然有明显的不足。为此,文献[6]在数据源描述中添加了空间语义;Gupta[7]等提出了基于 XML 优良的数据描述能力实现不同种类空间信息的统一描述的目标。但是,上述的方法同样未考虑诸如空间尺度等语义对于集成结果的影响。

空间尺度是空间数据重要的语义之一,对于以绘制为主的空间数据查询具有重要的影响。当显示尺度远小于查询尺度时,如果仍然向所有相关数据源发送完整的查询,则处理和传输对最终表达结果无贡献的数据将引起查询效率的降低。因此,面向以绘制为主的分布式空间查询,有必要结合尺度因素研究其数据源选择策略。本文将综合考虑空间尺度特征对分布式空间信息检索的影响,研究面向地图绘制的分布式查询中的数据源选择问题,并采用数据预取策略,进一步提高数据源选择的效率。

* 收稿日期:2004 - 09 - 29
基金项目:国家 863 高技术研究发展计划基金项目(2002AA134010;2002AA134020)
作者简介:陈荦(1973—),男,博士生。

1 空间数据源的尺度选择模型

在面向以绘制为主的分布式空间查询中,影响数据源对最终结果贡献的因素主要有三个:数据源的尺度,查询窗口的尺度和结果显示窗口的尺度。基于这个认识,建立一个包含上述三个因素的数学模型,计算某数据源在某查询条件下对于最终结果的影响程度。下面明确要用到的几个基本概念:

定义1 视口(view port),指最终查询结果的显示窗口,记为 V,对于二维空间,定义其为矩形,如 128×128 的显示窗口。

定义2 查询窗口(query window),指显示窗口所对应的实际地理空间范围,记为 W,对于二维空间,定义其为矩形,如 $10\text{km}\times10\text{km}$ 的范围。

定义3 尺度函数(scale function),是计算某对象尺度的函数,记为 $SF(x)$,其中,x 为空间对象。尺度是衡量空间对象某几何特征量大小的量,对于二维空间对象,选择尺度函数为求面积函数,即 $SF(x)=Area(x)$。尺度函数的定义可以扩展到高维空间,其具体内容对形式表述无影响。

以往的研究认为,从纯几何角度考虑,影响空间对象在某一尺度下显示方式的因素主要是对象尺度与查询窗口尺度的相对比值,当该值小于某一门限 T_f 时,对象即被过滤。上述模型中门限是固定的,没有考虑到视口大小对对象表示的影响,而在实际中,视口大小对于空间对象表示的影响不可忽略。

通过对人眼视觉现象的观察,并结合视觉生理、心理学等方面的研究成果,发现:① 人眼对图像边缘区信息的失真很敏感;② 人眼对图像平滑区信息的失真比较敏感;③ 人眼对图像纹理区信息的失真不敏感[8]。视口尺度对空间对象过滤门限的影响主要体现在以下几个方面:

(1) 门限 T_f 的确定。对于包含相同数目对象的视口来说,当视口尺度扩大时,对象分布较稀疏,视口具有图像平滑区的视觉特征,对象的过滤对视觉效果的影响较大,T_f 应较小;当视口尺度缩小时,对象分布较密集,视口具有图像纹理区的视觉特征,对象的过滤对视觉效果的影响较小,T_f 应较大。因此,综合考虑人眼的视觉特性,在视口尺度较小的情况下,可以过滤更多的对象而对最终表现结果无大的影响。也就是说,过滤对象的门限 T_f 不再是固定不变的,而是随着视口的尺度大小而变化。

(2) 门限的变化率。门限的变化率也随着视口的大小改变,在保持视觉效果基本不变的前提下,视口尺度较大时,门限变化率较小;视口尺度较小时,门限变化率较大。

(3) 边界条件。根据对象显示的物理意义,当视口尺度趋向于无穷大时,无论尺度多小的对象都要显示,当视口尺度小于某一值 b 时,所有对象都被过滤(如视口尺度小于人眼分辨率时)。

设所研究的对象为二维空间对象,尺度函数为求面积函数,则可构造门限 T_f 的数学模型:

$$T_f(V)=\begin{cases}\dfrac{r}{SF(V)}=\dfrac{r}{w_v\cdot h_v}, & (w_v\geqslant b)\cap(h_v\geqslant b)\\ 0, & (b>w_v)\cup(b>h_v)\end{cases}\tag{1}$$

其中,r 是视觉分辨力常数,r 越大,过滤的数据对象越多,反之,过滤的数据对象越少;w_v 和 h_v 是视口 V 的两个边长。由此可看出,可以通过计算空间数据源与查询窗口尺度函数值的比,并与相应视口下的 T_f 值进行比较来确定该空间数据源的表现方式。由此,定义分布空间数据源的尺度选择性因子为:

定义4 空间数据源 s 的尺度选择性:

$$SS(s)=\begin{cases}\dfrac{SF(s)}{SF(W)\,T_f(V)}=\dfrac{SF(s)\cdot w_v\cdot h_v}{w_w\cdot h_w\cdot r}, & (w_v\geqslant b)\cap(h_v\geqslant b)\\ 0, & (b>w_v)\cup(b>h_v)\end{cases}\tag{2}$$

$SS(s)$ 的取值决定了其表现方式和处理方式,在实际应用中,空间数据源的表达具有多种细节层次,为讨论简单起见,本文只设定三个层次:不表达、简化表达、完全表达,则相应的尺度选择性门限为两个:①过滤门限:判定不表达与表达,记为 t_f;②简化门限:判定简化表达与完全表达,记为 t_s。

- 当 $SS(s)<t_f$ 时,该数据源被过滤,生成查询计划时将不对该数据源生成子查询;
- 当 $t_f\leqslant SS(s)<t_s$ 时,该数据源将简化表达,生成查询计划时将生成一个简化的子查询;
- 当 $SS(s)\geqslant t_s$ 时,该数据源被完全表达,将生成完整的子查询。

2　数据源选择算法与代价分析

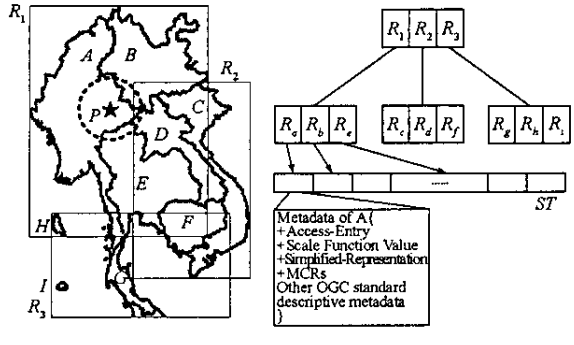
设 D 为所有数据源的集合,Q 为针对全局模式提交的查询,一般的数据源选择算法在 D 中选择出与 Q 相关的一个数据源子集 $D' = \{D_1, D_2, \cdots, D_N\}$,其中,$N$ 为相关数据源个数,C_i 表示在数据源 i 处理局部查询的代价,S_i 表示局部查询在数据源 i 产生的结果集大小,α 表示在网络中单位数据的传输代价,则查询 Q 的总体代价估计 C_Q 为:

$$C_Q = \sum_{i=1}^{N} (C_i + \alpha S_i) = \sum_{i=1}^{N} C_i + \alpha \sum_{i=1}^{N} S_i \tag{3}$$

尺度扩展的数据源选择算法在此基础上,计算 D' 集合中各元素的尺度选择性因子,剪除小于尺度门限的数据源,进一步减少参与查询的数据源个数。

算法首先扩展数据源尺度信息形成数据源元数据模型,建立扩展的 R 树空间索引,如图 1(a)所示。图中,$\{A, B, C, D, E, F, G, H, I\}$ 为分布的空间数据源,R_x 为数据源 x 的最小包围矩形,ST 是存储了所有数据源元数据信息的查找表。该 R 树包含两类结点记录:tree-entry 是非叶子结点上的记录,数据结构为 < MBR, child-pointer >,其中,MBR 是包含以该记录为根的子树的最小包围矩形,child-pointer 是指向其子树的指针;source-entry 是叶结点上的记录,数据结构为 < MBR, Source-metadata-entry >,其中,MBR 是包含该数据源区域的最小包围矩形,Source-metadata-entry 是指向 ST 中某记录的指针。

进行数据源选择时,在搜索树的每一层时,不但要进行矩形求交运算,判断搜索子树或者叶结点是否与检索范围相交,还要对相交的结果进行尺度选择性计算。我们采用两个链队列来存储搜索的中间结果:NormalRepSet 存储完全表达的数据源,SimpRepSet 存储简化表达的数据源。具体的选择算法 SSSE (source selection with scale extension)的步骤见图 1(b)。



(a)

```
SSSE(R:R-tree-root, Q,V,tf,ts)                    /*Source Selection
                                                    with Scale Extension*/
01  BEGIN
02    FOR(each Entry in R)DO
03      I=intersection(Entry.MBR,Q)
04      IF I≠ Φ THEN
05        SWITCH(Entry. Type)
06          CASE Source-Entry:
07              Calculate SS (Entry)
08              IF SS (Entry)>ts THEN
09                Add To NormalRepSet(Entry.Source-metadata-Entry)
10              ELSE IF SS (Entry)>ts and SS (Entry)<=ts THEN
11                Add To SimpRepSet(Entry.Source-metadate-Entry)
12              END-IF
13          CASE Tree-Entry:
14              SET Effective-Area=area(I)
15              Calculate SS(I)
16              IF SS(Entry)>tr THEN
17                SSSE(Entry.Child-Pointer, Q, V, tf, ts)
18              END-IF
19        END-SWITCH
20      END-IF
21    END-FOR
22  END
```

(b)

图 1　分布数据源的扩展 R 树索引与数据源选择算法

Fig. 1　Extended R-tree index on distributed data sources and source selection algorithm

由算法 SSSE 可以看出,如果搜索结点(树结点或者叶子结点)的空间范围与检索范围不相交或者相交的尺度选择性小于过滤门限,则该结点及其子树都将被剪枝。通过这种尺度扩展的数据源选择算法,既能剪除与检索范围不相关的数据源,又能够剪除因尺度太小而对最终查询显示结果无贡献的数据源,因此能够进一步减少参与全局查询的数据源数目,提高空间信息检索的整体查询效率。

设 N' 为进行完全表达处理的数据源数目,N'' 为进行简化表达的数据源的数目,则有 $N' + N'' \leqslant N$,另外,令 d_i 表示在数据源 i 上进行简化计算的代价,则基于 SSSE 的查询总体代价 $C_{Q'}$ 为

$$C_{Q'} = \sum_{i=1}^{N'} (C_i + \alpha S_i) + \sum_{j=1}^{N''} (C_j + \alpha S_j + d_j) = \sum_{i=1}^{N'+N''} C_i + \sum_{j=1}^{N''} d_j + \alpha \left(\sum_{i=1}^{N'} S_i + \sum_{j=1}^{N''} S_j \right) \tag{4}$$

$C_{Q'}$ 同样由两部分构成，一是计算代价，包括进行查询计算的代价和进行简化计算的代价；二是传输代价。与式(3)比较，当 $N'' = 0$ 时，有 $N' \leqslant N$，易知 $C_{Q'} \leqslant C_Q$；当 $N'' > 0$ 时，有 $N' \leqslant N$，$\sum_{i=1}^{N+N'} C_i \leqslant \sum_{i=1}^{N} C_i$，在网络速度相对较慢的情况下，有 $\sum_{j=1}^{N'} d_j + \alpha\left(\sum_{i=1}^{N'} S_i + \sum_{j=1}^{N'} S_j\right) < \alpha \sum_{j=1}^{N} S_j$，从而 $C_{Q'} \leqslant C_Q$。

综合上述两种情况，可以得出：在网络速度相对较慢的情况下，基于 SSSE 的查询总体代价将小于一般方法的总体代价。

为了提高分布式查询的整体响应性能，在集中器上对结果集进行预取(或具体化)是常用的方法，但是由于空间数据量很大，因此实际中有效的预取难以实现。通过对(4)式的进一步分析可知，空间数据源简化表达产生的结果集很小，因此在集中器上对之进行缓存及预取是可行的。我们采用的基本方法是扩展数据源元数据模型，将数据源的简化表达作为其元数据的一个复合属性，并注册到系统目录中去。在执行关于简化表达数据源的子查询时，不需要访问远端的数据源，而是直接访问系统目录中该数据源对应项的简化表达属性，从而在本地获得查询结果。基于这种方法，减少了远端数据源的查询代价和从远端数据源获取数据的网络传输代价。数据源的简化表达可以用任何支持空间对象多尺度表达的数据结构生成。为了能够支持连续尺度变化，我们采用了对多线进行递归二分的数据结构 BLG tree，并将之作为数据源的简化表达。

带简化表达预取的尺度扩展数据源选择查询总体代价估计为：

$$C_{Q''} = \sum_{i=1}^{N'}(C_i + \alpha S_i) + \sum_{j=1}^{N'} g_j = \sum_{i=1}^{N'} C_i + \sum_{j=1}^{N'} g_j + \alpha \sum_{i=1}^{N'} S_i \tag{5}$$

其中，g_i 是在本地执行数据源 i 简化表达的代价。通过与(4)式的比较易得 $C_{Q''} \leqslant C_{Q'} \leqslant C_Q$，从而可知，应用带预取的方法，可以进一步减少查询代价。

3 实验与结果分析

为了验证本文提出的方法的有效性，我们设计了一组实验，在模拟环境中比较三种数据源选择方法的性能。这三种方法是：(1) R-tree，即没有尺度选择和缓冲支持的数据源选择方法；(2) RSE，集成 R-tree 索引的尺度扩展数据源选择方法；(3) RSEP，带简化表达预取的集成 R-tree 索引尺度扩展数据源选择方法。实验数据取自 DCW(Digital Chart of the World)[9]，其中包括中国和南亚部分的数据，因为该部分数据的几何特征较丰富。实验将其中每个行政区模拟为一个分布的数据源，共模拟 359 个数据源。

我们仿真了在不同分辨力参数和视口尺寸下的查询情况。对每一种情况，构造不同尺寸的查询窗口进行查询，统计在该查询下三种方法所命中的必须远程访问的数据源个数，并记录相应的网络数据传输量。考虑到空间数据的分布并不均匀，在每一组查询测试中，将查询窗口放置在 10 个不同的位置进行测试，并将各结果的统计平均作为该组查询的结果。

图 2(a)表示了在 $r = 4$，$V = 200 \times 200$，$t_f = 1$，$t_s = 100$，$b = 100$ 情况下的测试结果。可以看出，在查询窗口较大时，RSE 方法需要访问的远程数据源数目要小于 R-tree 方法；但是在查询窗口较小时，二者差别不大，这是由于在查询窗口较大时，一些数据源的尺度选择性小于门限而被过滤了，而在查询窗口较小时，过滤的数据源比较少。RSEP 的综合性能是最好的，在查询窗口较大和较小时都保持了一个相对平均的远程数据源访问数目，这是因为在查询窗口较大时，不但有被过滤的数据源，还有被简化表达的数据源，这部分数据源由于采用了预取策略，其数据不必向远程数据源获取；而在查询窗口较小时，由于被过滤的数据和简化表达的数据源不多，需要完全表达的数据源较多，查询代价反而有所增加。

图 2(b)表示了在 $r = 4$，$V = 400 \times 400$，$t_f = 1$，$t_s = 100$，$b = 100$ 情况下的测试结果。可以看出，当视口变大时，需要远程访问数据源的数目增加了，相应的网络数据传输量也增大了。

图 2(c)表示了在 $r = 8$，$V = 200 \times 200$，$t_f = 1$，$t_s = 100$，$b = 100$ 情况下的测试结果。可以看出，当分辨力参数 r 变大时，远程访问数据源数目减少了，表明被过滤和简化的数据源数目增多。

由以上可以得出结论：RSEP 方法在不同尺度的查询窗口下均有良好的数据源选择性能；RSE 方法

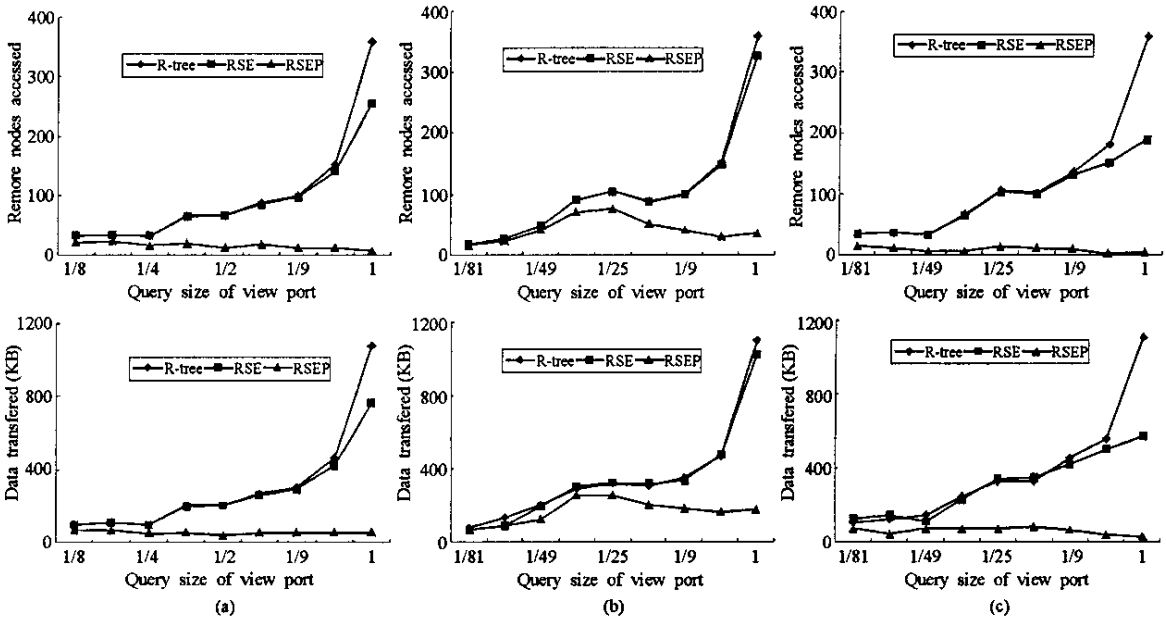


图 2　不同分辨力参数和视口尺寸下的实验结果

Fig.2　Experiment results in different resolution parameter and view port size

在大尺度查询窗口下可以有效地过滤小尺度的数据源,具有较好的数据源选择性能。

4　结束语

本文综合考虑分布式空间信息检索中人眼视觉特性与查询窗口、显示视口的关系,建立了一个空间信息资源尺度选择数学模型,提出了基于该模型的分布式空间数据资源选择策略。该策略能根据查询窗口的尺度和显示视口的尺度对分布的空间数据源进行动态选择,并对数据源的简化表达进行预取,进一步降低了分布式地图获取的代价。实验结果表明,本文提出的策略在大尺度查询和小尺度查询时均具有良好的系统响应性能。该方法带来的不足是,由于简化计算的影响,被简化数据区域的表现质量有所下降,对此,我们将进一步研究减少视觉质量损失的方法。目前,该策略与方法已被初步应用到面向城市空间信息应用的原型系统中,并期望在面向 Web 的分布式空间信息服务(如 WMS,WCS,WFS)中得到广泛的应用。

参 考 文 献:

[1] **唐宇,陈荦,何凯涛,等. 空间信息栅格 SIG 框架体系与关键技术研究[J]. 遥感学报,** 2004,8(5):425 – 433.

[2] Li C, Yerneni R, Vassalos V, et al. Capability-based Mediation in TSIMMIS[A]. Proceedings ACM SIGMOD[C], Seattle: ACM Press, 1998: 564 – 566.

[3] Yerneni R, Li C, Molina H G, et al. Computing Capabilities of Mediators[A]. Proceedings of ACM SIGMOD[C], Philadelphia: ACM Press, 1999: 443 – 454.

[4] **钟志农,李军,景宁,等. 数字图书馆中地理信息系统的设计与实现[J]. 国防科技大学学报,** 2002,24(6):87 – 90.

[5] Wang X L, Wen J R, Luan J F, et al. A Method to Query Document Database by Content and Structure[J]. Journal of Software, 2003,14(5): 976 – 983.

[6] Ishikawa Y, Ryu G, Kitagawa H. Integration of Spatial Information Sources Based on Source Description Framework [A]. Proceedings of the 7th DASFAA[C], Hong Kong, 2001: 160 – 161.

[7] Gupta A, Marciano R, Zaslavsky I, et al. Integrating GIS and Imagery through XML-based Information Mediation[A]. Proceedings of Integrated Spatial Databases[C], Portland: Springer, 1999: 211 – 234.

[8] Wang X Y, Yang H Y. A Fast Image Coding Algorithm Based on Human Visual System[J]. Journal of Software, 2003,14(11): 1964 – 1970.

[9] Digital Chart of the World Data Server[DB]. http://www.maproom.psu.edu/dcw/.