基于分布式GIS的一种矢量数据查询优化方法

尚艳玲[1,2], 徐旭东[2]
(1. 北京工业大学计算机学院, 北京　100022; 2. 安阳师范学院,安阳　455000)

摘要: 分布式GIS的海量数据和有限带宽的网络资源之间矛盾日益突出,远程矢量数据的查询和海量数据传输成为解决问题的关键。在分析海量数据查询特点的基础上,结合空间数据库中数据层的查询优化方法和代价函数,提出了一个在客户端定义并使用代价函数优化动态规划的矢量数据查询方法,同时在客户端实现数据拓扑差查询。实践证明,该方法能够有效提高查询速度和保持全局网络负载的良好性能。

关键词: 分布式GIS; 查询优化; 代价函数
中图分类号: TP 311　**文献标识码**: A　**文章编号**: 1001-070X(2008)03-0100-04

0　引言

作为获取、处理、管理和分析地理空间数据的重要工具、技术和学科,地理信息系统(Geographic Information Systems,GIS)自诞生以来就得到了广泛关注,并伴随着计算机科学技术的发展而迅猛发展。然而,数据信息的海量增长和有限带宽的网络资源之间矛盾日益突出,导致需求信息无法充分利用。

分布式GIS的数据表现、查询和存储逻辑部署在不同的终端。对于当前网络资源有限的环境,如何在不同的终端部署不同的查询方案和优化不同的查询算法及广域网环境下的数据传输,是海量数据应用的关键。查询方法和查询范围在很大程度上决定了空间数据库的应用程度和应用水平,定位空间对象和提取对象信息是地理信息系统进行高层次空间分析的基础。借助分布式GIS技术和数据查询优化技术可以有效地解决数据产生和应用之间的数据预览、分发环节。

1　GIS数据查询的特殊性

空间数据查询包含空间几何对象数据和属性数据的查询,查询结果集应同时满足这2个方面的要求。属性查询采用SQL查询即可;而几何数据的查询过程由于几何数据对象的不规则、多样性和大数据量的特点,在网络环境下的查询有其特殊性[1]:

(1)查询会话时间有限。在同一时刻,一个会话只能对应一个传输,即空间对象的一次查询请求对应一次完整的会话,基于因特网的会话时间是有限制的,不可以无限长;

(2)GIS会话是有状态的会话。即用户一次会话后与服务器的连接仍然存在,当GIS多次请求空间数据时,驻留在应用服务器上的代表用户的GIS会话组件只要从连接池中取得一个连接即可完成与空间数据库之间的会话请求,从而避免客户与服务器之间再次建立连接;

(3)事务过程难以把握。GIS数据查询实质上是一个会话、处理和执行事务的过程,目前由于空间数据库中空间实体之间的关系缺乏完整的描述,GIS空间数据模型与组织方面存在不少问题,事务处理时空间实体的封锁粒度甚难把握,基本上采用图层封锁的方法,但难以组织有效的原子事务和嵌套事务机制,直接影响了GIS复杂应用的操作难度[2];

(4)查询结果集表达的限制。由于网络带宽是有限的,而几何空间数据往往是海量的,往往导致查询结果集无法在用户可容忍的时间内完整显示。仅靠几何空间索引解决这种矛盾是远远不够的,需要寻找其它解决的办法。

2　空间数据库数据层空间矢量对象的优化

2.1　空间对象的查询处理

GIS系统一般采用表现层、服务层和数据层3层模式,表现层主要体现在客户端,服务层主要体现在GIS服务器,数据层主要体现在空间数据库[3]。目前在数据层,因为空间查询处理涉及复杂的数据类型,通常采用图1所示的2步算法来高效地处理

收稿日期: 2008-01-19; **修订日期**: 2008-03-19

这些空间对象。

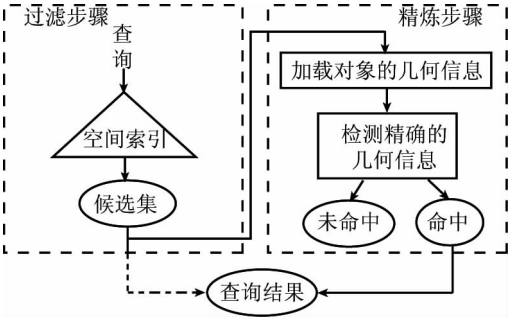

图 1　多步处理算法[4]

(1)过滤。主要利用最小外包矩形(Minimum Bounding Rectangle,MBR)对空间矢量对象作相同简单的近似。对于拓扑关系的求交计算,使用最小外包矩形来代替任意形状的不规则空间对象,计算代价小。近似过滤的结果是真实结果集的超集(候选集),且保证过滤中精确几何体最终结果中的元组包含于侯选集中,该阶段得到的结果包含满足原始查询条件的候选者。

(2)精炼。该阶段使用精确的几何条件对过滤结果进行处理,是一个计算代价高的过程。对经由过滤阶段得到的候选集中每个元素的精确几何信息和精确的空间谓词进行检查,这通常需要使用 CPU 密集型的算法。该步骤可以在空间数据库之外的应用程序中执行,故可以移植到客户端执行。

2.2　查询优化

查询优化器(Query Optimizer)是数据库软件中的一个模块,它产生不同的计算计划,从系统目录中获取信息,并结合一些启发式规则和动态规划技术以制定合适的策略[5]。查询优化器所承担的任务分为逻辑转换和动态规划。

(1)逻辑转换。查询优化器的模式如图 2 所示。由语法分析器来扫描高级声明性语句,并将语

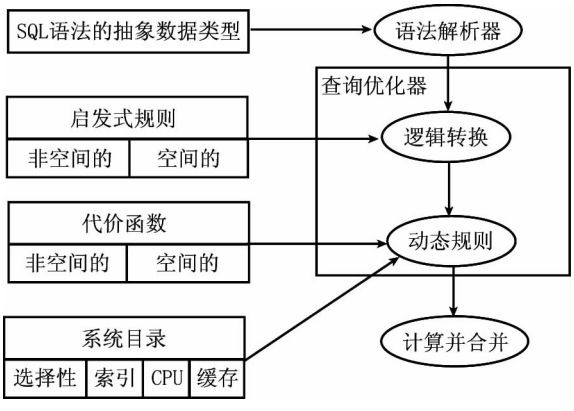

图 2　查询优化器的模式

句转换成一个查询树。在查询树中,叶结点对应着所涉及的关系,内部结点对应组成查询的基本操作,查询处理从叶结点开始自底向上处理,直到根结点上的操作完成。如果将语法分析器得到的查询树直接作为制定执行计划的基础,则其连接操作将会花费很大的代价,它的代价和所涉及关系大小的乘积有关[6],故应设法减少连接操作所涉及关系的大小。语法分析器利用从关系代数继承的一组形式化规则的结果,将得到的查询树转换为等价的查询树。逻辑转换是在在枚举出等价树之后,应用启发式规则过滤掉不是最终执行策略的候选者,它是删除那些不是最佳查询树的一条捷径。

(2)动态规划。是利用一种自底向上的优化策略,根据代价函数的推导从一组执行计划中确定最优执行策略的技术,能以较高的效率从可能的查询计划中得到代价较小的计划。在组建整个查询树的过程中,每个结点的不同处理策略构成计划空间。计划空间的基数通常很大,性能也可能相差几个数量级。在动态规则步骤中,关注的是查询树的每个结点,枚举处理该结点的所有可能的执行策略,对每个计划使用其代价函数来评估,选择代价最小的计划即是最优计划。

动态规划的核心在于利用代价函数评估每个执行策略,代价函数直接决定选择的计划。一个较好的代价函数必须考虑如下因素:①访问代价,即从二级存储搜索和传输数据的代价;②存储代价,即存储查询的执行策略所产生的临时关系的代价;③计算代价,即执行主存内运算的 CPU 代价;④通信代价,即在客户端与服务器之间传递信息的代价。

在传统的数据库中,I/O 代价在所有查询处理的代价中占主导地位。但对于空间数据库和分布式 GIS,通信代价在代价函数中占有的比重越来越大。

3　查询优化器植入客户端的查询方法

3.1　在客户端定义并使用代价函数优化动态规划查询

目前,查询优化技术主要分布在数据层,在代价函数选择上主要考虑计算代价、存储代价和 I/O 代价。在客户端的优化查询中,主要是对 SQL 语句执行的优化方法;在分布式系统中,通常采用逻辑部署不同操作与服务器端或客户端的技术,以及在不同的终端配置代理服务器的方式来提高查询操作速度,节省网络带宽资源。

本文结合分布式 GIS 的特点和目前数据层使用优化器的方法,提出了一个将数据层查询优化器的功能模块移植入客户端,定义并使用代价函数优化动态规划的矢量数据查询方法。

查询优化器针对空间数据库的第 $n(n \geq 1)$ 次动态规划的代价函数定义[7]为

$$DataLayerCost(n) = \mathrm{Exp}(records - examined) + K \cdot \mathrm{Exp}(pages - read) \quad (1)$$

式中,$DataLayerCost(n)$ 为在数据层第 n 次查询的代价; $\mathrm{Exp}(records - examined)$ 为预计读取的记录数,作为 CPU 时间的度量; $\mathrm{Exp}(page - read)$ 为预计从外存读取的页数,作为 I/O 时间的度量; 因子 K 度量 CPU 资源相对于 I/O 资源的重要程度。在数据层中,采用此代价函数,主要考虑访问代价和计算代价来选择查询树中代价最小的计划来执行查询。

但对于分布式系统,不仅要考虑到在不同物理终端实现不同查询优化方法,而且必须考虑到通信代价的比重。在不同的逻辑层部署不同的查询优化方法,在数据层主要考虑计算代价和访问代价。在客户层采用插件技术移植数据层的查询优化器到客户端时要考虑到通信代价,故定义客户层第 $n(n \geq 2)$ 次查询代价函数为

$$ClientCost(n) = DataLayerCost(n-1) - \sum_{i=1}^{n-2} DataLayerCost(i) \quad (2)$$

式中,$ClientCost(n)$ 为在客户层查询的第 n 次代价,这里 $n \geq 2$,使用数据层多次查询代价的差集来衡量; $DataLayerCost(n-1)$ 为数据层的第 $n-1$ 次查询代价; $\sum_{i=1}^{n-2} DataLayerCost(i)$ 为数据层查询的第 1 次到 $n-2$ 次查询代价总和,这里使用数据层的第 $n-1$ 次查询与前 $n-2$ 次查询代价的总和差来表示客户层的第 n 次查询代价。代价函数的值越小,通信数据越少,占用带宽越小。

实际应用中,首次在客户端查询时,首先利用数据层的过滤—精炼步骤处理空间矢量对象,并第一次生成查询的代价函数保存在客户端的缓存中,用户再次或者多次查询时,利用公式(2)生成客户端代价函数,在客户端优先使用动态规划函数,从客户层查询计划中选择最优的查询计划,从而减轻网络负载,提高查询速度。

客户端代价函数不仅在客户层的动态规划中用来有效选择一个较优的查询计划,同时也可作为数据层第 $n+1$ 次查询,修正数据层代价函数的一个参考值。把客户端代价作为通信代价来考虑,修正数据层代价函数,即

$$DataLayerCost(n+1) = \mathrm{Exp}(records - examined) + K_1 \cdot \mathrm{Exp}(pages - read) + K_2 \cdot ClientCost(n) \quad (3)$$

式中,$DataLayerCost(n+1)$ 是客户端使用动态代价函数后数据层的修正代价函数; 因子 K_1 和因子 K_2 表示相应代价的比重。在后续查询中,可以连续使用式(3)迭代修正代价函数。每次查询都在原来代价函数的基础上进行优化,在整个查询树的空间中,多次使用优化代价函数评估查询计划,在理论上得到近似最优的查询计划。

3.2 拓扑差集优化查询

在移植这种查询优化器技术过程中,需要在客户端根据 OPENGIS 规范重新实现数据库中底层矢量操作,这样客户机就能单独处理很多事务。能独自处理的事务由客户机自行完成; 客户机无法自行处理的事务,利用首次查询返回的结果集,暂时缓存于客户端,用户下次查询时,将需要查询集和暂存在本地的数据进行拓扑运算。拓扑运算后,无需从服务器读取数据则直接在客户端处理,反之,将计算查询数据差集,从服务器端直接取数据差集,从而减少网络传输量,减轻网络负载。

4 结论和展望

GIS 数据在当前网络环境查询中,因其海量信息和计算操作量无法估计的特点,空间信息的查询具有与众不同的特性。对于给定的一个查询,存在许多常见的执行查询策略,而采用查询优化器可以生成执行查询的计划并选择最优或近似最优的计划。在客户端移植数据层的查询优化器,使用新的代价函数来选择最优的计划,同时也实现数据拓扑差集的查询。该理论首次在“北京市电缆厂监控系统及生产管理系统”的项目应用,证明在井盖、隧道等空间数据查询和传输上,该方法能大大减轻网络的负载、有效提高查询速度。与此同时,如何利用智能代理技术,根据客户机和服务器的计算能力,部署不同操作于不同的终端的负载均衡的分布式 GIS 体系结构也是一个值得研究的问题。

参考文献:

[1] 陈建成,龚健雅,等. 基于 Internet 的矢量数据远程查询设计[J]. 测绘通报,2002,10:31 – 33.

[2] Gong Jian – ya. Design and Implementation of an Internet GIS [J]. Geo – spatial Information Science,2001,4(2) :1 – 7.

[3] 宋海朝,杨 钰. 分布式空间数据库的研究与设计[J]. 计算机工程与设计,2004,11:2046 – 2048.

[4] Brinkhoff T, Kriegel H P. Efficient Processing of Spatial Joins Using R – Trees[A]. Proc, ACM SIGMOD Conf. Management of Data[C]. 1993, 115 – 126.

[5] Beomseok Nam, PHenrique Andrade, PAlan Sussman. Multiple Range Query Optimization with Distributed Cache Indexing[A]. Proceedings of the 2006 ACM/IEEE Conference on Supercomputing[C]. Tampa Flaorda:ACM Press, 2006,11:100 - 110.

[6] 李立言,秦小麟. 空间数据库中连接运算的处理与优化[J]. 中国图象图形学报,2003,7:733 - 736.

[7] Shashi Shekhar, Sanjay Chawla. 空间数据库[M]. 谢昆青,马修军,杨冬青,等译. 北京:机械工业出版社,2004.

AN OPTIMIZATION METHOD FOR INQUIRY OF VECTOR DATA BASED ON DISTRIBUTED - GIS

SHANG Yan - Ling[1,2], XU Xu - Dong[2]

(1. *Department of Computer*, *Beijing University of Technology*, *Beijing* 100022, *China*; 2. *Anyang Normal University*, *Anyang* 455000, *China*)

Abstract: The contradiction between massive Distributed - GIS data and limited bandwidths of network resources is becoming more and more serious, and the inquiry of distant vector data and the transmission of massive data have become the key to the problem. Based on analyzing the method of massive data inquiry, combined with the query optimized method of spatial database and cost function, the authors propose a vector data inquiry method which is optimized and can be transplanted to the client and realize the data inquiry of topologic differences. Experiments have proved that this method is effective in improving the inquiry speed and maintaining a good network - load performance.

Key words: Distributed - GIS; Query optimization; Cost function

第一作者简介: 尚艳玲(1979 -),女, 助教,硕士研究生,主要研究方向为空间数据库和分布式地理信息系统。

(责任编辑: 李 瑜)

《西部开发重点区域遥感综合调查与监测报告》简介

2006 年 5 月 17 日,国务院西部开发办公室综合规划组对《"西部开发重点区域遥感综合调查与监测"项目建议的请示》的批复,对西部地区重点区域的国土资源和生态环境进行一次遥感综合调查研究,为国家编制《"十一五"西部开发重点区域发展规划》提供依据。

为了实施"西部开发重点区域遥感综合调查与监测"项目,中国遥感应用协会邀请西部 12 个省(区、市)和吉林省有关遥感单位代表,成立了项目领导小组、技术联络组和课题组,制订了充分利用已有遥感技术应用成果并集成有关资料的技术路线,按照《西部大开发"十一五"规划》规定的重点区域(包括成渝、关中—天水、环北部湾(广西)3 个重点经济区、11 个省会城市、6 个重点资源城市、13 个重点边境口岸城镇等 4 个部分)共设计 36 个课题,面积 40 多万 km^2,由于缺少云南省的资料,实际完成 30 个课题。

该报告由胡如忠、管海晏、王平担任主编,于 2008 年 3 月完成报告印刷并提交国务院西部开发办。

本项目是我国遥感界紧密结合国家西部大开发规划的需要,跨地区、跨部门、大规模组织起来的,在开发利用我国已经积累的海量遥感数据和丰富的遥感研究、开发、应用成果基础上,30 个课题组对课题调查范围内的各种国土资源和生态环境、自然灾害、开发区以及相关的优惠政策等进行了研究。国务院西部开发办综合规划组高度评价本成果"为编制西部开发重点区域规划提供了重要的参考资料","该课题依据遥感资料以及遥感应用技术分析提出的各区域和城市的发展趋势、资源开发、环境保护、灾害防治等规划建议,都对西部大开发工作具有直接的帮助"。并指出"该项目组织了 22 个单位和 80 多位各方面专家参与研究,课题成果凝聚了全体成员的大量的辛勤劳动。我们认为,该项目在利用遥感资料分析区域情况,应用遥感技术调查与监测区域发展方面,达到国内先进水平。"

(胡如忠、管海晏)