**地质时空大数据智能索引方法研究**

1. **立项依据**

现实世界是一个永恒变化的四维时空，当今中国社会面临的诸如环境污染、土地演化、海岸变迁、人口剧增、城市扩张、传染病转播、火灾传播、泥石流等自然灾害监控与防治等问题，无不与时空概念紧密相关，并且每时每刻都在持续不断地产生着大量的时空数据。时空数据可以帮助人类了解历史、掌握现在、预测将来，有助于提高人类对四维时空中各种存在与状态演变的洞察、感知与预测能力；对环境监测、矿山安全、城市内涝、土地管理、泥石流和地震等自然灾害的监测、预报、防治与救援等时空敏感性问题的求解具有重要意义。

2014年4月16～18日在北京召开的香山科学会议第491次学术讨论会——“中国‘玻璃地球’建设的核心技术及发展战略”（吴冲龙教授申办）， 李德仁、王家耀、李廷栋、赵鹏大、康玉柱等院士等40余名专家学者一致认为“玻璃地球”是地质时空大数据的有效载体，地矿工作信息化进入了大数据时代。然而，地质时空大数据时代如何有效地存储和管理这些海量时空数据，依然是以分布式协同、高性能计算、时空数据流处理为核心的多维地质信息系统必须解决的关键技术问题之一。

多维时空数据库目前仅处于研究阶段，海量时空数据管理中还存在着诸如多层次时空缓存技术、多维时空数据智能检索与调度、分布式环境下时空数据负载均衡等技术瓶颈问题，研究可支持时空大数据存储、分布式冗余备份、规模动态调整的智能时空索引技术是这些问题有效解决的基础。随着计算机硬件设备的发展，多核计算机已经成为常规计算设备，分布式多核计算环境在软硬件方面已经获得长足发展，并日趋成熟。在分布式多核计算环境中，如何构建合理有效的分布式并行时空索引体系结构，实现地质时空大数据规模动态调整、负载动均衡，降低频繁更新的时空索引中并发控制成本，充分发挥多核计算机及分布式计算优势，实现分布式多核计算环境下的细粒度的多维并行智能时空索引方法，也成为地质时空大数据领域亟需解决的关键科学技术问题之一，这也是本项目研究意义之所在。

1. **研究内容（研究目标、内容及技术路线）**

针对日趋发展的分布式多核并行计算环境中地质时空大数据高效访问需求，研究适合分布式并行计算环境下的智能时空索引的理论架构和基于间隔关系算子的多维并行时空索引方法，突破多维树形索引的层次结构对并行算法实现的限制，细化多维时空索引并行粒度的同时降低并发控制开销，提升分布式多核计算环境下多维并行时空索引性能，为分布式多核计算环境下的海量三维、四维或更高维的地质时空大数据的智能检索提供可行、通用、高效的分布式并行时空索引解决方案。

**1、主要研究内容**

项目将研究分布式并行智能时空索引的理论架构体系，研究各个网络节点之间的顾及数据均衡的时空划分方法和动态调整策略，然后再研究网络节点内部多核环境下的基于间隔关系算子的多维并行时空索引方法及其相关理论与技术问题。基于间隔关系算子的多维分布式并行时空索引方法，拟采用顾及数据均衡的网络节点时空划分，构建合理有效的分布式索引体系；在每个网络结点上将多维时空数据转化为多个独立的间隔数据集合，采用基于三角区域递归分解的虚拟二叉树实现各种间隔关系算子操作，将复杂多维时空查询问题转化成一系列简单的可并行的间隔关系算子的多核并行计算问题，实现单个多维时空数据索引内部的并行计算，充分发挥多核计算机性能优势，提高多维时空索引性能。其主要研究内容如下：

（1）分布式并行计算环境下的智能时空索引理论架构

针对不同的分布式环境的不同类型，研究同构和异构的分布式时空索引结构；研究各个节点之间顾及数据均衡的时空划分方法和动态调整策略。针对单节点计算机上多核并行计算环境，研究适合多核并行的细粒度的并行时空索引结构；在此两者基础上，研究提出分布式并行计算环境下的多层次的智能时空索引结构。

（2）基于间隔数据集的多维时空对象表达

针对多维时空对象表达和时空过程模拟过程中并行计算需求，研究间隔对象以及间隔对象之间的关系表达与计算问题；研究时间数据的间隔表达方式及其算子操作方法；研究空间数据与间隔数据集的相互转化方法；实现索引结构中基于间隔数据集的多维时空对象统一表达。

（3）间隔关系算子与间隔数据索引方法

针对间隔数据特点（开始值永远小于或等于结束值），研究间隔数据集的基于三角区域递归分解的二分策略和二叉树构建方法；研究消除二叉树内部节点（非叶节点）的虚拟二叉树的编码方法，减少节点访问次数；研究基于虚拟二叉树的各种间隔关系算子算法；研究可并行间隔关系算子在多核环境下的部署与并行计算问题。

（4）智能时空索引的查询、插入与删除算法

研究如何根据时空数据与间隔数据的相互转化方法，将复杂多维时空查询问题转化成一系列简单的可并行的间隔关系算子的多核并行计算问题；研究基于三角区域递归分解的二叉树索引的插入、查询与删除算法，以及多棵不同维度上的二叉树之间的并发控制问题；研究如何有效地整合每个维度的虚拟二叉树插入、查询与删除算法，实现高效的并行时空索引算法；研究并行时空索引中时空维度与计算机处理器或核的数目之间的关系及其对时空索引整体性能的影响。

**2、技术路线**

项目中分布式并行时空索引的基本实现思路是：首先采用顾及数据均衡的时空域划分方法将全局时空数据划分为多个时空域，将每个时空域动态映射到不同的服务器节点。节点之间通过自定义时空数据交换协议通信，构建基于主从结构或对等结构的分布式结构；节点内部的本地索引采用基于间隔关系算子的多维并行时空索引方法。

基于上述总体思路，本项目拟采用的总体技术路线是：充分利用地球空间信息科学领域和数据库领域的前沿技术成果，以国家自然科学基金青年基金项目“地上下一体化三维动态广义表空间索引方法”（41101368）、国家自然科学基金面上项目“基于间隔关系算子的多维分布式并行时空索引方法”（41572314），国家863计划重点项目课题“三维空间数据管理系统与分析组件研发”（2008AA121602）和“时空过程模拟与实时GIS系统”（2012AA121401）项目已有研究成果为基础，针对分布式多核并行计算环境下地质时空大数据高效访问需求，研究基于间隔关系算子的多维分布式并行智能时空索引方法及其相关理论与技术问题，细化多维时空索引并行粒度，提升分布式多核计算环境下多维并行时空索引性能；并结合两个大型城市时空数据集对所提出的理论、方法进行技术实现和验证。

1. **特色与创新点**

项目将时间数据索引泛化为间隔关系算子，并创造性地应用在四维时空大数据中，其的特色之处在于建立时空查询与间隔关系算子之间的关联关系，通过间隔关系算子并行计算来实现地质时空大数据的智能索引。项目的主要创新有以下两点：

1. 多维时空查询与并行的间隔关系算子集的转换理论与方法

针对多核并行计算环境下多维树形时空索引对细粒度并行计算的限制，建立时空对象关系与间隔关系算子之间的转换公式，通过这些转换公式将多维时空数据查询转化为可并行的间隔关系算子操作，然后将间隔数据集映射到可并行的不同维度的间隔点集平面，采用统一的二维平面元素求交运算实现各种间隔关系算子的高效计算，从而在分布式多核计算环境下以统一接口，智能地实现多种时空查询，降低并发控制成本。

1. 基于递归二分策略的间隔关系算子的高效计算方法

针对间隔对象的开始值小于等于结束值的特点，提出递归二分策略，构建虚拟二叉树，研究提出高效的间隔关系算子计算方法。该方法将间隔对象投影到间隔点集平面，形成等腰直角的上三角区域（简记为基础三角形区域），对该三角区域递归二分，以分解得到的三角形区域构建二叉树，并对递归二分得到的每个三角区域（二叉树节点）采用二进制编码，每个二进制编码可以转换成一个整数，这样只要知道基础三角形区域和一个代表节点编码的整数，就可以计算出该节点所覆盖的三角型区域范围。这样的特性使得二叉树可以只记录叶子结点，减少节点访问次数，提高二叉树索引查询性能。同时由于投影平面只需要考虑上三角区域而不用考虑整个平面范围，这也大大缩减了平面元素求交计算量。该方法在缩减计算量的同时有效降低了并行时空索引的并发控制成本。

1. **研究计划与预期成果**

**1、年度研究计划**

2015.11-2016.11：

在前期工作基础上，研究分布式并行时空索引的网络邻近度理论和多层次体系结构，构建索引框架结构；研究多维时空数据与可并行间隔数据集之间的转换理论和方法，构建多维时空数据查询与可并行间隔关系算子之间的转换公式。研究间隔数据集到间隔点集平面的映射以及间隔点集平面三角化区域递归分解方法；根据该分解方法，研究虚拟二叉树编码方法，消除二叉树内部节点（非叶节点），减少节点访问次数。参加国际学术会议1人次。发表相关论文1篇（T2）。

2016.12-2017.12：

研究基于虚拟二叉树的并行时空索引的查询、插入与删除算法；研究虚拟二叉树在单核和多核上的计算部署问题；采用C++语言和OpenMP设计并实现该索引算法；研究并行时空索引中时空维度与计算机处理器、核的数目之间的关系及其对并行时空索引整体性能的影响。在Linux平台上构建测试环境，分别针对三组不同实验数据进行对比实验与分析。派1名项目组成员到荷兰屯特大学地球观测与地学信息科学学院进行短期交流访问。发表相关论文1篇（T2）。

**2、预期研究成果**

（1）拟发表T2论文2篇

An efficient parallel algorithm for interval relationship operators

基于递归二分策略的间隔关系算子的高效计算方法，拟定发表在ISPRS Journal of Photogrammetry and Remote Sensing期刊。

Distributed indexing technique for spatio-temporal big data

针对地质时空大数据的基于间隔关系算子的分布式智能索引方法，拟发表在Computers & Geosciences期刊。

（2）培养2 名硕士研究生。

1. **申请人及团队介绍**

申请人：何珍文，1978年生，湖北潜江人，副教授，硕士研究生导师，博士，毕业于华中科技大学信息与通信工程专业，荷兰屯特大学对地观测与地学信息学院（ITC）时空数据处理方向博士后，国际数学地质协会终身会员；主要从事空间信息科学与技术、地学信息工程、计算机应用方面教学与科研工作；参与主持了国家自然科学基金项目“基于间隔关系算子的多维分布式并行时空索引方法”（41572314，项目负责人）和“地上下一体化三维动态广义表空间索引方法”(41101368，项目负责人)、国家“十一五”863重点项目课题“三维空间数据管理系统与分析组件研发”(2008AA121602，课题组副组长)、国家大型油气田及煤层气开发科技重大专项项目（2008ZX05051，子课题负责人）等多项国家级、省部级攻关课题；作为主要技术负责人参与主持了国内第一套针对地质矿产勘查开发信息化的三维可视化软件平台(QuantyView)，以及多个相关重要应用系统的研究与开发工作；获国家授权发明专利7项；作为副主编和主要成员出版行业重点教材3部；公开发表论文20余篇；获省部级技术发明一等奖1次、科技进步一等奖2次、二等奖2次，三等奖1次。

本项目依托中国地质大学（武汉）计算机学院吴冲龙教授和刘刚教授领导的地质信息科学与技术团队，项目主要参与人员如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **姓名** | **职称** | **学位** | **单位** |
| 田宜平 | 教授 | 博士 | 中国地质大学(武汉) |
| 张夏林 | 教授 | 博士 | 中国地质大学(武汉) |
| 孔春芳 | 副教授 | 博士 | 中国地质大学(武汉) |
| 毛小平 | 副教授 | 博士 | 中国地质大学(北京) |
| 翁正平 | 讲师 | 博士 | 中国地质大学(武汉) |
| 张志庭 | 讲师 | 博士 | 中国地质大学(武汉) |
| 李新川 | 讲师 | 博士 | 中国地质大学(武汉) |
| 李章林 | 讲师 | 博士 | 中国地质大学(武汉) |
| 彭诗杰 | 博士生 | 硕士 | 中国地质大学(武汉) |
| 张涛 | 硕士生 | 学士 | 中国地质大学(武汉) |
| 花松昌 | 硕士生 | 学士 | 中国地质大学(武汉) |
| 吕浙明 | 硕士生 | 学士 | 中国地质大学(武汉) |

1. **项目预算**

|  |  |  |
| --- | --- | --- |
| 支 出 科 目 | 金 额  （万元） | 计 算 根 据 及 理 由 |
| 合 计 | 6.0 |  |
| 1.科研业务费 |  |  |
| 设备费 | 2.0 | 购买小型数据库服务器，构建分布式并行索引测试环境 |
| 材料费 | 1.0 | 打印纸、硒鼓及电脑配件等 |
| 差旅费 | 1.5 | 科学考察、业务调研、学术交流等差旅费用 |
| 出版/文献/信息传播 | 1.5 | 版面费、查新检索费用等 |