



分布对等环境中的时空查询技术

FENBU DUIDENG HUANJING ZHONG DE SHIKONG CHAXUN JISHU

张翀 葛斌 肖卫东 郭澄 陈晓莹 · 著

国防科技大学出版社
National University of Defense Technology Press

分布对等环境中的时空查询技术

张 翀 葛 斌 肖卫东 郭 澄 陈晓莹 著

国防科技大学出版社

· 长沙 ·

图书在版编目 (CIP) 数据

分布对等环境中的时空查询技术/张翀等著. —长沙: 国防科技大学出版社, 2018. 11

ISBN 978 - 7 - 5673 - 0523 - 6

I. ①分… II. ①张… III. ①数据处理—研究 IV. ①TP274

中国版本图书馆 CIP 数据核字 (2018) 第 117756 号

国防科技大学出版社出版发行

电话: (0731) 84572640 邮政编码: 410073

责任编辑: 邹思思 责任校对: 周蓉

新华书店总店北京发行所经销

国防科技大学印刷厂印装

*

开本: 710 × 1000 1/16 印张: 13.5 字数: 250 千

2018 年 11 月第 1 版第 1 次印刷 印数: 1 - 400 册

ISBN 978 - 7 - 5673 - 0523 - 6

定价: 36.00 元

前 言

时空数据可以帮助人们掌握历史、现在甚至预测将来,对提高空间各种存在和状态演变的感知度、洞察力和预见性有重要价值。随着探测采集手段的不断进步以及网络的不断发达,时空数据不再是少量的、局部的存在,而是规模庞大,分散广泛,应用普遍。面对这种现状以及今后更加向综合化、普适化的发展趋势,集中式的时空索引必然不能满足各类用户的各种查询,其性能低下、单点失效的缺点使得时空索引势必向分布式发展。然而,以主从模式或层次模式为代表的分布式计算方式也无法满足规模海量、动态复杂的分布式时空对象环境,因为这样的分布式模式存在关键节点,是整个系统的性能瓶颈和脆弱环节,不利于系统扩展和灵活重组。

本书提出采用对等计算模式组织分布式时空索引的建立,主要从节点同构和异构两个角度出发,并结合用户的时空查询类型,从面向历史查询、面向将来预测查询和面向描述时空信息数据的查询出发,研究设计查询性能良好、更新能力稳定、支持半结构化描述查询的分布式时空索引机制,主要取得了如下成果:

(1) 深入全面地总结了相关工作。目前分布式索引研究较少,在总结仅有的几篇相关文献后,从集中式时空索引和基于对等计算的多维索引两个研究方向出发,梳理与本书相关的方法理论。以可

视化的方式展现相关工作和作者之间的关系是本书的创新之处,采用这种方式可以直观地洞察各类工作之间的脉络、研究思路和研究风格,为本书科学地研究基于对等计算的时空索引奠定了良好的理论基础。

(2) 描述了基于对等计算的分布式时空索引的问题模型,并提出索引的整体架构。由于目前开展的相关工作较少,需要从根本上界定问题的本质,因而提出基于 P2P 的分布式时空对象模型,采用空间划分的办法对分散的时空对象和分布服务器节点之间的关系建立模型,并对面向历史和面向将来查询的时空数据以及异构情况做了描述与统一。提出了一种双层的整体架构,双层分为本地和全局,每个节点首先完成本地的信息组织及索引建立,然后再参与全局的时空信息组织和索引构建,其整体框架描述了索引查询、建立和维护的基本协议,整体架构的合理性、完整性和可扩展性为后续工作奠定了研究基础。

(3) 提出了基于“双环”结构的面向历史查询的对等时空索引。针对历史时空数据规模庞大、查询复杂的特点,提出了“双环”结构,它由时间环和空间环组成,特点是能够独立地支持空间维和时间维查询,并通过建立直方图对“双环”进行选择性的估计,从而择优查询,这能够较大地提高查询效率。基于此原理,进一步提出了面向历史查询的对等时空索引 DRHSTI,设计了范围查询、kNN 查询和轨迹查询。“双环”在提高查询效率的同时增加了索引维护开销,因此,本书分别采取本地时空索引结点发布、动态分片索引时间环以及缓存技术降低对“双环”的维护代价,实验表明,基于“双环”建立的对等分布式时空索引能够在保持较低维护代价

的同时快速有效地支持面向历史的时空查询。

(4) 提出了基于邻域连接的面向将来查询的对等时空索引。面向将来的预测查询具有时空数据动态更新性强、索引压力大的特点,更新效率问题一直备受关注,利用物理空间邻接节点进行局部时空状态摘要交换,进而再利用距离较远的逻辑邻接节点可以掌握全局时空状态,从而打破 Hilbert 曲线降维带来的信息损失的限制,充分提高查询和更新的效率。基于此原则,进一步提出了索引 PeerFOX,利用邻域连接的概念设计了路由表、点路由规则和区域路由规则,并给出了面向将来预测的范围查询、kNN 查询和连续查询算法以及在动态环境下索引的维护算法。实验表明,PeerFOX 能够很好地处理更新问题,并且其查询效率较好。

(5) 提出了在多源异构环境中基于 XML 聚类面向半结构化查询的时空描述信息组织方法与索引建立机制。在异构环境下,各节点中时空数据格式、规范、度量标准不同,查询时需要基于描述信息进行,由于数据本身不统一,描述信息亦不相同,这给对等环境下的时空查询带来困难。因此提出了将描述信息 XML 进行聚类,从而将含有相似 XML 的节点聚集在一起并建立基于域的索引,这样描述信息被合理组织在同一个域内,用户的半结构化描述查询可以在域内进行搜索,大大降低了转发的代价,提高了查询效率。在设计 XML 聚类时,提出了基于簇核心的 XML 聚类方法,利用簇核心可以提高聚类精度、加快聚类速度并支持增量式聚类,从而提高整个系统运行效率。实验表明,基于簇核心进行 XML 聚类具有质量好、效率高的特点,基于 XML 聚类结果建立对等索引可以提高半结构化时空索引的效率。

(6) 应用研究。应用相应关键技术的研究成果,从同构和异构两方面出发,结合战场移动对象和航天时空资源信息应用,设计并实现了两个原型概念系统,验证了本书研究方法和技术的有效性和可行性。

张翀

2018 年 6 月

目 录

第 1 章 概 述	(1)
1.1 背景介绍	(1)
1.1.1 分布式时空索引	(2)
1.1.2 对等计算结构的时空索引	(3)
1.2 分布对等时空索引应用价值	(4)
1.3 主要内容	(6)
第 2 章 研究现状	(10)
2.1 技术脉络与发展趋势	(12)
2.1.1 集中式时空索引	(12)
2.1.2 基于 P2P 的空间索引技术	(12)
2.2 研究派别与研究风格	(14)
2.2.1 集中式时空索引	(15)
2.2.2 基于对等计算的空间索引技术	(17)
2.3 分门别类综述	(19)
2.3.1 集中式时空索引	(19)
2.3.2 基于 P2P 的空间 (多维) 索引技术	(25)
第 3 章 基于对等计算的时空索引理论架构	(37)
3.1 时空索引的相关因素分析	(37)
3.1.1 时空对象的特点分析	(38)

3.1.2	时空对象抽象表示·····	(39)
3.1.3	时空对象的查询方式·····	(40)
3.2	基于对等计算的时空索引问题建模·····	(42)
3.2.1	两种划分模式·····	(42)
3.2.2	基于 P2P 的时空对象模型 ·····	(46)
3.3	一种双层的基于对等计算的分布式时空索引整体架构·····	(52)
3.3.1	整体架构描述·····	(52)
3.3.2	整体架构的分析·····	(55)
3.4	本章小结·····	(56)
第 4 章	基于“双环”结构面向历史查询的对等时空索引 ···	(57)
4.1	引论·····	(57)
4.1.1	Chord 结构 ·····	(57)
4.1.2	Hilbert 曲线·····	(58)
4.2	基于“双环”的分布式历史时空索引结构描述 ·····	(59)
4.2.1	“双环”结构 ·····	(59)
4.2.2	构建 peer 数量分布直方图 ·····	(64)
4.3	时空查询算法·····	(64)
4.3.1	范围查询·····	(65)
4.3.2	kNN 查询 ·····	(67)
4.3.3	轨迹查询·····	(67)
4.4	索引的建立与维护·····	(68)
4.4.1	索引的建立·····	(68)
4.4.2	索引的维护·····	(69)
4.5	性能评价·····	(71)
4.5.1	数据集描述·····	(71)
4.5.2	实验结果与分析·····	(72)
4.6	本章小结·····	(80)

第 5 章 基于邻域连接面向将来预测查询的对等时空索引 ...	(81)
5.1 引论	(82)
5.2 PeerFOX 索引结构	(83)
5.2.1 拓扑结构	(83)
5.2.2 PeerFOX 节点间时空信息收集机制	(90)
5.3 路由规则	(95)
5.3.1 点路由规则	(95)
5.3.2 区域路由规则	(96)
5.4 查询算法	(99)
5.4.1 范围查询	(99)
5.4.2 kNN 查询	(100)
5.4.3 连续范围查询	(101)
5.4.4 连续 kNN 查询	(101)
5.5 索引建立与维护算法	(102)
5.5.1 索引的建立	(102)
5.5.2 索引维护过程	(103)
5.6 负载均衡	(104)
5.6.1 节点加入系统时的负载均衡策略	(104)
5.6.2 运行时的负载均衡策略	(105)
5.7 性能测试与分析	(105)
5.7.1 实验环境	(105)
5.7.2 实验结果及分析	(109)
5.8 本章小结	(120)
第 6 章 基于 XML 聚类面向半结构化查询的对等多源异构时空 信息组织与索引建立	(121)
6.1 引论	(122)
6.1.1 背景	(122)

6.1.2	问题举例	(123)
6.1.3	问题提出	(125)
6.2	相关工作	(128)
6.2.1	p, q -gram 距离	(128)
6.2.2	CXK ^[115] 分布式聚类算法	(129)
6.2.3	相关工作的不足	(130)
6.3	簇核心的概念	(131)
6.4	本地静态聚类	(134)
6.4.1	特征提取	(135)
6.4.2	构造特征关联度矩阵	(135)
6.4.3	生成簇核心	(135)
6.4.4	利用簇核心指导分类	(136)
6.5	分布式聚类及索引构建	(136)
6.5.1	分布式聚类过程	(136)
6.5.2	索引构建	(138)
6.6	本地簇核心动态维护及增量式聚类	(140)
6.6.1	生成特征更新集合	(141)
6.6.2	簇核心动态维护	(143)
6.7	全局簇核心动态维护及索引的维护	(146)
6.7.1	全局簇核心动态维护	(147)
6.7.2	索引的维护	(148)
6.8	查询路由机制	(149)
6.9	实验与结果分析	(151)
6.9.1	数据描述与参数设置	(151)
6.9.2	本地静态聚类效率	(152)
6.9.3	本地静态聚类的质量	(153)
6.9.4	本地增量聚类实验	(156)
6.9.5	分布式聚类与簇核心维护实验	(157)
6.9.6	索引查询与维护性能	(159)

6.10 本章小结	(161)
第7章 基于对等计算的分布式时空信息查询与管理系统设计 及应用研究	(162)
7.1 面向数字化战场移动对象的时空查询处理系统.....	(162)
7.1.1 设计思路.....	(164)
7.1.2 系统体系架构.....	(165)
7.1.3 系统运行流程.....	(167)
7.1.4 系统展示.....	(171)
7.2 面向时空查询基于对等计算的多源航天信息资源共享管理系统	(177)
7.2.1 设计思路.....	(177)
7.2.2 系统体系架构.....	(178)
7.2.3 系统运行流程.....	(180)
7.2.4 系统展示.....	(183)
7.3 本章小结.....	(189)
第8章 总结与展望	(190)
8.1 主要工作.....	(190)
8.2 未来研究工作.....	(191)
参考文献	(193)

第1章 概述

索引是数据库系统实现的重要关键技术之一。时空索引是时空数据库这一研究领域的热点问题。随着时空数据积累规模的扩大,分散程度的增强,用户即时进行时空检索的需求不断提高,研究分布式时空索引技术将从性能和共享范围上满足用户对时空对象搜索的需求,实现快速高效的查询能力,提高召回率,并在分布式GIS、智能交通、物联网等应用中发挥重要作用,特别在战场环境下,分布式时空索引对提高战场透明度、提高联合作战能力、实施高效的指挥决策具有重要意义,它将成为数字化战场建设、联合作战应用中极富挑战性的研究课题。

目前集中式时空索引技术已被广泛而深入地讨论与研究,形成了若干研究思路与发展方向。然而,分布式环境下的时空索引技术目前研究却非常少,正如 Ilari S 等在《ACM 计算概览》中所强调^[1]:“目前多数工作仅考虑了集中式的结构,未来的工作将会聚焦于分布式环境。”本书正是致力于研究分布式环境下如何索引时空数据这一目前较少探讨的问题。

本书提出了基于对等计算^[2](Peer-to-Peer Computing, P2P)技术建立分布式时空索引的方法,对等计算技术以“地位平等”思想管理计算节点之间的关系,充分满足了分布式环境下动态、自治、无中心化地管理大规模时空数据的需求。本章综述了研究背景和相关工作,并介绍了本书的主要工作与全文结构。

1.1 背景介绍

时空数据是一类特殊的数据类型,它所描述的空间状态随着时间推进而不断演变,利用时空数据可以帮助人们掌握历史、现在甚至预测将来,提高对空

间各种存在和状态演变的感知度、洞察力和预见性。由于时空数据的特殊性,如何对其进行索引成为有效管理时空数据的关键技术之一。

索引是提高查询处理能力的重要手段,其核心思想是从大量的数据中提取出反映概要信息的结构,根据一定排序规则分层、分级或分块存储,查询时由于使用了事先合理组织好的概要信息,可以将不符合查询条件的数据依排序规则提前排除,从而节省大量的查找时间,提高查询效率。由于时空数据的查询方式较复杂,如面向历史的 kNN 查询、轨迹查询以及面向将来预测的连续式查询,从而使时空索引的作用显得更为重要。

1.1.1 分布式时空索引

以往对时空索引的研究都限定在集中式环境下,即时空数据集中式存储,时空索引一般在单个节点上建立,然而随着各种探测手段的便捷化以及计算机网络的高速发展,以集中式手段对时空数据进行管理存在较多弊端,因而已不能满足在新形势下时空查询的需求。最典型的例子如未来我军维护国家统一的作战行动是信息化条件下的一体化联合作战,各军兵种、各级指挥部门都将参与军事行动的部署、决策与实施,战场中存在与战斗各种因素相关的时空对象,如战斗人员、武器装备、后勤设施以及气象条件等,如果仍然以集中式索引组织时空数据,将会带来如下问题:第一,联合作战的空间存在规模庞大的时空信息,随着战争的不断推进,海量的时空数据将一直不停地进入指挥控制系统,这将使得集中式的时空索引性能骤然下降,导致查询低效,严重影响指挥决策;第二,在联合作战条件下,诸军兵种共同参战,集中式的组织方式使得各单位、各级指挥部门必须链接到同一节点进行查询和更新,在庞大的用户规模下,集中式时空索引势必会产生性能瓶颈,指数级的消息量将会造成网络瘫痪;第三,战场环境动态多变,处处存在危险,集中式组织时空数据将带来单点失效这种毁灭性的打击,造成丧失掌握空间信息的能力,致使指挥处于瘫痪状态。

因此,研究分布式的时空索引对提高系统性能,消除单点失效,增强系统可靠性具有十分重要的意义。

1.1.2 对等计算结构的时空索引

分布式时空索引能够消除集中式时空索引在新形式下的若干不足，但采用何种分布式结构组织节点与节点之间、节点与时空数据之间的关系值得讨论。

以主从式 (Master-slave) 和层次式 (如分布式目录技术 LDAP^[3]) 为代表的分布式计算技术将无法满足大规模动态环境下灵活协作的时空索引构建与查询。原因在于：第一，性能低效，主从式或层次式中都存在关键性的中心节点，当数据量剧增、用户规模庞大时，中心节点将不堪重负，导致其他节点受到影响，系统总体性能下降；第二，扩展性差，中心节点的存在使得网络的规模受到限制，当大规模部署时，中心节点的负载将会是巨大的，这将是系统的瓶颈；第三，灵活性差，主从式或层次式部署方式笨拙，从建立到节点易动需要遵守大量严格的协议，当组织结构发生变化时，系统要重新组建，这在复杂动态的环境特别是战场环境中，将成为阻碍系统高效运行的重大障碍；第四，抗毁性弱，中心节点往往会成为整个系统的脆弱点，破坏者一旦发现并攻击，整个系统将会瘫痪。

本书提出采用对等计算 (P2P) 的模式设计分布式时空索引结构，这是因为对等结构能够满足分布式环境下高性能、大规模、灵活和抗毁的需求，P2P 计算将参与计算的节点在功能上视为一致，各个节点既是服务器也是客户端。基于 P2P 技术建立分布式时空索引，也就是采用“地位平等”的思想组织各个节点针对时空查询建立拓扑连接，形成具有一定结构的覆盖网 (Overlay Network)，在整个索引中各节点对等地贡献自己的时空查询能力，通过消息转发将查询投递到含有符合查询条件的数据的节点上。对等计算技术的独特优点将使分布式时空索引在纷繁复杂、动态多变的环境下体现重要价值。

基于对等计算的分布式时空索引是跨领域的，图 1.1 描述了基于对等计算的分布式时空索引与其他相关领域的关系，它是一项交叉领域，结合了数据库和分布式计算领域，是在集中式时空数据库索引和对等计算的不断发展下产生的。

表 1.1 显示了集中式时空索引与基于 P2P 的分布式时空索引的区别，其中，“节点”指参与计算的计算机，而“结点”指集中式索引中的数据结构，如 R-树的节点。

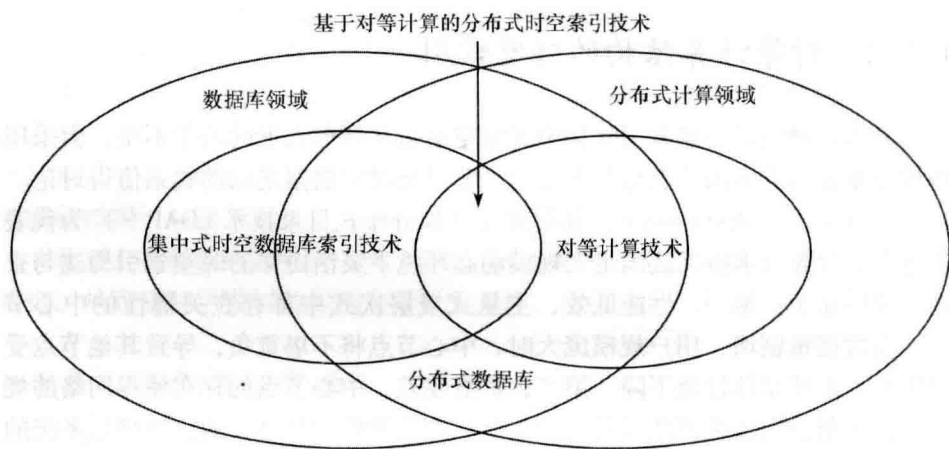


图 1.1 基于对等计算的分布式时空索引与其他相关领域的关系

表 1.1 集中式时空索引与基于 P2P 的分布式时空索引的比较

	集中式时空索引	基于 P2P 的分布式时空索引
节点（结点）	节点为计算机中的某一块内存或磁盘	节点为一台自治的计算机
剪枝目的	降低 I/O 次数	降低通信开销
剪枝途径	划分时间空间，设计合理索引结构和集中式搜索算法	划分时间空间，设计合理的拓扑结构和分布式搜索算法

1.2 分布对等时空索引应用价值

基于对等计算的分布式时空索引技术对提高时空数据的访问性能和拓展时空数据的应用领域具有较大的实践意义。它在数字化战场态势监控、云计算（Cloud Computing）、物联网（Internet of Things, IOT）以及传感器全球网（Sensor Web）中具有较大的应用价值。

(1) 数字化战场^[4]。1998 年美军提出以网络为中心的战场，战场中一切参战的对象都配备感知设备，使其对战场上任何行动都可进行监控。数字化战场使分布式时空数据的管理任务凸显重要性，作战单元或指挥单元都负责存储

本级或本单位的时空数据,这对跨部门、大范围的时空查询提出了挑战。基于对等计算的分布式时空索引技术从适应战场多变、动态的角度充分支持对全域内各单位时空数据的高效查询,各单位对等地共同参与构建全局分布式索引,每个单位不但可以自治地建立本级索引,还可以参与构建上级单位的索引,从而为全域内的查询提供服务。

(2) 云计算^[5]。狭义的云计算是指 IT 基础设施的交付和使用模式,指通过网络以按需、易扩展的方式获得所需资源;广义的云计算是指服务的交付和使用模式,指通过网络以按需、易扩展的方式获得所需服务。核心思想是将大量用网络连接的计算资源统一管理和调度,构成一个计算资源池向用户提供按需服务。利用云计算提供位置服务将充分提高位置服务的计算性能。采用 P2P 技术组织云服务器拓扑结构实现虚拟化, P2P 计算将各个节点既自治又联合地提供服务,恰当地满足了云计算虚拟化需求,在此基础上建立时空索引,可以充分确保用户对位置的透明式查询,即无地域差异和无服务节点差异的时空查询。

(3) 物联网^[6]。物联网是通过射频识别 (RFID)、红外感应器、全球定位系统、激光扫描器等信息传感设备,按约定的协议,把任何物品与互联网相连接,进行信息交换和通信,以实现对物品的智能化识别、定位、跟踪、监控和管理的一种网络。以对等模式组织物联网服务器,从广域上扩大了物联网的触及范围,从时间和空间上扩展管辖更多的感知事物,在此基础上建立起的分布式时空索引,使得各个运营商既可以自主地对本公司的物品进行查询,还可以从整个市场的角度按照一定权限查询物品的时空信息。这样可以做到对全局内的物品充分了解,增加战略敏锐感。

(4) 传感器全球网^[7]。无线传感器网络 (Wireless Sensor Networks, WSN) 是由大量的静止或移动的传感器以自组织和多跳的方式构成的无线网络,其目的是协作地感知、采集、处理和传输网络覆盖地理区域内感知对象的监测信息,并报告给用户。在多个地域部署 WSN,再将每个 WSN 的基站互联起来,这样就形成了传感器全球网 (Sensor Web)。基于 P2P 的分布式时空索引技术将增强传感器全球网构建的灵活性,新的 WSN 及基站的建立在不影响已经存在的网络的前提下,方便地加入传感器全球网,这充分满足了传感器全球网的动态性,提高了使用的简易性。