



中国计算机学会学术著作丛书

大规模图数据的 分布式处理

谷峪 于戈 鲍玉斌 著

清华大学出版社





中国计算机学会学术著作丛书

大规模图数据的 分布式处理

谷峪 于戈 鲍玉斌 著

清华大学出版社
北京

内 容 简 介

随着以社交网络为代表的图数据规模高速增长,复杂的查询需求不断涌现,处理这类大规模数据有许多理论问题需要解决。本书结合作者多年的研究积累,系统地介绍了大图分布式处理中基础的数据划分、组织和消息管理技术,以及三角形、最大 k 边连通子图、最小生成树、频繁子图、重叠社区发现等大图查询和分析算法的优化,并对系统实现技术进行了探讨。

本书适合高等院校计算机专业的教师、学生及计算机应用系统的研发人员阅读参考。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

大规模图数据的分布式处理/谷峪,于戈,鲍玉斌著.--北京:清华大学出版社,2015

(中国计算机学会学术著作丛书)

ISBN 978-7-302-42072-9

I. ①大… II. ①谷… ②于… ③鲍… III. ①分布式数据处理—研究 IV. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 263558 号

责任编辑:薛 慧

封面设计:傅瑞学

责任校对:刘玉霞

责任印制:沈 露

出版发行:清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址:北京清华大学学研大厦 A 座 邮 编:100084

社 总 机:010-62770175 邮 购:010-62786544

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者:北京密云胶印厂

经 销:全国新华书店

开 本:175mm×245mm

印 张:15.75

字 数:306 千字

版 次:2015 年 12 月第 1 版

印 次:2015 年 12 月第 1 次印刷

印 数:1~1500

定 价:58.00 元

产品编号:066763-01

评
审
委
员
会

中国计算机学会学术著作丛书

■ 名誉主任委员：张效祥

■ 主 任 委 员：唐泽圣

■ 副 主 任 委 员：陆汝钤

■ 委 员：（按姓氏笔画为序）

王 珊 吕 建 李晓明

林惠民 罗军舟 郑纬民

施伯乐 焦金生 谭铁牛

丛书序

第

一台电子计算机诞生于 20 世纪 40 年代。到目前为止，计算机的发展已远远超出了其创始者的想象。计算机的处理能力越来越强，应用面越来越广，应用领域也从单纯的科学计算渗透到社会生活的方方面面：从工业、国防、医疗、教育、娱乐直至人们的日常生活，计算机的影响可谓无处不在。

计算机之所以能取得上述地位并成为全球最具活力的产业，原因在于其高速的计算能力、庞大的存储能力以及友好、灵活的用户界面。而这些新技术及其应用有赖于研究人员多年不懈的努力。学术研究是应用研究的基础，也是技术发展的动力。

自 1992 年起，清华大学出版社与广西科学技术出版社为促进我国计算机科学技术与产业的发展，推动计算机科技著作的出版，设立了“计算机学术著作出版基金”，并将资助出版的著作列为中国计算机学会的学术著作丛书。时至今日，本套丛书已出版学术专著近 50 种，产生了很好的社会影响，有的专著具有很高的学术水平，有的则奠定了一类学术研究的基础。中国计算机学会一直将学术著作的出版作为学会的一项主要工作。本届理事会将秉承这一传统，继续大力支持本套丛书的出版，鼓励科技工作者写出更多的优秀学术著作，多出好书，多出精品，为提高我国的知识创新和技术创新能力，促进计算机科学技术的发展和进步作出更大的贡献。

中国计算机学会

2002 年 6 月 14 日

前言



数据能够有效地反映数据之间普遍存在的联系,具有丰富的表达力,在 Web、社会网络、生物和化学数据库等领域获得了广泛的应用。随着数据获取方式的多样化,图数据规模越来越大,应用也日趋复杂,传统的集中式图查询处理和分析挖掘方法满足不了日益增长的功能和性能上的需求。特别是近年来随着云计算和大数据等概念的兴起,分布式图处理计算也随之得到快速的发展,成为热点的研究领域。本专著系统综述了目前该领域的主要研究进展,并总结和整理了作者近年来在这方面的研究成果,内容囊括大规模图数据分布式处理的主要模型、技术和系统,包括执行机制、数据组织、代表性算法,以及系统实现和典型应用等各个方面。本书试图为读者系统地展现大数据技术高速发展和变革时代大图处理区别于传统数据管理和分布式计算的新技术、新思想、新系统和新挑战。

本书共分为 10 章,第 1 章主要介绍大规模图数据分布式处理的研究背景和问题;第 2 章介绍分布式图计算模型和执行机制;第 3 章和第 4 章分别介绍基础的数据组织问题,包括数据的划分以及存储和索引;第 5 章到第 9 章介绍代表性的大图复杂查询、分析和挖掘算法及其分布式实现技术,包括三角形查询、最大 k 边连通子图查询、最小生成树搜索、频繁子图挖掘和重叠社区发现;第 10 章对现有的主要分布式大图处理系统和典型应用进行综述。

本书涉及的研究课题得到国家重点基础研究发展计划(“973 计划”)项目(No. 2012CB316201)、国家自然科学基金项目(61472071、61272179、61433008)、教育部-中国移

动科研基金项目(MCM20125021)等资助。

作者指导的部分研究生参与了本书的撰写和相关课题的研发,他们是王志刚、刘金鹏、王文安、杨佳学、张天明、张楠、毕亚辉等,他们为本书付出了辛勤的劳动,在此一并表示衷心的感谢。

该专著主要作为从事图数据管理、分布式计算和大数据分析等相关领域研究开发和管理人员的参考书籍,也可作为高校计算机和大数据等相关专业研究生的补充教材和参考读物。

由于著者水平所限,而本书涉及很多新的技术,因此书中难免有疏漏和错误,恳请读者提出宝贵意见。

作 者

2015年5月7日

目录

第 1 章 大规模图数据处理：问题与挑战	1
1.1 大图数据处理的背景	1
1.2 图数据的表示	2
1.3 传统的大图数据管理方法	4
1.4 云计算环境处理大图数据的优势	6
1.5 新型大图计算系统面临的挑战	7
1.6 关键技术问题	8
第 2 章 大图分布式处理的计算模型和执行机制	11
2.1 大图分布式处理的基本计算框架	11
2.1.1 基于 MapReduce 的计算框架	11
2.1.2 基于 BSP 的计算框架与 GAS 模型	14
2.1.3 MapReduce 与 BSP 对比	16
2.1.4 其他处理框架	17
2.2 图查询处理的遍历模式	19
2.2.1 以顶点为中心	19
2.2.2 以子图为中心	19
2.2.3 以边和路径为中心	19
2.3 消息通信	20
2.3.1 消息发送时序控制	20
2.3.2 消息交换模式	20
2.3.3 网络通信平台	22
2.3.4 上层消息优化技术	23
2.4 同步控制	23
2.4.1 同步模式	24

2.4.2	异步模式	25
2.4.3	混合模式	25
2.4.4	跨步模式	26
2.5	容错管理	26
2.5.1	故障恢复技术	26
2.5.2	故障侦测技术	30
2.6	任务调度	30
2.7	可扩展性	32
第3章	大图数据划分技术	34
3.1	图数据划分技术综述	34
3.1.1	离线划分算法	35
3.1.2	在线划分算法	36
3.1.3	动态划分算法	38
3.2	大图划分定义	41
3.2.1	处理流程和定义	41
3.2.2	真实图的局部性分析	42
3.3	OnFlyP 划分算法	44
3.3.1	Range 划分	44
3.3.2	OnFlyP 划分	45
3.3.3	负载均衡控制机制	47
3.3.4	计算接口描述	49
3.3.5	动态调整机制	51
3.4	性能评价	52
3.5	小结	55
第4章	大图数据分布式存储与索引技术	56
4.1	大图数据的存储索引技术综述	56
4.2	图迭代算法的状态转换模型	60
4.3	大图的磁盘存储管理机制	62
4.3.1	基于列存储模型的静态 Hash 索引策略	62
4.3.2	基于状态转换的动态 Hash 索引策略	66
4.4	基于消息有序的磁盘迭代	71
4.4.1	消息有序迭代 MSI	71
4.4.2	OERSV 数据模型	72

4.4.3	两阶段计算过程	74
4.5	性能评价	75
4.6	小结	77
第 5 章	大图数据分布式三角形查询技术	78
5.1	大图三角形查询技术综述	78
5.1.1	集中式算法	79
5.1.2	分布式算法	81
5.1.3	近似算法	84
5.2	分布式大图三角形查询优化技术	84
5.2.1	存储结构	84
5.2.2	EN-Iterator 算法	86
5.2.3	消息优化	87
5.3	基于采样的近似处理技术	89
5.3.1	采样策略	89
5.3.2	算法描述	91
5.4	性能评价	92
5.5	小结	95
第 6 章	大图数据分布式最大 k 边连通子图查询技术	96
6.1	大图最大 k 边连通子图查询技术综述	96
6.2	分布式最大 k 边连通子图优化技术	103
6.2.1	顶点优化	103
6.2.2	剪枝策略	104
6.2.3	消息优化	106
6.3	基于采样的近似处理技术	107
6.3.1	采样策略	107
6.3.2	算法描述	108
6.4	性能评价	109
6.5	小结	112
第 7 章	大图数据分布式最小生成树查询技术	113
7.1	大图数据最小生成树综述	113
7.2	顶点驱动的并行 MST 算法	118
7.2.1	PB 算法(分区 Prim 算法+ Borůvka 算法)	118

7.2.2	算法正确性	119
7.2.3	双重索引	120
7.2.4	终止条件	121
7.2.5	索引维护	122
7.3	基于并行处理模型的 PB 算法	122
7.3.1	基于 MapReduce 模型的 PB 算法	122
7.3.2	基于 BSP 模型的 PB 算法	124
7.3.3	PB 算法代价分析	124
7.4	动态图的 MST 维护算法	126
7.4.1	MST 结果预处理	126
7.4.2	删除边维护	127
7.4.3	删除顶点维护	129
7.4.4	维护代价	129
7.5	性能评价	130
7.6	小结	132
第 8 章	大图数据分布式频繁子图挖掘技术	133
8.1	图数据频繁子图挖掘技术综述	133
8.1.1	图数据集中的频繁模式挖掘算法	134
8.1.2	单个大图的频繁模式挖掘算法	134
8.1.3	并行图频繁模式挖掘	136
8.2	基于最大团频繁计数的频繁子图挖掘	137
8.2.1	整体框架	137
8.2.2	挖掘频繁 1-子图	140
8.2.3	候选子图产生	141
8.2.4	频繁计数	143
8.3	频繁子图挖掘分布式处理的优化	146
8.4	基于 AMNI 频繁计数的子图挖掘	149
8.5	频繁子图挖掘的 BSP 实现	151
8.6	性能评价	152
8.7	小结	155
第 9 章	大图数据分布式重叠社区发现技术	156
9.1	复杂网络重叠社区发现技术综述	156
9.1.1	团渗透方法	158

9.1.2	边图与边划分方法	159
9.1.3	局部扩展最优化算法	159
9.1.4	模糊检测法	160
9.1.5	基于混合概率模型算法	161
9.1.6	基于非负矩阵分解算法	161
9.1.7	其他类型算法	162
9.2	分布式并行极大团枚举	162
9.2.1	问题描述	163
9.2.2	极大团枚举方法	164
9.2.3	极大团枚举方法优化	166
9.2.4	并行极大团枚举方法	169
9.2.5	复杂度分析	172
9.3	复杂网络中并行重叠社区发现	173
9.3.1	问题描述	173
9.3.2	GCE 基本算法	174
9.3.3	GCE 算法的优化	175
9.3.4	GCE 算法并行化	177
9.4	性能评价	183
9.5	小结	185
第 10 章	大规模图数据分布式处理系统和应用	186
10.1	基于 MapReduce 模型的大图处理系统	186
10.1.1	PEGASUS	186
10.1.2	HaLoop	187
10.1.3	Twister	187
10.2	基于 BSP 模型的大图处理系统	188
10.2.1	Pregel	188
10.2.2	Hama	188
10.2.3	Giraph	189
10.2.4	Giraph++	190
10.2.5	GPS	191
10.2.6	X-Pregel	191
10.2.7	Pregelix	192
10.2.8	MOCgraph	192
10.2.9	Kylin	193

10.3	其他代表性系统	193
10.3.1	PowerGraph	193
10.3.2	Trinity	194
10.3.3	GBase	194
10.3.4	Spark(GraphX)	195
10.3.5	GraphLab	195
10.3.6	Chronos	196
10.3.7	LFGraph	196
10.3.8	GraphChi、X-Stream 和 TurboGraph	197
10.4	BC-BSP 系统介绍	200
10.4.1	体系结构概况	200
10.4.2	图处理作业的执行流程	201
10.4.3	PageRank 算法示例	203
10.5	大规模图数据分布式处理的应用	207
10.5.1	Web 应用	207
10.5.2	社会网络应用	210
10.5.3	生物和化学领域应用	215
	参考文献	220

第 1 章

大规模图数据处理：问题与挑战

1.1 大图数据处理的背景

随着云计算等新技术的快速发展、社交网络等新型互联网应用的兴起和各种电子设备的日益普及,人类获取和存储数据的规模正以前所未有的速度爆炸式增长,与大数据相关的技术变革成为学术界和工业界的热点问题。《中华人民共和国国民经济和社会发展第十二个五年规划纲要》提出要重点研究海量信息处理及知识挖掘的理论和方法,从国家战略层面上强调了对大数据的研究。而大图是大数据研究领域的一个重要分支,具有广泛的理论研究和应用价值。作为计算机科学中最常用的一类抽象数据结构,图能够有效地表达对象之间广泛存在的联系,在结构和语义方面比线性表和树更为复杂,更具有一般性表示能力。

现实世界中的许多应用场景都需要用图结构表示,与图相关的处理和应用几乎无所不在。传统应用如最优运输路线的确定、疾病暴发路径的预测、科技文献的引用关系等;新兴应用如道路交通管理、社交网络分析、语义 Web 分析、生物信息网络分析等。如在蛋白质交互网络中,图的顶点对应着蛋白质,边对应着蛋白质之间的联系;再如化合物的分子结构就可以被抽象为无向标号图,其中图的顶点对应着原子而边则可以表示成原子间的化学键;在社交网络中,图的顶点表示具体的用户,边表示用户之间的好友或者关注关系,相关的属性可以记录在对应的点和边上。此外,信息科学中的资源描述框架(RDF)文件、XML 文件、文本检索,以及机械工程中技术图纸的基本对象建模和通信网

络中集成电路的布局布线等领域也都大量使用了图数据。针对这些快速增长、形式多样且语义丰富的图数据,如何开发有效并且高效的查询分析技术,成为具有重要应用价值的课题。

传统的图数据库和图数据管理技术通常针对彼此独立的“小图”分别进行处理,尽管图的数目可能较多,但通常不需要复杂的迭代过程,也不会产生大量的消息,算法的时间和空间开销一般较低。然而,近年来,随着互联网的普及和 Web 2.0 技术的推动,以及生物化学等科学数据采集手段的丰富,产生了众多规模巨大且结构复杂的“大图”甚至“超大图”。以互联网和社交网络为例,据报道,Google 索引的网页数目早已超过了 1 万亿(10^{12})幅,Facebook 2012 年的活跃用户已经超过 10 亿。特别地,这种发展的势头非常迅猛,据 CNNIC 统计,2010 年中国网页规模就已经达到 600 亿,年增长率 78.6%,国内如微信、微博等社会媒体的发展也异常迅猛。而在生物信息学领域,人脑级别的图建模已经达到了 10^{14} 的规模。

真实世界中实体规模的扩张,导致相应图模型的数据规模迅速增长,动辄有数十亿个顶点和上万亿条边。以搜索引擎中常用的 PageRank 计算为例^[1],网页用图顶点表示,网页之间的链接关系用有向边表示,一个网页的 PageRank 得分根据网页之间相互的超链接关系计算而得到。假设按邻接表形式存储 100 亿个图顶点和 600 亿条边,每个顶点及出度边的存储空间占 100 字节,那么整个图的存储空间将超过 1TB。值得注意的是,庞大的顶点和边数目构成的结构信息只是大图数据规模惊人的冰山一角,复杂应用中的图数据为了表达复杂的语义,因此在顶点和边上往往会附带各类属性信息,这些属性信息内容丰富,需要大量的空间开销。此外,除了静态的结构和属性信息,相比于基于属性的简单查询和搜索,大图上的统计分析算法往往需要基于图的结构进行循环和递归操作,直至达到收敛条件,因此需要频繁地处理并行迭代过程中由于通信交互产生的消息数据等中间结果。面对如此大规模的静态和动态数据,对其存储、索引、查找和分析等处理的时间开销和空间开销远远超出了传统集中式图数据管理的承受能力。对大规模图数据的高效管理和计算,已经成为急需解决的问题,也是一项极具挑战性的工作。

1.2 图数据的表示

作为数学的一个重要分支,图论以图作为研究对象,在简单图的基础上衍生出超图理论、极图理论、拓扑图论等,使图可以从多方面表达现实世界。当前大规模图数据管理,采用的数据模型有多种,按照图中节点的复杂程度分为简单节点图模

型和复杂节点图模型,按照一条边可以连接的顶点数目分为简单图模型和超图模型。不论是简单图模型、超图模型、简单节点模型还是复杂节点模型,它们的顶点和边都可以带有属性。

1. 简单图模型

这里所说的简单图,并不是图论中的简单图,是相对于超图而言的,一条边只能连接两个顶点,可以存在环路。其形式化表示,即 $G=(V, E)$ 。其中, $V=\{v_1, v_2, \dots, v_n\}$, $E=\{e_1, e_2, \dots, e_n\}=\{\{v_1, v_2\}, \{v_3, v_4\}, \dots, \{v_{n-1}, v_n\}\}$ 。简单图的存储和处理都比较容易,对于一般的应用,简单图的表达能力完全可以胜任,如 PageRank 计算、最短路径查询等。简单图模型的常用组织存储结构包括邻接矩阵、邻接表、十字链表和邻接多重表等多种方式。不同的系统根据目标不同采用不同的表示方式。

2. 超图模型

一条边可以连接任意数目的图顶点。此模型中图的边称为超边。基于这种特点,超图比上述简单图的适用性更强,保留的信息更多。形式化表示为 $G=(V, E)$,其中, $V=\{v_1, v_2, \dots, v_n\}$, $E=\{e_1, e_2, \dots, e_n\}=\{\{v_1, v_2, v_3\}, \{v_3, v_4, v_5\}, \dots, \{v_{n-2}, v_{n-1}, v_n\}\}$ 。例如,以图顶点代表文章,每条边代表两个顶点(文章)享有同一个作者。现有三篇文章 v_1 (作者 A、B), v_2 (作者 A、C), v_3 (作者 A、D),三篇文章的作者都有 A。图 1-1 的左图表示了简单图存储模式,边集 $E=\{e_1, e_2, e_3\}=\{\{v_1, v_2\}, \{v_1, v_3\}, \{v_2, v_3\}\}$,无法直接保留作者 A 同时是三篇文章 v_1, v_2, v_3 的作者这一信息。图 1-1 的右图代表了超图存储模式,边集 $E=\{e_1\}=\{\{v_1, v_2, v_3\}\}$,超边 e_1 中直接保留了 A 是三篇文章 v_1, v_2, v_3 的作者这一信息。对于具有复杂联系的应用,可以使用超图模型建模,例如社交网络、生物医学网络等。

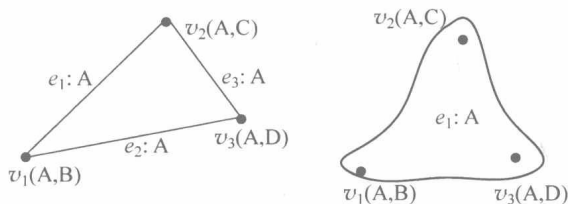


图 1-1 简单图(左)和超图(右)

超图模型的组织方式主要使用关系矩阵,它与邻接矩阵较为相似。图 1-2 展示了超图的关系矩阵表示方法。与邻接矩阵不同的是,关系矩阵的行和列分别表示图顶点编号和超边的编号,关系矩阵中,1 表示一条超边包含某个图顶点,或一个图顶点隶属于某条超边。

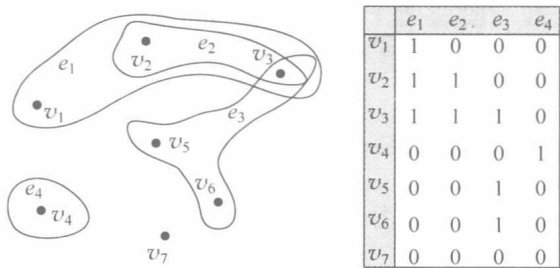


图 1-2 超图的关系矩阵表示方法

1.3 传统的大图数据管理方法

传统的分布式图处理技术主要基于 NoSQL 数据库,关注的是图数据、特别是小图数据集的管理问题,查询相对简单,具有高效的索引,可以支持数据的更新。当图数据更新时,需要解决在分布式环境下的一致性控制问题,提供事务功能。

特别地,NoSQL 数据库采用的数据模型主要有文档存储(Document Store)模型、列族存储(ColumnFamily Store)模型、Key-Value 存储模型、图形模型等几大类^[2]。

Key-Value 存储模型的存储模式简单,在高并发环境下可以提供高效的查询或遍历服务,能存储海量数据,非常适合通过主键进行查询或遍历,但对复杂的条件查询支持度不佳。从图处理的角度看,像 PageRank 计算等并不需要复杂查询,Key-Value 模型完全可以胜任。对于采用邻接表组织的图数据,可以将图顶点及其值作为 Key,将出边或出度顶点列表作为 Value。文献[3]结合语义 Web 和传统的 Key-Value 模型,提出 Key-Key-Value 模型。以社交网络为例,Key-Key-Value 模型将 Alice 和 Bob 之间的好友关系组织为一个三元组 $\langle \text{Alice}, \text{Bob}, \text{FriendShip} \rangle$ 。该模型存储的信息比传统的 Key-Value 模型更加丰富,可以据此进行数据迁移和合并,以提高时空局部性,使得在查询处理时能减少数据远程跨机读取的次数,因而可以提高数据读取效率。文档存储模型在存储格式方面十分灵活,比较适合存储系统日志等非结构化数据,对以邻接矩阵或邻接表组织的图数据来讲,意义不大。而且文档存储模型为支持灵活性所导致的处理效率的降低也会成为大规模图数据管理的瓶颈。列族存储模型比较适合对某一列进行随机查询处理,但是对于穷举式遍历,反而不如传统的面向行的存储模式。图模型的相关研究目前还不完善,只有少数分布式图数据库,如 Neo4j^[4]等采用这种模型存储图数据。文献[5]从管理数据的规模和模型的复杂性两个维度比较了这 4 种基本存储模型,见图 1-3。

在查询处理方面,传统的分布式图数据库和一些新型的大图在线查询系统往往只支持图的简单查询检索,返回用户感兴趣的信息,而通常不支持大图上的复杂