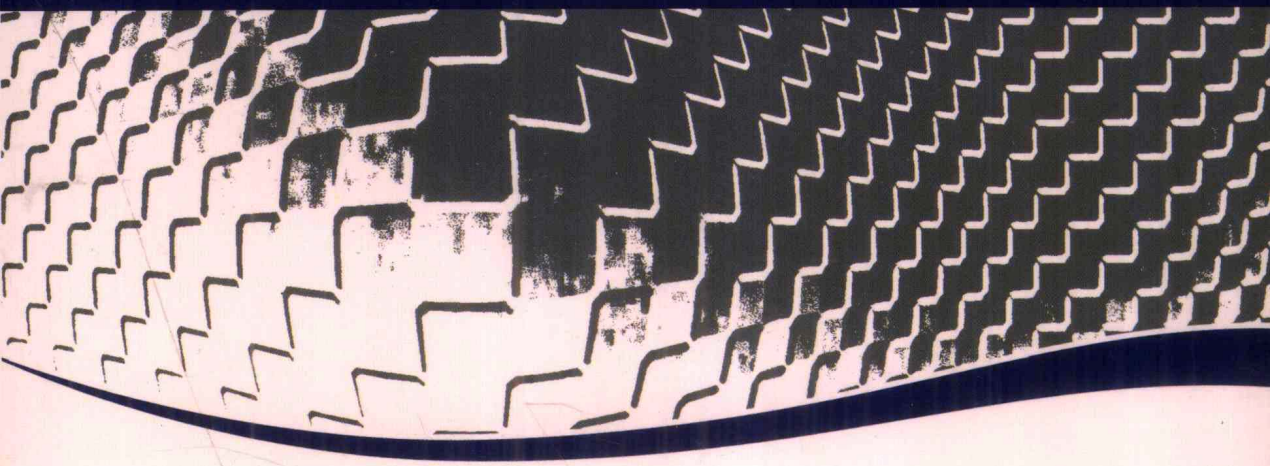
可扩展计算学术著作丛书



陈汉华　金海　著

大规模分布式内容检索技术

科学出版社

可扩展计算学术著作丛书

大规模分布式内容检索技术

陈汉华　金　海　著

科学出版社

北京

内 容 简 介

大规模分布式内容检索是近年来分布式系统方向的一个热点研究领域。本书全面地阐述了各种体系结构的分布式大规模内容检索系统的关键技术和核心理论，并对各项技术和理论的来龙去脉进行了详细深入的分析。本书通过丰富的文献资料和研究成果，从研究者的视角对大规模分布式内容检索技术进行了深入剖析，是分布式处理系统领域的学术专著。

本书可供高等院校计算机科学与技术相关专业的高年级本科生、研究生、教师、研究人员及工程技术人员阅读参考，也可作为相关专业的研究生教材。

图书在版编目(CIP)数据

大规模分布式内容检索技术/陈汉华，金海著. —北京：科学出版社，2011
(可扩展计算学术著作丛书)
ISBN 978-7-03-031417-8

Ⅰ. 大… Ⅱ. ①陈…②金… Ⅲ. 机器检索：情报检索 Ⅳ. G354.4

中国版本图书馆 CIP 数据核字(2011)第 106910 号

责任编辑：魏英杰 杨向萍 / 责任校对：林青梅
责任印制：赵 博 / 封面设计：鑫联必升

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销
*
2011 年 5 月第 一 版 开本：B5(720×1000)
2011 年 5 月第一次印刷 印张：19 1/2
印数：1—3 000 字数：372 000

定价：70.00 元

(如有印装质量问题，我社负责调换)

《可扩展计算学术著作丛书》编委会

序

从万维网的辉煌到对等系统的风靡，再到以在线社会网络应用为成功代表的云应用的崛起，互联网用户之间的信息共享不断向更便捷、更高效的形式发展。近二十年来，分布式计算系统的体系结构分分合合，网络应用层新概念、新事物层出不穷，技术飞速进步归根结底归功于摩尔定律所起的作用。在互联网兴起的初期，大多数连接到互联网上的普通用户由于计算机性能、资源等因素的限制，并没有能力对外提供服务。因而，互联网逐步形成了以少数服务器为中心的状况——客户机/服务器(C/S)模式成为 Web 和 FTP 等互联网应用的主要架构。时至今日，即使是廉价的普通机器也具备了相当的计算、存储和通信能力，甚至超过了老一代曾经风云一时的超级计算机。对等模型顺应了技术的发展而迅速兴起，并为开发大规模分布式内容共享系统奠定了核心基础。

分布式内容检索是构建大规模分布式应用系统的关键支撑技术。由于大规模分布式系统中往往不适合或难于搭建中心索引服务，因此大规模分布式内容检索的核心问题是如何建立高效的分布式索引，并基于分布式索引开发能透明支持复杂查询处理的协议和算法。《大规模分布式内容检索技术》一书围绕此核心问题展开，通过结合对等分布式模型的概念、结构、资源描述和组织、资源定位和路由选择、结果融合及排序方法等，从结构化对等网络、非结构化对等网络和混合式对等网络各自的特点出发，全面系统地阐述了各种环境下进行分布式大规模内容检索的策略。在每一类方法的讨论中，作者系统地对各项关键技术进行了深入的介绍。同时，对各项关键技术的来龙去脉进行了详细地剖析，为读者还原了各项创新性技术从思想萌芽到成熟应用的技术纵深。

该书内容结合了作者学术研究和系统开发的实践，介绍的许多技术源于作者在香港科技大学计算机科学与工程系访问期间的科研成果。

该书内容丰富、鞭辟入里，是分布式处理系统领域一本难得的学术专著，故向读者郑重推荐。

俊明选

香港科技大学讲座教授

2011 年 4 月于香港九龙清水湾

前　言

随着网络技术的迅猛发展和网络应用的迅速普及，互联网日益形成一个巨大的分布式信息库。互联网应用产生的超大规模信息对现有的网络数据管理基础设施提出了新的严峻挑战。互联网信息库的无限扩张性和与生俱来的分布式特性使非集中式的数据管理和共享机制的兴起成为一种必然趋势。大规模分布式内容检索技术具有重要的学术价值和应用价值。大规模分布式内容检索系统的核心问题是如何建立高效的分布式索引以支持大规模网络环境下的复杂内容检索。本书从此核心问题出发，通过扩展传统对等模式的概念、结构、资源描述与组织、资源发现与路由、结果融合与排序等，全面、深入、系统地论述了利用对等模型构建大规模分布式内容检索系统的解决方案和关键技术。

全书共 9 章，其中各章的主要内容安排如下：

第 1 章为绪论。本章介绍了大规模分布式内容检索面临的挑战。从对等网络的概念和特性入手，对现有大规模分布式内容检索技术进行了分类介绍。

第 2 章为分布式哈希表及单关键字全局索引。本章介绍了分布式定位的经典算法——分布式哈希表。首先介绍了构造分布式哈希表的理论基础——一致性哈希；然后介绍了四种经典的分布式哈希表算法，包括 Chord、CAN、Pastry 和 Tapestry，并对各种算法进行了对比分析；最后分别介绍了基于分布式哈希表构建的典型内容检索系统 eSearch 和 Minerva 的基本原理。

第 3 章为布隆滤波。本章系统介绍了布隆滤波这种时空高效的集合哈希编码技术。首先从传统的哈希编码技术入手，分析通过引入一定的误判率来实现高效的集合编码；然后介绍了布隆滤波的几种扩展形式，包括计数布隆滤波、压缩布隆滤波、动态布隆滤波；最后介绍了布隆滤波在网络和分布式系统中的应用。

第 4 章为基于分布式哈希表单关键字索引的搜索。本章介绍了如何在基于分布式哈希表的单关键字索引上进行多关键字复杂搜索。首先详细分析比较了几种分布式 Top-k 算法，然后重点介绍了一种利用优化布隆滤波技术来实现多关键字搜索的 PWEB 系统的设计。

第 5 章为多关键字全局索引及搜索。本章介绍了如何通过扩展标准分布式哈希表实现关键字集索引。首先介绍了经典关键字赋权模型，然后重点介绍了两个基于分布式哈希表关键字集索引的系统 HDK 和 TSS。

第 6 章为基于复制的联邦式对等搜索策略。本章介绍了如何通过复制策略来提高联邦式搜索系统。首先对各种复制策略进行了理论分析，然后重点介绍了

WP、BubbleStorm 和 BloomCast 三种有效的复制策略。

第 7 章为基于内容路由的联邦式搜索策略。本章介绍了如何重构路由表以提高联邦式搜索引擎的检索效率。分别介绍了基于语义小世界的模型、基于兴趣局部性的路由策略和一种基于语义覆盖网的搜索系统 SemreX。

第 8 章为混合式对等搜索策略。本章介绍了如何结合分布式哈希表和洪泛协议各自的特点设计混合搜索策略。分别详细介绍了基于预先探测的混合搜索策略 SimpleHybrid、基于 Gossip 的混合搜索策略 GAB 和基于查询难度感知的混合搜索策略 QRank。

第 9 章为大规模在线社会网络搜索。本章通过分析大规模社会网络的特点，结合传统分布式检索中的成熟技术，对大规模社会网络搜索中的关键技术进行了展望。

本书的出版得到了国家重点基础研究发展计划(973 计划)课题“基于语义网格的语义关联存贮模型及管理和通信平台(2003CB317003)”和国家自然科学基金委员会与香港研究资助局(NSFC/RGC)联合科研基金项目“因特网上基于对等网络的大规模实时视频系统：理论和实践(60731160630)”的资助。

作　者

2010 年 12 月于瑜珈山

目　录

第1章 绪 论

随着网络技术的迅猛发展和网络应用的迅速普及，互联网日益形成一个巨大的分布式信息库。互联网应用产生的超大规模信息对现有的网络数据管理基础设施提出了新的更为严峻的挑战。互联网信息库的无限扩张性和与生俱来的分布式特性使非集中式的数据管理和共享机制成为近年来的研究热点[1]。大规模分布式内容检索研究具有重要的学术价值和应用价值。

对等计算技术是分布式系统和计算机网络结合的产物，它在网络协议的应用层打破了传统的客户机/服务器(C/S)模式，以自主、平等的原则将处于网络边缘的计算、存储、通信、信息等各种资源有效地共享起来，形成协作网络[2,3]。自诞生以来的短短几年时间里，对等文件共享应用，如 Gnutella[4]、KaZaA[5]、BitTorrent[6]等都取得了极大的成功，并占据了当前互联网一半以上的网络流量。对等模式因其可扩展性、鲁棒性和动态自适应性等优点，在大规模互联网应用的数据管理和搜索领域日益展现出巨大的潜力。近年来，基于对等模式的大规模分布式数据管理和内容搜索系统如雨后春笋般涌现出来[7~13]。

本书将围绕大规模分布式内容检索应用展开，全面系统地阐述大规模分布式内容检索系统的关键理论和支撑技术。

1.1 对等网络概述

在互联网产生的初期，大多数连接到互联网上的普通用户由于计算机性能、资源等因素的限制，并没有能力对外提供服务。互联网逐步形成了以少数服务器为中心的状况，C/S 模式成为互联网应用的主要模式。C/S 架构对客户机的资源要求非常少，满足了早期互联网用户希望以非常低廉的成本接入互联网的需求。最近二十年来，计算机性能按照摩尔定律增长，采用廉价的计算机也能够搭建分布式高性能数据中心，提供互联网规模的服务。时至今日，即使是处于网络边缘的个人电脑也具备了相当的计算、存储和通信能力，甚至超过十多年前的超级计算机。这使得人们在互联网上进行更大规模、更广泛、更自主的信息共享成为可能。

1999 年，18 岁的美国波士顿东北大学的一年级新生 Fanning 在校园里编写了一个共享软件 Napster。Napster 能够很方便地让用户在网上交换最新的 MP3 音乐文件。与传统的提供音乐下载的网站不同，Napster 的服务器中没有一首音乐，它只维护一个轻量级的用户音乐文件索引提供查询功能，而音乐文件实际上是在

用户之间直接交换的，而且系统中的某用户在从网络中别的用户那里获取 MP3 文件的同时，也为其他用户提供下载服务。同年，Napster 程序被放到了互联网上供网友下载。Napster 一夜成名，突然成为人们争相转告的互联网“杀手锏”应用，它就像一个拥有无限货架的虚拟音像店[14]，令无数散布在互联网上的音乐爱好者美梦成真。短短一年半之间，Napster 就吸引了超过 5000 万的下载用户[15]，一度成为互联网上增长最快的应用。Napster[16]采用的对等模式随后引发了一场互联网地震，其背后隐藏的巨大的 P2P 世界也从此浮出水面。

在传统的基于 C/S 模式的互联网信息共享应用中，如 Web、FTP 文件服务等，用户之间如果要共享数据，必须经过服务器交换。在 FTP 应用中，用户先将文件上传到某个 FTP 服务器上，然后其他用户再到相应站点下载需要的文件。这种方式不但共享效率低、不可扩展，而且对用户而言极不方便。相比较而言，采用对等模式，数据可以自主分散在网络中，不像以往的 C/S 模式那样把数据存放在中心服务器上。采用对等模式，网络结点(Peer)既可以获取其他结点的资源或服务，又可以提供资源或服务，即兼具客户机和服务器的双重身份，这种特征使得互联网上的任意两台计算机之间直接共享文件成为可能[15]。Napster 无意间采用的对等模式打破了 C/S 模式的束缚，充分挖掘了蕴藏在互联网中的巨大的计算、存储和通信资源，高度共享，形成协作网络。对等计算作为新型的分布式计算模型，正深刻地影响着互联网用户发布、共享和获取信息的方式，并被称为改变互联网的新一代网络技术[2]。

一般来说，对等网络具有以下基本特征：

① 自主平等性。对等网络中的资源分散在所有结点上，信息交换直接在结点之间自主进行，无需中间环节的介入，避免了可能的性能瓶颈，提高了资源共享的效率。对等结点之间交换哪些信息，可由结点本身自主决定，而不必受限于其他网络结点。在对等模式中，虽然网络结点能力不同，但所有成员在功能、地位上都是平等的，没有谁拥有超越其他结点的特权，没有谁能控制或者限制其他结点。对网络模型而言，平等指的是打破传统的 C/S 模式，取消服务器这一特权结点的存在，让所有网络结点之间平等地交流信息。平等性是对等网络的工作基础，对等网络对网络带宽的高效利用、对网络结点潜力的充分开发以及可扩展性等，都是基于其平等性的[15]。

② 可扩展性。在传统的 C/S 架构中，系统能够容纳的用户数量和提供服务的能力主要受服务器的资源限制，为了支持互联网上的大量用户，需要在服务器端使用高性能的计算机和高带宽的网络。一般来说，当系统网络结点数目增加时，服务器的负载也随之线性增长，如果原来的服务器不堪重负，则必须购买增加新的服务器来分担原有服务器的负载。此外，C/S 架构往往需要集中式服务器之间的同步、协同处理，也会产生大量的开销，使系统难以扩展。相比之下，对等网络具有非常

高的可扩展性，当结点数目大量增加时，随之增加的存储、处理、通信开销被更多的结点分担，系统结点的平均负载不会增加太多。因为网络用户规模的增大也带来了系统整体资源和服务能力的同步扩充，所以系统无需增加额外的设备。对等网络这种良好的可扩展性优点已经在一些大规模实例系统中得到了证明[2]。

③ 鲁棒性和动态自适应性。互联网上随时可能出现异常情况，如网络中断、网络拥塞、结点失效等，都会影响系统的可用性和稳定性。在传统的集中式服务模式中，集中式服务器成为整个系统的要害所在，一旦发生异常，将影响到所有用户使用系统服务。相比较而言，在对等网络中，服务是由分散的各个结点协同提供的，部分结点或网络遭到破坏不会造成整个协同网络的瘫痪。对等网络是动态变化的，结点加入或离开，数据随时动态更新。对等网络协议一般针对这种动态性提供自适应拓扑调整、更新功能，保持网络的连通性和服务的可用性[2]。

④ 隐私性。随着互联网的普及和计算、存储及通信能力的飞速增长，收集隐私信息正在变得越来越容易。例如，目前的互联网攻击者往往通过监控用户的流量特征，获得 IP 地址，进而追踪到个人用户。实践中解决互联网隐私问题主要采用中继转发的方法，将通信双方隐藏在众多的网络实体中，实现匿名通信。在传统的匿名通信系统中，实现这一机制依赖于某些中继服务器结点。而在对等网络中，信息的传输分散在各结点之间进行，所有参与者都可以提供中继转发的功能，信息传输无需经过某个特定的集中环节，从而大大提高了匿名通信的灵活性和可靠性，能够为用户提供更好的隐私保护。另外，对等网络也采用分布式哈希表(DHT)[18~21]等技术保障其匿名性[22]。分布式哈希表采用一致性哈希[23]将用户信息(如 IP 地址等)、数据对象信息映射到一个没有任何意义的数值标识(ID)，用来唯一地标识用户和数据对象。一致性哈希具有单向不可逆性和抗冲突性等特性，这使得网络攻击者很难从 ID 破解出它实际所代表的真实信息[15]。

采用对等模式，网络结点在网络体系结构的应用层中建立虚拟连接，从而整个分布式系统中的所有结点互连组成了一个逻辑虚拟网络，简称覆盖网络(overlay network)。覆盖网络构建于底层物理网络之上，结点间通信依赖于底层物理网络通信协议的支持[24]。

根据覆盖网络的结构，对等网络可以分为集中索引对等网络、结构化对等网络和非结构化对等网络[24]。

(1) 集中索引对等网络

集中索引对等网络维护一个集中式的目录服务器用于注册和查找网络中的数据。用户访问数据时，首先搜索请求被发送到集中目录服务器进行资源定位，获取数据的地址，然后直接从地址所在结点获取数据，而不需要经过集中服务器。集中索引服务器对等模式，虽然避免了数据在服务器上集中放置，但集中式的索引服务器仍然是系统的核心。这自然使此类系统保留了集中结构的一些缺点：集中索引

服务器是系统潜在的性能瓶颈[25]，随着系统规模的增长，搜索请求和目录更新会造成巨大的负载，保障系统的可扩展性是这类系统面临的一个主要问题；集中索引也是系统潜在的单一失效点，索引服务器故障或者网络攻击，都有可能造成整个系统的崩溃。集中索引对等系统的典型代表是 Napster 和 Maze[26]。

(2) 结构化对等网络

结构化对等网络[27~29]一般使用分布式哈希表[18~21]技术对数据的放置进行严格控制。分布式哈希表以分布的方式维护逻辑上统一的哈希表，具有良好的可扩展性。另外，哈希函数的映射原理使结构化对等网络只提供精确查找功能，这虽然能有效保证系统查全率(recall)[30,31]，但不直接支持灵活的查询处理，如多关键字搜索和区间查询[32]等。研究人员在结构化对等网络方面进行了大量卓有成效的研究工作，下面简要介绍基于分布式哈希表的一些流行算法。

① Chord[18,33]是麻省理工学院和加州大学伯克利分校设计的一种基于分布式哈希表的查找和路由协议。Chord 采用统一的哈希函数将网络对象(key，value)中的 key 和网络结点信息(如 IP 地址)映射到相同的值空间，然后按照哈希值临近原则将对象(key，value)存储在最接近 key 的哈希值的结点上。Chord 协议采用了类似二分查找的方法，每次查找发送的消息数为 $O(\log_2 n)$，其中 n 为网络结点数。结点离开或加入网络时，拓扑维护所需网络消息开销为 $O(\log_2^2 n)$，因此一般认为当网络结点动态性很强时，Chord 维护成本比较高。总体来说，Chord 算法精妙，并具有扩展性好、查询效率高、负载均衡、自适应性拓扑维护等优点。然而，由于采用哈希映射，Chord 仅支持精确的对象查询，不直接支持复杂查询应用，这也是其他几种分布式哈希表算法的共同问题。

② CAN[19]是加州大学伯克利分校和 AT&T 设计的一种分布式查找和路由算法。CAN 将网络中所有结点映射到一个 n 维的笛卡儿空间中，并为每个结点尽可能均匀地分配一块区域。CAN 采用的哈希函数通过对(key，value)对中的 key 进行哈希运算，得到笛卡儿空间中的一个点，并将(key，value)对存储到该点所在区域对应的结点上。CAN 采用的路由算法简单有效，当前结点知道目标点的坐标后，就利用贪心算法的思想，将请求传给其四邻中坐标最接近目标点的结点，依次路由，直至查询达到目标结点。CAN 具有良好的可扩展性，其路由路径长度为 $O(\sqrt[d]{n})$，结点状态信息和网络规模无关，为 $O(d)$，其中 n 为结点数，d 为系统维数。

③ Tapestry[20]是加州大学伯克利分校 Zhao 等设计的面向广域网分布式数据存取的结构化对等系统。Tapestry 结点通过逐位匹配的方式路由查找请求，所以结点同样匹配自己标识符的每一个前缀相同而下一位不同的邻居信息。逐位匹配的思想类似于 Chord 的二分查找思想，有效提高了路由效率。系统中的每个结点有 $O(\log_2 n)$ 个邻居，由于每跳匹配一位，故路由路径为 $O(\log_2 n)$ 跳，其中 n 为结点数。网络中的结点能够通过选择最佳邻居结点保证最小化平均路径延迟。

Tapestry在构建拓扑时考虑了覆盖网络和底层网络的匹配问题，有效提高了系统效率。

④ Pastry[21]是微软研究院提出的一种可扩展的分布式查找路由协议，可用于构建大规模的对等系统。Pastry 源于 Plaxton 路由机制[34]，Plaxton 是第一个基于分布式哈希表的算法，但设计目标不是用于对等系统，所以只假设了静态环境，并且没有提供很有效的路由查找。在 Pastry 中每个结点分配一个 128 位的结点标识符号(nodeID)，所有的结点标识符被映射到环形的 nodeID 空间，范围从0～$2^{128}-1$，结点加入系统时通过哈希结点 IP 地址在 128 位 nodeID 空间中随机分配。Pastry 采用了类似 Tapestry 的逐位匹配思想，算法具有指数收敛性。在 Pastry 中路由一个消息需要 $O(\log_2 n)$ 步，每个结点需要维护 $O(\log_2 n)$ 个入口，其中 n 为网络结点数。

(3) 非结构化对等网络

非结构化对等网络[35~37]对覆盖网络拓扑结构以及系统中资源的放置没有严格的限制[38]，因此具有较好的鲁棒性，能够很好地应对网络中结点的高度动态变化。该类系统中的搜索主要是通过结点间的消息洪泛转发，协议简单且易于支持各类查询语法。然而，采用基于洪泛思想的算法，搜索消息数会在网络上指数增长并导致大量冗余的通信负载，从而限制了非结构化对等网络的可扩展性[4]。早期的非结构化对等网络不考虑结点之间的区别，后期的对等网络协议考虑了网络中结点之间存在的性能差异，一些具有较高网络带宽和较长在线时间的网络结点以超级结点的身份加入对等网络，每个普通结点至少与一个超级结点连接。超级结点在搜索过程中代理与之相连的普通结点执行搜索请求。研究表明，对于广泛分布的热点数据，非结构对等协议搜索性能很高，而对于分布非常稀少的数据，其搜索性能却很差[30,39,40]。非结构对等系统的典型代表是 Gnutella 和 KaZaA 等。

1.2 基于对等模式的大规模分布式文本内容检索

随着网络技术的迅猛发展和网络应用的迅速普及，互联网日益成为一个巨大的分布式信息库。特别是以 FaceBook、Wikipedia、MySpace、YouTube、Google Document 和 Flickr 等为代表的云计算应用兴起以来，越来越多的用户正以更广泛、更直接和更高效的方式参与到互联网信息的创造和共享行列中来。互联网信息库的无限扩张性和与生俱来的分布式特性使研究非集中式的数据管理和共享机制成为一种必然趋势。基于分布式模型的新一代内容检索技术已成为近年来国内外学术界和工业界的新热点。对等计算模型凭借其分布式、易扩展、容错性高等优点，日益在互联网信息搜索方面显示出巨大的潜力。

① 互联网信息的无限增长和集中式搜索引擎数据中心处理能力的局限性之