中国计算机学会学术著作丛书

P2P流量识别方法研究

孙知信 著

清华大学出版社

中国计算机学会学术著作丛书

P2P流量识别方法研究

孙知信 著

清华大学出版社

内 容 简 介

本书系统地阐述了 P2P 流量识别技术。首先综述了 P2P 流量识别技术现状和最新的研究成果，在此基础上详细阐述作者提出的 6 种 P2P 流量检测模型：基于滑动窗口机制的 P2P 流量识别模型、基于通信网络拓扑结构的 P2P 流量识别模型、基于 BP 算法的 P2P 流量识别模型、基于多重特征分类的 P2P 流量识别算法、基于 SVM 的 P2P 流量识别方法的设计与实现以及基于流特征描述的模糊识别算法。在理论研究的基础上，作者将上述模型和算法应用到具体的项目开发中，取得了良好的效果。

本书是作者多年从事科研项目研究的成果结晶，书中内容都来自具体的项目，有很好的工程基础，特色是学术与具体的工程应用相结合。本书可作为计算机科学与技术、网络与信息安全相关专业研究生及高年级本科生的教材，也可作为科研人员的参考书，同时可作为研究生、博士生及教师论文写作的参考书。

图书在版编目(CIP)数据

P2P 流量识别方法研究/孙知信著. --北京：清华大学出版社，2014
中国计算机学会学术著作丛书
ISBN 978-7-302-34072-0

Ⅰ. ①P…　Ⅱ. ①孙…　Ⅲ. ①计算机网络－流量－识别－研究　Ⅳ. ①TP393

中国版本图书馆 CIP 数据核字(2013)第 238087 号

责任编辑：闫红梅　王冰飞
封面设计：傅瑞学
责任校对：焦丽丽
责任印制：李红英

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
课 件 下 载：http://www.tup.com.cn，010-62795954

印 装 者：北京国马印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：17.5　　**字　　数**：423 千字
版　　次：2014 年 7 月第 1 版　　**印　　次**：2014 年 7 月第1 次印刷
印　　数：1～2000
定　　价：39.00 元

产品编号：042601-01

前言

P2P 网络服务的概念最早于 1969 年由 Steve Crocker 提出，每个参与 P2P 网络服务的主机都称为一个 Peer，由 Peer-to-Peer 连接构成的网络就称为 P2P 网络。P2P 网络不仅能提供快速高效的文件共享、低成本高可用的计算资源和存储资源共享，而且具有强大的网络连通性，以及更直接、更灵活的信息传送能力。然而，P2P 网络在提供高效快速资源共享的同时，也带来了众多的问题：P2P 对传输速度的需求高，且 P2P 应用的数据交换具有一对多、突发性和分布式特性，P2P 用户的超大容量下载消耗了大量带宽；其用户抢占了 60%～80%的网络带宽，容易引起企业及 ISP 瓶颈链路的阻塞；P2P 用户不分时段地进行高速下载，增大网络设备的负荷，容易造成高峰时段的链路拥塞等。随着 P2P 应用的快速发展，P2P 指数级增长的数据传输使得网络带宽资源更加紧张。因此，为了确保其他正常网络用户的服务，同时为了更好地发挥 P2P 技术的优势，对 P2P 流量进行分类、识别就成为目前业内研究的一个热点。

本书作者从 2004 年开始研究 P2P 流量识别的相关技术，先后得到国家自然科学基金、科技部中小企业创新基金、国家重点实验室基金、江苏省自然基金、江苏省高校自然基金项目及企业委托项目的资助，取得了以下成果。

(1) 立足于 P2P 协议最基本的特点，研究并总结出 P2P 流量的多个统计特性，提出了基于滑动窗口机制的 P2P 流量识别模型(SW-P2PIM)，并建立了 P2P 流量识别与控制仿真系统，详细内容在第 2 章中阐述。

(2) 根据 P2P 独有的通信网络拓扑结构特征，提出了基于通信网络拓扑结构的 P2P 流量识别模型(P2P-CNTIM)，该模型使用多主机特征以及通信对端类型特征对 P2P 流量进行识别，并将这两个特征有机地结合起来以提高识别的准确率和识别效率，第 3 章给出了详细分析。

(3) 将 BP 网络应用于 P2P 流识别，提出了一种基于改进 BP 算法的 P2P 流识别模型(IBPNN-P2PIM)。针对传统 BP 算法的缺陷，在总结了他人研究成果的基础上，对传统 BP 算法进行了改进，并将其应用于模型中流分类器模块的设计，与采用传统 BP 算法的流分类器相比，IBPNN-P2PIM 对 P2P 流的识别具有一定的有效性和优越性，第 4 章展示了其应用效果。

(4) 通过大量的实验对比分析 P2P 应用和非 P2P 应用，总结出 P2P 的连接特征、深层数据包特征和流量统计特征这些特征，并在此基础上提出了基于多重特征分类的 P2P 流量识别模型(MCC-P2PIM)，第 5 章中对该系统实现进行了剖析。

(5) 基于近几年对支持向量机(Support Vector Machines，SVM)技术的深入研究，致力于将 SVM 技术运用到 P2P 流量识别中，分别提出了基于改进 SVM 的 P2P 流量检测模型、基于 P2P 与 DPI(Deep Packet Inspection，深度包检测)的 P2P 流量识别方法以及基于 MSVM(Multidimensional Support Vector Machine，多维 SVM)的 P2P 流量识别模型，第 6

章描述了这3个模型及其系统的成果。

(6) 首次将模糊数学的理论运用到P2P数据流的识别中，提出了一种基于流特征描述的模糊识别方法(FCD)。该识别方法能够较好地识别网络流量中的某些网络应用流，对于其他的网络应用流量识别同样适用，具有较好的准确性和可扩展性，第7章对此进行了深入分析和阐述。

本书是项目组集体成果的结晶，他们是宫婧副教授、刘三民博士、焦琳硕士、姜举良硕士、陈松乐博士、卜凯博士、许刘兵硕士、余小芳硕士、颜小倩硕士、高同硕士、张玉峰硕士等，在此书交稿之际，谨向他们对本书做出的贡献表示衷心的感谢。另外，感谢宫婧副教授、刘三民博士、汪胡青博士、何丽萍博士在本书撰写过程中给予的帮助。

感谢我的爱人张娟和儿子孙翌博，他们是我写书的动力所在。

最后要感谢清华大学出版社的领导和编辑，没有他们的辛勤劳动，就没有本书的出版。

孙知信

2014.5 于南京邮电大学

目 录

第1章 绪 论

近年来，对等计算(Peer-to-Peer，P2P)技术迅速发展，日益受到计算机界的关注和青睐，迅速成为业界关注的热门话题之一。P2P 技术在文件共享、实时流媒体、视频点播和分布式计算系统等领域都有重要应用。P2P 技术的使用使得用户可获得的资源更广泛，内容更丰富，形式更多样，但 P2P 技术的广泛使用带来许多负面影响，如吞噬网络资源、知识侵权、网络安全等，这些问题也导致 P2P 流量识别研究迫在眉睫。同时 P2P 流量的准确高效识别是网络运营商或网络管理员开展网络活动的前提。

1.1 P2P 形成背景及概念

1.1.1 P2P 产生背景

随着通信技术的发展，Internet 已经融入到社会各个领域，成为人们获取信息的重要途径。同时，人们对 Internet 的要求也越来越高，已经不满足于浏览网页、聊天等基本功能，更希望获得丰富多彩的多媒体信息，如视频、音频等。相比于浏览网页、聊天等传统应用，网络多媒体应用具有持续时间长、数据量大、占用网络带宽高等特点。网络基础设施性能的增强和多媒体技术的发展，也使得在 Internet 上开展更为复杂的多媒体服务成为可能。

在传统的客户/服务器(C/S)模式中，服务器为所有客户提供服务，数据的上传、下载都要经过服务器的处理。客户请求服务，服务器提供服务，客户端都是主动与服务器建立连接，请求具体的资源或请求提供具体的服务，而服务器则被动地等待客户端发起连接，并且客户端之间不能互相通信。例如，Web 服务、邮件服务、FTP 服务等都是 C/S 模式的服务。C/S 模式网络结构如图 1-1 所示[1]。

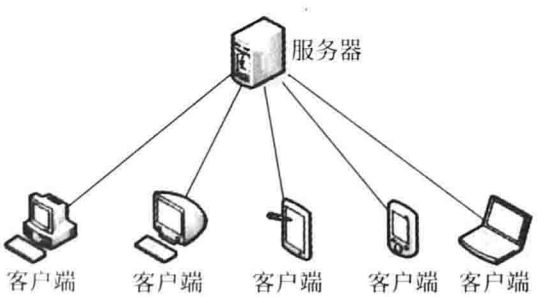

图 1-1 C/S 模式网络结构图

C/S模式中服务器的负载随着客户节点增多而线性增加，容易成为单点瓶颈。并且一旦服务器失效，则整个网络就会瘫痪，易形成单点失效问题。当网络规模达到一定程度时，C/S模式会影响系统的性能，降低系统的可扩展性，所以有必要寻求新的网络模式分散服务器功能，提高网络灵活性。同时计算机软硬件技术的发展，使得客户节点拥有越来越强的计算和存储能力，人们开始意识到可以将部分服务功能分散到客户机上。在C/S模式下，客户节点的计算和存储能力没有得到充分的使用，也很难将网络边缘节点的空闲资源使用起来。因此，P2P模式在这样的背景下应运而生。

和C/S模式相比，P2P模式能够充分利用边缘节点的资源，降低服务器的负载压力。在P2P模式中，所有参与节点是对等的，既是传统意义上的客户节点，也被赋予服务器的功能。节点发送请求的同时，也可以为其他节点提供服务，参与节点之间可以直接相互交换数据。P2P模式网络结构如图1-2所示。P2P模式和C/S模式相比，具有以下优势[2]。

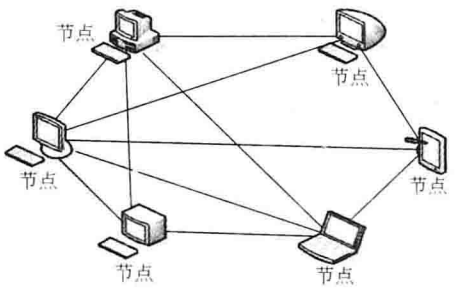

图1-2 P2P模式网络结构图

(1) 系统的可扩展性好、可靠性强，解决C/S模式下的单点瓶颈问题和单点失效问题。

(2) 有效地利用网络边缘节点闲置资源，包括计算、缓存和带宽等。

(3) 降低服务器端的负载，将部分负载从服务器转移到客户端和基础网络上来。

1.1.2 P2P概念

P2P是英文Peer-to-Peer的缩写，即对等网络或点对点网络。P2P网络中每个主机称为节点(Peer)，其中所有节点具有同等权利和义务，其相关概念最早是1969年由Steve Crocker提出。最近几年，P2P网络日益受到计算机界的关注和青睐，迅速称为业界关注的热门话题之一。但到目前为止，无论是学术界还是工业界对P2P都没有一个统一的定义，文献[2]列举了几种常见的P2P定义。

定义1.1 Peer-to-Peer is a type of Internet network allowing a group of computer users with the same networking program to connect with each other for the purposes of directly accessing files from one another's hard drives.

定义1.2 Peer-to-Peer networking (P2P) is an application that runs on a personal computer and shares files with other users across the Internet. P2P networks work by connecting individual computers together to share files instead of having to go through a central server.

定义1.3 P2P是一种分布式网络，网络的参与者共享他们所拥有的一部分硬件资源(处理能力、存储能力、网络连接能力、打印机等)，这些共享资源需要由网络提供服务和内容，能被其他对等节点(Peer)直接访问而无须经过中间实体。在此网络中的参与者既是资源(服务和内容)提供者(Server)，又是资源(服务和内容)获取者(Client)。

Intel将P2P技术定义为"通过系统间的直接交换达成计算机资源与信息的共享"，这些资源与服务包括信息交换、处理器时钟、缓存和磁盘空间等。

IBM 则把它看成是由若干互联协作的计算机构成的系统并具备如下特性[1]。

(1) 系统依存于边缘化(非中央式服务器)设备的主动协作,每个成员直接从其他成员而不是从服务器中受益。

(2) 系统中成员同时扮演服务器与客户端的角色;系统应用的用户能够意识到彼此的存在而构成一个虚拟或实际的群体。

虽然 P2P 网络定义有多种,但其本质是相同的。即在 P2P 网络中每个主机节点的地位是对等的,每个节点既充当服务器,为其他节点提供服务,同时享用其他节点提供的服务。

根据以上定义,可以总结出 P2P 网络有如下特点[3]。

(1) 非集中式:网络中的资源和服务分散在所有节点上,信息的传输和服务的实现都直接在节点之间进行,避免了可能的瓶颈。

(2) 可扩展性:在 P2P 网络中,随着用户的加入,不仅服务的需求增加,系统整体的资源和服务能力也在同步地扩充,始终能较容易地满足用户的需要。理论上,整个体系是分布的,不存在瓶颈。

(3) 健壮性:P2P 架构天生具有耐攻击、高容错的优点。由于服务是分散在各个节点之间进行的,部分节点或网络遭到破坏对其他部分的影响很小。P2P 网络一般在部分节点失效时能够自动调整整体拓扑,保持其他节点的连通性。P2P 网络通常都是以自组织的方式建立连接的,允许节点自由地加入和离开。P2P 网络还能够依据网络带宽、节点数、负载等变化不断地做自适应的调整。

(4) 高性能/价格比:性能是 P2P 被广泛关注的一个重要原因。采用 P2P 架构可以有效地利用互联网中散布的大量普通节点,将计算任务或存储资料分布到所有节点上。利用其中闲置的计算机能力或存储空间,达到高性能计算和海量存储的目的。通过利用网络中的大量空闲资源,可以用更低的成本提供更高的计算和存储能力。

(5) 隐私保护:在 P2P 网络中,由于信息的传输分散在各个节点之间进行而无须经过某个集中环节,用户的隐私信息被窃听和泄露的可能性大大减小。

(6) 负载均衡:P2P 网络环境下由于每个节点既是服务器又是客户机,减少了对传统 C/S 结构服务器计算能力、存储能力的要求。同时因为资源分布在多个节点,更好地实现了整个网络的负载均衡。

1.2 网络拓扑结构

1.2.1 网络拓扑结构概念

计算机网络拓扑结构是指从网络电缆的物理连接抽象出来的网络连接形式,是网上计算机或设备与传输媒介形成的节点与线的物理构成模式,它用几何形状代表网络电缆构成,用来表示网络服务器、工作站以及其他网络设备的连接关系。网络的节点有两类:一类是转换和交换信息的转接节点,包括节点交换机、集线器和终端控制器等;另一类是访问节点,包括计算机主机和终端等。线则代表各种传输媒介,包括有形的和无形的。网络拓扑图是理解和研究网络结构和分布的语言,它反映了网络连接关系的本质,不仅可以反映网络节点在结构中的位置,而且还排除了一些没有反映网络本质特征的细节,如网络连接所使用的

缆线类型和网络主机使用的操作系统等。在对网络结构进行设计时，必须依靠网络拓扑图反映网络主机在网络中所处的位置和连接关系，从而指导硬件设备实施和网络布线工程。从某种意义上说，网络拓扑图就是网络建设的蓝图。当网络投入使用后，网络拓扑图仍有着非常重要的意义。当网络结构需要进行改变时，需要参考网络拓扑图，并在结构改变后对网络拓扑图进行相应的修改。另外在网络维护方面，参考网络拓扑图，并结合网络操作系统的网络监控工具，可以很快地发现导致网络故障的原因[4]。

1.2.2 传统网络拓扑结构

传统计算机网络的拓扑结构主要有总线型、星形、环形、网状和混合型[5]。

1. 总线型拓扑

总线型拓扑结构用高速公用主干电缆作为传输介质，所有的节点都通过相应的硬件连接器接到传输介质上。总线两端连有终结器，其中一端接地。网络中任意节点发送的信号都沿着传输介质传播，且能被其他站点接收。节点接到信息时，先要分析该信息的目标地址与本地地址是否相同，相同则接收该信息，否则拒绝接收。总线型拓扑结构如图 1-3 所示。

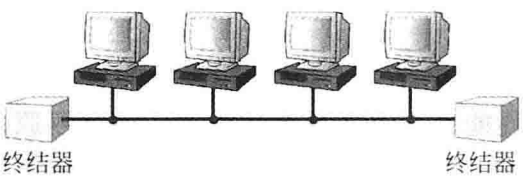

图 1-3 总线型拓扑结构

总线型拓扑结构的优点：一是所需电缆数量较少；二是结构简单，无源工作有较高可靠性且易于扩充。

总线型拓扑结构的缺点：一是总线传输距离有限，通信范围受到限制；二是主干总线对网络起决定性作用，总线故障将影响整个网络；三是总线拓扑不是集中控制的，所以故障检测需要在网上的各个节点进行，故障诊断和隔离比较困难；四是分布式协议不能保证信息的及时传送，不具有实时功能，节点必须有介质访问控制功能，从而增加了站点的硬件和软件开销。

总线型拓扑结构适用于计算机数目相对较少的局域网络，通常这种局域网络的传输速率为 100Mbps，网络连接选用同轴电缆。

2. 星形拓扑

星形拓扑结构中存在一个中心节点，其他节点直接与中心节点(Hub)相连构成的网络拓扑。星形拓扑网络属于集中控制型网络，整个网络由中心节点执行集中控制管理，各节点间的通信都要通过中心节点。星形拓扑结构图如图 1-4 所示。

星形拓扑结构的优点：一是控制简单，易于网络监控和管理；二是故障诊断和隔离容易；三是中心节点可以方便地对各个节点提供服务和网络配置。其不足之处主要在于中心节点负担过重，易形成“瓶颈”问题。

总的来说，星形拓扑结构相对简单、便于管理、建网容易，是目前局域网普遍采用的一种

网络拓扑结构。

3. 环形拓扑

环形拓扑结构中各节点通过环路接口连在一条首尾相连的闭合环形通信线路中，环路中各节点地位相同，环路上任何节点均可请求发送信息，请求一旦被批准，便可以向环路发送信息，如图 1-5 所示。环形网中的数据按照设计主要有单向传输和双向传输(双向环)。由于环线公用，一个节点发出的信息必须穿越环中所有的环路接口，信息流的目的地址与环上某节点地址相符时，信息被该节点的环路接口所接收，并继续流向下一环路接口，直到流向源节点为止。由于多个设备共享一个环，因此需要对此进行控制，以便决定每个节点在什么时候可以把分组放在环上。

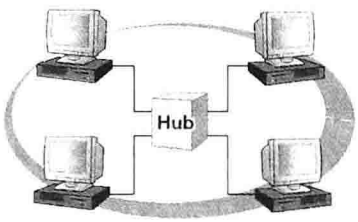

图 1-4 星形拓扑结构图

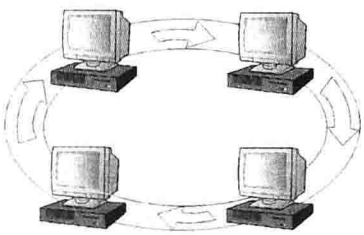

图 1-5 环形拓扑结构图

环形拓扑结构的优点：一是信息在网中沿固定方向流动，两个节点间仅有唯一的通路，简化了路径选择的控制；二是当某个节点发生故障时，可以自动旁路(由“中继器”完成)，可靠性较高；三是所需电缆长度比星形拓扑结构要短得多，同时不需像星形拓扑结构那样配制接线盒。

该拓扑结构不足主要体现在：一是扩充环的配置比较困难；二是由于信息是串行穿过多个节点环路接口，当节点过多时，影响传输效率，使网络响应时间变长；三是环上每个节点接到数据后，要负责将它发送至环上，这意味着要同时考虑访问控制协议。节点发送数据前，必须事先知道传输介质对它是可用的。

环形网结构比较适合于实时信息处理系统和工厂自动化系统。

4. 网状拓扑

网状拓扑结构指各节点通过传输线互联连接起来，并且每一个节点至少与其他两个节点相连，如图 1-6 所示。

网状拓扑结构的特点有：一是网状拓扑结构提供了通过网络的冗余路径，可靠性高；二是结构复杂，不易管理和维护；三是所需线路太多，造价很昂贵。

网状拓扑结构主要用在广域网内，局域网中较少使用。

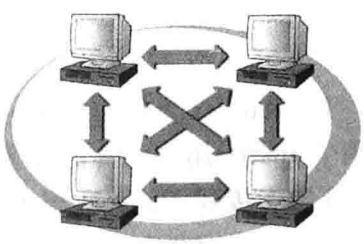

图 1-6 网状拓扑结构图

5. 混合型拓扑

将两种或几种网络拓扑结构混合起来构成的一种网络拓扑结构称为混合型拓扑结构。混合型拓扑结构可以是不规则型的网络，也可以是点-点相连结构的网络。例如，可以将星形结构和总线型结构的网络结合在一起，这样的拓扑结构更

能满足较大网络的拓展，解决星形网络在传输距离上的局限，同时解决了总线型网络在连接用户数量的限制。这种网络拓扑结构同时兼顾星形网络与总线型网络的优点。

1.2.3 P2P 网络拓扑结构

P2P 网络拓扑结构主要有中心化 P2P、全分布非结构化 P2P、全分布结构化 P2P、半分布式 P2P[1]。

1. 中心化 P2P

中心化拓扑结构最大优势是维护简单、资源查找效率高。资源查找依赖中心化的目录系统，其相关查找算法灵活高效并能够实现复杂查询。由于资源查找过分依赖于中心服务器，因此容易造成单点故障，出现访问“热点”现象。中心化 P2P 网络架构如图 1-7 所示，图中的服务器负责记录共享信息以及对查询的响应。这种形式虽具有中心化的特点，但它不同于传统意义上的 C/S 模式。传统意义上的 C/S 模式采用的是一种垄断的手段，所有资料都存放在服务器上，客户机只能被动地从服务器上读取信息，并且客户机之间不具有交互能力。而在中心化 P2P 网络中，服务器只保留索引信息，由对等节点负责保存各自提供服务的全部资料，此外服务器与对等实体以及对等实体之间都具有交互能力。在中心化 P2P 网络架构中，一台高性能的服务器保存着网络中所有主机的地址信息及其提供的共享资源目录信息。当需要查询某个文件时，对等节点会向服务器发出文件查询请求。服务器进行相应的检索和查询后，会返回符合查询要求的对等节点地址信息列表。查询发起对等节点接收应答后，根据网络流量和延迟等信息进行选择，选择合适的对等节点建立连接，并开始文件传输[6]。

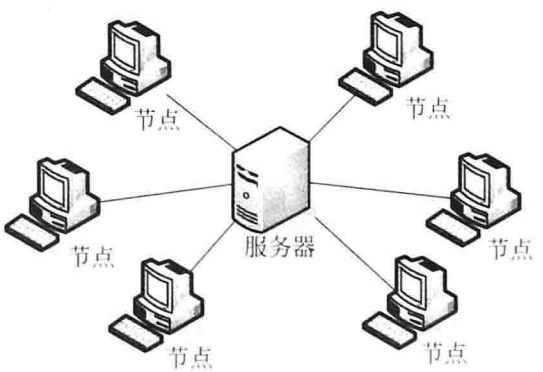

图 1-7 中心化 P2P 网络架构

以 Napster 为代表的第一代 P2P 采用中心化的拓扑结构。它通过一个中央服务器保存所有 Napster 用户上传的音乐文件索引和存放位置信息。当某个用户需要某个音乐文件时，首先连接到 Napster 服务器，在服务器上进行检索，并由服务器返回存有该文件的用户信息，再由请求者直接连接到文件的所有者传输文件。Napster 首先实现了文件查询与文件传输的分离，有效地节省中央服务器的带宽消耗，减少了系统的文件传输延时。这种方式最大的隐患在中央服务器上，如果该服务器失效，整个系统都会瘫痪。

中心化对等网络的优点：采用快速搜索算法，排队时间短，发现算法能够灵活高效地实现复杂的查询。

中心化对等网络的缺点如下。

(1) 中央服务器的瘫痪容易导致整个网络的崩溃,可靠性和安全性较低。

(2) 随着网络规模的扩大,对中央目录服务器进行维护和更新的费用将急剧增加,所需成本过高。

(3) 中央服务器的存在引起共享资源在版权问题上的纠纷,并因此被攻击为非纯粹意义上的 P2P 网络模型。

对小型网络而言,中心化 P2P 模型在管理和控制方面占一定优势。但鉴于其存在的种种缺陷,该模型并不适合大型网络应用[7]。

2. 全分布非结构化 P2P

该拓扑结构采用随机图的组织方式,节点度数服从幂律分布[8],能够快速发现目的节点,对网络的动态变化有较好的容错能力,同时可以支持复杂查询,如模糊查询、带有规则表达式的多关键词查询等。典型代表是 Gnutella 软件,它是一个 P2P 文件共享系统,与 Napster 最大的区别在于 Gnutella 没有索引服务器,是纯粹的 P2P 系统,它采用了基于完全随机图的泛洪(Flooding)发现和随机转发(Random Walker)机制。在 Gnutella 分布式对等网络模型中,所有的对等节点既是查询的发出者,又是搜索处理的执行者。为控制查询消息的传输,通过 TTL(Time To Live)减值实现。

全分布非结构化的 P2P 网络的优点:有较好的容错能力,能快速发现目的节点,可以支持复杂查询,由于不再使用中心服务器,各个节点之间是对等的,网络不会因为某一个节点的故障而整个瘫痪。

全分布非结构化的 P2P 网络的缺点如下。

(1) 随着联网节点的不断增多,网络规模不断扩大,由于每次搜索都是以广播方式进行,产生的网络流量剧增,从而导致搜索速度下降,排队响应时间长,网络中部分低带宽节点因网络资源过载而失效。

(2) 全分布非结构化网络的查询访问只能在网络的很小一部分进行,因此网络的可扩展性不好。

(3) 由于没有确定拓扑结构的支持,非结构化网络无法保证资源发现的效率。即使需要查找的目的节点存在,查找也有可能失败。

(4) 全分布非结构化 P2P 一般采用泛洪、随机漫步或有选择转发算法,因此直径不可控,可扩展性较差。

3. 全分布结构化 P2P

由于非结构化系统中随机搜索可扩展性不强,如今大量的研究集中在如何构造一个高度结构化的网络拓扑系统。目前研究的重点是如何有效地查找信息,最新成果都是基于分布式散列表(Distributed Hash Tables,DHT)的发现和路由算法方面。全分布结构化 P2P 采用 DHT 技术,由大量节点共同维护散列表,散列表分割成不连续的块,每个节点分配一个属于自己的块并管理该块。DHT 按照一定的方式为网络节点分配一个全局唯一的节点标识符(Node ID),资源对象通过散列运算产生一个唯一的资源标识符(Object ID),且将该资源存储在与之相等或者相近节点上。需要查找该资源时,采用同样的方法可定位到存储