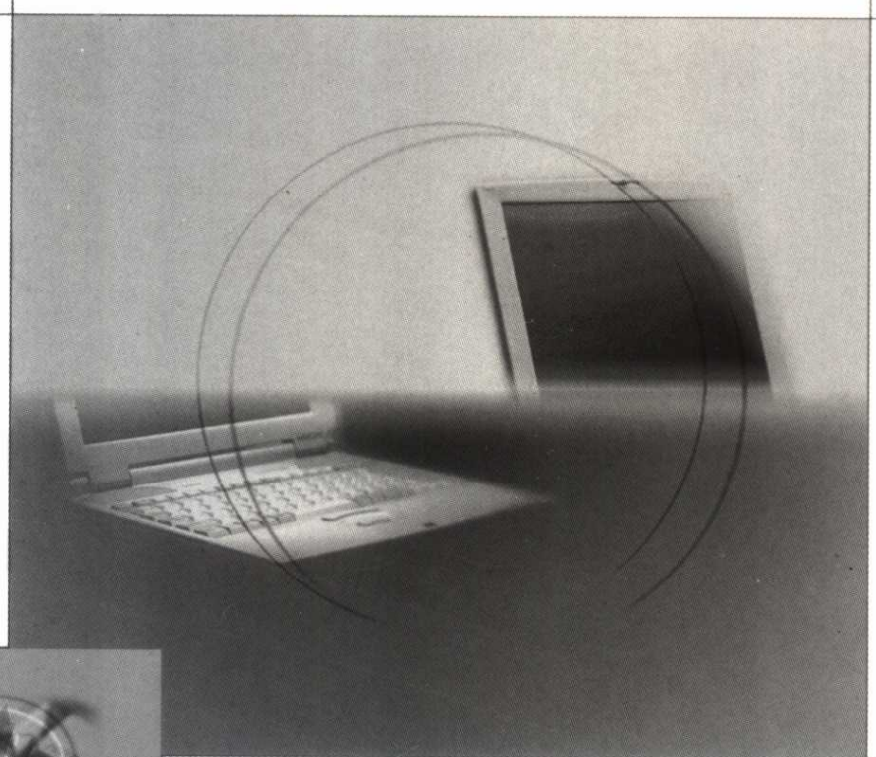
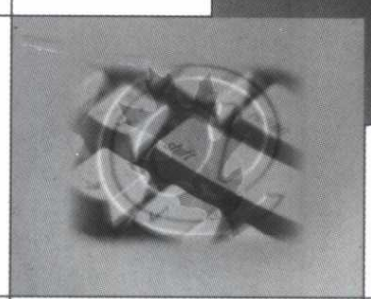
国家自然科学基金资助项目

网络链接分析与网站评价研究

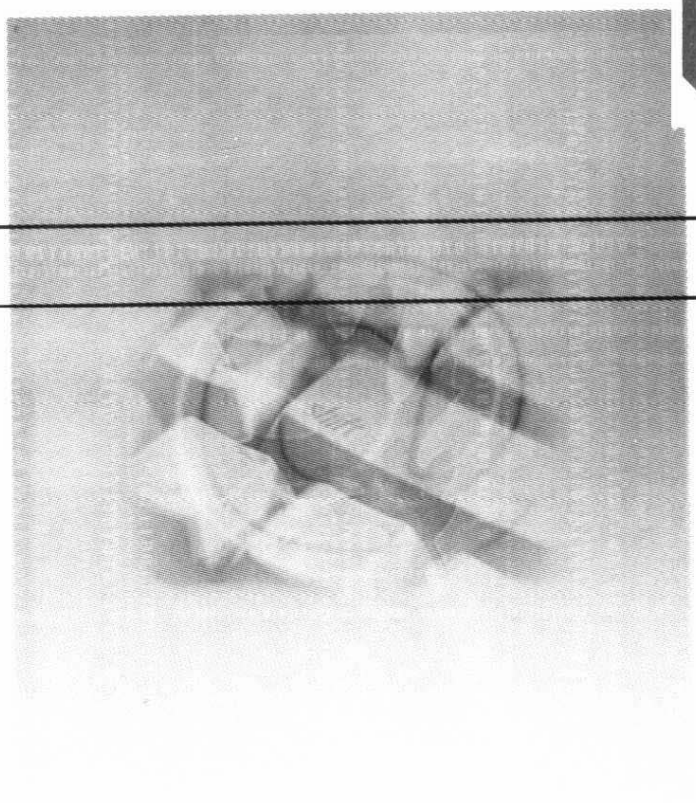




段宇峰 著

北京图书馆出版社

国家自然科学基金资助项目



网络链接分析与网站评价研究

段宇峰 著

北京图书馆出版社

图书在版编目(CIP)数据

网络链接分析与网站评价研究/段宇锋著. —北京:北京图书馆出版社,2005.6

ISBN 7-5013-2762-9

Ⅰ.网… Ⅱ.段… Ⅲ.①计算机网络—链接技术—研究②网站—评价—研究 Ⅳ.TP393.0

中国版本图书馆 CIP 数据核字(2005)第 023270 号

书名 网络链接分析与网站评价研究
著者 段宇锋 著

出版 北京图书馆出版社 (100034 北京西城区文津街 7 号)
发行 010-66139745 66175620 66126153
66174391(传真) 66126156(门市部)
E-mail cbs@nlc.gov.cn(投稿) btsfxb@nlc.gov.cn(邮购)
Website www.nlcpress.com
经销 新华书店
印刷 北京华正印刷厂

开本 880×1230 毫米 1/32
印张 9.75
版次 2005 年 6 月第 1 版 2005 年 6 月第 1 次印刷
字数 230(千字)

书号 ISBN 7-5013-2762-9/G·611
定价 25.00 元

前　言

20世纪90年代以来，网络已经成为推动科技、经济和社会发展的重要因素，网络化成为当今社会的显著特征。随着网络的日益普及，加强网络管理已经成为当务之急，而实施定量化管理则是其主要途径之一。在网络管理、网络经济和网络社会需求的推动下，“网络信息计量学(Webmetrics)”便应运而生。网络信息计量学是由网络技术、网络管理、信息资源管理与信息计量学等相互结合、交叉渗透而形成的一门交叉性边缘学科，也是信息计量学的一个新的发展方向和重要研究领域。其根本目的主要是通过对网上信息的计量研究，揭示其新的特征、数量关系和内在规律，为网上信息的有序化组织和合理分布，为网络信息资源的优化配置和有效利用，为网络管理的规范化和科学化提供必要的理论支持和定量依据，从而改善网络的组织管理和信息资源管理，提高其管理水平，促进其经济效益和社会效益的充分发挥，推动社会经济信息化、网络化的健康发展。

1997年，丹麦学者阿曼德(T. C. Almind)等人首次提出了“Webmetrics”这一概念，认为文献计量学的各种方法完全可以用于互联网的信息计量分析中。此后，许多研究者在不同方面对网络信息计量学问题进行了探讨。例如，Rousseau、Mckiernan根据文献计量学引文(Citation)的含义，提出了“Sitation”的概念，对网页的引用行为进行分析；Website. net仿照《科学引文索引》(SCI)的做法编制了一个“网络引证分析索引”(Web Citation Index，WCI)，用来统计分析网页的引用情况，研究网页链接之间的关系和规律，监视网页链接的变化情况等；Ingwersen提出可以把文献计量学的期刊影响因子应用到网页的评价中去，提出了“网络影响因子”(Web Impact Factor，WIF)的概念，可以用来分析一定时期内相对关注的网页平均被引情

况;Brin 等提出了“Pagerank”算法,根据一个网页链接其他网页的数量和质量来判断一个网页的质量和权威性;英国南安普顿大学的“开发期刊计划”(Open Journal Project)开发了一个网页自动链接工具,根据语义的相似性定量分析,将电子期刊有关的内容和有关的网页进行自动链接,并可以对有关文章的引证关系进行定量研究;2001年,美国伯克利加州大学信息管理与系统学院对网络各类型的信息依据一定的指标进行计量,分析其产生、分布、增长和过载等的原因和规律。可以说,自网络信息计量学产生之日起,网络链接就是其最重要的研究内容,迄今为止,几乎所有的研究都离不开对网络链接的分析。因此笔者认为,网络链接分析是网络信息资源管理的核心,网络链接研究的进展将直接推动网络的发展并使其充分发挥所蕴含的巨大潜力。这既是本研究的出发点,也是所期望达到的最终目标。

本着理论探讨与实证研究相结合的思路,笔者在分析和借鉴国内外已有研究的基础上,分九章,从3个方面对网络链接分析及其在网站评价中的应用进行了较为全面的探讨。第一部分理论篇在介绍网络链接概念的基础上,阐述了国内外网络链接研究进展、研究方法和研究工具;第二部分探索篇以美国著名商学院和医学院网站为样本,采用理论与实证研究相结合的方式,深入探讨了网站的链接特征、核心网站的确定方法以及网站中网络信息资源的分布规律;第三部分应用篇是在前两部分研究成果的基础上,发掘网络链接分析在大学评价和网站评价中的应用价值。

网络链接分析与网站评价研究是网络信息计量学和网络信息资源评价研究的组成部分,既是国际学术界的新兴研究领域,又对网络管理和产业发展具有重要的指导意义。尽管笔者在成书的过程中付出了艰苦的努力,但水平有限,不足之处在所难免,敬请专家与广大读者批评指正。笔者希望通过本书加强与同行之间的交流,携手共进,推动网络信息资源管理研究的不断深入。

目　录

前　言…… 1

理 论 篇

1　网络与网络链接 …… 3
1.1　网络 …… 3
1.2　网络链接的概念…… 48
2　网络链接研究的现状及趋势…… 70
2.1　网络链接研究的意义…… 70
2.2　网络链接研究的主要进展…… 73
2.3　网络链接研究目前存在的主要问题…… 87
2.4　网络链接研究的走向和趋势…… 95
3　网络链接研究方法…… 97
3.1　网络链接研究方法的建立原则和体系…… 97
3.2　网络链接研究的一般步骤 …… 100
4　网络链接研究工具 …… 126
4.1　抽样工具 …… 126
4.2　原始数据的获取工具 …… 129
4.3　网络链接解析工具 …… 137
4.4　统计分析工具 …… 140
4.5　搜索引擎 …… 144

探 索 篇

5　网站链接特征研究 …… 153
5.1　网站链接特征指标体系 …… 153

5.2 网站链接特征研究 …… 158
6 核心网站研究 …… 192
6.1 核心网站的测定方法 …… 192
6.2 核心网站的测定 …… 194
6.3 出链所指向网站被链接次数的频数分布规律 …… 202
6.4 研究结论 …… 206
7 网站分层研究 …… 207
7.1 网站分层的概念 …… 207
7.2 研究方法 …… 208
7.3 网站分层研究 …… 211

应 用 篇

8 链接分析在大学评价中的应用 …… 229
8.1 研究内容与方法 …… 229
8.2 数据分析及结果 …… 236
8.3 研究结论 …… 240
9 中、美大学网站链接特征的比较研究 …… 241
9.1 研究内容与方法 …… 242
9.2 数据分析及结果 …… 249
9.3 研究结论 …… 253
参考文献 …… 255
附录 WEBSTAT 的源程序 …… 263
致 谢 …… 306

理 论 篇

本篇是网络链接研究的基础。第一章旨在阐述网络链接的基本概念，在此基础上，第二章全面介绍了网络链接研究的现状、问题和发展趋势。针对现存的难点和问题，第三章和第四章讨论网络链接研究方法和工具，为进一步的理论探讨和实践应用提供依据和手段。

1 网络与网络链接

1.1 网络

1.1.1 计算机网络的概念和类型

1.1.1.1 计算机网络的概念

计算机网络是现代通信技术与计算机技术相结合的产物。一方面,通信技术为计算机之间的数据传递和交换提供了必要的手段;另一方面,数字计算技术也极大地改善了通信网络的性能。自上个世纪60年代迄今,现代计算机网络已经有40年的历史,特别是90年代中后期Internet的快速成长,使网络渗透到社会生活的各个环节。尽管网络在理论、技术和应用领域已经取得重大进展,但对于什么是"计算机网络"却一直没有统一的概念。

计算机网络是"一些互相连接的、自治的计算机的集合",这是对计算机网络最简单的定义。张公忠在其主编的《现代网络技术》第2版中指出,"计算机网络是一种地理上分散的,具有独立功能的多台计算机通过通信设备和线路连接起来,在配有相应的网络软件的情况下实现资源共享的系统"①;张国鸣则认为,"计算机网络就是把分布在不同地理区域的计算机与专用外部设备用通信线路互联成

① 张公忠主编. 现代网络技术教程(第2版). 北京:电子工业出版社,2004

一个规模大、功能强的计算机应用系统，从而使众多的计算机可以方便地互相传递信息，共享硬件、软件、数据信息等资源”。[①] 尽管对计算机网络定义的表述各不相同，但必然涉及以下4个方面：

① 至少包括两台计算机或外部设备；

② 通信设备和线路介质；

③ 网络软件；

④ 目的是为了实现在硬件、软件和数据资源等方面的共享。

1.1.1.2 计算机网络的类型

依据不同的标准，可以将计算机网络划分成多种类型。常见的划分方式包括以下几种：

(1) 按网络的作用范围分类

① 局域网(Local Area Network，简称 LAN)：指在较小的地理范围内(一般小于10km)由计算机、通信线路和网络连接设备组成的网络。在网络发展的初期，通常一个学校或企业只拥有一个局域网，但目前都拥有许多个局域网，因而又出现了诸如校园网、企业网等。

② 城域网(Netropolitan Area Network，简称 MAN)：指在一个城市范围内(一般小于100km)由计算机、通信线路和网络连接设备组成的网络。在技术和体系结构上，MAN 与 LAN 比较相似，可以将其看成一种大型的 LAN；从范围上来说，又非常类似较小的广域网。

③ 广域网(Wide Area Network，简称 WAN)：作用范围大的网络，通常在数十至数千公里的范围。广域网是互联网的核心部分，国际互联网也是世界上最大的广域网。

(2) 按网络的通信方式分类

① 广播式网络(Broadcast Network)：网络上的所有计算机共享一条通信信道。网上的分组或包(Packet)可以被任何机器发送并被

① 张国鸣主编. 网络管理员教程. 北京：清华大学出版社，2004

其他所有的机器接收。分组的地址字段指明此分组应被哪台机器接受。一旦收到分组,各机器将检查它的地址字段,如果是发送给它的,则处理该分组,否则将它丢弃。广播式系统通常也允许在它的地址字段中使用一段特殊的代码,以便将分组发送到所有的目标。使用此代码的分组发出以后,网络上的每一台机器都会接收它。这种操作被称为广播(Broadcasting)。

② 点到点网络(Point-to-Point Network):由一对机器之间的多条连接构成。为了能从源到达目的地,这种网络上的分组可能必须通过一台或多台中间机器。通常是多条路径,并且长度可能不一样,因此在点到点的网络中路由算法显得特别重要。

(3) 按网络的使用者分类

① 公用网(Public Network):也称为公众网,是指国家的电信公司出资建造的大型网络。“公用”的意思是所有愿意按电信公司的规定缴纳费用的人都可以使用。

② 专用网(Private Network):指某个部门为本单位的特殊业务工作的需要而建造的网络,它不向本单位以外的人提供服务。

(4) 按网络的交换功能分类

“交换”就是按照某种方式动态地分配传输线路的资源。常用的交换技术有电路交换、报文交换和分组交换,采用相应技术的网络也就被称为电路交换网、报文交换网和分组交换网。

① 电路交换(Circuit Switching):一种直接的交换方式,它为一对需要进行通信的装置(站)提供一条临时的专用通道,即提供一条专用的传输通道,既可是物理通道又可是逻辑通道(使用时分或频分复用技术)。这条通道是由节点内部电路对节点间传输路径经过适当选择、连接而完成的,是由多个节点和多条节点间传输路径组成的链路。由电路交换的通信包括电路建立(通过源站点请求完成交换网中对应的所需逐个节点的接续过程,以建立起一条由源站点到目的站的传输通道)、数据传输、电路拆除(在完成数据或信号的传

输后，由源站点或目的站提出终止通信，各节点相应拆除该电路的对应连接，释放由该电路占用的节点和信道资源）3 个阶段。电路交换具有以下特点：呼叫建立时间长，并且存在呼损；电路连通后提供给用户的是“透明通路”，但通信双方的收发速度、编码方法、信息格式、传输控制等必须一致；一旦电路建立，数据以固定的数据率传输，除通过传输链路的传播延迟外，没有别的延迟，在每个节点的延迟是可以忽略的，是用于实时大批量连续的数据传输；在整个传输过程中，信道是专用的，再加上通信建立时间、拆除时间和呼损，线路的利用率较低。

② 报文交换（Message Switching）：又称为存储转发（Store and Forward）。在报文交换网中，网络节点通常为一台专用机算计，带有足够的外存，以便在报文进入时进行缓冲存储。节点接收一个报文之后，报文暂存放在节点的存储设备之中，等输出线路空闲时，再根据报文中所附的目的地址转发到下一个合适的节点，如此往复，直到报文到达目标数据终端。在报文交换中，每一个报文由传输的数据和报头组成，报头中有源地址和目标地址。节点根据报头中的目标地址为报文进行路径选择，并且对收发的报文进行相应的处理，如差错检查和纠错、调节输入/输出速度进行数据速率转换、进行流量控制，甚至可以进行编码方式的转换等，所以报文交换是在两个节点间的一段链路上逐段传输，不需要在两个主机间建立多个节点组成的电路通道。报文交换的特点包括：源站与目标站在通信时不需要建立一条专用的通路；没有建立和拆除线路的等待和时延；线路利用率高，节点间可根据线路情况选择不同的速度传输，能高效地传输数据；要求节点具备足够的报文数据存放能力，一般节点由微机或小型机担当；数据传输的可靠性高，每个节点在存储转发中都进行差错控制。但是，由于采用了对完整报文的存储/转发，节点存储/转发的时延较大，不适用于交互式通信。

③ 分组交换（Packet Switching）：是以报文分组（Packet）为单位

进行交换传输,仍属于“存储/转发”交换方式。分组是一组包含数据和呼叫控制信号的二进制数,把它作为一个整体加以转接,这些数据、呼叫控制信号以及可能附加的差错控制信息是按规定的格式排列的。分组交换又可以分为数据报(Datagram)传输分组交换和虚电路(Virtual Circuit)传输分组交换。

数据报传输分组交换网是把进网的所有分组都当作单独的“小报文”来处理,作为基本传输单位的“小报文”被称为数据报。数据报都带有地址和分组序列,虽然它们不一定经过同一条路径,但最终都能到达同一目的节点,在此对数据报进行排序和重装。

虚电路传输分组交换是两个用户的终端设备在开始互相发送和接收数据之前通过通信网络建立逻辑上的连接,所有分组均沿着条连接传输,直至用户不需要发送和接收数据时清除这种连接。与电路交换相比,虚电路传输分组交换并不意味着实体间存在像电路交换方式那样的专用线路,而是选定了特定路径进行传输,分组所途经的所有节点都对这些分组进行存储/转发。

数据报和虚电路方式相比,虚电路方式对数据量较大的通信传输率高,分组传输时延短,且不容易产生数据分组丢失,但它对网络依赖性大;数据报方式对短报文数据通信传输比较合适,对网络故障的适应能力强,但时延大。

④ 混合交换:是在一个数据网中同时采用电路交换和分组交换。

(5) 其他类型

计算机网络依据网络的拓扑结构可分为总线型网络、星型网络、环型网络等;按照传输的信道可以划分为模拟信道网络和数字信道网络;按照通信传输介质可以分为双绞线网、同轴电缆网、光纤网、微波网、卫星网、红外线网等;按照信号频带占用方式又可分为基带网和宽带网。

1.1.2 客户和服务器①②

在计算机网络的术语中,与互联网相连的计算机由于位于网络的边缘,因而,通常被称为端系统(end systems)。因为它们容纳(即运行)诸如 Web 浏览器程序、Web 服务器程序、电子邮件阅读程序或电子邮件程序等应用程序,也被称为主机(host)。主机又被划分为客户机和服务器,客户和服务器都是指通信中所涉及的两个应用进程。

计算机的进程(Process)就是运行着的计算机程序。在网络环境下,许多问题的解决往往是通过位于不同主机中的多个进程之间的通信网络的通信和协同工作来完成的。这些为了解决具体的应用问题而彼此通信的进程就称为"应用进程"。而应用层的具体内容就是规定应用进程在通信时所遵循的协议。TCP/IP 的应用层协议使用客户(Client) - 服务器(Server)方式使两个应用进程能够进行通信。

客户 - 服务器方式所描述的是进程之间服务和被服务的关系。当 A 进程需要 B 进程的服务时就主动呼叫 B 进程,在这种情况下,A 是客户而 B 是服务器。可能在下一次通信中,B 需要 A 的服务,此时,B 就是客户而 A 就是服务器。这里最主要的特征就是:客户是服务请求方,服务器是服务提供方。在实际应用中,客户软件和服务器软件通常还具有以下一些主要特点:

客户软件:

① 在进行通信时临时成为客户,但它也可在本地进行其他的计算;

---

① James F. Kurose, Keith W. Ross 著, 陈鸣等译. 计算机网络——用自顶向下方法描述因特网特色(第二版). 北京: 人民邮电出版社, 2004

② 谢希仁编著. 计算机网络(第 4 版). 北京: 电子工业出版社, 2003

② 被用户调用并在用户的计算机上运行,在打算通信时主动向远地服务器发起通信;

③ 可与多个服务器进行通信;

④ 不需要特殊的硬件和很复杂的操作系统。

服务器软件:

① 是一种专门用来提供某种服务的程序,可同时处理多个远地或本地客户的请求;

② 在共享计算机上运行,当系统启动时即自动调用并一直不断地运行着;

③ 被动地等待并接受来自多个客户的通信请求;

④ 一般需要强大的硬件和高级的操作系统支持。

客户与服务器的通信关系一旦建立,通信就可是双向的,客户和服务器都可发送和接收信息。功能较强的计算机可同时运行多个服务器进程。客户和服务器从严格意义上来说都指的是进程,即计算机软件。但是,由于运行服务器进程的机器往往有许多特殊的要求,因此,人们经常将主要运行服务器进程的机器(硬件)不严格地称为服务器。于是,服务器一词有时指的是软件,但有时指的是硬件。

1.1.3 网络协议

网络协议(network protocol),简称协议,是为进行网络中的数据交换而建立的规则、标准或约定,它规定了网络数据的交换格式以及有关的同步问题。

网络协议主要由 3 个要素组成:

① 语法,即数据与控制信息的结构或格式;

② 语义,即需要发出何种控制信息,完成何种动作以及作出何种响应;

③ 同步,即事件实现顺序的详细说明。

1.1.3.1 网络的体系结构

(1) 网络的体系结构

所谓网络的体系结构(Architecture)就是计算机网络各层及其协议的集合。“分层”可将复杂的问题转化为若干个较小的局部问题,从而将计算机网络分解为若干个容易处理的子系统,使每一层实现一种相对独立的功能。早在 ARPANET 的设计时就提出了分层的方法,并在 1974 年 IBM 研制的系统网络体系结构(System Network Architecture,SNA)中得到充分体现。

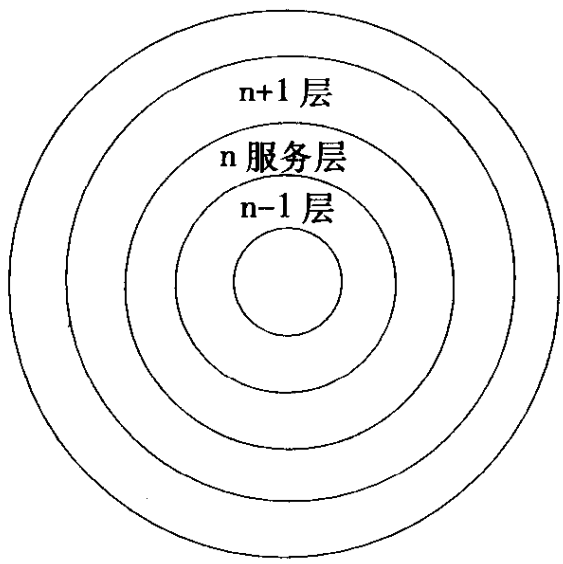


图 1-1 层次模型

图 1 - 1 所示的一般分层结构中,n 层是 n - 1 层的用户,又是 n + 1层的服务提供者。n + 1 层虽然只直接使用了 n 层提供的服务,实际上它通过 n 层还间接地使用了 n - 1 层以及以下所有各层的服务。

层次结构一般以垂直分层模型来表示(图 1 - 2)。从该模型我们可以看到,除了在物理媒体上进行的是实通信之外,其余各对等实体间进行的都是虚通信;n 层的虚通信是通过 n/n - 1 层间接口处 n - 1层提供的服务以及 n - 1 层的通信(通常也是虚通信)来实现的,并且,对等层的虚通信必须遵循该层的协议。