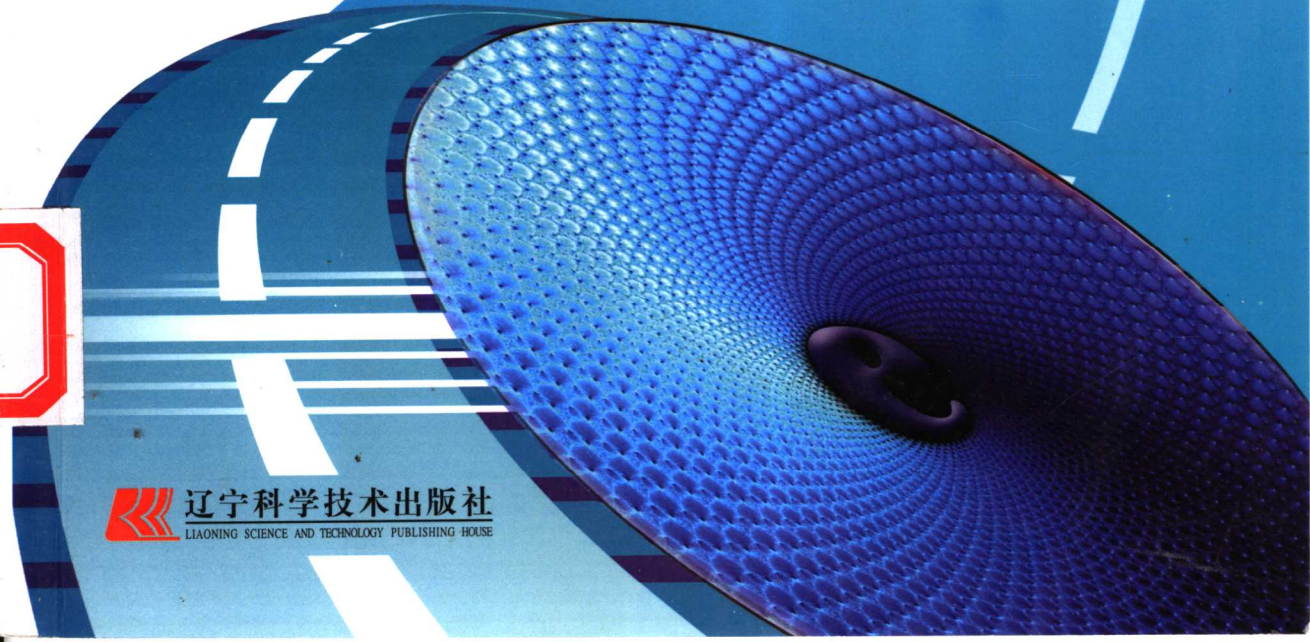
Internet信息检索系列

武汉大学信息管理学院
武汉大学信息资源研究中心
审定并推荐

看不见的网站——Internet专业信息检索指南

(美) C·谢尔曼　G·普赖斯　著
马费成　蔡东宏　李　勇　李　莹　译



辽宁科学技术出版社
LIAONING SCIENCE AND TECHNOLOGY PUBLISHING HOUSE

武汉大学信息管理学院
武汉大学信息资源研究中心 审定并推荐

Internet信息检索系列

看不见的网站——Internet专业信息检索指南

(美) C·谢尔曼　G·普赖斯 著
马费成　蔡东宏　李勇　李莹 译

辽宁科学技术出版社
沈阳

目 录
CONTENTS

第一章

Internet 与看不见的网站

为了广泛而深刻地了解网络，充分分享我和我的同事所持有的观点，必须了解网络是如何产生的。

——Tim Berners - Lee

大多数人趋向可互换地使用“Internet”和 Web(网络)，但是，它们不是同义词。Internet 是一个联网协议(一套规则)，允许各种类型的计算机在 Internet 上相互连接和通讯。Internet 的起源可追溯至研究人员和国防部成员分享信息的一种工具——由美国国防部高级研究机构(DARPA)于 1969 年发起的一项计划(Kahn, 2000)。

而 Web 则是一种在 Internet 上运作的软件协议，允许用户方便访问贮藏在 Internet 计算机上的文件。Web 是由 Tim Berners - Lee 于1990 年创建的。他是欧洲核研究组织的一名计算机程序师。在 Web 产生之前，访问 Internet 文件是一项挑战性的工作，需要专业化知识和技能。由于 Web 的出现，只要通过点击某一超文本链接这一简单机制，就能方便地在 Web 上检索各种不同类型的文件，包括文本、图像、声音、音像。

定 义

超文本

一个允许计算机化的对象（文本、图像、声音等）被链接在一起的系统。一个超文本链接指向某一具体对象或文本位置；点击该链接即可打开与该对象有关的文件。

这本书的核心就是 Web(网络)——更具体地说，是在搜索引擎

看不到的网络部分。为了充分地了解被称作看不见的网站的现象，重要的是要首先了解 Internet 和 Web 之间的基本差异。

该章将追踪一些早期 Internet 检索工具的发展，展示它们的局限性是怎样推动网络最终被普遍认同的。这一历史背景，虽然本身就颇具吸引力，却为理解为什么看不见的网站能够先出现奠定了基础。

Internet 的由来

在 20 世纪 60 年代中期之前，大多数计算机还都是独立工作的机器，同别的计算机互不连接或通信。1962 年麻省理工学院教授 J. C. R. Licklider 写了一篇文章，展望了连通全球计算机的“银河系网络”（Leiner 2000）。这种想法在当时很超前，但却引起 Larry Roberts——美国国防部高级研究工程代理处（ARPA）的工程经理的注意。1966 年 Roberts 向 ARPA 提交了一份建议，准备将代理处为数众多无联系的计算机连接在类似于 Licklider 的银河系网络的一个网络上。

Roberts 的建议被采纳，开始着手研究“ARPANET”，它最终成为我们现在所知道的 Internet。ARPANET 的最早节点是在 1969 年安装在洛杉矶加里福尼亚大学(UCLA)。逐渐地，整个 20 年纪 70 年代从事 ARPA 计划的大学和国防部承包者都开始与 ARPANET 连接。

在 1973 年，ARPA 启动了另一研究计划，让联网计算机通过多重链接透明地通信。ARPANET 仅仅是一个网络，而新的计划则是“网络的网络”。被公认为 Internet 之父的 Vint Cerf 认为：“这叫 Internet 计划，从这一研究计划涌现出来的网络系统就是人们所知的‘Internet’”（Cerf 2000）。

直到 20 年纪 80 年代中期，随着个人计算机使用的同步发展和被称作“传输控制协议/Internet 协议（TCP/IP）”的全球 Internet 通讯标准的被普遍采用，Internet 被希望与它连接的人所普遍接受。其他政府机构也纷纷建立专门承载 Internet 通信的基干网络，从而进一步促进了 Internet 的发展。到 20 世纪 80 年代后期，Internet 从最初的为数不多的由计算机组成的网络发展成为被世界各地政府和商业企业支持的一个发展迅猛的通信网络。

尽管提高了可检索性，但直到 20 世纪 90 年代早期，Internet

仍主要是一个服务于科研机构和政府承包者的工具。随着越来越多的计算机与 Internet 连接，用户开始需要能够在网络上检索和查找任何计算机上的文本和其他文件的工具。

早期网络检索工具

虽然复杂的搜索和信息检索技术可回溯至 20 世纪 50 年代晚期 60 年代早期，但这些技术主要被使用于封闭或具有独占权的系统。早期的 Internet 搜索和检索工具甚至缺乏最基本能力，这主要是因为人们认为传统的情报检索技术不会在一个开放、非结构化的信息世界如 Internet 上正常工作。

检索 Internet 上的一个文件，其过程分为两个部分。首先，需要使用被称作远程通信网(Telnet)的终端仿真程序，直接连接文本所在的那个远程计算机。然后，使用被称作文件传送协议（FTP）客户机的另一程序取来文件。多年来，为了检索文件，都必须知道计算机地址和正在查询的文件的准确地址和名称——过去没有像我们今天所熟悉的搜索引擎或其他查找文件的工具。

定　义

文件传送协议（FTP）

一套有关传送和接收与 Internet 相连接的计算机之间的各种类型文件的规则。

远程通信网

一套在你的计算机上运行的终端仿真程序，允许你通过一个传输控制协议/Internet 协议（TCP/IP）网络检索一个远程计算机、并在那个计算机上执行命令，就好像你直接与它连接一样。许多图书馆提供远程通信网(Telnet)检索其书目。

因此，“检索”总是意味着向电子邮件消息列表或讨论会发出帮助请求，并希望某个好心人对你所需要的详细资料做出反应，拿来你在查找的文件。随着“匿名”FTP 服务器的引入，情况有所改进。这种服务器乃是专题式文件服务器，专门用于共享文件。服务器是匿名的，因为它们不加设密码保护，任何人都能够进入系统或索取系统中的任何文件。

在 FTP 服务器上的文件都是按等级式目录排列的，这很像现

在的个人计算机系统文件按分层组织的文件夹。等级结构便于 FTP 服务器展示贮藏在服务器上所有文件的目录列表，但却需对 FTP 服务器的内容有比较深的了解。如果你在查找的文件不在你所进入的 FTP 服务器中，则说明你的运气不好。最早的用于检索贮藏在 FTP 服务器上的文件的真正的检索工具称作 Archie，它是 1990 年由位于蒙特利尔的 McGill 大学的系统管理人员和研究生组成的小组研制的。Archie 是当代搜索引擎的原形，但比较而言，还是很原始和有局限性的。Archie 漫游 Internet，搜索匿名服务器上可获得的文件，下载在每个匿名服务器上可找到的目录列表。这些列表贮藏在一个中央可搜索的数据库——McGill 大学 Internet 档案文件数据库上，并且每月更新一次。

尽管它代表了一个主要的前进方向，Archie 数据库还是极端原始的，只能检索某一特殊文件名或执行具体功能的计算机程序。尽管如此，Archie 还是盛行一时，在 20 世纪 90 年代早期，据 McGill 大学 Archie 小组组长 Peter Deutsch 称，约 50% 的蒙特利尔的 Internet 通信都与 Archie 有关。

“在 Archie 进入应用以后的较短时间，基于 Internet 的研究项目应运而生，它们包括 WWW、Gopher、WAIS 等等”（Deutsch，2000）。”

“每个人都探索不同领域的 Internet 信息问题空间，对怎样组织和部署基于 Internet 的服务都有独到的见解。” Deutsch 写到。该小组授权其他人使用 Archie，1992 年最早的影子网站在澳洲和芬兰启动。Archie 网络到 1995 年发展到顶峰，当时世界上有 63 个计算站。

Gopher（地鼠），Archie 的替代，由 Minnesota 大学的 Mark McCahill 和他的小组于 1991 年研制成功，以 Minnesota 大学的吉祥物“金地鼠”命名。

实际上，Gopher 合并了 Telnet 和 FTP 协议，允许用户点击超链接菜单检索所需信息，而不用借助于附加指令。使用一系列菜单，用户就可以逐级搜索更具体的类别，最终能够检索文献全文、图表甚至音乐文件，虽然在单一格式方面仍不完整。Gopher 的出现，使浏览 Internet 信息变得更加方便。

据 Gopher 的研制者 McCahill 称：“在 Gopher 之前，没有办法使用大的分布式系统，从一个机器上到另一个机器上的资料之间的指针没有缝隙。你必须知道这个机器的名称，而且如果你想越过这

里，也必须知道它的名称。

Gopher 为你保管所有资料。因此，航游 Gopher 极为容易。它是典型的点和爪，任何人都可用它来查询信息资料；它也很容易提供信息。所以，许多人开始自己经营服务器。它使用便利、不混乱、不忙乱，是第一台被人们到处寻找的查找信息的检索工具。它是个不是为技术而编写的工具。”

Gopher 的“不混乱，不忙乱”的接口是后来发展成知名网站目录如 Yahoo 的先驱！尤其在你设置接口时，你就能了解一组信息的一般结构并对其有一个总的看法，选择你感兴趣的类目，进入一个更加专指的领域，然后可以采用浏览某些文件的方法抑或采用提交检索式的方法检索某个类目。” McCahill 这样说道。

Gopher 存在的一个问题是用它提供在某一具体地点——如在 Minnesota 大学计算机上可获得的文件列表。当 Gopher 服务器搜索时，能搜索所有其他计算机中的专题式目录，而这些计算机都使用 Gopher 并接入 Internet 站，或所谓的“地鼠空间(Gopherspace)”。1992 年 11 月，在内华达大学系统计算服务组的 Fred Barrie 和 Steven Foster 解决了这个问题，研制出一套称作“Veronica(维朗妮卡)”的程序，它是一种检索 Gopher 的文件、像 Archie 一样的专题式检索工具。1993 年，另一称作“Jughead”的程序在 Gopher 的搜索上增加了关键词检索和布尔逻辑运算符的能力。

定 义

关键词

在提问框内输入词或词组，检索系统可将其与数据库里的文本文件匹配。

布尔逻辑

布尔逻辑运算符系统（AND、OR、NOT）对检索提问式执行“真假”运算、与关键词一起使用可扩检和缩检结果。

据通俗传奇，Archie、Veronica 和 Jughead 都是以卡通人物命名的。事实上，Archie 是 Archives 的缩写。Veronica 可能是以卡通人物命名的(她是 Archie 的女友)，是“Verg Easy Rodent-Oriented Net-Wide Index to Computerized Archires”(很容易以鼠标定向的计

算机化档案的网络索引)的正式首字母缩写词。而 Jughead(Archie 和 Veronica 的动画伙伴)是“ Jonzy's Universal Gopher Hierarchy Excavation and Display”(Jonzy 全球 Gopher 分级发掘和展示)的首字母简略词。而 Rhett“Jonzy ” Jones 则是其发明人，他在犹他大学计算机中心开发出该程序。

荒诞的说法：Web 和 Internet 是一样的

Internet 是世界上最大的计算机网络，由数百万计算机组成。它只不过是管道，能让各种各样的信息在世界各地的计算机上流动。

Web 是 Internet 上的许多接口中的一个，能检索来自计算机上的文本、图像和多媒体文件，而不必知道复杂指令。只要点击链接，就能神奇般地看到浏览器屏幕上展示的页面。

网络(Web)历史仅约 10 年，而 Internet 已有 30 年的历史了。在网络诞生之前，计算机之间就能通信，但是，其接口没有像在网络上使用起来那么灵巧和方便。这些更老的接口或“协议”还存在并提供许多不同的独特方式与其他计算机或人进行通信。

其他 Internet 协议或接口包括：

- 电子邮件
- 论坛和公告板
- Internet 邮件列表
- 新闻组
- 对等文件分享系统，如 Napster 和 GnutelIa
- 通过网络接口检索的数据库

正如人们所知，Internet 比 Web 内容丰富得多。事实上，上面最后一项——通过网络接口检索的数据库，是看不见的网站的主要组成部分。后面的章节将深入探讨迷人而又极其有用的网络可检索数据库世界。

在这个时间的前后，创建了第三大检索协议，就是广域信息服务系统（WAIS）。它是由研制思维机器的 Breuster Kahle 和他的同事创建的。WAIS 的作用如同现在的元搜索引擎。WAIS 客户机驻留于本地的计算机上，允许你运用自然语言而不是使用计算机指令检索别的 Internet 服务器上的信息。服务器本身负责解释提问式，并返回合适的结果，用户不必学习各个服务器的具体提问语言。

WAIS 使用了一个标准协议 Z39.50 扩充版。它在当时得到了

广泛使用。本质上，WAIS 只提供检索信息的单一计算机之间的协议。这种信息可以是文本、图像、声音或格式化文件。检索结果的质量取决于各个服务器怎样有效解释 WAIS 查询的直接结果。

所有的早期 Internet 检索协议都是对 Telnet 和 FTP 提供的蹩脚检索工具的巨大飞跃。尽管如此，它们仍须处理不连续数据对象的信息。这些协议缺乏结合不同类型信息如文本、声音、图像等的能力，不能形成概念链，因而不能将原始数据转换成有用的信息。虽然检索变得更加先进，但 Internet 上的信息仍未被公众所认识。20 世纪 80 年代末期，Internet 仍主要是科学家、大学教师、政府机构及其承包者的“活动场地”。

幸运的是，大约在相同的时间，一名瑞士的软件工程师编制了一个程序，最终产生了万维网。他将他的程序命名为“查询每件事情的相关信息(Enquire Within Upon Everything)”，它借用了“维多利亚女王时代的忠告”一书中的有关内容，即对每件事提供帮助信息——从去除污迹到投资。

查询每件事情的相关信息(Enquire Within Upon Everything)

> “我想，假如各地计算机贮藏的所有信息都相互连通，假使我能够给我的计算机编程序，创造一个空间让所有东西都能够相互连接，那么，欧洲核研究组织(CERN)甚至地球上的每个计算机上的所有信息对我和其他任何人都是可获得的。应该存在一个单一、全球化的信息空间。
>
> 一旦该空间上的一则信息被标上一个地址，我就能够让我的计算机得到它。能够同样方便地找到类似事情，计算机就能够表示好像不相关但事实上又有某种关系的事情之间的联系。信息网站就这样形成了。”
>
> ——Tim Berners - Lee

1990 年，Tim Berners - Lee 创建了网络，当时他在瑞士日内瓦的欧洲核研究组织(CERN)高能物理实验室担任合同程序设计员。网络是 Berners - Lee 承担的一个次要项目，但却帮助了他在礼仪上保持与 CERN 这样庞大的研究机构中的令人难以置信的各种不同的人、计算机、研究设备和其他资源的联系。CERN 科学家面临的主要

挑战之一正是它所具优势的这种差异性。实验室每年接待来自世界不同国家、说不同语言、从事各具特色的计算系统研究的人员达数千人。高能物理研究项目实验数据繁多，一套能够简化信息检索并能促进合作的程序简直是人们长期以来梦寐以求的东西。

Berners－Lee 在创立网络以前的差不多 10 年的时间里，一直在构思相对容易而又具分散链接能力的程序。他受到 Vannevar Bush 工作的影响，后者在第二次世界大战期间曾担任科学研究和发展办公室主任。在“诚如所思”这篇具里程碑意义的论文中，Bush 提出一个他称作 MEMEX 的系统——这是一种“用于个人贮藏自己的全部书籍、记录和信件的装置，通过设计，可以高速而灵活地查阅它。”（Bush，1945）

对贮藏在 MEMEX 中的材料当然可以标引，但 Bush 渴望超越简单的搜索和检索。当用户从一个文件转到另一个文件时，利用 MEMEX 可以建立概念“踪迹”，创造后来会检索到的 MEMEX 不同成分之间的持久联系。Bush 称其为相关标引。其基本概念，是一个条款，其中的任何选项可任意产生，从而自动选择另一个选项。这就是 MEMEX 的本质特征。将两个选项联系在一起的过程是一件重要的事情。

从 Bush 设想的著作中，我们能够看出现在被称作超文本的萌芽。但直到 1965 年，Ted Nelson 才实际描述了类似按 Bush 设想的方式运作的计算机化系统。Nelson 称他的系统为“超文本”，并在被他称之为 Xanadu 的系统中描述了下一代 MEMEX。

Nelson 的项目没有取得足够影响世界的能量。20 年后的 1985 年，Xerox 才完成了被称之为 Notecards 的最早的主流超文本程序。1 年以后，Owl 有限公司创建了一套“Guide”程序，其功能在很多方面类似现在的网络浏览器，但缺乏 Internet 连通性。

创建描写位图程序“MacPaint”的最知名的苹果计算机程序员 Bill Atkinson，于 1987 年创建了最早的真正意义上受欢迎的超文本程序。他的 Hypercard 程序专用于“苹果机”，也缺乏连通性。尽管如此，该程序还是受欢迎的。微软公司吸收了超文本的功能性和概念，并第一次出现在示窗软件的标准帮助系统中。

组织网络

在 Berners－Lee 开始他的网络组织以前，建立像万维网那样的