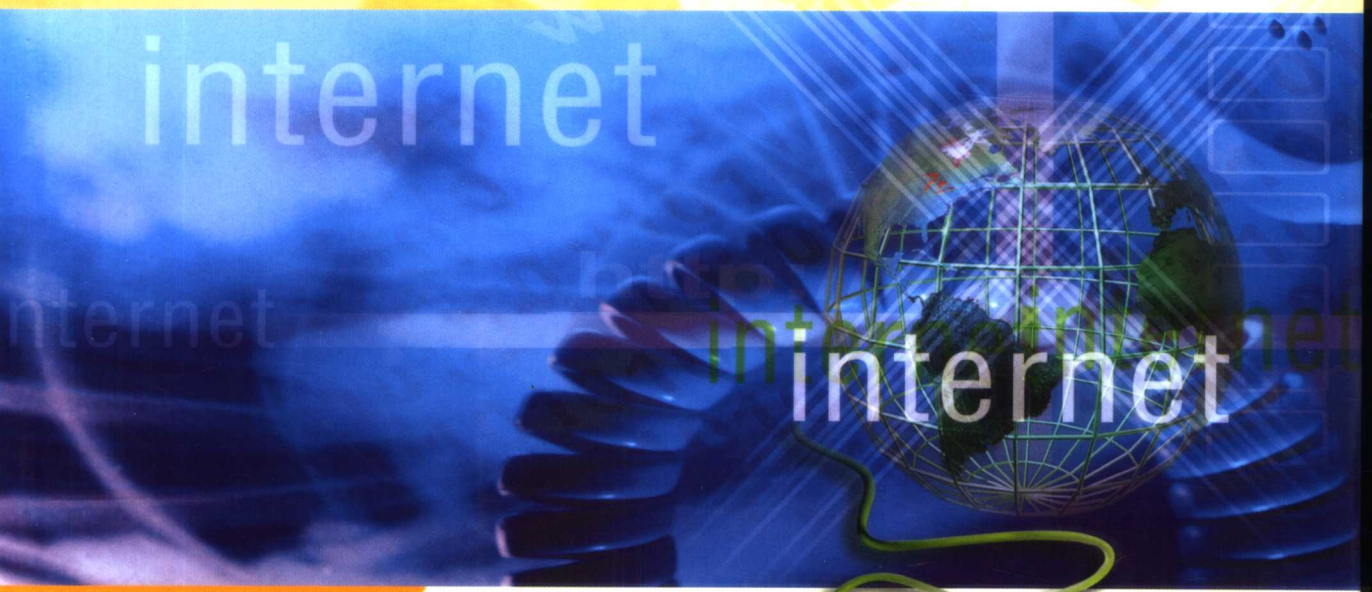
Internet信息搜索方法和技巧

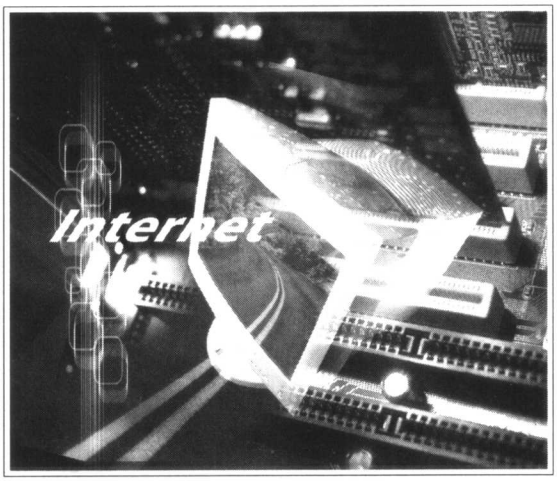


胡维治 主编

中国农业出版社

Internet
信息搜索方法和技巧

胡维治　主编



中国农业出版社

图书在版编目（CIP）数据

Internet 信息搜索方法和技巧/胡维治主编．—北京：中国农业出版社，2005.12
ISBN 7-109-10468-0

Ⅰ.Ⅰ… Ⅱ.胡… Ⅲ.①因特网—基本知识②网站—简介 Ⅳ.①TP393.4②TP393.092

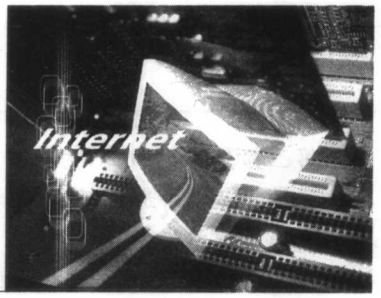
中国版本图书馆 CIP 数据核字（2005）第 131373 号

中国农业出版社出版
（北京市朝阳区农展馆北路 2 号）
（邮政编码 100026）
出版人：傅玉祥
责任编辑 杨天桥

中国农业出版社印刷厂印刷 新华书店北京发行所发行
2005 年 12 月第 1 版 2005 年 12 月北京第 1 次印刷

开本：787mm×1092mm 1/16 印张：18.25
字数：210 千字 印数：1～3 000 册
定价：30.00 元
（凡本版图书出现印刷、装订错误，请向出版社发行部调换）

参　编：李春子　马　力　熊　浪
杨　岿　叶颖泽



前言

21世纪的到来，意味着人类社会正在由工业社会向信息社会过渡，并由此产生一种新的经济形态——知识经济，知识经济时代的基本特征是信息的数字化和网络化、经济的全球化。信息化给各国的社会发展带来了新的机遇与挑战，并将对未来的社会发展产生深远的影响。信息是无形的财富，它是大至一个国家，小至一个企业，甚至个人的战略资源，这种观点已成为许多有识之士的共识。所以搜索和利用信息资源已变得十分重要，成为现代社会必须学会的技术。

由于Internet迅猛发展，使得信息的发布、采集、传播和利用，无论从发展规模还是网络速度上都达到了空前的水平。Internet已成为全球最大的信息资源库，是人类技术与文明的巨大财富。网上内容十分丰富，几乎囊括了商业、信息资讯、工农业生产、科技教育、娱乐休闲、文化艺术等人类活动的各个方面，是我们取之不尽用之不竭的信息宝库，而开启这个宝库的钥匙正是搜索引擎。搜索引擎以一定的策略在互联网中搜集、发现信息，对信息进行理解、提取、组织和处理，并为用户提供检索服务，从而起到信息导航的目的。

现在，Internet上已有成千上万个提供信息搜索服务的网站，除了众所周知的Google、Yahoo、百度、新浪等普通信息搜索的网站外，还存在大量的专业搜索引擎网站，它们中有些是综合性搜索引擎在原有基础上建立的具有专业特色的个性化引擎，这些搜索引擎收录了各方面、各学科、各行业的信息，建立了某一行业、某一主题或某一地区信息的专题性搜索引擎；有些是政府、大学、研究机构、某一行业或个人建立的专业性网站和个人专题网页，这些网站的内容往往非常专而精，最适合从事某一专业的研究者、工作者和学习者使用。

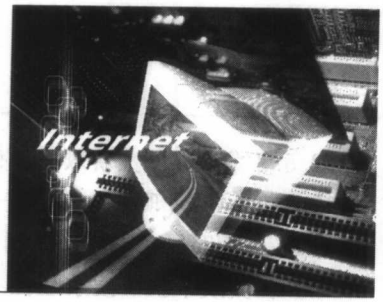
全书分为三个部分。第一部分(1～4章)通过图文系统通俗地介绍了搜索引擎的基本原理、各类搜索引擎的运行机制及互联网上的各种信息资源、以及作者长期进行网上搜索的实践经验与搜集的资料，列举实例，详实分析和总结了网上信息搜索实践中经常遇到的问题及其解决的方法和技巧。

第二部分(5～6章)为国内外著名中文和英文搜索引擎网站介绍，作者没有罗列他们的一般使用方法和随时可能变化的工具及功能，而是重点叙述其搜索技术的特点和进行精确搜索的高级搜索方法，有助于读者很快掌握常用搜索引擎的使用方法和诀窍。

第三部分(7～16 章)为作者在众多国外科技专业网站中精心选择,按自然科学和人文社会科学的各学科进行归类,并对每个网站配以简短说明,为从事某一专业的教学、科研和工作者提供搜索指南。

因此,本书既适合初学搜索的网上新手学习信息搜索技巧,又适合作为各学科研究者和专业工作者搜索专业资料时的网上导航,还可作为大专院校 Internet 信息检索课程的选用教材和学生自学材料。

由于我们水平有限,书中难免会有一些不严实之处。尤其是在第三部分国外科技专业网站介绍中,由于专业众多,限于我们的专业水平和外语水平,可能会有翻译错误或翻译不准确处,敬请读者批评指正。



目　录

第 1 章
搜索引擎的基本原理

在进入信息时代之前，人们普遍感觉到信息的匮乏，其主要原因是当时缺乏有效的信息交流工具和方式。Internet 的出现极大地丰富了人们的信息资源，但是人们仍然感到难以搜寻到所需要的信息，而 Internet 上却大量存在这样的信息。如何在 Internet 这个浩瀚的信息海洋中及时、准确地找到所需信息，“搜索引擎”便是承担这项任务的重要工具。

1.1 什么是搜索引擎？

事实上，几乎每个人在上网过程中都起始于几个最主要的搜索引擎中的一个。一般来说，搜索引擎网站会比其他的网站更有吸引力。

那么，搜索引擎是什么样的呢？引擎是英文“Engine”的音译词，代表发动机。搜索引擎的英文为“Search Engine”，即信息查找的发动机。一般将其定义为“是一种用于帮助 Internet 用户查询信息的搜索工具，它以一定的策略在 Internet 中搜集、发现信息，对信息进行理解、提取、组织和处理，并为用户提供检索服务，从而起到信息导航的目的”。搜索引擎提供的导航服务已经成为互联网上非常重要的网络服务，搜索引擎网站也被誉为“网络门户”。搜索引擎技术因而成为计算机业界和学术界争相研究和开发的对象。自十年前 Internet 第一次出现搜索引擎以来，搜索技术从简单目录搜索发展到现在具有能初步理解自然语言的复杂功能，从只能搜索文字、图形和音乐到可搜索各种多媒体信息。而搜索技术的发展又催生了各类搜索网站的诞生，知名的搜索网站从国外的 Lycos、Infoseek、Google、Yahoo 到国内的新浪、百度、搜狐、网易等等，人们早已耳熟目详。

搜索引擎是搜索引擎（Search Engine）和搜索目录（Search Directory）的统称，其实也是一个网站，只不过这样的网站专门提供信息“检索”服务，它使用特有的程序将 Internet 上的信息进行搜集、整理和归类，以帮助人们在浩如烟海的信息海洋中搜寻到各人所需信息。据中国电子信息产业发展研究院（CCID）2000 年调查，搜索引擎在网民上网经常参与的活动中位列第三，仅次于电子邮件与浏览新闻，可见搜索引擎的使用越来越受到网民的欢迎。Internet 上信息资源也在不断快速增加，因此搜索引擎对于那些在互联网上游弋、寻找信息的人们已经变得非常重要。

1.2 搜索引擎发展史

从 Internet 上出现第一个用于自动索引匿名 FTP 网站文件的程序 Archie 诞生，到今天 Internet 上已有了千万个各种各样的搜索引擎，仅仅走过了短短的十几年（表 1－1）。早期 Internet 上的搜索引擎与今天所使用的搜索引擎有所不同，早期的搜索引擎是把互联网上的资源服务器的地址收集起来，由其提供的资源类型的不同而分成不同的目录，再一层层地进行分类。人们要搜寻自己想要的信息可按照其分类系统，像剥竹笋一样层层进入，到达目的地后才找到自己想要的信息。这是最原始的方式，只适用于 Internet 信息不多的时代。今天 Internet 上信息如大海恒沙，如果使用这种方式查找一个信息就会花费很长时间。

表 1－1 搜索引擎发展简史

年份	发 生 的 事 件
1990	1. 蒙特利尔大学学生 Alan Emtage 和 Peter Deutsch 等开发出第一个用于自动索引互联网上匿名 FTP 网站文件的程序 Archie。
1991	1. 明尼苏达大学的 Mark McCahill 开发出分布式文件检索和获取系统 “Gopher”。
1992	1. 内华达大学 System Computing Services 开发出一个 Gopher 搜索工具 Veronica。
1993	1. 3 月，犹他大学 Rhett Jones 开发出另一个搜索 Gopher 的工具 “Jughead”，主要通过增加关键词搜索和增强布尔运算符的功能来提升搜索能力。 2. 6 月，Matthew Gray 开发出第一个机器人程序：World Wide Web Wanderer。 3. 10 月，Martin Koster 创建了相当于 Archie 的 HTTP 版本 ALIWEB。
1994	1. 1 月，第一个可搜索和浏览 Web 分类目录 Galaxy 发布。它支持网站、Gopher 和 Telnet 搜索。 2. 2 月，斯坦福大学杨致远和 David Filo 共同创办了超级目录索引 Yahoo!。 3. 4 月，华盛顿大学 Brian Pinkerton 创建第一个支持搜索文件全文搜索引擎：WebCrawler。 4. 7 月，Michael Mauldin 将 John Leavitt 的蜘蛛程序接入到其索引程序中创建了 Lycos。是第一个在搜索结果中使用网页自动摘要的搜索引擎。
1995	1. 2 月，Infoseek 公诸于世，同年 12 月成为 Netscape. 的默认搜索引擎。 2. 6 月，华盛顿大学 Eric Selberg 和 Oren Etizioni 创建第一个元搜索引擎 MetaCrawler。 3. 10 月，Excite 搜索引擎面市。 4. 12 月，第一个支持自然语言搜索和实现高级搜索语法的搜索引擎 AltaVista 面市。
1996	1. 2 月，Eric Brewer 和 Paul Gauthier 创建 Inktomi。 2. 5 月，HotBot 投放市场，并声称每天可索引 1 000 万网页。 3. 10 月，Web 站点列表分类目录的搜索引擎 LookSmart 面市。
1997	1. 4 月，使用自然语言提问的 Ask Jeeves 面市。 2. 发布点击付费的搜索引擎 GoTo 发布。 3. 8 月，Northernlight 正式现身，是第一个支持对搜索结果进行简单自动分类的搜索引擎。

（续）

年份	发 生 的 事 件
1998	1. 1月，台湾中正大学 GAIS 实验室创建 Open Directory 搜索引擎。 2. 9月，斯坦福大学的研究生 Larry Page 和 Sergey Brin 推出按网页等级评估相关度的搜索引擎 Google。 3. 9月，具有目录列表和使用 Inktomi 搜索结果的 MSN 搜索引擎开放。 4. 根据点击率排列网页相关度的 Direct Hit 搜索引擎发布。
1999	1. Disney 发布使用 InfoSeek 的搜索技术 Go Network。 2. 11月，NBC 推出提供 Internet 搜索和目录服务的 Web service Snap。 3. 5月，挪威科技大学发布 Fast（Alltheweb），是第一个可检索2亿 Web 页面的引擎。
2001	1. Ask Jeeves 收购全文搜索引擎 Teoma 。 2. NBCi 与 GoTo 全面合作。 3. 10月，GoTo 更名为 Overture Services，着重发展付费索引网站。
2002	1. 6月，Openfind 推出多元排序（PolyRank™）的搜索引擎。 2. 12月，雅虎收购 Inktomi。
2003	1. 2月，Overture 收购 FAST 的搜索部门。

现代意义上的搜索引擎最早是由美国蒙特利尔大学学生 Alan Emtage 于 1990 年发明的 Archie。虽然当时 World Wide Web 尚未出现，但网络中文件传输还是相当频繁的，而且由于大量的文件散布在各个分散的 FTP 主机中，查询起来非常不便，因此 Alan Emtage 想到了开发一个可以文件名查找文件的工具，于是便诞生了 Archie。

Archie 工作原理与现在的搜索引擎已经很接近，它依靠脚本程序自动搜索网上文件，然后对相关信息进行索引，供使用者以一定的表达式查询。由于 Archie 深受用户欢迎，受其启发，美国内华达 System Computing Services 大学于 1993 年开发了另一个与之非常相似的搜索工具，不过此时的搜索工具除了索引文件外，已能检索网页。

世界上第一个用于监测互联网发展规模的“机器人”程序是 Matthew Gray 于 1993 年 6 月开发的 World Wide Web Wanderer。电脑“机器人”（Computer Robot）是指某个能以人类无法达到的速度不间断地执行某项任务的软件程序。由于当时“机器人”一词在编程者中十分流行，因此人们将之称为“机器人”。这种专门用于检索信息的“机器人”程序像蜘蛛一样在网络间爬来爬去采集信息，所以搜索引擎的“机器人”程序就被称为“蜘蛛”程序。刚开始它只用来统计互联网上的服务器数量，后来则发展为能够检索网站域名。

与 Wanderer 相对应，美国麻省理工学院（MIT）的学生 Martin Koster 于 1993 年 10 月创建了 ALIWEB，它是 Archie 的 HTTP 版本。ALIWEB 不使用“机器人”程序，而是靠网站主动提交信息来建立自己的链接索引，类似于我们熟知的 Yahoo。

随着互联网的迅速发展，使得检索所有新出现的网页变得越来越困难，因此在 Matthew Gray 的 Wanderer 基础上，一些编程者将传统的“蜘蛛”程序工作原理作了一些改进。其设想是，既然所有网页都可能有连向其他网站的链接，那么从跟踪一个网站的链接开始，就有可能检索整个互联网。到 1993 年底，一些基于此原理的搜索引擎开始涌现，最具代表性的是 JumpStation、The World Wide Web Worm 和 Repository - Based Soft-

ware Engineering (RBSE) Spider。而 JumpStation 和 WWW Worm 只是以搜索工具在数据库中找到匹配信息的先后次序排列搜索结果，因此毫无信息关联度可言。而 RBSE 是第一个在搜索结果排列中引入关键字串匹配程度概念的引擎，其中的 RBSE 是第一个索引 Html 文件正文的搜索引擎，也是第一个在搜索结果排列中引入关键字串匹配程度概念的引擎。

1994 年 1 月，第一个可搜索和浏览的分类目录 EINet Galaxy（Tradewave Galaxy）面市，可支持网站、Gopher 和 Telnet 搜索。

1994 年 4 月，斯坦福大学电机工程系的两名博士生，大卫·费罗（David Filo）和美籍华人杨致远（Gerry Yang）共同创办了超级目录索引 Yahoo，最初他们是想建立自己的网络指南信息库，将其作为记录他们个人对互联网兴趣的一种方式。1995 年初，Netscape 公司邀请他们将其文件转移到 Netscape 公司提供的大型计算机上。随着访问量和收录链接数的增长，Yahoo 目录开始支持简单的数据库搜索，由于其数据是手工输入，所以不能真正称为搜索引擎，后来陆续使用 Altavista、Inktomi、Google 提供搜索引擎服务。2002 年 10 月，Yahoo 放弃自己的网站目录默认搜索，改为默认 Google 的搜索结果，成为一个真正的搜索引擎。并于 2002 和 2003 年分别收购了 Inktomi 和 Overture。

最早现代意义上的搜索引擎出现于 1994 年 7 月。当时 Michael Mauldin 将 John Leavitt 的蜘蛛程序接入到其索引程序中，创建了众所周知的 Lycos。这是搜索引擎史上又一个重要进步。Lycos 第一个在搜索结果中使用网页自动摘要，而且在当时其数据量远胜过其他搜索引擎。1999 年 4 月，Lycos 停止自己的搜索引擎，由 Fast 提供搜索引擎服务。

另一个对搜索史影响比较大的搜索引擎是 Excite，其特点是以概念搜索闻名于世。2002 年 5 月被 Infospace 收购，Excite 停止自己的搜索引擎，改用元搜索引擎 Dogpile。

互联网上第一个支持搜索文件全部文字的全文搜索引擎是 WebCrawler，于 1994 年正式发布。在它之前，用户只能通过 URL 和摘要搜索，摘要一般来自人工评论或程序自动取正文的前 100 个字。但 WebCrawler 后来陆续被 AOL 和 Excite 收购。

Infoseek 是另一个重要的搜索引擎，起初它只是一个不起眼的搜索引擎，但是它的友善用户界面、大量附加服务和较高搜索相关性使它声望日隆。1995 年 12 月，其与 Netscape 的联合，使它成为一个强势搜索引擎，但在 2001 年 2 月，Infoseek 停止了自己的搜索引擎，开始改用 Overture 的搜索结果。

第一个元搜索引擎是 1995 年由华盛顿大学的硕士生 Eric Selberg 和 Oren Etzioni 创立的 Metacrawler。用户只需提交一次搜索请求，由元搜索引擎负责转换处理后提交给多个预先选定的独立搜索引擎，并将从各独立搜索引擎返回的所有查询结果，集中起来处理后再返回给用户。

1995 年 12 月，第一个支持自然语言、实现高级搜索语法的搜索引擎 DEC 的 AltaVista 面世，用户可以用 AltaVista 搜索 Newsgroups（新闻组）的内容并从互联网上获取文章，还可搜索图片名称中的文字、搜索 Titles、Java applets、ActiveX objects 等。其最突出的优势是它的速度。2003 年 2 月，Altavista 被 Overture 收购。

Northernlight 是第一个支持对搜索结果进行简单自动分类的搜索引擎，它于 1997 年

8月发布，曾是拥有最大数据库的几个搜索引擎中的一个，它没有停用词问题，有出色的最新新闻，由7 100多种出版物组成的特殊搜集（Special Collection）栏目，而且其高级搜索语法比较出色。2002年1月，其公共搜索引擎关闭，随后被Divine收购。

目前，备受人们青睐的Google在1998年10月之前只是斯坦福大学的一个小项目。到2000年前，Google虽然以搜索准确性备受赞誉，但因其数据库较小，缺乏高级搜索语法，所以普及较慢。直到2000年中其数据库升级，又被Yahoo选为搜索引擎后，才被世人推崇。Google在网页等级（Pagerank）、动态摘要、网页快照、多文档格式支持、地图、股票、词典和寻人等集成搜索、多语言支持、用户界面等功能上有较大创新，像Altavista一样，被视为搜索引擎技术发展的新起点。

另一值得一提的搜索引擎是Openfind，它是由台湾中正大学吴升教授所领导的GAIS实验室于1998年1月创立，最初只做中文搜索引擎，鼎盛时期同时为新浪、奇摩、雅虎三大著名门户网站提供中文搜索引擎，但2000年后市场逐渐被Baidu和Google瓜分。2002年6月，Openfind重新发布基于GAIS30 Project的Openfind搜索引擎Beta版，推出多元排序（PolyRank™），宣布累计抓取网页35亿，开始进入英文搜索领域，此后技术升级明显加快。

1994年4月，中国科学院网首次与Internet网互联，即开始搜索引擎的研究，在短短10年中，中文搜索引擎的发展速度非常快。台湾和香港加入互联网的时间较早，建立和发展中文搜索引擎的历史较长，其发展速度也很快。在中国，大陆的中文搜索引擎以天网、搜狐、网易、新浪、百度搜索等为代表；台湾的中文搜索引擎以Openfind、奇摩、盖世引擎等为代表；香港的中文搜索引擎以茉莉之窗、网上行、悠游等为代表。国际上一些大型的搜索引擎公司也纷纷加入了中文搜索引擎市场，最具有代表性的是Alta Vista、Yahoo中文简体版和繁体版，还有Google、Excite等。

随着互联网规模的急剧膨胀，一家搜索引擎光靠自己单打独斗已无法适应目前市场需求，因此搜索引擎之间开始出现分工协作，并出现专业搜索引擎技术和搜索数据库服务提供商。如Inktomi，其本身并不是直接面向用户的搜索引擎，但向包括Overture、LookSmart、MSN、HotBot等在内的搜索引擎提供全文网页搜索服务。百度也属于这一类，搜狐和新浪就是使用它的技术。因此，从这个意义上说，它们是搜索引擎的搜索引擎。

由于目前各家搜索引擎的标准和功能不尽相同，给使用者造成了极大的麻烦，目前各家搜索引擎也在考虑统一信息搜索标准这个问题，探讨制定一个统一的行业标准的可能性。如果此事能成为现实，今后用户使用起来就会更觉方便了。

1.3 搜索引擎的工作原理

搜索引擎的原理起源于传统的信息全文检索理论，即计算机程序通过扫描每一篇文章中的所有词，建立以词为单位的排序文件，检索程序根据检索词在每一篇文章中出现的频率和概率，对包含这些检索词的文章进行排序，最后输出排序结果（图1-1）。