搜索引擎
原理与实践

袁津生　李　群　蔡　岳　编著

北京邮电大学出版社
www.buptpress.com

搜索引擎原理与实践

袁津生　李群　蔡岳　编著

北京邮电大学出版社
·北京·

内 容 提 要

随着搜索引擎技术的发展和不断完善，越来越多的人开始对搜索引擎原理和技术进行研究，越来越多的人喜欢上了搜索引擎。

本书从教学的角度出发，全面地阐述了搜索引擎的原理和实践，包括搜索引擎的基本原理与技术、搜索引擎的数据结构和搜索引擎的爬虫、多媒体信息检索技术以及搜索引擎开发技术。

本书适合高等院校计算机科学与技术专业及相关专业的高年级学生和研究生阅读参考，也适合相关领域的工程技术人员参阅。

图书在版编目(CIP)数据

搜索引擎原理与实践/袁津生，李群，蔡岳编著. —北京：北京邮电大学出版社，2008

ISBN 978-7-5635-1861-6

Ⅰ.搜… Ⅱ.①袁…②李…③蔡… Ⅲ.互联网络—情报检索 Ⅳ.G354.4

中国版本图书馆 CIP 数据核字(2008)第 155424 号

书　　名：搜索引擎原理与实践
作　　者：袁津生　李　群　蔡　岳
责任编辑：陈　瑶
出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路 10 号(邮编：100876)
发 行 部：电话：010-62282185　传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：北京忠信诚胶印厂
开　　本：787 mm×1 092 mm　1/16
印　　张：21.25
字　　数：529 千字
印　　数：1—3 000 册
版　　次：2008 年 11 月第 1 版　2008 年 11 月第 1 次印刷

ISBN 978-7-5635-1861-6　　定　价：36.00 元

前　言

网络的发展彻底改变了人们的思维、习惯与生活。一方面，它使我们更容易获取各种各样的信息，而另一方面，要想在数十亿网页的网络信息中精确地找到自己需要的信息简直就如“大海捞针”一般。那么在巨大的网络信息世界里，怎么样才能找到我们需要的数据呢？这就要靠搜索引擎。

面对浩如烟海的网络资源，搜索引擎就好像是航船的指南针，引领着人们在网络中冲浪。目前，搜索引擎已经成为信息检索最有效的工具。据统计，搜索引擎已经成为仅次于电子邮件的第二大网络应用服务，是用户获取信息的首要途径，在美国有超过84%的网民经常使用搜索引擎，在中国，这个数字每天都在增长。为了适应目前形势的发展，我们编写了这本书。

全书较为系统地阐述了搜索引擎的基本概念以及相关的技术，总共分为9章。第1章全面地介绍了搜索引擎的概念、搜索引擎的发展、分类及建立搜索引擎的关键技术。第2章讨论了搜索引擎的体系结构、工作原理以及元搜索引擎的概念。第3章讲述了信息处理技术，主要内容包括检索模型、文本处理技术、文本压缩技术以及Web信息处理技术。第4章介绍了信息检索技术，主要内容有顺排文档检索技术、倒排文档检索技术、布尔检索技术、加权检索技术、全文检索技术、超文本检索技术以及Web信息检索技术。第5章讨论了信息检索系统的性能评价问题，主要内容有相关性的评价、查全率和查准率等内容。第6章介绍了网络搜索引擎技术，主要内容有搜索引擎的基本结构、搜索引擎的数据结构、搜索引擎爬虫等。第7章介绍了多媒体信息检索的基本概念，主要内容有多媒体信息的知识、多媒体的基本概念、多媒体数据压缩、多媒体内容的理解以及多媒体信息检索的关键技术。第8章讨论了基于内容的多媒体，主要内容有基于内容的多媒体检索原理与特点、基于内容的音频检索、基于内容的图像检索以及基于内容的视频检索。第9章介绍了搜索引擎开发技术，主要内容有搜索引擎开发实例简介、环境的搭建与配置、网页搜集技术、网页预处理技术和查询服务。

我们编写本书的目的就是帮助读者对搜索引擎技术有一个全面的了解和应用上的提高，同时为更加深入地学习和研究搜索引擎打下良好的基础。我们

希望本书的出版能够对搜索引擎的设计者、Web 站点的管理员以及广大用户有所帮助，也希望它成为搜索引擎和信息检索有关领域学生学习的参考书。

本书是作者在多年的教学基础上，参考若干资料整理而成的。在本书的编写过程中，对基本概念、基础知识的介绍力求做到简明扼要；各章相互配合，又自成体系，每章都附有小结和习题，同时还有相关的实验。建议本课程为 40 学时，其中讲课 30 学时，实验 10 学时。

本书由袁津生、李群、蔡岳编写。由于作者水平有限，书中难免有许多错误和不当之处，请读者批评指正。

编　者

目　录

第 3 章 信息处理技术

第 4 章 信息检索技术

第 9 章　搜索引擎开发技术

第1章　搜索引擎概述

在浩瀚的网络资源中，搜索引擎(Search Engine)是一种网上信息检索工具，它能帮助用户迅速而全面地找到所需要的信息。我们可以这样对搜索引擎进行定义：搜索引擎是一种能够通过因特网接受用户的查询指令，并向用户提供符合其查询要求的信息资源网址的系统。多数网上用户使用搜索引擎来获得所需信息，据 CNNIC 的统计，用搜索引擎搜索仅次于电子邮件的应用。目前网上比较有影响的中文搜索工具有：Google、百度(Baidu)、北大天网、爱问(iask)、雅虎(Yahoo!)、搜狗(Sogou)等搜索引擎。英文的有：Yahoo!、AltaVista、Excite、Infoseek、Lycos、Aol 等。另外还有专用搜索引擎，例如，专门搜索歌曲和音乐的；专门搜索电子邮件地址、电话与地址及公众信息的；专门搜索各种文件的 FTP 搜索引擎等。

本章主要介绍搜索引擎的概念、搜索引擎的发展史、搜索引擎的分类以及一些著名的搜索引擎。

1.1　搜索引擎的概念

搜索引擎是指根据一定的策略、运用特定的计算机程序搜集互联网上的信息，在对信息进行组织和处理后，为用户提供检索服务的系统。

搜索引擎并不真正搜索互联网，它搜索的实际上是预先整理好的网页索引数据库。真正意义上的搜索引擎，通常指的是收集了互联网上几千万到几十亿个网页并对网页中的每一个词(即关键词)进行索引，建立索引数据库的全文搜索引擎。当用户查找某个关键词的时候，所有在页面内容中包含了该关键词的网页都将作为搜索结果被搜出来。在经过复杂的算法进行排序后，这些结果将按照与搜索关键词的相关度高低，依次排列。

现在的搜索引擎已普遍使用超链分析技术，除了分析索引网页本身的内容，还分析索引所有指向该网页的链接的 URL、Anchor Text，甚至链接周围的文字。所以，有时候，即使某个网页 A 中并没有某个词，比如“信息检索”，但如果有网页 B 用链接“信息检索”指向这个网页 A，那么用户搜索“信息检索”时也能找到网页 A。而且，如果有越多网页的“信息检索”链接指向网页 A，那么网页 A 在用户搜索“信息检索”时也会被认为更相关，排序也会越靠前。

搜索引擎的原理，可以分为 4 步：从互联网上抓取网页、建立索引数据库、在索引数据库

中搜索排序、对搜索结果进行处理和排序。

(1) 从互联网上抓取网页

利用能够从互联网上自动收集网页的蜘蛛系统程序,自动访问互联网,并沿着任何网页中的所有 URL 爬到其他网页,重复这过程,并把爬过的所有网页收集回来。

(2) 建立索引数据库

由分析索引系统程序对收集回来的网页进行分析,提取相关网页信息(包括网页所在URL、编码类型、页面内容包含的关键词、关键词位置、生成时间、大小、与其他网页的链接关系等),根据一定的相关度算法进行大量复杂计算,得到每一个网页针对页面内容中及超链中每一个关键词的相关度(或重要性),然后用这些相关信息建立网页索引数据库。

(3) 在索引数据库中搜索排序

当用户输入关键词搜索后,由搜索系统程序从网页索引数据库中找到符合该关键词的所有相关网页。因为所有相关网页针对该关键词的相关度早已计算好,所以只需按照现成的相关度数值排序,相关度越高,排名越靠前。最后,由页面生成系统将搜索结果的链接地址和页面内容摘要等内容组织起来返回给用户。

(4) 对搜索结果进行处理排序

所有相关网页针对该关键词的相关信息在索引库中都有记录,只需综合相关信息和网页级别形成相关度数值,然后进行排序,相关度越高,排名越靠前。最后由页面生成系统将搜索结果的链接地址和页面内容摘要等内容组织起来返回给用户。

1.2 搜索引擎的历史

在互联网发展初期,网站相对较少,信息查找比较容易。然而伴随互联网爆炸性的发展,普通网络用户想找到所需的资料简直如同大海捞针,这时为满足大众信息检索需求的专业搜索网站便应运而生了。

现代意义上的搜索引擎的祖先是 1990 年由蒙特利尔大学学生 Alan Emtage 发明的 Archie。虽然当时万维网(World Wide Web)还未出现,但是网络中文件的传输还是相当频繁的,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以文件名查找文件的系统,于是便有了 Archie。Archie 是第一个自动索引互联网上匿名 FTP 网站文件的程序,但它还不是真正的搜索引擎。Archie 是一个可搜索的 FTP 文件名列表,用户必须输入精确的文件名搜索,然后 Archie 会告诉用户哪一个 FTP 地址可以下载该文件。

由于 Archie 深受欢迎,受其启发,Nevada System Computing Services 大学于 1993 年开发了一个 Gopher(Gopher FAQ)搜索工具——Veronica(Veronica FAQ)。Jughead 是后来另一个 Gopher 搜索工具。

Robot(机器人)一词对编程者有特殊的意义。Computer Robot 是指某个能以人类无法达到的速度不断重复执行某项任务的自动程序。由于专门用于检索信息的 Robot 程序像蜘蛛(Spider)一样在网络间爬来爬去,因此,搜索引擎的 Robot 程序被称为 Spider(Spider FAQ)程序。世界上第一个 Spider 程序,是 MIT Matthew Gray 的 World Wide Web Wan-

derer，它用于追踪互联网发展规模。刚开始它只用来统计互联网上的服务器数量，后来则发展为也能够捕获网址(URL)。

与 Wanderer 相对应，1993 年 10 月 Martijn Koster 创建了 ALIWEB(Martijn Koster Announces the Availability of Aliweb)，它相当于 Archie 的 HTTP 版本。ALIWEB 不使用网络搜寻 Robot，如果网站主管们希望自己的网页被 ALIWEB 收录，需要自己提交每一个网页的简介索引信息，类似于后来大家熟知的 Yahoo!。

随着互联网的迅速发展，检索所有新出现的网页变得越来越困难，因此，在 Wanderer 基础上，一些编程者将传统的 Spider 程序工作原理作了些改进。其设想是：既然所有网页都可能有连向其他网站的链接，那么从一个网站开始，跟踪所有网页上的所有链接，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纷纷涌现，其中最负盛名的 3 个是：Scotland 的 JumpStation、Colorado 大学 Oliver McBryan 的 WWW Worm (First Mention of McBryan's World Wide Web Worm)、NASA 的 Repository-Based Software Engineering(RBSE)Spider。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。

Excite 的历史可以上溯到 1993 年 2 月，6 个 Stanford 大学的学生的想法是分析字词关系，以对互联网上的大量信息作更有效的检索。到 1993 年年中，这已是一个完全投资项目 Architext，他们还发布了一个供网站管理员在自己网站上使用的搜索软件版本，后来被叫做 Excite for Web Servers。Excite 后来曾以概念搜索闻名，2002 年 5 月，被 Infospace 收购的 Excite 停止自己的搜索引擎，改用元搜索引擎 Dogpile。

1994 年年初，华盛顿大学的学生 Brian Pinkerton 开始了他的小项目 WebCrawler(Brian Pinkerton Announces the Availability of WebCrawler)。1994 年 4 月 20 日，WebCrawler 正式亮相时仅包含来自 6 000 个服务器的内容。WebCrawler 是互联网上第一个支持搜索文件全部文字的全文搜索引擎，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。后来 WebCrawler 陆续被 AOL 和 Excite 收购，现在和 Excite 一样改用元搜索引擎 Dogpile。

1994 年 1 月，第一个既可搜索又可浏览的分类目录 EINet Galaxy(Tradewave Galaxy)上线。除了网站搜索，它还支持 Gopher 和 Telnet 搜索。

1994 年 4 月，斯坦福大学的两名博士生，美籍华人 Jerry Yang(杨致远)和 David Filo 共同创办了 Yahoo!。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。因为 Yahoo! 的数据是手工输入的，所以不能真正被归为搜索引擎，事实上只是一个可搜索的目录。Wanderer 只抓取 URL，但 URL 信息含量太小，很多信息难以单靠 URL 解释清楚，搜索效率很低。Yahoo! 中收录的网站，因为都附有简介信息，所以搜索效率明显提高。Yahoo! 以后陆续使用 Altavista、Inktomi、Google 提供搜索引擎服务；2002 年 10 月 9 日，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。

Lycos(Carnegie Mellon University Center for Machine Translation Announces Lycos)是搜索引擎史上又一个重要的进步。卡耐基·梅隆大学的 Michael Mauldin 将 John

Leavitt 的蜘蛛程序接入到其索引程序中，创建了 Lycos。1994 年 7 月 20 日，数据量为 54 000 的 Lycos 正式发布。除了相关性排序外，Lycos 还提供前缀匹配和字符相近限制，Lycos 第一个在搜索结果中使用了网页自动摘要，而最大的优势还是它远胜过其他搜索引擎的数据量，1994 年 8 月它已搜集了 394 000 个文档；1995 年 1 月搜集了 150 万个文档；1996 年 11 月已超过 6 000 万个文档。1999 年 4 月，Lycos 停止自己的蜘蛛程序，改由 Fast 提供搜索引擎服务。

Infoseek(Steve Kirsch Announces Free Demos Of the Infoseek Search Engine)是另一个重要的搜索引擎。Infoseek 沿袭 Yahoo! 和 Lycos 的概念，其有友善的用户界面和大量的附加服务，使它成为一个强势搜索引擎。当用户点击 Netscape 浏览器上的搜索按钮时，弹出 Infoseek 的搜索服务，而此前由 Yahoo! 提供该服务。Infoseek 后来曾以相关性闻名，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。

1995 年，一种新的搜索引擎形式出现了——元搜索引擎(A Meta Search Engine Roundup)。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。第一个元搜索引擎，是华盛顿大学硕士生 Eric Selberg 和 Oren Etzioni 设计的 Metacrawler。元搜索引擎概念上好听，但搜索效果始终不理想，所以没有哪个元搜索引擎有过强势地位。

1995 年 12 月，DEC 的 AltaVista 登场亮相，大量的创新功能使它迅速到达当时搜索引擎的顶峰。AltaVista 是第一个支持自然语言搜索的搜索引擎，AltaVista 是第一个实现高级搜索语法的搜索引擎，如 AND、OR、NOT 等。用户可以用 AltaVista 搜索新闻组(Newsgroups)的内容并从互联网上获得文章，还可以搜索图片名称中的文字、搜索 Titles、搜索 Java applets、搜索 ActiveX objects。AltaVista 是第一个支持用户自己向网页索引库提交或删除 URL 的搜索引擎，并能在 24 小时内上线。在面向用户的界面上，AltaVista 也作了大量革新。在搜索框下放了“tips”以帮助用户更好地表达搜索式，这些小提示经常更新，这样，在搜索过几次以后，用户会看到很多他们可能从来不知道的有趣功能。这系列功能，逐渐被其他搜索引擎广泛采用。1997 年，AltaVista 发布了一个图形演示系统 LiveTopics，帮助用户从成千上万的搜索结果中找到想要的结果。2003 年 2 月 18 日，Altavista 被 Overture 收购。

1995 年 9 月 26 日，加州伯克利分校 CS 助教 Eric Brewer、博士生 Paul Gauthier 创立了 Inktomi(UC Berkeley Announces Inktomi)，1996 年 5 月 20 日，Inktomi 公司成立，强大的 HotBot 出现在世人面前。它声称每天能抓取索引 1 000 万个网页，所以有远超过其他搜索引擎的新内容。Inktomi 于 2002 年 12 月 23 日被 Yahoo! 收购。

1998 年 10 月之前，Google 只是斯坦福大学的一个小项目——BackRub。1995 年博士生 Larry Page 开始学习搜索引擎设计，于 1997 年 9 月 15 日注册了 google.com 的域名，1997 年底，在 Sergey Brin、Scott Hassan、Alan Steremberg 的共同参与下，BachRub 开始提供 Demo。1999 年 2 月，Google 完成了从 Alpha 版到 Beta 版的蜕变。Google 公司则把 1998 年 9 月 27 日认作自己的生日。Google 在 Pagerank、动态摘要、网页快照、实时更新、多文档格式支持、地图/股票/词典/寻人等集成搜索、多语言支持、用户界面等功能上的革新，像 AltaVista 一样，再一次永远改变了搜索引擎的定义。在 2000 年以前，Google 虽然以

搜索准确性备受赞誉，但因为数据库不如其他搜索引擎大，缺乏高级搜索语法，所以推广并不快。直到 2000 年年中数据库升级后，又借着被 Yahoo! 选作搜索引擎的东风，才名声大震。Google 自 2000 年开始提供中文搜索服务。

1999 年 5 月，挪威科技大学的 Fast 公司发布了自己的搜索引擎 AllTheWeb。Fast 创立的目标是做世界上最大和最快的搜索引擎，Fast(Alltheweb)的网页搜索可利用 ODP 自动分类，支持 Flash 和 pdf 搜索，支持多语言搜索，还提供新闻搜索、图像搜索、视频、MP3 和 FTP 搜索，拥有极其强大的高级搜索功能。2003 年 2 月 25 日，Fast 的互联网搜索部门被 Overture 收购。

Teoma 起源于 1998 年 Rutgers 大学的一个项目。Apostolos Gerasoulis 教授带领华裔 Tao Yang 教授等人于新泽西 Piscataway 创立了 Teoma，2001 年春初次登场，2001 年 9 月被提问式搜索引擎 Ask Jeeves 收购，2002 年 4 月再次发布。Teoma 的数据库目前仍偏小，但有两个出色的功能：支持类似自动分类的 Refine；同时提供专业链接目录的 Resources。

Wisenut 由韩裔 Yeogirl Yun 创立。2001 年春季发布 Beta 版，2001 年 9 月 5 日发布正式版，2002 年 4 月被分类目录提供商 Looksmart 收购。Wisenut 也有两个出色的功能：包含类似自动分类和相关检索词的 WiseGuide；预览搜索结果的 Sneak-a-Peek。

Openfind 创立于 1998 年 1 月，其技术源自中国台湾中正大学吴升教授所领导的 GAIS 实验室。Openfind 起先只做中文搜索引擎，鼎盛时期同时为三大著名门户：新浪、奇摩、雅虎提供中文搜索引擎，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(PolyRankTM)，宣布累计抓取网页 35 亿个，开始进入英文搜索领域，此后技术升级明显加快。

北大天网是国家“九五”重点科技攻关项目“中文编码和分布式中英文信息发现”的研究成果，由北大计算机系网络与分布式系统研究室开发，于 1997 年 10 月 29 日正式在 CERnet 上提供服务。2000 年年初成立天网搜索引擎新课题组，由国家 973 重点基础研究发展规划项目基金资助开发，收录网页约 6 000 万个，利用教育网的优势，有强大的 FTP 搜索功能。

2000 年 1 月，两位北大校友，超链分析专利发明人、前 Infoseek 资深工程师李彦宏与好友徐勇(加州伯克利分校博士后)在北京中关村创立了百度(Baidu)公司。2001 年 8 月发布 Baidu.com 搜索引擎 Beta 版(此前 Baidu 只为其他门户网站，如搜狐、新浪、Tom 等提供搜索引擎)。2001 年 10 月 22 日正式发布 Baidu 搜索引擎，专注于中文搜索。Baidu 搜索引擎的其他特色包括：百度快照、网页预览、预览全部网页、相关搜索词、错别字纠正提示、MP3 搜索、Flash 搜索。2002 年 3 月闪电计划(Blitzen Project)开始后，技术升级明显加快。

1.3　搜索引擎的分类

搜索引擎的技术基础是全文检索技术，国外从 20 世纪 60 年代就开始对全文检索技术进行研究。全文检索通常指文本全文检索，包括信息的存储、组织、表现、查询、存取等各个方面，其核心为文本信息的索引和检索，一般用于企事业单位。随着互联网信息的发展，搜索引擎在全文检索技术上逐渐发展起来，并得到广泛的应用，但搜索引擎还是不同于全文检索。搜索引擎和常规意义上的全文检索主要区别有以下几点。

(1) 数据量

传统全文检索系统面向的是企业本身的数据或者和企业相关的数据，一般索引数据库的规模多在GB级，数据量大的也只有几百万条；但互联网网页搜索需要处理几十亿的网页，搜索引擎的策略都是采用服务器群集和分布式计算技术。

(2) 内容相关性

信息太多，查准和排序就特别重要，Google等搜索引擎采用网页链接分析技术，根据互联网上网页被链接次数作为重要性评判的依据；但全文检索的数据源中相互链接的程度并不高，不能作为判别重要性的依据，只能基于内容的相关性排序。

(3) 安全性

互联网搜索引擎的数据来源都是互联网上公开的信息，而且除了文本正文以外，其他信息都不太重要；但企业全文检索的数据源都是企业内部的信息，有等级、权限等限制，对查询方式也有更严格的要求，因此其数据一般会安全和集中地存放在数据仓库中以保证数据安全和管理的要求。

(4) 个性化和智能化

搜索引擎面向的是互联网的访问者，由于其数据量和客户数量的限制，自然语言处理技术、知识检索、知识挖掘等计算密集的智能计算技术很难应用，这也是目前搜索引擎技术努力的方向。而全文检索数据量小，检索需求明确，客户量少，在智能化和个性上更具有优势。

除了与全文检索系统有上述区别之外，搜索引擎按其工作方式主要可分为3种，分别是全文搜索引擎(Full Text Search Engine)、目录索引类(Search Index/Directory)搜索引擎和元搜索引擎(Meta Search Engine)。

1. 全文搜索引擎

全文搜索引擎是名副其实的搜索引擎，在国外具有代表性的搜索引擎有Google、AllTheWeb、AltaVista、Inktomi、Teoma、WiseNut等，国内著名的有百度、中文搜索、北大天网等。它们都是通过从互联网上提取的各个网站的信息(以网页文字为主)而建立的数据库中，检索与用户查询条件匹配的相关记录，然后按一定的排列顺序将结果返回给用户，因此它们是真正的搜索引擎。从搜索结果来源的角度，全文搜索引擎又可细分为两种：一种是拥有自己的检索程序，俗称蜘蛛程序或机器人程序，并自建网页数据库，搜索结果直接从自身的数据库中调用，如上面提到的引擎；另一种则是租用其他引擎的数据库，并按自定的格式排列搜索结果，如Lycos引擎。

全文搜索引擎有全文搜索、检索功能强、信息更新速度快等优点。但同时也有其不足之处，提供的信息虽然多而全，但可供选择的信息太多反而降低相应的命中率，并且提供的查询结果重复链接较多，层次结构不清晰，给人一种繁多杂乱的感觉。

2. 目录索引类搜索引擎

目录索引虽然有搜索功能，但在严格意义上算不上是真正的搜索引擎，仅仅是按目录分类的网站链接列表而已。用户完全可以不用进行关键词(Keywords)查询，仅靠分类目录也可找到需要的信息。目录索引中最具代表性的莫过于大名鼎鼎的Yahoo!，其他的还有Open Directory Project(DMOZ)、LookSmart、About等。国内的搜狐、新浪、网易搜索也都属于这一类。