做自己的搜索引擎

——搜索引擎精解案例教程

于天恩 编著

清华大学出版社
http://www.tup.com.cn
北京交通大学出版社
http://press.bjtu.edu.cn

做自己的搜索引擎

——搜索引擎精解案例教程

于天恩　编著

清华大学出版社
北京交通大学出版社
·北京·

内 容 简 介

本书对搜索引擎行业的发展形势及搜索引擎的相关技术作了精练、准确的介绍，同时对具体搜索引擎的配置和实现案例也进行了讲解，所有案例均可直接投入工程应用。全书分成三大部分：第一部分，搜索引擎概论，介绍了搜索引擎的历史及当前的发展状况，与搜索引擎相关的公司、技术，以及搜索引擎对人类生活的影响。第二部分，搜索引擎的原理和相关技术，介绍了搜索引擎的基本构造方法，实现索引的建立和搜索的基本算法。第三部分，基于数据库的全文检索，介绍了通常在采用数据库（MySQL，SQL Server，Oracle）的全文索引服务时，搜索引擎的基本搭建方式。本书对 Windows 索引服务，专门稍微细致地进行了介绍。

本书的所有源代码都放在出版社的网站上(http://press.bjtu.edu.cn)，读者可以免费下载。

图书在版编目(CIP)数据

做自己的搜索引擎：搜索引擎精解案例教程 / 于天恩编著. —北京：清华大学出版社；北京交通大学出版社，2007.10

ISBN 978－7－81123－141－0

Ⅰ. 做…　Ⅱ. 于…　Ⅲ. 互联网络－情报检索－案例－教材　Ⅳ. G354.4

中国版本图书馆 CIP 数据核字（2007）第 132668 号

责任编辑：谭文芳
出版发行：清 华 大 学 出 版 社　邮编：100084　电话：010－62776969
　　北京交通大学出版社　邮编：100044　电话：010－51686414
印 刷 者：北京交大印刷厂
经　　销：全国新华书店
开　　本：185×260　印张：18.75　字数：474 千字
版　　次：2007 年 10 月第 1 版　2007 年 10 月第 1 次印刷
书　　号：ISBN 978－7－81123－141－0/TP·382
印　　数：1～4000 册　定价：32.00 元

本书如有质量问题，请向北京交通大学出版社质监组反映。对您的意见和批评，我们表示欢迎和感谢。
投诉电话：010－51686043，51686008；传真：010－62225406；E-mail：press@bjtu.edu.cn。

前　言

说说搜索引擎

搜索引擎这几年热起来了。

作为世界上最大、最出名的搜索引擎，Google 在很多方面都发挥了重要的作用。

但是，当手中没有 Google 的搜索代码时，该如何搭建一个自己的搜索引擎呢？业界的人士说，全新开发一套完备的企业级搜索引擎要五年的时间。诚然，许多“业界”人士的话并不可信，不过，在搜索引擎这一块，真想要做好确实不容易。

开发搜索引擎耗费大量的时间和精力，所以有一些人开始研发独立的搜索引擎模块，并将其源代码开放，这样就可以给其他需要建立自己的搜索引擎的人提供一个基础平台。在这些开源搜索引擎模块的基础上做开发，可以节约非常多的时间和精力，大大减少了开发成本，缩短了产品投入市场的周期。而且，由于这些平台是开源的，可以亲自检查每一行代码，修改算法和显示格式等内容，这样的搜索引擎就相当于自己写的，用起来放心。

有时使用某些商业搜索模块，尽管搜索效果也很好，但是很难知道在单击“搜索”按钮的瞬间自己是否做了一些自己并不想做的事情，比如：给某个陌生人发送了一个特洛伊木马。

写这本书的动机

开源搜索引擎，对解决企业搜索等问题提供了可靠的二次开发平台（有的甚至不需要二次开发），大大提高了开发搜索引擎的效率，缩减了成本，好处多多。所以，需要有一些书来介绍如何使用开源搜索模块来提供搜索服务，而目前市面上这类书籍并不多。

我写了几本书，从搜索引擎的原理开始，讲解基于数据库的搜索引擎、基于文件系统的搜索引擎，内容覆盖各种主流的数据库（如 Oracle，SQL Server，MySQL）提供的全文搜索机制到各种开源搜索引擎（如 Lucene，DotLucene，Nutch，Hyper Estraier）实现的网络搜索平台所需的一切核心知识。当然，其中也不乏分词原理、搜索算法、Spider 抓取网页以及对各种不同格式文件（如 HTML，Word，PDF 等）的解析等知识。

实事求是地讲，做搜索引擎是一个门槛较高的行业，不是普通的程序员轻易就可以做的。但有了这几本书，即便是普通的程序员也可以轻松地搭建起自己的搜索引擎了。

这本书的特点

本书的内容以实践为主，并不深挖理论。用石志国博士的话来讲，就是“理论联系实际，并有所发展”。这是他写书的特点，同样适用于我。

这几本书中包含了必要的理论，但以实践为主。所讲的理论都不是纸上谈兵，而是可以立即付诸实践进行工程应用的。代码可以直接拿出来用（只是不要忘了输入信息验证等基本的安全检查），构建出搜索引擎。

书中含有大量的案例，由浅入深。浅，并不从“什么是程序设计”开始，所以读者需要具备一些编程的基础知识才能看懂。深，并没有深到“只可意会，无法言传”的地步，所以读者不需要担心无法看懂。

选择本书的理由

我始终认为翻译别人的书容易，只要照实翻就行。可是自己写书就不是很容易，一定要写有用的书，写好书。以下几点可以成为你看这本书的理由。

第一，搜索引擎很热，构建搜索引擎很有用。这本书可以使读者轻松掌握搭建搜索引擎平台所需的核心知识，并能轻松搭建起自己的搜索引擎平台。

第二，这本书讲解详细，书中涉及的代码全部提供。使用这些代码，即使不进行修改，也可以建立起一个中型的搜索平台了。

第三，网上有些文档，多半是英文的，还有其他国家语言的。对于外文不太好的读者来说，理解起来难免有误差和困难。这本书，至少比直接翻译外国文档要强一些，看起来会轻松些。

第四，即便网上所有的文档都不是外文的，而是中文的，我依然认为买一本书来看比在网上浏览要好得多。人要爱护眼睛，软件工程师尤其是这样。

这本书的内容和编写思路

这本书共包括 7 章，可以分成三个部分。介绍了搜索引擎相关的核心知识。

第一部分(第 1 章)：搜索引擎概论。介绍搜索引擎的历史及当前的发展状况，与搜索引擎相关的公司、技术，以及搜索引擎对人类生活的影响。

第二部分(第 2 章)：搜索引擎的原理和相关技术。介绍搜索引擎的基本构造方法，实现索引建立和搜索的基本算法。

第三部分(第 3～7 章)：基于数据库的全文搜索。介绍通常采用数据库的全文索引服务时搜索引擎的基本搭建方式。对于 Windows 索引服务，专门稍微细致地讲了一些。

这本书从当前搜索引擎市场的形势分析开始，陆续地介绍了与开发搜索引擎相关的理论和技术，这是为了使读者对搜索引擎有完全的认识，从整体上把握搜索引擎行业的走势及相关的技术情况，如果读者希望自己创立一个搜索引擎的公司或者在搜索行业有大作为，那么，这部分知识是必不可少的。

之后，介绍了常规的数据库搜索引擎的实现原理，这部分知识非常简单，大部分人对它都很熟悉。所以只做了简单的回顾，没有细致介绍。然而这部分知识又是需要的，它可以作为初级读者对搜索的新接触，也可以作为中级读者学习全文搜索引擎的过渡。介绍了这部分知识之后，再介绍主流数据库的全文索引服务和全文搜索支持就不会显得唐突了。

本书的最后介绍了目前世界上应用最广的三种数据库(Oracle，SQL Server，MySQL)提供的全文索引服务，并且介绍了有效的案例来实现应用程序层面的全文搜索模块。使用这些技术(只需在案例基础上修改)就可以轻松地搭建起企业搜索引擎。

另外，考虑到 Windows 的用户较多，专门介绍了 Windows 索引服务，在某些情况下，它是解决企业搜索的有效途径。

这样，通过这些知识的讲解，有关搜索引擎的理论、技术入门及基于数据库的全文搜索策略就阐述清楚了。对于想要涉足搜索行业的软件开发人员来说，这些知识足以将其领入门并

使其可以向前走一段路。

但是，如果读者想要再向前走一步，建议看看本书的兄弟篇《迅速搭建全文搜索平台——开源搜索引擎实战教程》。兄弟篇介绍了 Lucene，Lucene.net，DotLucene（现已更名为 Lucene.net），Lemur，Nutch，Hyper Estraier 等高性能的开源搜索引擎的原理和架设方法以及抓取网页，分词处理、中文支持等相关内容。

谅解和支持

本书从章节的安排到案例的编写，都经过了仔细揣摩，力图做到最好。然而，没有最好，只有更好。

本书尽力做到精练，且没有附加光盘以减少读者的购书成本。本书的所有源代码都放在出版社的网站上（http://press.bjtu.edu.cn），读者可以免费下载。

在这本书中发现任何问题，皆望能与我联系，以使本书臻于完善。我的 E-mail：yutianen@163.com。

衷心感谢

在我的成长过程中，得到过许多人的关心和鼓舞，他们启迪了我的思维，拓宽了我的视野，如果人生是在沙漠中旅行，他们就是眼前的足印和身上的水。

在本书的写作过程中，得到了许多人的支持和鼓励，他们是：哈尔滨工业大学语音处理研究室的李海峰老师，校部机关的蔡德彰、李新美、曲洪勤、黄峰、冯健、孔祥钰等老师，热能动力工程研究所的周逊老师，传统工业基地转型研究所的陈晓东老师，软件学院的田英鑫老师，网络与信息中心的杨庆海、何慧、李亚平、王宇航等老师，研究生院的彭远奎、朱群益、张思琦、王晓磊、雷稚蔷等老师，计时器研究所的王晓溪老师、图书馆的耿小兵老师，外国语学院的王桂芝、常巍（Sabrina）等老师，机器人研究所的蔡鹤皋院士，控制理论与制导技术研究中心的段广仁、尹航、周彬老师、科学园的宋斌、刘弋滢等老师，计算机学院的刘开昌老师，等等。

他们都是我的良师益友，是我心中的动力，每当想到他们，我总觉得自己应该放弃休息，去做更多有益的事，将真诚与善良传递下去。

在这里，对他们表示衷心的感谢！

同时，对哈工大天萌联合的一切成员表示感谢！那些曾跟我在一起的朋友，我会记得你们为我泡的每一杯咖啡和茶。那些始终保持独立的朋友，我也祝愿你们会有更加辉煌的未来。天萌联合永远是哈工大最强、最自由的社团，你们这些天萌的元老的名字，将永远铭刻在哈工大的历史上，铭刻在我的心里。

另外，需要特别感谢：

石志国博士，他的《ASP 精解案例教程》一书是我学习编程的开端。他是个高尚的人。这本书定名为“精解案例教程”就是表达对他的敬意。

顾倩萌。让我又怜又爱，时刻挂心。她的爱是股特殊的强大力量，让我找回记忆、重新开始唱歌、安静、宁神。她是我最爱的小月，是我唯一的轻松和仅存的快乐。没有小月，生命不该开始。没有小月，一切都没有意义。

于天恩

2007 年 7 月　哈工大　天人居

目　　录

第一部分　搜索引擎概论

第二部分　搜索引擎原理和相关技术

第三部分 基于数据库的全文搜索

第一部分

搜索引擎概论

第 1 章　搜索引擎概论

本章要点

本章介绍搜索引擎的发展史和当前形势，并对它未来的发展方向作了展望。本章是学习搜索引擎的引论，是一些非技术性知识。

1.1　什么是搜索引擎

搜索引擎(Search Engine)就是用来搜索的工具。如果要想从一组文件中查找符合要求的文件，就需要用到搜索引擎。

搜索的过程就是，搜索引擎接收用户提出的要求，然后进行处理，从文件组中筛选和提取出符合要求的文件。

很明显，这里涉及几个基本问题。

第一，如何使搜索引擎读懂用户的要求。这是用户和搜索引擎之间的接口问题。

第二，当搜索引擎读懂用户的要求时，如何来进行处理。这涉及不同文件格式的解析和如何最快地处理，以及如何揣摩用户的真正意图从而提供最符合需要的结果。这一点，在文件数量特别大的情况下尤为重要。

第三，如何把搜索的结果显示给用户。这看起来简单，其实也涉及很多技术问题。比如，当搜索出的结果有 300 000 条的时候，是否应该在一个页面或屏幕来显示？是否应该提供翻页的功能？你可能不认为这是个问题，认为当然要翻页了！但是，的确有一些大公司犯过这样的错误。类似的问题还有是否要存储网页快照，等等。

解决了这三个基本问题之后，一个搜索引擎就成型了。之后还要进一步考虑算法优化、程序性能问题，以及进行用户搜索偏好分析，提供与搜索相关的其他服务等。这些问题解决了，一个搜索引擎公司的技术方面就成熟了。

本书主要介绍两类搜索引擎：一类是本地搜索，就是在拥有数据的情况下进行搜索，这主要包括本地的文件系统和数据库搜索；另一类是网络搜索，就是先从局域网或互联网中抓取数据，然后对数据分析处理，之后进行搜索。前者，类似 Beagle 等桌面搜索引擎。后者，类似 Google 等网络搜索引擎。至于 FTP 搜索，技术上非常简单，本书不讲解。

两类搜索引擎的原理是互通的，理论上的差别并不显著，但实现时涉及一些技术细节的分歧，这是需要读者注意的。

本书着重讲解网络搜索，这也是目前最热的技术。

1.2　搜索引擎的发展

网络的发展极大地影响了我们的生活方式，它让我们在更容易获取信息的同时，也彻头彻尾地将我们陷入无边无际的信息海洋之中。每时每刻我们都要自觉或不自觉，被动或主动地

面对数十亿页面的网络信息，想找到自己需要的信息简直就是“大海捞针”。搜索引擎的横空出世，让我们有了探索信息海洋的指南针。随着技术的进步，这个指南针的功能也越来越强大，使用并接受它的人也越来越多。

1.2.1　历史中的搜索引擎

搜索引擎被业界公认为继广告、网络游戏、无线增值之后互联网的第四桶金。它也成为继电子邮箱之后，使用率最高的网络应用产品。

那么，今天风光无限的搜索引擎走过了怎样一段历史呢?

1990 年以前，没有人能搜索互联网。

现代意义上的所有搜索引擎的祖先，是 1990 年由加拿大魁北克省蒙特利尔的麦克吉尔(McGill)大学的学生 Alan Emtage、Peter Deutsch、Bill Wheelan 发明的 Archie。当时万维网(World Wide Web)还未出现。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，内华达系统计算服务大学于 1993 年开发了一个 Gopher 搜索工具 Veronica。后来又出现了另一个 Gopher 搜索工具 Jughead。

由于专门用于检索信息的机器人(Robot)程序像蜘蛛(Spider)一样在网络间爬来爬去，因此，搜索引擎的机器人程序被称为蜘蛛程序。世界上第一个蜘蛛程序，是麻省理工大学的互联网游荡者(World Wide Web Wanderer)，它被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，后来则发展为也能捕获网址。

与互联网游荡者相对应，1993 年 10 月，Martijn Koster 创建了 ALIWEB，它相当于 Archie 的 HTTP 版本。ALIWEB 不使用蜘蛛程序，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，这类似于后来大家熟知的 Yahoo。

1993 年底，一些基于此原理的搜索引擎开始纷纷涌现，其中最负盛名的三个是：苏格兰的 JumpStation、卡罗拉多大学 Oliver McBryan 的互联网蠕虫(World Wide Web Worm)、NASA 的 RBSE 蜘蛛。

1993 年 2 月，6 个斯坦福的大学生着手分析字词关系，以对互联网上的大量信息作更有效的检索，这就产生了 Excite。后来曾以概念搜索闻名，2002 年 5 月，Excite 被 Infospace 收购，由此停止自己的搜索引擎而改用元搜索引擎 Dogpile。

1994 年 1 月，第一个既可搜索又可浏览的分类目录 EINet Galaxy 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，美籍华人杨致远(Jerry Yang)和 David Filo 共同创办了雅虎(Yahoo!)。随着访问量和收录链接数的增长，雅虎目录开始支持简单的数据库搜索。因为当时雅虎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。

1994 年初，华盛顿大学的学生 Brian Pinkerton 开始了他的小项目网页抓取器(Web Crawler)。1994 年 4 月 20 日，网页抓取器正式亮相时仅包含来自 6000 个服务器的内容。网页抓取器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，用户只能通过网页地址和摘要搜索，摘要一般来自人工评论或程序自动取正文的前 100 个字。后来网页抓

取器陆续被美国在线(AOL)和 Excite 收购,现在和 Excite 一样改用元搜索引擎 Dogpile。

Lycos 是搜索引擎史上又一个重要的进步。卡内基梅隆大学的 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 还提供了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,而它最大的优势还在于远胜过其他搜索引擎的数据量:1994 年 8 月,394 000 个文档;1995 年 1 月,150 万个文档;1996 年 11 月,超过 6000 万个文档。(注:1999 年 4 月,Lycos 停止自己的蜘蛛,改由 Fast 提供搜索引擎服务。)

Infoseek 是另一个重要的搜索引擎,虽然该公司声称 1994 年 1 月已创立,但直到 1994 年年底它的搜索引擎才与公众见面。Infoseek 起初只是一个不起眼的搜索引擎,它沿袭雅虎和 Lycos 的概念,并没有什么独特的革新。但是它的发展史和后来受到的众口称赞证明,起初第一个登台并不总是很重要。Infoseek 友善的用户界面、大量附加服务(例如:UPS 跟踪,新闻等)使它声望日隆。而 1995 年 12 月与 Netscape 的战略性协议,使 Infoseek 成为一个强势搜索引擎:当用户单击 Netscape 浏览器上的搜索按钮时,就会弹出 Infoseek 的搜索服务,而此前是由雅虎提供该服务。(注: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,出现了一种新的搜索引擎形式——元搜索引擎。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 的元抓取器(Meta Crawler)。但元搜索引擎的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。

DEC 公司的 AltaVista 在 1995 年 12 月登场亮相,大量的创新功能使它迅速达到当时搜索引擎的顶峰。AltaVista 最突出的优势在于它的速度。而 AltaVista 的另一些新功能,则永远改变了搜索引擎的定义。AltaVista 是第一个支持自然语言搜索的搜索引擎,是第一个实现高级搜索语法(如 AND,OR,NOT 等)的搜索引擎。用户可以用 AltaVista 搜索新闻组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、搜索标题、搜索 Java Applets、搜索 ActiveX 控件。另外,AltaVista 也声称是第一个支持用户自己向网页索引库提交或删除网页路径的搜索引擎,并能在 24 小时内上线。AltaVista 最有趣的新功能之一,是搜索有链接指向某个网页路径的所有网站。在面向用户的界面上,AltaVista 也作了大量革新。它在搜索框区域下放了提示以帮助用户更好地输入表达搜索式,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的有趣功能。这一系列功能,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信息。

随后到来的搜索引擎是 HotBot。1995 年 9 月 26 日,加州伯克利分校的助教 Eric Brewer、博士生 Paul Gauthier 创立了 Inktomi 公司,随后强大的 HotBot 出现在世人面前。它声称每天能抓取 1 千万页索引以上,所以有远超过其他搜索引擎的新内容。HotBot 也大量运用 cookie 储存用户的个人搜索喜好设置。Hotbot 曾是随后几年最受欢迎的搜索引擎之一,后被 Lycos 收购。

Northernlight 公司于 1995 年 9 月成立于马萨诸塞州剑桥,1997 年 8 月,Northernlight 搜索引擎正式现身。它曾是拥有最大数据库的搜索引擎之一,它没有忽略词(Stop Words),具有良好的高级搜索语法,第一个支持对搜索结果进行简单的自动分类。2002 年 1 月 16 日,

Northernlight 公共搜索引擎关闭，随后被 Divine 公司收购。

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 和 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 在页面等级、动态摘要、网页快照、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，所以使用价值不是很高，推广并不快。直到 2000 年其数据库升级后，又借被雅虎选作搜索引擎的东风，才一飞冲天。

Fast 公司创立于 1997 年，是挪威科技大学学术研究的副产品。1999 年 5 月，该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，经过几年的发展，已经几乎达到了这个目标。Fast 的网页搜索可利用开放式目录管理(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。

Teoma 起源于 1998 年 Rutgers 大学的一个项目。Apostolos Gerasoulis 教授带领华裔 Tao Yang 教授等人在新泽西创立了 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 的数据库偏小，但有两个新颖的功能：支持类似自动分类的再次提取，同时提供专业链接目录的资源。

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 也有两个新颖的功能：包含类似自动分类和相关检索词的智能向导；预览搜索结果的一瞥(Sneak-a-Peek)。

Gigablast 由前 Infoseek 工程师 Matt Wells 创立，2002 年 3 月展示 Pre-beta 版，2002 年 7 月 21 日发布 Beta 版。Gigablast 的数据库偏小，但也提供网页快照，一个特色功能是即时索引网页，网页一提交它就能搜索。

Openfind 创立于 1998 年 1 月，其技术源自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，曾经是最好的中文搜索引擎，但 2000 年后市场逐渐被百度和 Google 瓜分。2002 年 6 月，Openfind 重新发布基于 GAIS30 项目的 Openfind 搜索引擎 Beta 版，推出多元排序，宣布累计抓取了 35 亿网页，开始进入英文搜索领域，此后技术升级明显加快。

北大天网是中国国家“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 6000 万个，利用教育网优势，具有强大的 FTP 搜索功能。

2000 年 1 月，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)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 只为其他门户网站提供搜索引擎)，2001 年 10 月 22 日正式发布百度搜索

引擎。百度虽然只提供中文搜索,但它是最大的中文数据库。百度搜索引擎的其他特色包括:网页快照、网页预览/预览全部网页、相关搜索词、错别字纠正提示、新闻搜索、Flash 搜索、信息快递搜索。在 2002 年 3 月闪电计划(Blitzen Project)开始后,其技术升级明显加快。

1.2.2 搜索引擎的分类

1. 按工作方式分类

搜索引擎按其工作方式主要可分为三种:全文搜索引擎(Full Text Search Engine)、目录索引类搜索引擎(Search Index/Directory)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,国外具代表性的有 Google、Fast、AltaVista、Inktomi、Teoma、WiseNut 等,国内著名的有百度。它们都是从互联网上提取的各个网站的信息(以网页文字为主)存入数据库中,然后检索与用户查询条件匹配的相关记录,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(Indexer),俗称"蜘蛛"(Spider)程序或"机器人"(Robot)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 7 家引擎;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。

(2) 目录索引

目录索引虽然有搜索功能,但在严格来讲算不上是真正的搜索引擎,只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雅虎。其他著名的目录索引还有 LookSmart、About 等。

(3) 元搜索引擎

元搜索引擎在接受用户查询请求时,同时在其多个引擎上进行搜索,并将结果返回给用户。著名的元搜索引擎有 InfoSpace、Dogpile、Vivisimo 等,中文元搜索引擎中具代表性的有搜星搜索引擎。在搜索结果排列方面,有的直接按来源引擎排列搜索结果,如 Dogpile,有的则按自定的规则将结果重新排列组合,如 Vivisimo。

2. 非主流形式的搜索引擎

除上述三大类引擎外,还有以下几种非主流形式。

(1) 集合式搜索引擎

如 HotBot 在 2002 年底推出的引擎。该引擎类似 META 搜索引擎,但区别在于,不是同时调用多个引擎进行搜索,而是由用户从提供的 4 个引擎当中选择,因此称它"集合式"搜索引擎更确切些。

(2) 门户搜索引擎

如 AOL Search、MSN Search 等。虽然提供搜索服务,但自身既没有分类目录也没有网页数据库,其搜索结果完全来自其他引擎。

(3) 免费链接列表(Free For All Links,FFA)

这类网站一般只简单地滚动排列链接条目,少部分有简单的分类目录,不过规模比起雅虎等目录索引来要小得多。