Broadview
www.broadview.com.cn

走进搜索引擎

Stepping into Search Engine

梁斌 编著

打造优质搜索引擎的第一书!

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

走进搜索引擎

Stepping into Search Engine

梁斌 编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

在网络普及的今天，人们经常在信息海洋中彷徨，在万维网迷宫般的复杂与魅力之间挣扎。直到搜索引擎这一伟大的技术产生，才使得人们犹如找到了走出迷宫的灯塔，可以非常便捷地找到自己所需要的信息。

正是因为搜索引擎离我们越来越近，所以越来越多的人期待着能够揭开她神秘的面纱。其实搜索引擎并不是变幻莫测的大海，也不是高不可攀的山峰。请拿起本书，它就是引领你的火炬，它就是你身边的伙伴，它将带着你走进搜索引擎。在那里，你必将会被搜索引擎精致的设计和宏伟的架构所征服。

本书由搜索引擎开发研究领域年轻而有活力的科学家精心编写，作者将自己对搜索引擎的深刻理解和实际应用巧妙地结合，使得从未接触过搜索引擎原理的读者也能够轻松地在搜索引擎的大厦中遨游一番。

本书作为搜索引擎原理与技术的入门书籍，面向那些有志从事搜索引擎行业的青年学生、需要完整理解并优化搜索引擎的专业技术人员、搜索引擎的营销人员，以及网站的负责人等。

本书是从事搜索引擎开发的工程技术人员难得的参考书，也可作为大中专院校相关专业的教学辅导书。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

走进搜索引擎 / 梁斌编著.—北京：电子工业出版社，2007.10
ISBN 978-7-121-04922-4

Ⅰ. 走… Ⅱ.梁… Ⅲ.互联网络—情报检索 Ⅳ.G354.4

中国版本图书馆 CIP 数据核字（2007）第 132658 号

责任编辑：孙学瑛
印　　刷：北京智力达印刷有限公司
装　　订：北京中新伟业印刷有限公司
出版发行：电子工业出版社出版
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×980　1/16　印张：18.25　字数：258 千字
印　　次：2007 年 10 月第 1 次印刷
印　　数：5000 册　　定价：49.80 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

推荐序

搜索改变生活

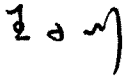
搜索引擎，改变了二十一世纪人类的生活方式。越来越多的人学会通过搜索，从海量的互联网信息中找到和分享全人类的经验与智慧。搜索引擎不仅成为人们最常使用的互联网应用，同时也开创了一种优秀的商业模式，引领着互联网技术与商业的发展。

搜索引擎有着对计算机科学与技术孜孜不倦的追求，将人工智能、信息检索、文本处理、系统结构、数据库技术等都发挥到了极致，以追求响应更快、结果更准确。如同制造原子弹或者登月计划一样，搜索引擎在科技方面的突破也会辐射到互联网的其他应用领域。

在 Google、Yahoo 等国外公司投入巨资进入搜索引擎应用领域之时，中国本土公司百度、搜狗、有道，也不断加大投入，努力为中国用户打造更优秀的中文搜索引擎。能够有能力、有信念从事这样研发的中国公司不是太多，而是太少了；而对技术有着狂热爱好，能够将搜索引擎研发作为事业来看待的人才也不是太多，而是太少了。

我经常在技术论坛上看到这样的提问："现在的搜索引擎应该怎样设计数据库的表结构呢？（言下之意，还在考虑通用的关系数据库）""搜索引擎是在用户搜索的时候，怎么能够这么快抓取其他网站的信息呢？（看起来还没有理解索引和镜像的概念）"。这些提问充分显示出大多数技术人员对搜索引擎工作原理的不理解。

市面上，对搜索引擎原理进行讲解的书非常稀少。搜索引擎公司更将各自研发中的关键技术和经验作为最高机密不予公开。**我希望借助这本书，不仅可以让大众技术人员通过了解搜索引擎的工作原理，提高对技术的理解，更为那些想以搜索引擎研发为未来事业的人，打开一扇窗户。**



搜狐 技术副总裁

2007-9-1

试读结束：需要全本请在线购买： www.ertongbook.com

前　言

随着互联网的蓬勃发展，建立在互联网之上的各种应用也层出不穷，其中最为成功的莫过于万维网（WWW）。万维网被称为“网中之网”，是互联网上最受欢迎的服务之一。它运用超文本技术为人们访问信息资源提供了巨大的方便，但也以非线性组织的构建方式使人们在信息海洋中彷徨。

在互联网彷徨无计之时，搜索引擎就像引领互联网走出迷雾的灯塔。在搜索引擎的帮助下，网民不再需要记住复杂的网址和复杂的路径（URL），而只需记住搜索引擎的入口，提交查询词即可直接找到想要找的信息。就这样，搜索引擎成为破解互联网迷宫的金钥匙。

随着搜索引擎技术和商业模式的成熟，一方面，越来越多的人对搜索引擎产生了浓厚的兴趣，需要深入了解和认识搜索引擎；另一方面，由于搜索引擎是最高端、最复杂的互联网技术之一，各家公司都将核心技术秘而不宣。在这样的背景下，本书犹如一本引领读者走进搜索引擎的导游图，可以带领读者进入搜索引擎的腹地，一探搜索引擎神秘面纱的背后。那里是搜索引擎精致的设计和宏伟的架构，必定让您不虚此行。

本书主要内容

本书共包括 7 章，每章的主要内容如下。

第 1 章“引言”介绍了搜索引擎的基本背景知识，通过介绍搜索引擎的历史回顾了搜索引擎的发展历程。

第 2 章“搜索引擎概貌”宏观上介绍了搜索引擎，以及搜索引擎的主要系统划分。

第 3 章“搜索引擎的下载系统”介绍了搜索引擎下载系统的背景知识、设计原理和技巧，以及网页库的设计等。

第 4 章“搜索引擎的分析系统”分别介绍了信息抽取、网页查重、中文分词，以及 PageRank 等分析系统子模块的计算原理和实现细节。

第 5 章“搜索引擎的索引系统”通过全文检索、文档编号、正排表和倒排表等基本概念，全方位介绍了搜索引擎核心的索引技术。

第 6 章“搜索引擎的查询系统”介绍了查询系统的两个主要功能模块，即检索模块和摘要提取模块的工作方法和设计技巧，解开搜索引擎查询准确的奥秘所在。

第 7 章“搜索引擎的其他话题”通过回答有关搜索引擎的常见问题系统地介绍了搜索引擎各个系统的相互关系，并展望了搜索引擎未来的发展。

附录部分为搜索引擎系统全图，它将有助于理解搜索引擎各个系统的相互配合和协作关系。

如何阅读本书

本书中带★的章节中介绍了具有一定深度的理论知识，没有学过数学分析和概率

论等基础知识的读者可以浏览这些章节。而不需要深究推导过程，只需要了解一些基本的结论即可。

在本书的主要章节，即第 3 章至第 6 章中，第 5 章介绍的索引系统最为复杂，较难理解。读者在第 1 遍通读时，不需要深入细节。待阅读查询系统后，再回头阅读索引系统会有更大地帮助。对于一些难以理解的地方可以阅读章节结束部分给出的参考文献，深入理解这些难点问题。

非技术类读者可以忽略第 5 章中的大部分内容 ,以及其他章节中技术性强的部分，而集中精力关注搜索引擎的基本策略和方法。

附录部分为搜索引擎全部系统的架构图，它将有助于宏观上理解搜索引擎各个系统相互配合的过程。

关于本书作者

作者毕业于南京大学软件学院，获得软件工程硕士学位。曾经发表过多篇论文，获得 1 项国家专利。作者主要的兴趣方向包括数据挖掘、Web 挖掘、搜索引擎和软件工程等，目前在清华大学信息科学与技术国家实验室从事搜索引擎相关研究工作。

致谢

我首先要特别感谢我的妻子伍绍连，正是她的无私支持，使我能够全身心投入到写作中。在此书完成后，她通读了全书并提出了大量宝贵意见，使得本书增色不少。

感谢电子工业出版社计算机图书事业部孙学瑛女士，她是推动本书的完成最为关

键性的人物。她参与了此书创作的全过程，为笔者提供了有关图书市场的宝贵信息，使得本书更加面向读者。

感谢此书参考文献的全部作者、搜索引擎研究界的杰出科学家们，以及其他为此书提出宝贵技术意见的业界同行，正是你们杰出的成就和无私的帮助，才能使本书达到了一定的学术水平。

由于作者水平有限，书中不足及错误之处在所难免，敬请专家和读者给予批评指正。

梁 斌

2007 年 8 月

目　录

第一章　引言 1

时至今日，万维网迷宫般的复杂和魅力还在继续。因为它每天都在不断地产生、更新或消失各种各样的网页。其魅力依然，然而复杂不在。正是由于诞生了搜索引擎（search engine）这样伟大的技术，万维网复杂的局面才被打破。搜索引擎成为带领人们走出迷宫的灯塔，帮助千百万的网民便捷地找到重要的信息。

第一节　什么是搜索引擎 2

第二节　搜索引擎的发展简史 5

搜索引擎的发展历史 5

第三节　搜索引擎大事快览 15

第四节　国内著名搜索引擎 17

百度（www.baidu.com） 17

中搜（www.zhongsou.com） 18

天网（e.pku.edu.cn） 19

搜狗（www.sogou.com） 20

参考文献 21

第二章　搜索引擎概貌 23

万维网的发展迫切地要求一种快速、全面、准确且稳定可靠的信息查询方法，由于搜

索引擎满足了这 4 个需求，所以才奠定了其在科学技术上的高度。有人甚至把搜索引擎和操作系统并列为当今最为复杂的系统软件。

第一节　搜索引擎的主要需求.....24
查得快.....24
查得全.....25
查得准.....25
查得稳.....27
第二节　搜索引擎的 4 大系统.....28
搜索引擎的体系结构.....28

第三章　搜索引擎的下载系统.....31

在搜索引擎的 4 大系统中，第 1 个系统是下载系统。和航天运载火箭系统的动力系统一样，下载系统是搜索引擎大厦的基础。搜索的数据均来自于下载系统的工作，其工作方式巧妙、合理且强大。爬虫（也称为“Crawler”，中文译为“爬虫”，或者“蜘蛛”）是其中最华彩的乐章。让我们从爬虫开始，逐渐进入闪烁着奇异光芒的领地。

第一节　爬虫的发展历史.....32
世界上第 1 个爬虫.....32
爬虫的发展历程.....33
第二节　万维网及其网页分析.....34
蝴蝶结型的万维网.....34
万维网的直径.....37
万维网的规模及变化特征.....39
网页的特征.....39
第三节　有关爬虫的基本概念.....41

试读结束：需要全本请在线购买：　www.ertongbook.com

爬虫 41
种子站点 41
URL 42
Backlinks 42
第四节 网页抓取原理 43
telnet 和 wget 43
从种子站点开始逐层抓取 44
不重复抓取策略 50
网页抓取优先策略 59
网页重访策略★ 61
Robots 协议 67
其他应该注意的礼貌性问题 69
抓取提速策略（合作抓取策略） 70
第五节 网页库 77
第六节 下载系统回顾及未来发展 82
参考文献 84
第四章 搜索引擎的分析系统 86
搜索引擎的 4 大系统中的第 2 个系统是分析系统，分析系统主要完成的工作包括信息抽取、网页消重、中文分词和 PageRank 计算等。
第一节 知识准备 87
HTML 语言 87
锚文本（anchor text） 87
半结构化数据（Semi-structured data） 88

第二节　信息抽取及网页信息结构化 89
网页结构化的目标 89
建立 HTML 标签树 93
通过投票方法得到正文 98
网页结构化过程回顾 103
第三节　网页查重 105
网页查重技术发展历史 105
网页查重实现方法 107
第四节　中文分词 113
什么是中文分词 113
通过字典实现分词 114
通过统计学方法实现分词 120
第五节　PageRank 121
PageRank 的来由 121
PageRank 的基本想法 122
PageRank 的计算公式 124
PageRank 的计算方法 ★ 129
第六节　分析系统结构图 134
参考文献 136
第五章　搜索引擎的索引系统 139
在搜索引擎的 4 大系统中，第 3 个系统称为“索引系统”。该系统就好像搜索引擎的数据大本营，在这里存储了并索引了数以亿计的网页。
第一节　知识准备 140

信息 140
索引 141
倒排索引、倒排表、临时倒排文件、最终倒排文件 141
其他概念 142
第二节　全文检索 143
全文检索 143
第三节　文档编号 146
编号的本质 146
文档编号的方法 147
游程编码 149
第四节　倒排索引 154
经典的倒排索引 154
正排索引（前向索引） 155
倒排索引 158
第五节　数据规模的估计 163
齐普夫法则 163
布尔检索模型下的索引规模估计★ 165
第六节　涉及存储规模的一些计算 170
正排表与倒排表的合并 170
多个临时倒排文件的归并 174
倒排索引分布式存储 179
倒排文件缓存 183
倒排索引词典统计信息的计算 183
第七节　倒排索引文件的创建过程 185

创建倒排表 185
计算统计信息 187
参考文献 189

第六章　搜索引擎的查询系统 191

在搜索引擎 4 大系统中，第 4 个系统称为“查询系统”。查询系统直接面对用户，在接收用户的查询请求后，通过检索、排序及摘要提取等计算，将计算结果组织成搜索结果页返回给用户。整个查询过程不仅要快，而且必须能够提供用户满意的查询结果。

第一节　知识准备 192
什么是信息熵 192
检索和查询的区别 196
检索词和查询词的区别 196
自动文本摘要(Automatic Text Summarization) 197
第二节　网页信息检索 198
早期的检索模型 198
向量空间模型（Vector Space Models） 201
关键词权重的量化方法 TF/IDF★ 207
搜索引擎采用的检索模型 213
多文档列表求交计算 215
检索结果排序 222
堆排序 223
第三节　中文自动摘要 230
自动摘要的发展历史 230
自动摘要的含义和实现 231

第四节　生成搜索结果页 239
 生成搜索结果页 239
第五节　搜索结果页的缓存 242
 搜索结果页的缓存 242
第六节　推测用户查询意图 245
 查询分类 245
 推测信息类、事物类的查询意图 247
第七节　查询系统的当前热点和发展方向 249
 查询系统的当前热点 249
参考文献 250

第七章　搜索引擎的其他话题 252

本书初步介绍了搜索引擎 4 大系统各自的运作原理，以及相互配合的关系。本节通过回答一些有关搜索引擎的基本问题，以从宏观上更好地理解和认识搜索引擎。

第一节　搜索引擎问与答 253
 为什么搜索引擎的搜索速度这么快 253
 为什么搜索引擎能够返回那么多的查询结果 255
 为什么搜索引擎总能返回最想要的结果 256
 搜索引擎如何大规模存储网页的 257
 什么是 SEO 259
 什么是元搜索引擎 260
 搜索引擎认为的作弊行为是哪些 261
 如何进一步学习和了解搜索引擎发展的最新成果 262
第二节　搜索引擎未来的发展 265

新兴的搜索产品 265
搜索技术的未来 268
参考文献 270
附录 A　搜索引擎系统结构全观图 271

试读结束：需要全本请在线购买：　www.ertongbook.com