Broadview
www.broadview.com.cn

走进搜索引擎

STEPPING INTO SEARCH ENGINE

潘雪峰　花贵春　梁　斌　编著

（第2版）

一本帮您轻松入门的搜索引擎技术书

新增内容：

搜索引擎性能调优

搜索引擎日志分析

基于学习进行 结果排序优化

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

潘雪峰　花贵春　梁　斌　编著

（第2版）

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

内 容 简 介

本书由搜索引擎开发研究领域三位年轻的博士生精心编写，作者们希望将自己对搜索引擎的理解和实际应用相结合，让未接触过搜索引擎原理和方法的读者也能轻松读懂该书的大部分内容。

本书在第 1 版的基础上，删除了搜索引擎历史等章节，并对错误和不足进行了修订和补充，同时增加了潘雪峰编写的第 6 章“搜索引擎日志分析”，花贵春编写的第 7 章“排序学习（Learning to Rank）”和梁斌编写的第 8 章“搜索引擎的性能调优”三个主要章节，变更的内容约占第 1 版的一半。

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。

图书在版编目（CIP）数据

走进搜索引擎 / 潘雪峰，花贵春，梁斌编著. —2 版. —北京：电子工业出版社，2011.5
ISBN 978-7-121-13104-2

Ⅰ. ①走… Ⅱ. ①潘… ②花… ③梁… Ⅲ. ①网络检索 Ⅳ. ①G354.4

中国版本图书馆 CIP 数据核字（2011）第 041946 号

策划编辑：孙学瑛
责任编辑：孙学瑛
印　　刷：北京中新伟业印刷有限公司
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：787×980　1/16　印张：18.75　字数：400 千字
印　　次：2011 年 5 月第 1 次印刷
印　　数：4000 册　定价：49.00 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。

作者序

本书第 1 版出版到现在已经 3 年了。在这段不长的时光里，搜索引擎技术有了进一步的发展。其中比较突出的是，随着数据规模进一步增大，为提升用户体验，搜索引擎性能进一步优化；在更广泛的用户参与下，增强了基于用户行为进行效果改进的能力。这也使得本书有了改版以适应这些重大变化的必要。

基于此，本书第 2 版增加了搜索引擎性能调优、搜索引擎日志分析，以及基于学习进行排序优化三方面的内容，希望能让读者跟上搜索技术的发展潮流，在这一领域的前沿真切地感受到它的勃勃生机。

当前，搜索技术已经不再局限于搜索引擎本身，它所建立的一套驾驭互联网级别海量数据的架构和理念正日益扩展到整个信息技术领域。而随着世界的日益信息化、数字化、网络化，这些理念的深远影响还会进一步显现。这又将是一次新的科技浪潮。

时光流逝，却有如轮回。信息技术产业，甚至整个科技界，正是在这样的浪潮更迭中不断进步。从 AT&T 的有线电话到 IBM 的大型机，到 Apple 的 PC 机，到 Intel 的 CPU，到 Motorola 的无线通信，到 Microsoft 的操作系统，到 Cisco 的路由器，到 Google 的搜索引擎，概莫能外。一次次浪潮，一个个产业巨擘，终将随自己的时代而去，但它们所带来的影响却将投射在人类文明的历史上，永不消逝。

至于搜索的浪潮究竟将持续多长时间，在整个 IT 史上留下怎样的一笔，只有时间才能告诉人们答案。此时此刻，置身其中，让我们打开书本，接受浪潮之巅的洗礼，走进搜索引擎。

关于本书作者

作者潘雪峰，毕业于中国科学院计算技术研究所，工学博士。研究兴趣包括多

媒体内容分析、机器学习和互联网数据挖掘，现从事搜索引擎领域相关工作。

作者花贵春，目前在清华大学信息科学与技术国家实验室攻读博士学位，研究兴趣包括机器学习及其在搜索领域的应用。

作者梁斌，目前在清华大学信息科学与技术国家实验室攻读博士学位，研究兴趣包括大规模数据处理、搜索引擎和软件工程等。

致谢

笔者首先要特别感谢他们的妻子，感谢她们在繁忙的工作和学习之余，包揽了家里家外大大小小的事务，还在笔者们有所懈怠的时候，从精神上给予莫大的支持和鼓励。正是她们无私的支持，才使本书得以面世。

感谢电子工业出版社计算机图书出版分社孙学瑛女士和邓彩屏女士，她们除了参与了此书的创作过程，还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，面向市场。

感谢本书参考文献的作者们、搜索引擎研究界的学者们，以及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，才使得本书有了写作的基础和必要。

由于笔者水平有限，加之搜索领域的发展日新月异，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。

潘雪峰、花贵春、梁斌

2011 年 2 月

目 录

第 1 章 引言 …… 1

1.1 搜索引擎概述 …… 2

1.1.1 目录式搜索引擎 …… 2

1.1.2 全文搜索引擎 …… 3

1.1.3 元搜索引擎（Meta-Search Engine） …… 3

1.2 搜索引擎的主要需求 …… 3

1.2.1 快 …… 4

1.2.2 全 …… 4

1.2.3 准 …… 4

1.2.4 稳 …… 5

1.2.5 省 …… 5

1.3 搜索引擎的 4 大系统 …… 6

1.3.1 搜索引擎的体系结构 …… 6

第 2 章 搜索引擎的下载系统 …… 8

2.1 爬虫的发展历史 …… 9

2.1.1 世界上第 1 个爬虫 …… 9

2.1.2 爬虫的发展历程 …… 9

2.2 万维网及其网页分析 …… 9

2.2.1 蝴蝶结型的万维网 …… 10

2.2.2 万维网的直径 …… 12

2.2.3 万维网的规模及变化特征 …… 12

2.2.4 网页的特征 …… 13

2.3 有关爬虫的基本概念 …… 13

2.3.1 爬虫 …… 13

2.3.2 种子站点 …… 14
2.3.3 URL …… 14
2.3.4 Backlinks …… 14
2.4 网页抓取原理 …… 14
2.4.1 telnet 和 wget …… 14
2.4.2 从种子站点开始逐层抓取 …… 15
2.4.3 不重复抓取策略 …… 19
2.4.4 网页抓取优先策略 …… 25
2.4.5 网页重访策略 …… 26
2.4.6 Robots 协议 …… 30
2.4.7 其他应该注意的礼貌性问题 …… 31
2.4.8 重要性网页优先抓取策略 …… 32
2.4.9 抓取提速策略（合作抓取策略） …… 34
2.5 网页库 …… 38
2.6 下载系统回顾及未来发展 …… 41
参考文献 …… 42
第 3 章 搜索引擎的分析系统 …… 44
3.1 知识准备 …… 45
3.1.1 HTML 语言 …… 45
3.1.2 锚文本（anchor text） …… 45
3.1.3 半结构化数据（semi-structured data） …… 45
3.2 信息抽取及网页信息结构化 …… 45
3.2.1 网页结构化的目标 …… 46
3.2.2 建立 HTML 标签树 …… 48
3.2.3 通过投票方法得到正文 …… 52
3.2.4 网页结构化过程回顾 …… 55
3.3 网页查重 …… 56
3.3.1 网页查重技术发展历史 …… 56
3.3.2 网页查重实现方法 …… 58
3.4 中文分词 …… 61
3.4.1 什么是中文分词 …… 61
3.4.2 通过字典实现分词 …… 61

3.4.3 基于统计的分词方法 …… 65
3.5 PageRank …… 67
3.5.1 PageRank 的来由 …… 68
3.5.2 PageRank 的基本想法 …… 68
3.5.3 PageRank 的计算公式 …… 69
3.5.4 PageRank 的计算方法 …… 73
3.6 分析系统结构图 …… 76
参考文献 …… 77
第 4 章 搜索引擎的索引系统 …… 79
4.1 知识准备 …… 80
4.1.1 信息 …… 80
4.1.2 索引 …… 80
4.1.3 倒排索引、倒排表、临时倒排文件、最终倒排文件 …… 80
4.1.4 其他概念 …… 81
4.2 全文检索 …… 81
4.3 文档编号 …… 82
4.3.1 编号的本质 …… 82
4.3.2 文档编号的方法 …… 83
4.3.3 游程编码 …… 84
4.4 倒排索引 …… 87
4.4.1 经典的倒排索引 …… 87
4.4.2 正排索引（前向索引） …… 88
4.4.3 倒排索引 …… 90
4.5 数据规模的估计 …… 92
4.5.1 齐普夫法则 …… 92
4.5.2 布尔检索模型下的索引规模估计 …… 94
4.6 涉及存储规模的一些计算 …… 97
4.6.1 正排表与倒排表的合并 …… 97
4.6.2 多个临时倒排文件的归并 …… 100
4.6.3 倒排索引分布式存储 …… 103
4.6.4 倒排文件缓存 …… 106
4.6.5 倒排索引词典统计信息的计算 …… 106

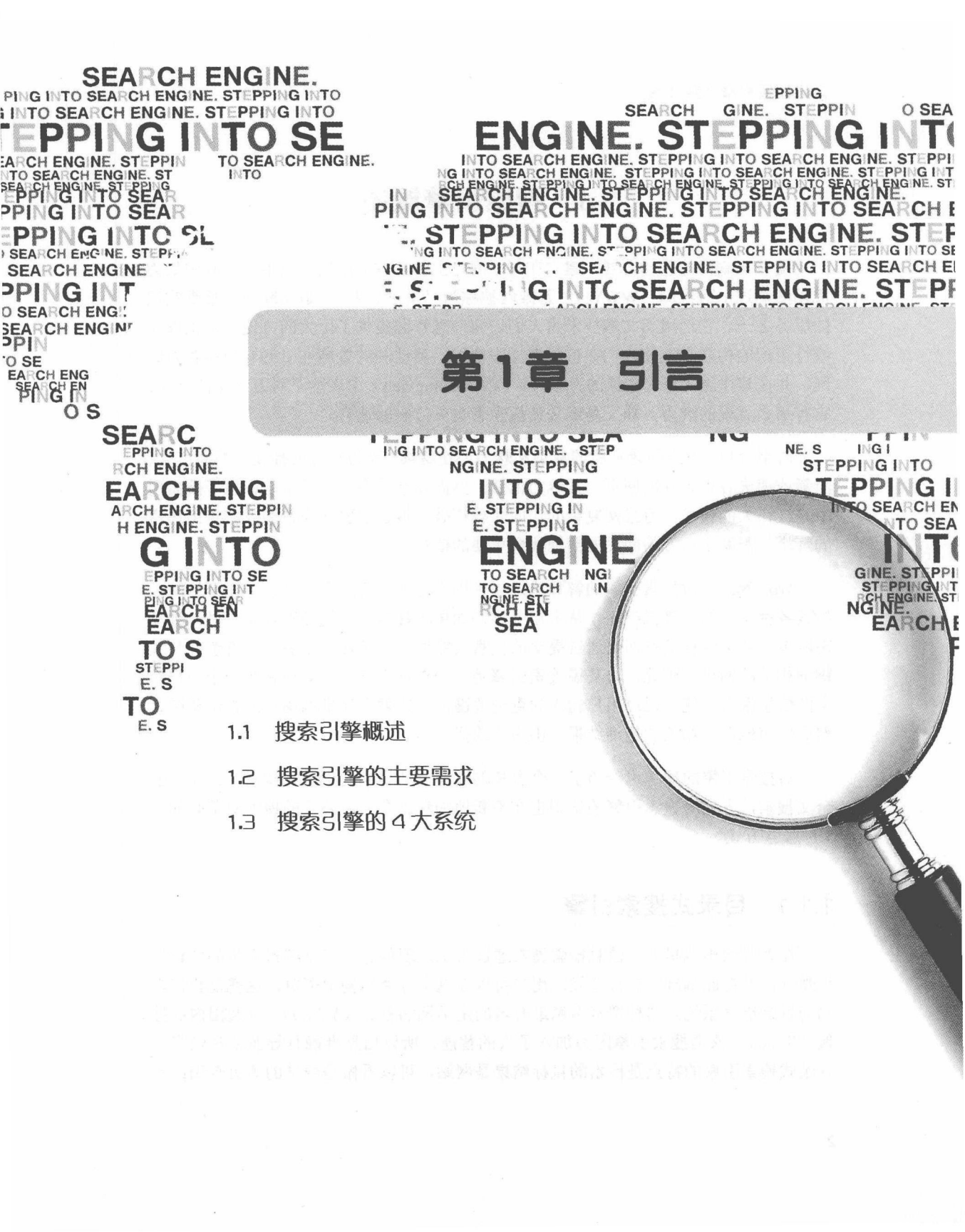
4.7 倒排索引文件的创建过程 107
4.7.1 创建倒排表 107
4.7.2 计算统计信息 109
参考文献 110
第 5 章 搜索引擎的查询系统 112
5.1 知识准备 113
5.1.1 什么是信息熵 113
5.1.2 检索和查询的区别 115
5.1.3 检索词和查询词的区别 115
5.1.4 自动文本摘要（Automatic Text Summarization） 116
5.2 网页信息检索 116
5.2.1 早期的检索模型 116
5.2.2 向量空间模型（Vector Space Models） 118
5.2.3 关键词权重的量化方法 TF/IDF 122
5.2.4 搜索引擎采用的检索模型 125
5.2.5 多文档列表求交计算 127
5.2.6 检索结果排序 132
5.2.7 堆排序 132
5.3 中文自动摘要 137
5.3.1 自动摘要的发展历史 137
5.3.2 自动摘要的含义和实现 137
5.4 生成搜索结果页 142
5.4.1 生成搜索结果页 142
5.5 搜索结果页的缓存 144
5.6 推测用户查询意图 145
5.6.1 查询分类 146
5.6.2 推测信息类、事物类的查询意图 147
5.7 查询系统的当前热点和发展方向 147
5.7.1 查询系统的当前热点 148
5.7.2 查询系统的发展方向 148
参考文献 149

第 6 章　搜索引擎日志分析 …… 150
6.1　简介 …… 151
6.1.1　人机交互的记录——日志 …… 151
6.1.2　分析搜索引擎日志的意义 …… 153
6.1.3　本章的主要内容 …… 154
6.2　知识准备 …… 155
6.2.1　二分图模型（Bipartite Model） …… 155
6.2.2　图模型（graphical model） …… 156
6.2.3　LDA（Latent Dirichlet Allocation）模型 …… 158
6.2.4　随机游走（Random Walk） …… 159
6.2.5　小结 …… 160
6.3　查询日志分析 …… 161
6.3.1　查询日志的内容 …… 161
6.3.2　查询词频统计 …… 162
6.3.3　查询词提示（Suggestion） …… 163
6.3.4　命名实体（Named Entity）类别识别 …… 165
6.3.5　小结 …… 167
6.4　点击日志分析 …… 167
6.4.1　点击日志的内容 …… 168
6.4.2　查询串提示（Suggestion）再分析 …… 169
6.4.3　查询和结果类别属性传递 …… 170
6.4.4　搜索结果相似性度量 …… 171
6.4.5　查询结果排序 …… 172
6.4.6　点击数据的稀疏性 …… 174
6.4.7　小结 …… 176
6.5　隐私问题 …… 177
6.5.1　日志的两面性 …… 177
6.5.2　日志的安全使用 …… 179
6.5.3　小结 …… 179
6.6　本章总结 …… 180
参考文献 …… 180

第 7 章　排序学习（Learning to Rank） …… 183
7.1　排序概述 …… 184
7.2　传统的排序模型 …… 186
7.2.1　查询相关的排序模型 …… 186
7.2.2　查询无关的排序模型 …… 188
7.3　排序学习简介以及研究现状 …… 190
7.3.1　排序学习简介 …… 190
7.3.2　排序学习问题的研究现状 …… 191
7.4　排序学习模型的应用实例 …… 192
7.5　排序学习方法的框架 …… 194
7.5.1　参数设置 …… 194
7.5.2　排序学习方法的框架 …… 195
7.6　评测数据集 …… 196
7.6.1　LETOR 数据集 …… 196
7.6.2　Microsoft Learning to Rank 数据集 …… 197
7.6.3　Yahoo Webscope 数据集 …… 198
7.7　排序学习模型简介 …… 198
7.7.1　实例 …… 199
7.7.2　Pointwise 方法 …… 199
7.7.3　Pairwise 方法 …… 204
7.7.4　Listwise 方法 …… 207
7.7.5　3 种排序方法的对比 …… 210
7.8　排序学习模型性能比较 …… 211
7.8.1　评测方法 …… 211
7.8.2　排序模型性能的比较 …… 215
7.9　排序学习的研究方向 …… 217
7.9.1　标准标注的自动构建 …… 217
7.9.2　排序特征 …… 217
7.9.3　半监督学习/主动学习 …… 218
7.9.4　查询相关的排序模型 …… 218
7.9.5　利用用户行为特征 …… 218
7.10　总结 …… 219
参考文献 …… 219

第 8 章　搜索引擎的性能调优 …… 223
8.1　系统调优概述 …… 224
8.2　瓶颈识别 …… 225
8.3　涉及 CPU 的优化方法 …… 226
8.3.1　上下文切换问题（context switching） …… 227
8.3.2　中断和轮询 …… 228
8.3.3　CPU 的 Affinity 问题 …… 229
8.3.4　流水线问题 …… 229
8.4　涉及内存的优化方法 …… 235
8.4.1　概述 …… 235
8.4.2　对换区 …… 236
8.4.3　cache line …… 240
8.4.4　false sharing 问题 …… 245
8.4.5　内存的锁问题 …… 247
8.4.6　内存库的使用 …… 257
8.5　涉及磁盘的优化方法 …… 262
8.5.1　磁盘 IO 的调度 …… 262
8.5.2　其他常见磁盘参数调优 …… 264
8.5.3　磁盘读写方式 …… 265
8.5.4　文件缓存问题 …… 267
8.5.5　5 分钟法则 …… 269
8.6　涉及网络的优化方法 …… 271
8.6.1　搜索首页，结果页提速方法 …… 271
8.6.2　Web Server 的架构选择 …… 274
参考文献 …… 284

第1章 引言



1.1 搜索引擎概述

1.2 搜索引擎的主要需求

1.3 搜索引擎的 4 大系统

1.1 搜索引擎概述

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。奥地利的鲁施在 1994 年接触万维网，并在其作品《令人吃惊的万维网》（aMAZEingweb）中表达了对万维网的感受：它有那么可观的潜力，却又是经常使探索者丧失方向的迷宫。

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。

WordNet 上对搜索引擎的解释是：一种用来在计算机网络，特别是在万维网上检索各种文件的计算机程序。从本质上讲，如果将搜索引擎的搜索结果看做一种动态网页，那么这种动态网页通过提交的检索关键词聚合了各种重要、有价值并与关键词相关的网页。因此，与其说搜索引擎是一个查询系统，不如说它是一个用户定义的信息聚合系统。通过用户输入的查询关键词，搜索引擎推测用户的查询意图，然后快速地返回相关的查询结果，供用户选择。

对搜索引擎的理解也经历了一个漫长的过程，从早期的目录式搜索，到今天的全文搜索，人们对搜索引擎的认识也在不断地加深。今天，公认的搜索引擎有如下 3 种服务方式。

1.1.1 目录式搜索引擎

在万维网出现早期，信息检索通常通过人工发现信息，依靠编辑人员的知识进行甄别，并在此基础上进行分类。用户可以在这个分类结构中浏览，这就是我们熟知的目录检索系统。这种搜索引擎最有名的是早期的雅虎（Yahoo），以及国内的搜狐（Sohu）。该类搜索引擎因为加入了人的智能，所以信息准确且导航质量较高。目录式搜索引擎的特点是检索的目标结果是网站，可以看做是网站的黄页查询；不

足是数据量有限、更新不及时，并且人工维护成本较高。

1.1.2　全文搜索引擎

全文搜索引擎是针对万维网所有网页进行全文检索的搜索引擎，由下载系统以某种策略自动地在万维网上搜集和发现信息，由索引系统为搜集到的信息建立索引，由查询系统根据用户的查询输入检索索引库，并将查询结果返回给用户。服务方式是面向网页的全文检索服务。该类搜索引擎的优点是信息量大、更新及时，并且无须人工干预；缺点是返回信息过多，有很多无关信息，用户必须从结果中筛选。其代表是谷歌（Google）及百度等第二代商用搜索引擎。

1.1.3　元搜索引擎（Meta-Search Engine）

这类搜索引擎没有自己的数据，而是将用户的查询请求同时向多个搜索引擎递交。然后将返回的结果进行重复排除及重新排序等处理后，作为自己的结果返回给用户。服务方式为面向网页的全文检索。这类搜索引擎的优点是返回结果的信息量大；缺点是不能够充分使用原搜索引擎的功能，用户需要做更多的筛选，其代表是WebCrawler。

上述3种搜索引擎共经历了不到20年的发展历程，然而就是在这短短的时间里，在一代代的搜索技术精英不断地努力下，成就了今天伟大而卓越的搜索引擎技术。其中的很多技术成果也用到了其他领域，创造了巨大的价值。

搜索引擎作为一个系统，它解决的是怎样的问题，主要设计目标是什么，主要分为哪几个部分？接下来我们在搜索引擎总体上进行一些讨论。

1.2　搜索引擎的主要需求

随着万维网上信息爆炸性地增长，传统的搜索方法无法为网民提供有效的搜索服务。万维网的发展迫切地要求一种快速、全面、准确、可靠且代价低廉的信息搜索方法，而具有全文检索的搜索引擎正满足了这5个需求，所以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，下面我们将对搜索引擎中最为主要的5个需求，同样也是搜索引擎的主要特点加以说明。

1.2.1 快

一方面，随着信息化社会的到来，信息可以说是无处不在，人们的日常生活离不开这些有价值的信息；另一方面，人们的生活节奏也在不断地加快，人们应该能够平等地获得这些公众信息。这就要求搜索引擎必须能够存储这些无处不在的信息，并且能够快速地进行信息搜索，满足网民的信息检索需求。

有调查表明，当今公开的搜索引擎的查询速度都在“秒”这个量级以下，商用搜索引擎的查询速度达到毫秒级，并且能够支持大规模用户的同时访问。

影响速度的原因很多，例如分词的效果、索引库的效率、分布查询的处理能力和查询缓存的命中率等，这些将在第 3 章和第 4 章中详细介绍。

1.2.2 全

在传统信息检索（Information retrieval）中，将查全率（Recall）作为衡量检索是否全面的度量指标（查全率也称作召回率），查全率是查询出的相关网页数和全部相关网页数的比率。例如在搜索引擎中查询“XML”，如果世界上包含“XML”这个关键词的网页数为 M，而实际该搜索引擎检索出这 M 条中的 N 条网页，那么查全率为 $N/M \times 100\%$。

是否能查得全，主要取决于网页索引库的大小。如果网页库只包含了2 条XML 的查询结果，即便都检索出来了，查全率也是极低的。可见，索引的网页数越多，越有助于提高查全率。

1.2.3 准

在传统信息检索中，应用查准率（Precision）作为衡量检索是否准确的指标，查准率是检索出的相关文档数与检索出的文档总数的比率。例如在搜索引擎中查询“XML”，在实际检索出的网页数 N 中，只有 P 个网页是与查询“XML”相关（Relavant）的，那么查准率为 $P/N \times 100\%$。

通过图 1-1，可以全面理解查全率和查准率的关系。

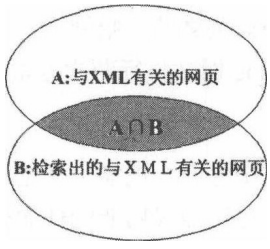


图 1-1　查全率和查准率的关系

查全率=$\frac{|A \cap B|}{|A|}$，其中对集合取||运算的结果表示集合的数量。

查准率=$\frac{|A \cap B|}{|B|}$。

在搜索引擎这种特殊的检索实践中，查全率往往是不重要的。衡量的意义也不大，因为没有一个用户会把所有与查询相关的网页都浏览一遍。一般情况下，用户最为关注的仅仅为搜索结果中的前几条。而查准率在很大程度上决定了搜索的质量，在前10 条搜索结果（搜索结果首页）中满足用户的查询目的，这是搜索引擎查准率的主要体现。

是否能查得准，主要取决于网页排序。常见的有PageRank 等排序方法，在第 3 章中将介绍这方面的内容，在第 7 章中也会做详细介绍。

1.2.4　稳

毫无疑问，搜索引擎必须是一个能够长期并稳定地提供服务的系统，因此系统的稳定运行是很重要的需求。特别是商用搜索引擎，其稳定性被提高到了相当的高度。在任何情况下可以牺牲检索质量和检索速度，但必须能够提供持续的信息检索服务。

对于搜索引擎来说，查询来自四面八方，查询词也千差万别，同时进行的查询量也非常巨大。稳定地满足这些查询需要，需要在系统的结构上做出权衡，在文件存储方式、查询系统和索引系统设计等方面都需要考虑稳定性的因素。

1.2.5　省

由于搜索引擎处理数百亿的网页信息，同时每天接受来自数十亿用户的搜索请