Broadview
www.broadview.com.cn

SEARCH ENGINE'S Principle·Practice·Application

SEARCH ENGINE'S
Principle·Practice·Application

搜索引擎
原理、实践与应用

SEARCH

卢 亮 张博文 编著

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

SEARCH ENGINE'S

Principle·Practice·Application

搜索引擎

原理、实践与应用

卢 亮 张博文 编著

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

内 容 简 介

本书是搜索引擎业界资深的搜索引擎专家长久以来经验的积累与研究工作的心路历程。本书从搜索引擎的历史和现状开始展开，为广大读者展现了搜索引擎简单界面背后的复杂技术、原理和应用，从数据挖掘到搜索引擎的数据结构，从搜索引擎爬虫到分布式搜索引擎的设计均是作者精心研究的结果和过程，对研究搜索引擎的人士颇有实用价值和参考意义。

本书适合刚从事搜索引擎行业乃至互联网行业的从业人员、如网站设计者、程序员、个人网站的站主，本书还适合作为大中专学院相关专业及培训机构培训的参考书籍。

图书在版编目（CIP）数据

搜索引擎原理、实践与应用 / 卢亮，张博文编著. —北京：电子工业出版社，2007.9
ISBN 978-7-121-04810-4

Ⅰ. 搜… Ⅱ. ①卢… ②张… 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 119283 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
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：787×980　1/16　印张：19.25　字数：210 千字
印　　次：2007 年 9 月第 1 次印刷
印　　数：4000 册　定价：42.00 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。

前　言

搜索引擎目前已经成为我们日常上网必备的工具之一，通过搜索引擎我们可以检索到需要的信息，检索可以说是通向互联网世界的大门。Google，百度等这些优秀的搜索引擎已经是我们耳熟能详的名字了。然而搜索引擎拥有较高的技术门槛，目前世界上也只有为数不多的公司能提供搜索引擎服务，不仅如此，在各行各业中，随着企业数据库规模的不断增涨，对搜索引擎技术的需求也日益旺盛。

为什么写作本书

搜索引擎技术是一门融合各方面知识于一体的技术，简单的搜索框背后隐藏着无限的复杂，作者致力于将搜索引擎技术平民化作为目标，撰写此书的目的就是使徘徊在搜索引擎技术大门外的读者入门。

关于本书作者

本书由卢亮先生与张博文先生合著，作者长久以来合作从事搜索引擎研究工作，并且是业界资深的搜索引擎专家，对搜索引擎产业有相当深入地了解，并且拥有丰富的研发经验。此书的灵魂便是作者长久以来经验的积累与研究工作的心路历程。

本书主要内容

第 1 章“搜索引擎的历史和现状” 介绍了搜索引擎的历史背景，使读者能够对搜索引擎现今的发展有一个明确地了解。

第 2 章“数据挖掘”介绍了数据挖掘技术，数据挖掘技术是搜索引擎技术的基础。

第 3 章“搜索引擎的数据结构”介绍了搜索引擎在存储与索引方法方面的基本原理，并且介绍了检索的一般规则。

第 4 章“搜索引擎的基本结构”是本书重要的一个章节，它介绍了搜索引擎的基本结构与组成部分，使读者能够了解搜索引擎技术的全貌，接下来的章节都是围绕此章作深入地研究。

第 5 章“搜索引擎爬虫”是搜索引擎采集数据的手段，本章介绍了搜索引擎爬虫的两种基本策略，并且给出了一个搜索引擎爬虫实例。

第 6 章“搜索引擎索引系统”介绍了搜索引擎中最核心的部分，即索引技术，此章揭示了搜索引擎为何能瞬间检索到数据的原理。并给出了一个索引系统的实例。

第 7 章“分布式搜索引擎设计”介绍了搜索引擎在分布式环境下的设计策略，现今搜索引擎的技术门槛在于大数据量，大访问量的情况下处理问题的能力，所以本书也着力于研究分布式搜索引擎的解决方案，如果读者已经对搜索引擎的实现原理有所了解，那么分布式搜索方案中的一些观点能够对这部分读者起到良好的进阶作用。

第 8 章“Google 搜索引擎的结构”介绍了世界著名搜索引擎 Google 是如何设计的，Google 是我们学习的榜样，学习它的分布式设计对我们也有相当大的启迪作用。

第 9 章“中文分词”。中文分词我们单独作为一章来讲，因为中文分词是实现中文搜索引擎过程中最为基础最为重要的技术。本章介绍了几种常用的分词策略，并给出了一个分词实例供读者学习。

第 10 章“分类与聚类”介绍了文本自动分类与自动聚类技术，这是数据挖掘技术中重要的两项应用，并且给出了一个分类模型的实验，目前互联网上较大的搜索引擎都采用了此两项技术。

第 11 章“内容消重和 SPAM 消除”阐述了屏蔽互联网上垃圾信息的方法，互联网上有大量的垃圾信息，垃圾信息太多会直接的影响搜索引擎的效率与用户的使用感受，并给出了一个典型的 SPAM 案例。

第 12 章“图像搜索引擎”将会介绍的一门较为新颖的检索技术——图像检索，本章将读者的视野从文字扩展到图像，对读者起到较好的进阶作用。

附录提供了本书中用到的系统化数据。

在一些重要章节之后，我们会介绍一些实例，这些实例能帮助作者更好的理解此章节的内容。比如在第 5 章“搜索引擎爬虫”章节中将给出一个实例 Larbin，在第 6 章“搜索引擎索引系统”章节中给出实例 Lucene(一个高效的索引库)和 Booso.com(基于 Lucene 的博客搜索系统)，在第 9 章“中文分词”环节中给出了一个中文自动分词的实例。

无论是初学者还是长久以来关注搜索引擎技术的读者，本书都会帮助你入门和提高。现在就让我们一起开启搜索引擎技术的大门吧！

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。

卢亮　张博文

2007 年 7 月

目　录

第 1 章 搜索引擎的历史和现状 01

长久以来，人类的发展离不开知识获取与发现的过程，进入互联网时代后，信息量开始爆炸性地增长，大量信息扑面而来，然而人们的接受能力却十分有限，此时人们急需一种技术手段，能够使信息的获取更加简单，准确，并且直接找到人们想要的信息，去除杂乱的干扰信息。在此情况下，搜索引擎诞生了，并且经过数十年的发展，目前已经成为人们日常生活中必不可少的工具。

1.1 搜索引擎的历史

众所周知，最早的互联网站点大部分是建立在大学校园里的。1990 年，由于当时并未出现 WWW（World Wide Web，万维网），所以 FTP（File Transportation Protocol，文件传输协议）软件成为共享文件的主要工具。要共享文件，就必须建立一个 FTP 服务器。而要检索 FTP 数据的人就必须使用 FTP 客户端，这样就造成数据零零散散地分布在各个不同的地方，搜索引擎的祖先——Archie——就应运而生，它是由 Montreal 的 McGill University 的学生 Alan

Emtage、Peter Deutsch、Bill Wheelan 发明的。人们使用 Archie 时必须输入精确的文件名进行搜索，Archie 才能告诉用户哪一个 FTP 地址可以下载该文件。

很快，第一个互联网爬虫——Mathew Gray 所开发的 World Wide Web Wanderer——出现了，爬虫(Spider)是用于抓取互联网信息的程序。其实，他最初是想测量 Web 的增长速度而开发这个爬虫来计算 Web 上活动的站点的个数的。但很快，他又升级了这个爬虫来抓取实际的 URL。

1993 年 2 月，Excite 由 6 个斯坦福的学生创建，Excite 从 Architext 项目衍生而来。他们想使用静态统计的方法来分析词之间的关系来使搜索更具效率。

1993 年 10 月，Martijn Koster 创建了 ALIWEB，ALIWEB 允许用户提交他们自己的网页简介信息，这意味着 ALIWEB 不需要通过网络来“爬”数据，也不会使用大量的带宽。然而有许多人不知道如何用 ALIWEB 来提交他们的站点。

1993 年 12 月，三个羽翼丰满的反馈式搜索引擎出现了，分别是 JumpStation，World Wide Web Worm 和 RBSE。JumpStation 与 World Wide Web Worm 类似，基本都是收集页面的标题和头部信息，然后使用简单的线性查找来进行检索。然而随着 Web 规模的迅速增长，JumpStation 由于太慢而终止了。而 RBSE 是第一个能够索引 Html 文件正文的搜索引擎。

1994 年 4 月，David Fili 与 Jerry Yang 创办了 Yahoo，Yahoo 分门别类地收集了一些站点供人们查询，这些站点都是手工收录的。

1994 年 4 月 20 日，华盛顿大学的 Brian Pinkertion 创建了

WebCrawler。WebCrawler 是第一个抓取整个网页全文的爬虫程序。很快，它非常流行，最后 AOL 收购了 WebCrawler，并且让它运行在自己的网络上。1997 年，Excite 购买了 WebCrawler，并且 AOL 开始使用 Excite 来加强自己的 NetFind。WebCrawler 为随后的许多应用打开了大门，在它登场的一年后，Lycos，Infoseek 和 OpenText 便出现了。

Lycos 的出现是搜索引擎历史上重要的一步，它由 Carnegie Mellon 大学于 1994 年 7 月设计。Michale Mauldin 负责搜索引擎的工作，直到现在，他还担任 Lycos 公司的首席科学家。1994 年 7 月 20 日，Lycos 发布了条目数数量为 54 000 个文档的搜索引擎，Lycos 还提供了前缀匹配和字符相近限制，到 1994 年 8 月，Lycos 已经能够搜索 394 000 个文档，到 1995 年，已经有一百五十万的文档，而 1996 年，Lycos 就已经索引了六千万的文档。

Infoseek 是另外一个重要的搜索引擎，起初它与 Lycos，Yahoo 没有多大区别，但其随后增加了很多附加功能，而后又与 Netscape 合作，即为 Netscape 提供搜索服务，当用户点击 Netscape 上的搜索按钮时，直接利用 Infoseek 进行搜索，而以前一直是由 Yahoo 提供该服务的。Infoseek 拥有先进的搜索技术，并且它能够允许用户在 24 小时内任意添加或删除自己的 URL。

1998 年，在今天仍产生巨大影响的 Google 公司诞生了。Google 使用了先进 PageRank 技术。Google 随后开始非常流行，并开始为 AOL 和 Yahoo 提供搜索服务。MSN Search 也在同年诞生。

2000 年，Teoma 搜索引擎诞生，它能够将网站按照其主题的流行程度进行分类，在 2001 年，Ask Jeeves 购买了 Teoma。

2003 年，Google 发布了在线广告 AdSense，让人们把相关的

广告放置在相关页面的任何地方。随后，Google 又陆续推出了更多的服务，详情可见 1.4 节的介绍。【1，2】

知名搜索引擎出现的时间表如表 1.1 所示，大家可对搜索引擎的发展历史一目了然。

表 1-1 知名搜索引擎出现的时间表

年 代	名 称
1990 年	Archie
1993 年 2 月	Excite
1993 年 10 月	ALIWEB
1993 年 12 月	JumpStation，World Wide Web Worm 和 RBSE
1994 年，4 月	Yahoo
1994 年 7 月	Lycos
1998 年	Google
2000 年	Teoma
2001 年	Teoma 被 Ask Jeeves 收购
2003 年	Google 发布 AdSense

1.2 搜索引擎的分类

搜索引擎按其工作方式主要可分为三种：全文搜索引擎、目录索引类搜索引擎和元搜索引擎。

全文搜索引擎是目前流行的大型搜索引擎普遍采用的形式。所谓全文，就是用户可以去搜索一篇文章的任何部分，不论是标题，还是正文，用户得到了更大的自由度。Google（其界面如图 1-1 所示）就是优秀的全文检索引擎。

目录式搜索引擎并不是严格意义上的搜索引擎，它只是将网站