Broadview®
www.broadview.com.cn

这就是搜索引擎
核心技术详解

改变全世界人们生活方式的“信息之门”

GO

张俊林 著

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

这就是搜索引擎

核心技术详解

张俊林 著

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

内 容 简 介

搜索引擎作为互联网发展中至关重要的一种应用，已经成为互联网各个领域的制高点，其重要性不言而喻。搜索引擎领域也是互联网应用中不多见的以核心技术作为其命脉的领域，搜索引擎各个子系统是如何设计的？这成为广大技术人员和搜索引擎优化人员密切关注的内容。

本书的最大特点是内容新颖全面而又通俗易懂。对于实际搜索引擎所涉及的各种核心技术都有全面细致的介绍，除了作为搜索系统核心的网络爬虫、索引系统、排序系统、链接分析及用户分析外，还包括网页反作弊、缓存管理、网页去重技术等实际搜索引擎必须关注的技术，同时用相当大的篇幅讲解了云计算与云存储的核心技术原理。另外，本书也密切关注搜索引擎发展的前沿技术：Google 的咖啡因系统及 Megastore 等云计算新技术、百度的暗网抓取技术阿拉丁计划、内容农场作弊、机器学习排序等。诸多新技术在相关章节都有详细讲解，同时对于社会化搜索、实时搜索及情境搜索等搜索引擎的未来发展方向做了技术展望。为了增进读者的理解，全书大量引入形象的图片来讲解算法原理，相信读者会发现原来搜索引擎的核心技术的理解比原先想象的要简单得多。

本书适合所有对搜索引擎技术感兴趣的人们，尤其对于相关领域的学生、对搜索引擎核心技术感到好奇的技术人员、从事搜索引擎优化的相关人员及中小网站站长等更有参考价值。

图书在版编目（CIP）数据

这就是搜索引擎：核心技术详解 / 张俊林著. —北京：电子工业出版社，2012.1
ISBN 978-7-121-14865-1

Ⅰ. ①这… Ⅱ. ①张… Ⅲ. ①互联网络—情报检索 Ⅳ. ①G354.4

中国版本图书馆 CIP 数据核字（2011）第 214200 号

责任编辑：付　睿
印　　刷：北京东光印刷厂
装　　订：三河市皇庄路通装订厂
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：20　　字数：416 千字
印　　次：2012 年 4 月第 2 次印刷
印　　数：4001~7000 册　　定价：45.00 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到 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

互联网产品形形色色，有产品导向的，有营销导向的，也有技术导向的，但是以技术见长的互联网产品比例相对小些。搜索引擎是目前互联网产品中最具技术含量的产品，如果不是唯一，至少也是其中之一。

经过十几年的发展，搜索引擎已经成为互联网的重要入口之一，Twitter 联合创始人埃文•威廉姆斯提出了“域名已死论”：好记的域名不再重要，因为人们会通过搜索进入网站。搜索引擎排名对于中小网站流量来说至关重要。了解搜索引擎简单界面背后的技术原理其实对很多人都很重要。

为什么会有这本书

最初写本搜索引擎技术书籍的想法萌生于两年前，当时的场景是要给团队成员做搜索技术培训，但是我找遍了相关图书，却没有发现非常合适的搜索技术入门书籍。当时市面上的书籍，要么是信息检索理论方面的专著，理论性太强不易懂，而且真正讲搜索引擎技术的章节并不太多；要么是 Lucene 代码分析这种过于实务的书籍，像搜索引擎这种充满算法的应用，直接分析开源系统代码并不是非常高效的学习方式。所以当时萌生了写一本既通俗易懂，适合没有相关技术背景的人员阅读，又比较全面，且融入最新技术的搜索引擎书籍，但是真正动手开始写是一年前的事情了。

写书前我给自己定了几个目标。首先内容要全面，即全面覆盖搜索引擎相关技术的主要方面，不仅要包含倒排索引、检索模型和爬虫等常见内容，也要详细讲解链接分析、网页反作弊、用户搜索意图分析、云存储及网页去重，甚至是搜索引擎缓存等内容，这些都是一个完整搜索引擎的有机构成部分，但是详述其原理的书籍并不多，我希望能够尽可能全面些。

第二个目标是通俗易懂。我希望没有任何相关技术背景的人也能够通过阅读这本书有所收获，最好是不懂技术的同学也能大致看懂。这个目标看似简单，其实很不容易达到，我也不敢说这本书已经达到了此目的，但是确实已经尽自己所能去做了。至于具体的措施，则包含以下三个方面。

- 一个是尽可能减少数学公式的出现次数，除非不得已不罗列公式。虽说数学公式具简洁之美，但是大多数人其实对于数学符号是有恐惧和逃避心理的，多年前我也有类似心理，所以但凡可能，尽量不用数学公式。
- 一个是尽可能多举例子，尤其是一些比较难理解的地方，需要例子来增进理解。
- 还有一个是多画图。就我个人的经验来说，尽管算法或者技术是很抽象的，但是如果深入理解其原理，去繁就简，那么一定可以把算法转换成形象的图片。如果不能在头脑中形成算法直观的图形表示，说明并未透彻了解其原理。这是我判断自己是否深入理解算法的一个私有标准。鉴于此，本书中在讲解算法的地方，大量采用了算法原理图，全书包含了超过300幅算法原理讲解图，相信这对于读者深入理解算法会有很大的帮助。

第三个目标是强调新现象新技术，比如Google的咖啡因系统及Megastore等云存储系统、Pregel云图计算模型、暗网爬取技术、Web 2.0网页作弊、机器学习排序、情境搜索、社会化搜索等在相关章节都有讲解。

第四个目标是强调原理，不纠缠技术细节。对于新手一个易犯的毛病是喜欢抠细节，只见树木不见森林，搞明白了一个公式却不了解其背后的基本思想和出发点。我接触的技术人员很多，十有七八会有这个特点。这里有个“道术孰优”的问题，何为“道”？何为“术”？举个例子的话，《孙子兵法》是道，而《三十六计》则为术。“道”所述，是宏观的、原理性的、长久不变的基本原理，而“术”则是在遵循基本原理基础上的具体手段和措施，具有易变性。技术也是如此，算法本身的细节是“术”，算法体现的基本思想则是“道”，知“道”而学“术”，两者虽不可偏废，但是若要选择优先级的话，无疑我会选择先“道”后“术”。

以上四点是写书前定下的目标，现在书写完了，也许很多地方不能达到最初的期望，但是尽了力就好。写书的过程很辛苦，起码比我原先想象的要辛苦，因为工作繁忙，所以只能每天早早起床，再加上周末及节假日的时间来完成。也许书中还存在这样那样的缺点，但是我可以无愧地说写这本书是有诚意的。

这本书是写给谁的

如果您是下列人员之一，那么本书就是写给您的。

1. 对搜索引擎核心算法有兴趣的技术人员

- 搜索引擎的整体框架是怎样的？包含哪些核心技术？
- 网络爬虫的基本架构是什么？常见的爬取策略是什么？什么是暗网爬取？如何构建分布式爬虫？百度的阿拉丁计划是什么？
- 什么是倒排索引？如何对倒排索引进行数据压缩？
- 搜索引擎如何对搜索结果排序？
- 什么是向量空间模型？什么是概率模型？什么是 BM25 模型？什么是机器学习排序？它们之间有何异同？
- PageRank 和 HITS 算法是什么关系？有何异同？SALSA 算法是什么？Hilltop 算法又是什么？各种链接分析算法之间是什么关系？
- 如何识别搜索用户的真实搜索意图？用户搜索目的可以分为几类？什么是点击图？什么是查询会话？相关搜索是如何做到的？
- 为什么要对网页进行去重处理？如何对网页进行去重？哪种算法效果较好？
- 搜索引擎缓存有几级结构？核心策略是什么？
- 什么是情境搜索？什么是社会化搜索？什么是实时搜索？
- 搜索引擎有哪些发展趋势？

如果您对其中三个以上的问题感兴趣，那么这本书就是为您而写的。

2. 对云计算与云存储有兴趣的技术人员

- 什么是 CAP 原理？什么是 ACID 原理？它们之间有什么异同？
- Google 的整套云计算框架包含哪些技术？Hadoop 系列和 Google 的云计算框架是什么关系？
- Google 的三驾马车 GFS、BigTable、MapReduce 各自代表什么含义？是什么关系？
- Google 的咖啡因系统的基本原理是什么？
- Google 的 Pregel 计算模型和 MapReduce 计算模型有什么区别？
- Google 的 Megastore 云存储系统和 BigTable 是什么关系？
- 亚马逊公司的 Dynamo 系统是什么？

- 雅虎公司的 PNUTS 系统是什么？
- Facebook 公司的 Haystack 存储系统适合应用在什么场合？

如果您对上述问题感兴趣，相信可以从书中找到答案。

3. 从事搜索引擎优化的网络营销人员及中小网站站长

- 搜索引擎的反作弊策略是怎样的？如何进行优化避免被认为是作弊？
- 搜索引擎如何对搜索结果排序？链接分析和内容排序是什么关系？
- 什么是内容农场？什么是链接农场？它们是什么关系？
- 什么是 Web 2.0 作弊？有哪些常见手法？
- 什么是 SpamRank?什么是 TrustRank?什么又是 BadRank？它们是什么关系？
- 咖啡因系统对网页排名有何影响？

最近有一批电子商务网站针对搜索引擎优化，结果被 Google 认为是黑帽 SEO 而导致搜索排名降权，如何避免这种情况？从事相关行业的营销人员和网站站长应该深入了解搜索引擎反作弊的基本策略和方法，甚至是网页排名算法等搜索引擎核心技术。SEO 技术说到底其实很简单，虽然不断发生变化，但是很多原理性的策略总是相似的，万变不离其宗，深入了解搜索引擎相关技术原理将形成您的行业竞争优势。

4. 作者自己

我的记性不太好，往往一段时间内了解的技术，时隔几年后就很模糊了，所以这本书也是为我自己写的，以作为技术备查手册。沈利也参与了本书的部分编写工作。

致谢

感谢博文视点的付睿编辑，没有她也就没有本书的面世，付编辑在阅稿过程中提出的细致入微的改进点对我帮助甚大。

感谢翻开此书的读者，如果您在阅读本书的过程中发现一些纰漏或者错误，或者是意见建议，希望您能够不吝让我知晓，我会守在 mailjunlin@gmail.com 这个信箱旁敬候您的来信，如果给我微博发信也非常欢迎 http://www.weibo.com/malefactor。

特别感谢我的妻子，在近一年的写作过程中，我几乎把能用的所有业余时间都投入在本书

的写作上，她为了不让我分心，承担了所有的家务，不介意没有时间陪她，这本书的诞生且算是送她的一个礼物吧。

于我而言，这本书的写作是一个辛苦而欣喜的过程，有如旅人远行，涉水跋山之际抬头远眺，总能看到曾经忽略的旖旎丽景，若您在阅读本书的过程中也能有此体会，那就是我的荣幸了。

张俊林

2011 年 6 月

目　录

第 1 章　搜索引擎及其技术架构

“天地玄黄　宇宙洪荒
日月盈昃　辰宿列张
寒来暑往　秋收冬藏
闰馀成岁　律吕调阳
云腾致雨　露结为霜
金生丽水　玉出昆冈”

《千字经》

搜索引擎已经发展为每个人上网都离不开的重要工具，但是为何搜索引擎有着如此重要的地位？其技术发展历程是怎样的？其基本目标是什么？核心问题又是什么？基本技术架构如何？本章内容即给出上述问题的答案，以使读者对搜索引擎有个宏观的理解。

1.1　搜索引擎为何重要

搜索引擎依托于互联网，互联网的蓬勃发展是搜索引擎产品与技术逐步成熟的大背景。离开互联网，搜索引擎将无从谈起。

1.1.1　互联网的发展

20 世纪 90 年代初期是互联网后期获得大规模发展的起爆点，之所以如此，是有其技术背景和社会背景的。

1991 年，Tim Berners-Lee 将超文本的概念引入互联网，同时推出了 WWW 雏形、配套的 HTTP 传输协议及相应的 Web 服务器技术。1993 年，第一个图形浏览器 mosaic 诞生，网页浏览客户端趋于成熟。这些技术与产品为互联网的快速普及和发展做好了技术准备，互联网用户开始从最初的军队和高校等科研机构普及到普通个人用户，为接下来互联网的商业化大规模发