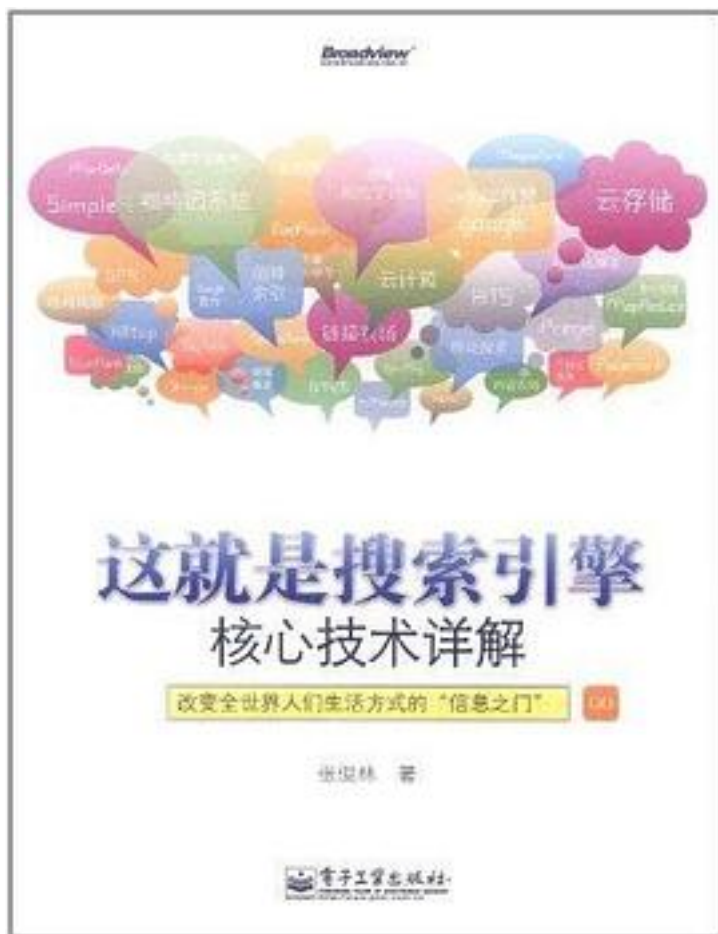


## 这就是搜索引擎



这就是搜索引擎\_下载链接1\_

著者:张俊林

出版者:电子工业出版社

出版时间:2012-1-1

装帧:平装

isbn:9787121148651

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理解起来比原先想象的要简单得多。

作者介绍:

张俊林：本科毕业于天津大学管理学院，2004年于中科院软件所直接获得博士学位并留所从事科研工作，研究方向为搜索引擎与自然语言处理。2005年在CSDN博客发布系列博文“搜索引擎设计实用教程：以百度为例”，在网络上获得了广泛转载与良好口碑。2006年作为联合创始人建立了智能信息聚合网站“玩聚网”，曾先后于阿里巴巴搜索技术中心任资深搜索技术研究员、房价网首席研究员，现任职于新浪微博，从事微博搜索与语义分析及推荐方面的研发工作。

目录: 目 录

第1章 搜索引擎及其技术架构 1

1.1 搜索引擎为何重要 1

1.1.1 互联网的发展 1

1.1.2 商业搜索引擎公司的发展 3

1.1.3 搜索引擎的重要地位 3

1.2 搜索引擎技术发展史 4

1.2.1 史前时代：分类目录的一代 4

1.2.2 第一代：文本检索的一代 5

1.2.3 第二代：链接分析的一代 5

1.2.4 第三代：用户中心的一代 5

1.3 搜索引擎的3个目标 6

1.4 搜索引擎的3个核心问题 7

1.4.1 3个核心问题 7

1.4.2 与技术发展的关系 8

1.5 搜索引擎的技术架构 9

第2章 网络爬虫 12

2.1 通用爬虫框架 12

2.2 优秀爬虫的特性 15

2.3 爬虫质量的评价标准 18

2.4 抓取策略 19

2.4.1 宽度优先遍历策略（Breadth First） 20

2.4.2 非完全PageRank策略（Partial PageRank） 21

2.4.3 OCIP策略（Online Page Importance Computation） 23

2.4.4 大站优先策略（Larger Sites First） 23

2.5 网页更新策略 23

2.5.1 历史参考策略 24

2.5.2 用户体验策略 24

2.5.3 聚类抽样策略 24

2.6 暗网抓取（Deep Web Crawling） 26

2.6.1 查询组合问题 27

|  |    |
|--|----|
| 2.6.2 文本框填写问题                              | 29 |
| 2.7 分布式爬虫                                  | 30 |
| 2.7.1 主从式分布爬虫 (Master-Slave)               | 31 |
| 2.7.2 对等式分布爬虫 (Peer to Peer)               | 31 |
| 本章提要                                       | 34 |
| 本章参考文献                                     | 34 |
| 第3章 搜索引擎索引                                 | 36 |
| 3.1 索引基础                                   | 36 |
| 3.1.1 单词—文档矩阵                              | 37 |
| 3.1.2 倒排索引基本概念                             | 37 |
| 3.1.3 倒排索引简单实例                             | 39 |
| 3.2 单词词典                                   | 42 |
| 3.2.1 哈希加链表                                | 42 |
| 3.2.2 树形结构                                 | 43 |
| 3.3 倒排列表 (Posting List)                    | 44 |
| 3.4 建立索引                                   | 45 |
| 3.4.1 两遍文档遍历法 (2-Pass In-Memory Inversion) | 45 |
| 3.4.2 排序法 (Sort-based Inversion)           | 46 |
| 3.4.3 归并法 (Merge-based Inversion)          | 49 |
| 3.5 动态索引                                   | 50 |
| 3.6 索引更新策略                                 | 51 |
| 3.6.1 完全重建策略 (Complete Re-Build)           | 51 |
| 3.6.2 再合并策略 (Re-Merge)                     | 52 |
| 3.6.3 原地更新策略 (In-Place)                    | 55 |
| 3.6.4 混合策略 (Hybrid)                        | 57 |
| 3.7 查询处理                                   | 57 |
| 3.7.1 一次一文档 (Doc at a Time)                | 58 |
| 3.7.2 一次一单词 (Term at a Time)               | 59 |
| 3.7.3 跳跃指针 (Skip Pointers)                 | 60 |
| 3.8 多字段索引                                  | 62 |
| 3.8.1 多索引方式                                | 62 |
| 3.8.2 倒排列表方式                               | 63 |
| 3.8.3 扩展列表方式 (Extent List)                 | 64 |
| 3.9 短语查询                                   | 64 |
| 3.9.1 位置信息索引 (Position Index)              | 65 |
| 3.9.2 双词索引 (Nextword Index)                | 66 |
| 3.9.3 短语索引 (Phrase Index)                  | 67 |
| 3.9.4 混合方法                                 | 67 |
| 3.10 分布式索引 (Parallel Indexing)             | 68 |
| 3.10.1 按文档划分 (Document Partitioning)       | 69 |
| 3.10.2 按单词划分 (Term Partitioning)           | 70 |
| 3.10.3 两种方案的比较                             | 72 |
| 本章提要                                       | 73 |
| 本章参考文献                                     | 73 |
| 第4章 索引压缩                                   | 76 |
| 4.1 词典压缩                                   | 76 |
| 4.2 倒排列表压缩算法                               | 78 |
| 4.2.1 评价索引压缩算法的指标                          | 79 |
| 4.2.2 一元编码与二进制编码                           | 79 |
| 4.2.3 Elias Gamma算法与Elias Delta算法          | 81 |
| 4.2.4 Golomb算法与Rice算法                      | 81 |
| 4.2.5 变长字节算法 (Variable Byte)               | 83 |
| 4.2.6 SimpleX 系列算法                         | 84 |
| 4.2.7 PForDelta算法                          | 86 |

|  |     |
|--|-----|
| 4.3 文档编号重排序 (DocID Reordering)                 | 89  |
| 4.4 静态索引裁剪 (Static Index Pruning)              | 93  |
| 4.4.1 以单词为中心的索引裁剪                              | 94  |
| 4.4.2 以文档为中心的索引裁剪                              | 96  |
| 本章提要   | 97  |
| 本章参考文献   | 97  |
| 第5章 检索模型与搜索排序                                  | 99  |
| 5.1 布尔模型 (Boolean Model)                       | 101 |
| 5.2 向量空间模型 (Vector Space Model)                | 102 |
| 5.2.1 文档表示                                     | 102 |
| 5.2.2 相似性计算                                    | 104 |
| 5.2.3 特征权重计算                                   | 106 |
| 5.3 概率检索模型                                     | 108 |
| 5.3.1 概率排序原理                                   | 108 |
| 5.3.2 二元独立模型 (Binary Independent Model)        | 110 |
| 5.3.3 BM25模型                                   | 113 |
| 5.3.4 BM25F模型                                  | 115 |
| 5.4 语言模型方法                                     | 116 |
| 5.5 机器学习排序 (Learning to Rank)                  | 119 |
| 5.5.1 机器学习排序的基本思路                              | 120 |
| 5.5.2 单文档方法 (PointWise Approach)               | 121 |
| 5.5.3 文档对方法 (PairWise Approach)                | 122 |
| 5.5.4 文档列表方法 (ListWise Approach)               | 123 |
| 5.6 检索质量评价标准                                   | 125 |
| 5.6.1 精确率与召回率                                  | 126 |
| 5.6.2 P@10指标                                   | 127 |
| 5.6.3 MAP指标 (Mean Average Precision)           | 128 |
| 本章提要   | 129 |
| 本章参考文献   | 129 |
| 第6章 链接分析                                       | 131 |
| 6.1 Web图                                       | 131 |
| 6.2 两个概念模型及算法之间的关系                             | 133 |
| 6.2.1 随机游走模型 (Random Surfer Model)             | 133 |
| 6.2.2 子集传播模型                                   | 135 |
| 6.2.3 链接分析算法之间的关系                              | 136 |
| 6.3 PageRank算法                                 | 137 |
| 6.3.1 从入链数量到PageRank                           | 137 |
| 6.3.2 PageRank计算                               | 138 |
| 6.3.3 链接陷阱 (Link Sink) 与远程跳转 (Teleporting)     | 139 |
| 6.4 HITS算法 (Hypertext Induced Topic Selection) | 140 |
| 6.4.1 Hub页面与Authority页面                        | 140 |
| 6.4.2 相互增强关系                                   | 141 |
| 6.4.3 HITS算法                                   | 142 |
| 6.4.4 HITS算法存在的问题                              | 144 |
| 6.4.5 HITS算法与PageRank算法比较                      | 145 |
| 6.5 SALSA算法                                    | 146 |
| 6.5.1 确定计算对象集合                                 | 146 |
| 6.5.2 链接关系传播                                   | 148 |
| 6.5.3 Authority权值计算                            | 150 |
| 6.6 主题敏感PageRank (Topic Sensitive PageRank)    | 152 |
| 6.6.1 主题敏感PageRank与PageRank的差异                 | 152 |
| 6.6.2 主题敏感PageRank计算流程                         | 153 |
| 6.6.3 利用主题敏感PageRank构造个性化搜索                    | 156 |
| 6.7 Hilltop算法                                  | 156 |

|   |     |
|---|-----|
| 6.7.1 Hilltop算法的一些基本定义                      | 157 |
| 6.7.2 Hilltop算法                             | 158 |
| 6.8 其他改进算法                                  | 162 |
| 6.8.1 智能游走模型 (Intelligent Surfer Model)     | 162 |
| 6.8.2 偏置游走模型 (Biased Surfer Model)          | 163 |
| 6.8.3 PHITS算法 (Probability Analogy of HITS) | 163 |
| 6.8.4 BFS算法 (Backward Forward Step)         | 163 |
| 本章提要  | 164 |
| 本章参考文献                                      | 164 |
| 第7章 云存储与云计算                                 | 166 |
| 7.1 云存储与云计算概述                               | 167 |
| 7.1.1 基本假设                                  | 167 |
| 7.1.2 理论基础                                  | 168 |
| 7.1.3 数据模型                                  | 170 |
| 7.1.4 基本问题                                  | 170 |
| 7.1.5 Google的云存储与云计算架构                      | 171 |
| 7.2 Google文件系统 (GFS)                        | 173 |
| 7.2.1 GFS设计原则                               | 174 |
| 7.2.2 GFS整体架构                               | 174 |
| 7.2.3 GFS主控服务器                              | 176 |
| 7.2.4 系统交互行为                                | 178 |
| 7.3 Chubby锁服务                               | 179 |
| 7.4 BigTable                                | 181 |
| 7.4.1 BigTable的数据模型                         | 181 |
| 7.4.2 BigTable整体结构                          | 183 |
| 7.4.3 BigTable的管理数据                         | 184 |
| 7.4.4 主控服务器 (Master Server)                 | 186 |
| 7.4.5 子表服务器 (Tablet Server)                 | 187 |
| 7.5 Megastore系统                             | 191 |
| 7.5.1 实体群组切分                                | 192 |
| 7.5.2 数据模型                                  | 193 |
| 7.5.3 数据读写与备份                               | 195 |
| 7.6 Map/Reduce云计算模型                         | 195 |
| 7.6.1 计算模型                                  | 196 |
| 7.6.2 整体逻辑流程                                | 197 |
| 7.6.3 应用示例                                  | 198 |
| 7.7 咖啡因系统——Percolator                       | 199 |
| 7.7.1 事务支持                                  | 200 |
| 7.7.2 观察/通知体系结构                             | 202 |
| 7.8 Pregel图计算模型                             | 203 |
| 7.9 Dynamo云存储系统                             | 206 |
| 7.9.1 数据划分算法 (Partitioning Algorithm)       | 207 |
| 7.9.2 数据备份 (Replication)                    | 208 |
| 7.9.3 数据读写                                  | 208 |
| 7.9.4 数据版本控制                                | 209 |
| 7.10 PNUTS云存储系统                             | 210 |
| 7.10.1 PNUTS整体架构                            | 211 |
| 7.10.2 存储单元                                 | 211 |
| 7.10.3 子表控制器与数据路由器                          | 213 |
| 7.10.4 雅虎消息代理                               | 213 |
| 7.10.5 数据一致性                                | 214 |
| 7.11 HayStack存储系统                           | 215 |
| 7.11.1 HayStack整体架构                         | 216 |
| 7.11.2 目录服务                                 | 218 |

|                                    |     |
|------------------------------------|-----|
| 7.11.3 HayStack缓存                  | 219 |
| 7.11.4 HayStack存储系统                | 219 |
| 本章提要                               | 222 |
| 本章参考文献                             | 222 |
| 第8章 网页反作弊                          | 224 |
| 8.1 内容作弊                           | 224 |
| 8.1.1 常见内容作弊手段                     | 225 |
| 8.1.2 内容农场 (Content Farm)          | 226 |
| 8.2 链接作弊                           | 227 |
| 8.3 页面隐藏作弊                         | 230 |
| 8.4 Web 2.0作弊方法                    | 231 |
| 8.5 反作弊技术的整体思路                     | 232 |
| 8.5.1 信任传播模型                       | 233 |
| 8.5.2 不信任传播模型                      | 234 |
| 8.5.3 异常发现模型                       | 234 |
| 8.6 通用链接反作弊方法                      | 236 |
| 8.6.1 TrustRank算法                  | 237 |
| 8.6.2 BadRank算法                    | 238 |
| 8.6.3 SpamRank                     | 239 |
| 8.7 专用链接反作弊技术                      | 240 |
| 8.7.1 识别链接农场                       | 240 |
| 8.7.2 识别Google轰炸                   | 241 |
| 8.8 识别内容作弊                         | 241 |
| 8.9 反隐藏作弊                          | 241 |
| 8.9.1 识别页面隐藏                       | 241 |
| 8.9.2 识别网页重定向                      | 242 |
| 8.10 搜索引擎反作弊综合框架                   | 242 |
| 本章提要                               | 244 |
| 本章参考文献                             | 244 |
| 第9章 用户查询意图分析                       | 246 |
| 9.1 搜索行为及其意图                       | 246 |
| 9.1.1 用户搜索行为                       | 246 |
| 9.1.2 用户搜索意图分类                     | 248 |
| 9.2 搜索日志挖掘                         | 250 |
| 9.2.1 查询会话 (Query Session)         | 250 |
| 9.2.2 点击图 (Click Graph)            | 251 |
| 9.2.3 查询图 (Query Graph)            | 252 |
| 9.3 相关搜索                           | 253 |
| 9.3.1 基于查询会话的方法                    | 253 |
| 9.3.2 基于点击图的方法                     | 254 |
| 9.4 查询纠错                           | 255 |
| 9.4.1 编辑距离 (Edit Distance)         | 256 |
| 9.4.2 噪声信道模型 (Noise Channel Model) | 257 |
| 本章提要                               | 257 |
| 本章参考文献                             | 258 |
| 第10章 网页去重                          | 259 |
| 10.1 通用去重算法框架                      | 261 |
| 10.2 Shingling算法                   | 262 |
| 10.3 l-Match算法                     | 265 |
| 10.4 SimHash算法                     | 268 |
| 10.4.1 文档指纹计算                      | 269 |
| 10.4.2 相似文档查找                      | 270 |
| 10.5 SpotSig算法                     | 272 |
| 10.5.1 特征抽取                        | 272 |

- 10.5.2 相似文档查找 273
- 本章提要 274
- 本章参考文献 274
- 第11章 搜索引擎缓存机制 276
  - 11.1 搜索引擎缓存系统架构 277
  - 11.2 缓存对象 279
  - 11.3 缓存结构 281
  - 11.4 缓存淘汰策略（Evict Policy） 283
    - 11.4.1 动态策略 284
    - 11.4.2 混合策略 284
  - 11.5 缓存更新策略（Refresh Policy） 285
- 本章提要 286
- 本章参考文献 287
- 第12章 搜索引擎发展趋势 288
  - 12.1 个性化搜索 288
  - 12.2 社会化搜索 290
  - 12.3 实时搜索 291
  - 12.4 移动搜索 293
  - 12.5 地理位置感知搜索 294
  - 12.6 跨语言搜索 296
  - 12.7 多媒体搜索 298
  - 12.8 情境搜索 299
  - • • • • [\(收起\)](#)

[这就是搜索引擎\\_下载链接1](#)

标签

搜索引擎

信息检索

互联网

数据挖掘

算法

搜索

计算机

评论

逻辑清晰,作者很用心; 火车上看的,三五六八是重点。

对咱们本行来说没有价值, 该知道的都知道了. 但对入门的人来说是极品.

挺好的入门书

先吐槽，不论装帧，书里边的图例和各种修饰图案就非常不适合出现在一本技术书当中，儒家思想都出来了，作者写好内容就可以了，其实大可不必往书里面加一些这么标新立异的东西，全是违和感好么。说完图就说排版，比word还难看，给文字内容造成的负面影响极大。不论内容，这就是本糟糕极了的书，可以作为出版界的反面教材。如果单论内容，怎么说呢我虽然才开始看相关的内容，但我觉得应该还是有干货的，即便二手的很多。把云存储单独成章我觉得有点突兀，介绍的时候和信息检索的联系又很少。好了负面评价这么多，那我还是给4星是因为好多人都给4星，技术方面又是新手，不敢妄自菲薄。不过还是支持吧。

: G254.4/1224

虽然是对各种技术的摘抄，但摘的不错，阅读起来很流畅，对于初学者来说，比起花大量时间到处查阅各种资料，花一天时间集中精力看完本书还是很划算的

非码农跳着看也能看懂，为数不多的注意英文前后空格的中文书，可惜估计找错出版社了……制作之糙，结尾都没一个。



就像是写综述，别想通过此书学会什么，只能作为一个各种技术的索引，知道有哪些研究领域。对我个人而言比较新颖的就是网页去重，原来没意识到这块也有很多工作。

-----  
图例丰富；隔靴搔痒

-----  
有点过时了，微信读书的排版很极品，但貌似目前也没有很好的搜索引擎中文科普了。

-----  
书丢飞机上了。看过大半

-----  
1

-----  
出乎意料的好

-----  
连科普都看不进去。。

-----  
一本入门书，涵盖工程及算法。

-----  
讲述太啰嗦了，不够精炼。

-----  
比较通俗，可惜对细节涉及不深，每章最后自带paper references

-----  
入门书都给人一种什么东西都是故弄玄虚其实秒懂的错觉

-----  
其实我是用这本书来看有关随机过程的应用滴~~嘿嘿，而且我跟作者在微博上互动很多。  
。。

-----  
科普读物，对于我这种门外汉最合适。

-----  
[这就是搜索引擎\\_下载链接1](#)

## 书评

个人一直都对搜索引擎有着比较高的兴趣，算上这本书，应该一共看过三本搜索引擎相关的书：1. 深入搜索引擎：海量信息的压缩、索引和查询 2. 自己动手写搜索引擎 3. 本书  
第二本书无须多说，我觉得自己动手系列简直就是一坨屎一样的存在，而第一本书非常详细地讲了关于搜索...

-----  
正如作者前言中提到，搜索引擎技术培训，发现缺少相关的入门书，要么太理论，要么太深入，初学者的门槛很高，其实对于初学者来说，最想快速了解到的，这是一个什么东西，做什么，大体流程是什么样，总体有了认识之后，再逐步挑自己喜欢的地方深入的学习和研究，到这时候，那些...

-----  
整本书的风格是广而不深，对搜索引擎相关的很多技术都进行了介绍，每个方面的介绍基本都在一页纸的篇幅，同时在每章后面附带了若干参考文献以供求职欲强的读者查询。  
作者基本做到了浅出，全书没有代码，基本也没有什么数学公式，全部用图来展现，读者即使没有相关基础...

-----  
最近几年一直坚持一个观点，那就是IT类尤其是IT技术类的书籍，只看英文原文的。原文还尽量找非扫描的PDF格式的，为的是可以尽情的标注、记笔记和搜索。后来还弄来一部9寸的kindle dxg。这本书改变了我的看法，行文流畅，插图生动，深入浅出。一周零零散散的读完，一句话概括...



的说明了一下。  
唯一的遗憾是我比较关心的与地理位置相关的搜索没有详细的说明，只在最后一章当做展望...

-----  
首先要说，印刷质量不错，纸张很厚。  
此书通俗易懂，插图也比较多，不过插图也起不到画龙点睛的效果，好像就是为了插图而插图。跟老外的比起来，写作方法还欠缺。  
可能是此书就是科普性质的，广而不精，可以晚上翻看，甚至在公交车上看，作为扩大知识面或者说入门还是很有...

-----  
此书总体上还可以，在讲一个算法是什么的时候下了挺大功夫，例如画了很多图，基本每个算法都举了简单的例子，这点还是不错的，值得鼓励。但是像很多国内的技术书一样，在讲why的时候不怎么给力，当然这个要求比较高，总体上还是值得推荐

-----  
除了索引压缩那一章，都看了。  
不过发现当时看懂了，看后边章节的时候有的前边的就忘了  
想不起细节，还是要和实践结合  
学知识最好的方法就是去用知识，不然看了不久就忘光了  
.....

-----  
[这就是搜索引擎 下载链接1](#)