

信息检索导论（修订版）



[信息检索导论（修订版）_下载链接1](#)

著者:普拉巴卡尔·拉格万 (Prabhakar Raghavan)

出版者:人民邮电出版社

出版时间:2019-7-1

装帧:平装

isbn:9787115514080

本书是信息检索的教材，旨在从计算机科学的视角提供一种现代的信息检索方法。书中从基本概念讲解网络搜索以及文本分类和文本聚类等，对收集、索引和搜索文档系统的设计和实现的方方面面、评估系统的方法、机器学习方法在文本收集中的应用等给出了

最新的讲解。

什么是排序SVM、XML、DNS和LSI? 什么是信息检索中的垃圾信息、隐藏页和门页? MapReduce和其他一些并行运算方法是如何实现由兆字节到百万兆字节的飞跃的? 这些问题你都能从本书中找到答案。本书首次将构建Web搜索引擎的复杂过程以一种清晰的全景方式展现给读者。——Peter Norvig, 计算机科学家, Google研发总监

本书对信息检索这个举足轻重、发展迅猛的领域进行了全面、准确的介绍, 是一本不可多得的教材。——Raymond Mooney, 得克萨斯大学奥斯汀分校教授

本书选材独特, 对信息检索的基础知识和发展方向进行了生动描述。——Jon Kleinberg, 康奈尔大学教授

作者介绍:

【美】 克里斯托夫·曼宁 (Christopher Manning)
计算机科学家, 斯坦福大学教授, 斯坦福大学人工智能实验室主任, ACM会士、AAAI会士、ACL会士。目前的研究目标为计算机如何智能地处理、理解和生成人类语言资料。曼宁博士是深度学习在自然语言处理应用方面的先锋人物, 在树递归神经网络、语义分析、神经机器翻译、深度语言理解等方面均有令业界瞩目的研究成果。

【美】 普拉巴卡尔·拉格万 (Prabhakar Raghavan)
Google高级副总裁, 目前负责谷歌的广告与商业产品、基础设施团队。之前作为Google App和Google Cloud的副总裁, 带领团队做出了突出业绩。在加入Google前任职于Yahoo!, 是Yahoo!实验室的创建者和负责人。拉格万博士毕业于加州大学伯克利分校, 长期担任斯坦福大学计算机科学系顾问教授, 主要研究方向是文本及Web数据挖掘、随机算法等, 是美国国家工程院院士、ACM会士、IEEE会士。

【德】 欣里希·舒策 (Hinrich Schütze)
德国慕尼黑大学信息与语言处理中心主任, 计算语言学家, 斯坦福大学博士。曾在美国硅谷工作多年。

王斌
博士, 小米公司AI实验室NLP方向首席科学家, 前中国科学院信息工程研究所研究员、博导, 中国科学院大学教授。

李鹏 博士, 中国科学院信息工程研究所高级工程师, 硕士生导师。

目录: 第1章 布尔检索 .	1
1. 1 一个信息检索的例子	2
1. 2 构建倒排索引的初体验 .	5
1. 3 布尔查询的处理	8
1. 4 对基本布尔操作的扩展及有序检索	11
1. 5 参考文献及补充读物 .	13
第2章 词项词典及倒排记录表	14
2. 1 文档分析及编码转换 .	14
2. 1. 1 字符序列的生成 .	14
2. 1. 2 文档单位的选择 .	16
2. 2 词项集合的确定	16
2. 2. 1 词条化	16
2. 2. 2 去除停用词	19

- 2. 2. 3 词项归一化 . 20
- 2. 2. 4 词干还原和词形归并 . 23
- 2. 3 基于跳表的倒排记录表快速合并算法 26
- 2. 4 含位置信息的倒排记录表及短语查询 28
 - 2. 4. 1 二元词索引 . 28
 - 2. 4. 2 位置信息索引 . 29
 - 2. 4. 3 混合索引机制 . 31
- 2. 5 参考文献及补充读物 . 32
- 第3章 词典及容错式检索 . 34
 - 3. 1 词典搜索的数据结构 34
 - 3. 2 通配符查询 . 36
 - 3. 2. 1 一般的通配符查询 . 37
 - 3. 2. 2 支持通配符查询的k-gram索引 . 38
 - 3. 3 拼写校正 39
 - 3. 3. 1 拼写校正的实现 . 39
 - 3. 3. 2 拼写校正的方法 40
 - 3. 3. 3 编辑距离 40
 - 3. 3. 4 拼写校正中的 k-gram索引 42
 - 3. 3. 5 上下文敏感的拼写校正 . 43
 - 3. 4 基于发音的校正技术 44
 - 3. 5 参考文献及补充读物 . 45
- 第4章 索引构建 . 46
 - 4. 1 硬件基础 46
 - 4. 2 基于块的排序索引方法 . 47
 - 4. 3 内存式单遍扫描索引构建方法 . 50
 - 4. 4 分布式索引构建方法 . 51
 - 4. 5 动态索引构建方法 . 54
 - 4. 6 其他索引类型 56
 - 4. 7 参考文献及补充读物 . 57
- 第5章 索引压缩 . 59
 - 5. 1 信息检索中词项的统计特性 . 59
 - 5. 1. 1 Heaps定律: 词项数目的估计 61
 - 5. 1. 2 Zipf定律: 对词项的分布建模 . 62
 - 5. 2 词典压缩 63
 - 5. 2. 1 将词典看成单一字符串的压缩方法 63
 - 5. 2. 2 按块存储 64
 - 5. 3 倒排记录表的压缩 . 66
 - 5. 3. 1 可变字节码 . 67
 - 5. 3. 2 γ 编码 68
 - 5. 4 参考文献及补充读物 74
- 第6章 文档评分、词项权重计算及 向量空间模型 76
 - 6. 1 参数化索引及域索引 76
 - 6. 1. 1 域加权评分 78
 - 6. 1. 2 权重学习 79
 - 6. 1. 3 最优权重 g 的计算 80
 - 6. 2 词项频率及权重计算 . 81
 - 6. 2. 1 逆文档频率 . 81
 - 6. 2. 2 tf-idf 权重计算 82
 - 6. 3 向量空间模型 83
 - 6. 3. 1 内积 83
 - 6. 3. 2 查询向量 86
 - 6. 3. 3 向量相似度计算 . 87
 - 6. 4 其他tf-idf 权重计算方法 . 88
 - 6. 4. 1 tf的亚线性尺度变换方法 . 88

6. 4. 2 基于最大值的tf归一化 .	88
6. 4. 3 文档权重和查询权重机制	89
6. 4. 4 文档长度的回转归一化 .	89
6. 5 参考文献及补充读物	92
第7章 一个完整搜索系统中的评分计算	93
7. 1 快速评分及排序 .	93
7. 1. 1 非精确返回前K篇文档的方法 .	94
7. 1. 2 索引去除技术 .	94
7. 1. 3 胜者表 .	95
7. 1. 4 静态得分和排序 .	95
7. 1. 5 影响度排序	96
7. 1. 6 簇剪枝方法 .	97
7. 2 信息检索系统的组成	98
7. 2. 1 层次型索引	98
7. 2. 2 查询词项的邻近性 .	98
7. 2. 3 查询分析及文档评分函数的设计 .	99
7. 2. 4 搜索系统的组成 .	100
7. 3 向量空间模型对各种查询操作的支持	101
7. 3. 1 布尔查询	101
7. 3. 2 通配符查询 .	102
7. 3. 3 短语查询	102
7. 4 参考文献及补充读物 .	102
第8章 信息检索的评价 .	103
8. 1 信息检索系统的评价 .	103
8. 2 标准测试集 .	104
8. 3 无序检索结果集合的评价 .	105
8. 4 有序检索结果的评价方法 .	108
8. 5 相关性判定 .	112
8. 6 更广的视角看评价: 系统质量及用户效用 .	115
8. 6. 1 系统相关问题 .	115
8. 6. 2 用户效用	115
8. 6. 3 对已有系统的改进 .	116
8. 7 结果片段 .	116
8. 8 参考文献及补充读物 .	118
第9章 相关反馈及查询扩展	120
9. 1 相关反馈及伪相关反馈 .	120
9. 1. 1 Rocchio相关反馈算法 .	122
9. 1. 2 基于概率的相关反馈方法	125
9. 1. 3 相关反馈的作用时机	125
9. 1. 4 Web上的相关反馈 .	126
9. 1. 5 相关反馈策略的评价	127
9. 1. 6 伪相关反馈 .	127
9. 1. 7 间接相关反馈 .	128
9. 1. 8 小结	128
9. 2 查询重构的全局方法 .	128
9. 2. 1 查询重构的词汇表工具	128
9. 2. 2 查询扩展	129
9. 2. 3 同义词词典的自动构建	130
9. 3 参考文献及补充读物 .	131
第10章 XML检索	133
10. 1 XML的基本概念	134
10. 2 XML检索中的挑战性问题 .	137
10. 3 基于向量空间模型的XML检索 .	140
10. 4 XML检索的评价	144

- 10. 5 XML检索：以文本为中心与以数据为中心的对比 . 146
- 10. 6 参考文献及补充读物 . 148
- 第 11 章 概率检索模型 150
 - 11. 1 概率论基础知识 . 150
 - 11. 2 概率排序原理 151
 - 11. 2. 1 10 风险的情况 151
 - 11. 2. 2 基于检索代价的概率排序 原理 152
 - 11. 3 二值独立模型 152
 - 11. 3. 1 排序函数的推导 . 153
 - 11. 3. 2 理论上的概率估计方法 155
 - 11. 3. 3 实际中的概率估计方法 156
 - 11. 3. 4 基于概率的相关反馈方法 157
 - 11. 4 概率模型的相关评论及扩展 158
 - 11. 4. 1 概率模型的评论 . 158
 - 11. 4. 2 词项之间的树型依赖 159
 - 11. 4. 3 Okapi BM25: 一个非二值的 模型 160
 - 11. 4. 4 IR中的贝叶斯网络 方法 161
 - 11. 5 参考文献及补充读物 . 162
- 第 12 章 基于语言建模的信息检索模型 163
 - 12. 1 语言模型 . 163
 - 12. 1. 1 有穷自动机和语言模型 163
 - 12. 1. 2 语言模型的种类 . 165
 - 12. 1. 3 词的多项式分布 . 166
 - 12. 2 查询似然模型 . 167
 - 12. 2. 1 IR中的查询似然模型 167
 - 12. 2. 2 查询生成概率的估计 167
 - 12. 2. 3 Ponte和Croft进行的实验 169
 - 12. 3 语言建模的方法与其他检索方法的 比较 . 171
 - 12. 4 扩展的LM方法 172
 - 12. 5 参考文献及补充读物 . 173
- 第 13 章 文本分类及朴素贝叶斯方法 175
 - 13. 1 文本分类问题 . 177
 - 13. 2 朴素贝叶斯文本分类 . 178
 - 13. 3 伯努利模型 . 182
 - 13. 4 NB的性质 183
 - 13. 5 特征选择 . 188
 - 13. 5. 1 互信息 . 188
 - 13. 5. 2 2 统计量 . 191
 - 13. 5. 3 基于频率的特征选择方法 192
 - 13. 5. 4 多类问题的特征选择方法 193
 - 13. 5. 5 不同特征选择方法的比较 193
 - 13. 6 文本分类的评价 . 194
 - 13. 7 参考文献及补充读物 . 199
- 第 14 章 基于向量空间模型的文本 分类 200
 - 14. 1 文档表示及向量空间中的关联度计算 . 201
 - 14. 2 Rocchio分类方法 . 202
 - 14. 3 k近邻分类器 205
 - 14. 4 线性及非线性分类器 . 209
 - 14. 5 多类问题的分类 . 212
 - 14. 6 偏差—方差折中准则 . 214
 - 14. 7 参考文献及补充读物 . 219
- 第 15 章 支持向量机及文档机器学习方法 221
 - 15. 1 二类线性可分条件下的支持向量机 221
 - 15. 2 支持向量机的扩展 . 226

- 15. 2. 1 软间隔分类 . 226
- 15. 2. 2 多类情况下的支持向量机 228
- 15. 2. 3 非线性支持向量机 228
- 15. 2. 4 实验结果 . 230
- 15. 3 有关文本文档分类的考虑 . 231
- 15. 3. 1 分类器类型的选择 232
- 15. 3. 2 分类器效果的提高 233
- 15. 4 ad hoc检索中的机器学习方法 . 236
- 15. 4. 1 基于机器学习评分的简单例子 . 236
- 15. 4. 2 基于机器学习的检索结果排序 . 238
- 15. 5 参考文献及补充读物 . 239
- 第16章 扁平聚类 . 241
- 16. 1 信息检索中的聚类应用 . 242
- 16. 2 问题描述 244
- 16. 3 聚类算法的评价 . 246
- 16. 4 K-均值算法 248
- 16. 5 基于模型的聚类 . 254
- 16. 6 参考文献及补充读物 . 258
- 第17章 层次聚类 . 260
- 17. 1 凝聚式层次聚类 . 260
- 17. 2 单连接及全连接聚类算法 . 263
- 17. 3 组平均凝聚式聚类 . 268
- 17. 4 质心聚类 269
- 17. 5 层次凝聚式聚类的最优性 . 270
- 17. 6 分裂式聚类 272
- 17. 7 簇标签生成 273
- 17. 8 实施中的注意事项 . 274
- 17. 9 参考文献及补充读物 . 275
- 第18章 矩阵分解及隐性语义索引 277
- 18. 1 线性代数基础 277
- 18. 2 词项—文档矩阵及SVD . 280
- 18. 3 低秩逼近 282
- 18. 4 LSI 284
- 18. 5 参考文献及补充读物 . 288
- 第19章 Web搜索基础 289
- 19. 1 背景和历史 . 289
- 19. 2 Web的特性 290
- 19. 2. 1 Web图 291
- 19. 2. 2 作弊网页 293
- 19. 3 广告经济模型 . 294
- 19. 4 搜索用户体验 . 296
- 19. 5 索引规模及其估计 297
- 19. 6 近似重复及搭叠 300
- 19. 7 参考文献及补充读物 . 303
- 第20章 Web采集及索引 . 304
- 20. 1 概述 . 304
- 20. 1. 1 采集器必须提供的功能特点 304
- 20. 1. 2 采集器应该提供的功能特点 304
- 20. 2 采集 . 305
- 20. 2. 1 采集器架构 . 305
- 20. 2. 2 DNS解析 . 308
- 20. 2. 3 待采集URL池 . 309
- 20. 3 分布式索引 311
- 20. 4 连接服务器 312

- 20. 5 参考文献及补充读物 . 314
- 第 21 章 链接分析 . 316
- 21. 1 Web图 316
- 21. 2 PageRank. 318
- 21. 2. 1 马尔科夫链 . 318
- 21. 2. 2 PageRank的计算 . 320
- 21. 2. 3 面向主题的PageRank 322
- 21. 3 Hub网页及Authority网页 325
- 21. 4 参考文献及补充读物. 329
- 参考文献 331
- 索引 . 356
- • • • • [\(收起\)](#)

[信息检索导论 \(修订版\) 下载链接1](#)

标签

搜索引擎

计算机科学

数据库

CS

评论

对于原理讲得很清晰，逐步引导读者实现搜索引擎，加深对信息检索的理解

作为导论可以打100分，其中有一些关于语音搜索的点非常有启发性，对业务价值很大。
Mark: <https://libindic.org/Soundex>

[信息检索导论 \(修订版\) 下载链接1](#)

书评

对于搜索引擎的初学者来说，本书是一本绝对值得阅读的书目。作者从最简单的布尔检索到一个完整的搜索引擎，逐步深入，逐步引导读者思考，对建造一个大型搜索引擎需要用到的架构和算法都有所涉猎，看完后会对搜索引擎有一个大概的认识，对其基本原理也会有所了解。搜索引擎并不...

stanford的IR入门书籍，cmu stanford都在用该书作为IR入门书籍，很nice。在某些章节如果你有统计的基础来看的话，会更容易些。

第一次看到这本书的时候，还是在前年，当时这本书还只是个草稿的电子版，基本上ir所涉及到的内容都有，讲的也比较全面。要是你英文阅读能力还好的话，推荐去读读这本书，肯定会对ir有一个较为全面的了解的。

最重要的收获，是对信息检索系统（搜索引擎）有一个宏观的认识，大体上说，需要从两个维度来看：
第一个是查询维度，它的核心，是两个索引结构；其一是字典，其二是倒排拉链和正排索引；字典的职责，是把 query 变成 term set；期间用到了多种技术，如：语义扩展（同义词、拼...

这本书不错。值得一看。Christopher D. Manning, 1989年毕业于澳大利亚国立大学，1995年获斯坦福大学语言学博士学位，曾先后在卡内基-梅隆大学、悉尼大学教授语言学，1999年起任斯坦福大学计算机科学和语言学副教授，其主要研究方向是统计自然语言处理、信息提取与表示，以及...

作为入门书籍，还不错。分别介绍了信息检索领域的几个重要概念：倒排索引、检索引擎；tf-idf权重计算技术；向量空间模型，信息检索的评价，有序检索结果的评价MAP，ROC曲线，NDCG等等；相关反馈技术，伪相关反馈；概率检索模型，BM25算法；基于语言建模的信息检索模型，各种文...
