

目 录

第1章 引言 (1)

- 1.1 信息检索与搜索引擎 (1)
- 1.2 搜索引擎的历史 (2)
- 1.3 搜索引擎的分类 (3)
- 1.4 搜索引擎的基本架构 (4)
 - 1.4.1 主要性能需求 (5)
 - 1.4.2 总体架构 (6)
- 1.5 搜索引擎的主要组件及其功能 (7)
 - 1.5.1 网络爬虫 (7)
 - 1.5.2 解析器 (8)
 - 1.5.3 索引器 (9)
 - 1.5.4 检索器 (9)
 - 1.5.5 用户交互接口 (10)
- 1.6 开源搜索引擎 (10)
- 本章小结 (12)
- 习题 (13)

第2章 信息采集 (14)

- 2.1 网络爬虫的概述 (14)
 - 2.1.1 网络爬虫的功能特点 (14)
 - 2.1.2 网络爬虫通用架构 (15)
 - 2.1.3 网络爬虫分类 (17)
- 2.2 分布式网络爬虫架构 (18)
 - 2.2.1 主从分布式结构爬虫 (master-slave) (18)
 - 2.2.2 对等分布式结构爬虫 (peer to peer) (19)
- 2.3 信息采集涉及的协议 (20)
 - 2.3.1 URL 规范和 HTTP 协议 (20)
 - 2.3.2 User Agent (21)
 - 2.3.3 Robots 协议 (22)
- 2.4 页面遍历 (23)
 - 2.4.1 宽度优先遍历策略 (23)
 - 2.4.2 深度优先遍历策略 (24)

2.4.3 重要度优先遍历策略	(24)
2.5 页面更新	(25)
2.5.1 网页更新策略	(26)
2.5.2 爬虫更新方式	(27)
2.6 深网抓取	(28)
2.7 开源网络爬虫	(30)
本章小结	(31)
习题	(32)
第3章 文本处理	(33)
3.1 文本信息提取	(33)
3.1.1 网页数据获取	(33)
3.1.2 非网页的数据获取	(36)
3.2 统计语言模型	(36)
3.2.1 N元模型(N-gram)的基本概念	(37)
3.2.2 数据平滑方法	(37)
3.3 英文分词	(39)
3.3.1 词素切分	(39)
3.3.2 词干提取	(40)
3.3.3 去除停用词	(41)
3.4 中文分词	(42)
3.4.1 中文分词概述	(42)
3.4.2 基于词典的机械分词法	(43)
3.4.3 基于统计的分词法	(45)
3.4.4 分词粒度	(46)
3.5 网页去重	(46)
3.5.1 通用去重算法流程	(46)
3.5.2 Shingling 算法	(47)
3.5.3 SimHash 算法	(48)
本章小结	(50)
习题	(51)
第4章 搜索引擎索引构建	(52)
4.1 倒排索引	(52)
4.1.1 倒排索引基础	(52)
4.1.2 词典结构	(54)

4.1.3 倒排表结构	(57)
4.2 建立索引方式	(58)
4.2.1 基于内存的索引构建	(58)
4.2.2 基于排序的索引建立	(60)
4.2.3 基于合并法的索引构建	(61)
4.3 索引更新	(61)
4.4 分布式索引	(63)
4.4.1 数据划分	(63)
4.4.2 冗余和容错	(64)
4.4.3 Elastic Search 的分布式索引	(65)
4.5 索引压缩	(66)
4.5.1 评价压缩算法的指标	(66)
4.5.2 Delta 编码 (D-Gaps)	(66)
4.5.3 无参数间距压缩编码	(67)
4.5.4 参数间距压缩	(69)
4.5.5 高查询性能的编码	(70)
本章小结	(72)
习题	(72)
第5章 基于文本内容的检索模型	(74)
5.1 检索模型概述	(74)
5.2 布尔模型	(75)
5.3 向量空间模型	(76)
5.3.1 文本表示	(76)
5.3.2 查询相关度计算	(79)
5.4 概率检索模型	(81)
5.4.1 概率检索模型概述	(81)
5.4.2 二元独立模型 (binary independent model)	(82)
5.4.3 BM25 模型	(84)
5.4.4 BM25F 模型	(86)
5.5 基于统计语言建模的检索模型	(87)
5.6 机器学习排序	(88)
5.6.1 机器学习排序概述	(88)
5.6.2 单文档方法 (pointwise approach)	(89)
5.6.3 文档对方法 (pairwise approach)	(89)

5.6.4 文档列表方法 (listwise approach)	(90)
5.7 检索质量评价标准	(92)
5.7.1 准确率和召回率	(92)
5.7.2 前 k 个文档的查准率 (P@k)	(93)
5.7.3 平均查准率均值 (mean average precision, MAP)	(94)
5.7.4 NDCG (normalize DCG)	(95)
本章小结	(96)
习题	(96)
第 6 章 基于链接的检索模型	(98)
6.1 Web 图	(98)
6.2 Page Rank 算法	(99)
6.2.1 基于简单模型的 Page Rank 算法	(99)
6.2.2 基于随机冲浪模型的 Page Rank 算法	(102)
6.2.3 主题敏感的 Page Rank	(103)
6.3 HITS 算法	(105)
6.3.1 HITS 算法基本思想	(105)
6.3.2 HITS 算法流程	(107)
6.3.3 HITS 的优势与缺陷	(108)
6.4 SALAS 算法	(109)
6.5 通用链接反作弊方法	(111)
6.5.1 链接作弊方法	(111)
6.5.2 反链接作弊思路	(112)
6.5.3 经典链接反作弊算法	(113)
本章小结	(115)
习题	(115)
第 7 章 查询处理与结果展示	(116)
7.1 查询纠错	(116)
7.1.1 查询纠错概述	(116)
7.1.2 英文纠错	(117)
7.2 搜索智能提示	(120)
7.3 不安全信息过滤	(122)
7.4 查询处理	(125)
7.4.1 “一次一文档”	(125)
7.4.2 “一次一词”	(127)

7.5 结果展示	(128)
7.5.1 页面摘要	(128)
7.5.2 查询结果聚类	(129)
7.6 查询缓存机制	(131)
本章小结	(132)
习题	(133)
第8章 相关反馈与查询扩展	(134)
8.1 相关反馈框架	(134)
8.2 显式相关反馈	(135)
8.2.1 Rocchio 相关反馈算法	(135)
8.2.2 概率相关反馈	(137)
8.2.3 相关反馈策略的评价	(138)
8.3 伪相关反馈	(138)
8.4 隐式反馈	(139)
8.5 查询扩展	(139)
本章小结	(141)
习题	(141)
第9章 分类与聚类	(142)
9.1 文本分类	(142)
9.1.1 文本分类框架	(142)
9.1.2 贝叶斯文档分类	(143)
9.1.3 支持向量机	(146)
9.1.4 特征选择	(148)
9.1.5 评价	(150)
9.2 聚类	(150)
9.2.1 划分聚类	(150)
9.2.2 层次聚类	(152)
9.2.3 评价	(153)
本章小结	(155)
习题	(156)
第10章 基于知识图谱的搜索引擎	(157)
10.1 概述	(157)
10.2 知识图谱的数据获取	(160)
10.3 信息抽取	(161)

10.3.1 实体抽取	(161)
10.3.2 关系抽取	(164)
10.3.3 属性抽取	(166)
10.4 知识融合	(167)
10.4.1 实体对齐	(167)
10.4.2 实体歧义分析	(168)
10.5 知识表示与知识推理	(169)
10.5.1 知识表示	(169)
10.5.2 知识推理	(171)
10.6 基于知识图谱的智能搜索引擎	(173)
10.6.1 基于知识图谱的搜索结构	(173)
10.6.2 查询理解	(174)
10.6.3 自动问答	(176)
本章小结	(177)
习题	(177)
参考文献	(178)