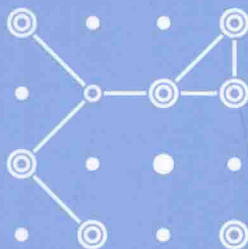


【面向实战，基于大数据分析、知识图谱、
人工智能构建现代搜索引擎】



原理分析及编程实现

刘凡平◎编著



中国工信出版集团



电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
<http://www.phei.com.cn>

大数据搜索引擎 原理分析及编程实现

刘凡平 编著

电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书向读者提供了一套完整的大数据时代背景下的搜索引擎解决方案,详尽地介绍了搜索引擎的技术架构、算法体系及取得的效果,以模块化的方式进行组织。着重介绍了机器学习在搜索引擎中的应用,包括中文分词、聚类、分类等核心的机器学习算法,并结合示例加以介绍和分析,使读者可以更好地理解机器学习在搜索引擎中的价值。还阐述了大数据给搜索引擎带来的新特性,结合目前大数据分析的主流工具,在搜索引擎中构建知识图谱,以及进行日志反馈学习机制,使得搜索引擎更加智能。

本书适合作为互联网行业从业者的技术参考书,也适合作为搜索引擎爱好者的参考读物。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

大数据搜索引擎原理分析及编程实现 / 刘凡平编著. — 北京:电子工业出版社,2016.7

ISBN 978-7-121-29164-7

I. ①大… II. ①刘… III. ①搜索引擎—程序设计IV. ①TP391.3

中国版本图书馆CIP数据核字(2016)第141781号

策划编辑:李 冰

责任编辑:李 冰

特约编辑:田学清 罗树利

印 刷:北京季蜂印刷有限公司

装 订:北京季蜂印刷有限公司

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:720×1000 1/16 印张:20.5 字数:525千字

版 次:2016年7月第1版

印 次:2016年7月第1次印刷

定 价:59.00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)88254888,88258888。

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。

本书咨询联系方式:libing@phei.com.cn。

前言

搜索引擎本身作为一门综合性的互联网技术，在行业中一直具备较高的关注度。随着最近几年大数据的发展，搜索引擎的关注度越来越高，原因在于搜索引擎技术是大数据应用最前线的领域，也是最容易产生价值的大数据应用。大数据存储、大数据计算都是从搜索引擎中衍生出的新领域。目前搜索引擎技术的发展不仅以大数据为基础，还利用分布式实时计算对数据进行高性能处理，以及利用机器学习将数据变得更具价值。在行业中吸引了包括搜索研发工程师、算法研发工程师、大数据分析工程师、自然语言处理工程师、计算平台架构师、数据挖掘工程师等行业各类精英的关注，这些工程师占据了整个互联网研发体系的 50%~60%，在 BAT 中，甚至超过 60% 的是研发人员。

这类群体薪资水平处于互联网研发人员中较高水平，以猎聘网公布的数据显示，北京地区搜索引擎研发工程师年薪为 35 万~60 万元，大数据工程师年薪为 20 万~25 万元，大数据架构师年薪为 40 万~70 万元，等等。也正是由于薪资水平处于较高位，很多互联网相关从业者也积极关注大数据搜索引擎领域动态。

本书通过介绍大数据下的自然语言处理框架、大数据存储引擎、搜索引擎的分布式实时计算、高性能可扩展爬虫，以及利用大数据构建知识图谱、基于大数据日志的搜索引擎反馈学习等相关信息，不仅使读者对当代搜索引擎研发体系有一定的认识，还可以使读者在搜索引擎领域及大数据领域进行深入思考。

本书特色

本书以当前搜索引擎主流技术为基础，密切结合前沿技术发展趋势，行文



通俗易懂，由初步的原理性了解到各模块应用示例，并结合分布式存储、实时计算等，向读者提供了一套完整的大数据时代背景下人工智能搜索引擎的解决方案。

（1）内容循序渐进、行文有条不紊地介绍搜索引擎知识。

本书充分考虑了不同层次的读者对搜索引擎的理解程度，因此本书由简入深、独特的技术写作视角符合广大读者对于技术类读物的理解需求，使得读者能够在掌握搜索引擎基础的情况下，不断按照搜索引擎的设计深入理解。

（2）技术前瞻性强，注重最新主流技术在现代搜索引擎中的应用。

本书充分利用了最新技术发展的应用成果，在自然语言处理的基础上不仅结合大数据分析，还包括分布式计算、机器学习、知识图谱等当前大数据应用与分析处理的主流技术，摒弃了传统过时的研发体系及算法。本书中相关研发成果在当前甚至在未来3~5年，都具有实际意义。

（3）将技术理论与应用范例结合，具备较高的商业实用价值。

本书内容紧密结合当前一线工程师工作研究成果，将众多的技术理论以实际工作经验的方式展示应用效果。本书介绍的内容也广泛结合工作中的应用示例，并以搜索引擎工程实践的脉络流程介绍技术要点，使读者在短时间内能够掌握当前搜索引擎研发的技术理论。

本书结构

本书按照由浅入深、循序渐进的顺序对现代搜索引擎原理和实现进行介绍。全书分为10章，各章的主要内容如下。

第1章针对搜索引擎发展的过去、现在、未来的相关概要介绍，以及现代搜索引擎与大数据、人工智能的相互关系，使广大读者能够在了解现代搜索引擎的背景之下，去了解本书的后续内容。

第2章是对搜索引擎原理与技术的初步分析,从模块方面大致介绍爬虫、索引、缓存等;从技术方面大致介绍自然语言处理、知识图谱技术、海量数据存储、分布式计算等。目的是使得读者对搜索引擎的体系结构、部分技术有一定认识,便于读者深入了解后续章节。

第3章从自然语言角度开始深入分析原理和实现,自然语言是搜索引擎进行文本处理的基础,其中包括分词、词性分析、语义分析、关键词抽取、核心句抽取、聚类分类等。读者将会从本章中获得当前主流的自然语言处理技术相关知识。

第4章主要是针对大数据存储引擎的介绍。大数据存储是搜索引擎最先遇到的问题,解决数据存储问题可以使搜索引擎在数据分析、索引构建、知识图谱等工作持续进行。读者在本章会了解到大数据存储引擎的架构体系、数据模型、数据压缩、负载均衡等。

第5章介绍了分布式实时计算。由于搜索引擎处理的是海量数据,数据分析必须依靠具有较强数据处理能力的计算平台,因此搜索引擎通过分布式实时计算去处理大数据并在尽可能短的时间内返回处理结果。本章中,读者会了解到分布式实时计算设计架构、负载均衡及通信设计等相关知识。

第6章对爬虫进行了深入分析。读者在本章中将会深入理解分布式可扩展爬虫的体系架构,以及对网页如何进行解析,并抽取出结构化的数据信息。本章还涉及链接去重、网页去重、广告识别等相关算法原理。

第7章详细介绍了知识图谱构建。知识图谱是智能化搜索引擎重要的组成部分,利用大数据分析构建出较为合理的知识图谱信息是当前主流的方式。读者将会从本章中深入了解到知识图谱的详细构建过程,以及利用机器学习原理对知识图谱中的实体抽取、关系抽取等相关技术进行。

第8章详细分析了索引构建机制。索引的设计与构造是搜索引擎能够进行



快速检索的核心要件，主要针对文件检索的倒排索引与用于智能提示的字典树索引。本章不仅对倒排索引做了深入分析，对倒排索引的压缩、分布式存储等也做了详细介绍。

第9章深入分析了搜索引擎的整个对外服务工作流程。包括大数据分布式缓存、搜索智能提示、个性化搜索、图片搜索、搜索与广告等。读者通过本章可以详细了解到文本纠错算法、动态摘要算法、网页排序算法及搜索引擎的评价体系。

第10章探讨和分析了搜索引擎日志与搜索引擎本身的关系。搜索引擎日志记录了用户与搜索系统交互的整个流程。通过日志挖掘，不仅可以发现用户的自有特征和行为规律，还可以有效地帮助搜索引擎提升性能和效果。日志作为搜索引擎的核心数据之一，一直使搜索引擎技术中的各类算法不断向前发展。读者通过本章将学会通过搜索引擎日志分析用户特征、用户的部分搜索意图等相关知识。

读者对象

- 适合对自然语言处理及机器学习应用领域有兴趣的读者。
- 适合对现代搜索引擎相关算法有兴趣的读者。
- 适合对大数据分析、数据挖掘应用有兴趣的读者。
- 适合互联网行业的不同层次从业者。
- 适合从事搜索引擎优化的网络营销读者。
- 适合高校中学习计算机、软件工程等相关专业的读者。

目 录

第 1 章 引论	1
1.1 搜索引擎的过去	1
1.2 搜索引擎的现在	2
1.3 搜索引擎的未来	4
1.4 大数据与搜索引擎	6
1.4.1 搜索价值提升	6
1.4.2 用户价值提升	7
1.5 大数据与人工智能	7
1.5.1 人工智能发展	7
1.5.2 人工智能技术	9
1.6 本章小结	11
第 2 章 搜索引擎原理与技术	12
2.1 基本工作原理	12
2.2 基本模块结构	13
2.2.1 爬虫服务	14
2.2.2 索引服务	15
2.2.3 缓存服务	16
2.2.4 搜索服务	17



2.2.5	日志服务	19
2.3	技术概要	20
2.3.1	自然语言处理	20
2.3.2	知识图谱技术	21
2.3.3	海量数据存储	23
2.3.4	分布式计算	25
2.3.5	搜索排序技术	26
2.4	本章小结	27
第 3 章	自然语言处理框架	28
3.1	英文分词	28
3.2	中文分词	30
3.2.1	中文分词概述	30
3.2.2	基于词库的分词技术	31
3.2.3	基于条件随机场的中文分词	33
3.2.4	分词粒度	41
3.3	词性标注	41
3.3.1	隐马尔科夫模型概要	42
3.3.2	隐马尔科夫模型与词性标注	43
3.4	语义相似度	51
3.5	依存句法分析	53
3.5.1	依存句法分析概要	53
3.5.2	依存句法分析实现	56

3.6	情感倾向分析	59
3.7	文档关键词抽取	61
3.7.1	关键词抽取概述	61
3.7.2	基于 TF-IDF 算法	62
3.7.3	基于 TextRank 算法	64
3.8	文档句子相似度分析	67
3.8.1	句子相似度	68
3.8.2	文档相似度	70
3.9	文档核心句抽取	71
3.10	聚类分类	74
3.10.1	文本分类	75
3.10.2	文本聚类	80
3.11	语种检测	84
3.12	本章小结	87
第 4 章	构建大数据存储引擎	88
4.1	架构体系	89
4.1.1	结构概要	89
4.1.2	服务器上线	92
4.1.3	服务器下线	92
4.1.4	数据读取	93
4.2	数据模型	94
4.3	数据压缩	96



4.4	负载均衡	97
4.5	数据存储逻辑视图	100
4.6	本章小结	103
第 5 章	构建分布式实时计算	104
5.1	概述	104
5.2	设计架构	106
5.2.1	设计思想	106
5.2.2	基本框架	108
5.3	运行模式	110
5.4	负载均衡	111
5.5	通信设计	112
5.5.1	基本方式	113
5.5.2	分布式远程服务调用	113
5.6	容灾恢复	114
5.7	数据容错原理	115
5.8	数据处理设计示例	117
5.9	本章小结	118
第 6 章	分布式可扩展爬虫	119
6.1	爬虫体系架构	119
6.1.1	主从分布式结构爬虫	120
6.1.2	对等分布式结构爬虫	120

6.1.3 基于分布式计算平台爬虫	121
6.2 网页解析	122
6.2.1 状态码处理	123
6.2.2 链接去重	123
6.2.3 广告识别	125
6.2.4 网站地图	128
6.2.5 非网页数据获取	129
6.2.6 网页去重	130
6.2.7 链接提取	134
6.2.8 爬虫协议	135
6.3 网页结构化	137
6.3.1 网页的编码信息	137
6.3.2 网页的正文信息	138
6.3.3 网站的关键词信息	142
6.3.4 网站的标题	142
6.3.5 网页的发布时间	144
6.3.6 网站语言检测	144
6.3.7 其他结构化数据	145
6.4 网页抓取策略	146
6.5 爬虫权限应对	147
6.6 深网抓取	150
6.7 抓取更新策略	151
6.8 本章小结	153



第 7 章 大数据构建知识图谱	154
7.1 概述	154
7.2 搜索引擎与知识图谱	155
7.3 可靠数据源选择	157
7.4 实体抽取	158
7.5 关系抽取	159
7.5.1 关系抽取概述	160
7.5.2 隐藏关系抽取	161
7.5.3 结构化确定关系抽取	164
7.5.4 非结构化确定关系抽取	166
7.6 知识图谱检测	171
7.6.1 实体关系修正	171
7.6.2 实体对齐整合	172
7.6.3 实体歧义分析	174
7.7 知识推理与计算	175
7.7.1 知识推理	175
7.7.2 知识计算	176
7.8 知识聚类	179
7.9 智能搜索实现	181
7.9.1 模式匹配	181
7.9.2 知识拆解	182
7.9.3 合并求解	184

7.10 智能搜索扩展	186
7.10.1 常识性智能搜索	186
7.10.2 实时信息智能搜索	187
7.10.3 可交互式智能搜索	187
7.11 本章小结	189
第 8 章 索引构建机制	190
8.1 倒排索引	190
8.1.1 倒排索引概述	191
8.1.2 索引结构	192
8.1.3 构建过程	194
8.1.4 排序规则	195
8.1.5 索引压缩	196
8.1.6 更新策略	202
8.2 分布式存储	202
8.2.1 存储划分方式	203
8.2.2 存储平衡策略	204
8.3 存储索引	209
8.3.1 二叉搜索树	210
8.3.2 B 树	211
8.3.3 B ⁺ 树	213
8.3.4 B ⁺ 树与文件索引	214
8.4 字典树索引	216
8.4.1 字典树索引概述	217



8.4.2	字典树索引构建	219
8.4.3	字典树查询优化	221
8.5	本章小结	221
第 9 章	搜索服务构建	223
9.1	概述	223
9.1.1	体系结构	223
9.1.2	七何分析法	224
9.1.3	搜索语法	225
9.1.4	相关性排序	227
9.1.5	不安全信息过滤	231
9.2	大数据分布式缓存	235
9.2.1	缓存结构设计	235
9.2.2	缓存更新策略	236
9.3	文本纠错算法	237
9.3.1	中文文本纠错	237
9.3.2	英文文本纠错	241
9.4	结果显示算法	242
9.4.1	动态摘要	243
9.4.2	关键词高亮算法	246
9.4.3	网页快照	250
9.5	搜索智能提示	250
9.6	网页排序	254

9.6.1	基于 PageRank 的网页重要性评价	254
9.6.2	基于 Hits 算法的网页权威性评价	257
9.6.3	Hilltop 算法	259
9.6.4	网页作弊评价	260
9.6.5	网页排序调试	263
9.7	个性化搜索	264
9.7.1	个性化搜索示例	264
9.7.2	人工神经网络与个性化搜索	265
9.7.3	地理位置搜索	266
9.8	图片搜索	271
9.8.1	基于内容的图片搜索	271
9.8.2	基于文本的图片搜索	272
9.9	搜索与广告	274
9.9.1	广告投放策略	275
9.9.2	基于 User-Based 协同过滤的广告投放	275
9.9.3	基于 Item-Based 协调过滤的广告投放	277
9.9.4	基于混合模式广告投放	278
9.9.5	广告投放评价	279
9.10	搜索引擎评价	282
9.10.1	搜索评价概述	282
9.10.2	基于准确率、召回率及 F 值评价	283
9.10.3	归一化折扣累计增益	285
9.11	本章小结	288



第 10 章 基于用户日志的反馈学习	290
10.1 基于用户搜索词语的分析	290
10.1.1 发现搜索词的价值	291
10.1.2 发现不明意图下的用户行为	292
10.2 基于用户点击日志的分析	293
10.2.1 时间与搜索意图的关系	293
10.2.2 地理位置与搜索意图的关系	294
10.2.3 点击日志与同义词	296
10.2.4 点击日志与词语权重	297
10.2.5 点击日志与新词分类	298
10.2.6 点击日志与知识图谱	300
10.2.7 点击日志与网页重排序	301
10.2.8 点击日志与网页评价	303
10.3 基于用户的特征分析	304
10.3.1 用户跟踪	305
10.3.2 用户群体特征	306
10.3.3 用户个体特征	308
10.4 本章小结	309