信息科学技术学术著作丛书

大数据搜索与挖掘

张华平　高　凯
黄河燕　赵燕平　著

科学出版社

信息科学技术学术著作丛书

大数据搜索与挖掘

张华平　高　凯　黄河燕　赵燕平　著

科学出版社

北　京

内 容 简 介

本书以作者团队十余年在大数据搜索与挖掘领域所作的研究与应用工作为内容，介绍大数据搜索与挖掘的研发成果，内容涵盖大数据处理概论、中文自然语言处理、网络信息预处理、网络情报挖掘(包括网络语言分析、新特征语言抽取、汉语词法分析、文本自动分类、自动聚类、自动摘要、关键词抽取)、网络精准搜索(信息检索模型、句子搜索)、人物搜索等方面的研究成果。从大数据这座金矿中挖掘有价值的信息，是本书的目的所在。全书体系完整，内容新颖，条理清晰，组织合理，理论与实践并重，突出原创的研究成果与实际应用。

本书可为高校计算机专业、计算机语言学专业和人工智能专业等师生的教学和科研工作提供帮助，也可为从事大数据搜索与挖掘、中文自然语言处理、信息检索与搜索引擎技术研发的工程技术人员和希望了解上述技术的爱好者等提供参考。

图书在版编目(CIP)数据

大数据搜索与挖掘/张华平等著. —北京：科学出版社，2014

(信息科学技术学术著作丛书)

ISBN 978-7-03-040318-6

Ⅰ. 大… Ⅱ. 张… Ⅲ. ①情报检索-研究②数据采集-研究

Ⅳ. ①G354②TP274

中国版本图书馆 CIP 数据核字(2014)第 060074 号

责任编辑：魏英杰 / 责任校对：彭 涛

责任印制：张 倩 / 封面设计：陈 敬

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新科印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2014 年 5 月第 一 版 开本：720×1000 1/16

2014 年 5 月第一次印刷 印张：19 1/2

字数：391 000

定价：90.00 元

(如有印装质量问题，我社负责调换)

作 者 简 介

张华平 1978年出生。工学博士，北京理工大学副教授。毕业于中国科学院计算技术研究所。汉语词法分析系统ICTCLAS创始人，ICTCLAS在国家973评测和第一届国际汉语分词大赛中综合得分均获得第1名。

主要从事大数据搜索与挖掘、自然语言处理、信息检索等方面的研究工作，主持或参与国家自然科学基金、863、973、242等十余项课题。曾先后获得2010年度钱伟长中文信息处理科学技术奖一等奖，中国科学院院长优秀奖、中国科学院计算技术研究所所长特别奖，是中国科学院计算技术研究所“百星计划”首批入选者。

高凯 1968年出生。工学博士。毕业于上海交通大学计算机应用技术专业，河北省重点学科“计算机软件与理论”中“信息检索与云计算”方向学术带头人。

主要从事大数据搜索与挖掘、自然语言处理、网络信息检索、社会网络计算等领域的研究工作。

黄河燕　1963 年出生。工学博士，教授、博士生导师，现任北京理工大学计算机学院院长、国家高技术研究发展计划（863 计划）主题专家组成员、教育部计算机专业指导委员会委员、中国人工智能学会副理事长、中国中文信息学会副理事长兼自然语言处理专业委员会主任。

主要从事自然语言处理和机器翻译、智能处理系统等领域的研究，承担了近 20 项国家级科研攻关项目和大型工程应用，以及国际合作项目，获得国家科学技术进步奖一等奖、国家经济贸易委员会九五技术创新优秀项目奖、中央国家机关十大杰出青年等荣誉和奖励。

赵燕平　1956 年出生。北京理工大学教授，国家人力资源和社会保障部职业技能鉴定中心电子商务专业委员会专家，中国电子学会健康物联专委会专家。北京理工大学大数据搜索与挖掘实验室副主任，曾任联合国开发计划署（UNDP）“中国可持续发展网络计划”项目专家。主持参与了多个科研和工程项目。

《信息科学技术学术著作丛书》序

21世纪是信息科学技术发生深刻变革的时代，一场以网络科学、高性能计算和仿真、智能科学、计算思维为特征的信息科学革命正在兴起。信息科学技术正在逐步融入各个应用领域并与生物、纳米、认知等交织在一起，悄然改变着我们的生活方式。信息科学技术已经成为人类社会进步过程中发展最快、交叉渗透性最强、应用面最广的关键技术。

如何进一步推动我国信息科学技术的研究与发展；如何将信息技术发展的新理论、新方法与研究成果转化为社会发展的新动力；如何抓住信息技术深刻发展变革的机遇，提升我国自主创新和可持续发展的能力？这些问题的解答都离不开我国科技工作者和工程技术人员的求索和艰辛付出。为这些科技工作者和工程技术人员提供一个良好的出版环境和平台，将这些科技成就迅速转化为智力成果，将对我国信息科学技术的发展起到重要的推动作用。

《信息科学技术学术著作丛书》是科学出版社在广泛征求专家意见的基础上，经过长期考察、反复论证之后组织出版的。这套丛书旨在传播网络科学和未来网络技术，微电子、光电子和量子信息技术、超级计算机、软件和信息存储技术，数据知识化和基于知识处理的未来信息服务业，低成本信息化和用信息技术提升传统产业，智能与认知科学、生物信息学、社会信息学等前沿交叉科学，信息科学基础理论，信息安全等几个未来信息科学技术重点发展领域的优秀科研成果。丛书力争起点高、内容新、导向性强，具有一定的原创性；体现出科学出版社“高层次、高质量、高水平”的特色和“严肃、严密、严格”的优良作风。

希望这套丛书的出版，能为我国信息科学技术的发展、创新和突破带来一些启迪和帮助。同时，欢迎广大读者提出好的建议，以促进和完善丛书的出版工作。

中国工程院院士
原中国科学院计算技术研究所所长
李国杰

序

随着以微博、Twitter 与 Facebook 等为代表的新型社交网络的迅猛发展，大数据已经成为新一轮科技革命中信息技术发展的新趋势，将对国家经济与社会产生重大的影响。大数据分为结构化大数据与非结构化大数据，结构化大数据在 20 世纪末期已经有较好的研究，产生了关系型数据库以及数据仓库等相对成熟的技术与产品，也造就了 IBM、Oracle 等跨国大企业。以文本、多媒体信息为主要元素的非结构化大数据，体量是结构化大数据的十倍甚至更大，但非结构化大数据的分析处理挑战更大。非结构化大数据的生产、搜索、挖掘与分析已经成为了当前研究的热点与难点。

《大数据搜索与挖掘》一书顺应了非结构化文本大数据搜索与挖掘的内在要求，作者结合十余年的科学研究与工程技术开发的积累，有针对性地分享了他们的研究成果与工程经验。该书是目前有关文本大数据搜索与挖掘领域比较全面和系统的著作，也是诸多大数据挖掘书籍中为数不多的穿插大量实践应用案例和场景的著作。该书全方位整理、总结、分享了学术思想与项目经验，并提供了相关程序代码和技术支持；通过案例分享，更好地为相关专业的学习者与研究者提供帮助。有理由相信，该书能对有志于从事大数据搜索与挖掘的学者与工程师提供理论和实践支撑。

张华平博士自主研发的 ICTCLAS 是目前最受欢迎、应用最广泛的中文词法分析系统，已经成为中文信息处理领域的标杆。张华平博士是国内大数据搜索与挖掘方面年轻活跃的学者，在学术界与产业界有着重要的影响，包括中央电视台在内的媒体均进行过报道。在这部著作中，张华平所在团队结合他们在自然语言处理、大数据搜索与挖掘、数据运营实践中积累的大量项目经验，围绕大数据文本挖掘中的思路、方法、技巧与应用，提出了知著、显微、晓意的大数据解决之道，与我多年在信息处理尤其是少数民族语言的信息处理方面有着异曲同工之妙。

这是一本难得一见的理论联系实际的学术著作，必将对国内大数据搜索与挖掘的研究和发展起到积极的推动作用。

中国工程院院士

吾守尔·斯拉木

前　言

随着互联网的迅猛发展和信息传播手段的丰富，人类已经进入一个信息爆炸的大数据时代。面对日益增长的网络大数据，高效快捷地获取有用信息，已经成为时代发展的迫切需要。本书以作者及其研发团队十余年来在大数据搜索与挖掘领域所作的研究与应用工作为主线，介绍大数据搜索与挖掘的基本原理，就网络信息智能处理技术中的 Web 数据挖掘、自然语言处理（如分词、词典管理、主题词抽取、摘要、分类、聚类、有意义串挖掘等）技术、信息检索与知识发现等进行阐述，并着重从中文自然语言处理的角度来提高信息检索与挖掘的质量，力争开发出性能优异的大数据搜索与挖掘系统。

全书突出原创性的研究成果，理论与实践并重，强调算法、技术实现与实际应用，其主要内容涉及作者团队近年来的研究成果，囊括了大数据搜索、中文自然语言处理、网络信息预处理、网络情报挖掘（包括网络语言分析、新特征语言抽取、汉语词法分析、文本自动分类、自动聚类、自动摘要、关键词抽取）、网络精准搜索（包括信息检索模型、句子搜索）等方面的研究成果，并从实际应用出发，介绍作者参与研发的科研成果及在相关部门的应用。

全书成果主要涉及张华平及其指导的研究生的科研成果，有些章节内容直接来自成果论文；高凯完成了有关算法、网络信息获取、搜索引擎等内容以及部分相关工作综述、扩展阅读等的撰写工作；黄河燕教授审阅了全书；赵燕平教授完成了最后的统稿工作。在这里，对导师以及相关工作的合作者白硕研究员、刘群教授、程学旗教授等表示衷心感谢。在本书的写作与相关科研课题的研究工作中，得到了多方面的支持与帮助，并参考了作者指导的部分研究生（贺敏、张京阳、王思力、黄玉兰、秦鹏、刘志华等）以及课题组成员于满泉等的博士学位论文和赵燕平老师一些学生的硕士学位论文，而有关信息检索模型部分系摘自由张华平等翻译、Grossman 和 Frieder 合著的信息检索专著 *Information Retrieval: Algorithms and Heuristics*（*Second Edition*）中的部分章节。另外，国内外众多的信息检索与数据挖掘方面的研究成果和相关网站亦为本书提供了帮助，本书的顺利完成也得益于参阅了大量的相关工作及研究成果，在此谨向这些文献的作者以及为本书提供帮助的人致以诚挚的谢意和崇高的敬意。

本书得到 2013 年国家自然科学基金（编号：61272362）、国家 973 重点基础研

究发展计划(编号:2013CB329606)、2012 年河北省科技支撑计划(编号:12213516D)、2012 年新疆维吾尔自治区高新技术计划(编号:201212124)、2013 年河北省自然科学基金(编号:F2013208105)的资助。在本书写作过程中,也得到了科学出版社的大力支持和帮助,在此一并表示衷心感谢。

作为本书成果的配套网站和相关资料的下载基地,课题组的官方网站(http://www.nlpir.org)提供实际成果演示与各类资源的下载。

限于作者的学识水平,书中不妥之处在所难免,恳请广大读者批评指正。

作 者

2014 年 2 月

目　　录

《信息科学技术学术著作丛书》序

序

前言

第 1 章　绪论 …… 1

1.1　大数据 …… 4

1.2　云计算及 Hadoop 简介 …… 5

1.3　Web 搜索、全文索引与 Lucene 简介 …… 6

1.3.1　Web 搜索 …… 7

1.3.2　全文索引 …… 9

1.3.3　Lucene 简介 …… 11

1.4　大数据挖掘 …… 13

1.5　本书主要内容及其知识点 …… 14

1.6　本章小结 …… 16

参考文献 …… 17

第 2 章　大数据搜索挖掘综述 …… 18

2.1　常用的信息检索模型 …… 18

2.1.1　传统布尔检索与扩展布尔检索模型 …… 18

2.1.2　向量空间模型 …… 19

2.1.3　概率检索模型 …… 23

2.1.4　语言模型 …… 24

2.2　自然语言理解与处理概述 …… 26

2.3　中文词法分析中的分词处理 …… 28

2.3.1　基于词典和规则的汉字分词 …… 30

2.3.2　基于大规模语料库的统计学习的分词方法 …… 30

2.3.3　规则和统计方法相结合的汉字分词方法 …… 32

2.4　未登录词及其识别 …… 32

2.4.1　命名实体及其识别 …… 33

2.4.2　未登录词与新词识别 …… 34

2.5 有意义串及其识别 …… 36
2.6 词典组织与管理 …… 37
2.6.1 基于 Trie 索引树的词典管理 …… 37
2.6.2 基于哈希表的词典管理 …… 38
2.7 文本分类 …… 39
2.8 文本聚类 …… 41
2.8.1 文本表示 …… 41
2.8.2 相似度度量 …… 42
2.8.3 聚类算法体系 …… 43
2.9 话题识别与跟踪 …… 46
2.10 句子及其检索 …… 48
2.10.1 传统的文档检索方法 …… 48
2.10.2 信息过滤方法 …… 48
2.10.3 分类方法 …… 49
2.10.4 语义比较方法 …… 49
2.10.5 隐马尔可夫模型方法 …… 50
2.10.6 自动文摘方法 …… 51
2.11 句子级新信息检测 …… 52
2.11.1 词重叠度 …… 52
2.11.2 最大区间相关度 …… 52
2.11.3 余弦冗余度 …… 52
2.11.4 命名实体触发方法 …… 53
2.11.5 统计机器翻译模型 …… 53
2.11.6 LexRank 方法 …… 54
2.12 本章小结 …… 55
参考文献 …… 56
第 3 章 大数据检索与分词 …… 62
3.1 概述 …… 62
3.2 分词对中文信息检索的影响 …… 63
3.3 分词精度与检索性能的关系 …… 66
3.4 大数据应用环境下中文信息检索的分词算法及其特点 …… 70
3.4.1 分词算法的时间性能要求高 …… 70

3.4.2　分词正确率的提高并不一定带来检索性能的提高 …… 70
3.4.3　分词切分粒度需在查询扩展层面进行相关处理 …… 70
3.4.4　未登录词识别的准确率要比召回率更重要 …… 71
3.5　基于双数组 Trie 树优化算法的词典 …… 72
3.5.1　双数组 Trie 树算法介绍及其优化 …… 72
3.5.2　利用优化的双数组 Trie 树算法组织词典 …… 74
3.5.3　实验结果与分析 …… 76
3.6　本章小结 …… 77
参考文献 …… 78
第 4 章　基于层次隐马尔可夫模型的浅层词法分析 …… 80
4.1　概述 …… 80
4.2　英文浅层分析的实现 …… 81
4.2.1　英文断句与词汇切分 …… 81
4.2.2　词性标注 …… 82
4.2.3　词干抽取与词形还原 …… 83
4.3　停用词处理与特征词选择 …… 84
4.3.1　停用词处理 …… 84
4.3.2　特征词选择 …… 85
4.4　基于层次隐马尔可夫模型的汉语浅层分析及其应用 …… 86
4.4.1　层次隐马尔可夫模型 …… 87
4.4.2　基于类的隐马尔可夫分词算法 …… 89
4.4.3　N 最短路径的切分排歧策略 …… 90
4.4.4　未登录词的隐马尔可夫识别方法 …… 91
4.5　汉语词法分析系统 ICTCLAS 性能实验与分析 …… 93
4.5.1　词法分析与层次隐马尔可夫模型 …… 94
4.5.2　ICTCLAS 在 973 评测中的测试结果 …… 95
4.5.3　第一届国际分词大赛的评测结果 …… 95
4.6　基于单字位置成词概率识别未登录词的算法 …… 96
4.6.1　字的位置成词概率 …… 96
4.6.2　局部二元串频统计 …… 98
4.6.3　有关未登录词识别的实验结果 …… 99
4.7　本章小结 …… 100

参考文献 …… 102
第 5 章 大数据语言新特征发现 …… 104
5.1 概述 …… 104
5.2 基于上下文邻接分析和语言模型的有意义串提取 …… 106
5.2.1 上下文邻接分析 …… 107
5.2.2 语言模型分析 …… 109
5.2.3 重复串发现及处理流程 …… 111
5.2.4 实验设计及结果分析 …… 115
5.3 基于局部性原理的低频有意义串提取 …… 120
5.3.1 有意义串的局部性 …… 121
5.3.2 局部性度量 …… 122
5.3.3 算法流程 …… 123
5.3.4 实验结果与分析 …… 124
5.4 基于伪相关反馈模型的有意义串提取 …… 127
5.4.1 算法的基本思想 …… 128
5.4.2 相关度的定义 …… 129
5.4.3 位置成词概率 PWP 的更新 …… 129
5.4.4 算法流程 …… 129
5.4.5 实验结果及分析 …… 130
5.5 本章小结 …… 133
参考文献 …… 135
第 6 章 大数据聚类与分类 …… 138
6.1 概述 …… 138
6.2 基于关键词提取的搜索结果聚类 …… 139
6.2.1 相关术语简介 …… 139
6.2.2 关键词提取 …… 139
6.2.3 基于关键词的检索结果聚类方法 …… 141
6.2.4 实验结果及分析 …… 142
6.3 基于 K-means 算法的有意义串主题聚类算法 …… 144
6.4 基于邻接串种类的有意义串语境聚类 …… 146
6.5 有意义串对分类的改进 …… 149
6.6 本章小结 …… 153

参考文献 …… 154
第 7 章　大数据文本自动摘要 …… 156
7.1　概述 …… 156
7.2　相关工作综述 …… 156
7.2.1　基于抽取的自动文摘 …… 158
7.2.2　基于理解的自动文摘 …… 160
7.3　基于关键词提取的自动摘要 …… 160
7.3.1　文本预处理 …… 160
7.3.2　停用词处理 …… 161
7.3.3　双数组 Trie 树 …… 162
7.3.4　关键词提取 …… 164
7.3.5　句子切分 …… 166
7.3.6　句子相似度计算 …… 166
7.4　面向主题的自动摘要 …… 167
7.4.1　改进的最大边缘相关度方法 …… 167
7.4.2　面向主题的词特征统计 …… 168
7.4.3　领域主题词表 …… 169
7.4.4　句子间的包含关系 …… 170
7.5　实验与分析 …… 171
7.5.1　稳定性测试 …… 171
7.5.2　时间性能 …… 171
7.5.3　文摘质量 …… 174
7.6　自动摘要应用场景分析及大数据搜索与挖掘软件应用示例 …… 174
7.7　本章小结 …… 176
参考文献 …… 176
第 8 章　JZSearch 大数据精准搜索引擎 …… 178
8.1　概述 …… 178
8.2　JZSearch 大数据搜索引擎系统架构 …… 178
8.3　JZSearch 索引关键技术 …… 180
8.3.1　索引字段类型 …… 180
8.3.2　索引词项的设计 …… 181
8.3.3　索引压缩技术 …… 181

8.3.4 内存交换 …… 184
8.3.5 增量索引 …… 184
8.3.6 数据库检索 …… 185
8.4 JZSearch 搜索技术 …… 187
8.4.1 JZSearch 排序算法 …… 187
8.4.2 JZSearch 结果格式 …… 188
8.4.3 JZSearch 检索语法说明 …… 188
8.5 JZSearch 搜索引擎管理 …… 193
8.5.1 搜索引擎可视化管理客户端 …… 193
8.5.2 客户端管理命令语法 …… 194
8.6 JZSearch 大数据搜索应用案例 …… 194
8.6.1 中国邮政集团名址信息中心首页的邮址垂直搜索 …… 194
8.6.2 河北省标准化研究院的标准搜索 …… 195
8.6.3 中国对外承包工程商会的知识搜索门户 …… 196
8.6.4 富基融通的商品比价搜索 …… 196
8.6.5 微博人物搜索 …… 196
8.6.6 维吾尔语搜索 …… 196
8.7 本章小结 …… 198
参考文献 …… 199
第 9 章 面向大数据的句子检索与新颖性监测 …… 200
9.1 概述 …… 200
9.2 句子检索的查询扩展方法 …… 201
9.2.1 语义扩展 …… 201
9.2.2 伪相关反馈扩展 …… 203
9.2.3 局部共现扩展 …… 204
9.3 语言模型检索 …… 206
9.3.1 概述 …… 206
9.3.2 句子级语言模型及其改进 …… 207
9.4 句子检索实验与分析 …… 207
9.4.1 浅层语言分析的贡献度 …… 207
9.4.2 三种句子检索模型的基准实验 …… 209
9.4.3 查询扩展实验 …… 211

9.5　新信息检测 …… 212
9.5.1　词重叠度及其扩展 …… 213
9.5.2　相似度比较方法 …… 214
9.5.3　信息增强评价方法 …… 215
9.5.4　其他方法 …… 215
9.5.5　新信息检测实验与分析 …… 217
9.6　监督学习条件下的句子检索与新信息检测 …… 219
9.6.1　监督学习环境下的参数调整与阈值设置 …… 219
9.6.2　基于分类的句子检索与新信息检测方法 …… 221
9.6.3　实验与分析 …… 222
9.7　本章小结 …… 224
参考文献 …… 225
第 10 章　人物追踪中的数据预处理与属性抽取 …… 227
10.1　概述 …… 227
10.1.1　研究背景 …… 227
10.1.2　人物追踪及其处理流程 …… 228
10.2　数据预处理 …… 228
10.2.1　数据预处理的主要流程 …… 229
10.2.2　网页正文提取与噪声过滤 …… 229
10.2.3　人名识别 …… 232
10.2.4　人名指代处理 …… 232
10.2.5　人物对应语段的确定 …… 233
10.2.6　时间和时序标签的确定 …… 234
10.3　人物属性抽取 …… 234
10.3.1　人物属性抽取的总体框架 …… 235
10.3.2　标注人物属性抽取语料 …… 235
10.3.3　分类器模型 …… 237
10.4　本章小结 …… 243
参考文献 …… 245
第 11 章　人物模型组织与基于事件的信息处理 …… 246
11.1　概述 …… 246
11.2　人物模型的特征表示 …… 247

11.2.1 属性特征的表示 …… 247
11.2.2 数值特征的表示 …… 247
11.2.3 各项特征的分布规律 …… 249
11.3 人物模型的相似度计算方法 …… 249
11.3.1 基本属性的相似度计算 …… 249
11.3.2 介绍性属性的相似度计算 …… 249
11.3.3 词场的相似度计算 …… 251
11.3.4 人物模型相似度计算 …… 251
11.4 人物模型的同一性判别与合并 …… 252
11.5 实验结果与分析 …… 253
11.5.1 数据集与评测方法 …… 253
11.5.2 实验结果 …… 254
11.6 基于宏观粒度的事件组织 …… 258
11.6.1 宏观粒度事件的特征 …… 258
11.6.2 针对事件特点的话题识别方法 …… 259
11.6.3 基于多层聚类的话题层次化组织方法 …… 261
11.6.4 实验结果与分析 …… 265
11.7 本章小结 …… 270
参考文献 …… 272
附录 A ICTCLAS/NLPIR 2014 汉语分词系统介绍 …… 274
附录 B NLPIR 大数据搜索与挖掘共享开发平台 …… 281