

《大数据搜索与挖掘》

图书基本信息

书名：《大数据搜索与挖掘》

13位ISBN编号：9787030403185

出版时间：2014-5

作者：张华平,高凯,黄河燕,赵燕平

页数：292

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：www.tushu000.com

《大数据搜索与挖掘》

内容概要

《信息科学技术学术著作丛书:大数据搜索与挖掘》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。

《大数据搜索与挖掘》

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。

书籍目录

- 《信息科学技术学术著作丛书》序
- 序
- 前言
- 第1章绪论
 - 1.1大数据
 - 1.2云计算及Hadoop简介
 - 1.3Web搜索、全文索引与Lucene简介
 - 1.3.1Web搜索
 - 1.3.2全文索引
 - 1.3.3Lucene简介
 - 1.4大数据挖掘
 - 1.5本书主要内容及其知识点
 - 1.6本章小结
- 参考文献
- 第2章大数据挖掘综述
 - 2.1常用的信息检索模型
 - 2.1.1传统布尔检索与扩展布尔检索模型
 - 2.1.2向量空间模型
 - 2.1.3概率检索模型
 - 2.1.4语言模型
 - 2.2自然语言理解与处理概述
 - 2.3中文词法分析中的分词处理
 - 2.3.1基于词典和规则的汉字分词
 - 2.3.2基于大规模语料库的统计学习的分词方法
 - 2.3.3规则和统计方法相结合的汉字分词方法
 - 2.4未登录词及其识别
 - 2.4.1命名实体及其识别
 - 2.4.2未登录词与新词识别
 - 2.5有意义串及其识别
 - 2.6词典组织与管理
 - 2.6.1基于Trie索引树的词典管理
 - 2.6.2基于哈希表的词典管理
 - 2.7文本分类
 - 2.8文本聚类
 - 2.8.1文本表示
 - 2.8.2相似度度量
 - 2.8.3聚类算法体系
 - 2.9话题识别与跟踪
 - 2.10句子及其检索
 - 2.10.1传统的文档检索方法
 - 2.10.2信息过滤方法
 - 2.10.3分类方法
 - 2.10.4语义比较方法
 - 2.10.5隐马尔可夫模型方法
 - 2.10.6自动文摘方法
 - 2.11句子级新信息检测
 - 2.11.1词重叠度

2.11.2最大区间相关度

2.11.3余弦冗余度

2.11.4命名实体触发方法

2.11.5统计机器翻译模型

2.11.6LexRank方法

2.12本章小结

参考文献

第3章大数据检索与分词

3.1概述

3.2分词对中文信息检索的影响

3.3分词精度与检索性能的关系

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

3.4.1分词算法的时间性能要求高

3.4.2分词正确率的提高并不一定带来检索性能的提高

3.4.3分词切分粒度需在查询扩展层面进行相关处理

3.4.4未登录词识别的准确率要比召回率更重要

3.5基于双数组Trie树优化算法的词典

3.5.1双数组Trie树算法介绍及其优化

3.5.2利用优化的双数组Trie树算法组织词典

3.5.3实验结果与分析

3.6本章小结

参考文献

第4章基于层次隐马尔可夫模型的浅层词法分析

4.1概述

4.2英文浅层分析的实现

4.2.1英文断句与词汇切分

4.2.2词性标注

4.2.3词干抽取与词形还原。

4.3停用词处理与特征词选择

4.3.1停用词处理

4.3.2特征词选择

4.4基于层次隐马尔可夫模型的汉语浅层分析及其应用

4.4.1层次隐马尔可夫模型

4.4.2基于类的隐马尔可夫分词算法

4.4.3N最短路径的切分排歧策略

4.4.4未登录词的隐马尔可夫识别方法

4.5汉语词法分析系统ICTCLAS性能实验与分析

4.5.1词法分析与层次隐马尔可夫模型

4.5.2ICTCLAS在973评测中的测试结果

4.5.3第一届国际分词大赛的评测结果

4.6基于单字位置成词概率识别未登录词的算法

4.6.1字的位置成词概率

4.6.2局部二元串频统计

4.6.3有关未登录词识别的实验结果

4.7本章小结

参考文献

第5章大数据语言新特征发现

5.1概述

5.2基于上下文邻接分析和语言模型的有意义串提取

- 5.2.1 上下文邻接分析
- 5.2.2 语言模型分析
- 5.2.3 重复串发现及处理流程
- 5.2.4 实验设计及结果分析
- 5.3 基于局部性原理的低频有意义串提取
 - 5.3.1 有意义串的局部性
 - 5.3.2 局部性度量
 - 5.3.3 算法流程
 - 5.3.4 实验结果与分析
- 5.4 基于伪相关反馈模型的有意义串提取
 - 5.4.1 算法的基本思想
 - 5.4.2 相关度的定义
 - 5.4.3 位置成词概率PWP的更新
 - 5.4.4 算法流程
 - 5.4.5 实验结果及分析
- 5.5 本章小结
- 参考文献
- 第6章 大数据聚类与分类
 - 6.1 概述
 - 6.2 基于关键词提取的搜索结果聚类
 - 6.2.1 相关术语简介
 - 6.2.2 关键词提取
 - 6.2.3 基于关键词的检索结果聚类方法
 - 6.2.4 实验结果及分析
 - 6.3 基于K—means算法的有意义串主题聚类算法
 - 6.4 基于邻接串种类的有意义串语境聚类
 - 6.5 有意义串对分类的改进
 - 6.6 本章小结
 - 参考文献
 -
- 第7章 大数据文本自动摘要
- 第8章 JZSearch 大数据精准搜索引擎
- 第9章 面向大数据的句子检索与新颖性监测
- 第10章 人物追踪中的数据预处理与属性抽取
- 第11章 人物模型组织与基于事件的信息处理
- 附录 A ICTCLAS/NLPIR2014 汉语分词系统介绍
- 附录 B NLPIR 大数据搜索与挖掘共享开发平台

《大数据搜索与挖掘》

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:www.tushu000.com