

## 内 容 提 要

本书共分7章, 主要包括: 大数据与数据分析、大数据存储、大数据分析工具、大数据与信息安全、基于二部图网络的电子商务推荐算法研究、基于位置的社交网络好友推荐算法研究、基于稀有类分类的信用卡欺诈识别研究。

本书可作为大中专院校计算机、电子商务相关专业的教材, 也可供渴望了解大数据知识的人士参考阅读。

## 图书在版编目(C I P)数据

大数据分析与应用 / 赵守香, 唐胡鑫, 熊海涛著

. -- 北京: 航空工业出版社, 2015. 12

ISBN 978-7-5165-0956-2

I. ①大… II. ①赵… ②唐… ③熊… III. ①数据处理 IV. ①TP274

中国版本图书馆CIP数据核字(2015)第305424号

## 大数据分析与应用

Dashuju Fenxi yu Yingyong

航空工业出版社出版发行

(北京市朝阳区北苑2号院 100012)

发行部电话: 010-84936597 010-84936343

三河市祥达印刷包装有限公司印刷

全国各地新华书店经售

2015年12月第1版

2015年12月第1次印刷

开本: 787×1092

1/16

印张: 20.25

字数: 468千字

印数: 1—3000

定价: 48.00元



第 1 章 大数据与数据分析	1
1.1 概述	1
1.1.1 大数据的含义	2
1.1.2 大数据的定义	3
1.1.3 大数据的特征	4
1.1.4 大数据与云计算	7
1.1.5 大数据与商业模式变革	8
1.1.6 大数据带来的改变	9
1.2 云计算与大数据	10
1.2.1 云计算的概念	11
1.2.2 云计算的特征	12
1.2.3 云计算的服务方式	13
1.2.4 云计算的应用	14
1.3 电子商务与大数据	15
1.3.1 电子商务催生大数据	16
1.3.2 数据分析给电子商务带来更多机会	17
1.3.3 网站分析与应用	19
1.4 物联网与大数据	20
1.4.1 物联网的含义	20
1.4.2 物联网与大数据的关系	21
1.4.3 美国物联网应用	22
1.5 移动互联网与大数据	23
1.6 大数据应用给企业带来的机会	25
1.7 大数据应用带来的挑战	27
1.7.1 大数据促使商业领域重新洗牌	27
1.7.2 三足鼎立的大数据公司	29
1.7.3 加速成长的大数据中间商	32
1.7.4 大数据给个人隐私带来威胁	35



1.7.5 大数据分析的不可靠性·····	36
1.7.6 大数据引发管理规范变革·····	40
1.8 大数据应用·····	41
1.8.1 大数据让互联网越来越智能·····	41
1.8.2 大数据在银行业的应用趋势·····	44
1.8.3 企业大数据创新的五大趋势·····	44
<b>第2章 大数据存储·····</b>	<b>46</b>
2.1 大数据对数据存储的要求·····	46
2.1.1 数据存储面临的问题·····	47
2.1.2 大数据存储不容忽视的问题·····	48
2.1.3 数据存储技术面临的挑战·····	51
2.1.4 存储技术趋势预测与分析·····	52
2.2 存储技术·····	54
2.2.1 存储概述·····	54
2.2.2 直接附加存储 (DAS)·····	56
2.2.3 磁盘阵列 (RAID)·····	57
2.2.4 网络附加存储 (NAS)·····	59
2.2.5 存储区域网络 (SAN)·····	60
2.2.6 IP 存储 (SoIP)·····	61
2.2.7 iSCSI 网络存储·····	63
2.2.8 存储技术比较·····	65
2.3 云存储技术·····	67
2.3.1 云存储概述·····	67
2.3.2 云存储技术与传统存储技术·····	68
2.3.3 云存储的优点·····	68
2.3.4 云存储的分类·····	69
2.3.5 云存储的技术基础·····	71
2.3.6 云存储系统的结构模型·····	72
2.3.7 云存储解决方案·····	74
2.3.8 云存储的用途和发展趋势·····	76
2.4 大数据存储解决方案·····	78
2.4.1 戴尔的流动文件系统·····	78
2.4.2 华为的集群存储系统·····	80
2.4.3 戴尔的自动分层存储·····	82

2.4.4	EMC 的闪存存储技术	84
2.4.5	虚拟化技术	87
<b>第 3 章 大数据分析工具</b>		<b>94</b>
3.1	数据分析概述	94
3.1.1	数据分析的概念	94
3.1.2	数据分析过程	96
3.1.3	数据分析框架的主要事件	98
3.2	数据挖掘	99
3.2.1	数据挖掘的概念	99
3.2.2	数据挖掘的任务	100
3.2.3	数据挖掘的过程	102
3.2.4	数据挖掘的主要算法	104
3.2.5	数据挖掘的应用领域	108
3.2.6	数据挖掘和 OLAP	109
3.3	关联分析	109
3.3.1	关联分析的概念	109
3.3.2	关联规则挖掘过程	110
3.3.3	关联规则的分类	112
3.3.4	关联规则的相关算法	112
3.3.5	关联规则的应用	113
3.4	Apriori 算法	117
3.4.1	Apriori 算法的挖掘	117
3.4.2	基于 Apriori 算法的数据挖掘应用实例	119
3.4.3	Apriori 算法的优缺点及优化思考	120
3.5	聚类分析	121
3.5.1	聚类分析的概念	121
3.5.2	聚类分析的应用	124
3.5.3	序列聚类	127
3.6	分类分析	127
3.6.1	决策树	127
3.6.2	朴素贝叶斯 (Naive Bayes)	130
3.6.3	神经网络	131
3.6.4	回归	132
3.6.5	其他分类算法	133



3.7 时间序列分析·····	134
3.7.1 时间序列的概念·····	134
3.7.2 时间序列的分类·····	136
3.7.3 时间序列分析方法·····	136
3.7.4 时间序列分析的步骤及用途·····	137
3.7.5 时间序列分析预测方法·····	138
3.8 确定性时间序列分析·····	141
3.8.1 移动平均法·····	141
3.8.2 指数平滑法·····	142
3.8.3 趋势预测·····	144
3.9 随机性时间序列分析·····	144
3.9.1 平稳随机时间序列分析·····	144
3.9.2 非平稳时间序列分析·····	146
<b>第4章 大数据与信息安全·····</b>	<b>147</b>
4.1 大数据带来的安全问题·····	147
4.1.1 大数据安全面临的问题·····	148
4.1.2 大数据安全需求·····	150
4.1.3 大数据安全的特征·····	152
4.2 大数据信息安全风险因素识别·····	155
4.2.1 大数据信息安全问题日益凸显·····	155
4.2.2 移动互联网/智能手机是个人信息泄露的重要渠道·····	157
4.2.3 物联网应用的安全问题·····	158
4.2.4 公民的信息安全意识薄弱带来的信息安全隐患·····	159
4.3 大数据安全策略·····	161
4.3.1 美国降低关键基础设施信息与网络安全风险的框架·····	162
4.3.2 确定关键信息基础设施·····	166
4.3.3 确定数据的访问权限·····	170
4.4 大数据安全与政策法规建设·····	171
4.4.1 国外大数据安全相关举措·····	171
4.4.2 树立隐私价值观·····	172
4.4.3 确定第三方数据的访问权限·····	173
4.4.4 制定大数据信息安全法律法规·····	175
4.4.5 大数据时代个人信息的法律保护·····	175

第 5 章 基于二部图网络的电子商务推荐算法研究 .....	178
5.1 概述 .....	178
5.1.1 研究背景 .....	178
5.1.2 研究目的及意义 .....	179
5.1.3 数据集介绍 .....	181
5.2 推荐算法概述 .....	181
5.2.1 推荐算法的起源及发展历史 .....	182
5.2.2 推荐算法的应用现状 .....	184
5.2.3 目前主要推荐算法 .....	186
5.2.4 推荐算法评测 .....	189
5.2.5 推荐算法评测结果的比较 .....	194
5.3 基于二部图网络的推荐算法 .....	194
5.3.1 复杂网络的演化过程 .....	195
5.3.2 复杂网络简介 .....	195
5.3.3 二部图网络简介 .....	196
5.3.4 基于二部图网络的推荐算法 .....	197
5.3.5 目前一些可行的优化算法 .....	204
5.4 基于二部图网络推荐算法的改进 .....	209
5.4.1 基于二部图网络的推荐算法的不足 .....	209
5.4.2 社会化标签 .....	210
5.4.3 引入社会化标签的二部图网络推荐算法 .....	212
5.5 仿真实验 .....	216
5.5.1 数据集 .....	216
5.5.2 实验思路 .....	218
5.5.3 实验结果及分析 .....	226
第 6 章 基于位置的社交网络好友推荐算法研究 .....	232
6.1 概述 .....	232
6.1.1 研究背景 .....	232
6.1.2 研究内容及组织结构 .....	235
6.1.3 研究目标与意义 .....	236
6.2 基于位置的社交网络 .....	236
6.2.1 基于位置的社交网络概述 .....	237
6.2.2 基于位置的社交网络研究现状 .....	238
6.2.3 基于位置的社交网络推荐算法分类 .....	240



6.3 实验数据集及其特征分析 .....	242
6.3.1 Brightkite 网站及实验数据集介绍 .....	242
6.3.2 数据清理与数据存储 .....	243
6.3.3 实验数据集的特征分析 .....	244
6.4 基于位置信息对好友推荐算法的改进 .....	250
6.4.1 实验方法 .....	250
6.4.2 评估方法 .....	251
6.4.3 基于局部信息的好友推荐算法 .....	253
6.4.4 基于随机游走的好友推荐算法 .....	256
6.4.5 基于路径相似的好友推荐算法 .....	259
6.5 时间信息对于位置信息作用的影响 .....	263
6.5.1 签到时间对位置信息的影响分析 .....	263
6.5.2 引入时间信息后的好友推荐算法改进 .....	267
6.6 总结与展望 .....	268
6.6.1 研究工作总结 .....	268
6.6.2 未来研究方向 .....	269
<b>第 7 章 基于稀有类分类的信用卡欺诈识别研究 .....</b>	<b>271</b>
7.1 概述 .....	271
7.1.1 信用卡行业发展 .....	271
7.1.2 信用卡欺诈风险 .....	272
7.1.3 信用卡欺诈识别研究 .....	273
7.1.4 国内外研究现状 .....	274
7.1.5 研究思路和步骤 .....	277
7.2 稀有类分类方法基本理论 .....	279
7.2.1 稀有类分类介绍 .....	279
7.2.2 稀有类分类的方法 .....	279
7.2.3 稀有类分类性能评估 .....	285
7.3 不均衡数据集的处理 .....	287
7.3.1 不均衡数据集的研究现状 .....	287
7.3.2 聚类方法的介绍 .....	288
7.3.3 聚类方法的选择 .....	293
7.4 基于 Adaboost 的稀有类分类算法 .....	296
7.4.1 Adaboost 算法介绍 .....	296
7.4.2 Adaboost 算法的研究现状 .....	297

---

7.4.3 Adaboost 算法的改进 .....	299
7.4.4 Adaboost 算法改进效果分析 .....	300
7.5 基于稀有类分类的信用卡欺诈识别模型 .....	303
7.5.1 信用卡欺诈识别模型介绍 .....	303
7.5.2 信用卡欺诈识别模型构建 .....	303
7.5.3 实验分析 .....	306
7.6 总结与展望 .....	308
7.6.1 研究工作总结 .....	308
7.6.2 后续研究展望 .....	309
参考文献 .....	311