大数据挖掘技术与应用

孟海东　宋宇辰　著

DASHUJU WAJUE
JISHU YU YINGYONG

冶金工业出版社
Metallurgical Industry Press

大数据挖掘技术与应用

孟海东　宋宇辰　著

北　京
冶 金 工 业 出 版 社
2014

内 容 提 要

本书针对大数据的海量性、高维性、异构性、动态性、多样性、多源性、多尺度性、时空性和模糊性等特征，对数据挖掘技术中的聚类分析和关联规则分析进行了系统的研究；研究与开发了基于密度和自适应密度可达聚类算法、动态增量聚类算法、并行聚类算法、高维多类型数据聚类算法、基于密度加权的模糊聚类算法、基于数据场的聚类和量化关联规则算法、基于距离的量化关联规则分析、基于云计算的大数据聚类算法，以及挖掘结果的可视化表达；给出了地球化学数据挖掘、基于数据挖掘的中国资源与区域经济发展关系的分析应用实例。

本书可供从事数据挖掘技术研究、应用和软件开发人员以及学习数据挖掘技术的本科生和研究生参考。

图书在版编目（CIP）数据

大数据挖掘技术与应用/孟海东，宋宇辰著．—北京：冶金工业出版社，2014.12

ISBN 978-7-5024-6780-7

Ⅰ.①大… Ⅱ.①孟… ②宋… Ⅲ.①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2014）第 276653 号

出 版 人 谭学余
地 址 北京市东城区嵩祝院北巷 39 号 邮编 100009 电话 （010）64027926
网 址 www.cnmip.com.cn 电子信箱 yjcbs@cnmip.com.cn
责任编辑 宋 良 王雪涛 美术编辑 吕欣童 版式设计 孙跃红
责任校对 石 静 责任印制 李玉山
ISBN 978-7-5024-6780-7
冶金工业出版社出版发行；各地新华书店经销；三河市双峰印刷装订有限公司印刷
2014 年 12 月第 1 版，2014 年 12 月第 1 次印刷
169mm×239mm；18.25 印张；368 千字；278 页
56.00 元

冶金工业出版社 投稿电话 （010）64027932 投稿信箱 tougao@cnmip.com.cn
冶金工业出版社营销中心 电话 （010）64044283 传真 （010）64027893
冶金书店 地址 北京市东四西大街 46 号（100010） 电话 （010）65289081（兼传真）
冶金工业出版社天猫旗舰店 yjgy.tmall.com

前　言

大数据必然隐含丰富的知识和高价值。无论大数据具有何种内涵和外延，如体量巨大、种类繁多、快速流动和低价值密度等，其本质特征是数据的海量性、高维性、异构性、动态性、时空性、多样性、多源性、多尺度性和模糊性。数据挖掘技术是实现数据向知识和价值转化的重要技术手段，但是要从大数据中挖掘出隐含的丰富知识和价值，传统的数据挖掘技术面临多方面的挑战。解决大数据挖掘问题的重要途径，就是根据大数据的本质特征研究与开发有效的数据挖掘算法。

聚类分析和关联规则分析是数据挖掘技术领域的重要研究内容。聚类分析已被广泛地应用于发现数据对象的全局分布模式，如模式识别、数据分析、图像处理、市场研究等。关联规则分析是用来研究事物与事物之间在一定约束条件下的依存性和关联性，以及数据对象属性取值之间存在的某种规律性和相关性，特别是当属性取值之间不能用某种数学函数关系表达时，量化关联规则能够客观地表达属性间的关联性。

大数据不仅来源于网络世界，而且也来源于大量的科学实验、天体探测、航天航空数字遥感、地球科学、医疗与生物、商业、金融与保险等领域，无论是大数据还是传统意义上的数据，聚类分析算法的有效性是数据挖掘领域研究的重要课题。通过利用复相关系数倒数对数据对象属性加权和数据对象的分布密度，在 K-means 算法的基础上提高了聚类分析的有效性。

大数据对象具有数据空间分布状态的复杂性，如数据空间分布不同大小、不同形态和不同密度数据对象的分布模式，为了能够有效地

在数据空间发现客观存在的复杂形态的数据对象分布模式，通过计算数据空间数据对象的分布密度，确定密度吸引点（极值点）和数据对象到密度吸引点的密度可达实现了不同大小、不同形态和不同密度数据对象分布的有效聚类。

数据的海量性是大数据的重要特征，如何实现大数据空间数据对象的有效聚类分析，是大数据挖掘技术研究的重要内容之一，也是实现大数据向知识与智慧和价值的转化需要解决的重要问题。根据数据空间数据对象密度可达与子簇特征相似定义，研究与开发了动态增量聚类分析算法，为解决海量数据聚类分析算法的可扩展性问题提供了一种方法。

Gartner Group 的一次高级技术调查将数据挖掘和人工智能列为“未来三到五年内将对工业产生深远影响的五大关键技术”之首，并且还将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术的前两位。Gartner 的最新 HPC 研究表明，“随着数据捕获、传输和存储技术的快速发展，大型系统用户将更多地需要采用新技术来挖掘市场以外的价值，采用更为广阔的并行处理系统来创建新的商业增长点”。书中采用任务和数据并行技术，研究与开发了并行聚类分析算法。

大数据具有的高维性给数据挖掘带来了维度灾难，大数据对象属性的多样性（多类型）也给数据挖掘算法带来了挑战。通过数据挖掘技术使低价值密度的大数据转化为知识、智慧和价值，重要的研究课题之一是高维度、多类型属性数据对象的聚类分析。在研究维度对聚类分析有效性影响的基础上，通过属性加权和属性转换的方法，研究了高维度、多类型属性数据对象聚类分析的有效性。

模糊聚类分析是依据数据对象（客观事物）间的特征、亲疏程度和相似性，通过建立模糊聚类相似关系对数据对象（客观事物）进行分类。大数据隐含的知识与智慧的表达具有更大的模糊性。模糊聚类表达的聚类信息更加客观真实，通过数据对象分布密度加权，使模糊聚类分析结果更加客观、有效。

关联规则分析最早用来确定事务数据库中事务项之间的关联关系，这种关系是在支持度和置信度约束下的布尔型关联关系。在自然科学领域，更多是需要研究与确定数据项（属性）间的关联关系，而这种关系是不能够用线性或非线性函数关系来表达，只能用在一定约束条件下的量化关联规则来表达，例如，在地球科学、气象学、医学、经济学等领域，这种量化关联关系客观存在于大数据中，而且对于大数据分析更有意义。基于距离的量化关联规则算法为量化关联规则挖掘提供了更多的途径。

大数据的重要特征是数据量“大”，但是“大”并不表示数据对象在数据空间分布的“完备”。根据数据场理论，将数据对象扩展到完整的数据空间，得到完备的数据对象分布，使得基于数据场的数据挖掘结果更有意义、有价值。

云计算具有海量的存储能力和弹性化的计算能力，在大数据挖掘领域逐渐表现出其显著优势。Hadoop 平台是 Apache 推出的开源云计算平台，在 Hadoop 平台上基于 MapReduce 的数据挖掘技术在大数据挖掘中发挥着重要作用。

数据挖掘过程与挖掘结果的可视化表达是获取、评价和理解挖掘知识的重要手段。利用可视化技术将隐含的、有意义的挖掘结果进行可视化表达，能够有序地发现其中隐藏的特征、关系、模式和趋势等，便于发现新的知识并做出合理的决策。书中采用二维散点图、三维散点图、平行坐标图、圆环段、星形图等方式实现了聚类分析和关联规则分析结果的可视化表达。

地球科学的发展为矿产资源预测提供了丰富的地质、地球化学、地球物理和数字遥感等地球科学数据。大数据挖掘技术为地球科学数据处理提供了有效的技术手段。书中利用聚类分析算法，对地球化学数据进行聚类分析，确定可能存在的矿产资源赋存靶区；在此基础上利用相关分析和模糊聚类确定地球化学元素间的共生组合关系，根据元素组合关系推断靶区赋存的矿产资源类型。

另外，根据中国统计年鉴数据，通过选取地区能源，有色金属、

黑色金属、非金属矿产资源，污染物的排放量，污染治理费用，经济发展水平，固定资产投资，教育和科研项目指标，将我国的省、自治区和直辖市作为数据对象进行聚类分析，研究了不同指标体系下我国地区间表现出的资源与经济发展的相似性与相异性，通过分析产生这些相似性与相异性的原因，研究资源分布与区域经济发展的关系。

内蒙古科技大学信息工程学院、矿业工程学院和网络中心为本项目提供了计算资源和支持环境。算法研究与开发是在国家自然科学基金项目40762003和40764002、内蒙古自然科学基金项目200711020814和2012MS0611、教育部“春晖计划”合作项目Z2009-1-01041和Z2009-1-01055、内蒙古高等学校科学研究重点项目NJZZ11140和内蒙古科技大学矿业系统工程特色创新团队的资助下完成的。在算法研究与开发的过程中，内蒙古科技大学吕晓琪教授、张金山教授、李宝山教授、谭跃生教授、张晓琳教授，中国地质大学（北京）管建和教授，福州大学陈福集教授，澳大利亚维多利亚大学徐贯东博士，日本东北大学Dinil Pushpalal教授等提出了很多宝贵意见。在算法实现、验证和应用方面，研究生张玉英、宋飞燕、郝永宽、顾瑞春、王淑玲、申海涛、杨彦侃、张炼、马金徽、蔺志举、唐旋、刘小荣、马娜娜、李秉秋、李丹丹、吴鹏飞、孙家驹、管世明、娄建成、任敬佩、刘占宁和韩丽萍等做了大量的工作。参考了国内外学者的大量成果文献。在此一并表达诚挚的谢意。

研究和开发的算法存在的不足和缺陷，敬请广大读者提出改进意见，希望本书能够达到抛砖引玉的目的。

作　者

2014年8月

目　录

1 绪　论

随着数据库应用的普及，人们正逐步陷入“数据丰富，知识贫乏”的尴尬境地。而近年来互联网的发展与快速普及，使得人类第一次真正体会到了数据海洋的无边无际。面对如此巨量的数据资源，人们迫切需要一种新技术和自动工具，以便能够利用智能技术将这巨大的数据资源转换为有用的知识与信息资源，从而可以帮助我们科学地进行各种决策。于是，一个新的研究领域——知识发现应运而生。由于蕴藏知识的数据信息大多存储于数据库中，因此又称作数据库中的知识发现（Knowledge Discovery in Database，KDD）或数据挖掘（Data Mining，DM）。

早在1982年，趋势大师约翰·奈斯比（John Naisbitt）在他的首部著作《大趋势》（Megatrends）[1]中就提到：“人类正被信息淹没，却饥渴于知识”。计算机硬件技术的稳定进步为人类提供了大量的数据收集设备和存储介质；数据库技术的成熟和普及已使人类积累的数据量正在以指数方式增长；Internet技术的出现和发展已将整个世界连接成一个地球村，人们可以穿越时空般地在网上交换信息和协同工作。在这个信息爆炸的时代，面对着浩瀚无垠的信息海洋，人们迫切需要一个去粗取精、去伪存真的能将浩如烟海的数据转换成知识的技术。数据挖掘就是在这个背景下产生的。

数据挖掘作为一门新兴的学科，就是对观测到的数据集或庞大数据集进行分析，目的是发现未知的关系和以数据拥有者可以理解并对其有价值的新颖方式来总结数据。从技术上的角度考虑，数据挖掘的含义就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的，但又是潜在有用的信息和知识的过程。用数据挖掘工具进行数据分析，可以发现重要的数据模式，在商务决策、知识库、科学和医学研究等领域取得了一系列重要成果。

1.1　大数据

1.1.1　大数据概念

2012年3月，美国奥巴马政府宣布推出“大数据的研究和发展计划”。该计划涉及美国国家科学基金、美国国家卫生研究院、美国能源部、美国国防部、美

国国防部高级研究计划局、美国地质勘探局6个联邦政府部门，承诺将投资两亿多美元，大力推动和改善与大数据相关的收集、组织和分析工具及技术，以推进从大量的、复杂的数据集合中获取知识和洞见的能力。美国奥巴马政府宣布投资大数据领域，是大数据从商业行为上升到国家战略的分水岭，表明大数据正式提升到战略层面，大数据在经济社会各个层面、各个领域都开始受到重视[2]。

大数据是一个抽象的概念，不同的研究机构与学者对其有不同的定义。全球最具权威的IT研究与顾问咨询公司研究机构高德纳（The Gartner Group）认为，大数据是指需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。高德纳断言，到2015年，世界500强的组织中85%的财富将无法利用大数据作为竞争优势。维基百科将大数据定义为：大数据是所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策目的的资讯。全球最大的战略咨询公司麦肯锡的定义：大数据是指无法在一定时间内用传统数据库软件工具对其内容进行采集、存储、管理和分析的数据集合[3]。

大数据是基于多源异构、跨域关联的海量数据分析所产生的决策流程、商业模式、科学范式、生活方式和观念形态上的颠覆性变化的总和。

我国一些专家认为，大数据是指对海量数据进行智慧化处理和决策，这不仅是技术层面的问题，还涉及管理层面、互信机制等问题。建议在专门机构领导下，寻找大数据研究切入点，应对信息时代挑战。总的来说，大数据的概念应包含以下几个方面的内容。

A 具有战略性的智慧化数据处理和决策

据有关专家介绍，大数据是一个战略层面的概念，相较于其他数据分析、处理和研究，大数据具有战略导向性，具有更高的应用价值。

大数据不仅仅是指数据量大，而且也是指处理数据的能力与所能获得的数据量之间的差距。汪斌强教授指出："假如我一天可以处理两三个PB，产生的数据量只有几十兆，那么数据量再大也不算大数据，因为尽在掌握之中。"大数据技术手段相对以往的数据处理有根本性突破。以往通常是设置关键词，在海量数据库中搜索，然后请数据分析团队分析，通过人脑进行判断和预测。这种方法的问题在于，用来分析的数据来自关键词搜索，难以达到完备性。而大数据采取反向思路，控制了数据采集的源头，剔除掉数据库中的无用信息，屏蔽没有战略价值的数据，这是"大数据"处理与目前大海捞针式数据处理方式的本质不同。

大数据意味着数据处理从智能走向智慧。国内某专家介绍说，以前的海量数据处理，仅仅是信息资料收集过程，最终的决策和判断由另外的系统负责；而大数据的数据搜索和处理是一体化即时处理，需要数据可随时再找。同时，大数据