大数据元启发式算法教程

【法】Clarisse Dhaenens　Laetitia Jourdan◎著
康宁　宫鑫　刘婷婷◎译

METAHEURISTICS
FOR BIG DATA

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS
ISTE
WILEY

大数据元启发式算法教程

【法】Clarisse Dhaenens　Laetitia Jourdan◎著
康宁　宫鑫　刘婷婷◎译

METAHEURISTICS
FOR BIG DATA

人民邮电出版社
北京

图书在版编目（CIP）数据

大数据元启发式算法教程 /（法）克拉丽丝·达恩恩斯（Clarisse Dhaenens），（法）利蒂希娅·儒尔当（Laetitia Jourdan）著；康宁，宫鑫，刘婷婷译. -- 北京：人民邮电出版社，2017.8
ISBN 978-7-115-46526-9

Ⅰ. ①大… Ⅱ. ①克… ②利… ③康… ④宫… ⑤刘… Ⅲ. ①数据处理－启发式算法 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第178040号

版权声明

◆ 著　【法】Clarisse Dhaenens　Laetitia Jourdan
译　康　宁　宫　鑫　刘婷婷
责任编辑　李　强
责任印制　彭志环

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路 11 号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
固安县铭成印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：13.75　2017 年 8 月第 1 版
字数：160 千字　2017 年 8 月河北第 1 次印刷

著作权合同登记号　图字：01-2017-1802 号

定价：69.00 元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

致　谢

本书概述了大数据的元启发式算法，参考了大量文献。本书两位作者自2000年至今工作于法国里尔大学，CRIStAL实验室（计算机科学、信号和自动化研究中心），法国国家科学研究院，INRIA里尔诺德欧洲研究中心（法国国家计算机科学与应用数学研究所）。我们的学生及同事与我们一起完成本书的出版工作，在此向他们表示由衷的感谢。

我们特别感谢Aymeric Blot，Fanny Dufossé，Lucien Mousin和Maxence Vandromme，他们阅读了本书第一版，并更正了里面的错误。感谢Marie-Elénore Marmion仔细阅读并评论了本书内容。

我们还要感谢Nicolas Monmarché和Patrick Siarry，是他们建议撰写本书，并展示了极大的耐心！很抱歉，我们花了这么久时间才完成本书。

最后，我们要感谢我们家人给予的支持和爱。

Clarisse Dhaenens和Laetitia Jourdan

前　言

大数据：是流行语还是真正的挑战?

大数据既是流行语又是真正的挑战。一方面，尽管有人尝试定义“大数据”这一术语，但目前该词尚没有准确的定义。事实上，不同的人在使用大数据这个术语时，大数据一词的意义不尽相同。它可以被看作是一个流行语：每个人都谈论大数据，但没有人真正在应用大数据。

另一方面，大数据的特征往往用三个“V”来概括：Volume(规模化)、多样化(Variety)和快速化(Velocity)，这三项特征在大数据处理流程的不同阶段面临着大量的新技术挑战。大数据处理流程的不同阶段在图 I.1 中以非常简单的方式呈现出来。

从数据生成、存储和管理开始，数据分析有助于决策的制定。如果需要其他附加信息你可以重复该过程。而且每个阶段都会出现一些重要的挑战。

实际上，在数据的生成和捕获期间，一些挑战可能与技术相关，如实时数据的获取技术。然而，在这个阶段，挑战也与如何确定有意义的数据相关。

数据存储和管理阶段，会出现两个关键挑战：第一，数据存储及传输的基础设施；第二，提供用于分析的可用数据的概念模型。

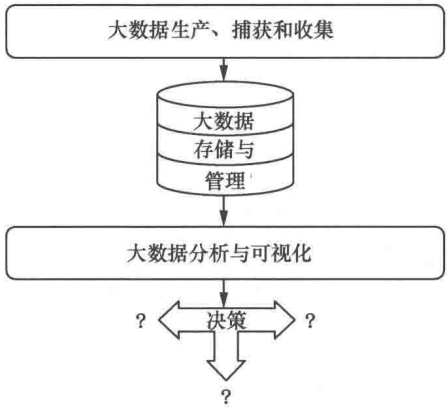

图 1.1　大数据处理流程的主要阶段

接下来，分析阶段的挑战是：如何处理异质化的海量数据，特别是在知识提取过程中，必然会遇到一些未知模式，所处理的数据性质使数据分析可能变得非常复杂，这是数据挖掘的核心。解决数据挖掘问题的一种方法是将问题建模为优化问题。在大数据环境中，大多数问题规模巨大。因此，元启发式算法似乎是解决这些问题的好办法。正如我们将在下文看到的，元启发式算法不仅适合解决大规模的问题，而且适用于处理大数据的其他方面，如多样性和速度。

本书的目的是介绍元启发式算法如何应对大数据环境（特别是在数据分析阶段）所带来的一些挑战。

本书由三部分组成。第一部分包括 3 章，目的是让读者更好地理解这部分之后的内容。

第 1 章“优化与大数据”，将介绍大数据环境中的主要问题。首先揭示大数据的特点，重点介绍数据分析阶段，更准确地说，是介绍数据挖掘任务。本章指出如何将数据挖掘问题视为组合优化问题，并向读者阐明为什么元启发式算法适合解决其中的一些问题。本章其中一节也专门讨论算法的性能评估。和数据挖掘过程一样，算法评估也必须遵循特定的规则。

第 2 章介绍元启发式算法。首先，本章提出元启发式算法相关的常见概念，

然后描述最广为人知的元启发式算法：区分基于单一解决方案和基于总体的算法。其中一节专门分析多目标元启发式算法，因为许多算法被建议用来处理数据挖掘问题。

第 3 章讨论并行优化，以及元启发式算法并行解决大型问题的方式。并行优化不仅可以处理大问题，而且能提供更好的质量解决方案。

本书第二部分为核心内容，由以下 4 章组成，每一章包含一项数据挖掘任务，并讲解如何使用元启发式算法来处理这些任务。

第 4 章是该部分的起始章，专门分析聚类算法。首先，介绍聚类任务，聚类的目的是将类似的对象进行归类，同时介绍一些经典的解决聚类任务的方法。其次，将聚类任务建模为一种优化问题，专注于通常使用的质量度量、多目标解析方法以及对元启发式算法中解的表示。最后，对多目标方法进行概述。本章最后分析聚类任务中的一个具体的难点：如何评估和验证聚类解决方案的质量。

第 5 章讨论关联规则。首先，本章描述相应的数据挖掘任务和经典算法，即先验算法；其次，指出如何将该任务建模为优化任务；最后，专门分析处理该任务的元启发式算法。根据所考虑的规则类型区分元启发式算法，规则类型包括分类关联规则、定量关联规则、模糊关联规则。本章最后用一张总表汇总了本领域的最重要文献。

第 6 章专门分析监督性分类法。数据挖掘非常重要，涉及通过已知类的观察信息，预测新观察到的类。本章首先介绍分类任务，以及标准分类方法；其次，从优化角度介绍其中一些方法，以及使用元启发式算法来优化这些方法。本章最后，专门分析如何使用元启发式算法搜索分类规则，分类规则被视为关联规则的特殊情况。

第 7 章探讨旨在减少属性数量和提高分类性能的分类特征选择。本章使用

了第 6 章提出的关于分类的几个概念。在介绍特征选择的一般概念后，本章将其建模为优化问题，然后呈现解决方案及其相关联的搜索机制的不同表示方法。本章最后概述用于特征选择的元启发式算法。

本书最后一部分由一章（第 8 章）组成，专门探讨用于数据挖掘的框架。每种框架都配有一个简短的对比调查。

通过浏览不同章节的内容，读者将会了解迄今为止元启发式算法应用于解决大数据环境中存在的问题的方式，重点是数据挖掘部分，为优化行业提供了许多具有挑战性的机会。

目　录


parameters.contains("age")){
and p.age
ypedQuery<Person> query
if(parameters.contains(
query.setParameter("name",

第1章

优化与大数据

“大数据”一词是指拥有不同来源的海量信息。因此，大数据不仅指数据量巨大，而且也指数据类型多样，处理速度快，频率高。本章试图从不同角度定义大数据的含义，并且侧重于大数据分析方法，这是大数据环境带来的主要挑战。

1.1 大数据环境

如图 1.1 所示，自 2011 年以来“大数据”一词在 Google 上的搜索请求次数呈指数级增长。

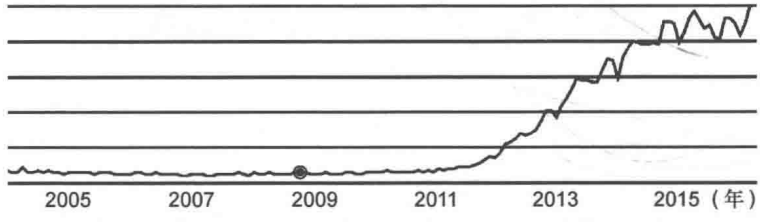

图 1.1 “Big Data”（大数据）一词在 Google 网站上请求次数的变化

我们如何解释人们对这个话题表现出来的日益增长的兴趣呢？对这一问题

的反应，有些可能成为一种模式，例如，我们知道世界上每天生成 2.5 万 MB 的数据，今天 90% 的数据是在最近两年内生成的。这些数据来自不同行业或组织的各个部门：传感器用于收集气候信息、社交媒体网站的帖子、数字图片和视频、购买交易记录和手机 GPS 信号等[IBM 16b]。收集到的这些数据被记录、存储和分析。

1.1.1 大数据环境示例

大数据出现在多种多样的环境中，生成了大量复杂的数据。每一种环境都存在挑战。在此，我们列举一些大数据环境的示例：

- 社交网络：社交网络中生成的数据可以用海量来形容。实际上，据估计，每月约 2 亿活跃用户发送了 120 亿条推文，在 YouTube 上观看了 40 亿小时的视频，在 Facebook 上分享了 300 亿条内容[IBM 16a]。而且这些数据具有不同的格式、类型。

- 交通管理：在创建智慧城市的背景下，城市内的交通成为一个重要的问题。近年来诸如智能手机、智能卡和各种传感器等科技产品被广泛采用，可以用来收集、存储和可视化城市活动（如人和交通流动）的各类信息，这些科技产品以使创建智能城市成为可能。然而，这也表明人们需要管理大量收集到的数据。

- 医疗保健：2011 年，医疗保健领域的数据全球规模估计为 150EB。医疗保健数据十分独特，难以处理，因为：（1）数据来源多样（不同的源系统，不同格式的文本和图像）；（2）数据为结构化和非结构化数据；（3）数据定义可能不一致（负责数据填充的人员不同，数据定义也可能不同）；（4）数据十分复杂（难以确定标准流程）；（5）数据受监管要求而变化[LES 16]。

- 基因组研究：随着 DNA 测序技术的迅速发展，现在我们可以鉴定超过 100 万个 SNP（遗传变异），大规模全基因组关联研究（GWAS）已经成为现实。基因组研究的目的是要跟踪“遗传”变异，这类变异有可能可以解释疾病的遗