隐私敏感移动性模式网络的净化方法研究

张海涛 著

科学出版社

隐私敏感移动性模式网络的净化方法研究

张海涛　著

科学出版社

北京

内 容 简 介

本书面向移动轨迹数据交易中的隐私及安全问题，提出了隐私敏感移动性模式网络的净化方法。主要内容包括：移动性模式网络构建方法，基于时空及网络特征的隐私敏感空间区域分类方法，基于隐私敏感移动性模式网络的推理攻击分析，隐私敏感移动性模式网络的净化方法及实现等。

本书面向的读者对象为 GIS 及相关专业的本科生或研究生，以及从事 LBS 相关应用开发和技术研究的工程技术人员。

图书在版编目（CIP）数据

隐私敏感移动性模式网络的净化方法研究 / 张海涛著. —北京：科学出版社，2019.2

ISBN 978-7-03-060463-7

Ⅰ. ①隐… Ⅱ. ①张… Ⅲ. ①互联网络－个人信息－隐私权－网络安全－数据保护－研究 Ⅳ. ①TP393.083

中国版本图书馆 CIP 数据核字（2019）第 014025 号

责任编辑：周 丹 沈 旭 邢 华 / 责任校对：樊雅琼
责任印制：张 伟 / 封面设计：许 瑞

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

北京九州逸驰传媒文化有限公司印刷
科学出版社发行 各地新华书店经销
*
2019 年 2 月第 一 版 开本：720 × 1000 1/16
2019 年 2 月第一次印刷 印张：9 插页：2
字数：190 000

定价：99.00 元

（如有印装质量问题，我社负责调换）

序

2009 年 2 月 6 日，15 名来自社会科学、计算机科学和物理学的重要科学家联名在 *Science* 上发表文章 *Computer social science*，宣告了计算社会科学的诞生。

计算社会科学是倡导使用大量新兴数据研究人类集体行为的研究范式，以前所未有的广度、深度及规模搜集与分析数据。与此同时，计算社会科学的兴起与发展也遭遇着重重阻碍。其中，最令人头疼的是数据获取和使用中涉及的个人隐私问题。一次偶然的违背个人隐私事故的发生，就可能使社会对信息共享深恶痛绝，甚至会颁布一些扼杀计算社会科学发展的法律条文。但是，计算社会科学的研究不能只集中在私人公司和政府部门内部，因为这会使得只有拥有特权的学术研究者才能使用独一无二的“秘密”数据，发表无法被别人评价和复制的论文。从长远来看，这不利于知识的积累、验证与传播。因此，制定合理的规章制度，研发新型的隐私保护技术，降低信息泄露风险，保留数据的研究价值，已成为计算社会科学发展的一个关键。

通信领域中电信运营商拥有的包含用户位置的移动轨迹数据，具有用户数量多、数据拥有者相对集中、时空尺度大等传统移动轨迹数据不可替代的优势。将电信运营商的移动轨迹数据与众多行业的专题数据进行集成、关联分析以及数据挖掘，可以发现具有语义隐喻信息的移动性模式。借助这些移动性模式，相关领域的科研工作者可以开展人口流动与资源、经济等动态变化特征之间关系的科学研究。但是，电信运营商拥有的移动轨迹数据主要从接近用户身体的移动设备（如智能手机等）在接入移动通信网络时发送的移动实时位置信令数据中获取，加重了隐私泄露的危险性。

南京邮电大学张海涛博士的科研团队在国家自然科学基金项目、江苏省自然科学基金项目、江苏省重点研发计划（社会发展）项目的资助下，开展了针对电信运营商的移动轨迹数据发布及共享中的隐私保护技术的研究。有别于传统的针对数据级别隐私问题的研究，该团队重点关注从移动轨迹数据中挖掘移动性模式的隐私保护技术。该书选取更有助于计算社会科学应用研究，但也更具挑战性的移动性模式网络为研究对象，研究可以保持移动性模式网络的结构特征，以用于了解复杂系统的宏观特征、发现隐藏在复杂系统中的机制规律，同时又能消除更具威胁性和隐蔽性的网络推理攻击的隐私保护技术。该书是作者在 2016 年出版的《基于时空关联规则推理的 LBS 隐私保护研究》的姊妹篇，书中提出的移动性模

式网络构建方法、基于时空及网络特征的隐私敏感空间区域分类方法、基于隐私敏感移动性模式网络的推理攻击分析、隐私敏感移动性模式网络的净化方法及实现等研究成果，对推动电信运营商开展电信大数据交易，促进地理信息科学领域、社会公共安全管理领域的学者开展隐私保护的知识挖掘与分析研究具有重要意义。

蒋国平
南京邮电大学副校长、教授、博士生导师
2018年9月6日

前　言

位置隐私是近年来学术界与工业界关注的热点。是数据开放共享促进创新应用，还是严格隐私保护保证数据安全，孰轻孰重？这是早期争论的焦点。大数据、人工智能等新兴技术的迅猛发展、国内外一系列隐私泄露事件的极大社会反响，使科研人员逐渐认识到研究同时实现隐私保护与数据可用性技术的重要性。

数据级别的位置隐私保护是国内外学者关注研究的热点，但是基于敏感知识的位置隐私推理攻击因具有预测性和普适性，往往更具攻击性与威胁性。为此，作者申请了国家自然科学基金项目（41201465，基于大时空范围 LBS 匿名集的推理攻击及隐私保护方法）和江苏省自然科学基金项目（BK2012439，对抗基于时空关联规则推理攻击的 LBS 隐私保护研究），在项目的资助下，研究了基于攻守双方对等感知信息级别的 LBS 隐私保护机制。在分析移动对象数据的时空关联规则推理与防护方法以及移动对象数据与匿名集数据不同特性的基础上，结合 LBS 长期、连续、在线服务的特点，研究了时空关联规则的概率化挖掘与推理攻击方法、基于动态阻止推理攻击的渐进式隐私保护方法，以及阻止推理攻击的匿名保护模型的量化评估与优化方法。并于 2016 年将这些成果在《基于时空关联规则推理的 LBS 隐私保护研究》专著中出版。

当前，复杂网络的研究与应用得到了蓬勃发展，从网络的视角研究复杂空间系统成为一个新的方向。挖掘移动轨迹数据得到的移动性模式网络，可以看成生成数据的复杂空间系统的拓扑抽象。分析移动性网络的结构特征，有助于了解如社交网络、城市系统、流行传染病等复杂系统的宏观特征，发现隐藏其中的机制规律。

但是，技术是一把“双刃剑”。基于移动性模式网络的推理分析，也可能会被攻击者用于对用户位置隐私的推断。例如，当移动性模式网络中节点对应的空间区域与具有敏感属性信息的外源专题地理数据（如发电厂、煤气站、油库、军事禁区、宗教、娱乐场所等所处的地理位置数据）产生交集时，移动性模式网络也就具有隐私敏感属性。隐私敏感移动性模式网络通常具有复杂的网络拓扑结构，使传统的针对单一模式的防护方法难以奏效。研究应对基于敏感移动性模式网络的推理攻击并保证网络可用性的防护方法更具挑战性。

为此，作者申请了江苏省重点研发计划（社会发展）项目（BE2016774，电信大数据中面向社会公共安全管理的敏感移动性知识的隐私设计方法研究），在

项目的资助下，研究可以保持移动性模式网络的结构特征又能消除网络推理攻击的隐私保护技术。本书内容涵盖了已经取得的成果。

在本书的撰写过程中，课题组的研究生朱云虹、武晨雪、汪佩佩、陈泽伟、刘钊、周欢、蒋继飞、胡志鹏、于晨光等参与了部分章节的图表制作、文字校对等工作，在此表示深深的谢意！本书的写作过程，得到了许多专家的支持和帮助，特别感谢南京师范大学的张书亮教授对书中内容提出的宝贵意见！在本书编写过程中得到了科学出版社的大力支持，周丹编辑做了大量的工作，使本书得以顺利出版，在此一并表示衷心的感谢！

本书涉及的知识领域广泛，而今科学技术发展日新月异，又由于时间和水平有限，书中难免有疏漏和不足之处，敬请读者批评、指正！

张海涛

2018 年 8 月 30 日

目　录

第1章 绪 论

1.1 研究背景

近年来，随着互联网、云计算、物联网和智能终端的迅速发展，数据处于爆炸式增长阶段[1]。中国信息通信研究院发布报告：大数据市场在 2010 年至 2015 年期间增长了 3 倍[2]。对于国家来说，数据是战略资源，地位堪比工业时期的石油资源，是衡量一个国家综合国力的标准之一；对于企业来说，数据是其核心竞争力，决定着企业的长远发展。

1.1.1 数据分析的价值

可穿戴设备、平板电脑、笔记本电脑、智能手机等类型的智能终端，其内置的位置感知设备及应用软件产生了大量的移动轨迹数据①[3]。接入电信运营商移动通信网络的智能手机产生的移动轨迹数据，具有用户数量多（截至 2016 年全球手机用户数接近 48 亿，截止到 2017 年 6 月底我国手机用户数量已达到 13.6 亿）、数据拥有者相对集中（目前，一个国家一般只有三四家电信运营商）、时空尺度大（接入移动通信网络即可从移动实时位置信令数据中获取用户的位置，且空间上都是全球覆盖）等卫星定位数据不可替代的优势。

电信运营商从内部管理的角度出发，通过对移动轨迹数据的分析以及数据关联，可实现系统优化、行业及个人客户的业务定制等增值服务[4-9]。但在应用驱动创新的需求下，电信运营商不再满足于只是提高企业自身业务，而是转向数据资产的平台化运营，即通过数据交易平台进行数据的业务化封装与运营。将电信运营商的移动轨迹数据与众多行业的专题数据进行集成、关联以及数据挖掘分析，可发现具有语义隐喻信息的移动性模式[10-14]。这些移动性模式不仅可以为相关行业应用提供一定的辅助决策[15]，促生个性化医疗、数字金融、精准营销等新型商业模式，而且可为相关领域的科学工作者开展智能交通[16-18]、城市规划[19-21]、疾病传播[22-24]、人口流动[25-31]等人口流动与资源、经济等动态变化特征之间关系的研究工作提供重要支撑。

① McKinsey 全球机构的报告显示，个人位置数据池在 2009 年的数据量为 1PB，并以每年 20%的速度增长。根据联合国全球地理空间信息管理指南的预算，每人每天生成 2.5MB 的数据，大部分数据产生于内置位置感知功能的智能终端设备。

1.1.2 隐私安全问题

开放电信运营商的移动轨迹数据在便于公众研究和使用的同时，也会带来隐私泄露的风险。用户的真实身份、特殊职业、宗教信仰、政治党派、性取向等隐私敏感信息的泄露，会给其生活、工作带来严重的干扰，甚至会产生人身安全风险。同时，电信运营商的移动轨迹数据主要从接近用户身体的移动设备在接入移动通信网络时发送的移动实时位置信令数据中获取，这更加重了隐私泄露的危险性。攻击者掌握移动轨迹数据和相关背景数据的数量、规模，以及进行数据分析能力的不同，对用户隐私安全产生威胁的程度也不同。隐私攻击的类别主要包括：数据级别和知识级别[32]。

（1）数据级别。电信运营商的移动轨迹数据产生于移动用户手机设备。这种基于直接测量的数据产生方式，使移动轨迹数据跟踪记录了用户在时空区间的运动过程。用户的移动轨迹数据包含了其在某个特定时刻的空间位置信息。用户轨迹信息的泄露，会产生隐私共享、隐私攻击的安全问题。因此，电信运营商的移动轨迹数据在提供给第三方进行集成分析时，必须进行去标识化的匿名处理。

但是，随着移动互联网、社交网络等相关技术的快速发展，用户的信息有了更加广泛的分布。攻击者可将匿名化处理的移动轨迹数据与外源数据进行连接分析，多源、多维的数据连接分析可以无限放大用户的个体独特性，从而造成对用户的重标识攻击。例如，文献[33]指出，对于任一用户的移动轨迹数据，只需要 4 个轨迹点的匹配运算，即可实现 95%的用户标识识别度。为此，一些学者提出了基于 Mixzone、PathCloaking 技术阻断攻击者对移动轨迹数据跟踪的方法。但是，这些方法通常面临基于速度等参数的推理攻击问题。此后，一些学者提出了轨迹 *k*-匿名的方法[34]。

（2）知识级别。轨迹 *k*-匿名的处理方法可以在一定程度上应对数据级别的隐私安全问题。但是，当攻击者使用数据挖掘等工具从电信运营商的移动轨迹数据中得到移动性模式，并将其与具有敏感属性信息的外源专题地理数据（如发电厂、煤气站、油库、军事禁区、宗教、娱乐场所等所处的地理位置数据）进行关联时，可发现具有隐私敏感属性的移动性模式。虽然隐私敏感的移动性模式反映的是大量用户群体的移动性规律，并不涉及特定个人的信息，但是，隐私敏感的移动性模式对于任何满足移动性模式前置条件（例如，移动用户的位置与移动性模式的位置相匹配时）的用户，都会产生位置推理的隐私威胁[35]。同时，隐私敏感的移动性模式具有预测性，攻击者还可以基于模式推理侵犯用户未来的位置隐私。因此，电信运营商在发布及共享其移动轨迹数据时，必须对数据进行净化处理，以消除隐藏其中的具有隐私敏感属性的移动性模式。

1.2 国内外研究现状

按照计算机科学与统计科学的分类，移动轨迹数据中的隐私敏感移动性模式的消除方法，隶属于数据挖掘领域中隐私保护数据挖掘（PPDM）和隐私保护数据发布（PPDP）的理论与方法范畴。采用失真和阻塞技术抑制敏感知识、实现敏感隐藏是其重要的实现方式[36-38]。目前，敏感知识隐藏方法的研究主要集中在关联规则、序列模式及分类模型 3 个方面。其中，在关联规则方面取得的成果最为丰富，国内外学者提出了包括启发式、边界修改及精确隐藏的系列方法[39]。Atallah 等[40]在 1999 年首次提出了基于启发式的关联规则隐藏方法，算法的基本思想是：从原始数据库中有选择性地清理事物，以实现敏感关联规则的快速隐藏。此后，文献[41]提出了系列的改进方法。文献[42]首次提出了边界修改方法，通过修改原始数据库中频繁项集和非频繁项集中的边界，并采用贪心算法进行数据修改，实现了敏感规则的隐藏和最小的边界修改。Menon 等[43]首次将频繁模式的隐藏转化为约束满足问题，并使用整数规划进行求解，开启了精确隐藏方法。目前，专门针对轨迹数据中敏感移动性模式的知识隐藏方法主要包括：轨迹数据发布、隐私感知的分布式分析及移动性序列模式的动态隐藏。

1.3 存在的问题

现有的针对轨迹数据中敏感移动性模式的隐藏方法存在的共性问题是：涉及的移动性模式，都是从时空数据库角度定义的关联规则、序列规则、序列模式等简单移动性模式。依据简单移动性模式间的共同项，将大量简单移动性模式连接在一起，可以构建以模式项为节点、模式项之间连接为方向边的移动性模式网络。基于移动性模式网络的分析，相对于基于单一的简单移动性模式分析，更能够发现模式间的关联关系和整体结构特性。

轨迹数据中包含了移动用户为完成特定目标任务，在不同空间区域间运动的信息。移动性模式网络可以看成产生轨迹数据的复杂空间系统的拓扑抽象。对移动性模式网络的结构特征（如聚集系数、节点度、中心性等）进行研究分析，有助于了解如社交网络[44-51]、城市系统[52, 53]、流行传染病[54-56]等复杂系统的宏观特征，发现隐藏其中的机制规律。

但是，技术具有中立性，攻击者也可分析敏感移动性模式，构建具有隐私敏感属性的移动性模式网络。敏感移动性模式网络通常具有复杂的网络拓扑结构，这使得基于敏感移动性模式网络的推理攻击更具有威胁性和隐蔽性。因此，分析基于敏感移动性模式网络的推理攻击并设计相应的防护方法，具有必要性，也更具挑战性。

本书结合作者承担的国家自然科学基金项目(41201465,基于大时空范围 LBS 匿名集的推理攻击及隐私保护方法)、江苏省自然科学基金项目(BK2012439,对抗基于时空关联规则推理攻击的 LBS 隐私保护研究)、江苏省重点研发计划(社会发展)项目(BE2016774,电信大数据中面向社会公共安全管理的敏感移动性知识的隐私设计方法研究)的研究成果,提出了通过分析网络拓扑结构特征对隐私敏感移动性模式网络进行净化以消除对应隐私推理攻击的方法。研究成果对于推动运营商开展电信大数据交易、促进地理信息科学领域、社会公共安全管理领域的学者开展隐私保护的知识挖掘与分析研究具有重要意义。

1.4　本书章节安排

本书共包括 7 章,基本内容如下。

第 1 章绪论,介绍了本书提出方法的相关背景、研究意义、解决问题的基本思路及章节的组织安排等。

第 2 章基本概念,介绍了 Spark 大数据平台、机器学习、复杂网络、隐私设计方法的基本概念,为后续章节内容的学习奠定基础。

第 3 章移动性模式网络构建方法,介绍了从移动轨迹数据中构建移动性模式网络的两种方法:基于序列模式挖掘的方法和基于图挖掘的方法,并通过实验对两种方法的性能进行了对比分析。

第 4 章基于时空及网络特征的隐私敏感空间区域分类方法,介绍了一种通过统计、分析移动性模式网络中所有节点对应空间区域的时空与网络特征,进行节点敏感属性判定的监督分类方法,并在 Spark 大数据平台上进行了实现。最后,实验分析了方法的性能。

第 5 章基于隐私敏感移动性模式网络的推理攻击分析,定义了基于隐私敏感移动性模式网络对用户位置隐私进行推理攻击的模型,并给出了基于网络连通性分析的推理攻击场景。

第 6 章隐私敏感移动性模式网络的净化方法及实现,介绍了一种包括数据可用性与网络安全性评估模型设计、网络类型判定、节点重要性分析等关键步骤的隐私敏感网络净化方法,并在 Spark 大数据平台上进行实现。最后,实验分析了方法的有效性和适用性。

第 7 章总结与展望,对本书的主要内容进行了总结,并对今后的研究方向进行了展望。

第2章 基本概念

本章介绍 Spark 大数据平台、机器学习、复杂网络、隐私设计方法的基本概念，为后续章节内容的学习奠定基础。

2.1 Spark 大数据平台

本节介绍 Spark 大数据平台的基本特点、运行模型、生态圈等，为后续相关算法的实现以及实验性能测试奠定基础。

2.1.1 Spark 简介

Spark 系统是一个通用、开源的快速并行计算框架，由加利福尼亚大学伯克利分校的 AMP 实验室开发。与 Hadoop 相比，Spark 系统具有以下特点[57, 58]。

（1）Spark 具有 Hadoop 的优点，也根据 MapReduce[59]算法执行分布式运算。但与 Hadoop 不同的是，Job 的中间输出及计算结果能够存储在内存上，不需要反复地读写 HDFS。因此，Spark 程序的运行速度远高于 Hadoop 程序。另外，Spark 更擅长迭代的 MapReduce 运算，这使其更适合于需要大量迭代运算的机器学习和数据挖掘算法。

（2）Spark 采用 Scala 语言开发实现。Scala 是以 Java 虚拟机为基础的新型程序设计语言，能在处理数据方面提供更高的可靠性和更快的计算性能。Spark 和 Scala 的紧密集成，使开发者能够采用类似于操作本地数据对象的方法去读写分布式环境中的数据集。

（3）Spark 拥有丰富的 API 接口。编程人员能够简易地实现相关算法。一般情况下实现相同功能的算法，Spark 的代码量是 Hadoop 的 1/10 或 1/100。因此，可大大提高程序开发的效率。

2.1.2 Spark 运行模型

1. 基本架构

Spark 部署应用运行在一个或者多个集群之上，Spark 的基本架构如图 2.1 所

示。其中，SparkContext 驱动 Spark 应用的运行，Cluster Manager 分配资源和任务调度，Executor 执行 Spark 调度的任务。各部分具体的功能如下：

（1）集群管理器（Cluster Manager）。Cluster Manager 核心任务为资源分发与管理。Cluster Manager 分发资源具有最高优先级。

（2）客户端驱动程序（Driver App）。Driver App 又称应用程序，用于将任务程序转换为弹性分布式数据集（resilient distributed datasets，RDD）[60]和 DAG（有向无环图），并负责与集群管理器之间的通信。

（3）工作节点（Worker Node）。Worker Node 的工作流程是，针对 Spark 的应用程序，先从 Cluster Manager 处获取分配的资源，然后创建 Executor，最后将获得的资源分配给 Executor。同时，还负责资源信息与 Cluster Manager 的同步。

（4）执行者（Executor）。Executor 是 Worker Node 上运行计算任务的一个进程。每个 Driver App 都有一个独立的 Executor。Executor 的核心是执行任务，同步 Worker Node 和 Driver App 之间的信息，以及存储数据资源到内存或磁盘。

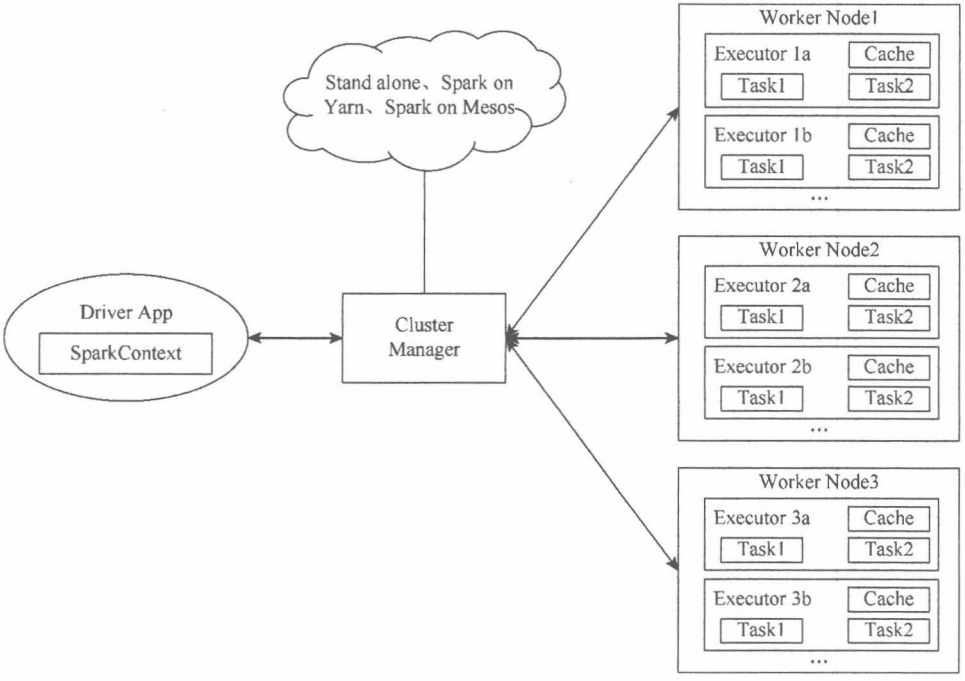

图 2.1　Spark 基本架构

具体执行流程如下：

（1）Driver App 的 SparkContext 与 Cluster Manager 进行通信。

（2）Cluster Manager 启动 Executor 进程。

（3）应用程序代码（如 JAR 文件或 Python 文件）经过 Cluster Manager 传送到 Executor。

（4）SparkContext 传送 Task 到 Executor，执行数据处理、计算和存储等操作。

2. 调度模式

Spark 的调度运行既可使用本地模式，也可使用分布式模式。分布式模式包括 3 个应用场景：Stand alone、Spark on Yarn 和 Spark on Mesos。Spark on Yarn 和 Spark on Mesos 分别使用 Yarn 和 Mesos 资源管理系统进行调度运行。Stand alone 模式，又称独立模式，不依靠额外的资源管理系统，即可完成资源管理和容错功能。Stand alone 模式的节点分为 Master、Client 及 Worker。Driver 程序能够运行在 Master 节点上，也能够运行在本地 Client 端。通过 Spark-Shell（交互式工具）提交 Spark 代码，Driver 程序运行在 Master 上。在 IDEA 和 Eclipse 等平台中利用 Spark-Submit 工具提交代码或利用“spark：//master：7077”方法执行 Spark 代码，则 Driver 程序运行在 Client 端。

3. 计算模型

RDD 是 Spark 进行数据处理和运算的核心，是在 MapReduce 基础上的一种扩展。在集群环境中 RDD 是一种内存抽象形式的数据集，运行在多个分布节点上，在并行计算阶段能够实现高效的数据共享。RDD 在迭代运算方面具有很好的表达能力，可以有效执行需要大量迭代运算的图数据处理，实现经典的图计算模型[61, 62]。

RDD 使用划分函数对数据集进行处理，实现内存的分布式控制。在程序运行时，Spark 系统根据可用资源有效地去放置可划分的数据，控制数据存储的位置（内存或磁盘）与分区，实现容错与并行，如图 2.2 所示。

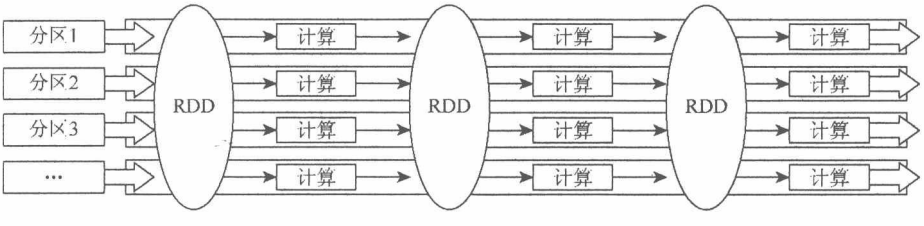

图 2.2 RDD 计算模型

RDD 定义了“变换 + 动作”的编程规范（如 join、map、filter 等变换和 collect、count 等动作），使得处理的数据在整个的计算流程中均可使用：首先根据依赖关系进行串联，然后按照顺序缓存每一个转换操作，最后调用 Action 执行计算功能。