DOI:10.3969/j.issn.1009-6868.2016.02.011
网络出版地址:http://www.cnki.net/kcms/detail/34.1228.TN.20160229.1613.004.html

应用驱动的大数据挖掘
Application-Driven Big Data Mining

中图分类号:TP393 文献标志码:A 文章编号:1009-6868 (2016) 02-0049-004

摘要: 认为大数据挖掘的核心和本质是应用、数据、算法和平台4个要素的紧密结合。从大数据的特点出发,结合大数据挖掘的案例,提出大数据挖掘中的平台架构、数据获取和预处理、算法的选择和集成都是应用驱动的。强调大数据挖掘的目标来自实际应用的真实需求,只有结合具体应用数据和适合应用的算法,利用高效处理平台的支撑,并将挖掘到的模式或知识应用在实践中,才能体现大数据挖掘的真正价值。

关键词: 大数据;数据挖掘;应用驱动;FIU-Miner;高端制造业

Abstract: The core of big data analysis is the combination of applications, data, algorithms and platforms. Big data mining platforms, algorithms, and big data itself are driven by applications. Big data mining tasks come from real applications. With specific application data and appropriate algorithms, using efficient processing platform, digging into the patterns or knowledge in practice, big data mining platform can show its true value.

Key words: big data; data mining; application-driven; FIU-Miner; advanced manufacturing

李涛/LI Tao[1,2]
刘峥/LIU Zheng[1]
周绮凤/ZHOU Qifeng[3]
(1. 南京邮电大学 计算机学院,南京 210023,中国;
2. 佛罗里达国际大学 计算机学院,迈阿密 33199,美国;
3. 厦门大学 自动化系,厦门 361005,中国)
(1. School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
2. School of Computing and Information Sciences, Florida International University, Miami 33199, USA;
3. Department of Automation, Xiamen University ,Xiamen 361005, China)

- 应用驱动的大数据挖掘能够有效处理大数据的复杂特征,真正体现大数据挖掘的价值
- 大数据的获取与预处理是应用驱动大数据挖掘的前提
- 应用驱动的大数据获取和预处理能够有效地连接企业业务需求和数据挖掘平台

1 大数据时代的发展

数字化变革推动信息技术(IT)和通信技术(CT)的飞速发展,人类社会所产出的信息总量呈爆发式增长。一方面,各行各业在日常运作中借助IT产生和存储了海量的运营数据,如商业运营、金融证券、健康医疗、科学研究等,分布在世界各地的10 000多家沃尔玛超市1 h需要处理百万条以上顾客的消费记录,数据量高达2.5 PB[1],欧洲的大型电子对撞机每天产生的记录有500 EB[2];另一方面,CT使得全世界数十亿用户通过互联网链接在一起。目前全球移动互联网的流量每月约4.2 EB,思科预计:2019年全球移动互联网的流量会增长到每年292 EB[3]。

这些海量数据被称为大数据。维基百科对大数据的定义是:"大数据是由于规模、复杂性、实时性而导致的无法在一定时间内用常规软件工具对其进行获取、存贮、搜索、分享、分析、可视化的数据集合"[4]。知名技术咨询公司Gartner对大数据的定义是:"大数据是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"[5]。

大数据技术的发展使得收集、处理、管理、分析在各行各业产生的海量数据成为可能:企业利用大数据技术理解客户的属性和行为,可以提供给客户更好的个性化服务,并可以利用大数据技术改善和优化商业流程,提高企业的运营效率;政府通过大数据技术来更智能的管理城市,包括公共交通、医疗服务、可持续性发展[7]等;超市可以向用户推销所需的商品;车险公司可以知道客户的驾驶水平;甚至2012年的美国总统大选,奥巴马的竞选团队也是依赖卓越的大数据分析取得胜利。大数据已经融入各行各业,大数据时代已经来临。

收稿时间:2016-02-03
网络出版时间:2016-02-29

2 大数据的特点与理解

2.1 大数据的特点

目前业界普遍用4V的特点来衡

量大数据所带来的挑战[7]，从数据本身的表现形式上描述了大数据与以往部分抽样的“小数据”的主要区别。

大量（Volume）：大数据的体量巨大，从TB级别跃升到PB级别；

多样（Variety）：大数据面对数据类型种类繁多，例如地理位置等结构化数据，事件日志等非结构化数据，还包括图片、视频等多媒体数据等；

高速（Velocity）：大数据产生和累计的速度快，要求处理速度快，做到实时分析，和传统的离线方式的数据挖掘技术有着本质的不同；

价值（Value）：大数据所蕴含的价值密度低，但有效价值高，合理利用低密度价值的数据并对其进行正确、准确的分析，将会带来巨大的商业和社会价值。

从现有的一些大数据挖掘应用案例出发[8]，大数据挖掘的流程可以总结为：

（1）准确定义大数据挖掘问题的目标；

（2）获取大数据，并对收集到的大数据进行数据清洗等预处理；

（3）选择合适的大数据挖掘平台架构和算法；

（4）进行大数据挖掘；

（5）理解所发现的模式或应用所产生的知识。

可以看到：只有应用才能体现大数据的价值。在大数据挖掘的流程和案例中，可以充分体现出实际应用中大数据所具有的以下一些新的4V的特点：

变化性（Variable）：不同的应用场景、不同的研究目标下，大数据的机构和意义均会发生变化，在大数据的实际应用和研究中需要考虑具体的上下文，从而体现大数据的价值。

真实性（Veracity）：大数据应用的基础是真实、可靠的大数据，它们是保证分析结果准确、挖掘知识有效的前提，只有真实而准确的大数据才能获取真正有意义的结果。

波动性（Volatility）：大数据本身往往含有噪音，加上有时分析流程的不规范，导致不同的算法、不同的分析流程、不同的衡量标准下，会得到不同的分析结果。

可视化（Visualization）：数据可视化可以在大数据应用中直观地阐述分析的结果以及数据的意义，帮助用户更好地理解、应用大数据。

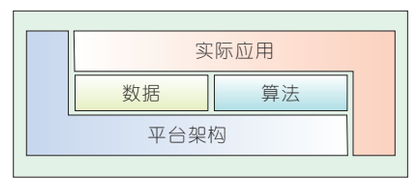
2.2 应用驱动的大数据架构

从上述大数据本身的表现形式上的4V特点出发，结合实际应用中大数据所具有的新4V特点，我们认为大数据的核心和本质是应用、算法、数据和平台4个要素的有机结合，如图1所示。大数据的基础是平台架构，数据和算法是大数据的核心，而实际应用是大数据的关键。上文所述的大数据挖掘的流程中，大数据挖掘的目标必须是来自实际应用的真实需求，只有结合具体应用数据和适合应用的算法，利用高效处理平台的有效支撑，并将挖掘到的模式或知识应用在实践中，才能提供量化、合理、可行、有价值的信息。这个应用、算法、数据和平台相结合的思想体现了大数据的本质和核心，可见大数据挖掘是应用驱动的，应用驱动的大数据挖掘能够有效处理大数据的复杂特征，体现大数据挖掘的价值。

3 应用驱动的大数据挖掘

3.1 应用驱动的大数据平台

一个高效的大数据平台可以有力地支撑海量数据的集成和数据挖掘算法，以及可视化的步骤执行，并可以利用规范的数据分析流程来保证结果的稳定性。传统的数据挖掘工具，如Weka、统计产品与服务解决方案（SPSS）等提供了友好的用户界面，但并不适合对海量数据进行挖掘分析。另外，最终用户很难对这些商业工具添加应用所需的合适算法。流行的数据挖掘算法库，如Mahout，提供了大量的数据挖掘算法，但需要数据挖掘专家来进行任务配置和算法集成，才能解决具体应用中的数据挖掘任务。最近出现的大数据挖掘产品，如Radoop等对于非基于Hadoop的算法支持有限，在多用户、多任务环境下的资源分配上也存在不足。



▲图1 大数据框架

应用驱动的大数据平台应该满足如下关键需求：

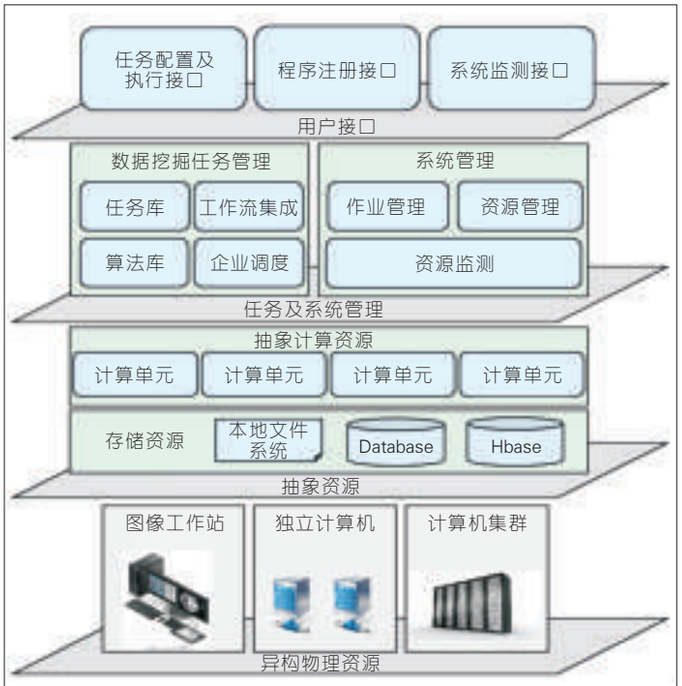
（1）人性化、友好的用户界面，快速任务配置；

（2）灵活的多语言，多算法集成；

（3）高效的分布式异构环境下的资源管理。

我们以一个快速、集成和用户友好的分布式数据挖掘系统（FIU-Miner）[9]为例介绍应用驱动的大数据平台如何满足这些需求。FIU-Miner友好的用户界面可以可视化地直接将现有算法配置成工作流，甚至无需编写任何代码，其他与挖掘任务无关的底层细节都由FIU-Miner进行管理。FIU-Miner不仅支持直接导入外部算法库来扩充分析工具集合，还会根据所导入算法的语言和运行环境自动分配对应任务到合适的计算节点。FIU-Miner可以支持各种异构的计算环境，包括PC、服务器、图形处理器（GPU）工作站等，同时根据算法实现、负载平衡、数据位置等因素来优化计算资源的利用率。

如图2所示的FIU-Miner的系统架构，包括用户界面层、任务和系统管理层、抽象计算资源层和异构物理资源层。抽象计算资源层屏蔽了不同物理环境给大数据挖掘带来的资源调度的复杂度，提高了分布式计算的效率；任务及系统管理层方便了不同数据挖掘算法的集成，多种分析任务的配置管理；友好的用户接口为基于FIU-Miner构建不同的大数据挖掘



FIU-Miner：分布式环境下数据挖掘的一个快速、集成和用户友好的系统

◀图2 FIU-Miner的系统架构

应用提供了极大的便捷，帮助数据分析人员方便有效地开展各项复杂的数据挖掘任务。

3.2 应用驱动的大数据获取与预处理

大数据的获取与预处理是应用驱动大数据挖掘的前提。以企业大数据挖掘为例，一个企业中所面临的大数据的任务多种多样，当确定大数据挖掘任务的目标时，企业对挖掘的对象和所能发现的知识往往缺乏理解，而大企业的业务流程复杂，具体业务逻辑和数据之间的对应关系十分琐碎，运营数据往往来自不同的数据源，具有不同的类型和格式，所以大数据通常无法预先规划和准备好，数据的获取是一个难题。在具体应用的大数据挖掘任务中，需要在数据的导入、整合上有很大的灵活性，只有通过业务人员和数据挖掘工程师的配合，不断尝试，才能有效地将企业的业务需求与数据挖掘的功能联系起来。在大数据获取过程中还需要根据应用需求注意数据聚合过程中的隐私保护，避免泄露用户的敏感信息。

由于大数据的多样性，所获取和整合的大数据通常还不能直接应用于数据挖掘算法，需要对数据进行预处理，结合具体应用处理数据的结构信息，抽象数据的语义信息等，并需要对所获得的大数据中的各种属性进行选择，剔除与应用无关的属性，或者引入额外的抽象测度等。大数据的质量是知识发现结果有效的保证，所以需要对数据中的噪音进行过滤，对缺失值进行处理。

3.3 应用驱动的大数据挖掘算法

数据挖掘领域中的很多算法都是从实际应用的具体需求衍生和发展出来的。从顾客交易数据分析到隐私保护数据挖掘，从文本数据挖掘到多媒体数据挖掘，从Web挖掘到社交网络挖掘，这些不同子领域的算法都是由应用推动的。数据挖掘是个交叉学科，融合了统计分析、数据库、信息检索、机器学习、模式识别、人工智能等领域的研究成果。大数据挖掘要以具体应用为驱动，根据应用数据特性，挖掘任务需求，选择、集成相应的数据挖掘和机器学习算法，并根据需要进一步进行研究，在实际问题中得到应用和验证。如基于关联规则和时间序列分析的分类算法就是关联规则发现和时间序列模式识别的有机结合；半监督学习和半监督聚类也是分类和聚类的融合结果。在处理高维、稀疏的数据时，数据的分布不明显，需要注意算法的可靠性。在处理复杂关系网络的数据时，需要根据应用的数据特征来研究能够处理异构信息网络的图挖掘算法。

4 应用驱动大数据挖掘的应用

4.1 高端制造业大数据挖掘挑战

高端制造业是指制造业中新出现的具有高技术含量、高附加值、强竞争力的产业，包括电子半导体生产、精密仪器制造、生物制药等。这些制造领域往往涉及严密的工程设计，复杂的装配生产线，大量的控制加工设备与工艺参数，精确的过程控制和材料的严格规范。随着信息技术在高端制造业中的普及，高端制造业中积累了大量的生成设计、机器设备、原材料、环境条件、生成流程等生产要素相关的历史数据，其中蕴含了对生产和管理有帮助的高价值信息。通过大数据挖掘，企业可以把隐藏在这些海量数据中有用的、深层次的信息挖掘出来，用来指导流程控制、生产调度、优化决策等方面，从而能够在实际应用中改进产品品质，提升产品性能和生产效率，最终达到提高企业行业竞争力的目的。

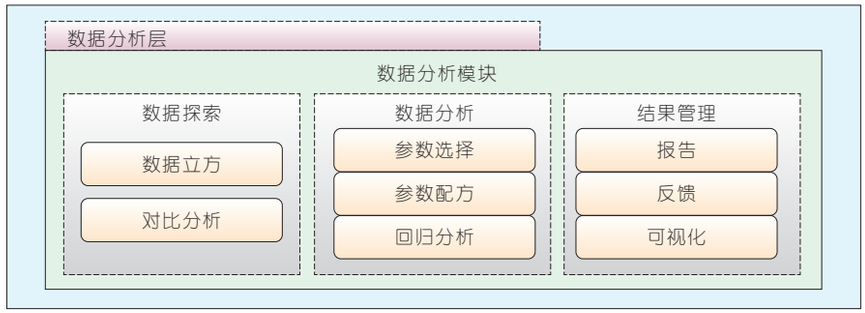
高端制造业中的数据挖掘面临很多挑战[10]，比如：如何有效分析大规模数据，如何保证对数据分析效率和分析结果的准确性等。在实际应用中，依靠传统信息系统从海量数据中进行查询和报警或单纯利用专家经验来分析和发现潜在有价值的信息已经变得不太现实。因此，企业需

要利用数据分析技术、工具或平台，智能地从大量复杂的生产原始数据中发现新的模式和知识作为改善生产过程的决策依据，系统性地提高生产效率。

4.2 等离子显示器制造中基于FIU-Miner 的大数据解决方案

四川虹欧显示器件有限公司就是利用大数据挖掘来提高等离子屏的生产良率。我们可以通过下面这个案例来阐述应用驱动的大数据挖掘。等离子显示器制造中大数据挖掘的难点是：自动化的生产方式中自动采集的数据急剧增长，需要强大的数据分析能力来支撑；大量的生成过程控制参数对高维数据分析的效率和结果的准确性提出了更高要求。这个过程本身就是对数据进行探索、分析和理解的一个循序渐进的迭代过程。因此，一个实用的系统应该提供一个集成的、高效率的分析平台来支持这个过程。

在平台方面，基于FIU-Miner，结合实际挖掘任务的具体需求和难点，我们在架构上增加了数据分析层，如图3所示。其中数据探索系统主要提供对数据的宏观理解和快速预览，以及敏感参数验证。利用联机分析处理（OLAP）技术帮助分析人员快速掌握挖掘任务相关数据的特性，指导后续的数据预处理，如属性选择和测度建立等。数据分析系统集成了根据实际大数据挖掘任务的需要所选择数据挖掘算法，包括参数选择、参数配置和回归分析。数据分析人员通过操作界面调用算法，聚焦具体的分析任务，并且算法对数据分析人员透明。结果管理系统基于业务分析结果产生分析报告，这些分析报告可以直接给决策者提供决策依据，同时报告系统也为领域专家提供收集反馈的接口。领域专家知识的引入对优化模型、改进算法具有很大的指导意义。



▲图3 数据分析层

5 结束语

大数据一词经常被用以描述和指代信息爆炸时代产生的海量信息，研究大数据的意义在于发现和理解信息内容及信息与信息之间的联系。文章从大数据本身的表现形式的4V特点出发，结合大数据挖掘的案例中体现的新4V特点，提出应用驱动的大数据挖掘思想，指出大数据的本质是应用、算法、数据和平台四个要素的有机结合。应用驱动的平台、应用驱动的数据获取和预处理、应用驱动的算法是大数据挖掘成功实施的关键。应用驱动的大数据挖掘在高端制造业的成功实施案例，验证了本文所提思想的正确性和可行性。未来，随着大数据挖掘技术的不断深入，应用驱动的大数据挖掘将会体现更大的价值和广泛的应用前景。

致谢

感谢南京邮电大学曾春秋、郑理老师在本篇文章的撰写过程中提出很多有意义的见解，并在相关工作中给予了很多帮助和贡献。

参考文献

[1] Data, Data Everywhere [EB/OL]. [2010-02-25]. http://www.economist.com/node/15557443
[2] HRUMFIEL G. High-EnergyPhysics: Down the Petabyte Highway [J]. Naure, 2011, 469 (19): 282-283
[3] BAMETT J T, SUMITS A, JAIN S, et al, Global Mobile Data Traffic Forecast, 2014-2019 [EB/OL].[2015-02-18]. http://www.ciscoknowledgenetwork.com/files/496_02-24-15_VNI_Mobile_Forecast_Prezo_for_CKN.pdf
[4] Big Data [EB/OL]. [2013-02-22]. https://en.wikipedia.org/wiki/Big_data
[5] GARTER. What Is Big Data [EB/OL]. [2014-10-20]. http://www.gartner.com/it-glossary/big-data
[6] 周绮凤, 李涛.大数据与计算可持续性[J]. 南京邮电大学学报, 2015(5): 20-31
[7] 严霄凤, 张德馨. 大数据研究[J]. 计算机技术与发展, 2013, 23(4): 168-172
[8] 李涛. 数据挖掘的应用与实践——大数据时代的案例分析[M]. 厦门：厦门大学出版社, 2015
[9] ZENG C, JIANG Y. ZHENG L, et al. Fiu-Miner: A Fast, Integrated, and User-Friendly System for Data Mining in Distributed Environment[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 13). USA: ACM, 2013: 1506-1509
[10] 李涛, 曾春秋, 周武柏等. 大数据时代的数据挖掘——从应用的角度看大数据挖掘[J]. 大数据, 2015, 1(4) :11-17

作者简介

李 涛，2004年7月获美国罗彻斯特大学计算机科学博士学位；现任美国佛罗里达国际大学计算机学院教授、博导，同时担任南京邮电大学计算机学院、软件学院院长，南京邮电大学大数据研究院院长；2006年获得美国国家自然科学基金委颁发的杰出青年教授奖，2009年获得佛罗里达国际大学最高学术研究奖，2010年获得IBM大规模数据分析创新奖；发表文章250余篇。

刘峥，南京邮电大学计算机学院讲师；主要研究方向为图数据挖掘与查询、网络数据挖掘等；已在国际知名会议发表多篇关于数据挖掘方面的论文。

周绮凤，厦门大学自动化系教授；研究方向为机器学习、数据挖掘及其在可持续发展等领域的应用。