大数据时代的数据挖掘——从应用的角度看大数据挖掘

李 涛[1,2]，曾春秋[1,2]，周武柏[1,2]，周绮凤[3]，郑 理[1,2]
1. 南京邮电大学计算机学院 南京 210023；2. 美国佛罗里达国际大学 迈阿密 33199；
3. 厦门大学自动化系 厦门 361005

摘要
介绍了大数据时代数据挖掘的特点、任务及难点，分析了大数据挖掘的核心架构，提出大数据的核心和本质，即应用、算法、数据和平台4个要素的有机结合。在此基础上介绍了本团队研究设计的大数据挖掘系统FIU - Miner。该系统是一个用户友好并支持在分布式环境中进行高效率计算和算法快速集成的数据挖掘系统平台，使得数据分析人员能够快速有效地进行各类数据挖掘任务。最后，介绍了基于FIU - Miner的3个典型的成功应用案例：高端制造业数据挖掘、空间数据挖掘和商务智能数据挖掘。

关键词
大数据；数据挖掘；FIU - Miner；高端制造业；空间数据挖掘；商务智能

doi: 10.11959/j.issn.2096-0271.2015041

Data Mining in the Era of Big Data: From the Application Perspective

Li Tao[1,2], Zeng Chunqiu[1,2], Zhou Wubai[1,2], Zhou Qifeng[3], Zheng Li[1,2]
1. School of Computer Science & Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
2. School of Computer Science, Florida International University, Miami 33199, USA;
3. Department of Automation, Xiamen University, Xiamen 361005, China

Abstract
The technical characteristics, tasks, and difficulties of data mining in big data era were introduced. The system architecture of large-scale data mining was analyzed. Then, the developed FIU-Miner which is a fast, integrated, and user-friendly system for data mining, was introduced. FIU-Miner supports user-friendly rapid data mining task configuration, flexible cross-language program integration, and effective resource management in heterogeneous environments. Finally three successful real-world applications of FIU-Miner: advanced manufacturing data mining, spatial data mining, and business intelligence data mining, were presented to demonstrate its efficacy and effectiveness.

Key words
big data, data mining, FIU-Miner, advanced manufacturing, spatial data mining, business intelligence

1 对大数据的理解和认识

大数据（big data）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。研究大数据首先要理清和了解大数据的特点及基本概念，进而理解和认识大数据。

1.1 大数据的特点“4V+4V”

从数据的表现形式看，业界普遍认为大数据具有如下的“4V”特点[1]。

- volume（大量）：数据体量巨大，从TB级别跃升到PB级别。
- variety（多样）：数据类型繁多，如网络日志、视频、图片、地理位置信息等。
- velocity（高速）：处理速度快，实时分析，这也是和传统的数据挖掘技术的本质上的不同。
- value（价值）：价值密度低，蕴含有效价值高，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。

上述“4V”特点描述了大数据与以往部分抽样的“小数据”的主要区别。然而，实践是大数据的最终价值体现的唯一途径。从实际应用和大数据处理的复杂性看，大数据还具有如下新的“4V”特点。

- variable（变化性）：在不同的场景、不同的研究目标下数据的结构和意义可能会发生变化，因此，在实际研究中要考虑具体的上下文场景。
- veracity（真实性）：获取真实、可靠的数据是保证分析结果准确、有效的前提。只有真实而准确的数据才能获取真正有意义的结果。
- volatility（波动性）：由于数据本身含有噪音及分析流程的不规范性，导致采用不同的算法或不同分析过程与手段会得到不稳定的分析结果。
- visualization（可视化）：在大数据环境下，通过数据可视化可以更加直观地阐释数据的意义，帮助理解数据，解释结果。

1.2 对大数据的理解

国内外不同的专家和学者对大数据有不同的理解，中国科学院计算技术研究所李国杰院士认为：大数据就是“海量数据”加“复杂数据类型”[2]。维基百科对大数据的定义是：“大数据是由于规模、复杂性、实时性而导致的使之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”[1]。Gartner咨询公司给出的定义是：“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”[2]。而互联网数据中心将大数据定义为：“为更经济地从高频率、大容量、不同结构和类型的数据中获取价值而设计的新一代架构和技术”[3]。

结合上述大数据的“8V”特征，笔者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。大数据是应用驱动的，大数据来源于实践，海量数据产生于实际应用中。

数据挖掘源于实践中的实际应用需求，用具体的应用数据作为驱动，以算法、工具和平台作为支撑，最终将发现的知识和信息用到实践中去，从而提供量化、合理、可行、能够产生巨大价值的信息。另外，挖掘大数据所蕴含的有用信息，需要设计和开发相应的数据挖掘和机器学习算法。算法的设计和开发要以具体的应用数据为驱动，同时也要在实际问题中得到应

1 https://en.wikipedia.org/wiki/Big_data

2 http://www.gartner.com/it-glossary/big-data

3 https://www.idc.com/prodserv/4Pillars/bigdata

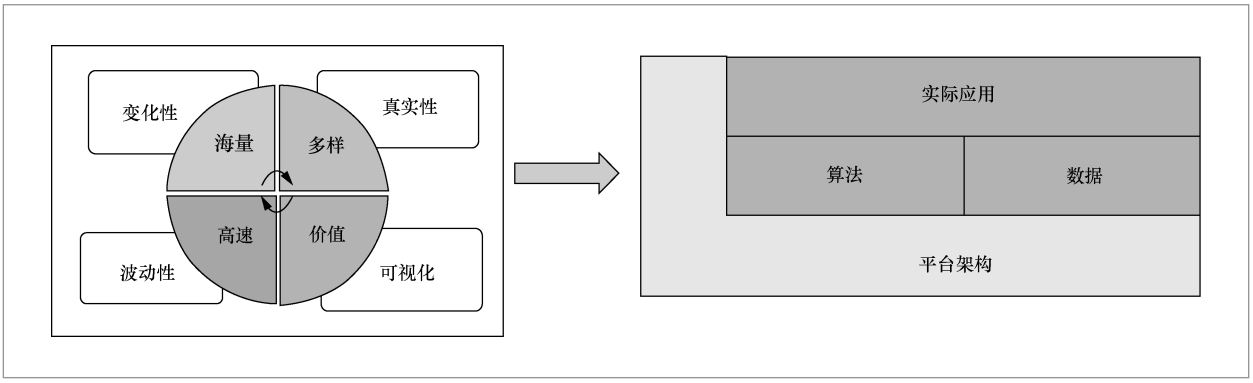


图1 大数据架构

用和验证，而算法的实现与应用需要高效的处理平台。高效的处理平台需要有效地分析海量的数据及对多源数据进行集成，同时有力支持数据挖掘算法以及数据可视化的执行，并对数据分析的流程进行规范。总而言之，这个应用、算法、数据和平台相结合的思想是对上述大数据的理解和认识的一个综合与凝练，体现了大数据的本质和核心。建立在此架构上的大数据挖掘，能够有效处理大数据的复杂特征，挖掘大数据的价值。

本文在此框架下，从应用的角度探讨了大数据时代的数据挖掘的机遇与挑战，介绍了研究团队开发的大数据挖掘平台FIU-Miner以及成功应用该平台实现的高端制造业数据挖掘、空间数据挖掘和商务智能3个大型、复杂数据挖掘案例。

2 大数据时代的数据挖掘

2.1 数据挖掘

在大数据时代，数据的产生和收集是基础，数据挖掘是关键。数据挖掘是大数据中最关键也最有价值的工作。通常，数据挖掘或知识发现泛指从大量数据中挖掘出隐含的、先前未知但潜在的有用信息和模式的一个工程化和系统化的过程。数据挖掘可以用以下4个特性概括[3]。

（1）应用性：数据挖掘是理论算法和应用实践的完美结合。数据挖掘源于实际生产生活中应用的需求，挖掘的数据来自于具体应用，同时通过数据挖掘发现的知识又要运用到实践中去，辅助实际决策。所以，数据挖掘来自于应用实践，同时也服务于应用实践。

（2）工程性：数据挖掘是一个由多个步骤组成的工程化过程。数据挖掘的应用特性决定了数据挖掘不仅仅是算法分析和应用，而是一个包含数据准备和管理、数据预处理和转换、挖掘算法开发和应用、结果展示和验证以及知识积累和使用的完整过程。而且在实际应用中，典型的数据挖掘过程还是一个交互和循环的过程。

（3）集合性：数据挖掘是多种功能的集合。常用的数据挖掘功能包括数据探索分析、关联规则挖掘、时间序列模式挖掘、分类预测、聚类分析、异常检测、数据可视化和链接分析等。一个具体的应用案例往往涉及多个不同的功能。不同的功能通常有不同的理论和技术基础，而且每一个功能都有不同的算法支撑。

（4）交叉性：数据挖掘是一个交叉学科，它利用了来自统计分析、模式识别、机器学习、人工智能、信息检索、数据库等

诸多不同领域的研究成果和学术思想。同时，一些其他领域如随机算法、信息论、可视化、分布式计算和最优化也对数据挖掘的发展起到重要的作用。数据挖掘与这些相关领域的区别可以由前面提到的数据挖掘的3个特性来总结，最重要的是它更侧重于应用。

具体而言，实际应用的需求是数据挖掘领域很多方法提出和发展的根源。从最开始的顾客交易数据分析（market basket analysis）、多媒体数据挖掘（multimedia data mining）、隐私保护数据挖掘（privacy-preserving data mining）到文本数据挖掘（text mining）和 Web 挖掘（Web mining），再到社交媒体挖掘（social media mining）都是由应用推动的。工程性和集合性决定了数据挖掘研究内容和方向的广泛性。其中，工程性使得整个研究过程里的不同步骤都属于数据挖掘的研究范畴。而集合性使得数据挖掘有多种不同的功能，而如何将多种功能联系和结合起来，从一定程度上影响了数据挖掘研究方法的发展。比如，20世纪90年代中期，数据挖掘的研究主要集中在关联规则和时间序列模式的挖掘。到20世纪90年代末，研究人员开始研究基于关联规则和时间序列模式的分类算法（如classification based on association），将两种不同的数据挖掘功能有机地结合起来。21世纪初，一个研究的热点是半监督学习（semi-supervised learning）和半监督聚类（semi-supervised clustering），也是将分类和聚类这两种功能有机结合起来。近年来的一些其他研究方向如子空间聚类（subspace clustering）（特征抽取和聚类的结合）和图分类（graph classification）（图挖掘和分类的结合）也是将多种功能联系和结合在一起。最后，交叉性导致了研究思路和方法设计的多样化。

2.2 从数据挖掘应用的角度看大数据

大数据是现象，核心是要挖掘数据的价值。结合数据挖掘的各种特性，尤其是其应用性，从应用业务的角度对大数据提出如下两点的认识[3]。

首先，大数据是“一把手工程”。在一个企业里，大数据通常涉及多个业务部门，业务逻辑复杂。一方面，要对大数据进行收集和整合，需要业务部门的配合和沟通以及业务人员的大力参与，这些需要企业决策人员的重视和认可，提供必要的资源调配和支持。另一方面，要对数据挖掘的结果进行验证和运用，更离不开相关人员的决策。数据挖掘的结果大多是相关关系，而不是因果关系，这些结果还可能有不确定性。另外，有时候数据挖掘的结果与企业运作的常识不一致，甚至相悖。所以，如何看待这些可能的不确定性和反常识的分析结论，充分利用好数据挖掘结果，必然离不开决策者的远见卓识。

其次，大数据需要数据导入、整合和预处理。当面对来自不同数据源的大量复杂数据时，具体业务逻辑复杂与数据之间的关系琐碎直接导致企业的业务流程和数据流程很难理解。因此，企业在实施大数据时可能并不清楚要挖掘和发现什么，对数据挖掘到底能帮助企业做什么并没有直观和清楚的认识。所以，很多时候都不可能先把数据事先规划好和准备好，这样在具体的数据挖掘中，就需要在数据的导入、整合和预处理上有很大的灵活性，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。

2.3 大数据时代应用数据挖掘的挑战

大数据时代的来临使得数据的规模和复杂性都出现爆炸式的增长，促使不同应用领域的数据分析人员利用数据挖掘技术对数据进行分析。在应用领域中，如医疗保健、高端制造、金融等，一个典型的数据挖掘任务往往需要复杂的子任务配置，整合多种不同类型的挖掘算法以及在分布式计算环境中高效运行。因此，在大数据时代进行数据挖掘应用的一个当务之急是要开发和建立计算平台和工具，支持应用领域的数据分析人员能够有效地执行数据分析任务。

现有的数据挖掘工具（如Weka[4]、SPSS和SQL Server等）提供了友好的界面，方便用户进行分析。然而，这些工具并不适合进行大规模的数据分析。同时使用这些工具时，用户很难添加新的算法程序。流行的数据挖掘算法库（如Mahout[5]、MLC++[4]和MILK[5]）提供了大量的数据挖掘算法。但是，这些算法库需要有高级编程技能才能在一个具体的数据挖掘任务中进行任务配置和算法集成。最近出现的一些集成的数据挖掘产品（如Radoop[6]和BC-PDM[7]）通过提供友好的用户界面来快速配置数据挖掘任务。然而，这些产品是基于Hadoop框架的，对非Hadoop算法程序的支持非常有限。此外，这些产品并没有明确地解决在多用户和多任务情况下的资源分配问题。

为了解决现有工具和产品在大数据挖掘中的局限性，开发了一个新的平台——FIU-Miner（a fast, integrated, and user-friendly system for data mining in distributed environment[8]），是一个用户友好并支持在分布式环境中进行高效率计算和快速集成的数据挖掘系统，该平台支持数据分析人员快速、有效地进行数据挖掘任务。

3 大数据挖掘系统FIU-Miner的研究设计

3.1 FIU-Miner平台介绍

与现有数据挖掘平台相比，FIU-Miner提供了一组新的功能，能够帮助数据分析人员方便并有效地开展各项复杂的数据挖掘任务。

具体而言，FIU-Miner 具有以下突出的优点。

（1）用户友好、人性化、快速的数据挖掘任务配置：基于“软件即服务”这一模式，FIU-Miner隐藏了与数据分析任务无关的低端细节。通过FIU-Miner提供的人性化用户界面，用户可以通过将现有算法直接组装成工作流，轻松完成一个复杂数据挖掘问题的任务配置，而不需要编写任何代码。

（2）灵活的多语言程序集成：FIU-Miner允许用户将目前最先进的数据挖掘算法直接导入系统算法库中，以此对分析工具集合进行扩充和管理。同时，由于FIU-Miner 能够正确地将任务分配到有合适运行环境的计算节点上，所以对这些导入的算法没有实现语言的限制。

（3）异构环境中有效的资源管理：FIU-Miner支持在异构的计算环境中（包括图形工作站、单个计算机、和服务器等）运行数据挖掘任务。FIU-Miner综合考虑各种因素（包括算法实现、服务器负载平衡和数据位置）来优化计算资源的利用率。

3.2 FIU-Miner系统架构

FIU-Miner的系统架构如图2所示。该系统分为4层：user interface（用户接口

4 http://www.sgi.com/tech/mlc

5 http://pythonhosted.org/milk

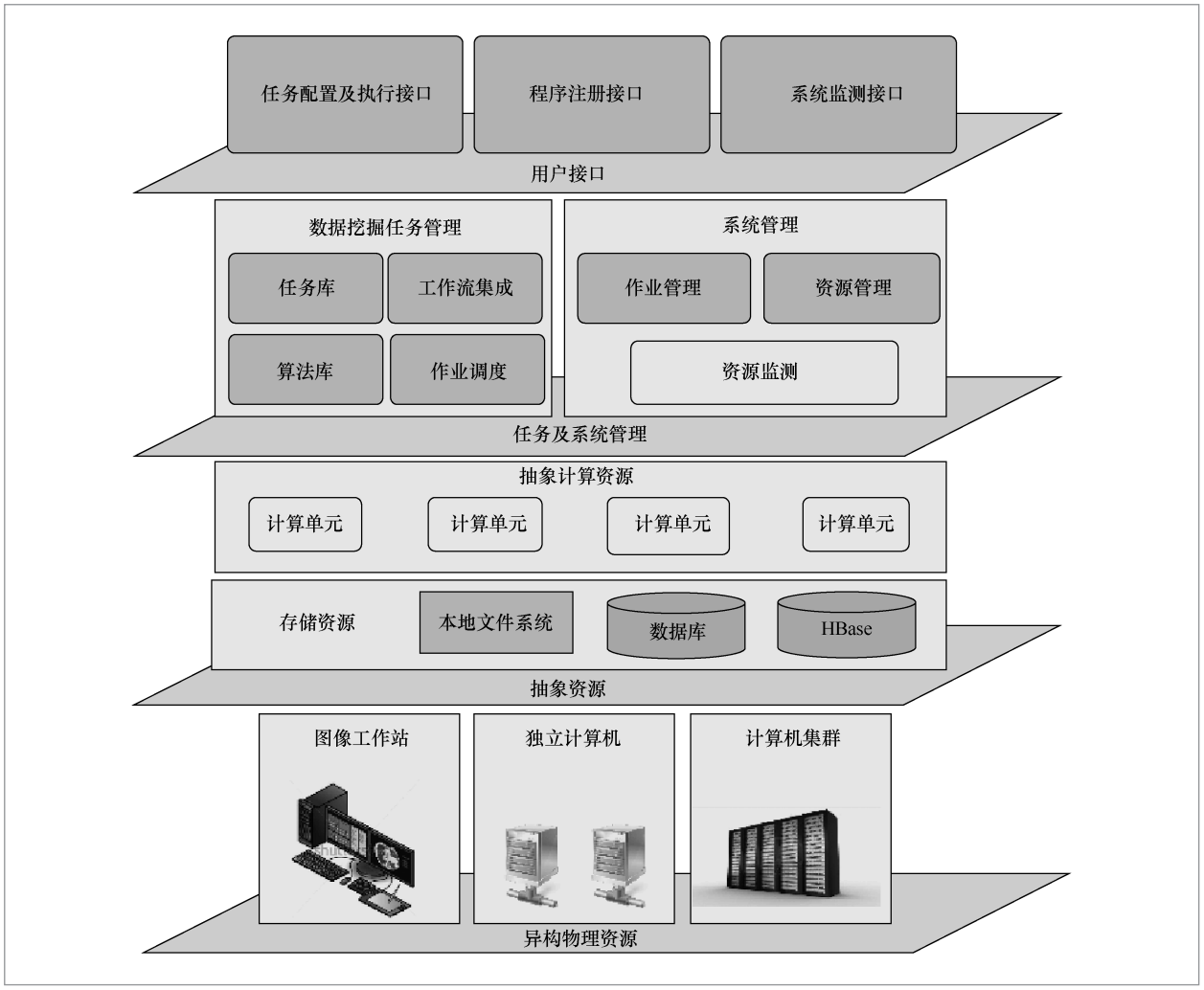


图 2 FIU-Miner 系统架构

层）、task and system management（任务与系统管理层）、abstracted resources（抽象资源层）和heterogeneous physical resource（异构物理资源层）。这种分层架构充分考虑了海量数据的分布式存储、不同数据挖掘算法的集成、多种分析任务的配置以及系统和用户的交互功能[6]。

6 http://datamining-node08.cs.fiu.edu/FIU-Miner

3.2.1 用户接口层

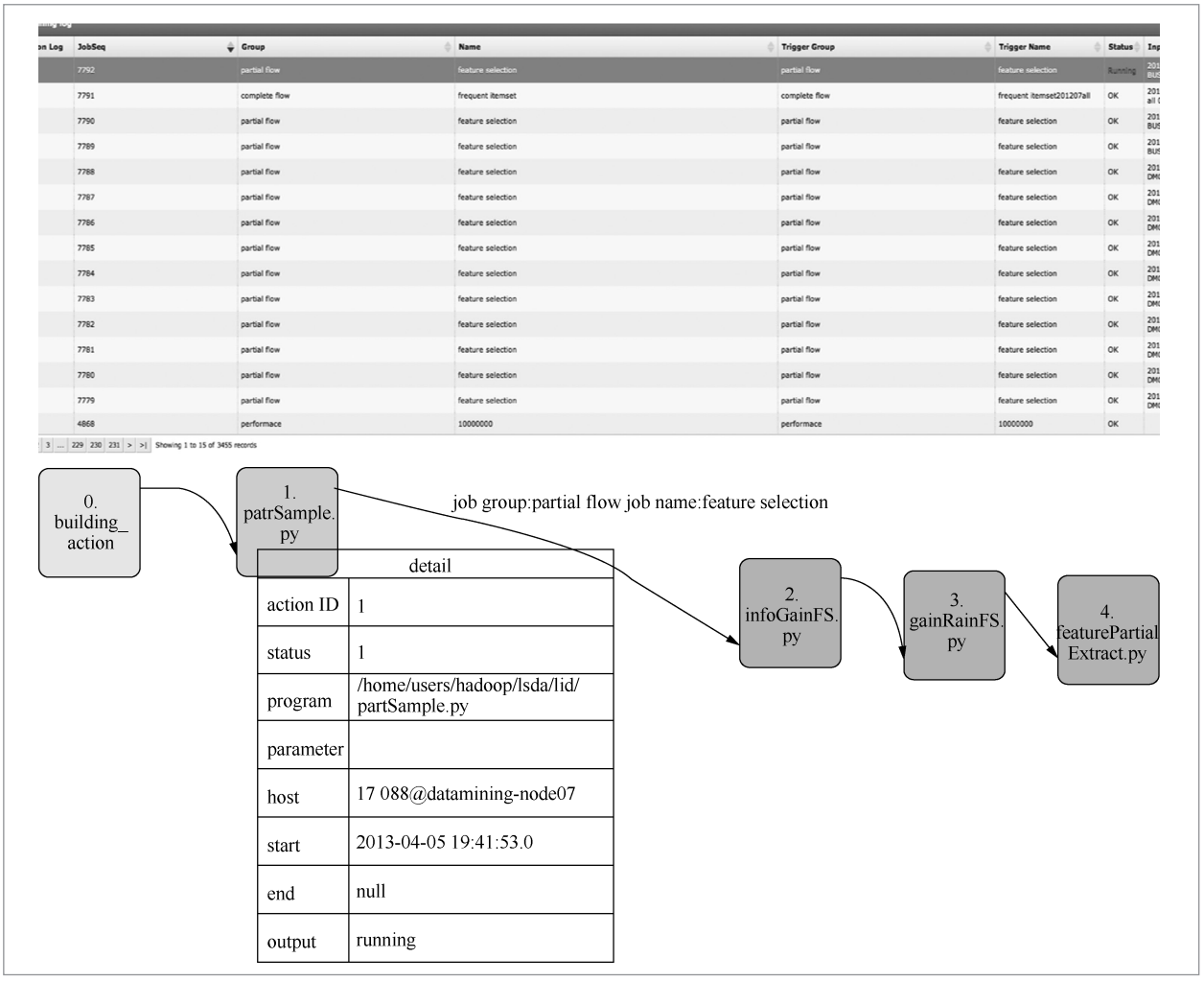
为了最大限度地提高系统的兼容性，用户接口层是完全用HTML5开发的Web应用程序。如图3所示，用户接口层有如下3个功能模块。

（1）任务配置和执行（task configuration and execution）

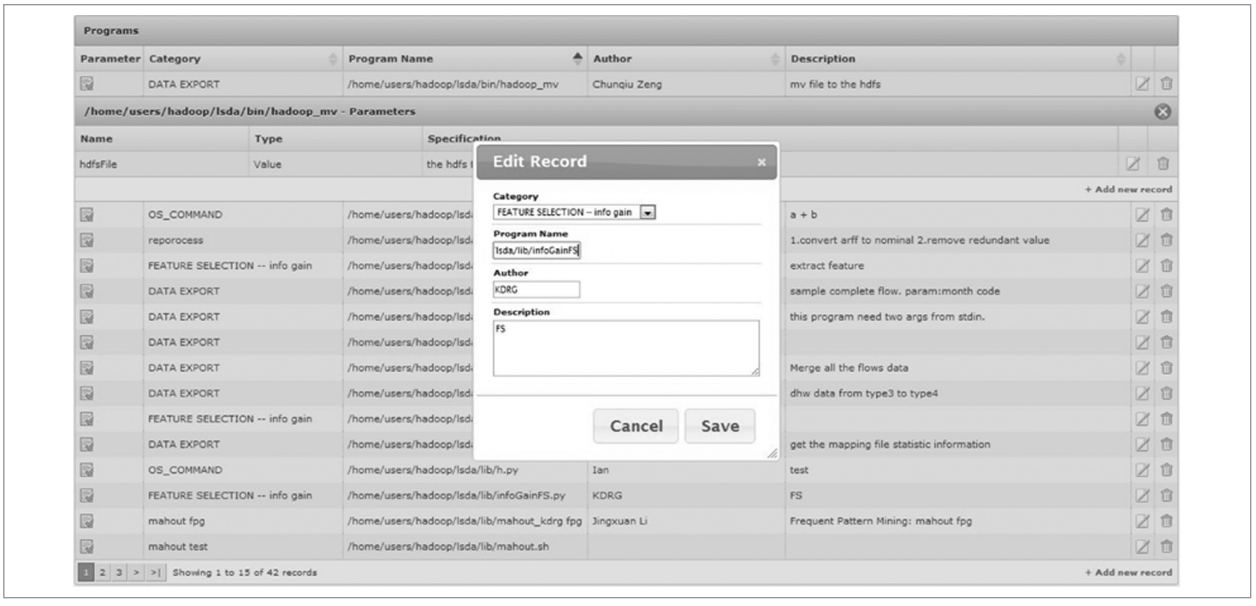
该模块支持面向工作流的数据挖掘任务配置。一个数据挖掘任务的工作流可以被表示为一个有向图，其中图的节点表示特定的算法，图的边表示算法中的数据相关性。在FIU-Miner 中，一个工作流程可通过图形用户界面来快速配置，而不需要编程。此外，用户可以设置数据挖掘任务的执行计划，包括程序的定时、循环、顺序等执行方式。

（2）程序注册（program registration）

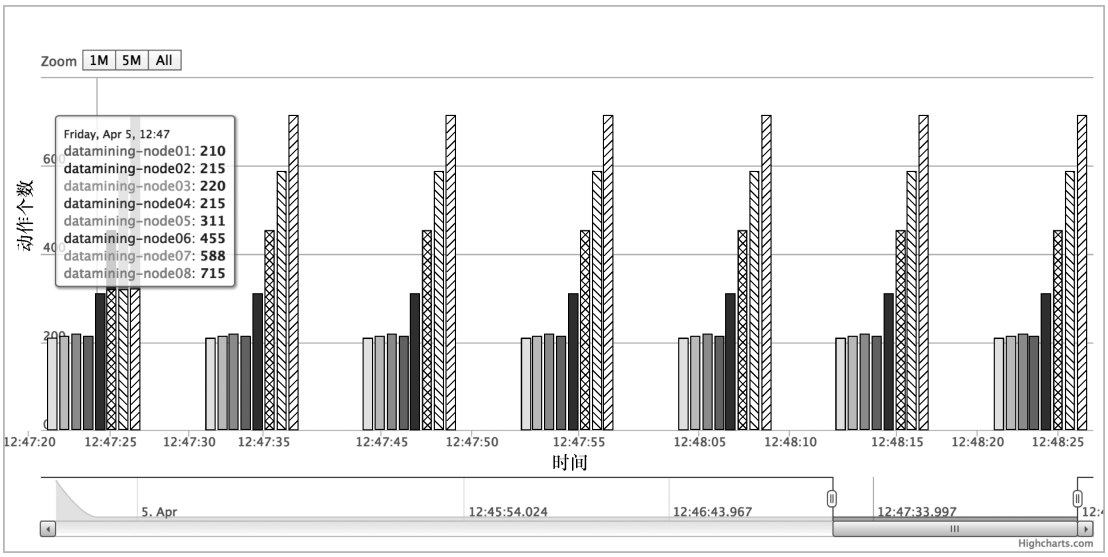
该模块可以让用户轻松地导入外部



（a）任务配置和执行



（b）程序注册模块



(c)系统监控模块

图 3 用户接口层功能模块

数据挖掘算法，充实FIU-Miner的算法库。如果要导入外部程序，用户需要上传可执行文件，提供详细的描述信息，包括程序的功能描述、需要的运行环境、程序和相关数据以及参数规范。导入的程序可以使用任何语言编写，只要后端服务器能支持它需要的运行环境。FIU-Miner目前支持Java（包括Hadoop的环境）、Shell、Python和C/C++等语言编写的程序，因此几乎所有实现的主流数据挖掘算法，如基于Weka、Mahout、MILK等数据挖掘和机器学习算法库的程序，都可以很容易地导入FIU-Miner。用户还可以将自己实现的算法导入系统中。

(3)系统监控(system monitoring)

该模块实时监测FIU-Miner 的资源利用率，并且动态跟踪系统中提交任务的运行状态。注意该模块只显示了抽象的资源（逻辑存储和计算资源包括数据库、文件系统、计算单元等），使底层物理资源对用户透明。

3.2.2 任务及系统管理层

任务及系统管理层包含了两个主要功能模块：任务管理和系统管理。

(1)任务管理

FIU-Miner允许用户动态配置数据挖掘任务，以满足他们的分析需求。用户可以选择在算法库(algorithm library)中注册的算法作为基本模块来构造工作流。工作流集成器(workflow integrator)负责工作流的任务集成和验证，同时发现和报告无效的流程。一旦新的数据挖掘任务集成和配置完成后，它将被自动添加到任务库（task library），可以随时被调度运行。作业调度器(job scheduler)负责分配计算资源及优化运行时间。FIU-Miner里的调度比较复杂。一方面，FIU-Miner支持不同编程语言实现的程序在异构的计算环境中运行。一个任务里的不同程序可能会有不同的运行环境要求。所以，简单地把任务分配到空闲的计算单元不一定可行。另一方面，将一个作业分成不同的步骤，让每

个步骤在不同的计算单元上运行，可能会增加I/O成本。如果再考虑多用户、多任务的情况，FIU-Miner里的调度就会变得更加困难和复杂。为了解决上面的难题，在实现FIU-Miner的调度时，综合考虑了如下因素：给定任务每一步的运行环境要求；每个计算单元支持的运行环境；每个计算结点的当前运行状态；输入数据的大小。

（2）系统管理

作业管理器（job manager）跟踪执行作业的运行状态。用户会收到作业的实时状态。除了作业监视，FIU-Miner还会跟踪计算单元以及相关计算资源的状态。资源监视器（resource monitor）监视计算单元并提供作业调度程序的运行状态，以帮助调度决策。资源管理器（resource manager）管理所有可用的计算单元。FIU-Miner的一个独特的优点是，它不需要人工登记可用物理资源。一旦计算单元部署在物理服务器上，它会将服务器的信息发送给资源管理器，自动将服务器在FIU-Miner里注册。

3.2.3 抽象资源层

抽象资源层包括存储和计算资源。存储资源建立在物理设备的基础上，包括传统数据库、本地文件系统、分布式文件系统（比如HDFS）等。计算单元是逻辑上的计算资源。平台的计算能力依赖计算单元的数量。通过扩展配置计算单元的数量，能有效地支撑上层的数据挖掘任务。

在FIU-Miner中，物理服务器的计算能力是由计算单元的数量和安排的数据挖掘任务来量化的。这种机制是一个系统虚拟化的简化版本，能够最大限度地提高计算资源的利用率。为了有效地管理计算资源，每个计算单元都包含详细规范的配置文件（信息包括计算能力、支持的运行环境、运行状态等）。一台物理服务器的存储（包括可用的数据库、HDFS和本地文件系统）由该服务器上布置的计算单元共享。

3.2.4 异构物理资源层

异构物理资源层亦称物理资源层，主要包括底层的物理设备。这些物理设备能有效地支撑数据存储和扩展。

3.3 FIU-Miner系统亮点评述

FIU-Miner 建立于分布式异构环境之上，大大减少了不同物理环境给构建数据分析任务带来的复杂度，充分利用分布式计算的能力提升数据分析的效率。另外，FIU-Miner的计算资源是可动态增减的，使其具备根据具体分析任务数量进行在线调整计算物理资源的能力。最后，友好的用户接口为基于FIU-Miner构建不同的大数据挖掘应用提供了极大的便捷。

4 FIU-Miner应用实例一：高端制造业

4.1 高端制造业大数据挖掘任务

制造业是指大规模地把原材料加工成成品的工业生产过程。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。典型的高端制造业[9]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。产量和品质极大地依赖流程管控和优化决策。因此，制造企业不遗余力地采用各种措施优化生产流程，调优控制参数，提高产品品质和产量，从而

提高企业的竞争力。

随着工艺、装备和信息技术的不断发展，现代制造业（特别是高端制造业）产生和积累了大量生产过程的历史数据。这些数据中蕴含对生产和管理有很高价值的知识和信息。高端制造企业利用这些技术能够更好地收集和管理生产流程数据，也使得企业累积的相关数据在日益增多的同时，也变得更加丰富、完备、准确。

这些采集的数据来源于实际生产，并与生产设计、机器设备、原材料、环境条件、生产流程等生产要素信息高度相关。通常情况下，工程人员通过人工分析很难察觉到参数间的关联模式和影响品质的重要生产要素等信息。然而，如何有效地利用这些数据优化生产过程，提升生产效率，成为了企业关注的焦点。因此，制造企业需要一种高效、可靠的分析方法及工具，把隐藏在海量数据中有用的、深层次的知识和信息挖掘出来，以提升高端制造业在控制、优化、调度、管理等各个层面分析和解决问题的能力。幸运的是，利用数据挖掘可以对这些数据进行有效的分析并转换成有价值的生产知识，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。因此，数据挖掘技术是解决制造业海量信息数据处理的关键技术之一。

4.2 高端制造业大数据挖掘挑战

高端制造业中的数据挖掘面临很多挑战，比如：如何有效分析大规模数据、如何保证数据分析效率和分析结果的准确性？在实际应用中，从海量数据中依靠传统信息系统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。

4.3 具体案例

FIU-Miner已经被成功地应用在四川虹欧显示器件有限公司，作为等离子屏制造过程的数据分析平台[3,10]。

4.3.1 等离子显示器制造

等离子显示器（plasma display panel，PDP）是一种利用气体等离子效应放出紫外线，从而激发三原色发光体独立发光，达到显示不同颜色和控制亮度的高端图像显示器。它具有亮度高、色彩多、面积大、视角广、图像清晰等众多优势，是大面积显示需求（如家庭影院、电子广告墙）的首选显示器。

四川虹欧显示器件有限公司是国内最大的等离子生产公司，每天生产超过1万张等离子显示面板，其生产线的一些指标包括[10]：20个大工序、151个小工序；1 000多台设备串联；工艺设备共计279台，设备种类达83种；2 225个物流单元，全长6 000 m；产品制造时间约76 h；单台产品涉及的过程设备参数超过1.17万个。

具体而言，在生产实践中，技术人员关注如何提高产品的良品率。实现这个目标，需要回答下面的一些问题：哪些是关键的工艺参数（它们对产品的良品率有显著的影响）、参数值的变动会怎样影响产品的良品率、哪些是有效的可以确保高良品率的工艺参数配方等。从PDP的数据特点来说，每天生产的数据存储量是10 GB以上，每月有3~5亿笔制造过程记录，在数量、维度和数据产生速度上具有海量大数据特

征。在生产工序复杂、设备参数众多、数据量大的背景下，人为分析PDP生产过程，以期达到提高生产质量的效果几乎是无法实现的。因此，迫切需要研究基于等离子显示屏制造过程的自动化流程和产品优化工具，从而提升制造过程参数管控能力和产品品质。

4.3.2 基于FIU-Miner 的解决方案

在过去的几年里，笔者的研究团队一直与四川虹欧显示器件有限公司的技术人员和工程师紧密合作，利用数据挖掘来提高等离子屏的生产良品率。在这个合作过程中，确定了如下两个主要的分析难点，并提出了相应的基于FIU-Miner 的解决方案。

- 7×24 h的自动化生产方式和新数据采集工具的使用，使得数据量急剧增长，需要强大的数据分析能力来支撑。
- 大量过程控制参数造成的数据高维特性对数据分析效率和分析结果的准确性提出了更高要求。生产数据分析是对生产工作流程的一个认知过程。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。

笔者的研究团队在FIU-Miner 的基础上，开发了离子屏制造过程数据挖掘系统（PDP-Miner）[10]来解决PDP数据分析的难题。PDP-Miner的架构如图4所示。具体而言，在FIU-Miner的基础上增加了数据分析层。

数据分析层提供具体分析任务的用户执行接口。以等离子屏数据挖掘系统为例，数据分析任务主要包括数据立方、对比分析、回归分析、参数选择、参数配方、操作平台、结果展示和报告管理。

其中，数据立方使分析人员能够对数据进行宏观理解和快速预览。数据立方子

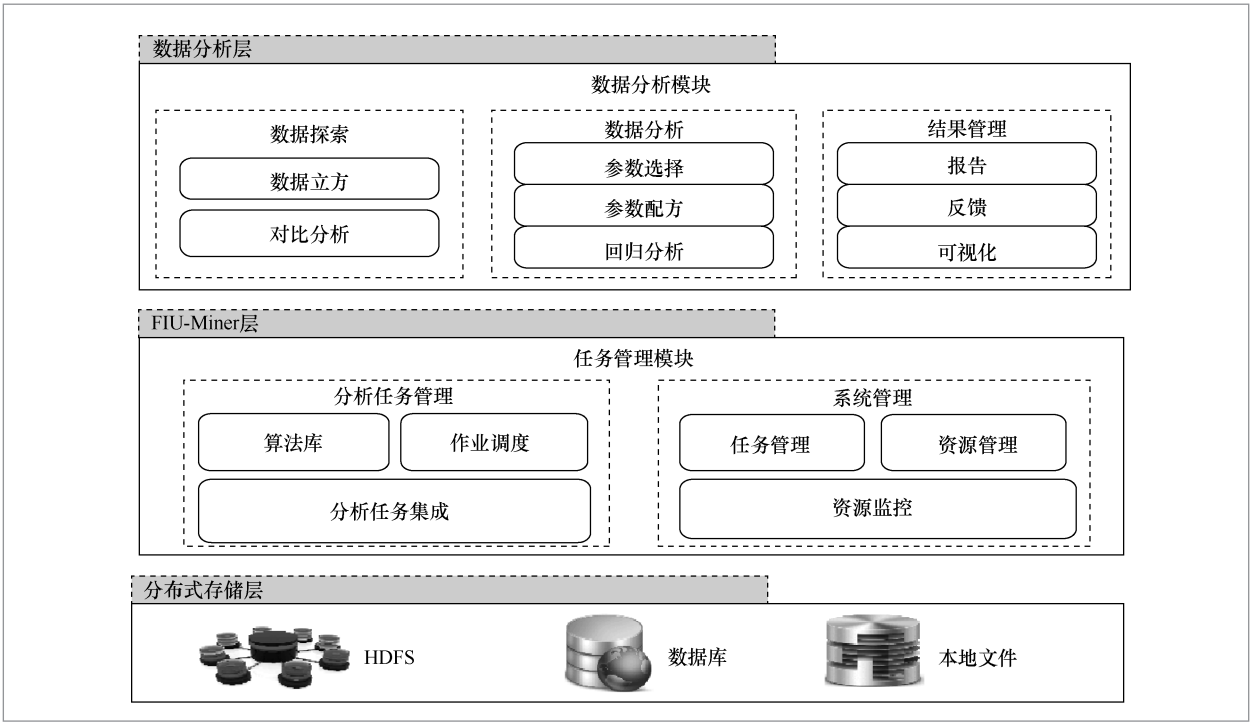


图 4 PDP-Miner 的系统架构

系统可以通过OLAP技术建立数据立方来帮助分析人员大致掌握数据特性。通过选择维度和建立测度来对数据集进行分析。通过数据立方操作（下钻、上卷等）实现对数据的多粒度、多角度的理解。

对比分析子系统，能快速发现敏感参数和验证重要参数，因此，在PDP生产系统中显得特别重要。通过比较参数在不同时期取值的统计特性，有效发现异常参数值，从而定位敏感设备或数据集。

数据分析子系统主要负责集成数据挖掘算法，提供业务操作接口。由于该系统面向非专业领域的操作人员，并聚焦到具体的分析业务，因此数据挖掘算法被合理封装到各个业务中，对操作人员透明。现在的挖掘算法主要支持回归分析、参数选择、参数配方等任务。

分析报告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5给出了两个具体PDP挖掘的工作流。其中第一个工作流（workflow 1）先集成多种特征选择的方法来选出影响PDP生产的重要工艺参数，然后利用回归分析来建立这些参数与产品质量的关系。第二个工作流（workflow 2）是利用频繁模式分析来挖掘重要工艺参数的关联关系，从而产生可能的参数配方。图6给出了工作流的配置界面。

使用等离子屏制造过程数据挖掘系统大大降低了对前台使用人员的要求，可以使得操作人员能够将精力聚焦到快速发现问题和解决问题上。

通过技术人员将数据挖掘研究的结果和平台进行有效应用，提高了对制造过程中所出现问题的分析和解决的效率（见表1），使PDP屏生产线的综合良品率及生产效率

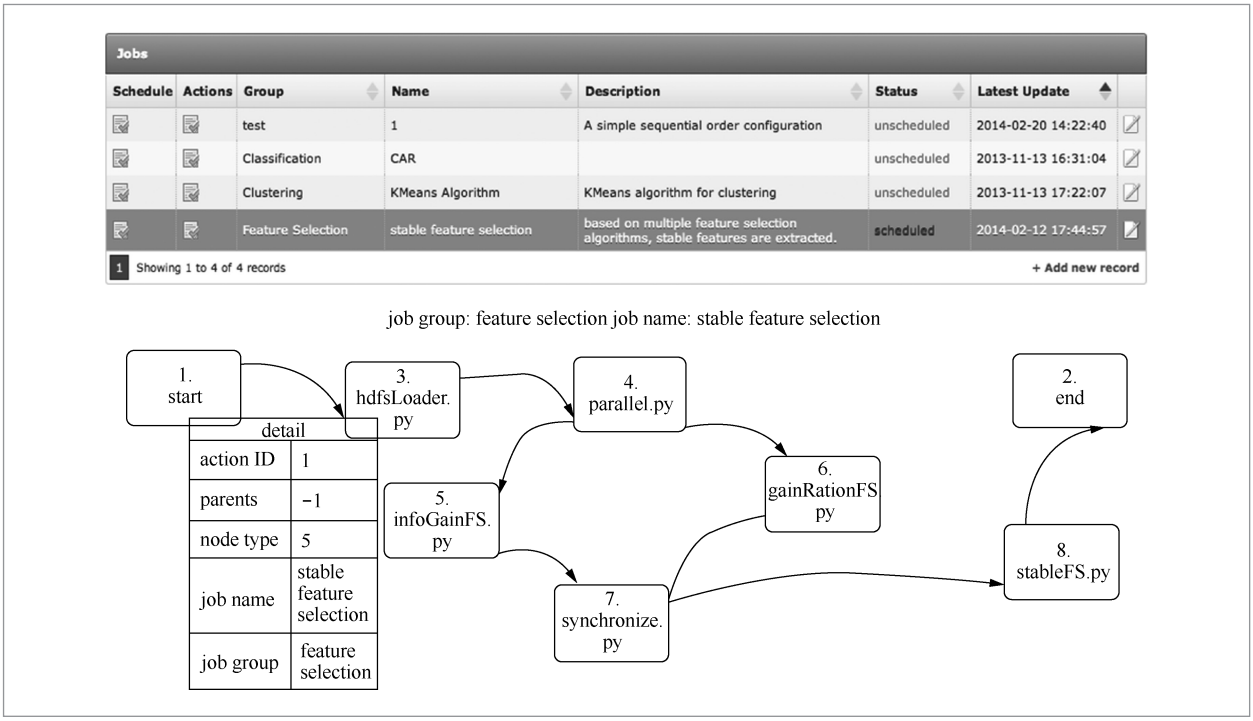


图 5 PDP-Miner 工作流程

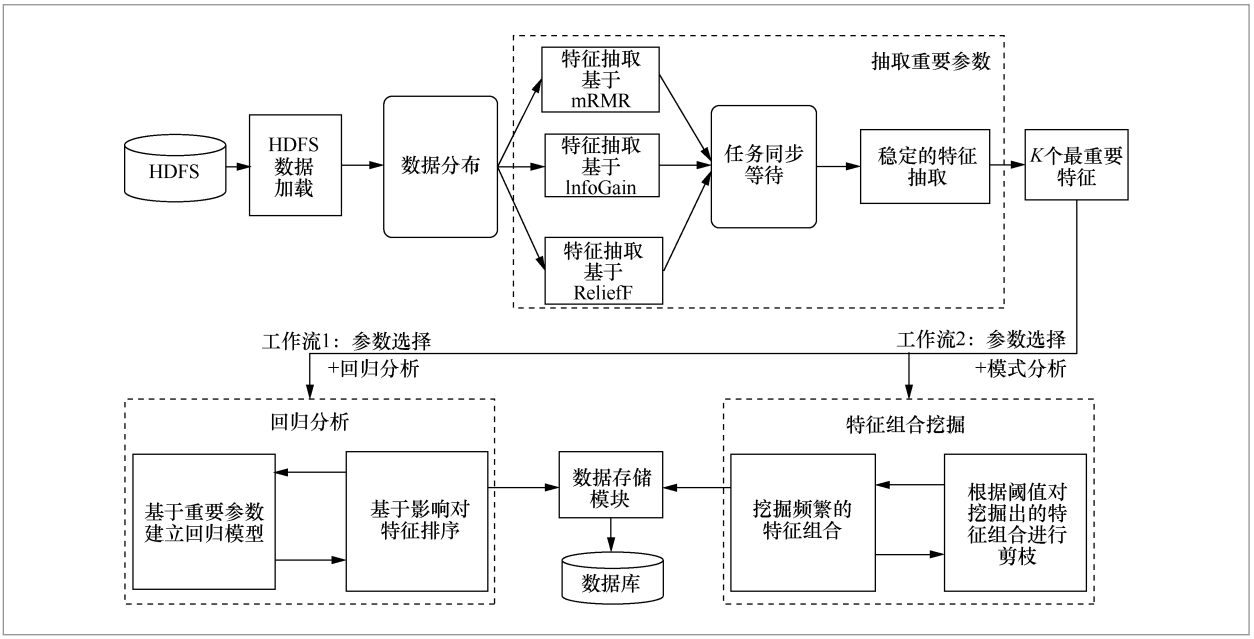


图 6 PDP-Miner 工作流配置界面

表 1 PDP-Miner 数据挖掘技术与传统数据挖掘技术比较

数据挖掘技术	评价标准	数据准备	流程配置	操作执行
传统数据挖掘技术工具库	评价内容	实现数据加载和存储流程	程序实现各个分析模块的依赖关系和调度顺序	实现监控模块，人工控制和检测失败任务
	时间成本	若干小时	若干天	若干天
	人力成本	需要具备在分布式环境下编程的数据分析人员		
等离子显示屏数据挖掘技术	评价内容	使用定制的用户界面加载和存储数据	通过后台程序来配置任务，对用户透明	监控任务的执行，支持失败任务诊断
	时间成本	简单点击配置	任务的自动调度和配置，若干分钟	简单点击查看若干小时
	人力成本	对业务人员不需要编程程序，具备基本的数据分析知识		

得到了快速提升。一方面，在显示器件制造业首次采用大数据挖据技术，实现了由传统离散型的试验设计方法到数据挖掘模型来进行制造过程参数管控的动态在线分析处理方法，降低了制造过程品质管控的试验成本。另一方面，通过数据挖掘平台，建立了等离子屏制造过程单工序/全工序的参数管控的主要数据挖掘分析模型，通过挖据结果的有效应用，促进了等离子显示屏的制造良品率和生产效率的提升。最后，利用平台挖掘方便快捷地指导技术人员进行参数管控的常态化螺旋式提升。在成果应用的这些年里，促进了PDP良品率和产能的快速提升，给公司带来了巨大的生产经济效益。图7给出了PDP-Miner的实际应用的主界面，该系统的功能模块包括数据探索（对比分析、数据立方）、数据分析（操作平台、参数选择、回归分析、判别分析）、结果管理（可视化、结果列表和反馈收集）。需要特别指

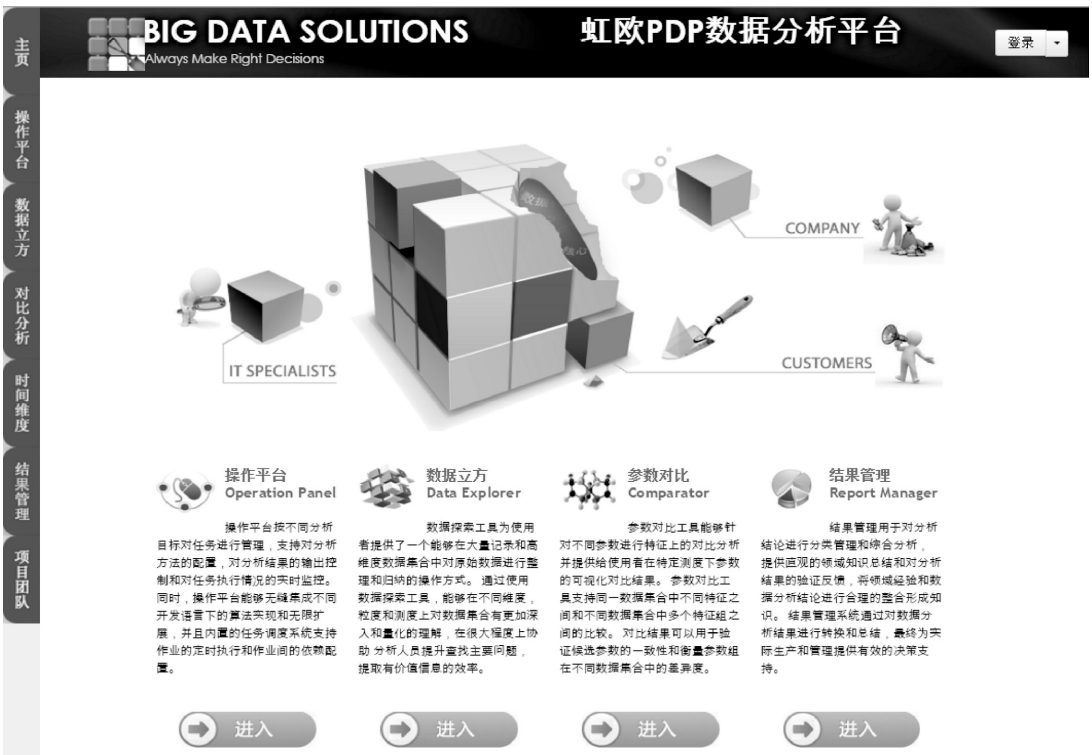


图 7 PDP-Miner 实际应用的主界面

出的是，等离子显示屏制造挖掘平台可方便地移植于液晶面板、OLED面板等其他平板显示领域，具备向整个平板行业推广的基础。

4.4 应用亮点评述

将FIU-Miner应用于高端制造业的实际案例，在国际上率先将数据挖掘技术应用于显示器件制造业，为四川虹欧显示器件有限公司构建了制造过程单工序/全工序数据挖掘分析模型，开发了基于数据挖掘的PDP-Miner平台，有效提升了生产效率和产品质量。该公司应用PDP-Miner平台后，产品综合良品率得到了很大提高，同时，生产效率的提升也带来了很大的经济效益。该研究获得2013年“中国制造业IT新兴技术应用最佳实践奖”[7]。

5 FIU-Miner应用实例二：空间数据挖掘

FIU-Miner 已被成功应用于TerraFlyGeocloud[11]，支持多种在线空间数据分析的平台。

5.1 空间数据挖掘

随着卫星科技的发展及移动设备的普及，获取一个对象实时完整的空间信息变得越来越容易。为了能够从中实时性地获取有用信息，需要有效的方法进行空间数据挖掘。空间数据挖掘是从大型空间数据库里发现有趣的、不知道的但非常有价值的模式的一个过程。但由于空间数据类型

7 http://news.e-works.net.cn/category146/news55123.htm

和空间关系的复杂性，从空间数据库里挖掘有趣和有价值的模式比从传统数据库里挖掘难度更大。

5.2 TerraFlyGeocloud介绍

空间数据挖掘可以应用在很多领域，包括水资源管理、交通管理、灾难管理、犯罪分析、疾病分析和房地产等。一个典型的空间挖掘系统应支持以下功能：在线的空间数据分析、空间数据可视化和空间数据查询。这里，介绍一个具体的空间数据挖掘系统：美国佛罗里达国际大学（FIU）计算机学院的高性能数据研究中心实验室开发的TerraFlyGeoCloud 系统。TerraFlyGeoCloud是建立在TerraFly系统之上的、支持多种在线空间数据分析的一个平台。图8和图9分别给出了TerraFlyGeoCloud的系统界面和工作流程。

为了方便使用，TerraFlyGeoCloud还提供了一种支持类SQL语句的空间数据查询语言MapQL。它不但支持类SQL语句，更重要的是可根据用户的不同要求，渲染和画图查询得到空间数据，比如学校周边一定距离内所有的开放住宅、离某条公路一定距离内所有的宾馆、特定地区的交通情况及不同邮政区域的平均收入情况等。MapQL的实现如图10（a）所示，其中MapQL语句是整个过程的输入，如图10（b）所示，输出则是通过MapQL引擎渲染得到的可视化地图，如图10（c）所示。

下面简要讲述一下使用MapQL的具体过程。如图10（a）所示，第一步语法检查，保证语法符合语法规则，不出现关键字拼写错误；第二步语义检查，确保MapQL将要访问的数据是正确并存在的。接下来，系统会进行语句解析并把包含样式信息的解析结果存入空间数据库中。样式信息包括“渲染什么”及“在哪

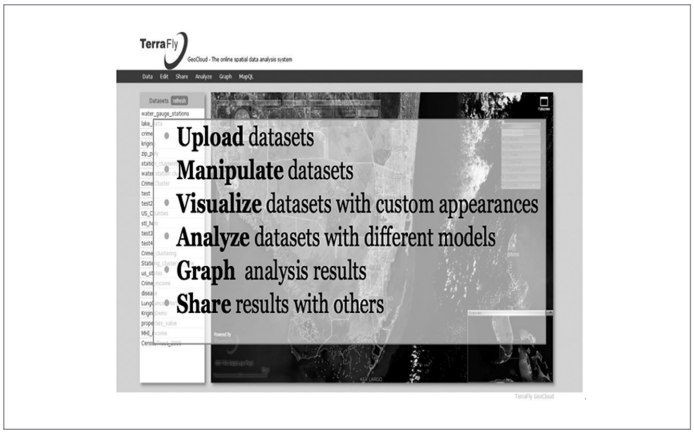


图 8 TerraFlyGeoCloud 系统界面

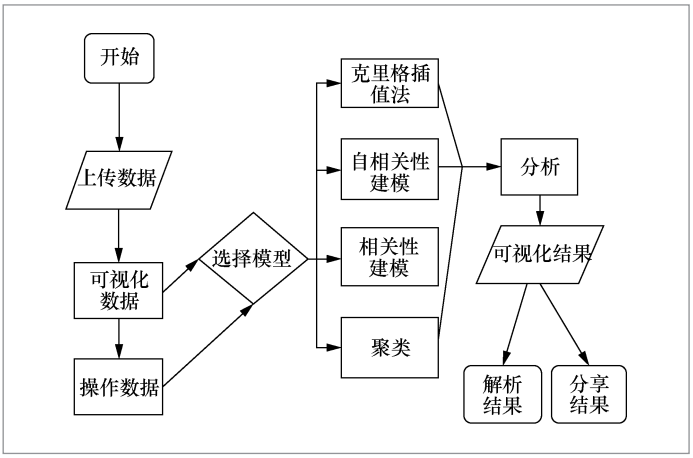
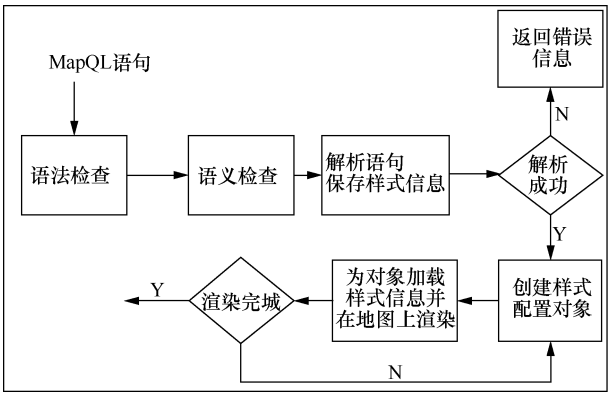


图 9 分析工作流程

渲染”。当所有的样式信息保存入库时，系统就会为接下来的渲染创建样式配置对象。最后，从空间数据库里加载样式信息，并根据样式信息为每个对象进行渲染。比如想查询佛罗里达国际大学周围的房价，可通过如图10（b）的MapQL语句查询，结果如图10（c）。

MapQL提供了一个比地理信息系统应用程序编程接口（API）更友好的界面，使得开发人员和终端用户能够便捷自如地使用TerraFly地图，同时能够灵活地创建自己的地图。

除了支持地理信息系统的各种应用外，TerraFly平台还有丰富的GIS数据集，包括美国和加拿大的道路数据、美国人口



(a) MapQL的实现

```
SELECT
  '/var/www/cgi-bin/house,png' AS T_ICON_PATH,
  r,price AS T_LAEEL,' 15'  AS T_ICON_SIZE,
  I,geo AS GEO FROM realtor_20121116 r WHERE
    ST_Distance(r,geo,GeomFromText('POINT
    (-80.276 283 25.757 228)'))<0.03;
```

(b) MapQL语句查询

(c) 可视化地图

图 10 MapQL 的实现、语句查询及可视化地图

普查和社会经济数据、1 500万企业的统计和管理记录、200万专业医生的数据、各种公共场所的数据集和全球环境数据等，用户可以通过TerraFlyGeoCloud浏览、使用和挖掘这些数据集。

5.3 TerraFlyGeocloud使用难点

通过对TerraFlyGeoCloud的进一步使用和研究，发现了如下几个问题。这些问题非常典型，普遍存在于这类空间数据挖掘系统中。

(1) 写MapQL查询语句的难度。虽然大多数开发人员熟悉SQL语句，可以很快地写MapQL查询。但对不熟悉SQL的用户而言，学习MapQL还是比较困难的。所以，对绝大多数用户而言，利用MapQL来完成空间分析任务仍然比较困难。

(2) 空间分析任务的复杂性。一个典型的空间分析任务往往涉及几个子任务。此外，这些子任务之间并不是完全独立的。其中一些子任务的输出往往是其他子任务的输入。根据这种依赖关系，一个空间数据分析任务可以自然地表示为一个工作流。但构造和管理这样一个复杂的工作流程是空间数据分析的一个难点。

(3) 顺序执行空间数据分析的工作流的效率往往很低。尽管一个工作流中的子

任务并不是互相依赖，但这些子任务只能由最终用户来顺序执行。这种顺序执行的方式没有充分利用分布式计算环境来并行执行独立的子任务和优化系统性能。

这3个问题给空间数据挖掘系统带来了很大的局限，限制了用户对系统的有效使用。将FIU-Miner 与TerraFlyGeocloud结合来解决这些问题。首先，根据序列模式挖掘算法从TerraFlyGeoCloud的MapQL查询日志中发现顺序查询模式[11]。然后利用这些顺序查询模式，在FIU-Miner里面构建空间数据分析任务的工作流。最后使用FIU-Miner来最大化子任务的并行执行，优化工作流的执行效率。

TerraFlyGeocloud+FIU-Miner系统架构如图11所示。主要有4层：用户界面层、地理空间服务层、计算服务层和空间数据存储和管理层。其中，从MapQL的查询日志中挖掘查询模式是一个关键的步骤，这个步骤发生在地理空间服务层。挖掘出的顺序查询模式可以用来产生查询模板和构造空间分析的工作流。序列模式里面的每个查询对应于工作流里面的一个子任务。FIU-Miner在计算服务层，主要负责工作流的构建、管理、调度和执行。

5.4 应用实例

利用FIU-Miner，系统可以通过构建空间数据分析的工作流来优化分析流程，提高分析效率。下面通过一个详细的房产投资案例来展示[12]。

房产投资案例的目的是要寻找具有良好升值潜力的房产。如果一栋房产本身价值很低，但它周围的房产却相对来说比其高，那么对此房产进行投资将是一个非常不错的选择。根据历史查询数据，通过序列模式挖掘，发现这个任务一般有下面几个步骤：

- 计算不同地区的平均价格，比较邻近地区的价格，确定感兴趣的地区；
- 对感兴趣的地区进行空间自相关分析，确定候选地区；

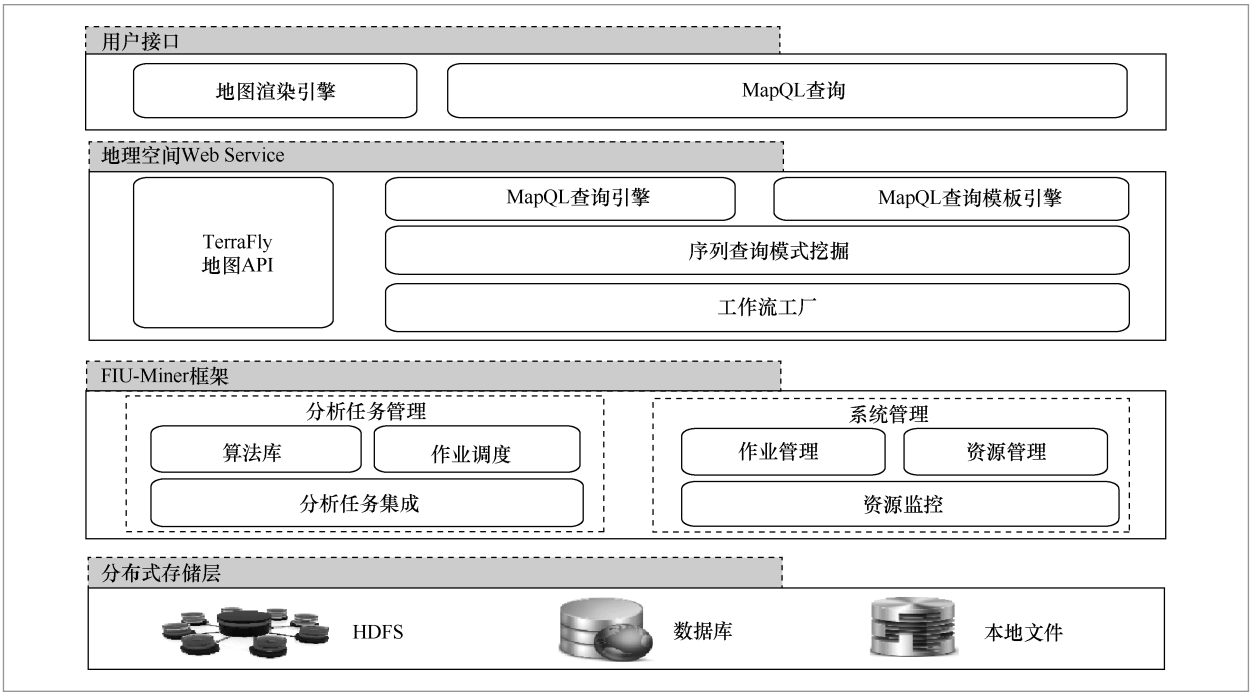


图 11 TerraFlyGeocloud+FIU-Miner 系统架构

● 验证候选地区罪案率和平均收入，确定选择结果；

● 在地图上对结果进行可视化。

这个任务的工作流如图12所示。工作流里面所有的子任务都是由FIU-Miner来调度并在分布式环境中执行的。

5.5 应用亮点评述

上述实际案例中，将FIU-Miner应用于空间数据挖掘，解决了空间数据挖掘中写MapQL查询语句困难、空间分析任务复杂性高及顺序执行空间数据分析工作流效率低这3个主要的难题。用户可以轻松地从TerraFlyGeoCloud的MapQL查询日志中发现顺序查询模式，并利用这些顺序查询模式，在FIU-Miner里面构建空间数据分析任务的工作流。最后使用FIU-Miner强大的分布式处理能力，提高工作流的执行效率。

基于 FIU-Miner的TerraFlyGeoCloud在线空间数据挖掘系统，已成功应用于地理（如国土边界、水位图等）、自然（飓风数据分析）、经济（如房产价格分析、人均收入等数据分析）、医疗（肝癌、关节炎等疾病数据分析）、社会（犯罪数据聚类等分析）等众多领域，受到政府、企业、研究机构及个人的极大重视。

6 FIU-Miner应用实例三：库存管理数据挖掘

FIU-Miner作为库存管理数据挖掘平台已被成功应用于企业，成为商务智能数据挖掘应用中一个典范[113]。

6.1 库存管理数据挖掘任务

库存管理是指对制造业或服务业生产、经营全过程的各种物品、产品以及其

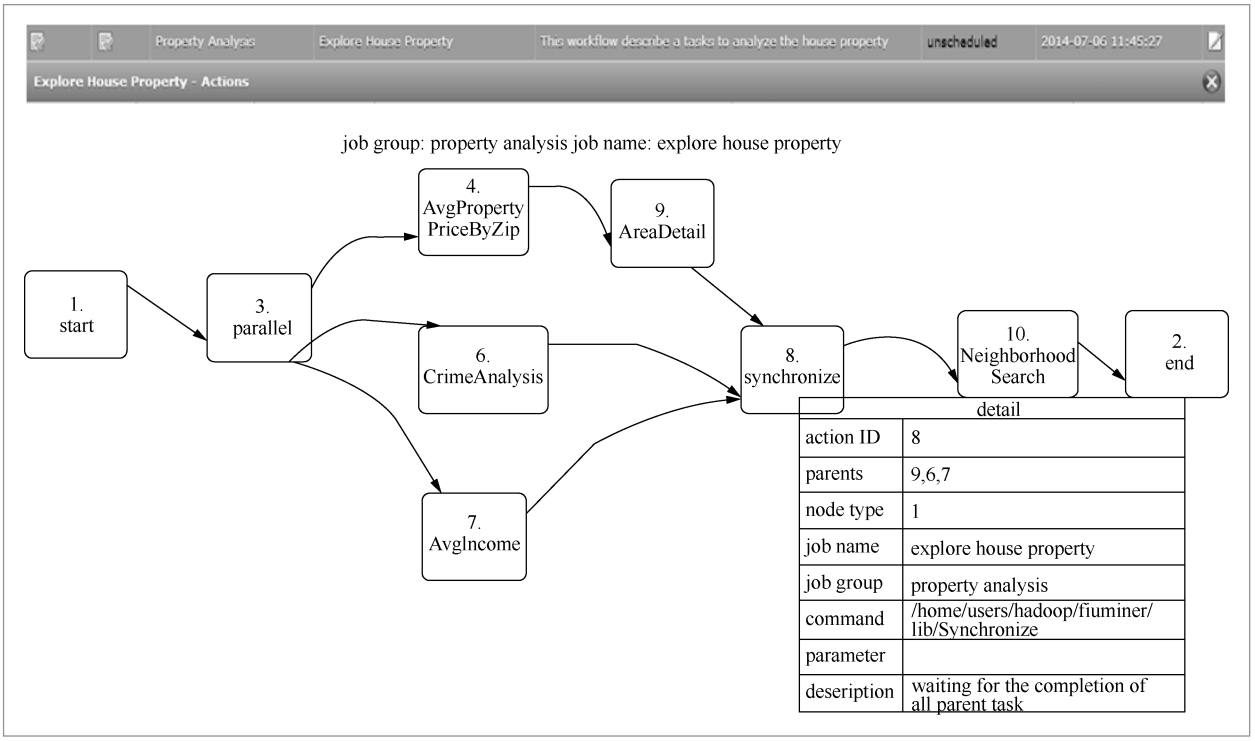


图 12 房产投资案例的工作流程

他资源进行管理和控制，使其储备保持在经济合理的水平上。高效、可靠的库存管理可以为制定合理的货物安全库存量和订货量提供可靠的依据，提高企业管理人员的决策质量，从而减小资金的占用和缺货损失，提高企业的经济效益。当今的零售业，供应商往往需要给不同的地区存储大量的货物，且交易活动复杂频繁，必须提前合理规划好库存方案。现有的库存管理系统（如InFlow和Inventoria）仅仅应用传统的统计分析方法分析现存的库存数据，对当前的库存信息分布进行跟踪监控。进行库存决策时仅考虑单一算法模型，而无法根据综合分析历史数据和市场的实际状况快速做出正确决策方案。因此，如何利用大数据挖掘技术开发智能库存管理平台，实现高效可靠的库存预测、库存异常检测及库龄分析等任务，成为当前大型零售企业亟需解决的问题。

6.2 库存管理数据挖掘挑战

随着库存管理数据日益庞大，库存管理系统处理问题的难度也在不断攀升。以国内某大型电子消费产品制造企业的两大类电视产品（液晶和等离子）交易为例，其库存管理数据挖掘面临的主要挑战如下。

（1）交易记录繁多：现代大型零售企业业务规模庞大，产生的交易记录繁多，从2011年1月到2013年12月有将近6 000万条，约50 GB数据。

（2）属性关系复杂：库存数据属性繁多，记录中包含种类众多的属性，有将近200个；数据层次繁多，在不同数据维度上，记录可属于不同的层次；库存数据和属性相关性复杂等。

（3）处理速度缓慢：现有数据分析工具大多基于内存，无法加载庞大数据集，对数据输入格式要求严格，适用性不强，运行速度慢，无法响应大数据的要求。

因此，现代库存管理需要采用大数据挖掘技术开发高效、可靠、能处理大规模数据的智能库存管理系统。

6.3 具体例子

笔者的研究团队开发了基于FIU-Miner的智能库存管理系统iMiner[13]，该系统为智能库存管理定制了专门的数据挖掘算法，实现了多个功能模块，开发了大规模的数据分析平台系统。

6.3.1 系统概况

图13展示了iMiner系统整体框架、各功能层次和模块。系统自底向上分为物理资源层、任务和系统管理层、数据分析层、用户界面层。该系统分析平台建立在支持高效数据分析的分布式系统——FIU-Miner中。这一分析平台可提供高效率的数据分析处理工作流，并且可以有效地集成多种数据分析工具和语言，如R、Weka、Python、Hadoop等。数据分析层包括了数据预处理和各类数据挖掘算法，其中关键因素提取算法有助于提取对入库/出库量产生较大影响的因素或者对物料异常情况有决定性影响的因素；分布式K近邻算法有助于查找入库/出库行为相似的物料；分布式回归分析有助于对大盘及具体物料的入库/出库量进行有效预测。

系统主要聚焦于库存预测、库存异常检测、库龄挖掘三大核心功能，通过综合评价和集成各种算法的输出使得分析结果更加稳定和准确。用户界面层囊括了多种库存分析结果的展示，用户可以通过属性选择来查看不同的分析结果，也可以通过对个别参数的修改来更新分析结果，实现实时的人机互动。展示结果不仅有列表显示，还提供了各种直观的图表显示，更有利

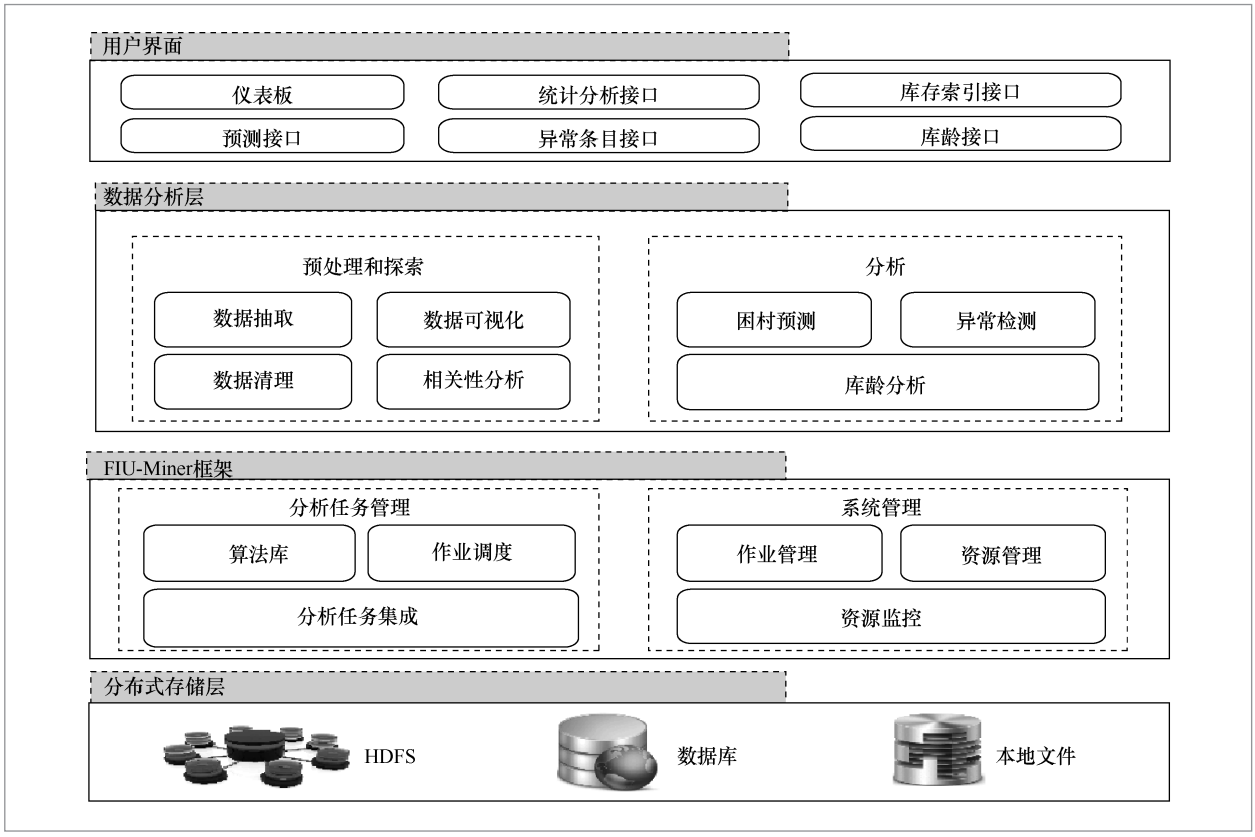


图 13 iMiner 系统架构

于用户接收到数据整体分布、趋势和关键信息点。

6.3.2 系统功能模块

iMiner主要包含库存预测(inventory forecasting)、库存异常检测(inventory anomaly detection)及库龄分析(inventory aging analysis)三大功能模块,如图14所示。

(1)库存预测

库存管理中,精确和可信的库存预测是关键。高效、可靠的预测可以大大减少库存负荷,降低额外的货物维护和损耗。库存数据为标准的时序数据,数据量大、时间跨度长、涵盖面广、规律性差。iMiner采用一种动态预测模型,首先根据历史数据对出库的基数进行预测,而后结合出库数据的长期趋势、周期性因素及事件性因素对基数进行动态调整,从而得到最终的预测结果。

(2)库存异常检测

对库存指标进行监控而达到异常检测的目的,是库存管理中不可或缺的部分。iMiner提供了多种库存指标的实时监控(如库存周转率、库存周转天数、存销比、周转提升率、库存资金周转率)和不同粒度下的指标查询(如按时间周期包括按周和按月、按指定公司和物料、按指定物料类别和公司、按指定物料类别等)。同时,系统从库存数据多个角度入手,及时、准确地发现库存的波动;采用相关物料的协同异常判定,使得对于异常结果的判定更有意义,系统还能够同时准确判定整体性指标变化和个别指标异常。

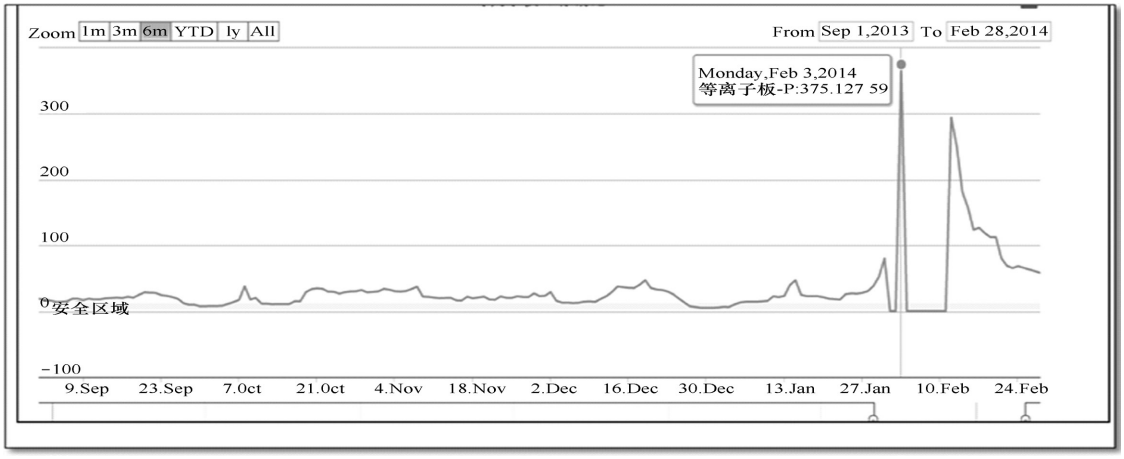
更新到数据库 刷新

提示：数字有下划线的是可配置参数，当配置完所有参数后点击更新，可以看到更新后的数据，数据即可保存到数据库。

物料选择 液晶屏板 等离子板

周期	需求	周期因素	云周期因素影响	平滑预测	趋势	平滑	覆盖(修正)趋势	考虑趋势调整的预测	考虑周期因素的预测	事件影响指标	覆盖(修正)预测指标	覆盖(修正)固定预测	最终预测结果
201 405	1 078 766	1	1 078 766	213 768	213 768	128 261	0.8	534 421	534 421	1	1	0	524 421
201 406	860 558	1	860 558	559 767	345 999	258 904	0	1 207 027	1 207 027	1	1	0	1 207 027
201 407	973 563	1	973 563	680 083	120 316	175 751	0	1 119 461	1 119 461	1	1	0	1 119 461

(a) 库存预测



(b) 库存异常检测

积压物料监控查询条件：直接属性

库龄一年以上库存量 | 库龄一年以上库存量占比

库龄一年以上库存金额 | 库龄一年以上库存金额占比

积压物料监控查询条件：间接属性

上月入库数量 200 | 上月出库数量 100

上次入库数量 | 上次出库数量

上次入库至今时间 | 上次出库至今时间

类型 | 尺寸

选取需要查看的指标然后点击提交 全部选取 提交

根据您所设置的指标得到如下结果

公司	物料名称	时间	一年以下库龄库存量	一年以上库龄库存量
总公司	物料名称1	2014-04-01 00:00:00	257	0
总公司	物料名称2	2014-04-01 00:00:00	236	0
总公司	物料名称3	2014-04-01 00:00:00	328	0

(c) 库龄分析

图 14 iMiner 主要功能模块

(3)库龄分析

库龄挖掘是为了防止货物积压，提前发现潜在积压货物，减小货物积压投资。iMiner系统利用统计回归模型实现库龄分析，并提供了库龄分析的基本工具和高级工具。基本工具允许用户可视化分析给定货物的库龄分布，比较不同货物中当前的和历史的库龄变化，高级工具能够帮助用户找到与积压相关的货物属性。iMiner系统中，库龄挖掘主要包含了库龄相关分类和标准、库龄计算、库龄金额计算以及安全库存的计算等功能模块。

6.4 应用亮点评述

iMiner是一种新的智能库存管理系统，该系统能够帮助大型供应商实现高效的库存管理，着力解决大数据时代现有库存管理面临的两大关键问题。

(1)大规模库存数据分析

iMiner系统分析平台建立在支持高效数据分析的分布式系统——FIU-Miner中。这一分析平台是在分布式环境中管理所有的交易数据，因此，iMiner能够自动配置和执行大规模库存数据预处理和数据分析任务。

(2)复杂库存任务管理

iMiner结合多种先进的数据挖掘算法来分析库存数据。在实践中，系统采用多种回归模型，结合时间序列分析方法来实现库存预测；运用情境感知异常检测算法来识别异常货物；利用统计回归模型来进行库龄分析。从而实现高效、准确的复杂库存任务管理。

基于FIU-Miner的iMiner商务智能库存管理平台已经应用于企业，成功解决了产品出库预测、指标异常检查、库龄挖掘等对企业产品生产和经济效益有重要影响的实际问题。

7 结束语

大数据的复杂特征对数据挖掘在理论和算法研究方面提出了新的要求和挑战。大数据是现象，核心是挖掘数据中蕴含的潜在信息，并使它们发挥价值。数据挖掘是理论技术和实际应用的完美结合。

本文通过目前业界对大数据的理解和认识，结合笔者及其研究团队多年来对大数据挖掘的深入理论研究及广泛的应用研究，综合凝练出大数据的核心架构，即大数据挖掘的本质是应用、算法、数据和平台4个要素的有机结合。在此架构下，从应用的角度重点介绍了研究团队开发的能够快速、有效地进行各类数据挖掘任务的数据挖掘系统FIU-Miner，并具体介绍了基于FIU-Miner的高端制造业数据挖掘、空间数据挖掘和商务智能数据挖掘3个典型的应用案例。FIU-Miner在这些领域的成功应用也说明了提出的数据挖掘核心架构的效用。

致谢

本文总结介绍了笔者研究团队近几年开展的与大数据相关的部分研究和成果。基于这些研究，给出了对大数据的理解和看法，希望能起到抛砖引玉的目的。在这些相关研究中，笔者研究团队得到了许多人的帮助和机构的资助，在此表示衷心感谢。

首先，要大力感谢长虹集团以及其相关科研人员Bing Duan、Ming Lei、Pengnian Wang、Jun Tang、 Dong Liu。他们不仅为笔者研究团队的科研提供了资助，而且其相关研究人员为笔者研究团队提供了非常多宝贵的专业领域知识指导。

其次，要深深感谢美国佛罗里达国际大学的Knowledge Discovery and Research Group（KDRG）研究组的成员：Dr Lei Li、Dr Yexi Jiang、Mr Wei Xue、Dr Jingxuan Li、Dr Chao Shen、Mr Hongtai Li、Dr Liang Tang、Mr Long Wang和Mr Longhui Zhang。他们在相关的研究及项目中付出了辛勤的劳动，提供了许多宝贵的反馈。

最后，要感谢美国佛罗里达国际大学的Naphtali Rishe教授以及其带领的High Performance Database Research Center（HPDRC）研究组里的成员：Mr Mingjin Zhang、Ms Huibo Wang、Dr Yun Lu、Mr Yudong Guang、Mr Chang Liu和Mr Erik Edrosa。他们在TerrayFlyGeocloud项目上与笔者研究团队开展了非常有成效的合作。

参考文献

[1] 严霄凤，张德馨. 大数据研究. 计算机技术与发展，2013，23(4)：168~172
Yan X F, Zhang D X. Big data research. Computer Technology and Development, 2013, 23(4): 168~172

[2] 李国杰. 对大数据的再认识. 大数据，2015001
Li G J. Further understanding of big data. Big Data Research, 2015001

[3] 李涛. 数据挖掘的应用与实践：大数据时代的案例分析. 厦门：厦门大学出版社，2013
Li T. Data Mining Where Theory Meets Practice. Xiamen: Xiamen Press, 2013

[4] Hall M, Frank E, Holmes G, *et al*. The Weka data mining software: an update. SIGKDD Explorations, 2009, 11(1): 10~18

[5] Owen S, Anil R, Dunning T, *et al*. Mahout in Action. Shelter Island: Manning Publications, 2011

[6] Prekopcsak Z, Makrai G, Henk T, *et al*. Radoop: analyzing big data with rapid miner and hadoop. Proceedings of RapidMiner Community Meeting and Conference, Dublin, Ireland, 2011

[7] Yu L, Zheng J, Wu B, *et al*. Bc-pdm: data mining, social network analysis and text mining system based on cloud computing. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 12), Beijing, China, 2012

[8] Zeng C Q, Jiang Y X, Zheng L, *et al*. Fiu-Miner: a fast, integrated, and user-friendly system for data mining in distributed environment. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 13), Chicago, Illinois, USA, 2013: 1506~1509

[9] Lei D, Hitt M A, Goldhar J D. Advanced manufacturing technology: organizational design and strategic flexibility. Organization Studies, 1996, 17(3): 501~523

[10] Zheng L, Zeng C Q, Li L, *et al*. Applying data mining techniques to address critical process optimization needs in advanced manufacturing. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 14), New York, USA, 2014: 1739~1748

[11] Zhang M J, Wang H B, Lu Y, *et al*. TerraFly GeoCloud: an online spatial data analysis and visualization system. ACM Transactions on Intelligent Systems and Technology (TIST), 2015, 6(3)

[12] Zeng C Q, Li H T, Wang H B, *et al*. Optimizing online spatial data analysis with sequential query patterns. Proceedings of the 15th IEEE International Conference on Information Reuse and Integration, San Francisco, CA, USA, 2014

[13] Li L, Shen C, Wang L, *et al*. iMiner: mining inventory data for intelligent management. Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, Shanghai, China, 2014

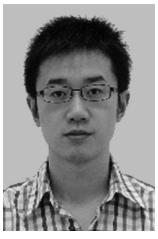
作者简介

李涛，男，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。2004年7月获美国罗彻斯特大学（University of Rochester）计算机科学博士学位，2004–2014年先后任美国佛罗里达国际大学（Florida International University）计算机学院助理教授、副教授（终身教授）、教授（full professor）、研究生主管（graduate program director）。由于在数据挖掘及应用领域成效显著的研究工作，曾多次获得各种荣誉和奖励，其中包括2006年美国国家自然科学基金委颁发的杰出青年教授奖，2010年IBM大规模数据分析创新奖，并于2009年获得佛罗里达国际大学最高学术研究奖。

曾春秋，男，美国佛罗里达国际大学计算机科学博士生，南京邮电大学计算机学院大数据项目组成员。2009年7月–2012年1月为阿里巴巴（中国）网络技术有限公司高级数据工程师。主要研究兴趣包括大规模分布式数据挖掘和系统管理，发表多篇顶级数据挖掘国际期刊和会议论文，参与多本数据挖掘相关应用领域书籍的编写工作。

周武柏，男，美国佛罗里达国际大学计算机科学博士生，南京邮电大学计算机学院大数据项目组成员。主要研究兴趣包括数据挖掘和计算机系统管理，发表多篇顶级数据挖掘国际期刊和会议论文，参与多本数据挖掘相关应用领域书籍的编写工作。

周绮凤，女，博士，厦门大学自动化系副教授。2002年起从事数据挖掘及智能系统方面的研究工作，2014–2015年在美国佛罗里达国际大学访学，主要研究兴趣包括机器学习、数据挖掘及其在可持续发展等领域的应用。



郑理，男，2014年在美国佛罗里达国际大学获得计算机科学博士学位，南京邮电大学计算机学院项目研究员。主要研究兴趣包括信息检索、推荐系统及灾难信息管理，发表多篇顶级数据挖掘国际期刊和会议论文，参与多本数据挖掘相关应用领域书籍编写。

收稿日期：2015–09–30

论文引用格式：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. 大数据，2015041

Li T, Zeng C Q, Zhou W B, *et al*. Data mining in the era of big data: from the application perspective. Big Data Research, 2015041