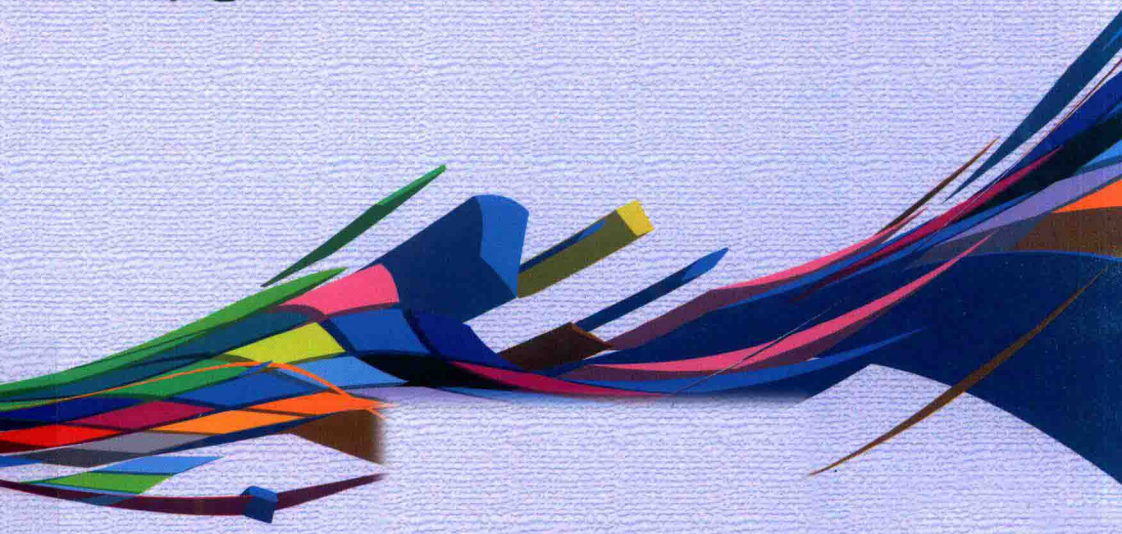
数据流

知识发现

潘 怡 何可可
叶 晖 刘华富 著



華中科技大學出版社
http://www.hustp.com

数据流知识发现

怡 何可可 叶 晖 刘华富 著

SHUJULIU ZHISHI FAXIAN

华中科技大学出版社
http://www.hustp.com
中国·武汉

内 容 简 介

面对“人们被数据淹没，却饥渴于知识”的挑战，数据挖掘和知识发现技术应运而生，并得以蓬勃发展。本书全面介绍了数据流知识发现相关领域的研究内容，涵盖了五个主题：海量知识发现平台架构分析、数据流关联规则知识挖掘、数据流分类知识挖掘、数据流聚类知识挖掘以及数据流序列知识挖掘。

本书可作为高等学校计算机专业的高年级本科生教材或研究生的教材，也可作为从事数据挖掘方面研究工作的科技工作者的参考资料。

图书在版编目(CIP)数据

数据流知识发现/潘怡 等著. —武汉：华中科技大学出版社，2016.12
ISBN 978-7-5680-0527-2

Ⅰ.①数… Ⅱ.①潘… Ⅲ.①数据采集 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2014)第 275437 号

数据流知识发现
Shujuliu Zhishi Faxian

潘 怡 何可可 叶 晖 刘华富 著

策划编辑：王红梅
责任编辑：余 涛
封面设计：三 禾
责任校对：张 琳
责任监印：朱 玢

出版发行：华中科技大学出版社(中国·武汉) 电话：(027)81321913
武汉市东湖新技术开发区华工科技园 邮编：430223

录 排：武汉楚海文化传播有限公司
印 刷：虎彩印艺股份有限公司
开 本：710mm×1000mm 1/16
印 张：13.75
字 数：246 千字
版 次：2016 年 12 月第 1 版第 1 次印刷
定 价：48.00 元

本书若有印装质量问题，请向出版社营销中心调换
全国免费服务热线：400-6679-118 竭诚为您服务

前　言

从20世纪60年代数据库技术诞生以来，以计算机和通信为代表的信息技术，对世界的经济、科技、军事、教育等方面产生了深刻的影响，许多行业积累了大量数据。例如，每天积累数以万计顾客购买交易数据，各种同步卫星每小时传回地球的遥感图像数据高达50 GB，证券市场的客户交易数据等。理解这些海量数据已经远远超出了人的能力，而传统的统计技术及数据管理工具面对如此宏大的数据库也无能为力。这种“数据海洋，知识孤岛”的局面迫切需要新的技术方法，从而能在这些海量数据中获取有用或感兴趣的知识。知识发现的方法便应运而生。

知识发现过程可粗略地理解为：数据准备、数据开采以及结果的解释评估。知识发现是从数据集中抽取和精化的新模式。知识发现的范围非常广泛，可以是工业、农业、军事、社会、商业、科学的数据或卫星观测到的数据。数据的形态有数字、符号、图形、图像、声音等。数据的组织方式也各不相同，可以是有结构、半结构或非结构的。知识发现的结果可以表示成各种形式，包括规则、法则、科学规律、方程式或概念网等。

知识发现的基本流程包括：①选定某个应用领域，包括选定要发现的知识和目标；②建立目标数据集，选择一个数据集或在多数据集的子集上聚焦；③数据预处理，去除噪声或无关数据，去除空白数据域，考虑时间顺序和数据变化等；④数据转换，找到数据的特征表示，用维变换或转换方法减少有效变量的数目或找到数据的不变式；⑤选定算法，用KDD过程中的准则，选择某个特定算法（如神经网络、SVM等）；⑥发现知识，搜索或产生一个特定的模式，或一个特定的函数；⑦解释，解释某个发现的知识，去掉多余的不切题意的模式，转换某个有用的模式，以使用户明白。

在海量数据中，数据流是最具有代表性的一种数据，这是一种实时、连续、有序的数据项序列（顺序由到达时间隐含地表示或显式地由时间戳指定），具有海量性、实时性和动态变化性三个基本特点，具体为：①海量性，由于数据流是随着时间而不停产生的，除非人为干预，否则这种状态会一

直持续下去，没有终点；②实时性，与海量性特点描述相似，由于数据的产生是随着时间而进行的，因此数据流具有时间属性，即实时性，在自然条件下，数据产生的速度和频率也仅与时间属性相关，不受人为因素影响；③动态变化性，由于数据流产生往往受到当前采集环境的影响，因此导致数据流会随着时间的变化而变化。数据流技术自 2000 年提出以来，就一直是国内外数据库及相关研究领域的热点之一。目前与数据流相关的研究工作可大体分为两个方面：数据流管理和数据流知识发现。前者侧重于数据流管理系统的研究，主要包括数据流查询语言、查询模型、操作调度、资源管理、负载控制等问题；而后者则更偏重于理论研究，其研究范围相对较广，主要有数据流统计分析、分类、聚类、变化/突变检测等数据流挖掘问题等。本书正是围绕数据流分析中若干具有深刻技术背景和广泛应用前景的热点问题展开研究的。

本书从数据流知识发现研究背景出发，以数据流为主要研究对象，在综述和分析国内外数据流数据分析和知识发现研究现状的基础上，重点介绍数据流分析中的五个关键问题。

（1）数据流知识发现系统架构的研究。本书第 2 章探讨了海量数据管理的历史，介绍了包括并行数据库结构、并行事务管理、基于云的数据挖掘架构等内容。重点探讨了海量数据知识发现硬件平台架构的发展历史，以及传统并行数据库事务管理的 TS-PRTTS 算法及事务并发控制 MRTT 算法。

（2）数据流关联规则知识发现。本书第 3 章探讨了与数据流关联规则知识发现有关的内容，如数据流关联规则知识发现算法的特点等。重点介绍了关联规则知识发现的核心问题之一——数据流频繁模式知识挖掘，针对现有算法需要消耗较多的 CPU 时间和内存，并不完全适合数据流知识发现中的频繁项集挖掘，为了兼顾解决频繁模式挖掘中时间和空间问题，作者提出一种称为前缀频繁模式树数据结构以及建立在其上的挖掘算法，它能紧凑完整地存放整个事务项集，同时挖掘算法采用一种遍历树的方式可以直接有效地挖掘出所有闭合频繁模式。同时，讨论支持度/可信度指标存在的问题，给出一种快速挖掘高效益项集算法 FHUI-Growth（fast high utility item），能够对关联规则加入了数量和利润来分析，将一些无意义关联规则挖掘变成一个有趣的、有价值的关联规则分析。

（3）数据流分类知识发现。本书第 4 章探讨了知识发现与机器学习研究领域的一个重要任务与课题——分类知识发现。针对周期性数据流概念漂移问题，作者提出了一种基于隐马尔可夫模型的周期性流数据分类算

法——HMM_SDC 算法，基于强大的统计基础的 HMM 和相应的计算效率，为数据流分类知识发现算法提供了有效的解决途径。

（4）数据流聚类知识发现。本书第 5 章系统地介绍了数据流挖掘中聚类分析的基本概念、支撑技术、发展历史和最新的发展水平，展示了数据流聚类分析的发展现状和整体蓝图；详细地探究了一些经典的、主流的聚类算法。

（5）数据流时序和序列知识发现。本书第 6 章介绍了时序和序列数据流挖掘的概念以及时序和序列数据流挖掘的目的；概要地介绍了时序和序列数据流挖掘中的几个主要问题，并对其中的一些问题列出了已有的解决方法。最后，本书还对这个领域的一些问题作了简要的讨论。

随着信息技术的飞速发展，数据流广泛应用于金融市场、网络监控、电信数据管理、传感器网络等领域中。数据流数据的处理已成为一个热点研究问题。本书针对数据流知识发现的硬件平台以及一些核心算法展开了讨论，提出了一些有效的解决方案，对于推进数据流知识发现的研究，具有较好的理论意义和应用价值。

著　者

2016 年 7 月

目　录

1 绪 论

1.1 什么是知识发现

数据是指关于事件的一组离散的客观事实，通常采用结构化的记录描述，它是构成信息和知识的原始材料，数据的价值在对数据进行验证或测试后才能被体现出来。信息是一种以文档或音频、视频交流形式表现的消息，人们通过对数据进行系统的组织、整理和分析，使其产生相关性，信息的价值必须反映出数据的准确性，应及时地发送和允许方便的访问，以满足用户需求。信息的另一重要特性是时效性，失去时效性的信息就是毫无意义的数据流。知识是人类在改造自然的实践中所获得的认识和经验的总和，它是“一种能够改变某些人或事物的信息，这既包括使信息成为行动的基础方式，也包括通过对信息的运用使事物的某个个体或机构有能力进行改变或进行更为有效的行为的方式”[1]。纵观人类科学的发展历程，人类的发展史就是知识的发展史，人类所掌握的知识虽然按照学科分类缤纷庞杂，但是概括起来却无非两种类型的知识：定义性的概括以及规律性的总结，它们分别代表了人类思维中最为基础和珍贵的学习能力——归纳及演绎。从数据到信息到知识的发展形成了如图 1-1 所示的金字塔体系。

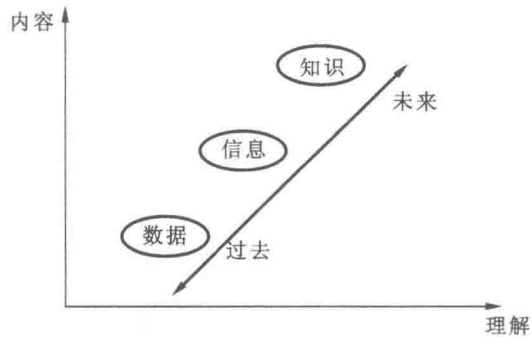


图 1-1　数据、信息和知识的金字塔体系

自 20 世纪 70 年代以来，数据库技术发展迅速，人类产生数据和获取数据的能力迅速提高，从制造业到服务业，从生物研究到空间探索，无时无刻不在产生着大量数据。这些数据涉及科学研究、医疗、金融、电信、互联网等各个领域。数据类型复杂多样，不但包括常见的数值型、符号型数据，在电信和互联网等网络相关邻域中还包括网络型数据，如电信用户呼叫图、互联网拓扑信息以及在实际应用中常出现的由各类数据组合而成的复杂型数据。这些数据在不同部门和地区之间进行着日益频繁的交换，逐步形成以数据库为核心的数据存储与管理模式。目前大型数据库系统已经能对海量数据进行高效的存取、排序、查询等管理工作，然而这些数据库操作都是被动的，人类主动处理和分析数据的能力却相当有限，互联网的兴起更加剧了“数据爆炸，知识匮乏”[2]这一趋势。例如，商业上条形码的普遍使用使得每天积累数以万计的顾客购买数据；各种同步卫星每小时传回地球的遥感图像数据就达 50 GB；世界著名咨询公司 Nielsen 公司 2008 年统计占全球 60%搜索市场的 Google 公司平均每天响应近 1.5 亿次的搜索请求；北京奥运会官方网站日访问量高达 1.83 亿次，创历届之最；而 2013 年每月多达 1.28 亿美国人在 YouTube 网站观看视频内容。人们渴望能够对数据进行较高层次的处理，从中发现知识和规律，以帮助人们更好地利用数据进行决策和研究。然而由于数据的繁杂，人工处理方式很难找出关于数据的较为全面的信息。因此，对智能、自动化的数据分析工具的需求越来越强烈，数据库中的知识发现(knowledge discovery)便应运而生。

1977 年，美国斯坦福大学计算机科学家费根鲍姆教授在第 5 届国际人工智能会议上提出了“知识工程”的概念，他认为“知识工程是人工智能的原理和方法，对那些需要专家知识才能解决的应用难题提供求解的手段。

恰当运用专家知识的获取、表达和推理过程的构成与解释，是设计基于知识的系统的重要技术问题”。1989 年 8 月在美国底特律召开的第 11 届国际人工智能联合会议的专题讨论会上，学者们首次提出了数据库中的知识发现一词，也即通常所说的数据挖掘[3]（knowledge discovery in database，KDD），这是一个多学科交叉研究领域，是指一个从大量的数据中提取出未知的、潜在有用的、可理解的模式的高级过程。其中，

数据：是用来描述事物的信息，是进一步发现知识的基础。

未知：经过数据挖掘提取出的模式必须是新颖的。

潜在有用：提取出的模式应该是有意义的，可以用某些标准来衡量。

可被人理解：数据集中隐含的模式要以容易被人理解的形式表现出来，从而帮助人们更好地理解数据中所包含的信息。

高级过程：一个多步骤的处理过程，多步骤之间相互影响、反复调整，形成螺旋式上升的高级过程。

知识发现融合了数据库（database）、人工智能（artificial intelligence）、机器学习（machine learning）、统计学（statistics）、知识工程（knowledge engineering）、面向对象方法（object oriented method）、信息检索（information retrieval）、高性能计算（high performance computing）以及数据可视化（data visualization）等技术的研究成果。以知识发现领域研究最核心的部分——算法为例，算法的好坏将直接影响到所发现知识。目前大多数的知识发现的研究都集中在算法和应用上。知识发现的算法与传统分析工具不同的是，运用的是模糊自适应的算法，解决传统方法所无法解决的高精度函数拟合，以及准确自动分类等数据之间的关联问题。它是一个反复的过程，通常需要将上述提到的步骤反复进行，最后得到用户可以理解的结论。而由于数据库中的数据被形象地喻为矿床，因此知识发现又经常被称为数据挖掘。1995 年，首届 KDD & Data Mining 国际会议在加拿大蒙特利尔召开，经过十几年的研究，产生了许多新概念和方法。特别是最近几年，一些基本概念和方法趋于清晰，研究重点逐渐从发现方法到专项系统应用，并且注重多种发现策略和技术等集成，以及各学科之间相互渗透，这些变化也体现在历届国际会议的主题中。例如，2012 年 KDD 大会主题为“大数据挖掘”，2013 年大会主题为“在生物、健康保护以及医学领域的 KDD 应用”，2014 年大会主题为“用数据科学造福社会”。会议吸引了来自统计、机器学习、数据库、国际互联网、生物信息学、多媒体、自然语言处理、人机交互、社会网络计算、高性能计算以及大数据挖掘等众多领域的专家和学者。会议讨论的主题既包含传统的知识发现问题，涵盖了图建模

和图挖掘、动态图分析、可扩展图算法、数据流、文本挖掘、推荐系统、排序推荐、主动学习、监督学习、迁移学习、特征工程、聚类算法、异常检测、话题建模、社区挖掘、国际互联网挖掘、降维算法等领域；也增加了不少新兴问题，如大数据统计，大数据可扩展算法，大规模问题优化和学习算法，社交媒体、社交网络和信息网络传播问题，商务应用，工业应用，政府工程，健康问题，安全问题，隐私问题，欺诈问题，环境问题，教育问题，医药学，地域服务，可解释性模型，监控与维护，广告与交通，群智与市场等，知识发现领域的新兴问题更偏重实际应用中所产生的大规模数据和非结构化数据，偏重解决实际问题。

1.2 知识发现的过程

知识发现是从数据中发现有用知识的整个过程，它不仅是面向特定数据库的简单检索、查询和调用，还需要对这些数据进行各类统计、分析、综合和推理，以指导实际问题的求解，以发现事件间的相互关联，或者利用已有的数据对未来的活动进行预测。

知识发现将人们对数据的应用，从最简单的低层次末端查询操作，提高到为各级决策者提供决策支持。它的基本流程如下。

(1) 选定某个应用领域：包括选定要发现的知识和目标，明确所要完成的数据挖掘任务和性质。

(2) 建立目标数据集：从数据库中提取与 KDD 相关的数据，选择一个数据集或在多数据集的子集上聚焦。

(3) 数据预处理：从与 KDD 相关的数据集合中除去明显错误的数据和冗余的数据，包括空白数据域，进一步精简所选数据中的有用部分，以使 KDD 更有效。

(4) 数据转换：找到数据的特征表示，考虑时间顺序和数据变化，用维变换或转换方法减少有效变量的数目或找到数据的不变式。

(5) 选定算法：遵守 KDD 过程中的准则，使用合适的算法和参数，寻求感兴趣的模式，并用一定的方法表达成某种易于理解的形式。

(6) 发现知识：搜索或产生一个特定的模式，或一个特定的函数。

(7) 解释：对发现的知识模式进行解释和评估，根据支持度、确信度、简洁性和新颖性等用户满意度指标，去掉多余的不切题意的模式，转换某个有用的模式，以使用户明白，必要时返回上述步骤中反复进行。

一个简化了的数据流知识发现体系如图 1-2 所示。其中，数据预处理层主要是对数据进行加工以改善原始数据的质量，为数据的连续查询和复杂分析打下基础，包括对数据进行降噪、压缩编码、修正，以减少存储空间和传输时间，也有研究者称完成此层主要功能的模块为封装器。信息提取层主要是对数据中感兴趣的目标进行检测和识别，该层处理的对象具有基本的语义单位（如关系数据库中的元组），处理后的结果是对数据的特点与性质进行描述的概要数据，如类别符号、小波系数、分位数或其他统计信息等。查询分析处理层主要完成各种基本查询和复杂分析。基本查询处理操作完成一般的选择、连接、投影和聚集等连续查询请求。复杂分析处理操作则是在信息提取层生成的概要信息基础上，进一步研究数据中各种因素的性质和相互之间的关系，比如直方图和采样用于分布式数据处理环境中的形成数据均分策略，小波大纲用于连续查询的近似解答。基本查询操作既可以直接在预处理过的数据流上执行，也可以在概要上执行。基于概要的复杂分析操作的抽象程度较高，基本上是对经抽象处理过的符号进行运算，其处理过程和方法与人类的思维推理有许多类似之处。接口层主要完成基本查询请求与复杂分析请求的提交，查询或分析请求的发出者可以是应用程序或最终用户，同时接口层也负责将查询结果和分析结果形成易于理解的知识表示返回给用户。

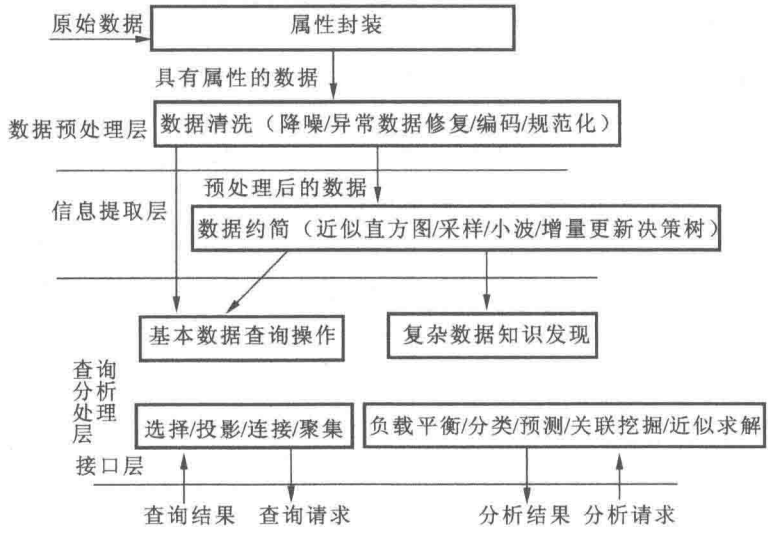


图 1-2 简化的知识发现体系

1.3 新型数据流应用

科学技术的高速发展和信息技术的广泛使用引发了计算机网络流量控制、网络安全监控、环境监测、股市交易和事务日志分析等新型应用，这些应用产生了一类新的数据类型——数据流。如果我们把传统的存储于数据库中的数据称为静止的数据，那么数据流就是动态的数据。在传统的数据库管理系统(data base management system，DBMS)中，数据可以独立于应用，由加载器集中存储到磁盘上的关系表或记录里面，存储的数据是一个固定的集合。用户可以随时输入所需的查询，DBMS 接受这样的查询，并且返回用户所关心的结果。虽然查询可能很频繁，但是因为数据比较稳定和持久，系统可以在同一时刻得到所需要的全部数据。在传统数据库中，如何保证数据的完整性和可维护性，以及如何提高系统的性能和可靠性等得到了充分的研究。尽管传统数据库技术在过去几十年中取得了巨大的成功，可是数据规模宏大、增长迅速的大量新型数据流应用，很难再遵循“存储—索引—查询”的数据库处理模式，它们对传统数据库技术提出了新的挑战，该类常见的典型应用如下。

1. 互联网应用

如前所述，互联网用户数量的增加直接导致网络通信量的急剧上升，互联网站点访问、即时消息通信、电子邮件和在线视频等各类应用都产生大量的数据。例如，网络数据包在转发的过程中连续、动态地流经网络节点(路由器、网关等)。针对这些海量的流式数据，用户是无法实时地对诸如网络访问日志、点击次数等信息进行精确查询的，但可以实时地进行流量估计和近似精确的分析，这有几个方面潜在的好处：维护网站的正常运行，避免受到分布式拒绝服务攻击或因访问量过大而引起网站的瘫痪；通过对访问某一站点的数据包进行监控，发现频繁的访问模式，检测带有恶意的网络入侵，对可疑地址源发送的数据包进行拦截和拒绝访问，以加强网络安全等；通过近似分析大量用户的行为，评估不同时段对不同类型商业广告的价值，当需要投放广告的时候，就可以最大化广告的性价比；分析不同网站之间用户点击率的相关关系，可以为客户推荐适当的产品，如根据搜索引擎中排名顺序和相关的零售商受关注的程度，可以为客户提供高质量的搜索结果，等等。

在2014年巴西世界杯期间，百度、微软、谷歌等国内外技术巨头和大投资银行高盛、德意志银行乃至彭博等，均推出了结果预测，互联网公司方面，除了雅虎，几乎全面大胜。百度、微软、谷歌正确预测了全部的十六强，以及八强；微软、百度预测对了全部的四强，谷歌在四强的预测中惜败（谷歌只预测了八强）；在半决赛中，百度和微软甚至还准确预测了巴西对德国的赛果。而在世界杯比赛开始前，德国足协就与知名的SAP公司合作，“私家定制”了一款名为“Match Insight”的足球解决方案，用以迅速收集、分析、处理球员和球队的技术数据，基于“数字和事实”优化球队配置，提升球队作战能力，并通过分析对手技术数据，找到在世界杯比赛中的“制敌”方式。通过这一数据工具，德国队教练可以迅速评估比赛状况、每个球员的特点和表现、球员的防守范围、对方球队的空当区等信息。通过这些信息，教练可以更有效地对球员上场时间、位置、技战术等情况优化配置，以提升球队表现。这款数据分析系统首先通过摄像头、传感器等工具捕捉到球员跑动速度、位置、控球时间、防御范围、动作细节等大量数据，并传入数据库，在10分钟内，10名球员用3个球进行训练，可产生超过700万个可供分析的数据点，使用工具可迅速对这些数据进行后台分析处理。

股票和基金等金融交易报价数据瞬息万变且规模庞大，单上海证券交易所2014年平均日交易次数为400万次，从数据规模上看，数据量极大且数据产生的速率非常快；从应用的角度来看，机构和股民都希望能够持续不断地分析交易数据，实时地获得查询结果，寻找获利机会。股市中典型的在线分析包括：跟踪股票报价的变化，实时监控各只股票的走势，根据以往的经验进行实时的模式匹配和趋势判断；实时比较并分析多只股票之间的关联关系，寻找交易机会；实时挖掘股票市场中的异常现象，及时提醒股民，包括投资者交易异常、交易违规、利润操纵、股价异常波动等情况；实时分析外部重大的新闻事件对股市的影响，预测股价的短期和长期走势，等等。银行信用卡日交易数据量也十分巨大，建立客户消费行为模式与实时鉴别信用卡欺诈的可疑消费行为，有利于保护银行的利益和客户的合法权益。

2. 传感器网络应用

传感器网络综合了传感器、无线通信、嵌入式计算和分布式信息处理等技术而被应用于地理环境检测、交通控制、空间探索、移动物体追踪等诸多领域。网络由部署在监测区域内大量廉价微型传感器节点组成，通过无线通信方式形成一个多跳的自组织网络系统，其目的是以协作方式感知、