華信經管 創新系列

Business Data Stream Mining:
Models, Methods and Applications

商业数据流挖掘
模型、方法及应用

琚春华　封　毅 □著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

華信經管 創新系列

Business Data Stream Mining:
Models, Methods and Applications

商业数据流挖掘
模型、方法及应用

琚春华　封　毅 □著

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

内 容 简 介

本书是针对商业数据流挖掘模型、方法及应用的学术研究专著。全书共分9章:第1章为绪论,综述了商业数据流挖掘的相关概念和研究进展并描述了全书的概貌,起到了导引的作用;第2章和第3章为模型篇,主要介绍了商业数据流管理模型与商业数据流概念漂移模型;第4、5、6章为方法篇,分别从商业数据流关联规则、分类、聚类三大方面对商业数据流的挖掘方法进行了详细阐述;第7章和第8章共同构成了应用篇,主要介绍了商业数据流挖掘的两方面应用案例;最后一章对商业数据流挖掘模型、方法及应用进行了归纳总结,并对商业数据流挖掘的未来发展做出展望。

本书适于从事数据挖掘和智能信息处理研发的科技工作者阅读使用,也可作为高等院校数据挖掘、智能信息处理、管理科学与工程等管理类和信息类相关专业研究生和本科生的教学参考书。

图书在版编目(CIP)数据

商业数据流挖掘模型、方法及应用/琚春华,封毅著. —北京:电子工业出版社,2016.6
ISBN 978-7-121-28965-1

Ⅰ.①商… Ⅱ.①琚… ②封… Ⅲ.①商业信息-数据采集 Ⅳ.①F713.51

中国版本图书馆CIP数据核字(2016)第123159号

策划编辑:王赫男
责任编辑:石会敏　　特约编辑:赵翠芝　侯学明
印　　刷:北京京海印刷厂
装　　订:北京京海印刷厂
出版发行:电子工业出版社
　　北京市海淀区万寿路173信箱　　邮编:100036
开　　本:787×1092　1/16　印张:13.75　　字数:352千字
版　　次:2016年6月第1版
印　　次:2016年6月第1次印刷
定　　价:39.00元

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010)88254888,88258888。

质量投诉请发邮件至zlts@phei.com.cn,盗版侵权举报请发邮件至dbqq@phei.com.cn。

本书咨询联系方式:(010)88254553,88254537。

前　言

21 世纪以来，数据已前所未有地成为国家和企业发展的重要战略资源，成为提高一个组织乃至一个国家战略竞争力的核心，也是实施科学管理与决策的基础。如何有效利用数据与发现知识，已成为各方关注的关键性问题。在这样的背景下，企业、政府和各类组织从爆炸性增长的海量数据中挖掘出有价值信息的需求变得更加强烈，宣告着大数据(Big Data)时代的来临，把基于海量数据的分析和挖掘推上了新的高度。正如《纽约时报》2012 年 2 月的一篇专栏中所称，大数据时代已经降临，在商业、经济及其他领域中，决策将日益基于数据和分析而做出，而并非基于经验和直觉。亚马逊首席科学家 Andreas Weigend 更直接提出“数据是新的石油”。数据，已逐渐成为和土地、人力、技术、资本并列的要素。如何更为有效地从海量数据中挖掘和提取出有价值的信息和知识，提升组织在大数据时代的竞争力，已成为企业、政府和各类组织极为关切的核心问题。

目前，数据挖掘技术及其延展的大数据技术所面向的分析对象，已从传统的静态数据集，转向以数据流为代表的复杂数据集。尤其是近年来随着物联网、云计算等信息技术的不断兴起和互联网应用的飞速发展，更为海量的数据以数据流的形式大量涌现，并且仍在以几何方式继续保持增长。典型的如商业领域的超市交易记录、网络购物数据、网络搜索请求、电信通话记录、银行 ATM 交易记录等，科学领域的天文观测数据、气象观测数据等。这些数据流具有数据量大、变化快、要求快速响应、适合于线性扫描、随机存取代价高等特点。与科学领域的数据流相比，商业领域的数据流来源更为广泛，出现得更为频繁、更为多样，且与大众的关系更为密切。更为重要的是，商业领域的数据流来源于企业和组织的实际业务流程，往往包含着企业和组织的运行状态、管理要求、影响因素、变化特征等价值极高的信息，并蕴含着企业和组织的运行规律、变化趋势等动态变化情况，值得开展针对性的数据挖掘研究。与此同时，商业数据流所具有的分布性明显、概念特征易漂移等内在特点，也给计算机带来了存储空间、计算速度和挖掘方法等方面的挑战，难以采用传统的数据挖掘模型和算法进行处理，需要探索和研究面向商业数据流的数据挖掘模型与方法。

本书即是针对商业数据流挖掘模型、方法及应用的学术研究专著。全书共分 9 章，涉及商业数据流挖掘的模型、方法及应用三方面，具体结构和内容主要是：第 1 章为绪论，综述了商业数据流挖掘的相关概念和研究进展，并描述了全书的概貌，起到了导引的作用；第 2 章和第 3 章为模型篇，其中第 2 章论述商业数据流管理模型，主要从商业数据流特点、商业数据流管理模型、商业数据流预处理模型三个方面进行了阐述；第 3 章论述商业数据流概念漂移模型，主要包括商业数据流概念漂移描述模型、特征提取模型和概念漂移检测模型；第 4、5、6 章为方法篇，分别从商业数据流关联规则、商业数据流分类、商业数据流聚类三大方面对商业数据流的挖掘方法进行了详细阐述，提出了一些具有实践意义的模型和算法；在此基础上，第 7 章和第 8 章共同构成了应用篇，主要介绍了商业数据流挖掘

的两方面应用案例：分布式零售数据挖掘与网购数据挖掘；最后一章对商业数据流挖掘模型、方法及应用进行了归纳总结，并对商业数据流挖掘的未来发展做出展望。本书由琚春华和封毅执笔，感谢许翀寰、郭飞鹏、邹江波等参与了写作。本书适于从事数据挖掘和智能信息处理研发的科技工作者阅读使用，也可作为高等院校数据挖掘、智能信息处理、管理科学与工程等管理类和信息类相关专业研究生和本科生的教学参考书。

大数据时代的号角已然响起，未来基于数据的研究和应用将在企业、组织甚至国家层面的竞争中发挥越来越重要的作用。衷心希望本书能为数据流挖掘的相关研究者和实践者带来点滴帮助，成为读者扬帆大数据时代的助力剂。感谢业内专家对本书内容的指导、推荐和帮助。由于作者水平有限，书中难免有疏漏和不妥之处，恳请读者批评指正。

琚春华　封毅

2016 年 2 月

目　录

Chapter 1

第1章 绪 论

本章综述了商业数据流挖掘的相关概念和研究进展，并描述了全书的概貌，起到了导引的作用。内容上首先对商业数据流挖掘的相关研究背景进行了概述，然后对国内外商业数据流挖掘的研究现状进行了综述，随后对商业数据流挖掘的基本概念进行了阐述，包括基本定义、流程，以及主要的模型和方法，在此基础上，介绍了本书中两个重点论述的商业数据流挖掘应用案例。最后，概述了全书的主要内容和结构。

1.1 背景概述

1.1.1 数据挖掘

在当今这个大数据时代，随着计算机、互联网技术的飞速发展，人们积累了大量的数据，这些数据记录了丰富的资料，如个人消费习惯、超市物品选择、网页浏览取向等。然而隐藏在这些数据后面的是具有潜在价值的信息，这些信息与知识已经成为国家和企业发展的重要战略资源，是提高一个组织乃至一个国家战略竞争力的核心，也是实施科学管理与决策的基础。如何获取信息与发现知识，尤其是如何快速高效地在动态变化和高维特征的海量数据流中获取信息和发现知识是政府部门、企业等关注的焦点问题，也是学术界研究的热点问题。

数据挖掘技术，作为从海量数据中提取有价值模式和知识的重要手段，兴起于20世纪90年代，并在近年来得到了持续高速的发展[1]。MIT出版社知名科技期刊 *Technology Review* 将数据挖掘技术评选为改变世界的十大新兴技术之一[2]。目前，数据挖掘已经成为子领域众多、内涵非常丰富的学科领域。数据挖掘是在大型数据存储库中，自动地发现有用信息的过程，它作为一个强有力的数据分析工具，可以发现数据中潜在的模式和规律(如一组规则、聚类、决策树、依赖网络或其他方式表示的知识)。数据挖掘技术广泛地应用于科技研究、商业智能处理等领域，如定向销售、顾客分析、超市商品分类、商店分布和欺诈检测等。

数据挖掘是一门融合人工智能、数据库、统计学、可视化等多学科的交叉学科，其主要功能是在大量数据中自动发现潜在的有用知识[3]。数据挖掘技术的出现为解决海量数据下的复杂分析任务提供了新的解决方法，例如在金融领域，数据挖掘技术为银行开展以信用卡客户细分为基础的针对性营销提供了有效工具。通过数据挖掘技术，在银行内部丰富的客户资料和历史消费数据中挖掘潜在的有用信息，对信用卡客户进行市场细分，将客户划分出一个可以识别的客户群类，并在此基础上进行金融产品创新，实现提供差别化的服务来提高银行自身的市场竞争力。数据挖掘中的决策树法、贝叶斯分类、神经网络法等都可以作为客户细分的基础技术。

1.1.2 数据流挖掘

目前，数据挖掘技术及其延展的大数据技术所面向的分析对象，已从传统的静态数据集，转向以数据流为代表的复杂数据集。尤其是近年来随着物联网、云计算等信息技术的不断兴起和互联网应用的飞速发展，更为海量的数据以数据流的形式大量涌现，并且仍在以几何方式继续保持增长。典型的如：商业领域的超市交易记录、网络搜索请求、电信通话记录、银行 ATM 交易记录等，科学领域的天文观测数据、气象观测数据等。这些数据流具有数据量大、可无限、变化快、要求快速响应、适合于线性扫描、随机存取代价高等特点。与科学领域的数据流相比，商业领域的数据流出现得更为频繁、更为多样，且与大众的关系更为密切。更为重要的是，商业领域的数据流往往蕴含着企业和组织的运行状态、管理要求、影响因素、变化特征等价值极高的信息，更能反映企业和组织的运行规律、变化趋势等动态变化情况，值得开展针对性的数据挖掘研究。与此同时，商业数据流所具有的分布性明显、概念特征易漂移等内在特点，也给计算机带来了存储空间、计算速度和挖掘方法等方面的挑战，难以采用传统的数据挖掘模型和算法进行处理，需要探索和研究面向商业数据流的数据挖掘模型与方法。

数据流的广泛应用、商业价值以及现有技术存在的诸多问题引起了国内外专家和学者的关注，数据库领域的专家、学者逐步开始关注数据流的管理和挖掘的研究工作。十多年来，机器学习、数据挖掘、人工智能等领域的权威期刊，以及 VLDB、SIGMOD 和 ICDE 等颇具声望的数据库国际会议发表了大量数据流方面的文章。

数据流挖掘(Data Streams Mining)在电子商务、移动商务和基于情境感知计算的信息推荐系统中具有重要的研究价值和广阔的市场前景。在处理诸如新的客户的知识发现、客户兴趣漂移、客户的动态层次划分等分析时，传统的基于静态数据的数据挖掘技术存在局限性，它依赖历史记录而不能使用最新的客户访问信息，从而限制了个性化、即时性的推荐服务。电子商务中涉及的知识作用强大，例如，客户关系管理系统中挖掘到的知识可用来调整服务策略，以给特定客户提供个性化、即时性的服务，从而获得最大利润，这就要求能够快速准确地处理数据流信息。

数据流挖掘方面的研究成果主要集中在数据流的频繁模式挖掘、动态分类和聚类等方面，可利用即时的客户点击、浏览、购买的行为信息实现客户分类、分层、客户背景分析、客户偏好分析等。移动商务是指各种具有商业活动能力的实体利用网络和先进的通

信技术进行的各项商务贸易活动。通过移动商务，用户可突破传统电子商务受时空限制的客观因素，随时随地利用通信终端查找、选择和购买商品和服务。情境感知计算[4]是移动商务服务的基础，这种分析地点、附近人或物体的身份以及这些对象发生的变化的方法，能够即时地将个体的情境信息传送到信息系统中，通过决策系统利用情境信息为任何地方的使用者提供与任务有关的信息和/或服务，即个性化推荐服务。通过情景感知技术获得以数据流形式出现的用户信息，可综合用户的既往历史资料和感知的用户情景提供实时的针对性服务。在移动商务用户接受研究时，个性化已作为接受模型要素出现。而如何挖掘隐含情境背景知识的数据流，识别人的行为模式及其变化，是情境感知服务的基础。

2015年7月最新发布的《中国互联网络发展状况统计报告》在网民属性方面给出了男女比例、网民年龄分布、网民学历、职业结构、收入分布结构、地区分布等因素，这些因素都是挖掘一个网络用户个体或群体的行为特征时需要考虑的背景信息。截至2014年6月，我国网民手机上网使用率已超越PC端，手机成第一大上网终端设备。手机等移动互联网设备的随身性和及时性，使得信息获取类应用在手机等移动终端上网应用中相对比较普及。在移动互联网领域，如何根据以数据流形式产生的移动终端用户的情境信息、用户行为信息，为用户提供及时、准确的“流式信息推荐服务”是一个值得研究的课题。

数据流的动态变化特征是人们感兴趣的概念，即目标概念常常发生根本性的改变[5]。例如，顾客购买偏好随购买时间、购买对象的范围、通货膨胀率、个体背景、环境情境等因素的改变而发生变化；天气预测规则随季节、地域等因素的改变而发生变化。这种由于数据流中上下文变化而导致所隐含的目标概念发生变化甚至是根本性改变的现象规律，称为数据流中的概念漂移(Concept Drift)[6]。

有效地挖掘概念漂移情境下数据流和信息推荐是近年来数据流挖掘领域的热点问题。数据流中概念漂移现象是客观存在的，由于数据生成过程中不可估计的变化(如个体情境的变化)，从历史数据中发现的“迟到”的知识数据通常是无用的。因此，在数据流挖掘工作中发现这些规律并应用于实际的商业场景中，可有效提高推荐准确率，为决策支持提供重要依据。

1.2 商业数据流挖掘主要研究概况

1.2.1 国外研究现状

目前商业数据流挖掘研究主要集中在商业数据流频繁模式挖掘研究、商业数据流分类挖掘研究和商业数据流聚类挖掘研究。

1. 国外商业数据流频繁模式挖掘研究

Manerikar 等[7]讨论了在动态零售市场环境下，客户行为的变化对数据流挖掘结果的影响，并提出了一种基于关联规则的挖掘方法，有助于管理者更好地制定市场决策。Tanbeer 等人[8]针对数据流实时性以及海量性特点，将传统权重频繁模式挖掘方法进行改进，提出一种基于滑动窗口的频繁项集挖掘方法。该方法首先将数据流静态化后进行单遍扫描，并且对不同数据块的重要性进行评估，而后针对重要的数据块从中提取出有价值知识，为频繁模式发觉提供帮助。Cormode 等人[9]对现存数据流频繁模式挖掘方法进行汇总，并对其进行实验，以验证何种方法对数据流频繁项集最为有效。Feigenblat 等人[10]使用多项式衰减函数，对频繁项集进行筛选，并最终找到真正的频繁项集，此种方法避免了数据流的干扰，通过实验验证所找到的频繁项集的正确性。Homem 等人[11]提出一种融合基于计算器和草图技术的数据流频繁项集挖掘方法，试图在海量数据中通过模糊的方式获取答案，减少使用精确算法所要消耗的时间，并且给出了所提出方法的误差下限证明以及错误估计等内容。Cafaro 等人[12]从并行计算角度对数据流频繁项集挖掘进行探索，通过将传统频繁挖掘算法并行化，从而大大提高频繁项寻找速度。Memar 等人[13]在原有技术的基础上，对滑动窗口技术进行了深入分析，找到了滑动窗口大小与频繁挖掘之间的关联关系。为了加快挖掘速度，使用锁定位序列技术的方法，能够从当前窗口中快速寻找和保存频繁项集。Unil Yun 等人[14]提出一种挖掘数据流最大频繁项集算法，该算法充分考虑了数据流权重因素影响，对数据流单次扫描，有效挖掘频繁模式。Lee 等人[15]提出一种基于滑动窗口模型的挖掘数据流最大频繁项集算法，该算法也考虑了数据流加权问题，进而有效挖掘频繁模式。

2. 国外商业数据流分类挖掘研究

VFDT(Very Fast Decision Tree)[16]是一种基于 Hoeffding 不等式建立决策树的方法，它通过不断地将叶节点替换为决策节点而生成。其中每个叶节点都保存有关于属性值的统计信息，这些统计信息用于计算基于属性值的测试。当一个新样本到达后，在沿着决策树从上到下遍历的过程中，它在树的每个节点都进行划分测试，根据不同的属性取值进入不同的分支，最终到达树的叶节点。当数据到达叶节点后，节点上的统计信息就被更新，同时该节点基于属性值的测试值就被重新计算。如果统计信息计算显示测试满足一定的条件，则该叶节点变为决策节点。新的决策节点根据属性的可能取值的数目产生相应数目的子女叶节点。决策节点只保存该节点的划分测试所需要的信息。Aboalsamh 等人[17]提出一种使用增量式学习方法的数据流分类模型，通过将分类模型学习过程进行增量式处理，能够加快模型自我更新速度，从而通过不断更新模型的方式解决分类问题。Torres 等人[18]提出一种新的方法——基于相似度的数据流分类算法，采用新型的插入和删除策略。

3. 国外商业数据流聚类挖掘研究

2009 年 Alex 等人[19]针对海量实时数据流在聚类过程中受到时间和空间局限的问题，

将两种聚类方法，即神经云(Neural Gas,NG)与自组织映射(Self-Organizing Map，SOM)算法进行改进，提出一种基于单次通过(One Pass)的NG和SOM模型，其主要思想是利用快速划分方法将动态数据流转变为静态数据块，进而使用数据块评估函数对这些数据块进行评估，即评估数据块是否对聚类结果产生帮助的程度，如果此数据块对聚类的帮助程度达不到标准则将其删除，否则被使用进行聚类。Antonellis等人[20]同样利用了上述思想，针对网页点击数据流进行实验和分析。对数据流聚类来说，除了使用纯数值方法进行相似性聚类外，还可以使用带有标签的数据进行聚类。Al-Mulla等人[21]针对多数据流聚类问题进行研究，提出使用增量式聚类方法和窗口缓冲区机制，对每一个数据流中的每一个滑动窗口中的样本进行聚类，并使用衰减函数，对聚类结果进行取舍。LEE等人[22]对高维数据流的聚类方法进行了深入的探讨，提出一种聚类统计树的方法。Luhr等人[23]提出一种增量式聚类方法。Beringer等人[24]通过对原始数据进行离散傅里叶变换得到数据流的低频分量，并比较各数据流低频分量的欧氏距离来测量数据流间的相似度。Rehman等人[25]提出一种新的数据流聚类算法——基于HyCARC的超椭圆聚类算法，采用滑动窗口技术处理到来的数据流，用马氏距离代替传统的欧氏距离计算相关性。Kim等人[26]提出一种间隔聚类方法，采用一种错误控制机制，将根据用户指定的允许误差作为阈值条件进行聚类。

1.2.2 国内研究现状

1. 国内商业数据流频繁模式挖掘研究

在国内商业数据流频繁模式挖掘方面，国内的Li H等人提出了DSMFI算法[27]，采取了基于前缀树的约减模式的表示方法，又将自顶向下的频繁项集发现策略进行了扩展，实现数据流上所有历史数据的频繁项集挖掘。Teng等人提出了FTPDS算法[28]，该算法能够在线地对事务流进行一次扫描生成候选频繁模式，然后用回归分析的方法表示约减的模式，数据流管理方法采用的是基于滑动窗口的技术，挖掘的结果是时序频繁模式，具有单遍扫描在线统计和基于回归分析的约减模式表示两大特征。卢琦蓓等人[29]认为传统的信息挖掘技术已经无法满足大数据环境下日益复杂的应用需求，而分布式数据挖掘技术是解决这个难题的一种手段，提出了基于改进型频繁模式树(FP-Tree)的分布式关联分类算法。该方法首先在各局部节点优化FP-Tree，生成局部条件模式树(CFP-Tree)，再通过各节点间传送CFP-Tree构建全局CFP-Tree；其次，在挖掘全局CFP-Tree时通过计算显著度来获取初始的全局显著分类规则；最后，利用剪枝策略选取一个较小规则集来构造全局的关联分类器。该算法能够有效降低网络通信量，提高信息挖掘效率，同时保证剪枝的质量和规则的统计显著性，提高分类的精确性。

2. 国内商业数据流分类挖掘研究

分类的基本方法有基于距离的、基于决策树的、基于贝叶斯分类的和基于规则归纳的分类算法等，是一种有监督学习算法，把新到来的数据项映射到给定类别中的某一个

类别。区别于传统数据，数据流变化非常快、具有非统计特性，因此建立在训练样本是随机采样且服从一定的统计分布的前提下的传统分类算法[30]不能很好地应用于数据流挖掘。

在数据流分类挖掘方面，王鹏[31]提出了一种基于频繁模式和 P-tree 的数据流分类算法，通过数据流中的频繁模式进行分类，在压缩数据的同时保存了数据中的分类信息，可以很好地处理训练集包含大量缺失取值的应用，提高了分类准确性。为实现对概念漂移数据流的分类，CVFDT 通过增加与新样例相关联的计数，减少与旧样例相关联样例来更新统计量，但是当靠近 CVFDT 树根的属性不再通过 Hoeffding 界，树的大部分必须重新生成。决策树并不是处理概念漂移的最自然的方法，因此出现了分类器系综或者叫集成分类器(Classifier Ensemble)的方法[32-33]。分类器系综是一种基于集成学习(Ensemble Learning)的算法，其主要思想是从数据流的相继块训练一个分类系综或一组分类器，个体分类器根据其在随时间变化的环境中期望分类准确率加权，截取准确率最高的 k 个分类器，基于这 k 个基础分类器加权投票得出分类结果。实验结果表明，系综方法比任何单个的分类器的准确率更高，许多基于集成学习的算法在隐含概念漂移的流数据分类问题上取得了较好的效果，因而成为这一领域的主流方法之一。

3. 国内商业数据流聚类挖掘研究

数据流广泛存在于现实生活中，如股市分析、传感网络、网络监控等。通过国内外学者的研究，提出了很多基于传统聚类算法的数据流聚类算法，也有将多种原有技术有机结合的新颖算法，其中基于网格的数据流聚类技术[34-35]、子空间聚类技术[34-35]、混合属性数据流聚类[36]这三类方法受到了广泛关注。

CAStream 算法[34]是采用子空间聚类的思想，提出了网格单元估计算法，改进了金字塔的时间结构用来记录子空间网格统计信息，然后在离线部分采用深度优先搜索方法进行聚类及其演化分析。该算法可以有效处理高维数据流，并可以发现任意形状的聚类。CluStream 是一种基于用户、联机聚类查询与演变分析的数据流聚类算法，该算法扩展了 BIRCH 算法的聚类特征，用来记录簇内所有数据的信息。CluStream 算法将聚类过程分为两个阶段：第一阶段为在线聚类阶段，首先提取数据流特征，然后对数据流进行初步聚类，并采用金字塔时间窗口存储相关信息；第二阶段为离线聚类阶段，此阶段会根据客户的请求，对在线阶段产生的聚类结果再次分析。在后来的研究中，这种处理数据流的双层聚类框架得到了众多学者的认可与借鉴[34-38]。

1.3 商业数据流挖掘的基本概念

1.3.1 商业数据流的基本定义

商业数据流与数据流形式类似，它是伴随着近年来商业领域数据采集技术的提高而形成的一种连续的有序数据元素集合。相比传统静态的关系型商业数据，它可以看作是

一系列连续而有序的二维元素集合(T, I)，其中，T为数据元素的时间标记，反映了数据的时间序列特征，如对应不同时间点的收银机商品销售流水单；I则对应商业数据流的实体要素，反映了数据流背后所体现出来的物流、商流、资金流等各类商业信息，如连锁零售企业不同时点的商品库存信息、商品销售信息、财务结算信息、商品采购信息和运输信息等。

1.3.2 商业数据流挖掘的基本流程

商业数据流挖掘的基本流程如图1-1所示，考虑到商业数据流是按时间顺序的、快速变化的、海量的和潜在无限的，对流过的数据进行存储或者挖掘时进行多遍扫描都是不可取的，因此在数据流挖掘时序设计缓冲区对时序到来的数据流进行缓存，并在缓存的基础上对数据流进行清理及集成，根据时间粒度不同，分为在线和离线存储两种方式，在选择或变换中形成流式概要数据，将该数据导入增量式的商业数据流挖掘引擎中进行数据流挖掘生成商业模型，最后对模型进行评估及可视化表示，产生支持商业决策的知识。

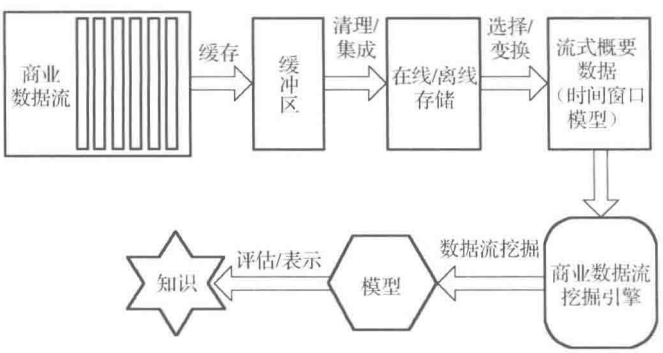

图1-1 商业数据流挖掘的基本流程

1.3.3 商业数据流挖掘的主要模型和方法

目前，网上交易、股票交易、工业生成控制等以流的形式数据成为一种常见数据模式，更促进了数据流挖掘的发展，面向数据流挖掘的模型已得到了广泛的研究，主要集中在商业数据流预处理模型、商业数据流频繁项挖掘模型、实时商业数据流聚类模型、商业数据流动态分类模型，以及数据流变化导致的概念漂移及离群点检测方面的研究，其主要模型和方法包括：基于粗集的事务项压缩方法、基于相关关系的属性压缩方法、基于小波变换的时间序列树状结构概要构造方法、基于密度的聚类降噪算法、基于时序轮盘模型的数据流频繁模式挖掘模型、基于模糊积分融合的数据流分类挖掘模型、基于增量存储树的集成贝叶斯分类挖掘模型、基于密度的数据流聚类算法、基于小波网络的多维时间序列耦合特征聚类方法等。本文后面的章节将对商业数据流的主要挖掘模型和方法进行详细阐述，如图1-2所示。