持续时态数据挖掘及其实现机制

Continuous Temporal data Mining and Implementation Mechanism

潘 定 /著

经济科学出版社
Economic Science Press

国家自然科学基金项目（70372024）阶段性研究成果
国家自然科学基金项目（70771044）阶段性研究成果

持续时态数据挖掘及其实现机制

Continuous Temporal data Mining and Implementation Mechanism

潘 定 /著

经济科学出版社
Economic Science Press

责任编辑：吕　萍　于海汛
责任校对：徐领柱
版式设计：代小卫
技术编辑：邱　天

持续时态数据挖掘及其实现机制
潘　定　著
经济科学出版社出版、发行　新华书店经销
社址：北京市海淀区阜成路甲28号　邮编：100036
总编室电话：88191217　发行部电话：88191540
网址：www.esp.com.cn
电子邮件：esp@esp.com.cn
北京汉德鼎印刷厂印刷
永胜装订厂装订
690×990　16开　14.25印张　180000字
2008年3月第一版　2008年3月第一次印刷
印数：0001—2000册
ISBN 978-7-5058-6903-5/F·6155　定价：20.00元
（图书出现印装问题，本社负责调换）

前　言

世间万物总是随着时间的流逝而演变、进化，其具体表现为相关属性数据的不断变化。同时，事物活动规则也随时间呈现出某种明显或隐含的变化趋势，使得已发现的规则不再生效，新的规则不断孕育，并有待我们进一步的探索和发现。对数据和规则变化趋势的研究可归结为对时态数据挖掘（Temporal Data Mining，TDM）理论与应用的研究。

在数据资源中存在着许多随时间变化且不同时间状态相互关联的数据，即时态数据，如股市交易指数、超市销售、Web 访问流、气象观测、临床数据等。近几年，急剧增加的时态数据已经远远超出了人类的直接理解能力，大量的处理需求使 TDM 成为数据挖掘领域的重要分支。但是，从实际应用情况来看却与学术研究极不相称，特别是应用中难以实现自治、持续的时态数据挖掘支持机制，而这种机制是归纳分析数据和规则变化趋势所必需的。本书试图在数据仓库环境中，以高阶数据挖掘和领域知识融合为研究线索，为实现持续的时态数据挖掘过程提供理论和应用方法，形成及时发现事物动态演化规律的有效机制。

本书首先分析、总结了现有的持续数据挖掘研究中涉及的主要思路、模型和体系结构，然后分类讨论时态序列分析的主要方法和技术，涉及相似性度量、序列抽象表示和搜索，以及挖掘任务和算法操作，研究了高阶数据挖掘的方法和技术。

现有 TDM 研究主要在于挖掘方法和算法方面，较少涉及理论

研究。作为后续研究的基础，本书采用形式化方法，基于一阶线性时态逻辑，形成持续时态数据挖掘的理论基础。在现实应用中，按分段有限模型可获得规则的度量值序列。基于信息扩散原理，作者研究了度量值序列的参数扩散估计方法，形成一种支持高阶数据挖掘的有效方法。

传统的数据挖掘过程局限于按照特定挖掘目的，组织数据集并执行挖掘的过程模型，其应用背景是独立的挖掘任务。为了提高数据挖掘的有效性和效率，本书研究一种能够对不断增量的数据集合实现自治、持续的数据挖掘过程模型和体系结构。过程模型着重于说明数据挖掘的逻辑处理流程，体系结构描述支持处理流程的各构件间的结构化关系和机制。通过设立一种通用的通信机制，利用本体服务封装领域知识，结合元数据的支持使挖掘过程能透明地存取异构数据源，支持逻辑和物理数据独立性。

元数据提供数据的上下文描述信息。元数据管理作为数据仓库的重要组成部分，为所有基于数据仓库的分析应用，如数据挖掘、多维分析提供有力的支持。为了在数据仓库的环境中构建持续时态数据挖掘的实现机制，有效的元数据管理是一个关键基础条件。作者简要介绍迄今为止相关元数据的重要研究项目，讨论数据仓库中元数据管理的内容、管理标准和体系结构及其实现技术，并比较了目前主要商用元数据管理产品的特性。

为了更好地通过本体共享语义信息，本书还讨论了一种轻量级的本体存储库及其服务接口。同时，为了解决元数据的互操作问题，及时、动态地获得数据源的物理属性和操作信息，提出一种基于公共仓库元模型（CWM）规范的实时分布元数据管理的体系结构，为持续的数据挖掘过程提供准确、透明的数据和知识资源描述。

本书总结了近年来作者在时态数据挖掘、元数据管理等方面的主要研究成果，研究工作得到国家自然科学基金项目（No. 70372024、70771044）、科技部创新基金项目（No. 2005V41C0311）、广州市科技攻关项目（No. 2004Z3 - D0351、2006Z3 - D3101）和暨南大学人文社

会科学发展基金项目（No. 006JSYJ013）的资助，在此一并表示感谢！

特别感谢西安交通大学电信学院的沈钧毅教授！沈老师的为人和治学精神，使作者感触颇深，不仅引导完成学业，且终身受益。

在本书的写作过程中，作者参阅了大量国内外的著作和学术论文。在此，对参考文献中提及和未提及的研究人员表示衷心的感谢！

衷心感谢经济科学出版社的吕萍主任为本书问世所提供的多方面支持和帮助！

本书中的疏漏与错误，敬请同行专家批评指正。

潘 定

2007 年 12 月于暨南大学惠全楼

目　录

第1章
绪　论

1.1 研究背景

数据是人们用各种工具和手段观察外部世界所得到的原始材料，它本身并没有什么直接的价值，有价值的是蕴藏在其中的信息和知识。随着信息技术的快速发展、互联网的广泛普及，人们获取、储存数据的手段和方式已变得非常便捷和廉价。数据的增长积累速度已远远超过数据总结和分析能力的提升速度，致使各行业的数据量以空前的速度急速增长。因而，一方面有大量的“数据过剩”；而另一方面却又严重地“信息匮乏”[①]。如何开发适宜于从海量数据中自动、高效地提取所需的有用知识，已成为众多学科共同关注的焦点。

数据库中知识发现（Knowledge Discovery in Databases，简称KDD）是适应这一现实要求而发展起来的一种数据分析技术。KDD是指从数据中识别出有效、新颖、潜在有用的和最终可理解的模式

① Han, J., Kamber, M. Data Mining Concepts and Techniques (2nd edition). Morgan Kaufmann Publishers. 2006.

或规则的非平凡过程①。KDD 是一个多阶段的处理过程，可能需要多次的反复循环和调整。这些典型的处理包括数据存储、目标数据选择、清洗、预处理、交换和缩减、数据挖掘、结果评价和解释等步骤。通常 KDD 可简要地概括为：数据准备、实施挖掘及结果评价和解释三个主要阶段。

KDD 是一个介于统计学、机器学习、模式识别、数据库技术、数据可视化和并行计算等领域的交叉新兴学科，也因此有了许多不同的术语和名称。除 KDD 之外，主要有“数据挖掘”、“智能数据分析”、“信息发现”、“探索式数据分析”，等等。鉴于本书讨论涉及的文献中，对有关时态数据的 KDD 通常使用“数据挖掘”一词，因此，在以下论述中，除非特别说明，“数据挖掘”皆为 KDD 的同义词。

知识发现自 1989 年提出以来，经历了从单纯结构数据挖掘到复杂类型（时态数据、Web 数据、多媒体数据等）数据挖掘的发展过程。在大量数据资源中存在着许多随时间变化且不同时间状态相互关联的数据，即时态数据。传统上，大多数数据挖掘问题不考虑时间因素。近几年，主要关注数据动态特性的时态数据挖掘（Temporal Data Mining，简称 TDM）成为学术界研究的热点之一。现实生活中，时态数据随处可见，如股市交易指数、超市销售、Web 访问、气象观测、临床数据等。

数据仓库是以统一形式存储、访问的中央共享数据库，是一个面向主题的、集成的、非易失的、随时间变化的用于支持管理人员决策的数据集合。数据仓库技术的出现为海量数据的存储和管理提供了新的操作平台，其中保存的时态数据集合构成反映事件变动演化的时态序列。数据仓库先期解决了数据清洗、数据变换、数据集成、数据载入和定期增量等问题，可以看做已完成了大部分的数据

① Hand, D., Mannila, H., Smyth, P. Principles of Data Mining. Cambridge, Mass.: The MIT Press, 2001.

挖掘预处理工作。作为数据挖掘的对象，数据仓库提供了统一、干净、高质量的数据源。另一方面，数据仓库本身无法发现时态数据中存在的各类知识（如关系、规则等），也不能根据现有的数据预测未来的发展趋势，因而，通过数据挖掘的应用，数据仓库才能真正发挥其支持管理决策的作用。

随着数据量的不断增长，特别是数据仓库和数据挖掘技术的普及应用，人们已不满足于发现数据中的静态规则，而更加关心数据和规则的动态演化趋势。因而，对规则的动态演化规律的归纳分析（动态演化规则挖掘）已成为数据挖掘及其应用领域中一个亟待解决的重要问题。例如，在现实生活中，财务报告的及时性和高质量一直是人们长期关注的焦点。绝大多数现有财务报告模式的技术手段因局限于解决静态、手工劳动的自动化问题，而难以形成实质性突破，远远不能满足经营、投资和监管的实时信息需求。如果能将动态演化规则挖掘的技术应用于财务报告供应链，探索智能化的财务信息实时动态监控手段，就可导出全新的主动（Active）财务报告模式，为建立智能化的企业经营活动全程监控机制奠定坚实基础。

动态演化规则挖掘的研究可归结为对 TDM 理论和应用的研究。在现实应用中，建立持续的 TDM 支持环境是实现动态演化规则挖掘、形成实时跟踪和发现动态演化规律的基本方法。对时态数据的知识发现已经有多方面的研究。统计学研究时间序列，主要集中在实数或离散数量的预测，时间序列分析方法可用于时态数据的表示、度量和预测；机器学习研究涉及离散序列的模式发现和预测；数据库研究涉及多维数据的存储、索引，以及对大数据集的相似性查询。

1993 年 Agrawal 等首先发表了关于时间序列相似搜索的研究论文[①]，此后，相关研究项目和研究者不断增加。IBM 公司的 Agrawal 和 UC Irvine 的 M. Pazzani 研究小组是较早且持续开展相关研究的团

① Agrawal, R., Faloutsos, C., Swami, A. Efficient Similarity search in Sequence Databases. Proc. of the FODO'93. LNCS 730, Heidelberg: Springer, 1993. pp. 69 – 84.

体，UC Riverside 的 E. Keogh 和 UI Urbana-Champaign 的 J. Han 研究小组是目前 TDM 界最活跃的群体。还有美国 UC Santa Barbara、Maryland、Massachus，以及南澳洲 Flinders 大学等也活跃着相应的研究小组。国内的相关研究大约从 2000 年开始起步，复旦大学、浙江大学、西安交大、中国科大等曾有相关的研究。

高维度、高特征相关性和大量噪音是时态数据的独特结构。这种特征使许多经典算法难以发挥作用，大大增加了挖掘算法的研究难度。目前，TDM 的研究主要以数据抽象表示、序列距离度量和任务描述为主线，以有效的搜索或优化算法为中心。近几年，TDM 引起了学术界的极大关注，已经有数百篇的论文对索引、分类、聚类和分割算法进行了讨论，并取得了长足进步，但也存在着一些明显的不足。Keogh 等[①]收集了多种高质量学术会议和期刊上发表的 360 篇数据挖掘论文，并选择考察了其中最常引用的 57 篇文章。通过用涉及金融、医学、生物、化学、网络等行业的 50 种时间序列数据集合对 25 篇文章中的算法进行彻底的测试后，他们发现由于实验中的瑕疵，这些算法存在不同程度的实现偏差和数据偏差，从而大大降低了算法的实用性。

从应用的观点看，动态演化规则挖掘的研究为解决许多实际问题提供了新的分析手段。例如，企业财务数据是一类典型的时态数据，财务数据挖掘一直就是 TDM 的重要应用领域。现有财务报告模式的不足及其变革问题已经引起广泛关注。随着互联网应用的普及和企业财务报告“按需定制”概念的逐渐成熟，Beattie[②] 结合企业信息保密问题提出“分层报告结构”和向不同类型用户预打包信

① Keogh, E. J., Kasetty, S. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. Data Mining and Knowledge Discovery. 2003, 7 (4): pp. 349 - 371.

② Beattie, V. Business reporting: The inevitable change? Edinburgh: Institute of Chartered Accountants of Scotland, 1999.

息的方法，Jensen 等[①]提出客户化报告模式，美国 AICPA 提出增强型报告模式[②]，李瑞生[③]提出需求决定模式等。这些研究虽然对财务报告的改进做出了有益的贡献，然而它们仅限于概念或理论框架范畴，缺乏相应技术支持手段的创新研究和实验。近两年的研究开始关注如何将 IT 前沿技术与公司报告模式变革相结合，如潘琰等[④]综合应用 XBRL 和 Web 服务等前沿技术提出柔性化的公司按需报告模式，试图对会计信息质量特征进行改进与创新。但是，它们仍属于仅利用计算机大容量、高速运算的特点，将交易事项保存在数据库中，在需要时能够灵活地“拉”出所需财务报告的模式，体现的是被动的信息服务理念，其实质仅仅解决手工劳动的自动化问题。

事实上，计算机具备强大的归纳和推理能力，有可能从大量数据中持续地归纳出动态演化规则，进而实时推断所发生的事项。主动财务报告模式的基本思想是从繁杂琐碎的经营活动过程中自动发现重要事项，并实时向指定的人、指定的地点自动“推”出所需财务报告，从而将被动报告转变为智能主动报告。这种模式需要利用动态演化规则挖掘的研究成果，建立智能化的监控机制，能部分解决脑力劳动的自动化问题。

20 世纪 90 年代至今，数据挖掘以归纳挖掘算法为中心，沿着任务描述、数据准备、执行挖掘和评价解释的经典挖掘过程主线展开研究，较少涉及对挖掘过程的内在机理的研究。从实际应用的情况来看也与学术研究状况极不相称。许多企业对知识发现的需求十分迫切，却迄今还无法建立能持续运行的挖掘应用。因而，如何建立有效的知识发现内在机制，实现自治、持续的数据挖掘过程环境

① Jensen, R. E., Xiao, J. Z. Customized financial reporting, networked databases and distributed file sharing. Accounting Horizons, 2001, 15 (3): pp. 209 - 222.

② AICPA. Quality and transparency in business reporting—A call for action…in the public interest. Dec. 2003. http://www.aicpa.org.

③ 李瑞生、续慧泓：《论网络环境下的会计报告模式》，载《会计研究》，2004 年第 1 期，第 60 ~ 64 页。

④ 潘琰、林琳：《公司报告模式再造：基于 XBRL 与 Web 服务的柔性报告模式》，载《会计研究》2007 年第 5 期，第 80 ~ 87 页。

是现实世界给学术界提出的新挑战。特别是面对客观世界的海量数据，还有许多现有结构模型和技术方法难以解决的问题，具体表现在：

（1）经典数据挖掘过程局限于按照特定挖掘目的，组织数据集并执行挖掘的过程模型，其研究背景主要针对独立、静态的挖掘任务。为了提高数据挖掘的有效性和效率，需要一种能够对不断增量的数据仓库环境实现自治、持续数据挖掘的过程模型。虽然增量挖掘是一个重要的研究方向之一，但其技术方法通常仍局限于特定的场景。客观世界中的数据是不断扩充和持续变化的，人们对规律的认识和评价也在不断发展。如果不能持续地跟踪、发现数据的变化趋势和隐含的有用知识，将降低数据挖掘的有效性和应用价值。

（2）较少研究涉及高阶规则（Higher Order Rule）或元规则（Meta-rule）的归纳。高阶规则描述直接由数据归纳获得的一阶规则的动态演化特征，可以更易理解、自然的方式对数据中蕴含的全局或局部规律进行描述。高阶规则归纳方法的研究对跟踪规则进化、解释客观现象能力的提升有极大帮助。

（3）人们已经发现，已知规则（以前发现的知识和领域知识）可以大大改进挖掘效率和效果，但目前缺乏通用的机制，用于支持将已知规则耦合到数据挖掘过程。知识的耦合应运用在对动态挖掘进程的结果进行理解、评价和估计等方面，需要有标准的规则存取接口和同步进化机制。

（4）现有的数据挖掘系统主要面向技术人员，要求操作用户必须熟练掌握数据挖掘技术及其过程。只有富有经验的数据挖掘专家才能有效制定挖掘方案、选择相关归纳算法，才能帮助明确解释、评价发现的结果，从而大大限制了挖掘系统的使用面。

（5）现有的数据挖掘系统难以灵活设置和变更挖掘任务、数据源、模型选择和算法参数；难以及时地掌握挖掘环境、数据源变更的情况；难以快速地对已有规则进行实时维护、更新。

这些不足使得数据挖掘技术在现实中难以推广应用，反过来也对数据挖掘的理论和算法研究发展形成了一定的障碍。因此，解决这些问题对有效地从时态数据中自动抽取隐含的和潜在有用信息具有重要的现实应用意义和理论价值。

基于上述背景，本书的研究在国家自然科学基金项目（70372024、70771044）、科技部创新基金项目（2005V41C0311）、广州市科技攻关项目（2004Z3－D0351、2006Z3－D3101）和暨南大学人文社会科学发展基金项目（006JSYJ013）的资助下，对持续时态数据挖掘及其实现机制中的有关理论和应用方法进行研究与探索，以期为在数据仓库环境中有效地实施动态演化规则挖掘，实时跟踪、发现时态序列中隐含的动态信息和知识提供基本理论与方法，促进数据挖掘应用研究和实践的发展。

1.2 研究内容及其目标

复杂的海量时态数据已经远远超出了人类的直观理解能力，大量的处理需求使 TDM 成为数据挖掘领域的一个重要的分支。从 1993 年开始的 TDM 研究在 21 世纪初形成学术界关注的一个重要研究课题。由于时态数据结构的复杂性，大大增加了挖掘过程的实现难度。本书试图在数据仓库环境中，以高阶数据挖掘和领域知识融合为研究线索，为实现持续的时态数据挖掘过程提供理论和应用方法，形成及时发现事物动态演化规律的有效机制。

具体来说，本书涉及以下几个方面的研究内容：

（1）现有持续数据挖掘的相关技术研究。经典数据挖掘过程规定了一个多步骤、多阶段的处理过程，为数据挖掘的研究内容和方向奠定了基础。但如果将这种挖掘处理过程简单地移植到现实的应用环境中，却往往难以满足用户的需求。从经营管理的角度看，企