学校编码：10384　　分类号_____密级_
学号：14220051300776　　UDC

厦门大学

硕 士 学 位 论 文

时态数据挖掘中关联规则的应用研究

Study on the Application of Association Rules in Temporal Data Mining

乐燕波

指导教师姓名：朱建平 教授
专 业 名 称：经济信息管理学

论文提交日期：2008 年 4 月
论文答辩时间：2008 年 月
学位授予日期：2008 年 月

答辩委员会主席：_____
评 阅 人：_____

2008 年 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（ ），在　　年解密后适用本授权书。

2、不保密（ ）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

摘要

20 世纪 80 年代末，数据挖掘作为一个全新的研究领域悄然出现并迅速发展。数据挖掘的研究目的是在大型数据集中发现那些隐藏的、人们感兴趣的具有特定规律的信息。作为数据挖掘对象之一的时态数据库是由随时间变化的一系列序列值或事件组成的数据库。时态数据挖掘的研究对商业、金融、医疗诊断、科学与工程等领域的数据分析具有重要意义，因而时态数据的挖掘方法也成为数据挖掘的一个研究热点。

关联规则一直是近年来数据挖掘和人工智能领域研究的热点课题，目前在客户关系管理、医学、生物等领域已有应用。传统的关联规则挖掘过程通常不考虑时间约束，如购物篮分析等。由于时态数据库规模不断壮大，重要性不断加强，如何将关联规则挖掘应用到时态数据库中以获得有价值的时态关联规则是一个非常值得研究的课题。另外关联规则在证券市场中的应用尚处于起步阶段，如何有效地将关联规则应用于证券交易系统数据库分析，也需要进一步的探索。

本文从研究所处的背景出发，详细阐述了数据挖掘技术及时态数据挖掘的研究现状，介绍了关联规则的相关理论及其在时态数据挖掘中的应用。在深入分析时态关联规则的基础上，本文改进了一般时态关联规则挖掘算法，提出了加权时态关联规则的概念，初步研究了它应用的可行性。用该方法挖掘时态数据库，挖掘结果能更好地反映客户购买习惯的变迁，为市场营销提供决策支持。文章还对不同关联规则挖掘模型得到的结果进行了比较和分析，验证了加权思想的有效性和合理性。

本文还创新性地提出了有时态约束的数值型关联规则挖掘方法，并将其引入股市技术分析指标有效性的检验中。选取相对强弱指标RSI，收集交易数据进行实证分析，得出了若干条有用的数值型关联规则，为技术分析的实际应用和投资操作实践提供了指导。

关键词：时态数据挖掘；关联规则；技术分析

Abstract

In the later 1980's, Data Mining-a new research field, appears gradually and develops rapidly. The study purpose of Data Mining is to find the regular information which hided in large data set and people interested in. As one of the mining objects of Data Mining, temporal data base is composed by series of sequence or affair. The study of temporal data mining plays an important role in the data analysis of Commerce, Finance, Science and Engineer. So the methods of temporal data mining became a research hotspot of Data Mining.

Association rules is always a hot research subject in Data Mining and Artificial Intelligence which has already been applied in the areas of CRM, Physic, Biology and so on. Traditional Association rules mining processes like Market Basket Analysis seldom consider the temporal constrain. Since the scale and importance of temporal data base ceaselessly grandness, how to apply Association rules mining into temporal data base became a subject which highly worth our research. Besides that, the application of Association rules in security market is still at the start stage. How to effectively use Association rules mining in the data analysis of security trade system also need our study.

Start form the background of the study this paper expatiate the technology of Data mining and status quo of temporal data mining. It also introduce the theory of association rules and its application in temporal data mining.On the basis of deeply analyze the temporal association rules mining, the paper improves the mining arithmetic of commonly temporal association rules, brings forward Weighted Temporal Association Rules and study its feasibility of application. Using the improved method to mine the temporal data base, the results can finely detect the diversification of purchasing custom of customers and offer supports to the marketing decision making. The paper also compares and analyzes several results by different association rules mining models and verifies the validity and rationality of weighted theory.

The paper also innovatively brings forward a method of Temporal Constrained Quantitative Association Rules Mining and uses it into the validity verifying of the

technical analysis indexes in stock market. The paper chooses the Relative Strength Index (RSI), collects trade data to do demonstration analysis and mines out several useful quantitative association rules which can guide technical analysis application and actual operation in investment.

Keywords: temporal data mining; association rules; technical analysis

目录

Catalogue

第一章 绪论

1．1 选题的研究背景与意义

数据挖掘又称知识发现，是从数据库或数据仓库中识别出有效的、新颖的、潜在有用的以及最终可理解的模式的非平凡过程[1]。知识的表现形式可以是描述数据属性的规则、频繁发生的模式、数据库中的对象分类等。由于数据库技术和计算机存储技术的迅速发展，数据存储变得越来越容易，数据存储量也越来越大，而与之伴随的数据分析却相对滞后。为发现隐藏在这些数据背后大量有用的知识，研究各种类型的数据挖掘技术成为当前科学研究领域的热门课题。

由于带有时态约束的时间序列数据是一类常见而重要的数据，自然受到数据挖掘研究者的广泛关注，并成为数据挖掘研究的一个重要分支——时态数据挖掘。尽管相对于数据挖掘较成熟部分而言，时态数据挖掘研究是较新的一个方向，但其在工程、医疗和金融等领域都已有很多重要的应用。

关联规则是目前数据挖掘领域中研究得最为广泛的课题。这一方面是因为挖掘素材的广泛性和可获得性，各种销售记录、股票交易记录以及气象记录等都可以成为关联规则数据挖掘的对象；另一方面是因为关联规则价值判断的直接性，挖掘出的关联规则是否有兴趣通过分析很容易判断，而且也容易获得实际应用。传统的关联规则通常挖掘的是不考虑时间约束的规则，如购物篮分析等。在时态数据库规模不断壮大和重要性不断加强的前提下，如何将关联规则挖掘应用到时态数据挖掘中以获得有价值的时态关联规则是一个非常值得研究的课题。

时态关联规则的实用性是显而易见的，例如在金融行业，金融机构的许多业务活动（如 CRM、投资决策、风险管理、价格预测等）都越来越依赖于对大量历史数据的分析。投资者与金融机构也越来越清楚地认识到分析金融数据、从中挖掘出有价值的信息是其实现科学化管理决策的必要手段与“ 基础核心 ”工作。证券市场是最为活跃的金融市场之一，其交易系统中记录了大量交易明细数据，如各种价格、成交量、持仓量、收益率等以时间序列的形式记录并保

存的数据。因此交易数据库中蕴含了金融系统许多客观规律信息。时态关联规则能从中“挖掘”出各种知识，更好地认识、掌握、并利用其规律，这无疑对金融投融资决策与客户关系管理等活动具有特别重要的意义。

1．2 数据挖掘理论概述

1．2．1 知识发现与数据挖掘

在过去数十年中，随着数据库、互联网技术的迅速发展以及管理信息系统(MIS)和网络数据中心(IDC)的推广应用，人类产生和收集数据的能力已经迅速提高，数据的存取、查询、描述统计等技术己日臻完善，但高层次的决策分析、知识发现等实用技术还很不成熟，导致了“信息爆炸”但“知识贫乏”的现象。需求是发展之母，数据库管理系统和人工智能中机器学习两种技术的发展与结合，促成了在数据库中发现知识（KDD）这一新技术的诞生。

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，首次提出知识发现（Knowledge Discovery in Database），简称 KDD。它是一门涉及机器学习、模式识别、统计学、智能数据库、知识获取、数据可视化和专家系统等多个领域的交叉性学科。KDD 的一个比较公认定义是：KDD 是从大量数据中提取出有效的、新颖的、可信的并能被人理解的模式的高级处理过程。KDD 作为一个高级处理过程，整个过程包括在指定的数据库中使用数据挖掘技术提取模型，以及围绕数据挖掘进行预处理和结果表达等一系列的计算步骤[2]（见图 1.1）。

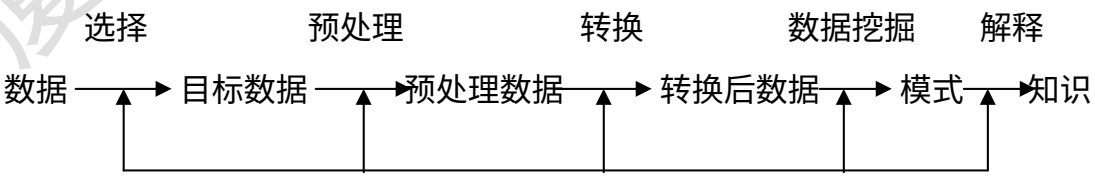

图 1.1：KDD 过程

由于 KDD 内涵极为广泛，理论和技术难度很大，从而使针对大型数据库的 KDD 技术一时还难以满足应用需要。于是，1995 年的(美)计算机学会(ACM)

会议提出了数据挖掘概念[3]，它形象地把大型数据库看成是存放有价值信息的矿藏，通过有效的知识发现技术，从中挖掘或开采出有用信息。由于数据挖掘是知识发现过程的关键步骤，因此，许多人不加区分地使用知识发现和数据挖掘这两个术语。数据挖掘 DM（Data Mining），又译作数据开采、数据采掘。它的定义有很多种，一种比较公认的数据挖掘定义是 W.J. Frawley、G. Piatetsky-Shapiro 等人提出的：数据挖掘就是从数据库中抽取隐含的、以前未知的、具有潜在应用价值的信息或知识的过程；提取的信息或知识表示为概念(concepts) 、规则 (rules) 、规律 (regularities) 、模式 (patterns) 等形式。

从经营管理角度出发，进入 21 世纪以后，全球经济一体化的进程日益加快，企业所面临的市场竞争压力日趋严重，经营管理者特别是决策者希望能够从企业积累的大量历史数据中找到经营管理中存在的根本问题，并从大量数据中快速挖掘出对经营管理有用的信息，以应对瞬息万变的市场竞争压力。通过数据挖掘工具可以从海量的原始数据中抽取出各种规则与知识，从而避免决策者陷入浩瀚的数据海洋中。因此数据挖掘技术是一个对管理决策者提供决策支持的有力工具。目前，在很多行业，尤其是在如银行、电信、保险、交通、零售等商业领域，数据挖掘技术得到了广泛的应用。它能解决的典型商业问题包括：数据库营销、客户群体划分、背景分析、交叉销售等市场分析行为，以及客户流失性分析、客户信用记分、欺诈发现等等。

1．2．2 数据挖掘过程和基本思想

数据挖掘不但能够学习己有的知识，而且能够发现未知的知识，得到的知识是“显式”的，既能为人所理解，又便于存储和应用，因此一出现就得到广泛的重视。对原始的业务数据进行挖掘，一般要经过以下几个步骤[4]：

(1)数据净化——消除数据库中的干扰数据和不一致数据(即噪声数据)；

(2)数据集成——对企业内不同部门的多数据源进行综合；

(3)数据聚焦——从数据库中检索出与分析任务相关的数据；

(4)数据转换——对数据条目进行分割或组合以满足数据挖掘的要求；

(5)模式发现——运用各种分析方法抽取出数据模式或规则(即知识)；

(6)知识呈现——将挖掘出来的模式与规则以图形化方式呈现给用户。

数据挖掘是从数据中识别出有效的、新颖的、潜在有用的、以及最终可理解的模式的高级过程。从上述定义中我们可以了解数据挖掘的基本思想：

(1)数据：是指一个有关事实 F 的集合（如股票信息数据库中有关个股的成交记录），它是用来描述事物有关方面的信息，是我们进一步发现知识的原材料。

(2)新颖：经过数据挖掘提取出的模式必须是新颖的，至少对系统来说应该如此。模式是否新颖可以通过两个途径来衡量：其一是得到的数据，通过当前得到的数据和以前的数据或期望得到的数据之间的比较来判断该模式的新颖程度；其二是通过其内部所包含的知识，通过对比发现的模式与已有的模式的关系来判断。通常我们可以用一个函数来表示模式的新颖程度 $N(E,F)$，该函数的返回值是逻辑值，是对模式 E 新颖程度的一个判断数值。

(3)潜在有用：提取出的模式应该是有意义的，这可以通过某些函数的值来衡量。用 u 表示模式 E 的有作用程度，$u=U(E,F)$。

(4)可被人理解：数据挖掘的一个目标就是将数据中隐含的模式以容易被人理解的形式表现出来，从而帮助人们更好地了解数据中所包含的信息。数据挖掘不同于以往知识获取技术的一个特点是发现的知识是人们（至少是领域专家）可以理解的，如“ If ...then... ”的形式，因此挖掘过程也是一个人机交互、螺旋上升的过程。

(5)模式：对于集合 F 中的数据，可以用语言 L 来描述其中数据的特性。表达式 $E \in L$，E 所描述的数据是集合 F 的一个子集 F_E。只有当表达式 E 比列举的所有 F_E 中元素的描述方法更为简单时，我们才可称之为模式。如：“ 如果股票涨幅在 5%~10%之间，则认为涨幅很大 ”可称为一个模式，而“ 如果涨幅为 5%, 6%, 7%,8%, 9%, 10%，则认为涨幅很大 ”就不能称之为一个模式。

(6)高级过程：数据挖掘是对数据进行更深层处理的过程，而不是仅仅对数据进行加减求和等简单运算或查询，因此说它是一个高级过程。

1．2．3 数据挖掘的任务和技术

数据挖掘不仅能对过去的数据进行查询和遍历，并且能够对将来的趋势和行为进行预测，还能发掘以前未发现的模式、知识，从而很好地支持人们的决

策。被挖掘出来的信息，能够用于信息管理、查询处理、决策支持、过程控制以及许多其它应用[5]。数据挖掘的任务主要包括以下几个方面：

1．关联分析。 关联分析是寻找数据库中值间的相关性，即若两个或多个数据项的取值重复出现且概率很高时，它们就存在着某种关联，可以建立起这些数据项之间的关联规则。在大型数据库中，这种关联规则是很多的，一般用“支持度”和“置信度”两个阈值来淘汰那些无用的规则。

2．分类。 分类要解决的问题是为一个事件或对象归类，是数据挖掘中应用最多的方法。分类是找出一个类别的概念或内涵描述，它代表了这类数据的整体信息，一般用规则或决策树模式表示。一个类的内涵描述分为特征性描述和区别性描述。特征性描述是对类中对象共同特征的描述，区别性描述是对两个或多个类之间区别的描述。在使用上，既可以用此模型分析已有的数据，也可以用它来预测未来的数据。

3．聚类。 聚类是把整个数据库分成不同的群组。它的目的是要使群与群之间的差别很明显，而同一个群之间的数据则尽量相似。聚类增强了人们对客观现实的认识，即通过聚类建立宏观概念。与分类不同，在开始聚类之前我们并不知道要把数据分成几组，也不知道怎么分（依照哪几个变量）。因此在聚类之后要有一个对业务很熟悉的人来解释这样分群的意义。很多时候还需要根据实际需要删除或增加变量以影响分群的方式，经过几次反复之后再最终得到一个理想的结果。

4．时序模式。 通过时间序列搜索出重复发生概率较高的模式。与关联分析相似，其目的也是为了挖掘数据项之间的联系，但序列分析的侧重点在于分析数据项之间在发生时间上的前后关系，这里强调时间序列的影响。例如，在所有购买激光打印机的人中，半年后有 80%的人再购买新硒鼓，20%的人用旧硒鼓装碳粉。序列规则也可记为 $A \Rightarrow B$，表示 A 发生以后将会发生 B。

5．偏差检测。 数据库中的数据常有一些异常记录，从数据库中检测出这些偏差很有意义。偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。偏差检测的基本方法是寻找观测结果与参照之间的差别。

6．预测。 预测是利用历史数据找出变化规律，即建立模型，并用此模型

来预测未来数据的种类、特征等。

数据挖掘作为一门交叉学科，受多个学科的影响，包括数据库系统、统计学、机器学习、可视化和信息科学等。因此数据挖掘研究产生了大量的、各种不同类型的数据挖掘方法和技术。数据挖掘的常用技术有以下几种：

1．关联规则：寻找数据库属性值间的相关性，即寻找在同一个事件中出现的不同数据项之间的相关性[6]，它是形式如下的一种规则，“ 在购买面包和黄油的顾客中，有 90%的人同时也买了牛奶 ”。

2．人工神经网络：它从结构上模仿生物神经网络，是一种通过训练来学习的非线性预测模型。可以完成分类、聚类和特征挖掘等任务。

3．决策树：用树型结构来表示决策集合。这些决策集合通过对数据集的分类产生规则，典型的决策树方法有分类回归树 CART、C4.5 等，其典型应用为分类规则的挖掘。

4．遗传算法[7]：是一种新的优化技术，基于生物进化概念设计了一系列过程来达到优化的目的。这些过程有基因组合、交叉、变异和自然选择等。遗传算法易于并行计算，并且已经应用于分类和其他优化问题。

5．粗糙集理论[8]：它是一种研究不确定性问题的数学工具，作为集合论的扩展，主要用于研究不完全和不完整信息描述的数据挖掘技术。可以用于分类，进行特征归约和最小属性子集归约。

6．模糊逻辑：通过隶属度函数定义分类系统的“ 模糊 ”阈值或边界，从而可以产生人们易于理解的分类规则。

7．最近邻技术：通过 K 个与之相近的历史记录的组合来辨别新的记录，也称为 K-最近邻技术。主要应用于分类、聚类和偏差分析等。

8．可视化：采用直观的图形方式将信息模式、数据的关联或趋势呈现给决策者，决策者可以通过可视化技术交互式地分析数据关系。

1．3 时态数据挖掘研究综述

在现实生活中，大量数据集都带有显式或隐式的时间特性，遍及经济、气象、通信、医疗等多个领域。随着数据库和计算机网络的广泛应用，加上先进

的自动数据生成和采集工具的使用，计算机中存储的数据量急剧增大。时态数据库就是由随时间变化的一系列序列值或事件组成的数据库。由于时间序列是常见且应用十分广泛的数据表征形式，时态数据库也自然成为最为重要的数据库之一，因而关于时态数据的挖掘方法也成为数据挖掘研究的一个热点。

相对于数据挖掘较成熟的部分而言，时态数据挖掘的研究是数据挖掘较新的一个方向。目前，国际上对于时态数据挖掘的研究逐渐成为一个新的热点，研究的重点主要集中在挖掘算法分布式和并行计算方面；国内在这方面的研究文献还主要集中在时态数据挖掘的理论框架上，比较重要的工作有 1998 年欧阳为民等从理论框架的角度对时态数据挖掘做过的介绍和分析[9]。本节结合国内外研究文献中时态数据挖掘方面的进展情况，对其研究内容进行归纳总结和分析，并对其主要分支加以讨论。

1．3．1 时态数据的定义

很多时态数据库中的数据项是按事件发生的先后顺序记录的。如商场的交易数据库，每位客户消费后数据库中都会记录交易发生的时间、购买商品名称、数量等信息。在传统的时态关联挖掘中，事务项的时间属性通常被忽略，而专注于挖掘布尔属性项之间的关联关系。事实上，在挖掘属性间关联规则时加上其固有的时态约束，能发现许多新的有价值的规则。

有些时态数据库中记录的数据是连续型的时间序列数值数据。时间序列是指随时间变化的序列值或事件，时态数据库是指由随时间变化的序列值或事件组成的数据库。这些值或事件通常是在等时间间隔测得的。以数学方式表达如下：

定义 1.3.1.1：一个时态数据库是指包含一系列记录$\{r_j\}_{j=1}^{N}$的数据库，N 为序列值的个数，其中每个记录为$m+1$维数据，即$r_j=\{a_1,a_2,\cdots,a_m,t_j\}$，$a_i$为特性值，可以是连续型数据也可以是离散型数据，而且可以是与时间有关联也可以没有。如果某特性值与时间有关，则该特性值为动态特性，否则为静态特性。传统关联规则挖掘对象是离散的布尔型动态特性，一般时间序列分析研究主要针对的是连续型动态特性。t_j是一个时间间隔的标志，例如天、月、年等。

以商场的单条交易记录说明布尔型动态特性：

$r_j = \{457838, 张三, 面包, 牛奶, 洗发水, 面巾纸, 20080218\}$

其中 457838 是客户代码，张三为客户姓名，接下来四个特性值为客户购买物品，最后是交易时间。其中的客户代码和姓名为静态特性，其余为动态特性。

以股票每日交易记录为例说明连续型动态特性：

$r_j = \{600036, 招商银行, 43.15, 43.4, 41.57, 41.81.367365, 156287.2, 20080115\}$

其中 600036 是股票代码，招商银行是股票名称，接下来分别是当日开盘价，最高价，最低价，收盘价，成交量，成交金额，20080115 代表所截取的研究日期为 2008 年 1 月 15 日。很显然前两个特性是与时间无关的，为静态特性，而其他特性值是与时间密切相关的，是动态特性。一般来说，对于静态特性研究的意义不大，我们通常挖掘的是记录中的动态特性。

定义 1.3.1.2：对于定义 1.1 中的动态连续特性值 a_i 可以定义为特性函数 f_i，其 f_i 是时间的函数，函数的系数可以从特性值 a_i 中得到，其函数表达为

$$f_i(t_x) = a_i \in r_j, 其中 \ni t_x \in t_j$$

当 a_i 连续时，该函数的系数通常可以通过经典的时间序列模型计算得到。基本的时间序列模型包括自回归移动平均 ARMA 模型和自回归条件异方差 ARCH 模型以及它们的扩展模型 ARIMA 和 GARCH 等等，这些模型通常用于短期预测。

由于实际应用中时间序列数据具有不规则、混沌等非线性特征。使得预测系统未来的全部行为几乎不可能，对系统行为的精确预测效果也难以令人满意。传统的时间序列分析方法的参数求解都基于坚实的数学基础，并要求很严格的假设，如果假设不合理，那么模型法将会严重失真。例如很多金融计量模型，常常基于平稳性假设、正态分布假设、线性假设等，但实际上金融时间序列具有信噪比低、非平稳、非正态、非线性的特点。另外，模型的构建也存在困难。

时态数据挖掘不同于传统的时间序列模型之处在于它是基于归纳推理的思维，直接以数据为驱动，因而它常常可以撇开一些假设条件。但它也需要面对一个很大的困难就是如何表达挖掘出来的知识，另外它对数据样本的数量和质量要求也比较高。