学校编码：10384　　分类号＿＿＿＿密级＿＿＿＿
学　　号：X200343030　　UDC＿＿＿＿＿＿＿＿＿＿＿＿

硕 士 学 位 论 文

时序关联规则挖掘研究

Research in Temporal Association Rules Mining

崔 晓 军

指导教师姓名：　薛永生　教授
申请学位级别：　工　学　硕　士
专 业 名 称：　计算机应用技术
论文提交时间：　2006 年 9 月
论文答辩时间：　2006 年 10 月
学位授予单位：　厦 门 大 学
学位授予日期：　2006 年　12 月

答辩委员会主席：＿＿＿＿＿＿
评　　阅　　人：＿＿＿＿＿＿

2006 年 9 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

数据挖掘又称数据库中的知识发现，是数据库研究最活跃的领域之一，这门技术自兴起以来因其广阔的应用前景和深远的现实意义受到学术界的广泛关注，而其中的关联规则挖掘问题，因其丰硕的研究成果和自身理论的逐渐成熟，正在形成一个比较完善的研究体系并带动整个数据挖掘技术快速发展。

自从Agrawal等学者于1993年首先提出了关联规则挖掘问题以来，诸多的研究人员对关联规则挖掘问题进行了大量的研究，提出了很多高效的算法，然而大多数方法都未考虑时间因素的影响。但在现实世界中，时间是数据本身固有的因素，在数据中常常会发现时序语义问题。时序数据的出现使得有必要在数据挖掘中考虑时间因素，在现实中，附加上某种时序约束的规则将可以更好地描述客观现实情况，因而也会更有价值，称这样的规则为时序关联规则。

现阶段对时序关联规则的挖掘主要集中在周期性关联规则挖掘、循环关联规则挖掘和日历关联规则挖掘几个方面。由于周期性和循环模式是建立在单一的时间粒度上，而日历模式是建立在多时间粒度上，这与实际生活中的年、月、日，时、分、秒等多粒度时间表示更加吻合，因此基于日历的时序关联规则挖掘研究更有实用价值。

本文主要研究基于日历的时序关联规则挖掘。首先在查阅国内外大量文献资料的基础上，对数据挖掘技术和关联规则挖掘技术进行了概述，对关联规则挖掘的典型算法进行了分析，并对时序关联规则挖掘的概化算法进行了描述。然后基于日历代数，提出了一种基于日历的时序关联规则挖掘算法BCTAR，该算法旨在发现基于给定的日历格的所有的时序关联规则，即发现所有的频繁项集和日历模式的匹配，算法只需扫描数据库两次。另一方面，基于模糊日历代数，提出了一种模糊时序关联规则的挖掘算法BFCTAR，该算法旨在发现用户指定的复杂日历下所有的频繁项集。实验结果分析说明，这两个算法是高效、实用的。

关键词： 数据挖掘;关联规则;时序;日历格;模糊日历代数

Abstract

Data mining, also known as knowledge discovery in database(KDD),is one of the most active fields in database. After existing, because of its wide application background and realistic significance, this technology has been drawing upon the attention of academic circle. Association rules mining is one of the important research aspects of data mining. Because it has rich research fruits and its theories become gradually mature,association rules mining is forming a perfect system and facilitate the development of the data mining on the whole.

After association rule mining proposed by Agrawal R,etc., which has received an extensive research and is one of more enrich and more active branches comparatively in achievements in data mining,but many mining techniques don't take the time factor into account.However time is the inherent attribute of data,so we should take time into account when mining association rules. In real world,rules with time restriction are more useful,we called this temporal association rules.

Research on mining temporal association rules have been focused on three forms, periodical association rules ,cyclic association rules, and calendric association rules. While periodical and cyclic patterns are basically in terms of a single time granularity, calendar patterns are based on a framework with multiple time granularities. Human activities are usually related to time granularities, e.g., months,days, and hours. Therefore, system support and reasoning involving calendars with multiple granularities have been recognized to be an important issue recently.

This thesis emphasizes on temporal association rules mining.At first,the background knowledge of data mining and association rules mining are introduced briefly. After that, some typical algorithms of association rule mining are discussed respectively and an general algorithm of temporal association rules is proposed.Furthermore,based on the theory of calendar algebra,a more efficient algorithm(BCTAR) to discover all calendar-based temporal association rules is studied. This method only scan database twice to find all frequent itemsets along with their frequent calendar patterns.Then with the study of fuzzy calendar algebra ,this thesis presents a algorithm to mining fuzzy temporal association rules,this algorithm is used to find all frequent itemsets with user defined complex calendar. Experiments proved that these algorithms are effective.

Keywords：Data mining;Association rules;Temporal;Calendar lattice; fuzzy calendar algebra

目　录

Contents

第一章　绪论

本章首先概述了数据挖掘技术的由来、市场现状、未来研究方向及热点，然后对论文的选题依据进行阐述，最后介绍了本文的主要内容及组织。

1.1 论文的研究背景

1.1.1 数据挖掘技术的由来

近十几年来，随着计算机软、硬件的飞速发展，人们利用信息技术产生和搜集数据的能力大幅度提高，数以千万计的数据库被用于商业管理、政府办公、科学研究和工程开发等方面。特别在一些领域采用集中或者分布式数据库存储技术，比如：

- 金融投资:股票指数和价格、利息率、信用卡数据、欺诈检验等；
- 卫生保健:医院管理系统保存了一些诊断信息等；
- 制造生产:过程优化、故障检测、DCS实时数据库等；
- 通信网络:呼叫模式、拥塞分析等；
- 科学领域:天文观察、基因数据等；
- 万维网(WWW)等。

收集工具的进步使人们拥有了数量庞大的数据。面对这些数据，急需一些新的工具和技术，能够智能化且自动地把这些数据转化为有用的信息和知识，从而解决“信息爆炸”所带来的问题--数据丰富，信息贫乏[1]。过去对于数据的分析主要依赖分析员来进行，从而对数据的分析工作也就变成了简单的根据专家知识从数据库进行查询和获取数据，并呈现给分析人员做出决策。这种对收集数据进行传统的数理统计和数据管理工具进行的分析不再适用。如何从海量数据中及时发现有用的知识，提高信息利用率，并将这些有用的信息和知识运用到实际工作中去成为一个迫切需要解决的问题。因此，数据挖掘(Data Mining，DM)越来越受到人们的重视[2]。各项技术的发展也激发了数据挖掘的开发、应用和研究的兴趣，Friedman[3]列举了数据挖掘发展的四个主要的技术理由:

- 超大规模数据库的出现，如商业数据仓库和计算机自动收集数据记录；
- 先进的计算机技术，例如更快、更大的计算能力和并行体系结构；

- 对巨大量数据的快速访问；
- 对这些数据应用精深的统计方法计算的能力。

因此，数据挖掘和知识发现可以说是数据库技术与信息技术发展的一个必然趋势，当人们不再为获取数据而烦恼时，如何分析、理解并利用这些数据就成为必然的要求。

1.1.2 数据挖掘的市场现状及发展前景

在国外，数据挖掘已经有不少成功案例。尽管数据挖掘的好处已经引起国内许多企业的重视，但实施的并不多，更多的企业是在观望和考虑。

目前国内企业实现数据挖掘的困难在于缺少数据积累、难于构建业务模型、各类人员之间的沟通存在障碍、缺少有经验的实施者、初期资金投入较大。而在国外，数据挖掘首先在金融、证券、电信、零售业等数据密集型行业实施，因为这些行业信息化程度比较高，数据库中已经保留了大量数据资源。

目前提供数据挖掘产品的厂商非常多，如著名的产品有SAS Enterprise Miner、NCR Teradata Warehouse Miner、SPSS Clementine 7.0、IBM DB2 Intelligent Mine、SQL Server 2000数据挖掘组件、Oracle9i Data Mining、CA CleverPath Predictive Analysis Server、德门软件DMiner等。这些产品各有特色：NCR、IBM、ORACLE等数据挖掘工具可以直接在数据库上进行挖掘；SAS提供了数据获取、取样、筛选、转换工具来构造要挖掘的数据集；SPSS针对具体应用领域推出了多个应用模版，以简化应用开发过程。

相关数据表明，二十世纪90年代以来，人类积累的数据量以每月高于15%的速度增加，如果不借助强有力的挖掘工具，仅依靠人的能力来理解这些数据是不可能的。

数据挖掘的前景被人们普遍看好。国际知名调查机构Gartner Group在高级技术调查报告中，将数据挖掘和人工智能列为“未来三到五年内将对工业产生深远影响的五大关键技术”之首，还将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术前两位。Gartner的调查报告预计：到2010年，数据挖掘在相关市场的应用将从目前少于5%增加到超过80%。美国银行家协会预测数据仓库和数据挖掘技术在美国商业银行的应用增长率是14.9%

1.1.3 数据挖掘的未来研究方向及热点

1.数据挖掘未来研究方向

当前，DM研究方兴未艾，其研究与开发的总体水平相当于数据库技术在70年代所处的地位，迫切需要类似于关系模式、DBMS系统和SQL查询语言等理论和方法的指导，才能使DM的应用得以普遍推广。预计在本世纪，DM的研究还会形成更大的高潮，研究焦点可能会集中到以下几个方面：

（1）发现语言的形式化描述，即研究专门用于知识发现的数据挖掘语言，也许会像SQL语言一样走向形式化和标准化；

（2）寻求数据挖掘过程中的可视化方法，使知识发现的过程能够被用户理解，也便于在知识发现的过程中进行人机交互；

（3）研究在网络环境下的数据挖掘技术（Web Mining），特别是在因特网上建立DM服务器，并且与数据库服务器配合，实现Web Mining；

（4）加强对各种非结构化数据的开采（Data Mining for Audio&Video），如对文本数据、图形数据、视频图像数据、声音数据乃至综合多媒体数据的挖掘；

处理的数据将会涉及到更多的数据类型，这些数据类型或者比较复杂，或者是结构比较独特。为了处理这些复杂的数据，就需要一些新的和更好的分析、建立模型的方法，同时还会涉及到为处理这些复杂或独特数据所做的费时和复杂数据准备的一些工具和软件。

（5）交互式发现；

（6）知识的维护更新。

但是，不管怎样，需求牵引与市场推动是永恒的，DM将首先满足信息时代用户的急需，大量的基于DM的决策支持软件产品将会问世。

只有从数据中有效地提取信息，从信息中及时地发现知识，才能为人类的思维决策和战略发展服务。也只有到那时，数据才能够真正成为与物质、能源相媲美的资源，信息时代才会真正到来。

2. 数据挖掘热点

就目前来看，将来的几个热点包括网站的数据挖掘（Web site data mining）、生物信息或基因（Bioinformatics/genomics）的数据挖掘及文本数据挖掘（Textual mining）。下面就这几个方面加以简单介绍。

（1） 网站的数据挖掘（Web site data mining）

随着Web技术的发展，各类电子商务网站风起云涌，建立起一个电子商务网站并不困难，困难的是如何让电子商务网站有效益。要想有效益就必须吸引客户，增加能带来效益的客户忠诚度。电子商务业务的竞争比传统的业务竞争更加激烈，原因有很多方面，其中一个因素是客户从一个电子商务网站转换到竞争对手那边，只需点击几下鼠标即可。网站的内容和层次、用词、标题、奖励方案、服务等任何一个地方都有可能成为吸引客户、同时也可能成为失去客户的因素。而同时电子商务网站每天都可能有上百万次的在线交易，生成大量的记录文件（Logfiles）和登记表，如何对这些数据进行分析和挖掘，充分了解客户的喜好、购买模式，甚至是客户一时的冲动，设计出满足于不同客户群体需要的个性化网站，进而增加其竞争力，几乎变得势在必行。若想在竞争中生存进而获胜，就要比您的竞争对手更了解客户。

在对网站进行数据挖掘时，所需要的数据主要来自于两个方面：一方面是客户的背景信息，此部分信息主要来自于客户的登记表；而另外一部分数据主要来自浏览者的点击流（Click-stream），此部分数据主要用于考察客户的行为表现。但有的时候，客户对自己的背景信息十分珍重，不肯把这部分信息填写在登记表上，这就会给数据分析和挖掘带来不便。在这种情况之下，就不得不从浏览者的表现数据中来推测客户的背景信息，进而再加以利用。

就分析和建立模型的技术和算法而言，网站的数据挖掘和原来的数据挖掘差别并不是特别大，很多方法和分析思想都可以运用。所不同的是网站的数据格式有很大一部分来自于点击流，和传统的数据库格式有区别。因而对电子商务网站进行数据挖掘所做的主要工作是数据准备。目前，有很多厂商正在致力于开发专门用于网站挖掘的软件。

（2） 生物信息或基因的数据挖掘

生物信息或基因数据挖掘则完全属于另外一个领域，在商业上很难讲有多大的价值，但对于人类却受益非浅。例如，基因的组合千变万化，得某种病的人的基因和正常人的基因到底差别多大？能否找出其中不同的地方，进而对其不同之处加以改变，使之成为正常基因？这都需要数据挖掘技术的支持。

对于生物信息或基因的数据挖掘和通常的数据挖掘相比，无论在数据的复杂程度、数据量还有分析和建立模型的算法而言，都要复杂得多。从分析算法上讲，更需要一些新的和好的算法。现在很多厂商正在致力于这方面的研究。但就技术和软件而言，还远没有达到成熟的地步。

（3） 文本数据挖掘（Textualmining）

人们很关心的另外一个话题是文本数据挖掘。例如，在客户服务中心，把同客户的谈话转化为文本数据，再对这些数据进行挖掘，进而了解客户对服务的满意程度和客户的需求以及客户之间的相互关系等信息。从这个例子可以看出，无论是在数据结构还是在分析处理方法方面，文本数据挖掘和前面谈到的数据挖掘相差很大。文本数据挖掘并不是一件容易的事情，尤其是在分析方法方面，还有很多需要研究的专题。目前市场上有一些类似的软件，但大部分方法只是把文本移来移去，或简单地计算一下某些词汇的出现频率，并没有真正的分析功能。

随着计算机计算能力的发展和业务复杂性的提高，数据的类型会越来越多、越来越复杂，数据挖掘将发挥出越来越大的作用。

1.2 论文的选题依据

关联规则挖掘用于发现大量数据中项集之间有趣的关联或相关联系，它在数据挖掘中是一个重要领域，最近已被业界所广泛研究。

Agrawal等于1993年首先提出了挖掘顾客交易数据库中项集间的关联规则问题[4]，以后诸多的研究人员对关联规则的挖掘问题进行了大量的研究。他们的工作包括对原有的算法进行优化，如引入随机采样、并行的思想等，以提高算法挖掘规则的效率；对关联规则的应用进行推广。

最近也有独立于Agrawal的频集方法的工作[37]，以避免频集方法的一些缺陷，探索挖掘关联规则的新方法。也有一些工作[6]注重于对挖掘到的模式的价值进行评估，他们提出的模型建议了一些值得考虑的研究方向。但这些传统方法均未考虑时间因素的影响。

然而在现实世界中，由于时间是数据本身固有的因素，例如超市交易记录中的交易时间，病历中的检查和诊断时间等，因此在数据中常常会发现时序语义问题。时序数据的出现使我们有必要在数据挖掘中考虑时间因素，在现实中，附加

上某种时序约束的规则将可以更好地描述客观现实情况，因而也会更有价值，称这样的规则为时序关联规则。

时序关联规则的经典例子有啤酒与尿布的故事，在对一个超市的事务分析中发现夏天的每个周末啤酒与尿布的销售量会大幅度上升。调查得知，许多年轻的爸爸们周末在为自己的宝宝买尿布的同时也不忘给自己买些啤酒。当时得出的一条关联规则是：啤酒⇒尿布，支持度是3%，置信度是87%，也就是说在所有事务数据中，有3%的顾客同时购买了啤酒与尿布，而在购买啤酒的顾客中有87%也购买了尿布。

下面再来对这些年轻爸爸们的购买行为进行仔细分析。

(1) 他们的购买行为多发生在周末，因此，如果把事务数据按照周末与非周末划分的话，就可能发现规则：啤酒⇒尿布的支持度在周末升到 10%以上。

(2) 众所周知，夏季是啤酒的销售旺季，因此，如果再把事务数据按照季节划分的话，就可能发现规则：啤酒⇒尿布的支持度在夏季的周末上升到 30%以上。

(3) 这个例子中年轻爸爸们的购买行为多发生在周末，因此，如果再把事务数据按照周末与非周末划分的话，就可能发现规则：啤酒⇒尿布的支持度在夏季周末的 6PM-9PM 上升到 50%以上。

(4)相反，如果当初把关联规则的支持度阈值定为 4%的话，这条经典的关联规则就可能被埋没掉了。

这样的例子还有很多，如圣诞火鸡、情人节巧克力等。这些商品都有自己的生命周期，没有必要存在于数据库的整个时间段内（如 1 年甚至更长）。但是传统的挖掘算法却可能找不到这些商品的关联规则，因为从其支持度的计算公式：$\sup port(X \Rightarrow Y) = \frac{|T : X \cup Y \subseteq T, T \in D|}{|D|}$

来看，规则 X⇒Y 在事务数据库 D 中的支持度是事务集中包含 X和 Y 的事务数与所有事务数之比，这里|D|是数据库中的所有事务数，不随商品的变化而变化，这对那些生命周期短的商品来说是不公平的，后果就是可能失去一些重要的规则。为弥补这一缺点，就必须给事务数据加上时序信息。

国内对于时序关联规则的研究较少，文献[7]尝试了将数据挖掘和时序数据相结合来讨论时序数据挖掘的有关问题，讨论了时间表达式的描述，并提出了有关时序模式和时序关联规则的若干定义；文献[8]讨论了某一时间段内的关联规则的有关概念及其有关性质，但二者都没有给出相应算法。文献[9]为关联规则模型增加了时序信息，即发现的关联规则包含该关联规则成立的时间范围，并给出相应的挖掘算法。但是关联规则的时间范围是按照支持该规则的交易时间逐步修正而确定的，其缺点是可能极少数(甚至个别)交易将规则的时间范围拉得太宽，使得规则的时序信息减弱。文献[10]充分考虑了元组（交易事务）存在的有效时间，提出了时间区间的延展与归并技术，结合 Apriori 算法，得到了一个新的能够处理具有时序约束的关联规则发现算法，是国内较早研究时序关联规则的文章之一。但是，该文中所考虑的时序语义较单一，仅考虑了时间区间的时序约束问题，对于呈周期性变化的时序约束问题尚不能解决。文献[11]从概化的角度提出了一个挖掘时序关联规则的算法。与已经存在的算法不同，该算法避开一些具体细节，使读者能从概化的角度对时序关联规则的挖掘有一个总体的了解。

在国外的研究中，文献[12]明确地阐述了不考虑时间维所带来的问题。文中假定数据库中的每个事务都有时间戳，并且由用户将数据划分为互不相交的时间区间，如年、月、日等。文献[13]提出周期性关联规则的挖掘，周期关联规则定义为在明确的有规律的时间区间内符合最小支持度和最小置信度的关联规则。应用这种定义后，一条规则在整个事务数据库中可能不具有高的支持度和置信度，但是可能在一个特定的周期性的时间区间数据内具有。其缺点是不能处理多粒度时间区间[14],例如，像“每个月的第一个工作日”这样的日历模式，周期关联规则就不能处理。文献[15][41]提出了循环关联规则挖掘，可以处理非周期性的规则，但同样不能处理多粒度时间区间。

文献[16]在有关工作的基础上介绍了日历代数的概念，用来描述关联规则中的一些感兴趣的现象。日历代数定义了发现关联规则的一套时间区间集，具体地说，一个日历 C 是一套时间区间的集合{(s1,e1), (s2,e2),…,(sk,ek)},其中，(si,ei)表示一个时间区间，si 表示开始时间，ei 表示结束时间。如果一个规则在一个日历包含的每一个时间单元内符合最小支持度和最小置信度的话，这样