学校编码：10384　　分类号_____密级_____
学号：24320071151840　　UDC _____

厦门大学

硕 士 学 位 论 文

增量式关联规则挖掘研究与应用

Research of Incremental Association Rules Mining and Applications

张 根 香

指导教师姓名：陈 海 山　教 授
专 业 名 称：计算机软件与理论

论文提交日期：2010 年 5 月
论文答辩日期：2010 年 月
学位授予日期：2010 年 月

答辩委员会主席：_____
评　　阅　　人：_____

2010 年 5 月

厦门大学学位论文原创性声明

本人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其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

随着信息科学、网络技术的发展，商业活动和工程实践中的数据量以惊人的速度在膨胀，数据分析与处理所面对的数据规模也随之急剧增长。关联规则挖掘作为数据挖掘分支之一，其任务是从大量的数据中发现信息，已经成为信息产业中热门的研究课题。本文正是对数据挖掘中的增量式关联规则提取算法及应用进行研究。

Apriori 算法是 Agrawal 提出的第一个经典关联规则挖掘算法，之后的大部分关联规则挖掘算法都是在此基础上不断地进行优化、改进，尽管这些算法各有优点，但在实际应用中都面临着这样两个问题：一是当数据频繁地增加或更新时，如何进行增量式规则提取？如果对于增加少量数据后的数据集重新运行一次算法，势必会造成效率的下降；二是当增量式关联规则涉及的数据量及数据维度都比较大时，若是完全基于频繁项、支持度-置信度框架进行处理，规则的质量难以得到保证。

基于以上两点的考虑，本文提出一种基于 IOGA 的增量式关联规则挖掘方法。该方法借助遗传算法的相关机理，同时模仿生物免疫系统中的自适应调节(免疫识别、免疫记忆、免疫调节)对遗传算法进行进一步优化，克服其存在的易陷入局部解、过早收敛现象。通过病毒文件检测实验表明，该方法应用于大规模数据集的增量式关联规则挖掘时，可以及时地感知规则的变化并发现有用的规则，减少了冗余规则的产生，同时挖掘效率也有明显提高。

关键字：遗传算法；免疫优化；增量式关联规则

Abstracts

As the Information Science and Network Technology develops, the data sets from business activities and engineering practice are expanding rapidly, data analysis and processing face large and large data scale. As an important subdomain of data mining, Association rules mining is to find hidden information from the data set, and it has caused more and more attention for its significance in theory and business.This paper is about incremental association rules mining algorithms the research and its applications.

Apriori algorithm is the first critical association rules mining algorithm raised by Agrawal,based which the later improved algorithms sprung up. Though all the algorithms has their own merits,they all face two problems: for one thing, when the data set is updated frequently, how can we mining the rules? If we run the algorithm repeatedly when new data added,the efficiency can absolutely be effected,for an other thing, for an other thing, when the data set is large and has many dimentions, if we mine rules totally based on frequent item set and support-confidence framwork, the rules' quality can't be promised.

Considering for the two points above, this paper proposes an IOGA approach for incremental association rules mining. The Algorithm follows genetic principles, and immitates the Self-Adaption in bionic immune system such as immune detection,immune memory and immune adjustment to overcome the shortcomings of local solutions falling and premature in Genetic Algorithm. Virus File Detection Experiment demonstrates the IOGA based incremental association rules mining's effectiveness and presents its good performance in perceiving rules' change, reducing redundant and finding interesting rules.

Key Words: Genetic Algorithm; Immune Optimization; Incremental Association Rules Mining.

目　录

Content

第一章 绪论

1.1 数据挖掘概述

1.1.1 数据挖掘产生的背景

随着计算机网络技术的飞速发展和社会各行业领域技术的完善，人类社会的信息化进程不断加快。人们在取得大量数据信息的同时也遭遇到前所未有的问题：面对海量的数据难以及时消化并从中提取有用的信息；数据表示形式错综复杂而难以统一处理；数据表面所携带的信息已经远远超越了人类的理解力、真假难辨使信息安全无法保证。海量的数据呈现出“数据丰富，但信息贫乏”现象，存储数据的数据库演变成了“数据坟墓”，尽管其中隐含着巨大的价值，但如何去开发这种价值，将未知变为可知，成为摆在眼前的一大难题，如某商务网站希望能够从已往的交易数据中发现客户的消费习惯和消费特征，从而预测潜在客户、刺激消费并创造更多的交叉销售的机会。再如银行的信贷机构，如何分析历史信贷纪录，以帮助贷款种类的设计、市场营销的策略及贷款的发放，从而降低不良贷款率，提高银行抗风险能力，是他们面临的一个重大难题。类似的需求体现在金融，零售，电信等社会的各行各业，不胜枚举。

人们希望通过对积累的数据进行更高层次的分析，找出数据内部潜在的关系和规则，这些必须借助于强有力的数据分析工具，这种工具首先应该能够收集快速增长的大量数据并对它们进行整理以特定的形式存放于数据库中，同时对数据进行去粗取精、去伪存真的处理，从浩如烟海的数据中提炼出有用的信息，帮助各种决策。 于是数据挖掘（Data Mining,DM）便在这样的背景应运而生。

人们最初使用的方法是通过机器学习实现自动决策支持[1]，把已知的并已被成功解决的问题作为训练集输入至计算机，机器在这些训练集的基础上进行学习，然后总结并产生相应的具有通用性的规则，以解某一领域的某类问题。

此后，随着神经网络技术的出现及发展[2]，人们逐渐把注意力从机器学习转向知识工程，知识工程不同于机器学习那样给计算机输入范例，让它生成规则，而是把规则代码化作为计算机的输入，通过这些规则解决某类问题，比如专家系

统[3]就是基于这种方法的成果，但它有投资大、效果不甚理想等诸多方面的不足。

20 世纪 80 年代，神经网络技术得到进一步的发展，人们开始在新神经网络的基础上又回到机器学习，且在大型商业及工程数据库中得到了应用[4]。到 20 世纪 80 年代末，知识发现（简称 KDD，Knowledge discovery in database）被人们普遍使用，以描述整个数据发掘的过程，即从起初业务目标的制定到最终结果的分析及决策[5]。1989 年 8 月，第十一届国际联合人工智能学术会议正式提出数据挖掘（DataMining, 简称 DM）这一概念，人们也把数据挖掘视为另一个常用术语 KDD, 也有人把数据挖掘看成是 KDD 中的部分，1995 年，第一届知识发现和数据挖掘国际学术会议举行，从此，每年一次的 KDD 国际学术会议把知识发现 KDD 和数据挖掘 DM 方面的研究不断推向前进，数据挖掘一词因此而开始“流行”[6]。

居于不同的应用需求，数据挖掘涉及到多方面的技术，因而在不同的领域得到相应的研究，如数据库技术方面、人工智能方面、数理统计方面、可视化技术、并行计算，还有人工免疫方面等，这些领域的学者和工程技术人员已形成一支新兴的队伍，投身到这数据挖掘不同的研究领域，使得数据挖掘成为了新的技术热点。

1.1.2 数据挖掘的过程

数据挖掘是从大量的、不完全的、有噪声的、模糊的数据中提取隐含在其中的、具潜在价值的信息和知识，完整的过程包括数据预处理（数据清理，数据集成，数据选择，数据变换）、数据挖掘、模式评估及知识表示[6]，如图 1.1 所示。

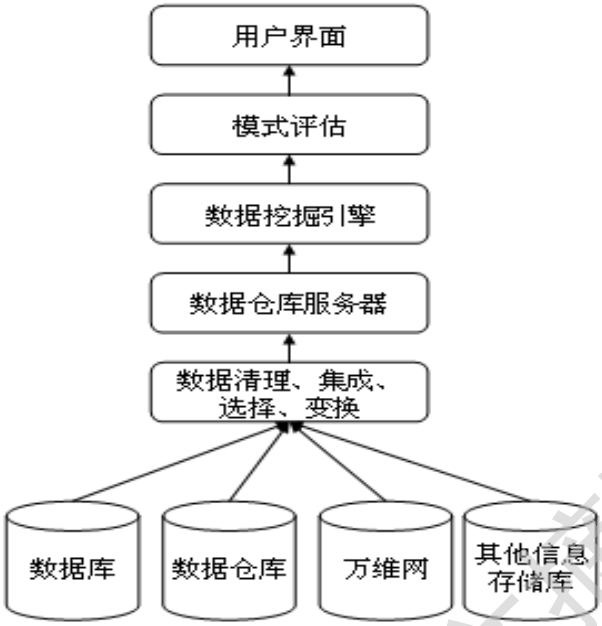

图 1.1：数据挖掘的系统结构

最底层的是数据存储库，不同的行业，不同的公司有不同的数据存储形式，如存储于不同的关系型数据库 Oracle 数据库、SQL Server 数据库、MySQL 数据库等；一些数据经过初步处理存入到数据仓库；此外，随着计算机网络技术的应用，万维网上每天都在产生海量的数据，且相当大一部分以文本形式进行存储，像最普遍的 Web 日志数据等。因此，数据挖掘面对的数据存储库囊括了关系型数据库、时间序列数据库、事务数据库、高级数据库系统、一般文件、数据流和万维网，其中高级数据库系统又包含对象关系数据库和面向特殊应用的数据库，如空间数据库、时间序列数据库、文本数据库和多媒体数据库等。

以上来源不同，表示形式不一致的数据无法对其进行直接的挖掘分析，首先要将其转化成统一的表示形式，补充缺失的数据，去除不必要的数据，这就涉及到数据清洗，集成，选择，变换，经由这些处理后由数据仓库负责提取相关数据。

数据挖掘引擎是整个系统的基本组成部分，也是整个系统的核心，通常由关联与相关分析，分类，聚类，离群点分析等功能模块组成，对于结构化数据集，如关系数据库，事务数据库或数据仓库中的数据，主要的处理方式包括关联规则提取，聚类，分类；对于非结构化或半结构化的数据如实时监控系统等动态环境产生的数据，则通常采用数据流挖掘、时间序列挖掘和序列数据挖掘[7]；随着计算机图形学，文本检索及 Web 分析的广泛发展及应用，图挖掘逐渐成为数据挖

掘领域的一项重要任务与研究课题；除了静态的图形图像，在数字文档，万维网，个人或专业数据库中还能获得以数字形式表示的视频和音频信息，且这类信息增长迅速，迫切需要针对视频音频的、有效的挖掘方法；在现实生活中文档数据库或文本数据库存储了大量的信息，对于这类非结构化的数据，传统的信息检索方式已经不适应日益增长的、大量文本数据处理的需要，因此，文本挖掘亦成为了数据挖掘当中的又一项重要作务[8]。

数据挖掘引擎会产生大量的规则与模式，对于不同的用户，在产生的模式中可能只有一小部分是他所感兴趣的，模式评估使用兴趣度量方式，通常与数据挖掘引擎集成，用于衡量及识别在特定的情况下哪些规则或模式是有用的或者是有趣的。

用户界面提供用户与系统之间的通信，允许用户进行相应的设置，说明挖掘的任务，提供信息以帮助搜索聚焦，还允许用户浏览数据库和数据仓库模式或数据结构，评估挖掘的模式，以不同的形式对模式进行可视化展现。

尽管市场上有许多“数据挖掘系统”，但并非所有的系统都可以进行真正的数据挖掘，很多只能称作机器学习系统[9]、统计数据分析工具或实验系统原型，只能够进行数据或信息检索，如在大型数据库中找出聚集值或回答演绎查询，实际上不能处理大量数据的数据分析。

1.2 国内外的研究现状

“在数据库中的知识发现（KDD:Knowledge Discovery in Database）” 这一术语是数据挖掘的前身[7]，在 1989 年 8 月十一届国际人工智能联合会议的专题研讨会上被首次提出，接下来的三年里又相继举行了 KDD 的专题讨论会，数据挖掘这一概念正式出现是在 1995 年在美国计算机年会（ACM）上。历经十多年的发展，数据挖掘越来越得到人们的普遍关注，同时各种数据挖掘软件也相继出现，极大地增强了人们对信息的掌握和处理能力，如今已成为具有广阔应用前景的热门研究方向之一[10]。

比较著名的数据挖掘软件有 IBM 公司的 Intelligent Miner V6[11]，它提供了一套分析数据库的挖掘过程、统计函数和查看、解释挖掘结果的可视化工具，是一种分别面向数据库和文本信息进行数据挖掘的软件系列，其目前的不足