学校编码：10384　　分类号_____密级_____
学　　号：x200431036　　UDC _____

厦　门　大　学

硕　士　学　位　论　文

ID3 算法的研究以及在成绩统计辅助决策系统中的应用

Research on ID3 algorithm and Application in Data Mining System of grade policy-making

张 凌

指导教师姓名：　罗　健 教授

江善贤 高工

专 业　名 称：　控　制　工　程

论文提交日期：　2007 年 4 月

论文答辩时间：　2007 年 7 月

学位授予日期：　2007 年 10 月

答辩委员会主席：_____

评　　阅　　人：_____

2007 年 4 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而产生的权利和责任。

声明人：

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

随着以计算机和网络为代表的信息技术的发展，越来越多的企业、政府组织、教育机构和科研单位实现了信息的数字化处理。数据库，特别是数据仓库已经被广泛地应用于企业管理、产品销售、科学计算和信息服务等领域，同时，信息量的不断增长也对数据的存储、管理和分析提出了更高的要求。剧增的数据中又有可能隐藏着许多重要的信息，人们希望能够对已经占有的信息更高层次的分析，以便更好地利用这些数据。目前的数据库系统虽然可以较好地实现数据的录入、查询和统计等功能，但尚不支持对数据背后重要信息的挖掘，从而导致了“数据爆炸，知识贫乏”的现象。

数据挖掘技术可以帮助人们从数据库，特别是数据仓库的相关数据集中提取出所感兴趣的知识、规律或更高层次的信息，并可以帮助人们从不同程度上去分析它们，从而可以更有效地利用数据库或数据仓库中的数据。数据挖掘技术不仅可以用于描述过去数据的发展过程，进一步还能够预测未来趋势。分类发现作为数据挖掘的一项重要任务，在科学实验，医疗诊断，气象预报，信贷审核，数据预测，案件侦破等领域有着广泛应用。

在高等院校的教务管理系统的成绩库中存储了大量的学生成绩，这其实是一个十分有用及丰富的数据库。本课题对 ID3 算法进行研究分析后，提出了一种改进的算法，对储存在教务管理系统中的数据进行挖掘分析。从而找出课程设置之间的关联，对高校的成绩统计辅助决策提供一些数据依据。

关键词：数据挖掘、ID3 算法、成绩统计辅助决策

ABSTRACT

With the development of information technology represented by computer and network, more and more companies, government organizations, academic institutions and research institutions realize the digitalization of information. The database, especially data warehouse has been widely used in management, distribution, scientific calculation and information services. At the same time, the increasing of information brings higher requirement for the storage, management and analysis of data. Much important information may hide in the rapidly increasing data, so that people hope to analysis the information in higher level in order to make better use of the data. Although the current database system can satisfy some functions such as entry, request and counting, it still does not support the excavating of some important information behind the data, which leads to the phenomenon of “exploding data, poor knowledge”.

The technology of data mining can help us find the knowledge, rules or information of high level from the related data in the Database , especially the Data Warehouse. Classified discovery is widely used in many fields such as scientific experiments, medical treatment, weather prediction, credit verification and case clearance.

Large numbers of data as students scores are stored in the college educational administration system. It is a helpful and abundant database indeed. This course will investigated and analysis the ID3 calculation method and raise an improved arithmetic which is able to deeply analysis the datas in this college educational administration system databse. Aim to find out the courses setting connection and propose the data base of students' scores which to support the college's decision making.

Keywords: data mining,ID3 algorithm, decision making support of grading

目录

content

第一章　绪论

1.1 课题研究的背景

近年来，随着数据库技术和计算机网络的广泛应用，数据挖掘技术已经吸收了许多学科的最新研究成果而形成独具特色的研究分支。数据的丰富带来了对强有力的数据分析工具的需求，大量数据被描述为“数据丰富，但信息贫乏”。海量数据存放在数据库中，没有强有力的工具就难以理解它们。这对数据挖掘技术的研究和应用就提出了很大的挑战。

现代教育事业的高速发展，使得各高等院校教务处始终把教师的教学效果和学生的学习效果的评价作为一项重要工作，所以，作为进行教学和学习效果分析评价的重要办法之一的成绩统计分析工作，也越来越受到教育工作者的关注。随着信息处理技术的发展，特别是校园网的建立，高校家务管理系统不断完善，教务系统中存储了大量的学生成绩历史数据，为数据挖掘和辅助决策提供了基础。

1.2 课题研究的意义

本文调研了了应用教广的本科教务管理系统的成绩管理子系统，各系统对成绩数据录入、信息查询、成绩数据汇总的功能使用得比较成熟。根据本文的调研，目前没有见到对成绩多角度挖掘的分析。近几年有关成绩统计与决策的文献，所涉及的研究主要集中在：探讨分数标准化，对学生综合成绩排序的比较，对教学班成绩正态分布的研究等方面。

针对现状，本文应用一种改进的 ID3 算法，结合教育信息的特点，对学生成绩和基本信息数据库进行挖掘分析，找出课程之间的相关性，并对挖掘结果进行比较分析，为学校的教务管理提供参考。

1.3 课题研究的创新

在目前的教学制订的过程中，制订课程先后顺序主要以教师经验为主要依据，这就缺乏相应的数据依据。而对于设置后的课程顺序是否合理，也缺少对其进行印证的方法。

本课题在对 ID3 算法进行改进之后应用到高校教务管理系统，尝试对大量储存在成绩库中的数据进行较深度的挖掘，找出各个课程之间设置的内在联系，为今后的教学计划的制订寻找有力的数据依据。

第二章　数据挖掘与知识发现

数据挖掘(Data Mining)是一个多学科交叉研究领域，它融合了数据库（Database）技术、人工智能（Artificial Intelligence）、机器学习（Machine Learning）、统计学（Statistics）、知识工程（Knowledge Engineering）、面向对象程序设计（Obiect-Oriented Method）、信息检索（Information Retriecval）、高性能计算（High-Performance）等最新技术的研究成果。经过十几年的研究，产生了许多新概念和方法。特别是最近几年，一些基本概念和方法趋于清晰，它的研究正向着更深入的方向发展。数据挖掘之所以被成为未来信息处理的骨干技术之一，主要在于它以一种全新的概念改变着人列利用数据的方式。二十世纪，数据库技术取得了决定性的成果并且已经得到了广泛的应用。但是，数据库技术作为一种基本的信息存储和管理方式，仍然以联机事务处理（OLTP:On-Line Transaction Processing）为核心应用，缺少对决策、分析、预测等高级功能的支持机制。众所周知，随着数据库容量的膨胀，特别是数据仓库（Data Warehouse）以及 Web 等新型数据源的日益普及，联机分析处理（OLAP:On-Line Analytic Processing）、决策支持（Decision Support）以及分类（Classing）[1]、聚类（Clustering）[2]等复杂应用成为必然。面对这一挑战，数据挖掘和知识发现（Knwuledge Discoversy）技术应运而生，并显示出强大的生命力。数据挖掘和知识发现使数据处理技术进入了一个更高级的阶段。它不仅能对过去的数据进行查询，并且能够找出过去数据之间的潜在联系，进行更高层次的分析，以便更好地做出理想的决策、预测未来的发展趋势。

从数据库中发现知识(Knowledge Discovery in Database，KDD)是 20 世纪 80 年代末开始的，KDD 一词是在 1989 年 8 月十美国底特律市合并的第一届 KDD 国际学术会议上正式形成的。KDD 研究的问题有：(1)定性知识和定量知识的发现；(2)知识发现方法；(3)知识发现的应用等[3]。

1995 年在加拿大召开了第一届知识发现和数据挖掘(Data Mining，DM)国际学术会议[4]。由于把数据库中的“数据”形象地比喻成矿床，“数据挖掘”一词很快流传开采。

数据挖掘是知识发现中的核心工作，主要研究发现知识的各种方法和技术。数据挖掘首先要确定挖掘的任务或目的。确定了挖掘任务后，就要决定使用的挖掘算法。选择实现算法有两个考虑因素：一是不同的数据有不同的特点，因此需要用与之相关的算法来挖掘，二是用户或实际运行系统的要求。选择了挖掘算法后，就可以实施数据挖掘操

作，获取有用的模式。

数据挖掘作为知识发现过程的一个特定步骤，它是一系列技术及应用，或者说是对大容量数据及数据间关系进行考察和建模的方法集。它的目标是将大容量数据转化为有用的知识和信息。

一般情况下，数据挖掘的对象定义为数据库，而更广义的说法是，数据挖掘意味着从一些实事或观察数据的集合中寻找模式。数据挖掘的对象不仅是数据库，也可以是文件系统或者其他任何组织在一起的数据集合，例如 Internet 信息资源、数据仓库等。数据挖掘广义定义：数据挖掘是从存放在数据库、数据仓库或其它信息库中的大量数据中挖掘有趣知识的过程[5]。

与数据挖掘和知识发现关系密切的研究领域包括归纳学习（Inductive Learning）、机器学习（Machine Learning）和统计（Statistics）分析，特别是机器学习被认为和数据挖掘的关系最密切[6][7]。

数据挖掘的技术基础就是人工智能，它利用了人工智能中的诸多算法进行挖掘，为用户提供有用信息[8]。在很大程度上，数据挖掘是人工智能的某些成熟的技术(人工神经网络、遗传算法、决策树)在特定的应用系统中具体而微小的应用，但是其问题的规模和难度大大降低。

除了人工智能之外，数据挖掘还结合了传统的统计分析、模糊数学以及科学计算可视化技术，以数据库和数据仓库为研究对象，形成了数据挖掘方法和技术[9][10][11]。

在现实生活中，数据挖掘技术被广泛应用，尤其是在决策支持系统中，常常利用数据挖掘技术从数据库系统中获得有用信息。让更高一级的用户根据挖掘结果做出更明智、更正确的决策。

2.1 数据挖掘的概念

2.1.1 数据挖掘的定义

数据挖掘（Data Mining）[12-15]就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程，与数据挖掘相近的同义词有数据融合、数据分析和决策支持等。这个定