

高等教育 评估决策支持系统

瞿斌 著



中国水利水电出版社
www.waterpub.com.cn

高等教育 评估决策支持系统

瞿斌 著



中国水利水电出版社
www.waterpub.com.cn

内 容 提 要

本书理论与实践相结合,系统地探讨了基于数据仓库的高等教育评估决策支持系统的基本理论,详细阐述了评估数据仓库的数据模型,全面分析了基于评估数据仓库与联机分析处理的技术选型,深入研究了基于评估数据仓库的数据挖掘算法,在此基础上,给出了基于数据仓库的评估决策支持系统的结构,并以全国优秀博士论文评选为例,进行了实证研究,证实该理论的可行性和有效性。

本书既可供管理科学与工程、计算机科学与工程、教育管理、系统工程、工商管理 etc 等学科领域的决策支持系统的学者、政策制定者和管理人员阅读,也可供高等院校相关专业师生参考。

图书在版编目(CIP)数据

高等教育评估决策支持系统 / 瞿斌著. — 北京 :
中国水利水电出版社, 2011.9
ISBN 978-7-5084-9037-3

I. ①高… II. ①瞿… III. ①高等教育—教育评估—
决策支持系统—研究—中国 IV. ①G649.2

中国版本图书馆CIP数据核字(2011)第197413号

书 名	高等教育评估决策支持系统
作 者	瞿斌 著
出版发行	中国水利水电出版社 (北京市海淀区玉渊潭南路1号D座 100038) 网址: www.waterpub.com.cn E-mail: sales@waterpub.com.cn 电话: (010) 68367658 (发行部)
经 售	北京科水图书销售中心(零售) 电话: (010) 88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排 版	中国水利水电出版社微机排版中心
印 刷	北京市兴怀印刷厂
规 格	145mm×210mm 32开本 6印张 161千字
版 次	2011年9月第1版 2011年9月第1次印刷
印 数	0001—1500册
定 价	28.00元

凡购买我社图书,如有缺页、倒页、脱页的,本社发行部负责调换

版权所有·侵权必究

前 言



高等教育评估是教育评估的重要组成部分，也是国内外普遍关注的前沿研究领域。本书以国家自然科学基金资助项目“研究生教育评估现代技术”这一科研课题为依托，主要从理论和实证的角度，研究如何利用数据仓库、联机分析处理和数据挖掘等信息技术，来构建高等教育评估决策支持系统，以期推动和促进教育评估理论和实践的不断创新。本书的主要研究工作如下。

1. 对基于评估数据仓库与联机分析处理进行了研究

给出了评估数据仓库的定义，指明了其数据特点，分析了其数据组织结构。进而设计了评估数据仓库的数据模型，其中包括确立了设计评估数据仓库的数据模型的指导思想，和设计了评估数据仓库的概念模型、逻辑模型和物理模型等，阐述了其实施过程中的若干关键技术——数据抽取、数据集成和数据刷新。同时，在此基础上，构建了基于评估数据仓库的联机分析处理的体系结构，选取了其相应的技术类型——混合型联机分析处理。另外，还研究了基于评估数据仓库的联机分析处理的模型——立方体，说明了其所支持的各种多维分析功能——切片、切块、旋转、下钻和上卷，等等。

2. 对基于评估数据仓库的数据挖掘进行了研究

分析了基于评估数据仓库的数据挖掘的运行过程，在此

基础上设计了其相应的框架结构,该框架结构具有很好的可扩展性。研究了概念树和分类树方法,并给出了具体的模型和算法。采用粗集方法,利用条件属性约简的方式,从更高层次和更一般意义去审视了指标变量之间的相互依赖关系。给出了算法 BestFeatures,该算法简单易懂,容易理解、操作和实施,具有很好的实用性。对原有的关联规则挖掘算法——Apriori 算法和 AprioriTid 算法进行了改进和完善,提出了新的算法——Tid - option 算法。该算法吸取了 Apriori 算法和 AprioriTid 算法的优点,并且只需要扫描表一次,可以更加快速有效地挖掘关联规则。

3. 对基于数据仓库的评估决策支持系统进行了研究

分析了基于数据仓库的评估决策支持系统开发的背景,明确了系统开发的目标,介绍了系统开发的方法开发要遵循为评估决策者服务这一原则。结合评估工作的实际情况,设计出了系统的逻辑结构和功能结构,说明了系统的运行过程。在该系统的逻辑结构中,模型库集模型与方法于一体,节约了系统的资源,提高了系统的可接受性,满足了评估决策者的实际需要。另外还从先进性、集成性和实用性三个方面出发,确定了系统开发的环境,阐明了系统实现的架构。

以优博评选为应用案例,进行了实证研究,在实践中检验了理论。根据各种数据标准,在优博评选数据集市的数据模型的基础上,经过处理实现了优博评选数据集市。设计了优博评选分析立方体。对优博评选分析立方体执行了联机分析处理,从多角度、多层次地观察和分析了优博评选的得分情况,更加全面和深入地认识了我国博士生教育中所存在的普遍现象和共性本质。应用分类方法,对优博评选中的入选论文和落选论文这两类评估事物进行了分析,得出了相应的分类模型及其准确度,并据此提出了各种改进措施。通过条

件属性约简的方式，考察了优博评选指标体系中各项指标之间的相互依赖关系，发现每项指标都不能舍弃，说明该指标体系具有很好的适应性。进而应用关联规则挖掘方法，实施了数据挖掘，获得了若干条关联规则，认识了优博评选中各项指标得分与总体定性评价得分之间的关系。通过实证研究，取得了令人满意的结果，表明理论能很好地应用于实践。

本书是一本高等教育评估方面的专著，对高等教育评估研究学者和专家有着重要的参考价值。当然，由于基于数据仓库的高等教育评估决策支持系统研究仍处于探索阶段，许多理论和观点是第一次提出，同时本人学术水平有限，很多建议尚不够成熟和完善，恳请各位学者和专家不吝赐教。

华北电力大学 瞿斌

2011年7月20日

目 录



前言

第一章 绪论	1
第一节 研究背景	1
第二节 研究意义	3
第三节 国内外研究现状综述	5
第四节 研究内容及其结构框架	31
小结	33
参考文献	33
第二章 评估数据仓库与联机分析处理	38
第一节 评估数据仓库的定义及其数据特点	38
第二节 评估数据仓库的数据组织结构	40
第三节 评估数据仓库的设计	43
第四节 评估数据仓库实施的关键技术	55
第五节 基于评估数据仓库的联机分析处理的体系结构	58
第六节 基于评估数据仓库的联机分析处理技术类型的选取	59
第七节 基于评估数据仓库的联机分析处理的模型——立方体	63
第八节 基于评估数据仓库的联机分析处理的功能	66
小结	70
参考文献	71
第三章 基于评估数据仓库的数据挖掘	74
第一节 基于评估数据仓库的数据挖掘的运行过程	74

第二节	基于评估数据仓库的数据挖掘框架结构	78
第三节	概念树方法研究	81
第四节	决策树方法研究	85
第五节	基于粗集的关联规则发现方法研究	88
小结		97
参考文献		97
第四章	基于数据仓库的评估决策支持系统	101
第一节	基于数据仓库的评估决策支持系统建设的背景与目标	101
第二节	基于数据仓库的评估决策支持系统的开发方法	104
第三节	基于数据仓库的评估决策支持系统的结构	106
第四节	基于数据仓库的评估决策支持系统的运行流程	114
第五节	基于数据仓库的评估决策支持系统的开发环境	116
小结		119
参考文献		120
第五章	基于数据仓库的评估决策支持系统应用探讨 (一)	
	——评估数据仓库的实现与联机分析处理的应用	122
第一节	评估数据仓库的实现	122
第二节	联机分析处理方法的应用	129
小结		160
第六章	基于数据仓库的评估决策支持系统应用探讨 (二)	
	——数据挖掘的应用	161
第一节	分类方法的应用	161
第二节	关联规则挖掘方法的应用	170
小结		180
第七章	结论	181

第一章 绪 论

第一节 研究背景

教育评估研究是按照一定的标准，对受教育者的发展变化以及构成变化的诸因素所进行的价值判断，已经成为当今世界教育科学研究的三大领域（即教育基础理论研究、教育发展研究和教育评估研究）之一^[1]。高等教育评估（以下简称评估）是教育评估的重要组成部分，它是以高等教育为对象，依据评估标准，利用可行的评估技术和手段，系统地收集评估信息，通过定量和定性分析，对评估客体所给予的价值判断^[2]。评估在加强高等教育的宏观管理，促进高等教育的改革和发展，检验、保障和提高高等教育质量，建立高等教育和社会、政治、经济、文化、科技相适应的内在机制等诸多方面都发挥着至关重要的作用。

要准确地判断评估对象的价值，就要实现评估的科学化、现代化。数据库技术是评估中常用和重要的工具之一，它促进了评估理论、方法和技术的创新与发展，也为评估理论研究和实践活动注入了新的活力。数据库技术使得大批量地处理评估数据能够更加准确、迅速；使得在数学建模过程中调用评估数据的方法能够更加容易实现；使得对评估数据、评估结果进行统计分析、提供各种报表的工作能够更加高效。另外，数据库技术还为数据输入和存储提供了极大的方便^[3]。随着数据库技术的应用和广大评估工作人员的不懈努力，评估信息管理工作取得了一批令人欣喜的成果，基于数据库的评估信息管理软件逐步建起来了，大量的评估数据逐渐积累起来了，信息共享在一定程度上和范围内实现了，工作效率和管理水平也提高了。同时培养和锻炼了一批既有业务知识又有一定计算机



水平的人才。这就从信息资源和人才队伍等方面为进一步提高评估信息管理和服务水平打下了良好的基础，也提供了必要的条件。

数据库技术丰富了评估理论，改进和创新了评估的方法和手段，提高了评估实践工作的效率，但是以数据库技术构建的信息系统功能有一定的局限性，它们基本上还停留于评估组织日常事务处理这一层面，也就是说，这些系统基本上都属于评估事务处理系统的范畴。随着外部环境的急剧变化和评估机构信息需求的不断增长，如何更加有效地管理评估组织内部日常事务处理过程中所产生的数据资源，和更好地整合评估组织外部的数据资源，以从中提取（检索、查询等）科学决策所需要的信息，就显得越来越重要了。这种需求既要求联机服务，响应时间要短，又要求涉及大量用于评估决策的数据，遍历范围要大。数据库技术虽然在评估事务处理方面取得了巨大的成功，但是面对新形势，它在很大程度上已经难以满足这种日益增长而且愈加迫切的需求了，主要表现在以下几个方面^[4]。

（1）评估决策必须以大量的数据为基础。这些决策数据多种多样，不但有评估事务处理系统的内部数据，而且也有大量与之相关的外部数据；不但有系统操作的细节性数据，而且也有许多分析所需要的综合性数据；不但有系统保存的当前数据，而且也有系统遗留下来的大量历史数据。而数据库技术无法提供对这些数据的支持。

（2）辅助决策的信息涉及多个评估事务处理系统的数据，这些系统相互独立，可能造成数据无序冗余的问题。这种无序冗余的存在不仅不能提高系统的整体效率，而且可能会因为系统之间的数据不一致性造成彼此功能的混乱，从而给不同系统的数据集成带来很大的难度。同时，这些评估决策所需要的数据来自不同的数据源，它们的定义、组织方式也不尽相同，评估决策者为获取有关的数据需要熟悉、理解不同的系统环境和数据定义，这样既不利于数据交流和提取，又对评估决策者提出了难以企及的要求，影响了信息的



使用和传播。

(3) 数据库技术的数据访问能力不足，它适合对单个数据记录的访问，而对多个数据记录的访问，性能就会明显下降。数据库技术的这种特点在评估事务处理环境中，针对用户多访问单个数据记录，并且操作数据频率高的这种行为模式是比较合适的。但是在评估分析处理环境中，用户的行为模式则有很大的不同，用户在评估决策时需要访问大量的数据记录，并且分析数据的频率较低。若再盲目地用数据库技术来进行设计，则必将事半功倍，加大评估决策支持系统规划、开发和实施的难度。

(4) 随着评估事务处理系统自身的演进和外部环境的变化而会有历史数据被遗留下来的可能，实际上这些数据也可以用于分析和比较，找出事物和现象发展变化的趋势，为评估组织将来的决策提供参考依据，但是由于这些数据的语义是由原来的应用程序来解释的，应用程序废弃以后，它的语义也就会随之部分或者全部地丢失了，于是无形之中增加了这些数据重新利用的难度。

应用数据库技术来存储数据，使得评估决策分析中的模型和方法缺乏相应的数据基础，影响和阻碍了评估决策支持系统的进一步发展。同时，对评估决策者而言，数据应用远比数据存储有意义。不断暴露出的问题和日益迫切的需求从两个方面呼唤着我们运用先进的信息技术来构建支持评估决策的新型信息系统，以为评估决策者在分析时提供有效的辅助手段，助其从数据访问模式向信息访问模式过渡。

第二节 研究意义

本书以国家自然科学基金项目“研究生教育评估现代技术”为依托。该项目的研究内容是应用各种现代科学技术对传统的评估理论、方法和技术进行改进和创新。本书作为该项目的一部分，将主要从理论和实证的角度，研究如何利用数据仓库、联机分析处



理和数据挖掘等技术来构建评估决策支持系统，以期推动和促进评估理论和实践的创新。

综观国内外相关文献，评估决策支持系统的理论研究和实际应用还很少，但是时代对评估决策科学化的要求比以往任何一个时候都更高、更强烈。因此，深入地研究基于数据仓库的评估决策支持系统的基本理论、方法和技术，并加以应用，无疑是一件很有意义和颇具价值的开创性工作。

从理论的角度来看，基于数据仓库的评估决策支持系统是将高科技时代的各种信息技术，如数据仓库、联机分析处理和数据挖掘等先进信息技术与评估技术进行有机结合，从而丰富评估理论，更新和发展评估方法和手段。自从数据库技术引入评估领域内以来，评估理论和方法得到丰富与发展，评估实践活动的效果和效率得到了改善与提高。然而，在信息技术飞速发展的今天，这些着眼于局部的评估事务处理系统已经日益暴露出其不足之处，决策支持功能微乎其微，信息孤岛无处不在，数据危机一触即发。基于数据仓库的评估决策支持系统正是以此为出发点所做的开拓性工作。它以系统科学理论为指导，以数据仓库、联机分析处理和数据挖掘等技术为主要工具，着重从整体和个体、内部和外部等关系中去揭示评估领域内各种事物和现象的联系与规律。它为我们提供了一种新型评估决策支持系统理论研究的思路和范式，为我们应用高科技时代的信息技术去改造和创新传统的评估理论、方法和技术这一方面的研究开辟了道路，因此，它具有重要的理论探索意义。

从实践的角度来看，随着信息技术的飞速向前和高等教育的不斷发展，在新的形势下，要保证高等教育质量，评估的决策者们迫切需要相应的信息系统来为他们高效、方便地提供相关信息以辅助分析。但是目前分析还多半是依靠评估决策者的经验来进行定性判断，缺乏定量的信息来给予相应的支持，因此很容易受到评估决策者的知识、经历、性格和精力等人为因素的干扰。本书研究开发的基于数据仓库的评估决策支持系统，是一个以高科技时代先进的信



息技术，如数据仓库、联机分析处理和数据挖掘等技术构建的决策支持平台，它将具备评估数据集成和分析的功能，可以促进各种评估数据资源能够得到合理有效的开发和利用，评估现象和问题能够得到更加全面而深刻的认识，评估决策能够更加科学、客观、高效，从而最终极大地推动高等教育健康、有序、可持续地发展。因此，它具有很大的应用价值。

第三节 国内外研究现状综述

一、数据仓库（Data Warehouse，简称 DW）

1. 数据仓库产生的背景

数据库技术作为一种数据管理手段，主要用于事务处理，并且取得了巨大的成功。在数据库中已经保存着大量的日常事务数据，而传统的决策支持系统也一直是建立在数据库基础上的。数据库技术力图使自己能够胜任各种类型的信息处理任务。然而，不同的数据处理类型有着不同的特点，以单一的方式进行数据组织的数据库技术不能反映这种差异，满足不了数据处理类型多样化的要求，并且还会发生明显的冲突。近年来，随着数据库技术的广泛应用和日益普及，人们对数据处理的这种多层次特点有了更清晰的认识。经过总结，将数据处理大致划分为两大类：事务处理和分析处理。事务处理也称为操作处理，是指对数据库进行日常的操作，为组织的特定应用服务，通常，它对一个或一组记录进行查询或者修改。分析处理则是指对决策人员和高层管理人员的分析请求所提供的信息服务，通常它要访问大量的数据^[5]。事务处理和分析处理之间的巨大差异，划清了事务处理环境与分析处理环境之间的界限，使得彼此的分离成为必然。这也告诉我们，要提高决策分析的效率及有效性，直接使用事务处理环境是不合适的，必须把用于决策分析的数据从事务处理环境中提取出来，按照分析处理的需要进行重新组



织，建立单独的分析处理环境来提供相应的支持。为此，人们专门为分析处理建立一个数据中心，它的数据可以从事务处理环境下的数据库、脱机的历史事务数据集合和外部数据源中得到，这个数据中心就是数据仓库。

2. 数据仓库的定义

20 世纪 90 年代初期，被誉为“数据仓库之父”的 Prism Solution 公司副总裁 Inmon 在其里程碑式的经典著作《Building the Data Warehouse》中提出了“数据仓库”这一概念，并在以后的一些著作中陆续对它加以完善和发展^[6-10]。他指出，数据仓库是“面向主题的（Subject Oriented）、集成的（Integrated）、稳定的（Time Variancy）、随时间变化的（Non - volatility）一系列用于管理和决策的数据集”。同时，他还对数据仓库作了如下的描述：数据仓库是 90 年代信息技术构架的新焦点，它提供集成化的和历史化的数据；它集成种类不同的应用系统；数据仓库从事物发展和历史的角度来组织和存储数据，以供信息化和分析处理之用。由于 Inmon 本人在数据仓库发展中的重要作用，他的上述论说在技术文献中不断地被引用，相对地成了一种权威的定义。

自从 Inmon 提出数据仓库这一概念以后，数据仓库在信息技术领域逐渐得到了广泛的关注，成为了非常热门的话题，很多学者从不同的侧面对数据仓库进行了探讨。如斯坦福大学的 WHIPS 小组认为，数据仓库是“一种主动存取多个分布的异构数据库的方法”^[11]。Vidette Poe 认为，为了避免对数据仓库的误解，可以把数据仓库定义为“作为决策支持系统基础的只读型分析数据库”。在她的著作中，把数据仓库和分析型数据库互为同义词而可以换用^[12]。在 Widom 等学者的文章中，把数据仓库看作为“用于查询和分析的集成数据的存储库。在这个存储库中，数据从数据源提取，转换为公共数据模型，并与仓库中已有的数据集成在一起，这样，就可直接在仓库中找到相关的信息”^[13]。Karlalalem 在 1998 年数据仓库设计和联机分析处理技术国际会议的主题发言中，把数



据仓库界定为“一个决策支持技术的集合，旨在使知识工人（主管、经理和分析专家）能很快地做出较好的决策”^[14]。这些提法从本质上说与 Inmon 的定义是基本一致的，只是从某些方面说得更具体、更易懂些。

3. 数据仓库的一个相关概念——数据集市

数据集市是一个和数据仓库紧密相关的概念，是指支持某一部门管理和决策，或者满足特定分析处理需求的数据集合^[15]。数据集市中的数据仍具有数据仓库中的数据的特点，只不过数据集市中的数据是专门为某一个部门或某个特定需求定制的而已。数据集市和数据仓库之间的关系相当于子系统和系统之间的关系，数据集市相当于子系统，数据仓库相当于系统。数据集市建设周期短（只需几个月），投资小，见效快，它的投入使用可以解决数据仓库资源竞争的紧张程度，提高数据处理的速度。

4. 数据仓库的特征

数据仓库主要有以下四个特征^[16]。

(1) 数据仓库中的数据是面向主题的。主题是一个抽象的概念，是在较高层次上对数据进行归类、综合并加以分析利用的标准。每个主题对应于一个宏观分析领域，都是用户使用数据仓库进行决策时所关心的重点。例如，在评估中，教材、优博，等等都可以看作为主题。数据仓库中的数据是面向主题来进行组织的，这种数据组织方式可以在较高层次上对分析对象的数据进行完整的描述。

(2) 数据仓库中的数据是集成的。数据仓库中的数据是从原有分散的各个数据源中抽取而来的，但绝不是原始数据的简单拷贝，而是经过系统地加工、汇总和整理而来的。这样就保证了数据仓库中的数据是关于整个组织的全局性和一致性的信息，从而实现了数据资源的增值。

(3) 数据仓库中的数据是稳定的。数据仓库中的数据主要供组织决策分析之用，所涉及的数据操作主要是查询，因此在一般情况



下是不允许直接对数据进行修改的。尽管外部环境是变化的，数据源中的数据也在发生相应的变化，但是数据经过集成进入数据仓库后，是很少或者根本不更新的。

(4) 数据仓库中的数据是随时间变化的。虽然数据仓库的用户进行分析处理时是不进行数据更新操作的，但这并不意味着数据仓库中的数据是一成不变的。随着时间的变化，数据仓库需要不断增加新的数据和删除过期的旧数据。同时，数据仓库中还包含大量的综合数据，这些综合数据中有很多也都跟时间有关，需要做相应的变动。在数据仓库中，数据是按照时间顺序来追加的，以码键来标明数据的历史时期。

5. 数据仓库与数据库的比较

数据仓库与数据库有着较大的区别，主要表现在以下几个方面^[17]。

(1) 应用的方法和对象不同。数据库技术是为了解决在文件系统应用中数据共享性差，而导致的大量数据冗余、数据潜在不一致性等缺陷而发展起来的。数据库中的数据主要反映的是某一时点上正在处理的事务状态，而不记录某一时段的事务变化状况。也就是说，它仍然是一种事务处理记录。数据库可以快速地响应事务处理过程中频繁的增加、删除、修改等操作，能很好地满足操作人员与基层管理人员的需要。数据仓库是为了构建分析处理环境而出现的一种数据存储和组织技术，其中保存的数据是对组织运营过程的记载，它已经超越了事务处理记录。数据仓库中的数据基本上是只读的，很少或者根本不做增加、删除、修改等操作，这与它支持的是分析处理功能有关。决策人员和高层管理人员在数据仓库可以围绕着某一分析需求而提炼和抽取出事物之间的相互作用关系，也可以对事物发展的全过程作趋势预测，还可以从大量数据中找出规律性的知识，这样就能够更加合理有效的制定决策。

(2) 数据集成的广度与深度不同。数据库是面向事务处理的，事务处理的目的是使日常工作自动化，需要的数据一般只与本部门



有关,而很少考虑整个组织范围内的集成应用。从广度上来说,由于事务处理应用的分散、“蜘蛛网”的存在、数据的不一致性、外部数据和非结构化数据的缺乏等诸多原因,使得在当前绝大多数组织内,数据库中数据的真正状况是分散而非集成的。从深度上来说,数据库中只有当前的细节性数据,大量的历史的细节性数据即使保存下来,也被束之高阁,而未得到充分的利用。数据仓库是面向分析处理的,分析处理的目的是为决策支持提供服务,需要的数据是集成的。全面而正确的数据是有效决策的首要前提,相关数据收集得越完整,分析处理的结果也就越可靠。从广度上来说,数据仓库充分考虑到造成数据库中数据分散而非集成的原因所在,进行了相应的整理,使得其中的数据从整个组织范围来看,是一个完整而统一的整体。从深度上来说,数据仓库中包括当前细节数据、历史细节数据、轻度综合数据和高度综合数据,这些数据的充分利用有助于满足决策分析的要求。不难看出,数据仓库对数据集成的广度和深度都要远远超过数据库。

(3) 数据集成方式不一样。数据集成有动态集成与静态集成两种方式。在以数据库为核心的事务处理应用中,采用静态集成的方式,即仅仅在开始时对所需的数据进行集成,以后就一直以这部分集成的数据作为分析的基础,不再与数据源发生联系。静态集成的最大缺点在于,如果在数据集成后数据源中数据发生了改变,这些变化将不能反馈给决策者,导致决策者使用的是过时的数据。对于决策者来说,虽然不要求随时准确地探知任何数据变化,但也不希望他所分析的数据是早已过时的数据,因此,在数据仓库中采用动态集成的方式,即以一定的周期(例如 24h)对数据仓库进行刷新,这样就可以保证分析处理的需要。

(4) 系统性能特性要求不同。数据库对应于事务处理,在事务处理环境中,用户的行为特点是数据存取操作的频率高,而每次操作处理的时间短,数据库可以满足用户高效地进行处理事务的要求。数据仓库对应于分析处理,在分析处理环境中,用户的行为特