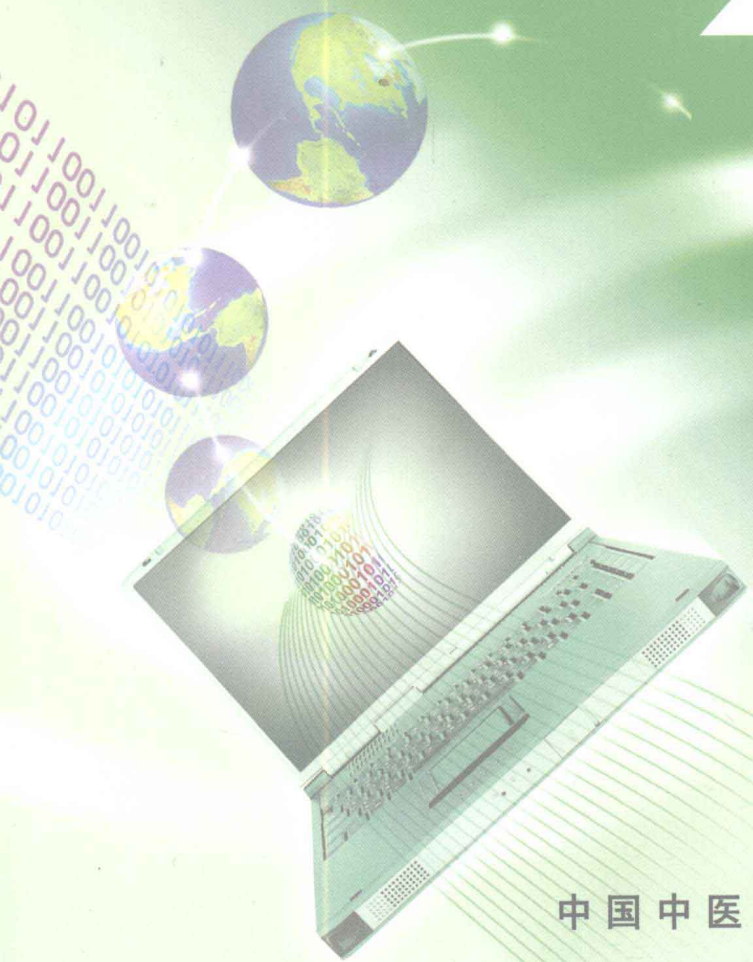
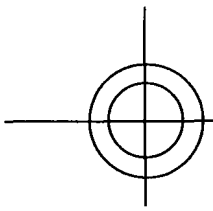
新世纪全国高等中医药院校规划教材

医学数据仓库与数据挖掘

主编　张承江



中国中医药出版社



新世纪全国高等中医药院校规划教材

医学数据仓库与数据挖掘

主　编　张承江（黑龙江中医药大学）

副主编　（以姓氏笔画为序）

杜建强（江西中医学院）

杨　明（云南中医学院）

赵文光（广州中医药大学）

中国中医药出版社

·北　京·

图书在版编目（CIP）数据

医学数据仓库与数据挖掘/张承江主编．—北京：中国中医药出版社，2008.10
新世纪全国高等中医药院校规划教材
ISBN 978-7-80231-479-5

Ⅰ.医… Ⅱ.张… Ⅲ.①医学-数据库系统-中医学院-教材②医学-数据采集-中医学院-教材 Ⅳ.R-05 TP311.13 TP274

中国版本图书馆 CIP 数据核字（2008）第 115306 号

中国中医药出版社出版
北京市朝阳区北三环东路 28 号易亨大厦 16 层
邮政编码 100013
传真 64405750
北京鑫正大印刷有限公司印刷
各地新华书店经销
*
开本 850×1168 1/16 印张 15 字数 346 千字
2008 年 10 月第 1 版 2008 年 10 月第 1 次印刷
书 号 ISBN 978-7-80231-479-5
*
定价 20.00 元
网址 www.cptcm.com

社长热线 010 64405720
读者服务部电话 010 64065415 010 84042153
书店网址 csln.net/qksd/

全国高等中医药教材建设
专家指导委员会

新世纪全国高等中医药院校规划教材

《医学数据仓库与数据挖掘》编委会

前　言

新世纪全国高等中医药院校计算机课程规划教材是依据国家教育部关于普通高等教育教材建设与改革的意见的精神，在国家中医药管理局的规划指导下，由全国中医药高等教育学会、全国高等中医药教材建设研究会组织，全国高等中医药院校教师联合参加编写，中国中医药出版社出版的高等中医药院校本科系列行业规划教材。

目前，计算机课程在全国各高等中医药院校均开设多年，计算机课程的开设对于提高中医药人才的综合素质，培养实现中医药现代化的人才有着重要的意义，因此各校对于计算机课程教学的重视程度越来越高。尽管近年来各校已经陆续开始招收计算机专业的学生，但目前全国各高等中医药院校计算机课程教学的主体对象是非计算机专业的学生。各高等中医药院校非计算机专业学生学习计算机知识时由于教学计划以及培养目标与普通院校的学生有所不同，因此就决定了高等中医药院校的计算机课程教学与普通院校有所不同。自全国各高等中医药院校开设计算机课程教学以来，由于所用教材大多是由综合性院校编写的，而且版本众多，所以一直没有较统一的教学计划，在教学上难以体现高等中医药教育的特色。基于以上现状，全国高等中医药教材建设研究会在进行充分调研的基础上，应各高等中医药院校一线教师以及教学主管部门的呼吁，于2005年开始了编写全国中医药院校计算机课程规划教材的准备工作。

按照国家中医药管理局关于行业规划教材建设的精神，本套教材的编写组织工作仍然采用了“政府指导，学会主办，院校联办，出版社协办”的运作机制，对教材进行了整体规划。全国高等中医药教材建设研究会于2005年7月在北京召开了“全国高等中医药院校计算机课程教学与教材建设研讨会”，会上来自全国多家高等中医药院校计算机教学的专家以及管理人员一致认为编写一套适合教学的计算机课程规划教材是十分必要和急需的，并初步提出了规划教材目录。之后全国高等中医药教材建设研究会组织有关专家对规划教材的目录进行了多次讨论，最终确定了12门新世纪全国高等中医药院校计算机课程规划教材，其中大部分是供非计算机专业教学使用的计算机教材，也有部分供计算机专业教学使用并能体现中医药特色的教材。本套教材的具体书目为：《SAS统计软件》《SPSS统计软件》《多媒体技术与应用》《计算机基础教程》《计算机技术在医疗仪器中的应用》《计算机网络基础与应用》《计算机医学信息检索》《计算机应用教程》《网页制作》《医学数据仓库与数据挖掘》《医学图形图像处理》《医院信息系统教程》。

本套教材在组织编写过程中，严格贯彻国家中医药管理局提出的“精品战略”精神，从教材规划到教材编写、专家论证、编辑加工、出版，都有计划、有步骤地实施，层层把关，步步强化，使“精品意识”、“质量意识”贯彻全过程。每种教材均经历了编写会、审稿会、定稿会的反复论证，不断完善，重在提高内在质量。注意体现素质教育和创新能力、实践能

力的培养，为学生知识、能力、素质协调发展创造条件；同时在编写过程中始终强调突出中医药人才的培养目标，在教材中尽量体现中医药特色。

本套教材从开始论证到最后编写工作的完成，始终得到了全国各高等中医药院校各级领导和教学管理部门的高度重视，各校在人力、物力和财力上均给予了大力支持。广大从事计算机教学的一线教师和管理人员在这套教材的编写工作中倾注了大量心血，充分体现了扎实的工作作风和严谨的治学态度。在此一并致以诚挚的谢意！

新世纪全国高等中医药院校计算机课程规划教材的编写是一项全新的工作，所有参与工作的教师都充分发挥了智慧和能力，通过教材建设工作对教学水平进行总结和提高，并进行了积极的探索。但是，一项创新性的工作难免存在一些不足之处，希望各位教学人员在使用过程中及时发现问题并提出宝贵意见，以便我们重印或再版时予以修改和提高，使教材质量不断提高，逐步完善，更好地适应新世纪中医药人才培养的需要。

全国中医药高等教育学会
全国高等中医药教材建设研究会
2007 年 8 月

编写说明

数据处理和分析技术越来越广泛且深入地应用于医学领域。数据库、数据仓库、分析与挖掘技术的进步极大地提高了医学信息的分析与处理能力，从海量医学信息中提取有价值的医学知识已由期望变为现实，医学决策过程将更加科学和高效。

近年来，数据仓库技术、数据分析与挖掘技术取得了很多新的进展，特别是非结构化数据和复杂类型数据的处理与分析。

医学数据仓库是数据仓库技术面向医学领域数据的具体实现。与其他企业数据仓库相比，医学数据仓库的数据来源、数据类型和数据特征都有其特殊之处。建立医学数据仓库是医学数据分析处理的基础，是医学信息技术发展的必然，对于医学、医疗卫生、药物学和医学管理等领域的研究与应用都有巨大的推动作用。

医学数据挖掘所面临的数据对象类型十分丰富，包括文本、图形、图像等；数据来源也非常广泛。临床医疗和医学研究已积累了大量的信息，如何有效地存储、检索、处理和分析医学数据，为医学决策提供支持，已为医学工作者和信息技术工作者强烈关注。该领域的分析与挖掘技术极富前景，也极具挑战性。

本书力图从两个角度观察和分析医学数据处理与分析技术。一方面从信息技术的角度介绍数据仓库及数据分析与挖掘的基本原理、技术和发展前景；另一方面从医学科学的角度介绍医学信息与医学数据以及相关处理技术的特殊性和最新的研究成果。

本书可作为医学院校的本科生和研究生教材。也适合于医学领域从事数据处理的专业技术人员阅读。

本书分3篇共13章。第0章和第3章由张承江编写；第1章由杨明编写；第2章由王珍编写；第4、6章由赵文光编写；第5、9章由闫朝升、张柯欣、马俊编写；第7章由杨玉凤编写；第8章由闫朝升编写；第10章由李志敏编写；第11章由杜建强、廖春华编写；第12章由李志敏、廖春华编写；全书最后由张承江统稿。

欢迎读者对本书的缺点和不足批评指正。

E－mail：zcj@hljucm.net

编　者

2008年7月

目 录

第 1 篇 数据仓库

第 2 篇 数据挖掘

第3篇 数据仓库与数据挖掘应用

0 绪 论

医学信息及数据的分析和处理技术已进入了飞速发展的时代。信息论、人工智能、数据库、计算机信息管理系统、计算机决策支持系统、运筹学、统计学等学科的发展与集成，极大地提高了数据处理能力。相反，存储数据的爆炸性增长也提出了新的需求，刺激相关数据处理技术的发展。医学数据仓库技术、医学数据挖掘技术作为多学科领域，近年来取得了令人瞩目的成就，在医学领域发挥着越来越重要的作用。

0.1 医学信息技术概述

0.1.1 信息与信息技术

伴随着人类文明社会的发展，人们对信息的依赖性越来越强。随着科学技术的发展，人类处理和分析信息的能力也越来越强。

1. 信息

信息是事物运动的状态和方式。人们通过感知事物的运动状态和方式来认知事物。“事物”包括一切可感知和研究的对象，可以是客观存在，也可以是主观反映。“运动”包括自然的、社会的和思维的等形式。

信息论的奠基人香农把信息定义为一个可量化的名词：信息是一个事件发生概率的对数的负值。

香农给出如下公式：

$$I = -\log_2 P \tag{0-1}$$

P 为某一事件的发生概率，I 是信息量的二进制表示的位数。该公式说明事件发生的概率越小，表示信息的位数就越大。例如：某一事件发生的概率为 1/16，则信息的表示需要 4 位，即 $-\log_2 2^{-4}$。若发生的概率为 1/2，则 $I = -\log_2 2^{-1} = 1$（位）。

2. 信息的特征

（1）*语法特征*：信息的语法特征包括信息的语法、存储和传递的描述。语法特征与信息的组成结构相关联，描述信息载体的行为规则，例如编码方式。语法特征对应的即是“数据”。

（2）*语义特征*：信息的语义特征表述的是信息的具体含义。对语义的理解需要背景知识。有一个非常典型的英文句子可以有 10 种以上的解释：“Time flies like an arrow”。

（3）*语用特征*：信息的语用特征是针对信息的目的性而言，为一定的目的服务，以减少不确定性。语用特征决定采取何种措施，在信息处理过程中通过应用程序来实现。

3. 信息技术

信息技术是指实现信息的获取、传递、加工、再生和施用功能的技术。信息技术是感测、通信、计算机和控制技术的整体。

本书所关注的是计算机科学技术中的数据库技术和人工智能技术。

0.1.2 医学信息与医学知识

1. 医学信息

医学信息是医学科学领域的信息，涉及医学、药物学、卫生学和医学管理学等专门知识。依据信息的语义特征和语用特征，医学信息可以解释为：其一，医学信息是以医学、医疗卫生、药物学和医学管理学为信息内容的；其二，医学信息是以医学、医疗卫生、药物学和医学管理为应用领域的；其三，医学信息的处理依赖于以计算机技术为核心的信息技术。

2. 医学知识

数据是信息的载体，是信息的语法表述。经过解释的数据演化为信息，而对信息进行加工，集成为知识。反过来，知识又指导数据解释。医学知识有两种类型，一种来自于医学文献，称为科学知识；另一种来自于临床专家，称为经验知识。

下面以一个临床医疗过程说明医疗数据、医疗信息和医疗知识的关系。

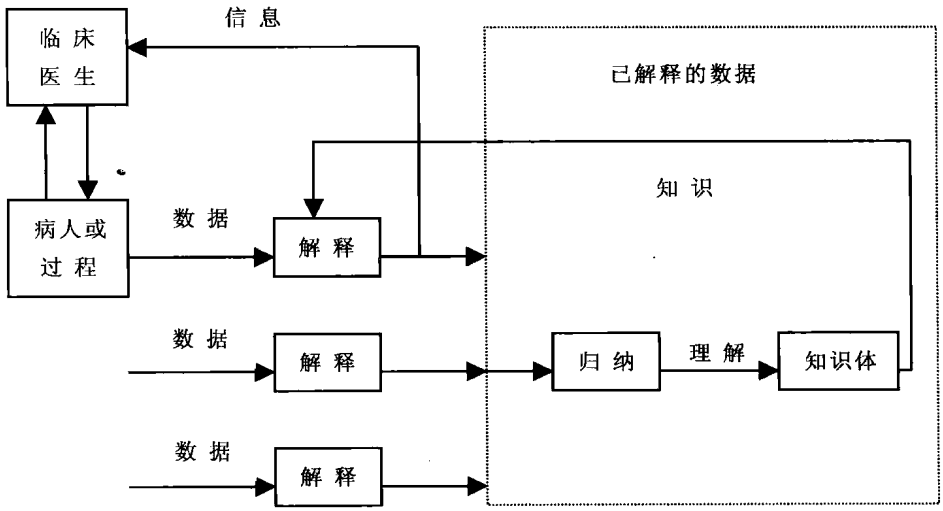


图 0-1 临床数据处理的循环过程

以上过程可借助计算机系统来实现，可使用数据仓库存储数据，利用数据分析与挖掘技术进行归纳推理，解释数据，从而达到医学决策支持的目的。

0.1.3 医学信息技术

现代信息技术的成果，极大地促进了医学信息处理的研究与应用。同时，医学信息的特点也需要有针对性的、创新的、计算机科学领域的技术、手段和方法。医学信息处理需要采

集技术、存储技术、检索技术、数据仓库技术、数据挖掘技术、流媒体技术、可视化技术、图像处理技术、决策支持技术等。

J. H. Van Bemmel 认为在信息学中，可以分出 3 个不同的研究层次：

- 基础计算机科学
- 应用方法信息学
- 应用信息学

医学信息系统的开发研究主要属于第 3 个层次。本书所介绍的数据仓库及数据挖掘是医学信息系统的重要组成部分。

0.2 数据库技术的演化

经过几十年的演变，数据管理经历了从简单的文件管理到复杂的多种数据库集成的发展过程。见图 0－2。

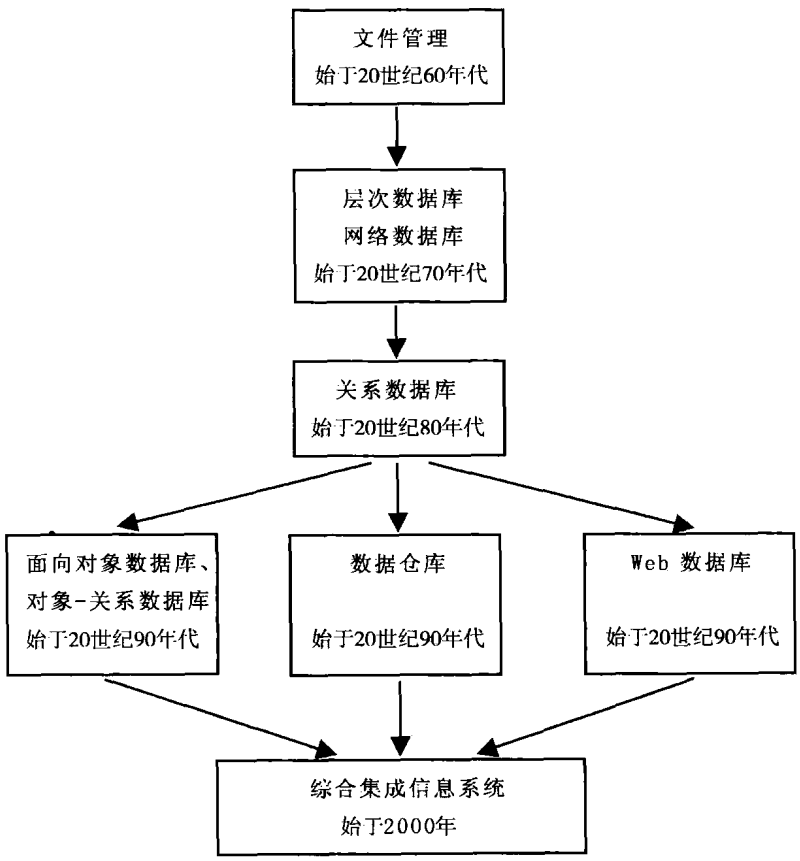


图 0－2 数据库技术的演化过程