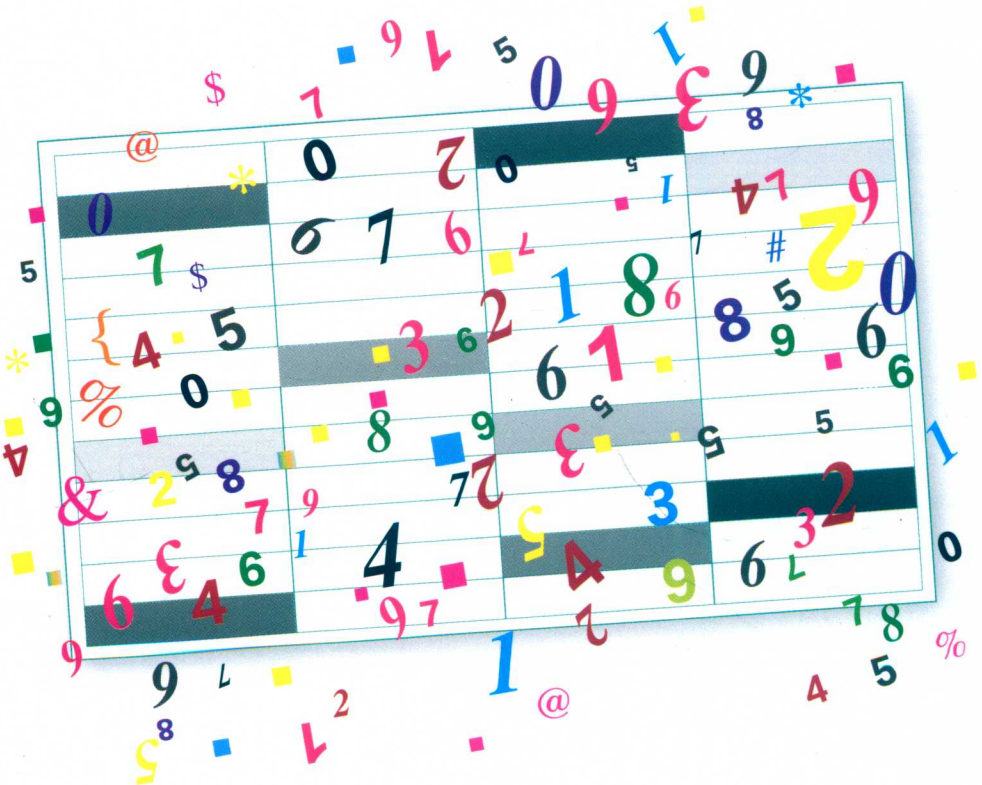
大数据系列丛书



审计分析从关系到大数据

董　东　王艳君　陈玉哲　编著

清华大学出版社

大数据系列丛书

审计分析：从关系到大数据

董　东　王艳君　陈玉哲　编著

清华大学出版社
北京

内 容 简 介

本书针对计算机在审计中数据分析所需要的技术、方法和工具，按照技术发展的脉络，介绍基于关系数据库的以结构化查询语言(SQL)为工具的查询分析，基于数据仓库的可视化途径的多维分析，基于模型训练的机器学习途径的挖掘分析，以及基于大数据的相关分析，力图使读者能够应用本书介绍的方法、工具和技术完成审计目标。

全书共有7章。第1章介绍审计数据分析的基本概念；第2章以T-SQL为例介绍结构化查询语言，包括基本的DR、DDL、DML、DCL等，以及结构化查询语言在财务审计与业务数据结合的财务审计中的应用技术；第3章介绍数据库用户以及授权、数据导入导出等技术，以及这些技术在审计中的应用；第4章介绍高级查询分析技术，包括游标、触发器、视图、索引等；第5章介绍如何通过多维分析发现审计线索；第6章介绍数据挖掘途径的审计数据分析；第7章介绍基于大数据的审计分析。

本书可作为计算机审计工作者的技术参考书。

图书在版编目(CIP)数据

审计分析：从关系到大数据/董东，王艳君，陈玉哲编著. —北京：清华大学出版社，2019
(大数据系列丛书)
ISBN 978-7-302-52052-8

Ⅰ. ①审…　Ⅱ. ①董… ②王… ③陈…　Ⅲ. ①审计学—研究　Ⅳ. ①F239.0

中国版本图书馆CIP数据核字(2019)第009619号

责任编辑：张　玥　常建丽
封面设计：常雪影
责任校对：时翠兰
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：11.5　　**字　　数**：263千字
版　　次：2019年7月第1版　　**印　　次**：2019年7月第1次印刷
定　　价：49.00元

产品编号：079897-01

前　言

PREFACE

2000年的春天，河北省审计厅陈金如副厅长来到河北师范大学，就计算机教育与培训问题与我进行了第1次“聊天”。当时没有想到，那次“聊天”却是十八年合作的开始。

当时，河北省审计系统面临如何提升全省审计工作者的计算机审计水平，以适应信息技术的迅猛发展问题。从事审计的人员如果不懂信息技术，那么在信息化环境下就进不了被审单位的门，打不开被审单位的账。审计署指出，开展计算机审计要靠审计人员素质的提高，靠审计机关知识结构的改善。根据审计署的要求，河北省审计厅认为计算机审计的发展和审计事业的发展已经密不可分，应从事关全省审计事业发展的高度思考计算机审计问题，强力推进计算机审计。应切实增强危机感和紧迫感，下大力量将计算机辅助审计工作搞上去。

2000—2005年，我应邀为审计厅的几次短期培训授课。从2006年开始，河北省审计厅每年都与河北师范大学联合举办“河北省审计机关计算机审计中级培训”，目前仍然在进行。参训学员大部分已成为各单位审计信息化工作骨干，并多次在审计署组织的AO应用实例评选和计算机审计专家经验评选中获奖。

在培训项目中，我不仅负责组织工作，而且主讲了“数据库应用技术”“审计数据高级分析技术”两门课程。为了把审计业务与计算机技术结合起来，让非计算机专业的审计师能够理解、掌握和应用计算机技术，在省审计厅安排下，我还深入审计现场，实地调研，学习审计业务，了解审计人员的需求；钻研审计一线人员提出的技术问题，提出解决方案；不断研究数据仓库、数据挖掘、大数据等技术在河北省计算机审计业务中的应用等问题，并将研究结果应用于培训。2014年，受河北省审计厅计算机中心委托，编写了“技巧系列丛书”中的《数据挖掘技巧》一书的初稿，并根据审计署、省审计厅的意见进行了修改，于2017年出版。十八年来累积的讲义以及其他技术资料经过整理，形成了本书。本书的内容不仅涉及与计算机审计有关的技术，还包括如何规避常见的技术错误。

本书初稿由河北师范大学的董东老师完成。河北师范大学的王艳君、陈玉哲、张朝昆等老师在历年培训过程中对前4章进行过修订。王艳君对本书的章节组织以及第1～4章内容提出了修改意见；陈玉哲对第5～7章提出了修改意见。董东最后对全稿进行了精简。

本书得到河北师范大学应用开发基金(L2013K01)和河北省审计厅2018年重点科研课题(201805)资助，在此表示感谢。

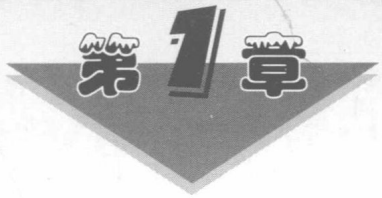
囿于学识，在思想、方法或者技术等方向定有不当之处，望读者批评指正。

董　东

2018年12月于河北师范大学

目　录

CONTENTS



审计数据分析

数据分析在计算机科学领域和统计学领域已经被广泛关注并取得了大量研究成果，并在各种领域已经得到广泛应用以解决本领域的具体问题，如商品推荐、交通管理、舆情分析等。

审计是对业务实体经济活动的真实性、合法性和效益情况的验证过程。计算机审计是以计算机为基本工具，以业务系统、财务收支系统电子数据为主要对象，通过计算机软件工具的功能与审计人员经验判断的结合，实现对业务实体经济活动进行真实、合法、效益的信息化审计过程。在“金审工程”推动下，我国计算机审计的研究和应用发展较快。但是，如何应用数据分析技术有效地实施计算机审计仍然是一个热点问题。

审计数据分析活动主要是验证(validation)活动，即按照政策、法规对反映经济活动数据项或者数据处理进行检查，以查证是否为完全按照政策、法规执行的活动，即查证被审单位是否正确进行了经济活动。

随着信息技术的普及和广泛应用，企事业单位每天产生的经济业务活动均能够以不同形式被各种各样的信息系统记录下来。那么，从这些纷繁复杂而且大量的数据中完成审计任务和达到审计目标成为当前审计领域关注的重要问题之一。本书讨论从数据中发现审计线索的相关数据分析概念、技术和方法。

1.1 审　　计

国务院设立审计机关，对国务院各部门和地方各级政府的财政收支，对国家的财政金融机构和企业事业组织的财务收支，进行审计监督。审计机关通过客观地获取和评价有关经济活动与经济事项认定的证据，以证实这些认定与既定标准的符合程度，并将结果传达给有关使用者。

审计的主体是从事审计工作的专职机构或专职的人员，是独立的第三者，如国家审计机关、会计师事务所及其人员。审计的客体是被审计单位的财政、财务收支及其他经济活动。审计的基本工作方式是搜集证据，查明事实，对照标准，做出好坏优劣的判断。审计的主要目标不仅要审查评价会计资料及其反映的财政、财务收支的真实性和合法性，而且还要审查评价有关经济活动的效益性。

1.2 计算机审计

计算机审计是以计算机为基本工具，以被审单位的财务收支电子账和业务系统的数据库为主要对象，在计算机软件工具的辅助下结合审计人员经验，对业务实体经济活动进行的真实、合法、效益的信息化审计过程。计算机审计丰富了传统审计中各阶段的内容。一个计算机审计项目主要包括3个阶段：准备阶段、实施阶段和报告阶段。

在审计准备阶段，审计人员和信息技术人员需要到被审计单位了解信息系统、业务过程、电子数据等状况；调查系统结构、业务处理流程、数据存储结构及数据处理流程；从被审单位系统采集审计相关的会计核算数据和业务数据；将数据导入审计软件；验证数据的合法性、有效性与完整性。

在审计实施阶段，根据审计人员经验和相关技术方法设计模型，对审计数据进行查询分析、多维分析或者发掘分析，发现审计线索。然后寻找疑点、落实线索形成证据。

在审计报告阶段，形成审计报告，送达有关使用者。还将被审单位系统、数据情况，以及操作流程加以归档，供以后审计使用。具体包括：被审单位信息系统开发及应用情况；业务流程及数据流程文档；与审计数据相关的数据结构资料；本次审计所建立、应用的审计模型，发现问题的方法思路及编写的程序或操作步骤。

1.3 审计数据分析

根据技术和模型的不同，审计数据分析分为4类：基于数据库的审计分析、基于数据仓库的审计分析、基于数据挖掘的审计分析和基于大数据的审计分析。

基于数据库的审计分析也称为查询型分析，指建立模型、设计结构化查询语言(Structured Query Language，SQL)语句查询数据库、分析查询结果的过程。查询型分析的主要对象是关系数据库管理系统中的表，涉及单表查询、多表查询等技术。一个成功的查询分析既需要定义一个很好的业务问题(查什么)，也需要设计很好的SQL语句(如何查)。

当需要从不同角度可视化地观察某个时期、某个主题的历史数据时，就需要应用基于数据仓库的分析技术。多维分析模型描述了如何在多个维上观察业务事实。事实是一组与业务事务或者事件相关的数据项。例如，会计事务中的一笔凭证就是一个事实，这个事实中含有日期、凭证类型、部门、科目、借贷金额等数据项。维是一组对事实进行分析所使用的属性，如在日期和科目属性上对凭证事实进行分析。

很多情况下，审计活动是基于经验的。经验是审计人员在长期实践中经过归纳、推演发现的一些特征或者规律。基于数据挖掘的分析指先由计算机发现隐藏在数据中的、不为审计人员预先所知的规则、规律、模式或者趋势，然后审计人员对这些发现的计算机“经验”进行甄别和利用。

大数据指数据量大得超出了常规软件的管理能力的数据。大数据分析的策略不同于常规数据。这些策略有基于总体而不是基于抽样，基于相关而不是基于因果，等等。

基于关系数据库的审计分析、基于数据仓库的审计分析、基于数据挖掘的审计分析和基于大数据的审计分析一起构成了审计数据分析，但它们有各自的适用范围。基于关系数据库的审计分析是已知数据结构的情况下，对操作型数据的访问，比较容易建立模型；基于数据仓库的审计分析需要对被审数据重新组织，需要从不同角度观察数据，建立模型较为复杂；而基于数据挖掘的审计分析需要审计人员根据具体业务问题选取合适的挖掘模型，从而在数据挖掘工具帮助下发现有用模式和趋势，处于高级的分析层次；基于大数据的审计分析的关键是获取数据的总体，有了总体，就可以用简单的模型解决复杂的问题。

1.3.1 基于关系数据库的审计分析

从集合角度看，关系(relation)是元组(tuple)的集合。一个元组中有若干属性(attribute)值。在关系中能唯一标识一个元组的属性集称为关系的超键(super key)。不含多余属性的超键称为候选键(candidate key)。一个关系中可能存在多个候选键，用户选作元组标识的候选键称为主键(primary key)。

例如，有关系：医生(医生号，姓名，性别，年龄，职称，身份证号，部门号)，其中“医生”是关系的名称；医生号、姓名、性别、年龄、职称、身份证号、部门号是该关系的属性。这些属性中，医生号和姓名两个属性一起可以作为超键，姓名和身份证号这两个属性一起也是超键。而医生号是候选键，身份证号也是候选键。可以把“医生号”作为主键，也可以把“身份证号”作为主键，但一个关系上只能有一个主键。

关系的属性值不可分解，从而不允许表中有表；一个关系中的元组不可重复；关系是元组的集合，集合的元素是无序的；属性之间也无序。

在关系上可以施加3类完整性约束：实体完整性(entity integrity)、参照完整性(referential integrity)和用户定义的完整性。

1. 实体完整性

若属性A是关系R主键中的属性，则属性A不能取空值。一个元组对应现实世界的一个实体，一个关系对应现实世界的一个实体集。现实世界中的实体是完整的，即每个实体都具有属性值。而在关系中不可能穷尽所有的属性，只能有感兴趣的部分属性。无法要求这部分属性都必须有属性值，但可以要求主键必须有值，即不能取空值(NULL)。所谓空值，就是“不知道”或“无意义”的值。如果主键中的属性取空值，就说明存在某个不可标识的实体，即存在不可区分的实体。例如，医生(医生号，姓名，年龄，身份证号)中，如果医生号为主键，则该属性不能取空值。

2. 参照完整性

设F是关系R的一个或一组属性，但不是关系R的键，如果F与关系S的主键对应，则称F是关系R的外键(foreign key)。例如，医生(医生号，姓名，部门号)和部门(编号，名称，级别)两个关系，“部门号”是医生关系模式的外键，对应于“部门”关系的属性“编号”。

若属性(或属性组)F是关系R的外键，它与关系S的主键对应(关系R和S可能是

同一个关系），则对于 R 中每个元组在 F 上的值都必须或者取空值（F 的每个属性值均为空值），或者等于 S 中某个元组的主键值。

例如，对于“医生”关系中的每个元组，其在“部门号”上的取值要么是空值，要么是“部门”关系中某一元组在“编号”属性上的值。也就是说，某个医生所在的部门必须是存在的一个部门。

在“医生”和“部门”这两个关系中，为了维护参照完整性，删除“部门”关系中的元组时可以采用 3 种策略：级联删除、级联受限删除和置空值删除。例如，删除“部门”关系中的 10101 部门，级联删除是将关系“医生”中所有部门号（外键）为 10101 的元组删除，接着将关系“部门”中键为 10101 的元组也删除；受限删除是当关系“医生”中没有任何元组的部门号（外键）为 10101 时，才执行删除操作，否则拒绝删除；置空值删除则将关系“医生”中部门号（外键）为 10101 的元组的部门号置空，然后删除关系“部门”中键为 10101 的元组。

如果要在关系“医生”中插入元组，其“部门号”在“部门”关系中存在，则插入操作可以顺利执行；如果不存在相应的元组，则可以有两种策略：受限插入和递归插入。前者拒绝在“医生”关系中插入元组；后者首先向“部门”关系中插入相应的元组，其编号（主键）值等于“医生”关系插入元组的部门号（外键）值，然后向参照关系中插入元组。

3. 用户定义的完整性

用户定义的完整性就是针对某一具体关系数据库的约束条件，它反映某一具体应用涉及的数据必须满足的语义要求。关系数据库的实现提供定义和检验这类完整性的机制，以便用统一的、系统的方法处理它们，而不要由应用程序承担这一功能。例如，约束“年龄”属性是 0～200 的一个整数。

数据库（database）是数据的汇集，它以某种组织形式存储在介质上。数据库管理系统（Database Management System，DBMS）是管理数据库的系统软件。从操作系统角度看，数据库表现为一个或一组特定扩展名的文件。例如，一个 Access 数据库是一个扩展名为.mdb 的文件，再如，一个 SQL Server 2008 的数据库在最简单的情况下由两个文件构成，一个文件的扩展名为.mdf，另一个文件的扩展名为.ldf。

按照关系模型实现的数据库关系系统称为关系数据库管理系统。目前广泛使用的数据库管理系统有甲骨文公司的 Oracle、Microsoft 公司的 SQL Server、IBM 公司开发的 DB2 以及开源数据库 MySQL 等。由于大多数的数据库管理系统产品都是关系模型的实现，所以后文中的“数据库管理系统”默认指“关系数据库管理系统”。

图 1-1 展示了关系模型中的术语与关系数据库中的术语对应关系。

从图 1-1 中可以看到，一个关系对应数据库中的一个表；若干关系组成关系模式，若干表形成数据库模式；一个元组对应表中的一行，元组在某个属性上的值对应行在某个列上的值。

数据库管理系统的主要功能有：定义数据库、操纵数据、控制数据库和维护数据库。

1）定义数据库

DBMS 提供数据描述语言（Data Definition Language，DDL），定义数据的结构、数据与数据间的关系、数据的完整性约束等。数据库中的主要对象是表（table），数据组织在

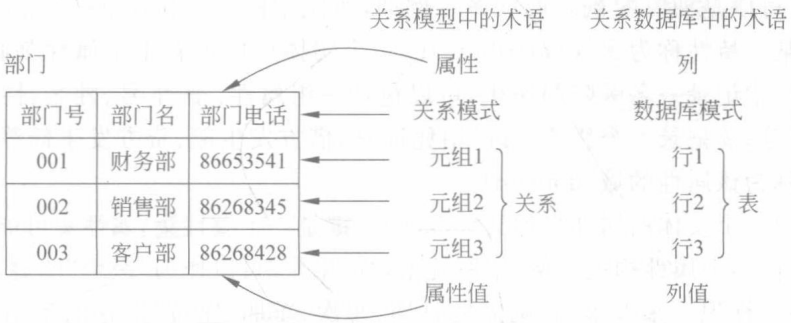


图 1-1 关系模型中的术语与关系数据库中的术语对应关系

表中。数据的完整性约束既可以定义在表上，也可以定义在表之间。数据库中的表、约束等对象合起来称为数据库模式(schema)。

2) 操纵数据

DBMS 提供数据操纵语言(Data Manipulation Language，DML)，可实现对数据的查询、插入、删除和修改等操作。DML 有两种用法：一种方法是把 DML 语句嵌入到高级语言中；另一种方法是交互式地使用 DML 语句。对于第一种方法，DBMS 必须提供预编译程序，预处理嵌入 DML 语句的源程序，识别 DML 语句，转换为相应高级语言能调用的语句，以便原来的编译程序能接受和处理它们。

3) 控制数据库

数据库的控制功能包括并发控制、数据的安全性控制、数据的完整性控制和权限控制，保证数据库系统正确有效地运行。

4) 维护数据库

已经建立好的数据库，在运行过程中需要进行维护。维护功能包括数据库出现故障后的恢复、数据库的重构、性能的监视等。这些功能大部分由实用程序完成。

数据字典(Data Dictionary，DD)中存放着数据库体系结构的描述。对于用户的一个处理请求，DBMS 都要查阅数据字典。

当应用程序需要处理数据库中的数据时，首先向数据库管理系统发送一个数据处理请求，数据库管理系统接收到这一请求后，对其进行分析，然后执行数据操作，并把操作结果返回给应用程序。由于应用程序直接与用户打交道，而数据库管理系统不直接与用户打交道，所以前者常被称为“前台”，后者常被称为“后台”。

由于应用程序是向数据库管理系统提出服务请求，通常称为客户机(Client)程序，而数据库管理系统是为其他应用程序提供服务，通常称为服务器(Server)程序，所以又将这种实现模式称为客户机/服务器(C/S)模式。

对于一个应用系统，需要有哪些表？表之间是什么关系？即如何设计数据库模式？数据库模式一般从概念模型得到。

概念模型用于信息世界的建模。概念模型不依赖于具体的计算机系统。概念模型可以转换为计算机上某一 DBMS 支持的特定数据模型。

在概念模型中，客观存在并可相互区别的事物称为实体(entity)。实体可以是具体的

人、事、物，也可以是抽象的概念或联系。例如，医院中的一名医生、一个部门都是实体。实体具有的某一特性称为属性(attribute)。一个实体可以由若干个属性刻画。例如，要在数据库系统中记录一名医院的医生，可以包括一组属性：医生号、姓名、性别、年龄、职称、身份证号等；要记录一个凭证，可以用凭证号、借方发生额、贷方发生额等属性。属性的取值范围称为该属性的域(domain)。

唯一标识一个实体的属性集称为键(key)。键是一个属性集，属性集可能由单个属性构成，也可能由多个属性构成。例如，医院中，“医生号”可以作为“医生”的键，这时键由单个属性构成；要标识一张火车票，则需要日期、车次、车厢、座位号，这时键由一组属性构成。键的确定是一个语义范畴的问题。实体可能同时存在多个键。例如，“医生号”可以作为医生的键，而“身份证号”也可以作为医生的键。当实体有多个键时，选其中一个键作为主键。

用实体名及其属性名集合抽象和刻画同类实体，称为实体型(entity type)。例如，医生的实体型可以表示为：医生(医生号，姓名，性别，年龄，职称，身份证号)，属性之间用逗号隔开。同型实体构成的集合称为实体集(entity set)。

现实世界中，事物内部以及事物之间的联系在信息世界中反映为实体内部的联系和实体之间的联系(relationship)。两个实体集之间的联系有3种类型：一对一联系、一对多联系和多对多联系。

如果对于实体集A中的每一个实体，实体集B中至多有一个实体与之联系，反之亦然，则称实体集A与实体集B具有一对一联系，记为1∶1。例如，一份审计报告只涉及一个审计项目，而一个审计项目最终产生一份审计报告。

如果对于实体集A中的每一个实体，实体集B中有n个实体($n\geqslant 0$)与之联系，反之，对于实体集B中的每一个实体，实体集A中至多只有一个实体与之联系，则称实体集A与实体集B有一对多联系，记为1∶n。例如，处室和工作人员的联系。一个处室中有若干工作人员，一个工作人员至多在一个处室中工作，由处室到工作人员是一对多联系。需要注意的是，一对多联系不是对称的，由处室到工作人员是一对多联系，反过来，由工作人员到处室是多对一联系。

如果对于实体集A中的每一个实体，实体集B中有n个实体($n\geqslant 0$)与之联系，反之，对于实体集B中的每一个实体，实体集A中也有m个实体($m\geqslant 0$)与之联系，则称实体集A与实体集B具有多对多联系。记为m∶n。例如，审计小组和工作人员的联系。一个审计小组由多个工作人员组成；一个工作人员可以参加多个审计小组。再如，一个审计小组可以负责多个审计项目；一个审计项目可以由多个审计小组负责。

实体-关系图(E-R图)提供了表示实体型、属性和联系的可视化方法。在E-R图中，使用矩形表示实体型，矩形框内写明实体名。例如，表示医院的医生和部门：

用椭圆表示属性，并用无向边将其与相应的实体连接起来。例如，医生和部门的属性表示如图1-2所示。

使用菱形表示联系，菱形框内写明联系名，并用无向边分别与有关实体集连接起来，

同时在无向边上标记联系的类型(1∶1、1∶n 或 m∶n)。例如,医生和部门的联系如图 1-3 所示。

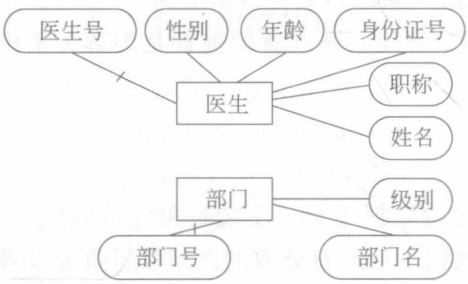


图 1-2 属性

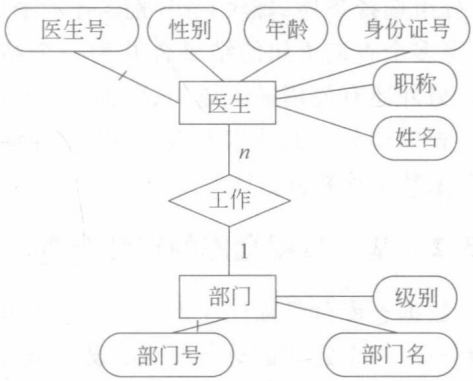


图 1-3 联系

联系本身也是一种实体型,也可以有属性。如果一个联系具有属性,则这些属性也要用椭圆形表示,并用无向边与该联系连接起来。

联系存在于两个实体之间,也可以存在于 3 个实体之间。例如,一次住院收费涉及 3 个实体:医生、患者和收费项目,可以用图 1-4 描述这三者之间的联系。

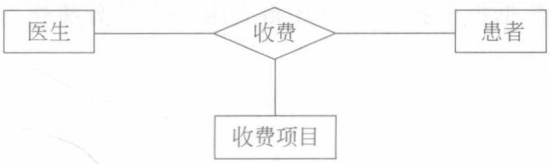
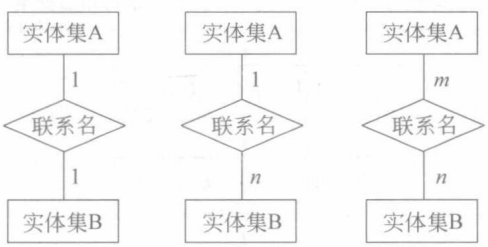


图 1-4 住院收费之间的联系

图 1-5 展示了两个实体集之间每种联系的 E-R 图表示方法。



(a) 1∶1 联系 (b) 1∶n 联系 (c) m∶n 联系

图 1-5 3 种联系的表示

通常将实体集及实体集间联系的上述表示模型称为实体-关系(relationship)模型;从分析用户项目涉及的对象及对象之间的联系出发,到获取 E-R 图的这一过程称为概念模型设计。

查询型分析指通过结构化查询语言(Structured Query Language, SQL)形成 SQL 语句交互式查询数据库的过程。查询型分析的主要对象是数据库管理系统中的表。运用 SQL 语句进行有关数据采集、数据预处理和数据分析在我国面向数据的计算机辅助审计

中的应用十分广泛，出现了大量成功的案例。数据查询语句，包括单表查询、面向多表的连接查询和嵌套查询是实现查询分析的主要技术。其他技术还有：使用行选择条件、分组、分组选择条件、排序、选择前若干行查询结果等；使用统计函数进行全表统计和分组统计；将多个查询语句的结果合并为一个结果；如何将查询结果保存在永久表和临时表中等。另外还有使用插入语句、数据更新语句和数据删除语句进行数据操纵。

查询型分析的技术要点是 SQL。而应用 SQL 解决实际问题的前提是理解关系模型和关系数据库模式。

1.3.2 基于数据仓库的审计分析

数据仓库就是面向主题的、集成的、相对稳定的数据集合。主题是关注的问题，一个主题对应一个宏观的分析领域。数据仓库中的数据是从原有分散的面向联机事务处理的数据库中抽取出来的。由于数据仓库的每一主题对应的源数据在原有分散的数据库中可能有重复或不一致的地方，因此数据在进入数据仓库之前必须经过数据的抽取、清洗和转换。数据仓库中存放的是历史数据，而不是日常事务处理产生的数据，数据进入数据仓库后极少甚至根本不被修改。

多维数据模型提供了多角度、多层次的分析应用，如基于时间维、地域维等构建星形模型或者雪花模型，可以实现在各时间维度和地域维度的交叉查询。多维分析技术以及分类技术、聚类技术等数据挖掘技术已经开始应用于审计分析中。多维分析途径的计算机审计过程如图 1-6 所示。

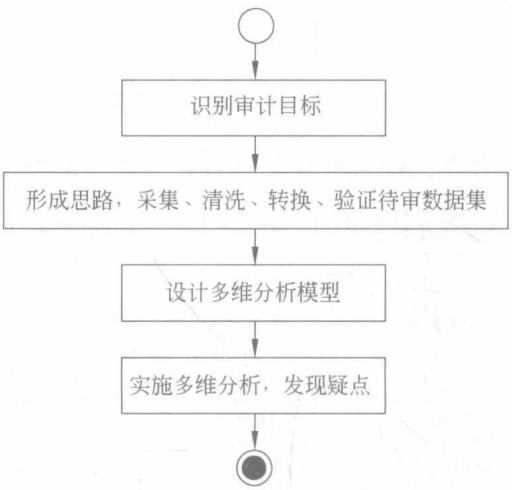


图 1-6　多维分析途径的计算机审计过程

例如，税务机关可以批准企业延期纳税，一个可能的审计目标是查证有无违法批准延期纳税的疑点。于是审计人员设想，能否从若干税种中选定一个常见的税种，如“企业所得税”，然后从若干税务机关中选择一个或者若干个，观察这些税务机关在某年某月某日的纳税额，从企业缴纳的时间、数额特征中寻找线索，这就是审计思路。

有了思路，就按照这个思路从税务机关采集、清洗和转换数据，存储到数据库或者数

据仓库中，整理成事实表。然后设计多维分析模型。想法中的“税种”“税务机关”和“日期”就是多维分析模型中的3个“维”；而纳税额就是多维分析模型中的“度量值”。

建立好多维分析模型后，应用软件工具以多维方式浏览数据，观察可疑模式，发现审计疑点。

1.3.3 基于数据挖掘的审计分析

从审计角度看，数据挖掘就是根据事先明确的审计目标，对被审单位的大量业务数据进行分析，揭示其中潜在的逻辑关系和规律，进而形成明确而有效的审计思路的过程。例如，银行贷款五级分类真实性情况审计。由于种种原因，某些事实上的不良贷款（次级、可疑或损失）被商业银行人为划归正常贷款分类（正常或关注）中。按照传统的审计思路，需要与被审单位了解贷款五级分类规则，然后根据被审单位提供的规则逐一核实各笔贷款的分类是否正确。挖掘型分析途径是先让计算机从被审单位提供的大量无错误分类的贷款数据中学习到贷款五级分类规则，然后应用该规则到被审贷款数据集上，从而发现审计线索。

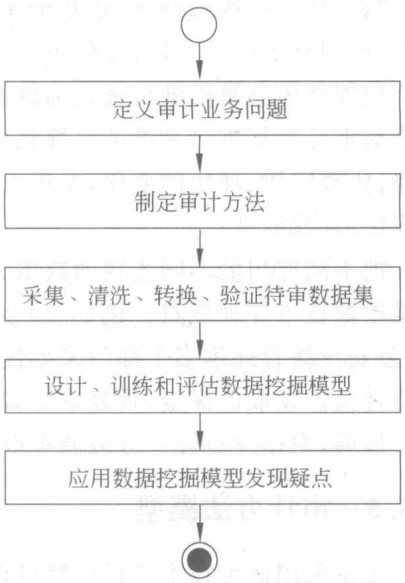


图1-7 数据挖掘途径的计算机审计过程

数据挖掘途径的计算机审计过程如图1-7所示。

“对某省某商业银行商业贷款进行五级分类真实性核实”就是一个审计目标。为了完成这个目标，首先采集、清洗和转换商业贷款以及五级分类情况的数据。因为这是数据挖掘中的“分类”问题，所以选择某种分类学习算法，如决策树算法，在具有分类标签的无错误分类的数据集上学习分类规则，并对学习到的分类规则进行评估。如果评估认为机器学习到的分类规则与实际的分类规则具有较高的一致性，则应用分类规则到可能具有错误分类的商业贷款事例上，即应用规则重新对这些商业贷款事例进行分类。重新分类得到类标签与原来的类标签不一致的事例就形成了审计疑点。

1.3.4 基于大数据的审计分析

审计数据获取、数据可视化技术、审计人员能力提升是当前审计行业共同关注的问题。如今，大数据环境为审计工作提供了总体数据，使得审计全覆盖成为可能。

大数据背景下，审计工作的展开是数据导向的，将会形成自顶向下、逐渐细化的工作模式：总体把握→发现疑点→分散核实。在大量数据中进行可视化分析，以把握总体、数据挖掘分析，以发现要点，从大量数据中发现审计重点、疑点，更加有针对性地开展现场审计，提升审计效率。

当数据集的规模超出了传统数据库软件的管理和分析能力，通常达到几十TB，甚至