第5章 审计数据预处理

本章学习目标

- 理解审计数据预处理的重要性。
- 理解数据质量、审计数据质量问题；掌握审计数据预处理的意义以及审计数据预处理的内容。
- 结合应用实例熟悉审计数据预处理的基本方法。
- 理解审计数据预处理阶段数据验证的重要性，熟悉审计数据预处理阶段数据验证的内容和方法。
- 了解一些大数据审计数据预处理方法。

5.1 概述

审计数据预处理是电子数据审计中的重要一环。目前，在审计数据采集过程中常常会遇到以下问题。

（1）审计不可能采集被审计单位的所有数据，在采集数据时，往往来不及对被审计单位的信息系统做详细的了解与分析，因此并不知道哪些数据重要，哪些数据不重要。通常是确定一个范围后把数据全部采集过来，再根据审计的需要进行整理和筛选。

（2）考虑到数据的全面和丰富，以及数据采集的风险，在采集数据时一般都宁多勿缺，故采集到的审计数据往往会有许多重复，且数据量巨大。

（3）采集到的被审计数据存在数据质量问题，在进行数据分析之前需要预处理，例如，有些数据属性的值不确定，在采集数据时，无法得到该数据属性的值，从而造成数据不完整。

由以上几点可以看出，由于被审计单位数据来源种类繁杂，采集到的数据存在一些数据质量问题，不能满足后续审计数据分析的需要。另外，这些问题的存在将直接影响后续审计工作所得出的审计结论的准确性。因此，完成审计数据采集后，审计人员必须对从被审计单位获得的原始电子数据进行预处理，从而使其满足后续审计数据分析的需要。

5.2 审计数据预处理理论分析

5.2.1 数据质量

1. 数据质量的概念

为了更好地理解审计数据预处理的必要性,本节首先介绍数据质量的相关概念。

目前,数据质量问题已引起广泛的关注。什么是数据质量呢?数据质量问题并不仅仅是指数据错误。有的文献把数据质量定义为数据的一致性(consistency)、正确性(correctness)、完整性(completeness)和最小性(minimality)这四个指标在信息系统中得到满足的程度,有的文献则把"适合使用"作为衡量数据质量的初步标准。

2. 数据质量评价指标

一般说来,评价数据质量最主要的几个指标如下。

1) 准确性

准确性(accuracy)是指数据源中实际数据值与假定正确数据值的一致程度。

2) 完整性

完整性(completeness)是指数据源中需要数值的字段中无值缺失的程度。

3) 一致性

一致性(consistency)是指数据源中数据对一组约束的满足程度。

4) 唯一性

唯一性(uniqueness)是指数据源中数据记录以及编码是否唯一。

5) 适时性

适时性(timeliness)是指在所要求的或指定的时间提供一个或多个数据项的程度。

6) 有效性

有效性(validity)是指维护的数据足够严格以满足分类准则的接受要求。

3. 可能存在的数据质量问题

当建立一个信息系统时,即使进行了良好的设计和规划,也不能保证在所有情况下信息系统中数据的质量都能满足用户的要求。用户录入错误、企业合并以及企业环境随着时间的推移而改变,这些都会影响所存放数据的质量。信息系统中可能存在的数据质量问题有很多种,总结起来主要有以下几种。

1) 重复的数据

重复的数据是指在一个数据源中存在表示现实世界同一个实体的重复信息,或在多个数据源中存在现实世界同一个实体的重复信息。

2) 不完整的数据

由于录入错误等原因,字段值或记录未被记入数据库,造成信息系统数据源中应该有的字段或记录缺失。

3）不正确的数据

由于录入错误，数据源中的数据未及时更新，或不正确的计算等原因，导致数据源中数据过时，或者一些数据与现实实体中字段的值不相符。

4）无法理解的数据值

无法理解的数据值是指由于某些原因，导致数据源中的一些数据难以解释或无法解释，如伪值、多用途域、古怪的格式、密码数据等。

5）不一致的数据

数据不一致包括了多种问题，例如，从不同数据源获得的数据很容易发生不一致；同一数据源的数据也会因位置、单位以及时间不同产生不一致。

在以上这些问题中，前三种问题在数据源中出现得最多。根据数据质量问题产生的原因，数据质量问题可分成单数据源问题和多数据源问题两个方面，其分类如图 5.1 所示。

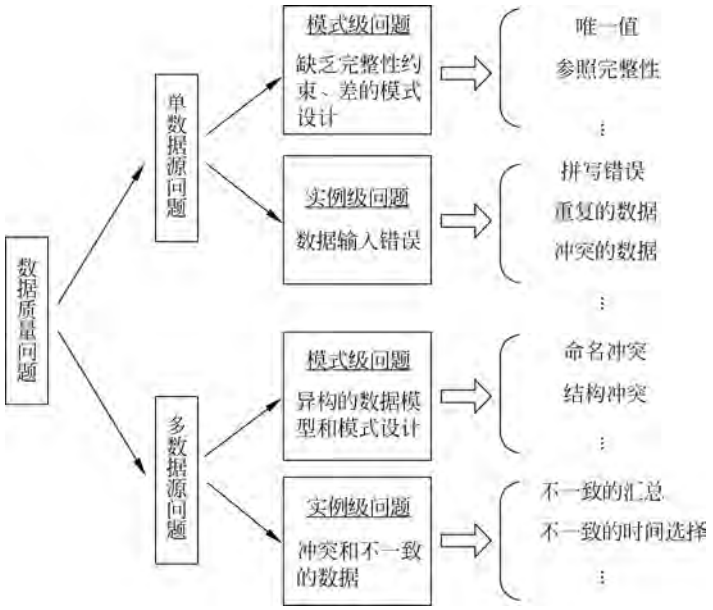


图 5.1　数据质量问题的分类

5.2.2　单数据源数据质量问题

单数据源数据质量问题可以分成模式级和实例级两类问题进行分析，如图 5.1 所示。一个数据源的数据质量很大程度上取决于控制这些数据的模式和完整性约束的等级。没有模式的数据源，例如文本文件数据，它对数据的输入和保存没有约束，于是出现错误和不一致的可能性就很大。因此，出现模式相关的数据质量问题是因为缺少合适的特定数据模型和特定的完整性约束，例如差的模式设计，或者因为仅定义了很少一些约束来进行完整性控制。特定实例问题相关错误和不一致，例如拼写错误，不能在模式级预防。另外，不唯一的模式级特定约束不能防止重复的实例，例如关于同一现实实体的记录可能会以不同的字段值输入两次。无论是模式级问题还是实例级问题，都可以分成字段、记录、记录类型和数据

源四种不同的问题范围,分别说明如下。

(1) 字段:这类错误仅局限于单个字段的值。

(2) 记录:这类错误表现在同一条记录中不同字段值之间出现的不一致。

(3) 记录类型:这类错误表现在同一个数据源中不同记录之间的不一致关系。

(4) 数据源:这类错误表现在数据源中的某些字段值和其他数据源中相关值的不一致关系。

四种不同情况的举例如表 5.1 和表 5.2 所示。

表 5.1 单数据源中模式级的数据质量问题

范围	问题	脏 数 据	原 因
字段	不合法值	出生日期=1970.13.12	字段值超出了域范围
记录	违反属性依赖	年龄=22,出生日期=1970.12.12	年龄=现在年-出生年
记录类型	违反唯一性	供应商 1:Name="新疆轴承总厂",No="G02002" 供应商 2:Name="西安汽车修配厂",No="G02002"	供应商编号不唯一
数据源	违反引用完整性	供应商:Name="新疆轴承总厂",City="102"	编号为 102 的城市不存在

表 5.2 单数据源中实例级的数据质量问题

范围	问题	脏 数 据	原 因
字段	空值	电话号码=(9999)999999	该值为缺省值,可能数值未输入或已丢失
	拼写错误	City="书州"	一般是数据录入错误
	含义模糊的值或缩写词	职位="DBProg."	不知道"DBProg."的意思
	多值嵌入	Name="西安汽车修配厂 710082 西安"	一个字段中输入了多个字段的值
	字段值错位	City="江苏"	某个字段的值输入另一个字段中
记录	违反属性依赖	City="南京",Zip="650093"	城市和邮政编码之间不匹配
记录类型	重复的记录	供应商 1:("西安汽车修配厂","西安",…) 供应商 2:("陕西省西安市汽车修配厂","西安",…)	由于数据输入错误,同一个供应商输入了两次
	冲突的值	供应商 1:("新疆轴承总厂","4",…) 供应商 2:("新疆轴承总厂","3",…)	同一个供应商被不同的值表示
数据源	引用错误	供应商:Name="新疆轴承总厂",City="12"	编号为 12 的城市存在,但该供应商不在这个城市

5.2.3 多数据源集成时数据质量问题

当多个数据源集成时,发生在单数据源中的这些问题会更加严重。这是因为每个数据源都是为了特定应用,单独开发、部署和维护的,这就很大程度上导致数据管理系统、数据模

型、模式设计和实际数据的不同。每个数据源都可能含有脏数据，多数据源中的数据可能会出现不同表示、重复、冲突等现象。

在模式级，模式设计的主要问题是命名冲突和结构冲突。命名冲突主要表现为不同的对象可能使用同一个命名，而同一对象可能使用不同的命名；结构冲突存在很多种不同的情况，一般是指在不同数据源中同一对象有不同表示，例如不同的组成结构、不同的数据类型、不同的完整性约束等。

除了模式级的冲突，很多冲突仅出现在实例级上，即数据冲突。由于不同数据源中数据的表示可能会不同，单数据源中的所有问题都可能会出现，例如重复的记录、冲突的记录等。此外，在整个数据源中，尽管有时不同的数据源中有相同的字段名和类型，仍可能存在不同的数值表示。例如，对性别的描述，一个数据源中可能用“0/1”来描述，另一个数据源中可能会用“F/M”来描述；或者对一些数值的不同表示，例如一个数据源中度量单位制可能用美元，另一个数据源中可能会用欧元。此外，不同数据源中的信息可能表示在不同的聚集级别上，例如一个数据源中信息可能指的是每种产品的销售量，而另一个数据源中信息可能指的是每组产品的销售量。

5.2.4 审计数据质量问题实例

为了便于理解审计数据的数据质量问题，以采集来的某税收征收电子数据（文件名为“税收征收.mdb”，数据表名为“征收表”，数据表结构见本书附录 A）为例，其可能存在的部分数据质量问题分析如下。

1. 不完整数据

在图 5.2 中，“实纳税额”字段中存在部分空值；在图 5.3 中，最后几条记录为空记录。空值并不等同于“0”，因而在进行数据分析时，不能参加如查询、筛选、汇总等数据分析，在审计数据分析过程中会被遗漏，所以必须对“征收表”中的空值进行处理。

2. 不一致的数据

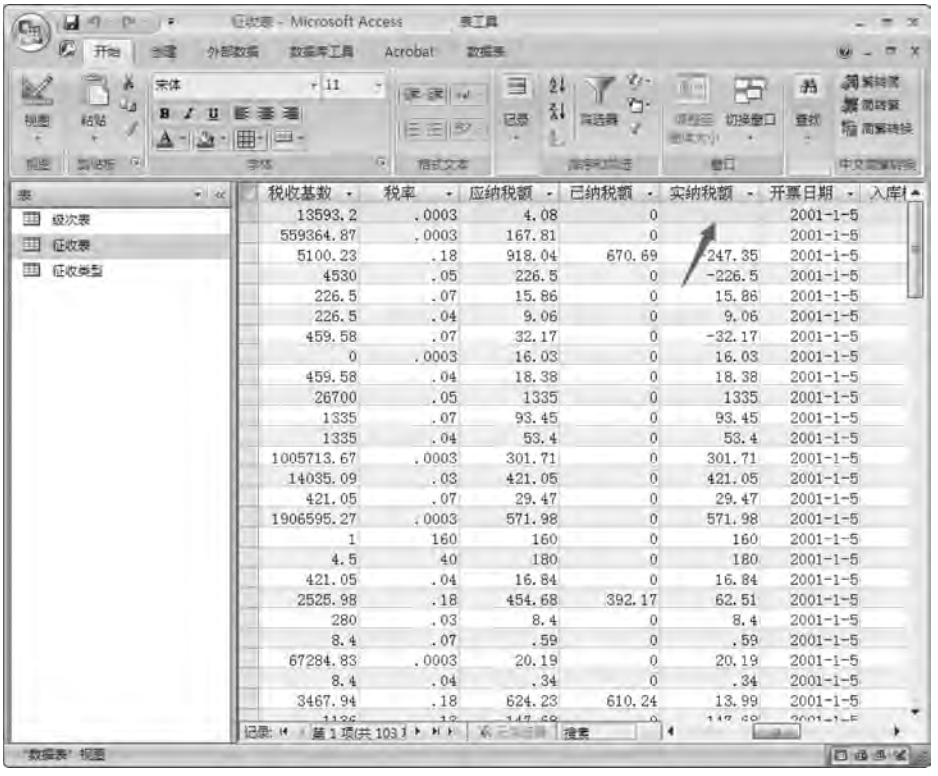
在图 5.4 中，“级次”字段中存在不一致的数据，即该字段中有的数据值为代码，有的数据值为实际的值，为方便后续的审计数据分析，需要转化成统一的格式来表示。

3. 不正确的数据

在图 5.5 中，“实纳税额”字段中有的数据值为负值，这些数据可能为错误的数值，为方便后续的审计数据分析，需要审计人员对该值进行确认，并对错误的数据值进行处理。

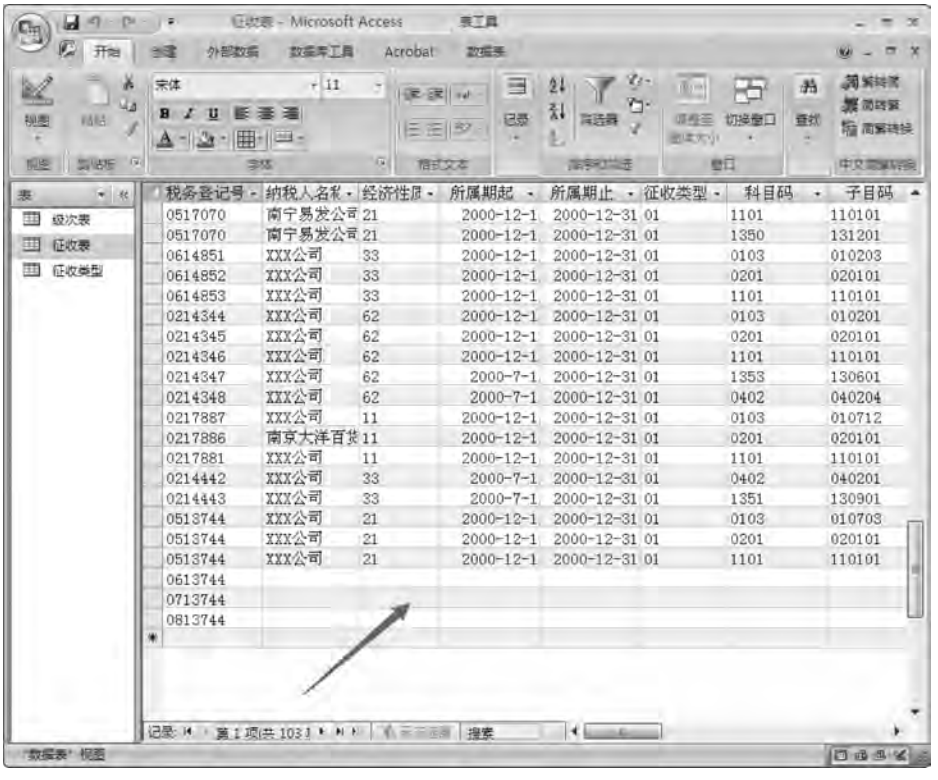
4. 重复的数据

在图 5.6 中，“税务登记号”为“0517070”数据存在多条，这些数据可能为重复的数据，为了保证审计数据分析结果的准确性，需要审计人员对这些重复的数据进行确认，找出造成数据重复的原因，并对重复的数据进行处理。



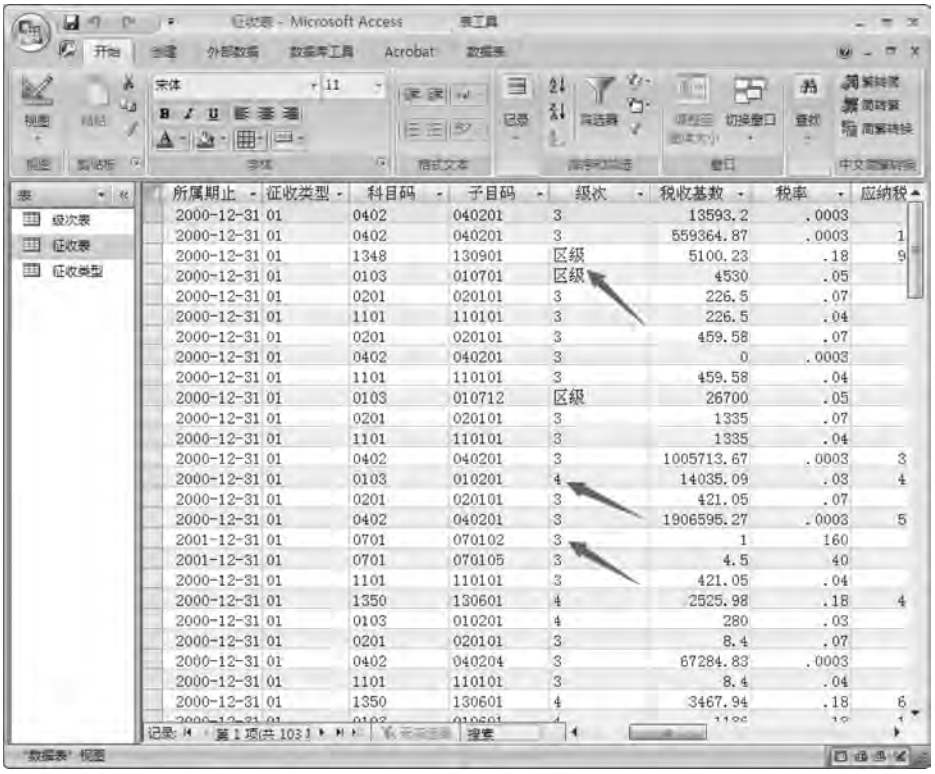
税收基数	税率	应纳税额	已纳税额	实纳税额	开票日期	入库
13593.2	.0003	4.08	0		2001-1-5	
559364.87	.0003	167.81	0		2001-1-5	
5100.23	.18	918.04	670.69	247.35	2001-1-5	
4530	.05	226.5	0	-226.5	2001-1-5	
226.5	.07	15.86	0	15.86	2001-1-5	
226.5	.04	9.06	0	9.06	2001-1-5	
459.58	.07	32.17	0	-32.17	2001-1-5	
0	.0003	16.03	0	16.03	2001-1-5	
459.58	.04	18.38	0	18.38	2001-1-5	
26700	.05	1335	0	1335	2001-1-5	
1335	.07	93.45	0	93.45	2001-1-5	
1335	.04	53.4	0	53.4	2001-1-5	
1005713.67	.0003	301.71	0	301.71	2001-1-5	
14035.09	.03	421.05	0	421.05	2001-1-5	
421.05	.07	29.47	0	29.47	2001-1-5	
1906595.27	.0003	571.98	0	571.98	2001-1-5	
1	160	160	0	160	2001-1-5	
4.5	40	180	0	180	2001-1-5	
421.05	.04	16.84	0	16.84	2001-1-5	
2525.98	.18	454.68	392.17	62.51	2001-1-5	
280	.03	8.4	0	8.4	2001-1-5	
8.4	.07	.59	0	.59	2001-1-5	
67284.83	.0003	20.19	0	20.19	2001-1-5	
8.4	.04	.34	0	.34	2001-1-5	
3467.94	.18	624.23	610.24	13.99	2001-1-5	

图 5.2　字段中存在空值数据质量问题的税收征收数据



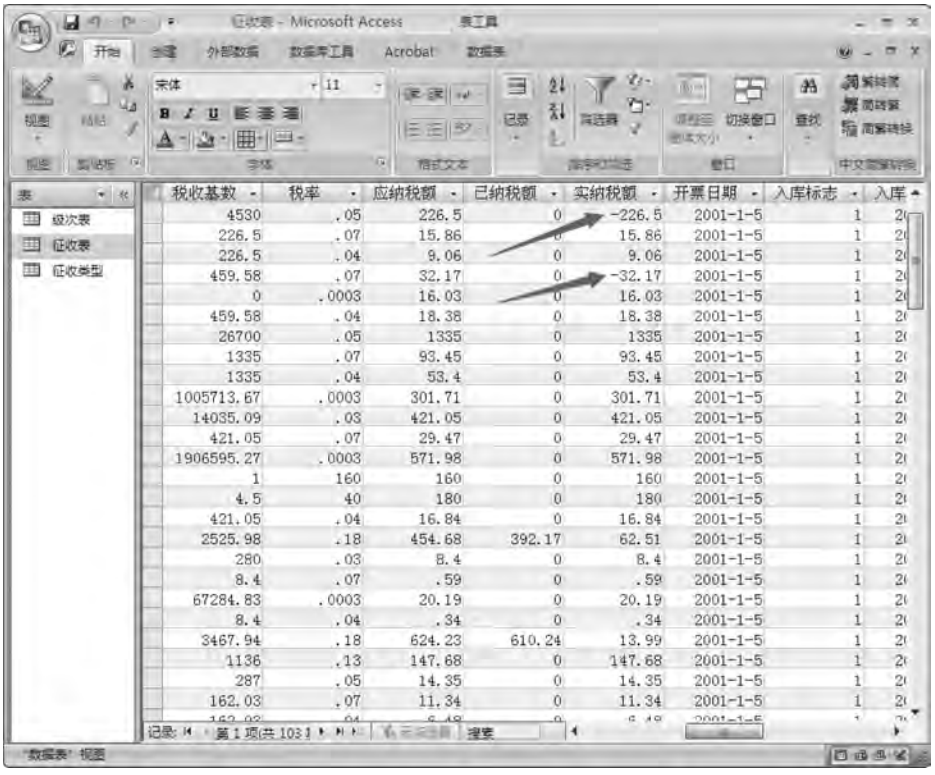
税务登记号	纳税人名称	经济性质	所属期起	所属期止	征收类型	科目码	子目码
0517070	南宁易发公司	21	2000-12-1	2000-12-31	01	1101	110101
0517070	南宁易发公司	21	2000-12-1	2000-12-31	01	1350	131201
0614851	XXX公司	33	2000-12-1	2000-12-31	01	0103	010203
0614852	XXX公司	33	2000-12-1	2000-12-31	01	0201	020101
0614853	XXX公司	33	2000-12-1	2000-12-31	01	1101	110101
0214344	XXX公司	62	2000-12-1	2000-12-31	01	0103	010201
0214345	XXX公司	62	2000-12-1	2000-12-31	01	0201	020101
0214346	XXX公司	62	2000-12-1	2000-12-31	01	1101	110101
0214347	XXX公司	62	2000-7-1	2000-12-31	01	1353	130601
0214348	XXX公司	62	2000-7-1	2000-12-31	01	0402	040204
0217887	XXX公司	11	2000-12-1	2000-12-31	01	0103	010712
0217886	南京大洋百货	11	2000-12-1	2000-12-31	01	0201	020101
0217881	XXX公司	11	2000-12-1	2000-12-31	01	1101	110101
0214442	XXX公司	33	2000-7-1	2000-12-31	01	0402	040201
0214443	XXX公司	33	2000-7-1	2000-12-31	01	1351	130901
0513744	XXX公司	21	2000-12-1	2000-12-31	01	0103	010703
0513744	XXX公司	21	2000-12-1	2000-12-31	01	0201	020101
0513744	XXX公司	21	2000-12-1	2000-12-31	01	1101	110101
0613744							
0713744							
0813744							

图 5.3　记录中存在空记录数据质量问题的税收征收数据



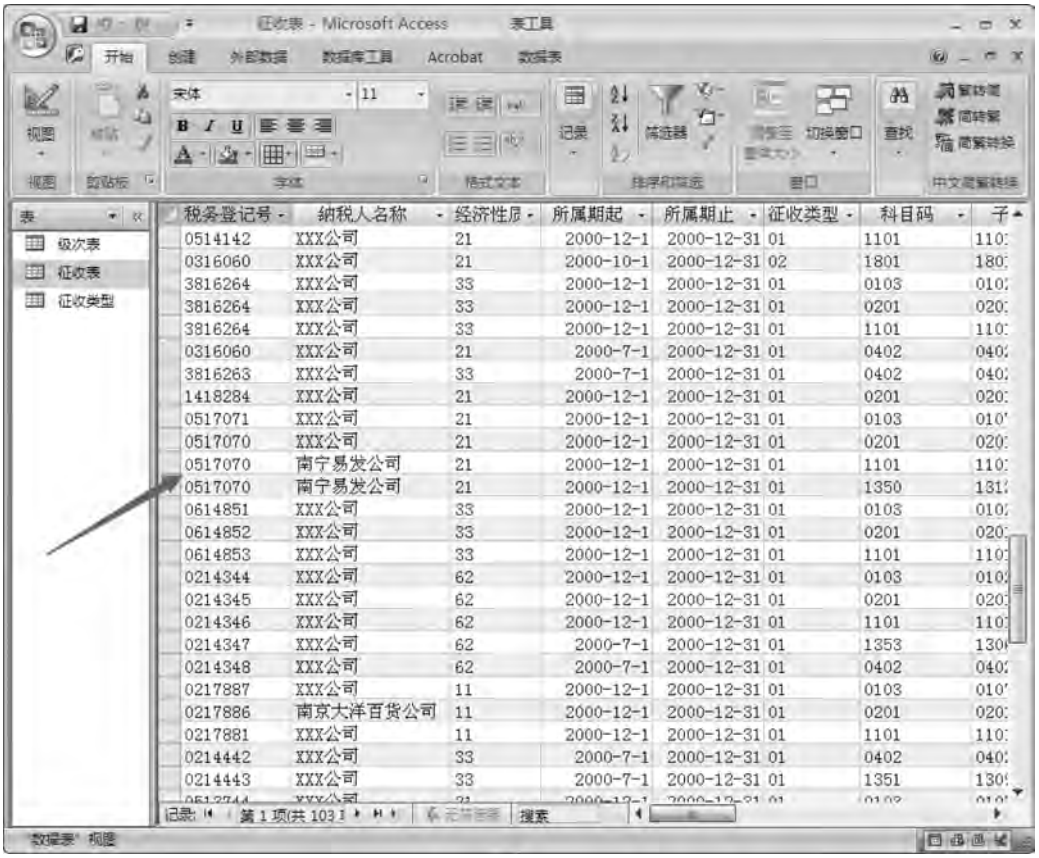
所属期止	征收类型	科目码	子目码	级次	税收基数	税率	应纳税
2000-12-31	01	0402	040201	3	13593.2	.0003	
2000-12-31	01	0402	040201	3	559364.87	.0003	1
2000-12-31	01	1348	130901	区级	5100.23	.18	9
2000-12-31	01	0103	010701	区级	4530	.05	
2000-12-31	01	0201	020101	3	226.5	.07	
2000-12-31	01	1101	110101	3	226.5	.04	
2000-12-31	01	0201	020101	3	459.58	.07	
2000-12-31	01	0402	040201	3	0	.0003	
2000-12-31	01	1101	110101	3	459.58	.04	
2000-12-31	01	0103	010712	区级	26700	.05	
2000-12-31	01	0201	020101	3	1335	.07	
2000-12-31	01	1101	110101	3	1335	.04	
2000-12-31	01	0402	040201	3	1005713.67	.0003	3
2000-12-31	01	0103	010201	4	14035.09	.03	4
2000-12-31	01	0201	020101	3	421.05	.07	
2000-12-31	01	0402	040201	3	1906595.27	.0003	5
2001-12-31	01	0701	070102	3	1	160	
2001-12-31	01	0701	070105	3	4.5	40	
2000-12-31	01	1101	110101	3	421.05	.04	
2000-12-31	01	1350	130601	4	2525.98	.18	4
2000-12-31	01	0103	010201	4	280	.03	
2000-12-31	01	0201	020101	3	8.4	.07	
2000-12-31	01	0402	040204	3	67284.83	.0003	
2000-12-31	01	1101	110101	3	8.4	.04	
2000-12-31	01	1350	130601	4	3467.94	.18	6

图 5.4　存在不一致数据质量问题的税收征收数据



税收基数	税率	应纳税额	已纳税额	实纳税额	开票日期	入库标志	入库
4530	.05	226.5	0	-226.5	2001-1-5	1	2(
226.5	.07	15.86	0	15.86	2001-1-5	1	2(
226.5	.04	9.06	0	9.06	2001-1-5	1	2(
459.58	.07	32.17	0	-32.17	2001-1-5	1	2(
0	.0003	16.03	0	16.03	2001-1-5	1	2(
459.58	.04	18.38	0	18.38	2001-1-5	1	2(
26700	.05	1335	0	1335	2001-1-5	1	2(
1335	.07	93.45	0	93.45	2001-1-5	1	2(
1335	.04	53.4	0	53.4	2001-1-5	1	2(
1005713.67	.0003	301.71	0	301.71	2001-1-5	1	2(
14035.09	.03	421.05	0	421.05	2001-1-5	1	2(
421.05	.07	29.47	0	29.47	2001-1-5	1	2(
1906595.27	.0003	571.98	0	571.98	2001-1-5	1	2(
1	160	160	0	160	2001-1-5	1	2(
4.5	40	180	0	180	2001-1-5	1	2(
421.05	.04	16.84	0	16.84	2001-1-5	1	2(
2525.98	.18	454.68	392.17	62.51	2001-1-5	1	2(
280	.03	8.4	0	8.4	2001-1-5	1	2(
8.4	.07	.59	0	.59	2001-1-5	1	2(
67284.83	.0003	20.19	0	20.19	2001-1-5	1	2(
8.4	.04	.34	0	.34	2001-1-5	1	2(
3467.94	.18	624.23	610.24	13.99	2001-1-5	1	2(
1136	.13	147.68	0	147.68	2001-1-5	1	2(
287	.05	14.35	0	14.35	2001-1-5	1	2(
162.03	.07	11.34	0	11.34	2001-1-5	1	2(

图 5.5　存在不正确数据质量问题的税收征收数据



税务登记号	纳税人名称	经济性质	所属期起	所属期止	征收类型	科目码	子
0514142	XXX公司	21	2000-12-1	2000-12-31	01	1101	110
0316060	XXX公司	21	2000-10-1	2000-12-31	02	1801	180
3816264	XXX公司	33	2000-12-1	2000-12-31	01	0103	010
3816264	XXX公司	33	2000-12-1	2000-12-31	01	0201	020
3816264	XXX公司	33	2000-12-1	2000-12-31	01	1101	110
0316060	XXX公司	21	2000-7-1	2000-12-31	01	0402	040
3816263	XXX公司	33	2000-7-1	2000-12-31	01	0402	040
1418284	XXX公司	21	2000-12-1	2000-12-31	01	0201	020
0517071	XXX公司	21	2000-12-1	2000-12-31	01	0103	010
0517070	XXX公司	21	2000-12-1	2000-12-31	01	0201	020
0517070	南宁易发公司	21	2000-12-1	2000-12-31	01	1101	110
0517070	南宁易发公司	21	2000-12-1	2000-12-31	01	1350	131
0614851	XXX公司	33	2000-12-1	2000-12-31	01	0103	010
0614852	XXX公司	33	2000-12-1	2000-12-31	01	0201	020
0614853	XXX公司	33	2000-12-1	2000-12-31	01	1101	110
0214344	XXX公司	62	2000-12-1	2000-12-31	01	0103	010
0214345	XXX公司	62	2000-12-1	2000-12-31	01	0201	020
0214346	XXX公司	62	2000-12-1	2000-12-31	01	1101	110
0214347	XXX公司	62	2000-7-1	2000-12-31	01	1353	130
0214348	XXX公司	62	2000-7-1	2000-12-31	01	0402	040
0217887	XXX公司	11	2000-12-1	2000-12-31	01	0103	010
0217886	南京大洋百货公司	11	2000-12-1	2000-12-31	01	0201	020
0217881	XXX公司	11	2000-12-1	2000-12-31	01	1101	110
0214442	XXX公司	33	2000-7-1	2000-12-31	01	0402	040
0214443	XXX公司	33	2000-7-1	2000-12-31	01	1351	130

图 5.6 存在重复数据质量问题的税收征收数据

5.2.5 审计数据预处理的意义

由以上分析可知,正是由于采集到的被审计数据中存在上述数据质量问题,所以需要对采集到的电子数据进行预处理,处理有数据质量问题的数据,为后续的审计数据分析打下基础。概括起来,进行审计数据预处理的意义如下。

1. 为下一步的审计数据分析提供准备

采集到的被审计数据不一定能完全满足审计数据分析的需要,因此,通过对有质量问题的被审计数据进行预处理,从而为后续的审计数据分析做好准备。

2. 帮助发现隐藏的审计线索

通过对被审计数据进行数据预处理,可以有效地发现被审计数据中不符合数据质量的数据,但是,审计人员不能简单地把有质量问题的数据删除,因为这些存在质量问题的数据中可能隐藏着审计线索。需要做的是:对发现的审计数据质量问题进行分析,找出造成数据质量问题的原因,发现隐藏的审计线索。

3. 降低审计风险

有质量问题的被审计数据会影响审计数据分析结果的正确性，造成一定的审计风险。因此，通过对有质量问题的审计数据进行数据预处理，从而降低审计风险。

4. 通过更改命名方式使数据便于数据分析

通过名称转换这一审计数据预处理操作，可以把采集到的数据表以及字段名称转换成直观的名称，便于审计人员的审计数据分析。

5.2.6　审计数据预处理的内容

根据审计工作的实际情况，审计数据预处理可简单地分成数据转换和数据清理两部分内容。

1. 数据转换

简单地讲，数据转换就是把具有相同或相近意义的各种不同格式的数据转换成审计人员需要的格式相对统一的数据，或把采集到的原始数据转换成审计人员容易识别的数据格式和容易理解的名称，如名称转换、数据类型转换、代码转换等。

2. 数据清理

数据清理也称数据清洗(data cleaning)。简单地讲，数据清理就是利用相关技术，如数理统计、数据挖掘或预定义的清理规则等，从数据中检测和消除错误数据、不完整数据和重复数据等，从而提高数据的质量。数据清理的原理可总结为如图 5.7 所示。

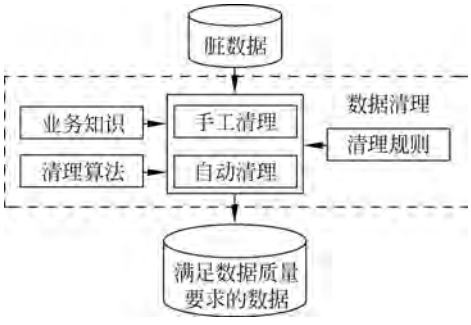


图 5.7　数据清理原理

5.3　审计数据预处理应用实例

目前，根据一般审计人员的技术能力和审计工作中的具体要求，并考虑到审计数据预处理方法的经济性和可操作性，一般进行的审计数据预处理内容包括名称转换、数据类型转换、代码转换、横向合并、纵向合并、空值处理等。例如，通过名称转换这一审计数据预处理

操作,可以把采集到的数据表以及字段名称转换成直观的名称,便于审计人员的审计数据分析;同样,其他的审计数据预处理也是便于审计人员的审计数据分析。常用的一些数据库产品也可以完成审计数据预处理功能,本节通过实例详细介绍如何使用 Access 2007 和 SQL Server 2008 来完成审计数据预处理。

5.3.1 基于 Access

本节以名称转换和空值处理为例,详细介绍如何使用 Access 来完成审计数据预处理。

1. 名称转换

在大多数情况下,采集到的被审计数据的命名并不直观,为了便于审计人员进行数据分析,需要对数据表和字段的名称进行调整。例如,采集到的数据表名称和字段名称有时采用拼音的缩写表示,这时如果将其转换成汉字表示,则便于审计人员进行审计数据分析。使用 Access 完成数据表名称和字段名称转换的操作如图 5.8 和图 5.9 所示。

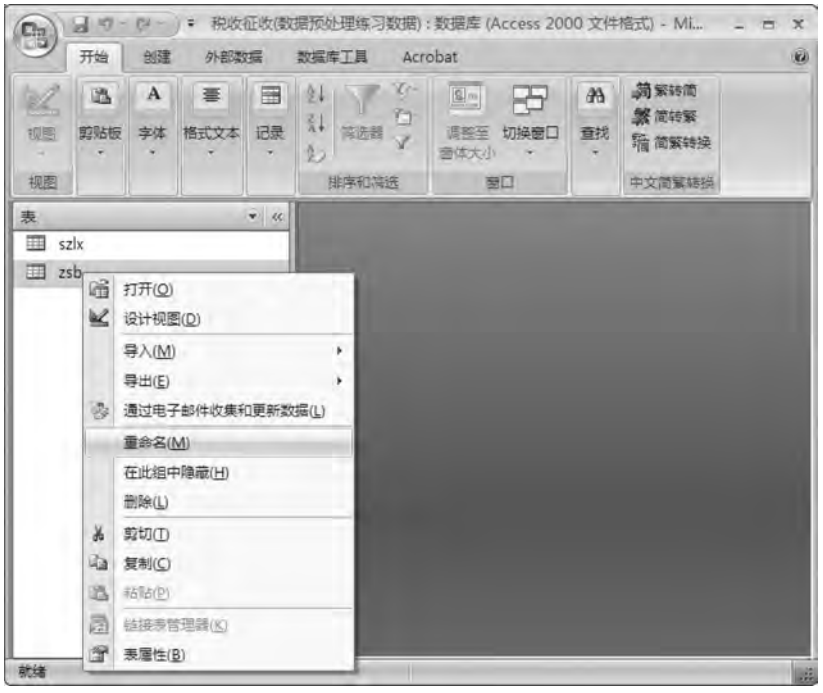


图 5.8 数据表名称转换操作实例

2. 空值处理

如 5.2 节所述,采集到的被审计数据中经常会出现一些“空值”,“空值”是字段的一种特殊状态,在数据库中用一个特殊的值 NULL 来表示,意味着该字段不包含任何数据,它不同于零值和空白。由于空值参与任何运算的结果都是空值,所以会对审计数据分析带来一些不便之处,因此,在审计数据预处理阶段,审计人员需要根据实际情况对空值数据进行处理。在实际操作中,审计人员可以使用 Access 来完成空值处理。

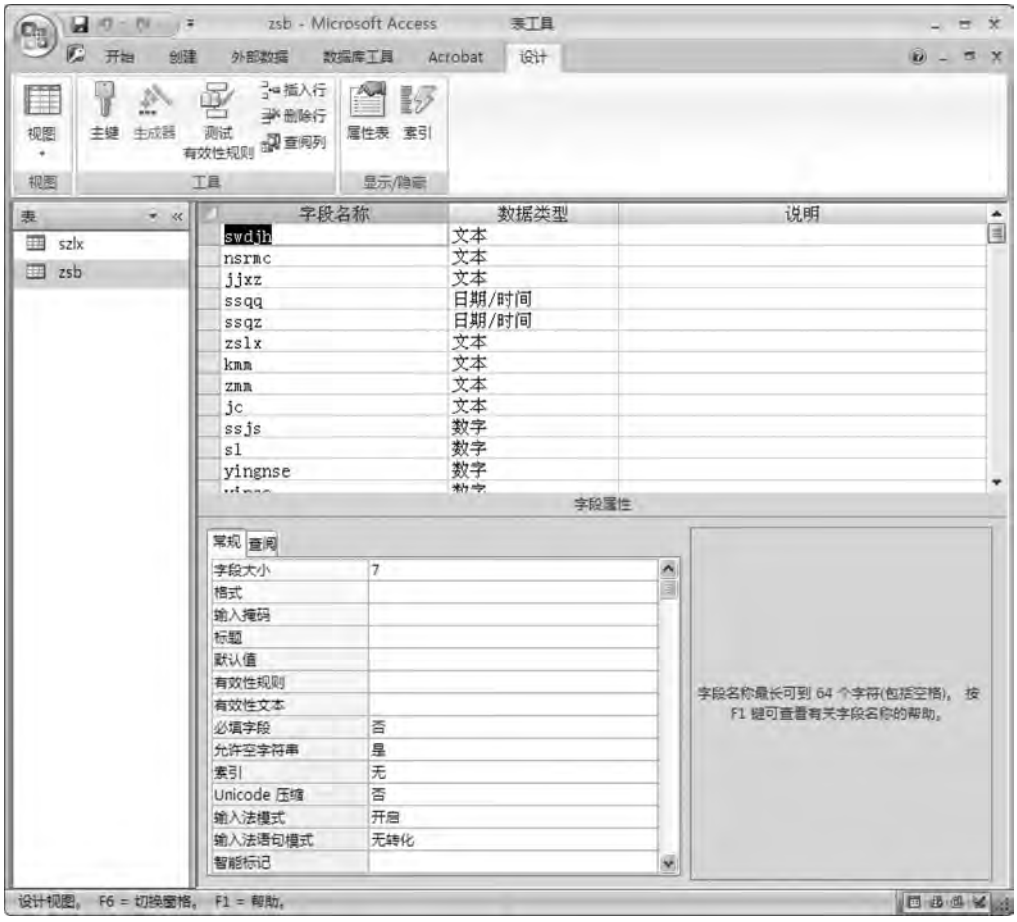


图 5.9　字段名称转换操作实例

例 5.1　某税收征收数据的空值处理

现有某税收征收电子数据(文件名为“税收征收.mdb”,数据表名为“征收表”),表结构见附录 A。假设已完成数据表名称转换和字段名称转换,要求对其进行审计数据预处理,把“征收表”中“实纳税额”字段中的空值变成“0”。

通过对税收征收电子数据的分析,对“实纳税额”字段进行空值处理的 SQL 语句如下。

```
UPDATE 征收表
SET 实纳税额 = 0
WHERE 实纳税额 Is Null;
```

通过运行以上 SQL 语句,可以很容易地把“征收表”中“实纳税额”字段中的空值设置成“0”。下面介绍如何使用 Access 来执行以上 SQL 语句,完成空值处理。

假设税收征收数据已被采集到 Access 中,如图 5.10 所示。

完成税收征收数据中“实纳税额”字段空值处理的操作步骤如下。

(1) 在 Access 中单击“创建”选项卡,如图 5.11 所示。

(2) 在图 5.11 中选择“查询设计”命令,然后单击“确定”按钮,弹出如图 5.12 所示“显示表”对话框。

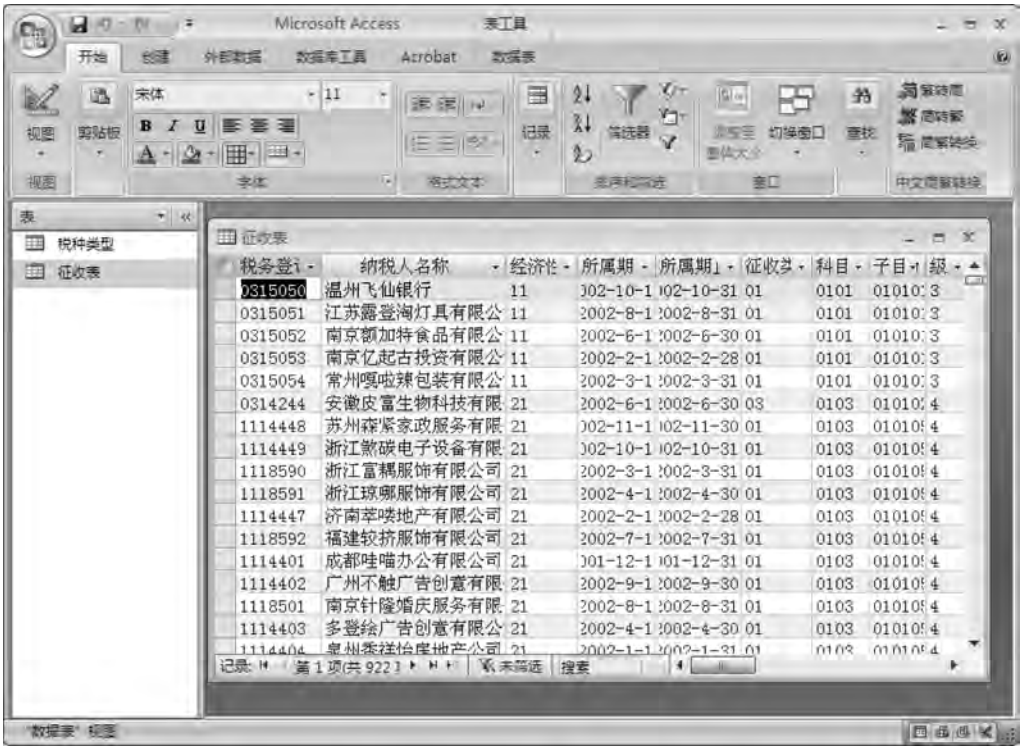


图 5.10 含有税收征收数据的 Access 数据库

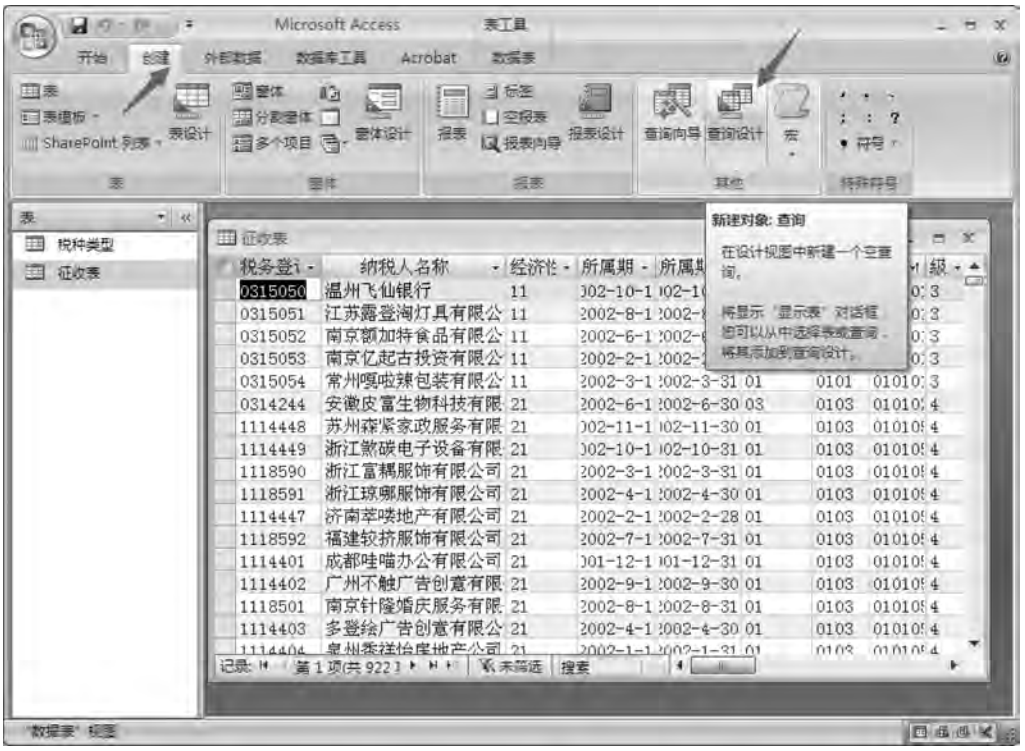


图 5.11 选择新建查询的类型

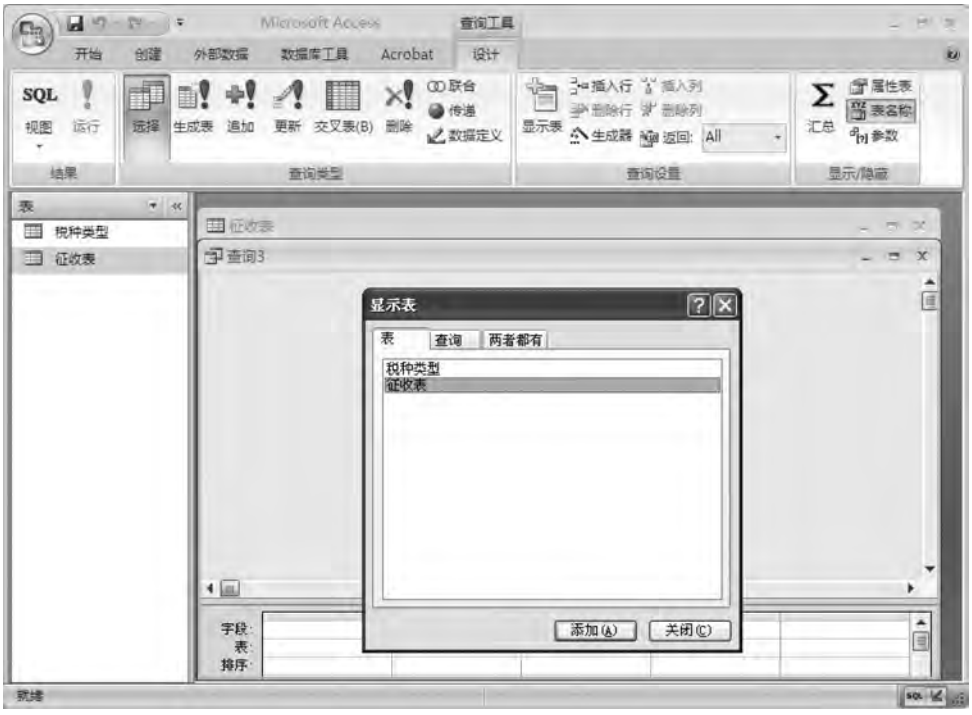


图 5.12　选择查询的对象

(3) 在图 5.12 中单击“关闭”按钮,弹出如图 5.13 所示窗口。

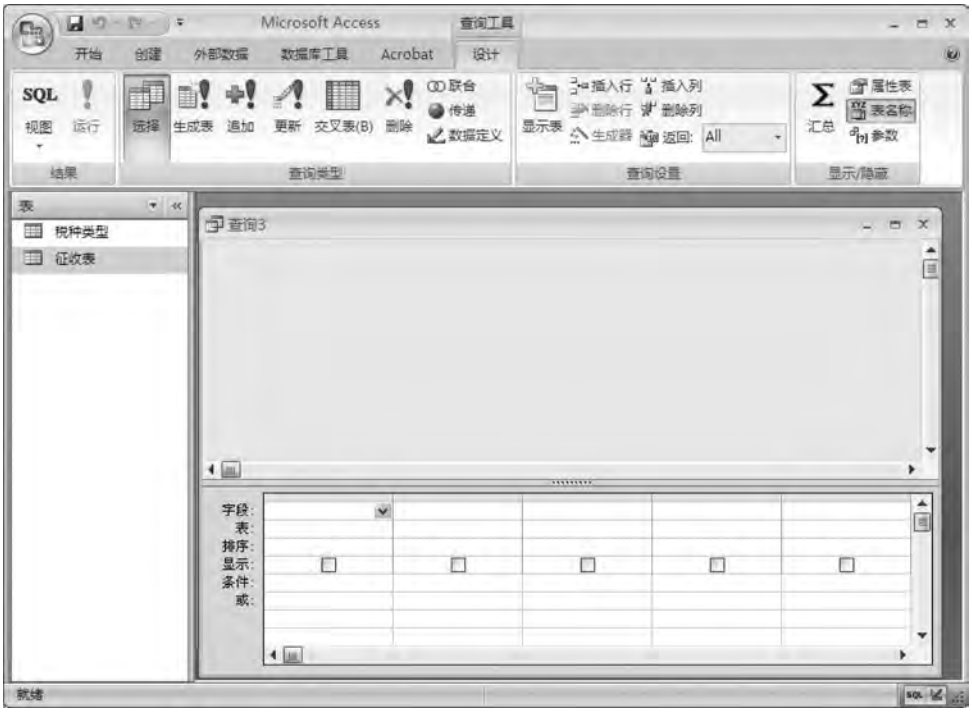


图 5.13　Access 的设计视图

(4) 切换到“SQL 视图”,其操作如图 5.14 所示。

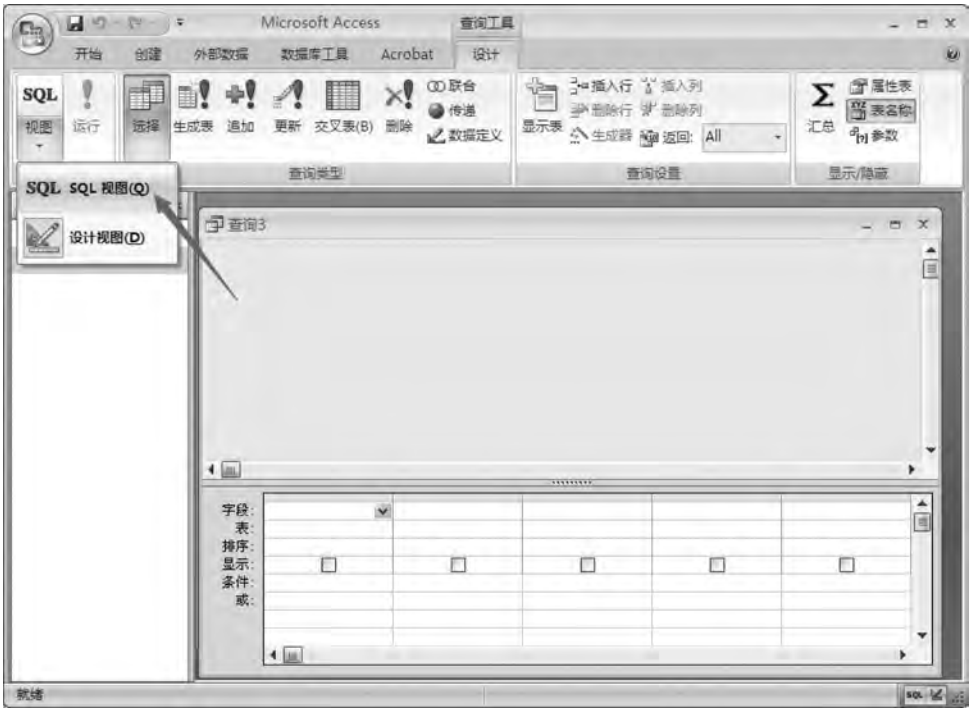


图 5.14 SQL 视图切换菜单

(5) 在图 5.14 中输入相应的 SQL 语句,如图 5.15 所示。

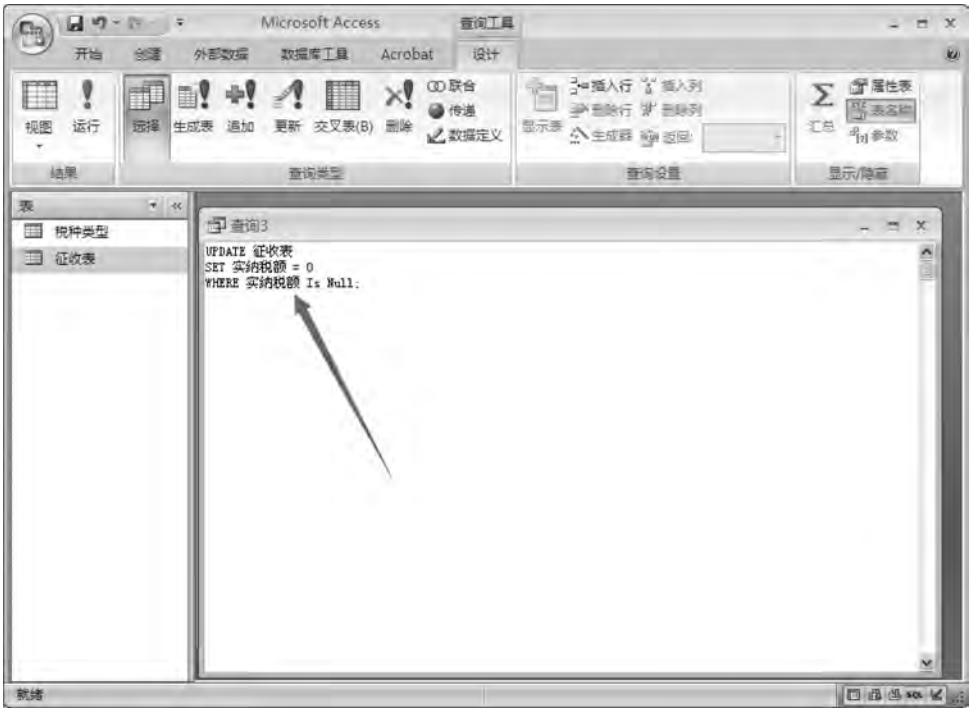


图 5.15 输入 SQL 语句之后的 SQL 视图

（6）在图 5.15 中单击“运行”按钮，则“实纳税额”字段中空值处理的结果如图 5.16 所示。

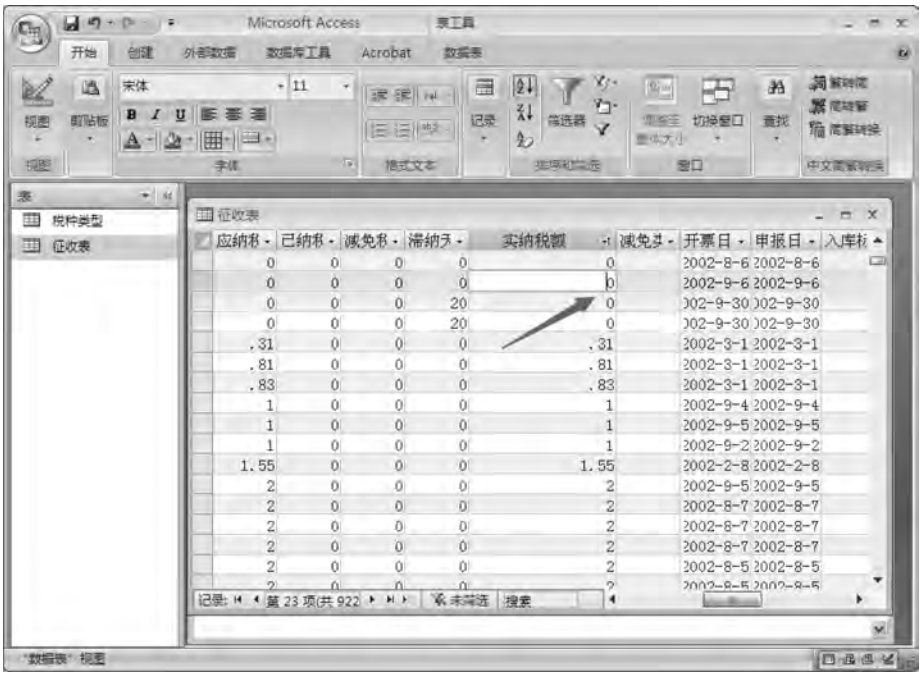


图 5.16　空值处理的结果界面

如果审计人员对 SQL 语句不熟练，也可以在“设计视图”中选择、输入相关参数，完成空值处理，如图 5.17 所示。

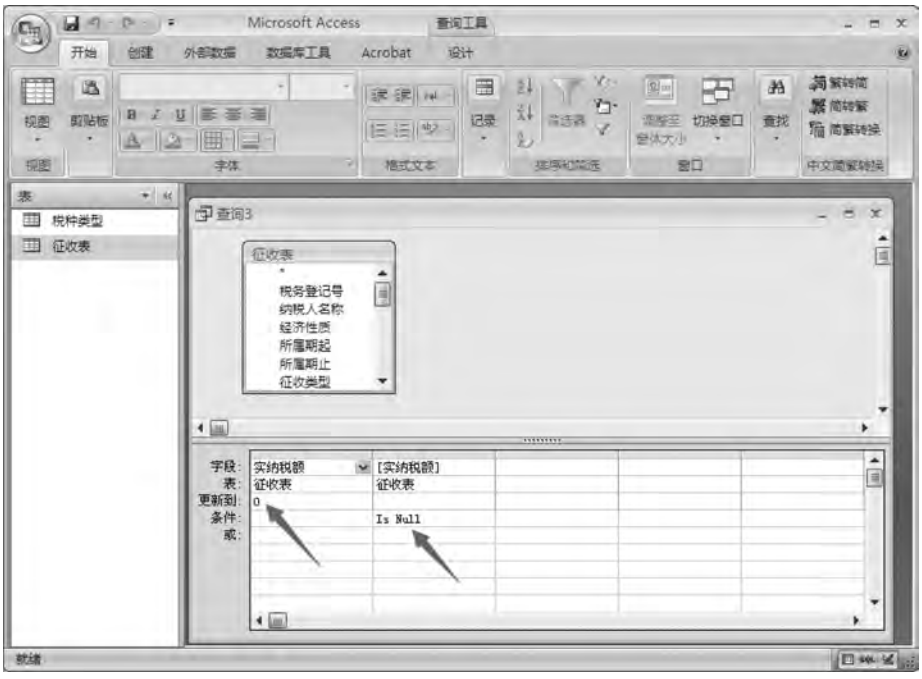


图 5.17　在设计视图中设置处理条件

5.3.2 基于 SQL Server

本节以名称转换和空值处理为例,详细介绍如何使用 SQL Server 完成审计数据预处理。

1. 名称转换

假设被审计数据已被采集到 SQL Server 中,使用 SQL Server 完成数据表名称转换的操作如图 5.18 所示。使用 SQL Server 完成字段名称转换的操作分别如图 5.19 和图 5.20 所示。

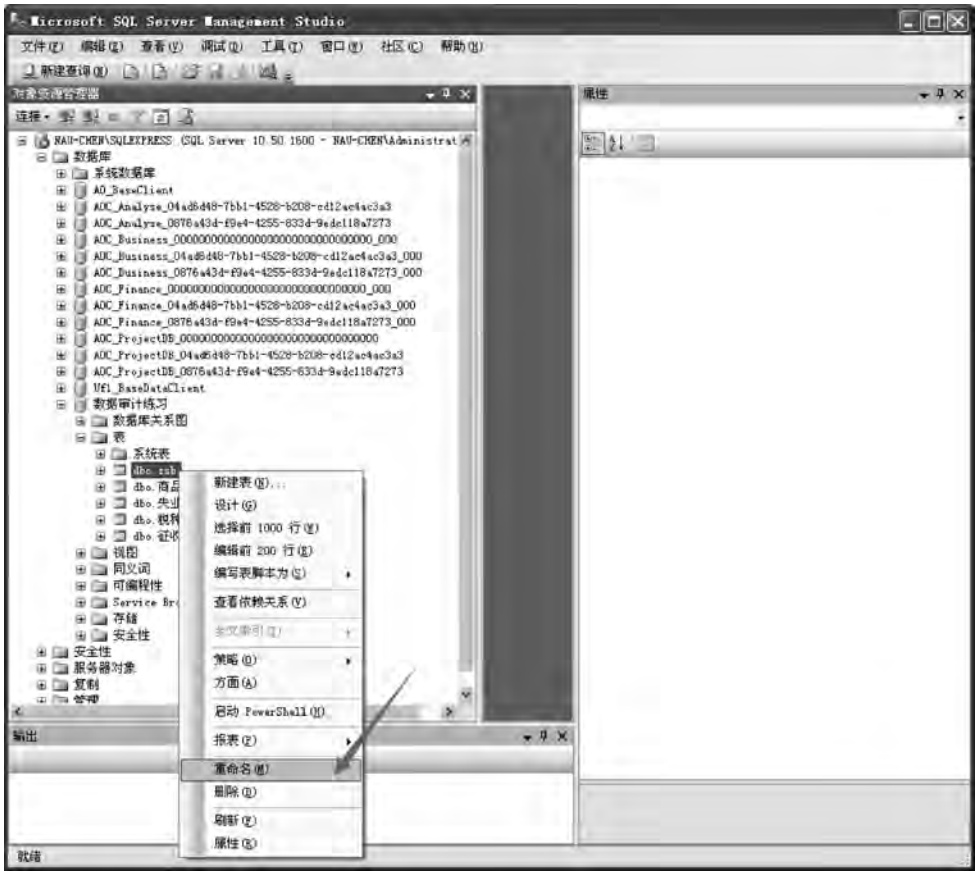


图 5.18 数据表名称转换操作实例

2. 空值处理

在实际操作中,审计人员可以使用 SQL Server 来完成空值处理。假设税收征收数据已被采集到 SQL Server 中,在完成数据表名称和字段名称的转换的基础上,单击“新建查询”,在查询界面中输入相应的空值处理 SQL 语句,如图 5.21 所示,即可完成税收征收数据中“实纳税额”字段空值处理。

图 5.19　进入字段设计界面

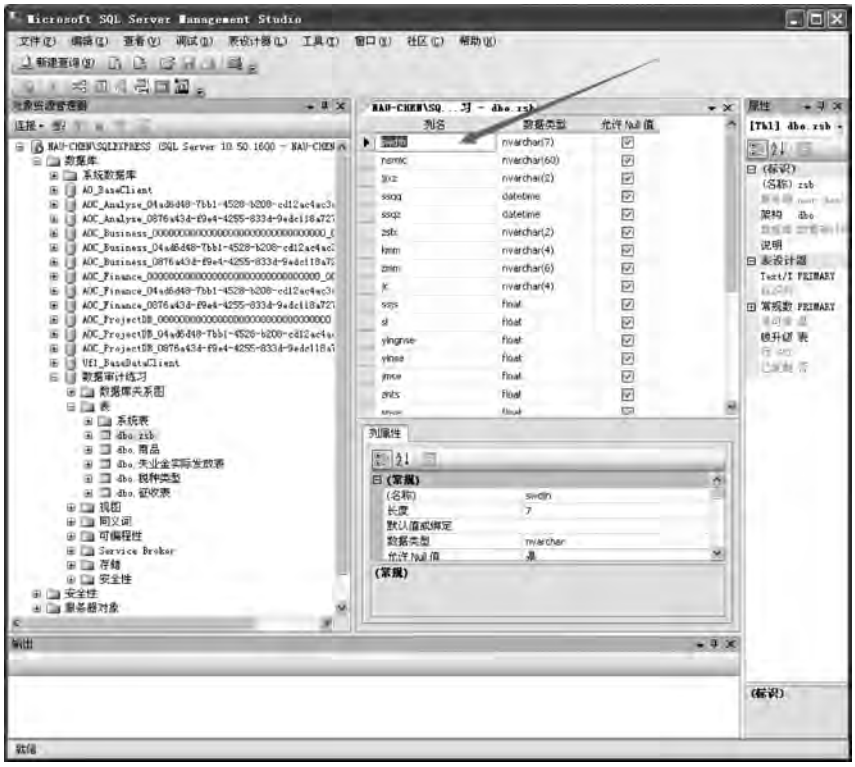


图 5.20　字段名称转换操作实例

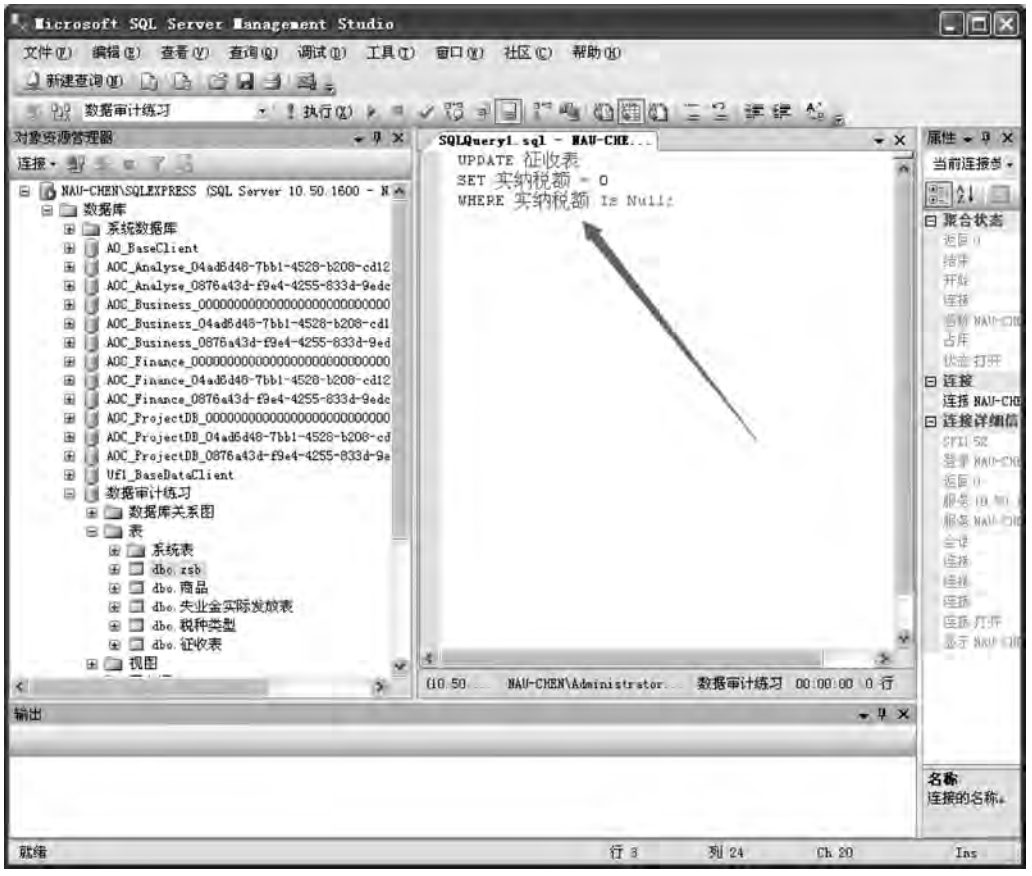


图 5.21 空值处理操作实例

同理,使用 SQL 语句或数据库工具中的其他工具也可以完成其他审计数据预处理工作。

5.4 大数据预处理方法简介

面对大数据环境下需要预处理的审计数据,本节对一些高效、自动的数据预处理方法进行简单介绍,以供进行审计数据预处理操作时参考。

5.4.1 不完整数据清理

在采集数据时,由于无法得到一些数据属性的值,从而造成数据的不完整。为了满足审计数据分析的需要,要对数据源中的不完整数据进行清理,不完整数据清理的原理如图 5.22 所示。

不完整数据清理的主要步骤说明如下。

1. 不完整数据检测

要清理数据源中的不完整数据,首先要做的就是把数据源中的不完整数据检测出来,以

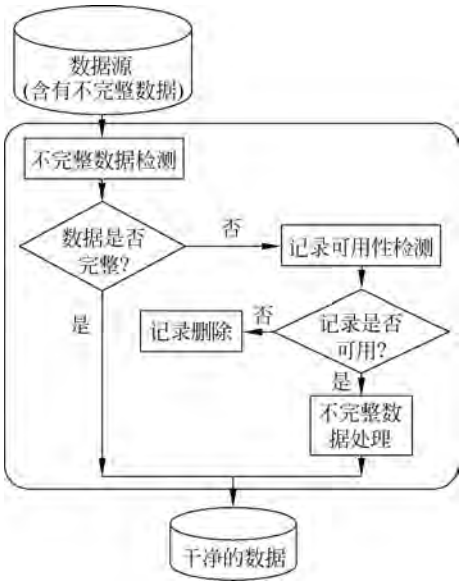


图 5.22 不完整数据清理的原理

便于下一步的处理,不完整数据检测就是完成这一工作。

2. 数据可用性检测

数据可用性检测是不完整数据清理过程中的一个重要步骤。如果一条记录属性值丢失得太多,或者剩余的属性值中根本就不包含关键信息,就没有必要花费精力去补全该记录。因此,要解决数据的不完整问题,判断记录的可用性非常重要。判断记录的可用性就是根据每一条记录的不完整程度以及其他因素,来决定这些记录是保留还是删除。对于记录的可用性检测,一般采用的方法如下。

先评估一条记录的不完整程度,也就是先计算一条数据记录中丢失属性值的字段的百分比,再考虑其他因素,例如数据记录剩余的属性值中关键信息是否存在,然后决定记录的取舍。由于当一条记录某属性取值为缺省值时,意味着该属性值已丢失,所以,一般把属性值为缺省值的也作为丢失值来处理。评估一条记录不完整程度的方法如下。

假设一条记录可表示成:

$$R=\{a_1,a_2,\cdots,a_n\}$$

$a_1,a_2,\cdots,a_n$ 表示记录 R 的 n 个属性,$R_i(a_j)$ 表示记录 R_i 第 j 个属性 a_j 的值,$a_j(\text{default})$ 表示记录第 j 个属性 a_j 的默认值,m 表示记录 R 中属性值丢失的数目(包括属性值取默认值的字段),AMR 表示记录 R 中属性值丢失的比率,λ 为记录 R 中属性值丢失比率的阈值,如果:

$$\text{AMR}=\frac{m}{n}<\lambda,\quad \lambda\in[0,1]$$

则表示该记录比较完整,应保留记录 R;否则,删除记录 R。

在进行不完整数据清理时,λ 的值由域专家根据对具体数据源的分析来确定其取值,并定义在规则库中,供系统调用。数据表中各个属性 a_j 的默认值 $a_j(\text{default})$ 也定义在规则

库中,供系统计算 m 值时调用。

此外,在决定记录取舍时,除了评估每一条记录的不完整程度外,有时还需要考虑该记录中关键的属性值是否存在,关键属性要由域专家根据对具体数据源的分析来确定。如果不完整数据中关键属性值存在,即使 AMR>λ,也应该保留记录。需要指出的是,在删除数据时一定要慎重。

3. 不完整数据处理

不完整数据处理是指在完成数据可用性检测之后,对那些要保留的不完整数据记录,要采取一定的方法来处理该记录中缺失的属性值,一般采取以下几种处理方法。

1) 人工处理法

对于一些重要数据,或当不完整数据的数据量不大时应该采用这种方法。

2) 常量替代法

常量替代法就是对所有缺失的属性值用同一个常量来填充,例如用“Unknown”或“Miss Value”,这种方法较为简单,但是由于所有的缺失值都被当成同一个值,容易导致错误的分析结果。

3) 平均值替代法

平均值替代法就是使用一个属性的平均值来填充该属性的所有缺失值。

4) 最常见值替代法

最常见值替代法就是使用一个属性中出现最多的那个值来填充该属性的所有缺失值。

5) 估算值替代法

估算值替代法是最复杂,也是最科学的一种处理方法,采用这种方法处理缺失的属性值过程为:首先采用相关算法,如回归、判定树归纳等算法预测该属性缺失值的可能值,然后用预测值填充缺失值。

以上给出了常用的几种处理记录中缺失属性值的方法,至于在执行不完整数据的清理过程中采用什么样的处理方法,要根据具体的被审计数据源来确定。

5.4.2 相似重复记录清理

1. 相似重复记录清理的原理

为了减少采集到的电子数据中的冗余信息,相似重复记录清理是一项重要任务。相似重复记录是指那些客观上表示现实世界同一实体,但由于在格式、拼写上有些差异而导致数据库系统不能正确识别的记录。相似重复记录清理的原理如图 5.23 所示。

相似重复记录清理的过程可描述如下。

首先,把数据源中需要清理的数据调入系统中;然后,执行数据清理,记录排序模块从算法库中调用排序算法,执行记录之间的排序。在记录已排序的基础上,记录相似检测模块从算法库中调用相似检测算法,作邻近范围内记录间的相似检测,从而计算出记录间的相似度,并根据预定义的重复识别规则,来判定是否为相似重复记录。为了能检测到更多的重复记录,一次排序不够,要采用多轮排序,多轮比较,每轮排序采用不同的键,然后把检测到的

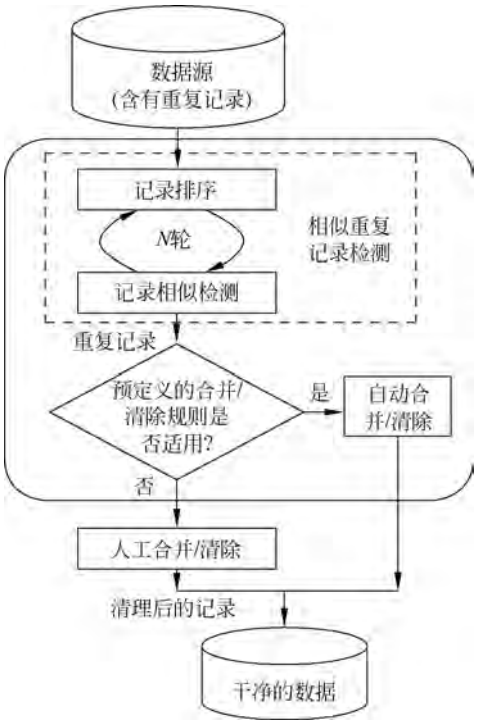


图 5.23　相似重复记录清理的原理

所有相似重复记录聚类到一起，从而完成相似重复记录的检测；最后，对所检测出的每一组相似重复记录根据预定义的合并/清除规则，完成相似重复记录的合并处理。

2. 相似重复记录清理的关键步骤

由图 5.23 中可以看出，相似重复记录清理的关键步骤可总结为：记录排序→记录相似检测→相似重复记录合并/清除，其作用分别说明如下。

1）记录排序

为了能查找到数据源中所有的相似重复记录，必须比较每一个可能的记录对，如此一来，检测相似重复记录是一个很烦琐的操作。当采集的电子数据的量很大时，这会导致是一个无效和不可行的方案。为了减少记录之间的比较次数，提高检测效率，常用的方法是仅比较相互距离在一定范围的记录，即先对数据表中的记录排序，然后对邻近记录进行比较。

2）记录相似检测

记录相似检测是相似重复记录清理过程中的一个重要步骤，通过记录相似检测，可以判断两条记录是不是相似重复记录。

3）相似重复记录合并/清除

当完成相似重复记录的检测之后，对检测出的相似重复记录要进行处理。对于一组相似重复记录，一般有两种处理方法。

（1）把一组相似重复记录中的一条记录看成是正确的，其他记录看成是含有错误信息的相似重复记录。于是，任务就是删除数据库中的相似重复记录。在这种情况下，一些常用

的处理规则如下。

① 人工规则。人工规则是指由人工从一组相似重复记录中选出一条最准确的记录保留,并把其他相似重复记录从数据库中删除,这种方法较为简单。

② 随机规则。随机规则是指从一组相似重复记录中随机地选出一条记录保留,并把其他相似重复记录从数据库中删除。

③ 最新规则。在很多情况下,最新的记录能更好地代表一组相似重复记录。例如,越接近当前日期的信息准确性可能越高,经常使用账户上的地址要比不常使用的账户上的地址权威一些。基于这种分析,最新规则是指选择每一组相似重复记录中最新的一条记录保留,并把其他相似重复记录从数据库中删除。

④ 完整规则。完整规则是指从一组相似重复记录中选择最完整的一条记录保留,并把其他相似重复记录从数据库中删除。

⑤ 实用规则。因为重复率越高的信息可能越相对准确,例如,如果三条记录中两个供应商的电话号码是相同的,那么重复的电话号码可能是正确的。基于这种分析,实用规则是指从一组相似重复记录中选择与其他记录匹配次数最多的一条记录保留,并把其他相似重复记录从数据库中删除。

(2) 把每一条相似重复记录看成是信息源的一部分。于是,目的就是合并一组相似重复记录,产生一个具有更完整信息的新记录。该方法一般要由人工进行处理。

5.4.3 PDF 格式文件转换成文本格式文件

大数据环境下,不仅需要分析结构化数据,还需要分析非结构化数据。为了便于对 PDF 格式的文件进行分析,可以把 PDF 格式文件转换成文本格式文件。

例 5.2 把 PDF 格式文件转换成文本格式文件

现有某 PDF 格式文件数据,如图 5.24 和图 5.25 所示。现需要将该数据转换成文本格式文件,文件命名为“基于大数据可视化技术的审计线索特征挖掘方法研究.txt”,并保存到计算机 F 盘中。

图 5.24 文件夹中需要转换的 PDF 文件

图 5.25　需要转换的 PDF 文件示例

以 R 语言为例，实现代码如下。

```
install.packages("pdftools")
library(pdftools)
txt <- pdf_text("F:\\大数据审计实验\\基于大数据可视化技术的审计线索特征挖掘方法研究.pdf")
write.csv(txt,file = "F:\\大数据审计实验\\基于大数据可视化技术的审计线索特征挖掘方法研究.txt")
```

在 RStudio 中运行以上代码，如图 5.26 所示。

图 5.26　基于 RStudio 的转换操作示例

运行结果如图 5.27 和图 5.28 所示。

图 5.27 文件夹中转换后的文本文件示例

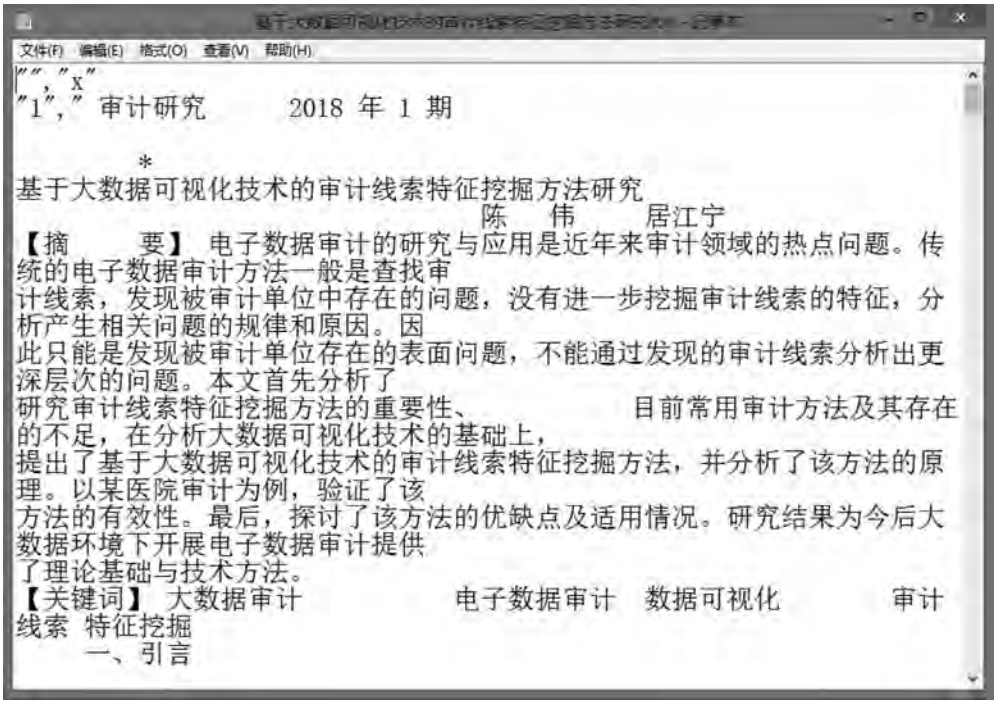


图 5.28 转换后的文本文件示例

5.5 审计数据预处理阶段的数据验证

5.5.1 审计数据预处理阶段数据验证的重要性

在开展电子数据审计的过程中，审计人员必须不断进行数据验证，以保证电子数据的真

实性、正确性和完整性。在审计数据预处理过程中，审计人员会将原始电子数据中表名、字段名、记录值代码以及表与表关联的经济含义明确标识出来，这需要进行大量的数据查询、数据修改、数据删除、数据添加等操作；另外，要对电子数据进行错误数据处理、空值处理、相似重复数据处理、不一致数据处理等操作，以提高审计数据质量，为下一步的审计数据分析做好准备。在审计数据预处理过程中可能存在一些问题，举例如下。

(1) 目标数据模式设计不合理。

(2) 审计数据预处理方法不当。

(3) 审计数据预处理工具使用不合适。

(4) 审计数据预处理过程不规范，没有日志记录。

根据以上分析，每一步审计数据预处理工作都有可能影响到审计数据的完整性和正确性，所以在这一阶段进行数据验证也是很有必要的。

5.5.2 审计数据预处理阶段数据验证的内容和方法

1. 数据验证的主要内容

在这一阶段，数据验证主要是确认上述审计数据预处理工作没有损害数据整体的完整性，保证审计数据的正确性。对审计数据预处理过程进行验证主要包含以下两方面内容。

1) 确认审计数据预处理的目标实现

为了确认审计数据预处理的目标得以实现，必须针对转换前存在的数据质量问题和转换要求逐一进行核对。

2) 确认审计数据预处理工作没有损害数据的完整性和正确性

要确认审计数据预处理工作没有损害数据的完整性和正确性，就必须确认审计数据预处理过程中没有带来新的错误。

2. 数据验证的方法

在审计数据预处理阶段，审计人员可以根据实际情况，采用核对总金额、保持借贷平衡、钩稽关系、审计抽样等数据验证方法来完成审计数据验证。

思考题

1. 为什么需要对被审计数据进行审计数据预处理？
2. 什么是数据质量？常见审计数据质量问题有哪些？
3. 如何对被审计数据进行数据预处理？
4. 大数据环境下审计数据预处理方法有哪些变化？
5. 数据预处理阶段为什么也需要数据验证？