统计数据的批量快速审核及审核引擎实现

周晓云　（徐州师范大学计算机科学与技术学院 221008）

摘要：为了保证统计调查采集的数据的有效性，对统计表内部的各项指标进行审核是行之有效的办法。而如何对大量的统计数据进行快速审核是一个难题。我们使用逻辑表和物理表的概念，通过 cursor 技术，优化的公式执行引擎技术非常满意地解决批量审核的响应时间要求。

关键词：统计数据　批量审核　公式　公式引擎

1　统计数据的结构和特点

统计可以分为一次性调查和周期性的统计，比如家庭收入构成调查是一次性统计，而工业企业的财务状况年度统计则是周期性统计。不管是一次性的调查或者周期性的统计，统计者希望从被统计调查对象获得足够的原始信息，以此作为进一步加工和处理的基础，那么就必须设计合理的表格，把每个被统计对象的信息填写在上边。

在表格里，根据指标之间的关系，我们可以把指标分为 3 类：

(1) 1 维指标：1 维指标是一系列的被统计对象的单个的指标，这些指标是平面式地展开的，比如被调查者的年龄、性别、文化程度等。

(2) 2 维指标：2 维指标在调查统计表里面表现为一个 2 维子表。比如表 1 就是一个 2 维子表。

这个 2 维的表格是对某个省份的不同地区的企业科技投入调查，2 维表的左边表示需要对每个地区的企业类型紧进行分类，同时还有不同企业类型的合计，2 维表的上边表示需要对科技投入的来源组成进行划分，分为政府基金和企业基金，并且求合计。

表 1　各地区科研资金投入调查表(2 维指标实例)

2 维表名称(d01)	合计 (01)		
		政府基金 (02)	企业基金 (03)
大型企业 (a01)	30 =10 +20	10	20
中型企业 (a02)	70 =30 +40	30	40
合计 (a03)	100(总计)	40 =10 +30	60 =20 +40

我们分别对该表的行和列进行命名，比如行的名称是 a01、a02、a03，而列的名称是 01、02、03，则各个单元格的计算关系可以简单表示为(单元格的命名请参考第 3 部分指标的命名)：

d01. [a01, 01] = d01. [a01, 02] + d01. [a01, 03]：表示大型企业合计数等于政府投入大型企业基金和大型企业自身投入基金的合计；

d01. [a02, 01] = d01. [a02, 02] + d01. [a02, 03]：表示中型企业合计数等于政府投入中型企业基金和中型企业自身投入基金的合计；

d01. [a03,02] = d01. [a02,02] + d01. [a01, 02]：表示政府投入基金数等于政府投入大型企业基金与政府投入中型企业基金的合计；

d01. [a03,03] = d01. [a02,03] + d01. [a01, 03]：表示企业自身投入基金数等于大型企业自身投入基金与中型企业自身投入基金的合计；

d01. [a03,01] = d01. [a03,02] + d01. [a03, 03]以及 d01. [a03,01] = [a02,01] + d01. [a01, 01]：表示对总计的审核关系。

(3) 3 维以上的指标：3 维以上的指标简单的说是表格中套子表格。这里举一个简单的例子，比如我们要调查某个地区的家庭状况。在调查表格的设计里，我们希望了解每个家庭的成员的基本情况，家庭成员可能有多个，每个都有基本情况，所以在表格里面包含 2 维以上的指标，同时我们希望家庭成员的基本情况包括姓名、性别、年龄和工作情况，而对于某个家庭成员来讲，他(她)的工作经历可能有多条，对于每个条目，我们需要了解工作起止时间、工作地点、工作内容、月薪等信息。这就需要在基本情况子表格里面套工作

情况子表。如下面的表 2 所示。

表 2　家庭成员基本情况与工作情况调查表
（3 维指标实例）

家庭成员	姓名	性别	年龄	工作情况				
成员 1				工作	时间	地点	内容	薪水
				工作 1				
				工作 2				
成员 2				工作	时间	地点	内容	薪水
				工作 1				
				工作 2				

整个统计表格里面，可能包含 1 维指标、2 维指标、3 到多维的指标，也可能都包含。每类指标之间具有内部的关系，比如全年总产量等于各个月份产量的总和等，为了保证统计调查的有效性，必须对统计表格里面的各类指标的相互关系进行确认，这就是审核。

计算机具有强大的信息处理能力，统计行业的信息化是大势所趋。我们设计并实现了统计信息收集和汇总系统，并且成功解决了大量统计数据的快速批量审核问题。

2　批量审核的实现

各个被统计调查对象的统计数据统一地放置在一个数据库里。由于统计表格的指标可能很多（超过 256 个），那么这些统计数据是被分配在数据库的不同表格里面的，也就是一个逻辑表格可能对应多个数据库物理表，如图 1 所示。

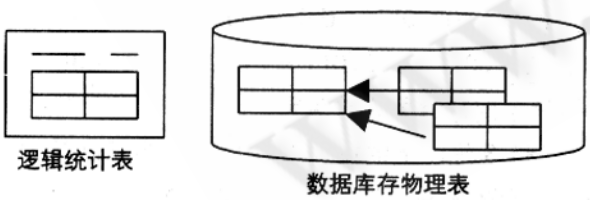

图 1　逻辑表和物理表的关系

一个逻辑表对应一个主物理表，其它的物理表通过主键与之进行关联。

为了快速地对数据库里众多单位的统计数据进行审核，我们设计了如下的算法（参阅图 2）：

（1）首先打开某个逻辑表对应的所有物理表的 cursor[2]；

（2）从各个 cursor 读取一行或者多行，在内存里面装载某个单位的逻辑统计表；

（3）执行审核的所有公式；

（4）登记审核出错的信息。

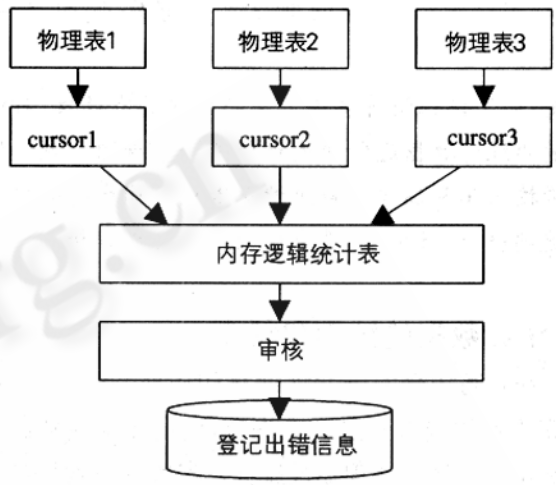

图 2　批量审核流程

3　审核执行引擎的实现

为了对统计数据进行审核，我们设计了对各类指标的命名方式。对于 1 维的指标，我们用一个英文字母开头的标识来表示，比如 a01，a02 等等。对于 2 维指标，我们首先给 2 维子表的行和列进行命名，比如行命名为 a01，a02，列命名为 b01，b02，同时对该 2 维子表给一个名字比如 d02。那么 2 维子表的一个单元格指标记为 d01.［a01，b01］，一行指标记为 d01.［*，b01］，一列指标记为 d01.［a01，*］，整个 2 维子表的所有指标记为 d01.［*，*］，3 维以及 3 维以上的子表的命名方式和 2 维子表类似，可以精确地命名一个单元格、一行、一列、多行、多列、整个集合等。仍然以各个地区科研资金投入调查表为例子，命名见表 3。

表 3　指标的命名

d01 表格名称	合计 01		
		政府基金 02	企业基金 03
大型企业 a01	d01.［a01,01］	d01.［a01,02］	d01.［a01,03］
中型企业 a02	d01.［a02,01］	d01.［a02,02］	d01.［a02,03］
合计 a03	d01.［a03,01］	d01.［a03,02］	d01.［a03,03］

审核公式的作用是审核表格内单元格的计算关系是否符合业务要求，比如审核公式 d01.[a01, 01] = d01.[a01, 02] + d01.[a01, 03]，表示大型企业合计数等于政府投入大型企业基金和大型企业自身投入基金的合计。

为了加强审核公式的计算审核能力，我们设计的审核公式执行引擎，可以对六种数据类型（布尔型 bool，整型 int，浮点型 float，短文本（小于 256 字符）text，长文本 memo，以及日期时间 datetime）进行四则运算（+，-，*，/，%，^）、关系运算（>，>=，<，<=，==，!=）和逻辑运算（not and or），还支持包括数学函数、字符串函数和日期时间函数在内的 40 多个函数，值得一提的是，我们还实现了从其它逻辑表或者从上期数据中进行摘抄的公式。

这样的功能设计，完全可以满足统计业务中对统计报表内数据的关系的审核要求。

审核公式执行引擎的结构[1]如图 3 所示。

（1）文法分析器：根据 1、2、3 维和多维指标的命名规则，分析指标名称，分析各种运算符号和函数，并且把分析的一系列符号传递给语法分析器。

（2）语法分析器：根据函数和运算的优先级别生成语法树。

（3）语义检查模块：检查操作数类型是否匹配，函数的参数个数是否符合定义等。

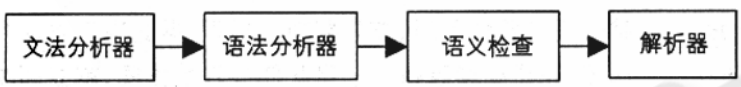

图 3　公式执行引擎

（4）解析器：负责对语法树进行解释执行。在执行过程中，重要的活动之一是，不断地从内存逻辑表中取得指标的具体的值。由于我们可以对逻辑表中的 1、2、3 维和多维指标进行计算，数据的交换接口包含两个参数，一个是指标的名称，一个是指标的行列子集合。具体如下：

```
IdataExchanger
{
    void getValueObject ( CnameObject& noTemp,
CvalueObject& voTemp);
    void setValueObject ( CnameObject& noTemp,
CvalueObject& voTemp);
};
```

由于指标的数据类型可以取 bool，int，float，text，memo，datetime 等六种类型，所以我们在实现行列子集时，实现了通用类型 Cvariant，而行列子集就是 Cvariant 的 2 维数组。

通用数据类型 Cvariant 在运算方面会比较慢，为了适应统计数据主要是 int 和 float 等数值，我们针对数值计算重新设计了行列子集，使得数值行列子集的计算有了很大的提高。如果行列子集的所有单元都是 int 型或者 float 型数值，那么行列子集用一个 double 数组内部表示，否则用一个通用类型 Cvariant 数组来内部表示。

4　实验结果

我们对统计数据的批量审核进行了实验，取得了满意的效果。

实验的硬件环境是：

CPU	内存	硬盘	操作系统	数据量（表格与字段数量）	数据量（总量 MB）
Pentium 4 2.0G	256M	60G	Windows 2000 Server	一个逻辑表（1005 行） 包含 3 个物理表（28140 行） 数据项或字段数量是 377880	24.2MB

我们使用一个逻辑表，包含 15 个 1 维指标，两个 2 维子表，其中一个 2 维子表包含 20 * 17 个单元格，另外一个 2 维子表包含是 7 * 3 个单元格。被调查单位是 1005 个。其中一张最大的物理表的行数是 20010 行。

审核公式总共有 25 个，其中涉及算术运算、逻辑运算和 sum()，if()等函数计算。

在上述配置的机器上，执行如此规模的统计数据审核，用时 3.15 秒，基本满足了统计业务的实际应用。

参考文献

1　杜淑敏编著，《编译程序设计原理》，北京大学出版社，1990。

2　袁鹏飞编著，《SQL Server 数据库应用开发技术》，人民邮电出版社，1998。

3　http://www.netscape.com