统计数据查询系统的设计与实现

何荣刚,陈启安
(厦门大学 信息科学与技术学院,福建 厦门 361005)

摘 要:以统计机构日常统计数据为基础,对原始数据进行清理整合,构建一个满足系统需求的数据库。在系统中创建包括跨年度、跨行业指标关联的指标管理体系,提出以指标为基础、法人单位和时间为限定条件并可以附加多种过滤条件的查询方式,使统计查询更具灵活性和多样性,查询结果以表格形式展示,具有清晰明确的特点,并且后期处理更加容易。

关键词:统计数据;指标关联;查询
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)02-0187-03

Framing and Implementation of Statistical Data Query System

HE Rong-gang, CHEN Qi-an
(School of Information Science and Technology, Xiamen University, Xiamen 361005, China)

Abstract: Take the statistical agency daily statistical data as a foundation, cleaning up the primary data to construct the database which satisfy the system requirements. Create framework which includes index associate by years and index associate by industry. Proposed a method of query based on index, limited corporation and time, also can add filtrate condition if it is needed. It makes statistical query more easier and multiplicity. The query result using the tabular form has the clear explicit characteristic, and the later period processes is easier.

Key words: statistical data; index associate; query

0 引 言

随着我国经济的迅速发展,统计工作中的统计数据越来越复杂,数据量越来越庞大,为提高统计效率,加快统计行业的信息化建设越来越迫切[1]。早在朱镕基总理视察国家统计局时就说过"统计部门是专门的信息部门,掌握着十分丰富的信息资源,迫切需要利用现代信息技术进行有效的开发和科学的管理。应当把统计信息化建设作为我国电子政务建设(政府信息化)的一个重点,使统计部门在信息化建设方面走在前面"。温家宝总理也曾深刻指出:"统计建设的重点是信息化,核心是信息化"。可见我们国家对统计信息化的重视。目前虽然各地统计部门的计算机硬件基础设施建设已取得极大的进展,但是各地统计部门的软件系统开发却严重滞后,成为统计工程信息化的一个瓶颈。统计数据的查询是统计工程信息化中最为基础最为核心的功能,该系统就是在这个背景要求下建立的基于区县级的统计数据查询系统。

收稿日期:2008-05-31

作者简介:何荣刚(1985-),男,江西南昌人,硕士研究生,主要研究领域为数据挖掘,移动数据库;陈启安,硕士生导师,教授,主要研究领域为人机界面、多媒体技术、嵌入式开发。

1 系统设计目标

目前统计行业中的问题主要表现在以下几点:

(1)统计数据格式缺乏统一的标准,规范性差;

(2)统计数据比较分散,各个科室各自保管各科数据;

(3)统计指标因每年需求不同而不断发生变化。

基于以上一些问题,大部分现有统计软件的设计思想是提供一个接口将所有原始的统计表整合到一起,再通过表来查询数据。这样虽然克服了统计数据格式的不规范和分散的缺点,但是只能进行一些简单的查询,远远不能满足平时统计人员的查询工作。比如没有办法进行多样化、跨年度、跨行业的综合性查询[2]。为解决这个问题,本系统提出了以指标为基础、法人单位和时间为限定条件并可以附加多种过滤条件的查询方式,极大地提高了查询的灵活性和多样性。

2 系统的构建与实现

2.1 统计数据的初步清理与整合

由于下级统计单位送上来的统计数据格式不统一,有些是文本格式,有些是 foxpro 格式,还有些是 access 格式,甚至其他格式。同时统计数据的内容包含各行各业,统计指标多,数据量大,为方便后面的查询

工作,必须建立一个数据库供系统使用。这个数据库不能只是简单的原始表的集合,还要对数据进行初步分析清理,将数据按行业和年份区分,分别存储在各个表中。

统计数据分为统计年报和统计定报[2]。年报统计的是年度数据,定报统计的是月度、季度或者半年的数据,在该系统数据库中也按年报、定报的划分格式将数据表分为年报类和定报类。所以在将一张表导入数据库之前要确定该表是年报表还是定报表。再根据统计表所属的行业以及时间将表中数据导入到数据库中的对应表中。为建立以指标为基础的查询,需要对原始数据表进行转化,读取原始数据表中所有数据指标,并做进一步优化,然后依次将每个法人单位的所有指标以及该指标下的值一条一条存入到新的数据表中。例如存在一张 B103 的原始数据表(如表 1 所示),读取其指标 B01,B02。在指标前面加上表标识 B103 变成 B103:B01,B103:B02。将指标进行行转列的变换。即对每一个法人单位的每一个指标生成一行。如表 2 所示。

表 1 原始数据 B103 表格式

UnitCode	UnitName	B01	B02
1028451＊＊	＊＊有限公司	65.078	1234

表 2 系统数据库中数据表格式

UnitCode	UnitName	IndexCode	DataVualue	其他属性
1028451＊＊	＊＊有限公司	B103:B01	65.078	○○○○○○
1028451＊＊	＊＊有限公司	B103:B02	1234	○○○○○○

这样的转换可以将多个原始表整合到一张表中,虽然单张表的数据量增加很大,但表的数量减少许多,使查询更加集中,并有利于多样化的查询。

2.2 建立统计数据指标管理系统

由于该系统的查询功能主要是以数据指标为基础的查询,所以建立一个强大的指标管理体系就能大大提高系统的查询能力。统计数据指标的管理主要是包括设计指标表结构、跨年度指标之间的关联和跨行业指标之间的关联。

2.2.1 设计指标表结构

指标表的设计是指标管理的基础。根据统计指标的特点,将指标分为基本属性指标和数据指标[3],基本属性指标是指所有法人单位基本的、共有的指标,包括行政区代码、行业代码、登记注册类型代码、控股情况、联属关系、营业状态、机构类型、表种类别、统计管理部门代码。数据指标是指根据统计需求而定的指标。数据指标数量相当庞大,而指标代码只是用字母和数字简单组合而成,因此会存在很多指标代码相同含义却不同的指标,比如在工业中的 B103 表有个 B12 指标表示"流动资产合计",而在运输邮电业中的 D1021 表中 B12 指标却表示"货运量"。如果不对指标进行改进而直接导入表中,可能会使数据指标混乱,导致指标管理十分困难,因此需要为数据指标添加标识使其唯一。设计思想如下:将数据指标按年份来存放,如 2007 年指标表命名为 Index_2007。该表包括该年的所有行业的所有数据指标。将表名加在该数据指标前面来唯一标识一个指标。形如:"B103:B12"和"D1021:B12"。这样数据指标所属的年份、所属的行业表就能清晰地展现了。

2.2.2 跨行业指标之间的关联

原始统计表是按行业划分的,包括农业(A)、工业(B)、建筑业(C)、运输邮电业(D)、批发零售及餐饮业(E)、服务业(F)、金融业(J)、房地产业(X)、劳动工资(I)。在多个行业中,会有相同性质的数据指标[3]。有时候需要查询该类数据指标在各个行业下的表现情况。在查询时,需要将表示该含义的所有表中的数据指标都加到查询条件中去,十分烦琐。为简化操作,使查询功能更方便易用,需要建立一个跨行业的指标关联。如果有多个行业共同拥有某数据指标,用户可以将其添加为综合指标。首先如果某行业的数据指标是第一次关联则新建一个行业 K 来存放综合指标以及综合指标数据。如果不是第一次关联则在添加为综合指标的同时将这些指标的数据导入到综合的数据表中。例如:"b103"表中的"b64"指标代码表示"营业利润"。在其他行业的表中发现也有表示"营业利润"的指标。需要将其添加到综合指标中。同时新建行业 K 的数据表,将"b103:b64"指标的数据导入行业 K 的数据表中。同时将其他表示"营业利润"的数据指标与该综合指标进行关联。关联的同时会将关联指标的数据导入到行业 K 的数据表中。因此查询综合指标的数据源来自与行业 K 的数据表[4]。

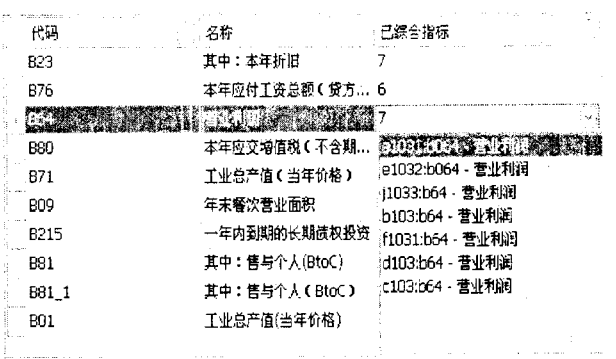


图 1 跨行业指标关联图

图 1 给出了跨行业指标关联展示,在这张跨行业指标关联图中,列出了综合指标代码、综合指标名称以及跟这个综合指标关联的普通指标个数。点击数字区

域就可以显示该综合指标关联的具体指标。

2.2.3　跨年度指标之间的关联

统计行业中经常要分析一个数据指标在两年甚至多年中的变化情况,可见查询多年的数据是一个很常用的查询,为提高查询的效率,有必要将多年的相同数据指标进行关联。这样就可以选择这个公共指标查询多年数据,而不需要添加每年的数据指标。为此在系统数据库中建一个 Index 表做为公共指标表,它是一个中间表,用来关联各个年份的数据指标。首先如果有两个以上年份同时有某个数据指标,可以将这个数据指标添加到公共指标表中,再将其他年份的该指标与公共指标进行关联。关联方法是在年份指标表中增加一个公共指标关联标识列 IndexAtt。如果某个指标有跨年度关联,则在该列填上相关联的公共指标[5]。这样公共指标就作为中间项实现了数据指标的关联。

图 2 给出跨年度指标关联展示,默认列表示公共指标,该图关联了 2004 年、2005 年、2006 年三个年份的指标。在数据表中的体现是在这三年的指标表中的 IndexAtt 列中写入相关联的公共指标。

默认	2004	2005	2006
B103:B11 - 一、年初存货	B103:B11	B103:B11	B103:B11
B103:B12 - 流动资产合计	B103:B12	B103:B12	B103:B12
B103:B13 - 短期投资	B103:B13	B103:B13	B103:B13
B103:B14 - 应收帐款(净额)	B103:B14	B103:B14	B103:B14
B103:B15 - 存货	B103:B15	B103:B15	B103:B15
B103:B16 - 其中:产成品	B103:B16	B103:B16	B103:B16
B103:B17 - 流动资产年平均余额	B103:B17	B103:B17	B103:B17
B103:B18 - 长期投资合计	B103:B18	B103:B18	B103:B18
B103:B19 - 固定资产合计	B103:B19	B103:B19	B103:B19

图 2　跨年度指标关联图

3　统计数据查询

统计行业最主要的工作就是收集数据并在庞大的数据中查询出自己想要的数据。

首先选择查询的时间,再确定查询对象的范围,可以是全部法人单位或者是满足某种条件的法人单位集合,并在这里选择是汇总还是明细查询。汇总查询是显示查询到的法人单位的数量,明细查询则是列出所有查询到的法人单位。然后选择所关心的数据指标,同时选择过滤条件使得在查询中对法人单位进一步筛选,查询得到满足条件的结果,并以表格的形式展现出来[5],以方便统计人员对数据编辑发布。

该查询系统能满足统计部门所要求的大部分查询工作,列举部分查询如下:

(1)查询在各种分组条件下的所有法人单位数。选择要查询的某个年份,查询对象选择该年下所有法人单位并选择汇总查询。选择一个或多个基本属性指标做为分组条件,过滤条件项如果没有可以不填,如果要限定某个条件则可以添加过滤。例如要设定查询的法人单位属于某个行业,则在过滤条件中增加行业代码等于该行业的条件[4]。图 3 展示了查询"按隶属关系分组以及按表种类型分组条件下的所有法人单位数"的结果。

全部法人单位

项目	2006
隶属关系分组	2165
中央	49
市	111
区属工业	420
区直工业	194
镇村工业	226
其他	1586
表种类型分组	2105
第一产业	158
第二产业	1153
工业	786
第三产业	794

图 3　查询案例一

(2)查询连续几年内某指标变化情况。选择多个年份,并在所有法人单位中通过搜索法人单位代码或者名称找到需要的法人单位选择明细查询或者汇总查询,统计指标中选择关注的数据指标。如果需要对法人单位做进一步限定,可以在过滤条件中添加[4]。

图 4 展示了查询"行政管理部门分组条件下工业企业现价总产值在 2005 和 2006 年的变化情况"的结果。

工业企业现价总产值

项目	2006	2005	增减%
行政管理部门分组	24465675	22729891	7.6
街道小计	1579304	2378972	-33.6
兴丰街道办事处	1235297	1942559	-36.4
林校路街道办事处	185379	283934	-34.7
清源街道办事处	158628	152479	4
镇村小计	19728166	17677549	11.6
亦庄地区办事处	1018777	1059650	-3.9
黄村地区办事处	5641453	4954400	13.9
旧宫地区办事处	1951454	1833961	6.4
西红门地区办事处	2452351	23880303	2.7
青云店镇	762032	696044	9.5
采育镇	855910	752934	13.7
安定镇	489289	279240	75.2
礼贤镇	239612	228952	4.7
榆垡镇	1263658	1059952	19.2

图 4　查询案例二

(下转第 192 页)

Q 为二次误差测度工二次误差矩阵，$Q(v)$ 为二次误差值。

网格中每边的误差矩阵为边的两个端点的误差矩阵之和：$Q = Q_1 + Q_2$，每次应选择当前误差值 $Q(v)$ 最小的边进行折叠。

1.2.2 线性插值

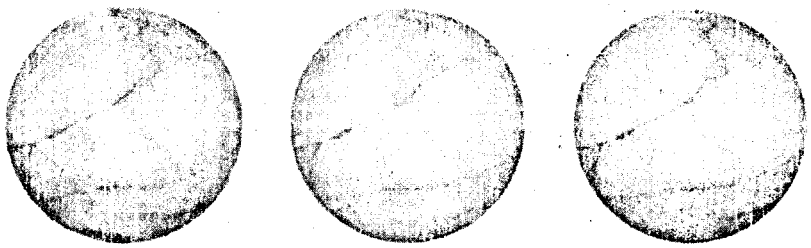
在远程接收三维模型的计算机上，接收完一批压缩数据并将 M_i 细化成 M_{i+1} 时，用动态变化的 M_G 代替 M_{i+1}，演示模型的渐变过程。具体的线性插值方法如下所示：设 v_i 是 M_i 中的顶点，v_{i+1} 是 M_{i+1} 中的顶点，v_s 是 M_i 中的分裂顶点，其分裂产生的顶点为 v_t，$V_G \in M_G$，有

$$V_G = \begin{cases} (\sigma)v_1 + (1-\sigma)v_3 \cdots v_{i+1} = v_t \\ v_1 \cdots v_{i+1} \neq v_t \end{cases}$$

其中：σ 取值[0,1]，从0开始以 $\triangle\alpha$ 的间隔增加，用 V_G 不断刷新显示 M_G 中的顶点数据，直到 $\alpha = 1 - \triangle\alpha$。假设 $M_i \rightarrow M_{i+1}$ 按网络带宽预估的传输时间为 T_i，模型显示计算机所能实现的显示刷新时间间隔 $\triangle t$，有 $\triangle\alpha = \triangle t / T_i$。$\triangle t$ 会随着模型数据的加大而增加。为简单起见，以最开始的 $\triangle t$ 计算，以 M_0 的顶点数为基础，按每批增加上一批 50% 顶点数来估算 T_i，在完全接收 M_{i+2} 的压缩数据时停止 M_G 的演示，直接显示 M_{i+1}。随着细化数据的使用，模型复杂度的加大，几何形变的变得细微，M_G 的演示应适可而止。

2 实验结果

经过实验得证文中提出算法不仅对网格模型几何模型信息得到有较好的保留，其颜色信息也得到了有效的保留，并且 MACPM 技术相对 PM 技术在网络传输上总体效率提高。图 3 是一组实例。



(a) 2088个面片 (b) 4804个面片 (c) 7328个面片

图 3 多属性网格简化传输显示实例

3 结束语

提出并实现了一种解决多属性三维网格模型的压缩与传输问题的技术，该技术利用扩展二次误差测度表示简化网络与初始网格的几何特征与其它属性信息的匹配，同时利用线性方法简化压缩算法，提高三维网格数据在网络上的传输效率，且保证三维数据解压显示效果。

以后的工作将尝试将该技术应用于海量三维地型数据流式传输方面，相信会有更大的扩展空间。

参考文献：

[1] Taubin G, Gueziec A, Horn W, et al. Progressive forest split compression[C]// SIGGRAPH'98 Proceedings. [s. l.]: [s. n.], 1988:123-132.

[2] 陶志良，潘志庚，石教英. 支持快速恢复的可逆递进网络及其生成方法[J]. 软件学报，1999,10(5):503-507.

[3] 卢章平，赵 泉. 基于"边折叠"的累进网络生成算法的研究[J]. 工程图学学报，2004,24(1):37-41.

[4] Garland M. Simplifying Surfaces with color and Texture using Quadric Error Metrics[C]// Proceedings of IEEE Visualization'98. [s. l.]: [s. n.], 1998:263-269.

[5] 郝齐辉，谭同德，李润知，等. 多属性递进网格快速生成算法研究[J]. 计算机技术与发展，2006,16(11):44-46.

(上接第 189 页)

4 结束语

文中所述系统为区县级的统计部门提供了一个查询解决方案，基本能满足统计人员平时工作所需。数据库设计简洁清晰，维护简单，查询操作简单却多变。查询结果以表格的形式展现，使结果编辑、发布更简单。

如果该系统与地理信息系统(GIS)以及分析系统相结合，将能提供更强大的结果展示功能。

参考文献：

[1] 杨立勋. 县乡(镇)统计管理体制模式的选择[J]. 统计与决策，2005(10):35-36.

[2] 张建平. 浅谈统计管理的自动化[J]. 甘肃科技，2007,23(10):150-151.

[3] 吴 飞，张世杰，罗 旭. 基于B/S可变条件数据库统计查询设计与实现[J]. 铁路计算机应用，2003,12(1):1-3.

[4] 罗远模. 完全掌握 SQL Server 2000[M]. 北京：北京邮电大学出版社，2001.

[5] Simon R, Nagel C. Professional C#[M]. 北京：清华大学出版社，2005.