人口数据库智能应用的实例研究

张荣肖　张力峰　（中国科学院北京软件工程研制中心）

摘要：本文首先分析了商用 DBMS 的局限性，然后以计算机人口管理数据库应用为实例，探讨数据库智能应用的技术与方法。

关键词：数据库　信息系统　智能应用　实例研究

一、商用 DBMS 的局限性

商用数据库管理系统已发展成熟，能很好的支持以检索查询数据为主的应用，当与知识信息结合使用时，出现一系列有待解决的问题，主要有：

1.数据类型：传统商用数据库管理系统基本是面向记录存贮的用字符表示的格式化数据为主，不能支持如图形、图象、语音、视频、动画音乐等各种类型数据。

2.存取信息：现在的数据库管理软件仅能提供取用那些已存入数据库中具体特定的信息，不能存取非直接存贮的不太确定和较为模糊的信息。

3.查询效率：从数据库中查询数据的响应时间难以保证，尤其当数据间关系复杂时，连接操作可能使查询响应时间很长，甚至使用户难以忍受。

4.用户界面：传统的数据库管理系统在用户界面一级均有缺陷，如不直观，没有标准的人机交互方式，使用不便。

5.适应性差：传统的数据库管理系统，只能机械地存贮数据，再经开发应用程序完成需求，被动地提供数据信息服务，不能直接地适应具体应用环境，不能为高层次管理决策者理解数据的整体特性服务，没有加工处理知识及演绎推理能力。

随着社会信息化，不仅信息量增加，对数据库技术也提出了一系列新要求，如多媒体信息服务要求数据库能适应多样化的数据类型；能提供同时交叉访问各类数据（图形、图象、声音、文本等）的能力；用户有多种手段即不仅可用命令、语言等方式，还可以用国标，表格甚至声音存贮、检索、操纵和管理数据；随着数据库的内容丰富，数据库通过网络同时为许许多多用户服务，这就涉及数据库的安全性和保密问题；高层次管理人员或决策者希望数据库能支持分析和预测，具有推理、类比能力，从中导出知识性数据，主动提供信息服务等。

二、计算机人口信息管理系统

计算机人口信息系统是为人口管理服务的，我国自从在全国范围内实行居民身份证制度以来，对人口数据作为社会基本信息，逐步实现现代化的计算机管理提供了条件，计算机在我国人口管理领域的应用从 80 年代后期已开始，以人口数据作为信息资源，建立管理信息系统与其它领域中的管理信息系统（MIS）有共同之处，也有其特殊性。若把人口的管理分为操作层管理，中层管理和高层管理三个层次，则可用图 1 表示。

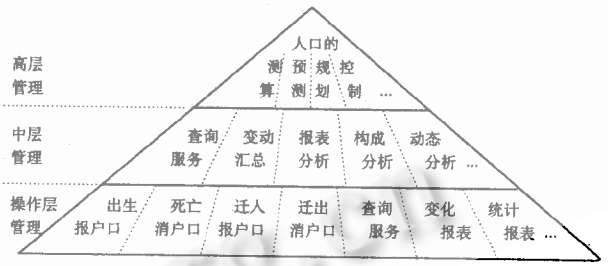

图 1

第一层：面向基层日常人口事务处理的操作层管理。主要内容是日常的人口登记，申报人口的出生、死亡、迁出、迁入等变动；为社会各方面提供有关户口及相关的查询服务。

第二层：对人口数据进行常规统计与分析的中层管理，主要内容是进行人口静态和动态统计，产生定期和不定期报表。如有关区域范围内的总人口数，人口自然构成，人口地区分布，人口经济文化构成等反映人口基本情况的各类报表，人口出生、死亡、迁移、婚姻生育等变动情况的报表。

第三层：计算机辅助人口知识应用的高层管理。主要内容有依据统计结果数据，反映人口自然构成状态和文化经济构成状态等指标数据进行人口分析，评价一个地区发展的现代化水平，辅助进行人口测算，人口预测，

制定城市或区域的人口规划。

上述三个层次间及各层次内部都是由相互关联，相互依存和制约的多个功能部分组成。设计并实现包含三个层次的人口管理应用系统是一个复杂庞大的系统工程，它涉及到计算机技术、通讯技术、软件工程、数据工程和知识工程，是现代科技与人口领域相关的管理科学领域等多方面的知识与技术功能的集成。图 2 给出了这种应用软件系统的组成以及与现实世界和用户之间的相互关系的示意。其中：

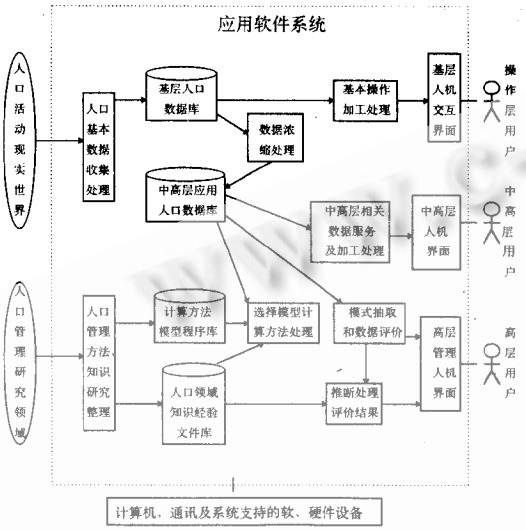

图 2

1.人口基本数据的收集处理：将现实世界中，手工用的人口管理薄上的数据进行整理，录入计算机中，由于人口数据量大，为了节省空间占用，一般需要用编码，建立编码标准和字典，经过审查核实后，放入人口基本数据库中，作为基础数据源。

2.人口管理的方法与知识整理：在人口管理和人口学研究领域有长期积累的许多专门的方法和知识，收集那些能与人口基本库中的数据相匹配，为解决专有问题为目标的计算方法和分析预测算法模型等，以可以直接调用的子程序方式放在方法模型库中，将评价人口的知识及相关的指标形成评价标准文件，存放在知识和经验库中。

3.数据加工与处理：这是一组计算机程序，可以分别完成各类具体应用需求，图中各方块都代表一类功能组。

4.人机交互界面：人机交互界面是供用户操作系统工作，完成具体需求任务而设计，一般应具有如下功能。

.提供多种多样的对话及显示，如窗口、菜单、选择功能等。

.输入/输出功能，是人和机器间能够互相理解，实现人机对话操作。

.接受用户指挥，调用加工处理功能程序模块运行，以便得到用户需求的结果。

实现中下层管理用户需求的软件设计过程与常规的以事务处理为中心的数据库应用系统的设计相类似，它们是：

.分析数据本身特性及数据使用的需求特性。

.设计数据库结构和划分数据处理功能。

.收集源数据，建立支持系统工作的数据库。

.依据对数据的使用需求，逐步细化处理功能，设计数据加工处理模块，如数据的检索处理，数据报表处理等。

若要求系统能支持高层管理应用，就需要扩充应用软件，但要注意两点，即不变更原来的数据库结构设计和不变更原来的应用处理程序。所扩充的功能主要围绕高层管理者的需求进行，这类需求设计的内容很多，下面仅就结合人口领域专家评价一个国家或地区的人口年龄构成特征和评价一个国家或地区的人口再生产类型的过程给出相应的设计实例。

三、与领域知识相结合的应用设计

任何领域中都有丰富的专门知识，同一领域中，不同的应用需求，涉及的知识内容也不相同。当数据库与领域知识相结合进行应用设计时其指导原则就是让计算机能模拟领域专家用知识和数据进行工作，图 3 给出了人

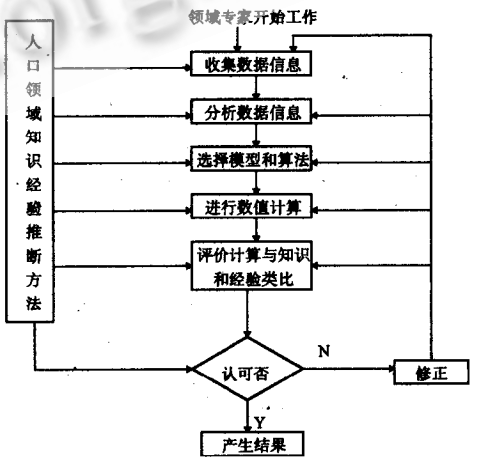

图 3

口领域专家进行人口预测或分析工作的过程。其中工作

过程的每一步都有领域的背景知识、经验和方法在做参照。

(1)收集数据信息:收集历史上的人口数据和当前相关的人口信息数据,同时要考虑影响数据变化的因素。

(2)分析数据信息:研究数据变化特征及演变趋势,相关因素带来的影响及作用。

(3)选择模型和方法:根据分析的信息,寻求合适的模型和计算方法。

(4)进行数值计算:用模型和数据计算参数,再用参数和模型进行计算以获得结果。

(5)与知识和经验类比,评价计算:这一步是检验数值计算结果的准确性和可靠性,研究因素变化产生的影响,再据知识和经验类比推断,对结果进行评价。

(6)修正:用专家的领域知识及宏观见解衡量结果,如对结果不满意,可以进行修正,返回前面的任何步骤,重复进行,直到获得认为合理的结果为止。

在前面图 2 中已简要描述了支持高层管理应用的软件框架,应用设计者首先要考虑如何解决数据和知识的存放,数据和知识如何获取,数据和知识的相联类比与推理的过程及编码方案。高层管理用户通过人机交互界面操纵系统,辅助领域专家进行工作。现在结合具体问题给出实用设计。

1. **问题描述**

瑞典人口学家桑德巴氏根据现有人口年龄构成与未来的人口出生率、自然增长率的关系,提出两个评价地区人口特征的国际通用标准。

标准 1:一个国家或地区的人口年龄构成特征分为年轻型、成年型和老年型三种,具体参照不同年龄段的人数的比例及年龄中位数确定。有表 1:

表 1

国际标准	少年人口系数(0-14 岁)	老年人口系数(65 岁以上)	年龄中位数
年轻型	40%以上	5%以下	20 岁以下
成年型	30-40 %	5-10%	20-30 岁
老年型	30%以下	10% 以上	30 岁以上

其中年龄中位数的意义如下:按年龄自然顺序排列的总人口构成一个连续的变量数列,这个连续变量数列的中间值就是年龄中位数,它把总人口数分为两半,一半在年龄中位数以上,一半在年龄中位数以下。

标准 2:从人口年龄构成情况可表明人口再生产类型是增加型,还是稳定型或减少型。标准见表 2:

问题:以某地区的计算机人口数据库为基础,依据国际通用标准 1,评价地区的人口年龄构成特征;依据国际通用标准 2,评价地区的人口再生产类型。

表 2

国际标准	0-14 岁(%)	19-45 岁(%)	50 岁以上(%)
增加型	40	50	10
稳定型	26.5	50.5	23
减少型	20	50	30

2. **应用分析**

实现此系统要求备有能提供年龄的人口基本信息数据库,能按年龄段进行组合的计算方法,衡量人口构成标准的知识文件,在此基础上设计供用户操纵使用系统的界面程序和计算程序。其中:

(1) 界面功能:供用户操纵使用系统,要求:从界面上可看到进行评价的标准表;从界面上可以启动系统使之开始计算评价;可以看到该区域的人口年龄构成表;可以看到该区域人口年龄金字塔图;从界面上可以得知评价结果。

(2) 计算处理功能:实现界面的显示及内部数据存取和类比,要求:从人口基本库中计算各年龄段人口数据及年龄中位数;用各年龄段人口数据计算并求出年龄金字塔;用各年龄段人口数据和评价标准进行类比形成结果矩阵;由结果矩阵推断结果并在屏幕上显示。

实现上述具体应用的功能模块关系见图 4。

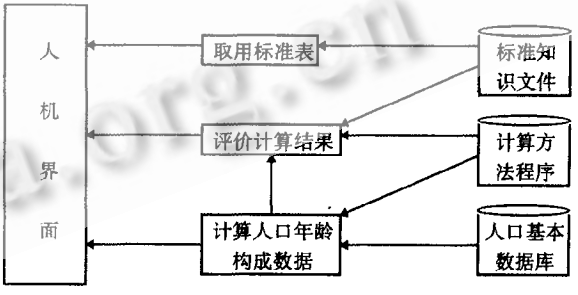

图 4

3. **实现设计**

下面以评价地区的人口年龄构成为例,说明设计与实现的细节。

将人口年龄构成评价标准表 1 变换成为等价的形式(表 3):

表 3

国际标准	0-14 岁人口(%)	65 岁以上人口(%)	年龄中位数
年轻型 (Y)	(40,100)	(0,5)	(0,20)
成年型 (M)	(30,40)	(5,10)	(20,30)
老年型 (O)	(0,30)	(10,100)	(30,100)

表 3 数据部分是个 3(3 的矩阵,称为标准矩阵,记为 S=|(aij,bij)| (i,j=1,2,3),其中 aij 表示第 i 行标准所对应的第 j 列年龄组人数百分比的下限值,bij 为其上限值。另设一个 3(3 矩阵为判定矩阵,记为 R=|rij|(i,j=1,2,3),rij 取值或为 1 或为 0,取值算法解释如下:

从人口基本库中使用计算程序可求得年龄在(0-14)岁之间人数百分比和 65 岁以上人数百分比,分别记为 x 和 y,年龄中位数记为 z。即可获得一个含有三个值的数组(x,y,z),且 x,y,z 三者必然分别落在区间(ak1,bk1),(al2,bl2),(am3,bm3)中(1(k,l,m(3)。设有 x 值落在(ak1,bk1)中则记 rk1=1,ri1=0(1(i(3 且 i(k),y 值落在(al2,bl2)中,则记 rl2=1,ri2.=0 (1(i(3 且 i(l),z 值落在(am3,bm3)中,则记 rm3=1,ri3=0(1(i(3 且 i(m)。因而可获得矩阵 R=|rij|的具体表示值。

判定法则:R=|rij|(i,j=1,2,3)矩阵元素由 0 和 1 组成,若某一行含有 2 个或 3 个"1"元素,则该行对应的类型就是该地区的年龄构成类型。若每行均含有一个元素"1",则看年龄中位数相应 rm3=1 的 m 值所对应的行即为所求类型。

用类似的方法可以评价地区人口的再生产类型。

数据库及智能应用的内容极为丰富,80 年代末国际上开始了在数据库中发现知识(KDD-Knowledge Discovery in Database)的研究与系统开发,探讨满足对数据库高层应用需求的方法,从 89 年到 94 年开过 4 次 KDD 的国际研讨会,已取得一批富有成效的研究成果,集成综合的应用数据库、网络、多媒体和领域知识,必将对信息化社会产生深远的影响。

参考文献

[1] C.J. Date An Introduction to Datebase System 1990.

[2] 翟振武 等著《现代人口分析技术》1989 中国人民大学出版社