构筑面向决策支持的统计数据挖掘系统面临的几个问题

柳小鹏,杨　帆,米　红
(厦门大学 模式识别与智能系统研究所,福建 厦门 351005)

摘　要:面向决策支持的统计数据挖掘系统在我国统计信息化进程中具有重要意义,特别是在区县级统计管理部门,必须将日常的统计业务与决策支持相衔接。统计数据的异构性和计算机技术应用的多样性是造成统计部门中信息孤岛的原因。文章认为,要解决该问题,必须设计统一的统计数据仓库和集成多种计算机技术,建立面向决策支持的统计数据挖掘系统,同时,为了保持前瞻性和有效性,必须保持系统的可扩展性,并融入当前最新的统计技术和数据挖掘技术。

关键词:决策支持;数据挖掘
中图分类号:TP182　　文献标识码:A　　文章编号:1002-6487(2008)12-0016-03

0　引言

随着我国社会主义市场经济的发展,人们的思想观念不断完善,政府的职能也不断转变。统计工作是一项政府行为,统计信息资料是社会公共产品,因此统计工作的开展、目标、内容和定位都依赖于政府职能和体制。随着"小政府,大社会"的观念逐渐深入人心,政府的基本职能、管理体制和工作方式将渐渐改变,官方统计工作也将需要做进一步调整和完善,向着更加科学、有效、公正、公开的方向发展。

一般认为,我国统计工作"主要矛盾是日益增长的统计需求与传统落后的统计工作方式之间的矛盾"。统计工作的需求是不断变化、日新月益的,统计内容无限扩张,有时需求还显得很不合理,这都对统计工作的效率和内容提出了挑战:是被动地接受新的需求而不断增添工作内容,还是主动适应变化的需求,以不变应万变?这是统计工作方式改革和发展的方向之一,统计信息化是一个解决手段。

信息化手段用计算机来代替人的手,可以提高统计工作的效率和准确性,对于缓解日益增长的新需求有积极意义。人们越来越重视统计信息化,从软件到硬件上的投入也越来越大。但是,目前常用的信息化手段偏重于数据的采集和整理以及初步分析,对于新的需求还是必须开发新的系统,于是统计部门也有了自己内部的座座信息孤岛。

随着时代的发展,统计部门的职能也将走向转变,笔者认为面向公众服务和提供直接的决策支持是其中两个重要方向。但是现有的统计信息化方案都没有很好的解决这两个问题,仅仅是单个或多个种类的业务数据的流动,缺乏对数据的分析,没有数据与区域社会经济发展之间的有机联系,也就形成不了知识。数据,放在数据库里永远只能是数据,既不能为公众服务,也不能为政府决策提供直接服务。只有面向决策支持的信息化方案才能创造更大的价值。

其实,从本质上来说最重要的不是信息化手段,而是统计理念,只有科学的理念才能指导统计工作的正确开展。统计必须科学、有效、公正、公开,不能靠人的主观观念和不断变化的需求来指导统计工作,不能迫使统计数据迎合某个指标、符合某个概念、达标达线。在获得统计数据后,统计分析中的主观性和随意性只能通过科学客观的方法才能消除或降低,而目前的统计信息化方案在这些方面做得远远不够。

还有一点必须指出,在构建统计信息系统时,必须考虑到该系统的应用层次,是应用在较高层次的宏观决策上,还是应用在基层单位的统计业务上,本文主要针对区县级统计业务部门在构筑适合区情的信息系统时所应注意的问题。因为区县统计部门的工作承上启下,区县作为政府行政和管理的基本单元,在社会经济生活中扮演一个不可分割的"原子"角色。很多经济生产、社会文化以及其他人类活动和现象都在区县的层面上表现出整体性和系统性;可以说,区县应该是我国社会、经济和人口复杂巨系统的最小层次的子系统,其上层是市、省乃至全国的宏观系统,其下是更小的乡镇和农村微观单元。区县应该是综合统计、管理经济、制定计划和宏观调控的最小单元,再往下,其经济运行和社会生活都不能成为一个整体、一个小系统,因此区县扮演了双重的角色,也是最基本的角色。

从这个角度上说,区县统计工作起着承上接下的关键作用,既需要按照上级管理部门的要求提供统计信息资料,又需要科学地分析本地社会经济和生产生活的运行状况,为监管监测、宏观调控提供科学支持和智力支撑,并为全民提供

基金项目:国家社会科学重点资助项目(07ASH009);"985"工程智能化国防信息安全技术科技创新平台资助项目(0000-X07204)

可信的参考信息，随着政府职能在市场经济中的转变，很多政府行为将退居市场之后，‘看不见的手’将更多地主导地方经济的运行，这时，统计工作就显得尤为重要，成了政府的触手来把握经济运行、社会发展的脉搏。因此，可以形象地称区县域的统计工作是‘小区域，大统计’。

为此笔者从区县级统计部门的需求入手，结合多年来统计业务实践，提出以下观点：

(1)从区域经济系统的整体角度来考虑统计工作。一个区域，其社会经济总离不开地理空间和时间，因此一个可行的信息化方案是赋予每一个统计指标以三个维度：时间维，空间维，分组描述维。通过这三个维度，每个报表都可化解成一个个指标的集合。这样新的需求带来的业务报表可以通过ETL过程加入统计数据仓库中，也可以从数据仓库中生成任意格式的报表。

(2)必须结合区域经济系统运行特征，设计一套完整科学的统计分析方案，满足以下分析要求：①可以从各个维度上分析，包括地理空间上的对比分析、时间上的分析、不同分组下的分析，以及多维度交叉分析；②涵盖常用统计方法、机器学习和人工智能方法。既有统计描述、相关分析多元回归和综合评价，又有神经网络、支持向量机、模糊聚类等智能分析方法弥补前者的不足；传统统计方法需要很强的假设和大容量的样本，而现实中的决策支持常常缺乏大量的数据，或者数据属性很多，这时机器学习方法就能充分发挥其在‘高维、小样本’数据分析中的优势；③面向主题，既有总的区域经济指标的分析(如区域经济分析，包括产业结构、地理联系率等)，又可以对人口普查、经济普查、农业普查以及其他自定义专题进行各种分析，例如模型生命表方法在人口普查模块中的应用；④具有很好的扩展性，将统计业务与前沿的分析方法自然地融为一体，使得统计数据的分析更加科学客观，并能产生知识。

(3)分析结果和知识不仅有丰富的图表展现手段，更可以存储成知识库的形式，并向公众发布。

1 统计信息化现状

1.1 统计数据的基本问题

虽然各地统计部门的计算机网络等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，成为统计信息化工程的一个瓶颈。特别是面对大量繁杂而分散的数据资源，如何安全有效地管理和重组数据，提炼出统计综合数据信息，以供政府部门和社会各界利用，成为目前统计行业所面临的一个亟待解决的难题。具体表现如下：

(1)统计数据规范性差，影响了统计信息标准的统一。目前各地统计部门的统计数据处理软件门类众多、运行平台各异、存贮格式多样。有的部门使用上级统计部门下发的系统，有的部门使用自主开发的系统，各系统相对独立；有些部门使用FoxPro、Access等格式数据，有些部门使用SARP自定义模式数据等；容易出现数据不齐全、不一致或重复现象。

(2)数据来源多，但存放相对分散，缺乏统一管理，影响了统计信息的综合开发。统计数据一般分别存放在各个统计专业科室的数据库中，而且大多只保存近期数据，缺乏集中存放和管理不同专业、不同时期统计数据的有效手段。

(3)统计业务涉及到各行各业，指标多、数据量大、应用少。各级统计局除了能将这些数据汇总成为统计报表、统计年鉴、市情手册或经济卡片外，缺乏对专业统计数据进行深层次分析、综合、提炼、挖掘和展现的工具，很难对丰富的统计资源进行二次开发和利用，辅助决策的有效信息更少。

(4)统计信息开发力度不够，统计服务方式单一，不能很好地适应社会公众的需要。当前，统计局的职能主要是报送统计报表，忽视了为各级政府决策和企业决策提供有效的信息服务；统计信息开放程度低、共享困难。

(5)统计分析方法单一，缺乏指标的深度及关联度分析，没有与其他行业数据和地理空间数据进行关联。

1.2 关键技术的集成问题

目前统计部门常用的信息系统软件，要不就是用于联机事务查询处理(OLTP)的统计数据库软件，要不就是管理信息系统(MIS)，或者是决策支持系统(DSS)，人们为了满足某种特定需要，不得不针对相同的数据集开发多种特定类型的软件系统，缺乏一个整体的统计数据处理和分析框架，既造成了资源的浪费，又加剧了信息孤岛效应。

常用的统计软件不具备空间查询和分析功能，而随着空间经济学和人口学的发展，新的统计分析的思想和方法不断产生，必然会为我国统计部门的日常业务带来新的需求，因此要求将地理空间信息纳入统计和分析框架中。GIS可以实现地图上的指标查询和分析，即在一个或多个图层上选定一个或多个任意形状的区域，对选中的指标进行汇总和比较，对于进行区域经济分析和空间计量经济分析有着重要的现实意义。

但是，不同于一般的社会经济地理信息系统，应用型数据挖掘系统(统计应用)，以数据仓库思想来存储管理统计数据，把GIS作为为其中一种知识表达手段，整个系统的设计思路力图遵循知识发现过程，按照数据导入和整理、数据查询和展示、数据分析和挖掘的流程，不仅满足日常的统计需要，更要为政府的科学决策提供知识支撑。

值得一提的是，除了增加人口普查、农业普查和经济普查以及其他专题以外，区域经济分析功能也非常重要，能够提供经典的区域经济指标分析，如，对地区产业经济、社会和人口发展的均衡性、多样性和协调性进行分析，针对所选地区、行业和时段的指标进行各种分析，进一步地服务于政府决策部门，而不需要业务人员亲自进行复杂的运算。

正是由于需求的多种多样，人们才开发了各种各样的信息系统。要避免产生新的信息孤岛，必须实现各种技术的集成，成为一个完整的统计业务解决方案。具体来说，必须全部或部分实现下列计算机技术的集成：

①地理信息系统，Geographical Information System；②管

理信息系统，Management Information System；③决策支持系统，Decision Supporting System；④统计数据库，Statistics Database；⑤数据仓库，Data Warehouse；⑥数据挖掘，Data Mining；⑦空间分析，Spatial Analysis；⑧空间数据挖掘，Spatial Data Mining；⑨社会经济地理信息系统，Socioeconomic GIS。

系统结构					
数据维护工具	统计业务			信息发布	
系统功能模块					
数据维护模块	统计查询模块	挖掘分析模块	GIS 模块	信息发布模块	用户管理模块
系统数据库					
年报数据库		定报数据库	普查数据库	综合数据库	

2 构筑面向决策支持的统计数据挖掘系统须考虑的几个问题及解决方案

通过前文的分析，笔者认为，为了更好的面对统计部门的职能转变，有效应对不断变化的、前瞻性的需求，同时解决信息孤岛问题，一个有效的途径是设计一个统一的数据仓库，并集成多种计算机技术，构筑一个面向决策支持的统计数据挖掘系统。在设计时，需要仔细斟酌如下问题。

2.1 建设目标是什么

该系统的建设目标应是一个包括人口和社会经济统计数据、地理空间数据的智能化查询分析平台，既拥有统计报表的整理、维护和查询功能，又有着比一般社会经济地理信息系统强大的数据分析功能，而由于考虑了地理空间数据，也有比一般的统计专业软件数据展示手段更加丰富的空间可视化功能，更有空间分析功能，为专业统计数据的整合和分析提供时间、空间和主题三个维度。吸取了不同软件系统的优势，但仍需保持小而精。

2.2 元数据库如何设计

可以运用数据仓库技术思想，设计方便存储和查询的统计数据库，将地理图形数据、统计报表、人口普查、经济普查和农业普查数据，以及其他专题数据统一存储管理。面对复杂的数据来源和数据结构，为使得查询方式更加灵活和多样化，可以通过时间、空间和主题三个维度来定位一笔统计数据。元数据库的设计，必须应对统计报表制度的不断改变，使得今后的修改越小越好，这需要进行不同年份同种含义指标的智能匹配。

2.3 一笔统计数据的工作流程是什么

各种各样、不同时期的统计报表，如历年的年、月、定报、普查数据等能够智能地导入，并实现服务器端的元数据库维护；在查询时，又需快捷地查询年报、月报和定报，并满足用户自定义的查询条件，实现统计数据按其时间维、空间维和主题维（即 101 表分组属性）之间的灵活查询、汇总、比较和展示。

2.4 应该包括哪些数据分析手段

考虑到业务人员进行数据分析时不同层次的需求，需要提供经典的描述性统计、相关分析、线性回归和主成分分析等，又需统计技术的发展，适当引入综合评价、灰关联分析、BP 神经网络预测和模糊聚类等数据挖掘算法，为用户在不同条件、不同深度分析和评判数据提供多样化的选择；同时，为适应区域经济分析需要，纳入常用的社会经济指标分析方法，为研究区域经济发展的均衡性、协调性和多样性提供方便；最后，考虑到空间经济学和空间人口学等学科发展对统计业务带来的前瞻性需求，应借助 GIS 功能，集成若干空间统计分析技术。

2.5 如何处理人口统计

随着老龄化时代的到来，人口统计应该作为一个重点关心的内容，可以从区域人口的可持续发展出发，精心设计集预测、预警、预报功能的基础模型生命表分析平台，包括人口数量、素质、结构和分布等要素及其组合信息指标在未来给定年度的变化趋势的预测，甚至可以进行区域适度人口容量的测算。

2.6 扩展性问题

构建适合于各级统计部门应用需要的、基于插件的功能可扩展的平台，有利于应用平台的统一、资源共享和互操作，可以根据实际情况，灵活地定制所需要的功能，确保系统的可持续发展。

参考文献：

[1]蒋正华.全国和分地区人口预测[M]. 北京：中国人口出版社，1998.

[2]张羚广，蒋正华.人口分析与信息处理技术[M]. 北京：中国统计出版社，1996.

[3]陈述彭，鲁学军，周成虎.地理信息系统导论[M].北京：科技出版社，1999.

（责任编辑/李友平）