

网络统计数据分析

● 石远志

(成都信息工程学院 计算机系, 四川 成都)

摘要: 讨论网络统计数据分析的发展、原理与作用, 介绍网络化统计数据处理技术、网络统计数据的分析工具和分析方法。

关键词: 计算机网络; 网络统计; 数据分析

中图分类号: C81 **文献标识码:** B **文章编号:** 1005-5762(2004)02-0052-03

一、统计数据分析的发展

计算机网络技术既是一种新技术, 同时也是一种重要的生产力。计算机网络技术正改变着人类的生产、生活以及社会经济事务活动, 各种基于计算机网络的网上办公、网上采购、网上交易、电子商务、网络银行、电子结算、网上海关、网络学院等蓬勃发展。反映和监控这些活动的统计数据分布广泛, 数据的规模、数据之间的关系比以往任何时候都更庞大、更纷繁复杂。传统的统计工作和统计数据的处理分析方法已显得跟不上时代发展的步伐, 统计面临着革命性的变革。统计改革和发展的重点就是将传统统计与计算机网络技术紧密结合, 逐步实现网络化统计。

计算机数据库技术及数据处理技术的发展和运用, 尤其是计算机网络技术的广泛应用, 使网络数据库技术与数据挖掘技术迅速发展和成熟。传统的数据处理已由手工处理转为数据库处理和网络数据处理, 作为反映社会经济及运行情况重要信息的统计数据当然也不例外。统计数据的采集、存储、分类、汇总计算处理各个方面都越来越多的通过计算机网络完成。统计数据的存储由传统的单机集中式数据库存储逐步转化为网络分布式数据库存储, 统计数据分析也由本地分析扩展到对远程的同构或异构数据库数据的分析。另一方面, 统计预测和决策分析往往要求数据是综合数据, 它不仅需要本单位内部各部门的相关数据, 还需要有关的外单位数据, 而这些数据大都分布存储在异地的不同计算机系统之中。决策分析不但需要反映当前情况的数据, 而且还需要历史数据, 以便分析变化趋势, 进行决策。再者, 人们利用各种统计

数据与信息进行生产流程控制、市场预测, 决策支持的要求越来越高, 这就从客观上要求统计分析必须基于计算机网络, 把大量相关而又分散的统计数据通过计算机网络集合到一起进行近程或远程的全面、高效和准确分析。这种基于计算机局域网和广域网对处于本地或远程的统计数据进行的分析即为网络统计数据分析。网络统计分析是统计现代化的标志之一, 传统的统计数据处理和数据分析正逐步迎来网络化时代。

二、网络统计数据分析的作用

以计算机通讯和网络为平台, 构成了现代统计工作运行的基本物质技术基础。通过计算机网络进行统计数据的收集、存储、整理、汇总、维护, 以及建立面向应用的大型数据库和数据仓库是网络统计数据处理的基本任务, 而网络统计数据的分析及决策支持则是网络统计数据处理的根本目的和重点。通过网络进行统计数据的处理和分析, 在数据的时效性、准确性和完整性等各方面的质量提高很多, 从而使统计数据反映的社会经济情况将更加及时、全面、准确和深刻。网络统计数据的分析对于深度挖掘统计数据中蕴涵的广博信息, 将统计数据转换为对社会经济及人们生活有用的信息资源, 帮助人们监控经济及生产活动, 进行市场预测, 辅助决策支持等方面都起着非常重要的作用。

三、网络统计数据分析的原理

网络统计数据分析离不开通过计算机网络进行的统计数据收集、整理和分类工作, 离不开面向应用的分布式统计数据库或数据仓库的建立, 这是数据分析的基础。计算机网络则是网络统计数据分析的介质和载体, 是数据

收稿日期: 2003-06-16

作者简介: 石远志(1965-), 男, 成都信息工程学院计算机系讲师, 研究方向: 计算机软件设计与开发。

和信息通讯的管道,是网络统计工作的平台。网络统计数据进行分析必须严格依照科学的统计分析理论和方法,利用目前占主流地位的统计分析软件所提供的各种统计分析过程和方法,通过计算机网络来对处于本地和异地、大量相关而又分散的、即时的和历史的数据进行综合全面的分析,并将分析结果以各种分析报告和图表的形式呈现在前台或远端平台,为人们进行决策提供科学的有力支持。

四、网络化的数据处理技术

目前,网络化的数据处理大致可分为两大类:操作型处理和分析型处理。操作型处理也叫联机事务处理(ON-Line Transaction Processing 简称 OLTP),是指在单机尤其是联网计算机上对数据库连机的日常操作,通常完成数据的查询和修改、维护等基本操作,这是数据处理的基本形式。为了充分发挥统计辅助决策的功能,完成数据的深度信息挖掘,适应信息社会化的要求,人们更加需要对数据进行分析型处理,分析型处理是数据处理的高级形式。分析型处理也叫联机分析处理(On-Line Analysis Processing 简称 OLAP),在这种高级的处理中为了使数据分析的结果及时、准确、全面有效,就需要访问网络上的大量分散在同构或异构数据库尤其是网络数据库中的即时和历史数据。在这种分析处理中,不仅传统的信息、咨询、监督功能得到了体现,而且统计的预测与辅助决策功能更是得到了充分的体现。

五、网络统计分析

统计信息系统是社会信息系统的一个重要组成部分,对信息的建立、管理和应用程度是衡量一个国家、社会或单位信息化程度高低的重要标志,也是衡量其管理水平是否现代化的标志。从微观的角度来讲,信息系统的应用任务一般分为三个部分:数据访问、应用系统处理逻辑和结果呈现。而在一个多平台的系统内,依照以上三项任务在系统内各平台划分不同的策略可有五种形式,这五种形式也是实现网络统计分析可采用的基本形式。

1. 分布式数据访问模式

在网络远端以文件服务器方式提供数据集中管理,而部分数据则保存在本地的前台系统,数据处理和呈现都在本地进行。这种方式可用在一些小型信息系统中,使部门数据在本地编辑修改和查询,而将高层次数据保存在远端,在进行数据分析和生成报告时既访问本地数据也访问远端数据。

2. 远端数据访问模式

该方式将所有数据都放在网络远端系统中,由远端系统进行数据管理,而所有的处理和结果呈现都由前台来完成。如在集中式的数据存放信息系统中,可将所有部门的数据都集中地编辑修改并存储在远端,在进行分析时,将数据从远端载入前台并暂存于前台或直接通过远端库服务功能来访问数据,从而完成数据的处理和分析。

3. 分布式处理逻辑模式

该方式将网络远端系统用作文件服务器,同时也使其执行部分处理逻辑,而主要的处理逻辑和结果呈现都在本地实现。如在信息系统的中心数据的预处理中,集中编辑和管理的数据由远端系统作预处理,前端逻辑只对经远端系统预处理过的数据作最后处理及结果呈现,这样就减少了数据的传送量,节省了时间和资源。

4. 远端结果呈现模式

在该方式下将数据访问和处理放在远端完成,本地只完成结果呈现。该方式常用在慢速网络数据处理中。这种在远端完成系统分析和汇总数据的处理方式,能大大减少在网络上传送的数据量,从而极大地节省网络资源。

5. 分布式呈现模式

该方式将数据访问和处理、部分结果的呈现交由远端完成,数据分析的结果呈现在前端的 GUI 界面,部分输出传送到远端的打印机或绘图仪上。其它的数据呈现则由本地完成。

在进行网络统计数据处理和分析时,必须根据信息系统中数据的模型和特点以及数据分析的实际情况,有效选择数据的访问和分析方法,合理高效地实现网络统计分析结果的呈现与应用。

六、网络统计分析的工具和方法

网络统计分析可采用的工具软件有:统计分析及决策支持系统 SAS、社会统计分析软件包 SPSS 以及 BMDP 系统、STATISTICA 系统、P-STAT、TSP、MINITAB、RDAS 等系统。

在进行网络统计分析时,可首选 SAS 系统。使用 SAS 系统进行网络统计分析分析与决策支持的主要操作步骤如下:

1. 远程连接的建立

用户通过网络进行远程的数据处理和分析时,应先用 SAS/CONNECT 连接两个 SAS 进程(每个进程都运行在各自的平台上),也可同时连接到多个远程 SAS, SAS/CONNECT 提供多远端连接功能,即一个客户机可以连接到多个服务器。通过连接到不同服务器,用户可以访问到在不同平台上的不同数据。当连接建立之后,就能访问到两个平台所提供的服务。建立本地和远端连接是网络数据处理与分析的基础。

2. 远程计算与处理

在远程连接建立后可通过 SAS/CONNECT 模块进行远程数据计算与处理。在 SAS 系统的远程计算服务过程中,可充分利用远程计算资源:包括硬件、软件和数据资源。在进行处理时,数据可以就地处理,也可以将应用的任何部分或全部放在远程系统上运行,而处理的结果则返回到本地。还能服务到连接在远端的硬设备,访问到远端机上的软件资源,可让图形程序由远端机执行而生成的图形结果

统计学研究对象与方法辨析

——兼与熊健益同志商榷

● 栗方忠

(东北财经大学 统计系, 辽宁 大连 116025)

摘 要: 在研讨统计学研究对象问题上, 必须明晰统计理论、实践、对象和方法这几个相关的概念及其关系。统计理论与统计实践研究对象是一致的。统计学研究对象不是统计方法, 而应是社会经济现象总体的数量方面。

关键词: 统计理论; 统计实践; 研究对象; 方法

中图分类号: C8

文献标识码: A

文章编号: 1005-5762(2004)02-0054-02

《统计教育》2002年第6期发表的熊健益同志题为“关于社会经济统计学的存在性及其研究对象的思考”的文章(以下简称“熊文”), 读后颇受启发, 许多观点是合适的。但是, 关于统计学(指社会经济统计学, 下同)研究对象的表述却不能苟同。“熊文”称“社会经济统计学的研究对象可以表述为社会经济现象总体数量资料的搜集、整理、分析的原理、原则和方法”。这一表述的实质是统计学研究对象是“统计方法”。从其整个论述中可见“熊文”混淆了统计理论、实践、对象和方法的概念, 所以, 在研讨统计学研究对象问题上必须明晰统计理论、实践对象和方

法这几个相关的概念及其关系。

众所周知, 统计实践即统计工作, 其过程是对社会经济现象总体数量方面进行调查占有材料, 加以科学整理, 进行综合分析, 做出正确结论, 上升到对事物本质和规律性的认识的活动过程。在统计工作的实践活动过程中, 除获得对所研究的对象有本质和规律性认识之外, 还同时得到如何研究对象的一系列原理、原则和方法的感性认识。对这些感性认识进一步加以整理, 形成概念、判断和推理, 得到深层次的认识, 形成合乎逻辑的结论, 上升到理性认识, 并使之系统化形成统计科学, 它再回到实践中

收稿日期: 2003-09-01

作者简介: 栗方忠(1932-), 男, 辽宁鞍山人, 教授, 研究方向: 统计学理论、方法及其应用。

则显示在本地。

3. 计算结果的呈现

在进行远程数据分析时, 利用 SAS/CONNECT 软件和 SAS/GRAPH 软件, 可以在远端运行 SAS/GRAPH 程序而把图形结果显示在本地的 GRAPH 窗口。通过图形化的结果显示, 可使决策者更加形象地分析数据处理的结果, 高效、准确的进行决策。

4. 远程数据的上传与下载

在远程网络数据分析时, 可根据需要通过 SAS/CONNECT 模块提供的完整的数据传输工具实现数据的远程传输, 即使用 SAS/CONNECT 的 UPLOAD 将数据或数据的结果与分析结果从本地上传到网络远端, 或使用 SAS/DOWNLOAD 从网络远端下载数据到本地, 再由本地的

SAS 负责计算、处理和分析。

5. 远程库访问

在远程网络统计分析过程中, 如果需要访问远程 SAS 数据集、同构或异构的数据库, 可通过 SAS/CONNECT 或 SAS/SHARE 模块来实现。这两个模块可实现对远端 SAS 数据集, 外部数据库管理系统(如 ORACLE, SYBASE 等), SAS Catalog 和外部文件的透明访问。还可实现在本地定义库名指向远端机的 SAS 库服务, 实现在本地 SAS 中透明访问远端数据。

在网络统计分析中, 采用其它的分析工具: 如 SPSS 以及 BMDP、STATISTICA 等系统。其分析方法与操作过程与采用 SAS 系统大致相同。