李永平 著

数据处理方法与技术

国防工业出版社
National Defense Industry Press

数据处理方法与技术

李永平　著

国防工业出版社

·北京·

图书在版编目(CIP)数据

数据处理方法与技术/李永平著. —北京:国防工业出版社,2009.4
ISBN 978-7-118-06253-3

Ⅰ.数... Ⅱ.李... Ⅲ.数据处理 Ⅳ.TP274

中国版本图书馆 CIP 数据核字(2009)第 035679 号

※

国防工业出版社出版发行
(北京市海淀区紫竹院南路 23 号 邮政编码 100048)
北京奥鑫印刷厂印刷
新华书店经售
*
开本 710×960 1/16 印张 10 字数 180 千字
2009 年 4 月第 1 版第 1 次印刷 印数 1—4000 册 定价 20.00 元

(本书如有印装错误,我社负责调换)

国防书店:(010)68428422 发行邮购:(010)68414474
发行传真:(010)68411535 发行业务:(010)68472764

前 言

随着人类社会的发展，日常事务处理的难度越来越大，每天都有大量新的数据出现，等待着人们去处理。使用计算机作为工具进行数据处理已经成为现实，目前出现大量的计算机数据处理工具，如电子表格软件 Excel、Lotus-123；小型数据库 VFP、Access；大中型数据库 SQL Server、Oracle；通用统计软件 Spss、Sas、Dps；还有针对医学、农业等研究用的统计软件。人们在进行数据处理工作过程中需要掌握数据处理的基本概念与工作要素，需要应用一些方法与技术，以便提高数据处理的速度与质量。

本书作者从事过 20 多年的数据处理工作，主持过几十次的大、中型数据处理项目的实施工作，实践了很多数据处理的方法与技术，同时积累了大量的数据处理工作经验与体会。在这些项目实施过程中，将已有的数据处理方法及技术与自己一些新的想法与实现技术结合起来，使项目实施达到令用户满意的结果，受到用户的好评。本书在编写的过程中，结合一些简单的例子和众多的图表，去化解一些比较抽象的概念，尤其对初学者来说更加容易阅读。

本书分为 7 章，共安排了数据处理概论、数据、编码与数据库设计、数据输入、数据编辑、常用统计方法及实现、典型例子数据处理分析与设计等内容。第 1 章主要描述数据处理的基本概念、硬软件结构、数据处理方式；第 2 章主要描述数据的表示方法、计算机中数的表示及数据存储方式；第 3 章主要描述编码的作用、编码的原则、编码的技术、编码的实现、数据库的设计方法与步骤；第 4 章主要描述原始数据组织与输入流程的设计、输入界面的分析与设计、如何保证数据输入的正确性；第 5 章主要描述数据编辑的工作原理、数据编辑的规则制定、数据编辑程序的编制；第 6 章主要描述常用的统计方法及如何用计算机工具进行实现；第 7 章通过一个真正的实例，“满意不满意”选票统计分析与设计，描述了整个项目实施的工作过程。

本书适合于计算机编程人员、统计专业技术人员及从事数据处理的工作人员阅读。

在数据处理项目实施的过程中认识了很多的朋友，每一个项目的成功实施都包含了他们的智慧与辛勤的汗水，在此一并表示感谢。同时，感谢统计学专家程跃秋老师，为本书提供了与统计知识相关的素材。

由于作者水平有限，加之时间仓促和工作繁忙，书中的不妥之处在所难免，恳请广大读者不吝赐教。

李永平

2009 年于温州职业技术学院

目 录

第 1 章　数据处理概论

1.1　数据处理实例介绍

1.1.1　商贸城投标

某市一个大型的综合性商贸城投标，共有 500 个摊位，参加投标者 5342 人。投标地点在一所中等专业学校的大操场上，投标时间为上午 9:00—10:00。10:10 开标并用计算机进行录入统计排序，按投标金额的大小选定前 500 名，并要求于 11:45 在投标现场将结果公布于众。

虽然投标的数据很简单，每张投标单只有两个数据，一个是投标号码，一个是投标金额，但要在短短的 1 个半小时内将 5342 张票进行整理、装订、编号、输入计算机，并要求绝对正确，是一个很难的问题。

所做的工作主要有以下几方面。

(1) 每 100 张投标单装订成本，并编制本号和流水号(原投标号已经混乱)。

(2) 按本输入每份投标单的投标号和投标金额。

(3) 进行复录并核对、校正。

(4) 数据编辑，核对是否有相同投标号。

(5) 合并文件并按投标金额进行从大到小排序。

(6) 打印投标结果。

由于初次进行这方面的数据处理，一直至下午 13:30 才出来结果，造成现场出现一些混乱局面。主要原因包括以下几方面。

(1) 事先没有对整个工作流程做周密的安排。

(2) 对参加管理和输入的人员没有进行事先培训。

(3) 核对是否有相同投标号的算法不好，运行十几分钟后还没有出结果，最后按新的算法重编程序。

(4) 录入员是临时抽调的学生，输入速度慢和误码率高，使录入和校对时间增加。

(5) 设备落后(8088CPU 机器，无网络，只有一台 9 针打印机)。

(6) 经验不足。

1.1.2 公务员录用考试考务安排

近几年，党政机关都要通过公务员录入考试选拔后才能进人，公务员录用考试的热度很高，一个地区将近有30000多人报名参加录用考试。通常，从报名到考试只有不到一个月的时间，在多达100多个职位、几十个专业、不同类别、不同性别等情况下，需要对各种有关证件(毕业证书、身份证、户口本、就业协议书等)进行审核，保证考生的合法性。专业考试内容不同、专业要求不同，因此对报名数据的正确性和考务安排的时间有很严格的要求。由于数据量大(每个考生需录入的资料多达32项)，分点工作多(各市县自行完成报名和数据录入工作)，其录入考生资料的正确性很难控制。录入一次加人工较对，其差错率仍达 3%以上，且不易发现。录入两次加计算机自动校对，也有一定的差错。经过多次机器和手工的校对修改后，在分发准考证时，仍有几个人的专业对不上，甚至还有遗漏，考生连准考证也拿不到。虽然可以对数据进行修改，重新输出准考证，但考场已经安排，试卷已经征订，所有的考务资料都已经印刷完毕。重新修订不仅工作量大，而且如果个别资料没有重新修改考务资料，在考试时可能还会引起混乱。

造成以上失误的原因如下。

(1) 对原始报名表审核不严格，报名表内容本身有错误。

(2) 两位考生的报名资料，由于回形针的原因，夹在一起，录入员认为只有一位考生，造成遗漏，校对时也当作只有一位考生的资料。

(3) 时间过于急迫，校对工作没有做到位。

(4) 数据来自多方，管理难度大。

(5) 工作人员对这项工作的重视度不够。

当前，已经采用网上报名、网上缴费、考生自己从网上下载打印准考证的方法，避免了很多环节中可能造成的错误。

因此，大型的数据处理看起来很简单，但实际上实施起来有一定的困难，假如没有充分的估计和预料，对这项工作的重要性认识不够，可能会出现不可收拾的场面。同时，在整个数据处理的过程中，使用一种严密的管理制度和良好的工作方法才能保证数据的正确性与时效性。

1.2 数据处理的基本内容和特点

数据处理是指使用电子计算机对大量的原始数据或资料进行录入、编辑、汇总、计算、分析、预测、存储管理等的操作过程。

1.2.1 数据处理的基本内容

数据处理的基本内容如下。

(1) 对所需的数据进行收集整理，按一定的格式输入，并保存在存储介质上。

(2) 在输入数据过程中，对原始数据进行检查、逻辑判断、查错、修改和简单的算术运算。

(3) 对录入的数据进行分类、合并、逻辑校正、插入、更新、排序检索等操作。

(4) 对数据汇总、分析、制表打印、存档等。

(5) 建立信息数据库，便于今后使用。

1.2.2 数据处理的特点

数据处理在很多的场合都有应用，如考务安排、成绩统计、选票统计、投标、会计业务处理、人口普查、超市管理、银行存款和取款等，其特点如下。

(1) 数据量大。

(2) 算法简单，主要是加、减运算、排序、分类及汇总。

(3) 数据要绝对正确。

(4) 事先的工作要准备充分。

(5) 牵涉面广，经手的人多。

(6) 有一定的时间性。

1.3 数据处理系统的硬件和软件构成

1.3.1 数据处理系统的硬件构成

在数据处理中，计算机硬件设备是一个必备条件，它是快速处理数据的保障。为了能够满足大型的数据处理，应该采用局域网或用多级局域网形式进行全国性的数据处理和统计。

一般的数据处理可采用集中式数据存储方式进行，其结构如图 1.1 所示。

集中式数据存储方式处理适合于原始资料比较集中、数据量适中的数据处理，其结构管理比较方便，容易控制。大部分的数据处理都可采用这种网络结构。

对于数据量很大、原始资料分散、工作点多的情况(如人口普查)，就采用多级局域网结构方式，如图 1.2 所示。

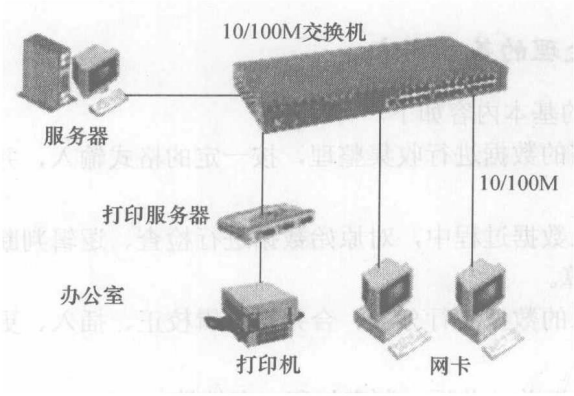

图 1.1　集中式数据处理网络结构图

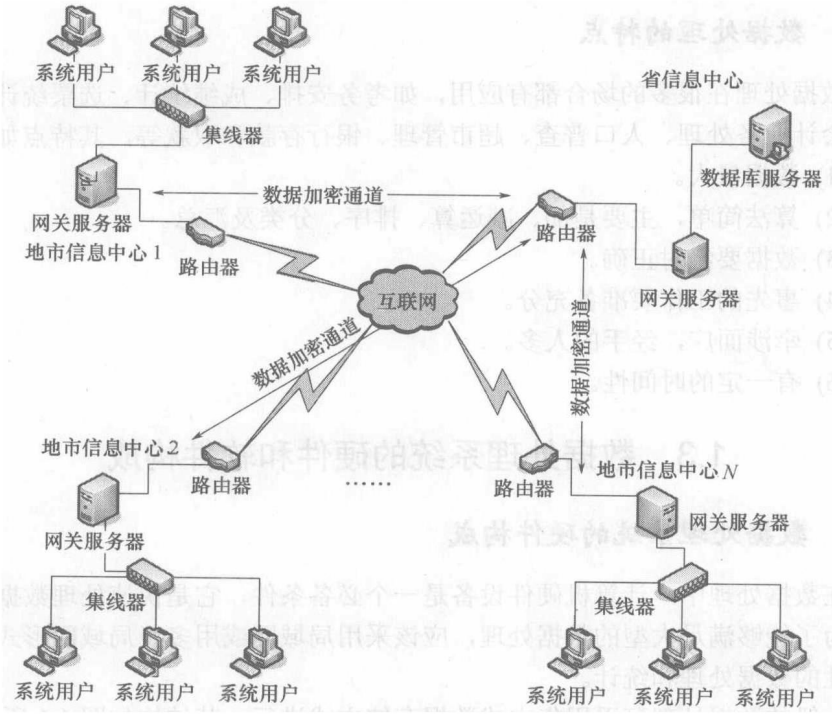

图 1.2　多级数据处理网络体系结构

1.3.2　数据处理系统软件平台

数据处理系统的软件平台，主要是操作系统和数据库。操作系统和数据库的选择主要根据具体的数据处理项目的大小、安全性要求等，还要考虑用户对

所选软件的熟悉程度和现有的硬件设备条件限制以及软件的价格等方面的因素，用户可根据自己的需要选择合适的操作系统和数据库系统。

1. 操作系统选择

目前，常用的操作系统有 Microsoft 系列操作系统、Unix/Linux 操作系统。

(1) Microsoft 系列。微软系列操作系统主要包括 Windows Server 2003 和 Windows 2000 Server，主要用于局域网，结合.net 技术，以浏览器/服务器(B/S)的方式进行数据处理。

微软系列操作系统都是图形界面的操作系统，具有友好的人机交互界面，可以建立安全可靠的数据库系统，具有各种安全防护和容错功能，保证信息的安全性和有效性。

(2) Unix/Linux 操作系统。Unix 是一个多用户、多任务的分布式网络操作系统，也是目前使用最广泛的操作系统之一，安全性能好。

Unix 是适用于各种机型的主流操作系统。它具有丰富的应用支持软件，良好的网络管理功能，能够提供真正的多任务和多线程服务，具有优异的内存管理、任务管理性能及 I/O 性能，是所有操作系统中的首要选择。尤其是许多客户选择客户机 / 服务器(C/S)模式的数据库管理系统(如 Oracle)时，均选用 Unix 操作系统。

Linux 也是一个多用户、多任务的操作系统，它是 Unix 操作系统的变种，兼容于 Unix。Linux 具有免费性、开放性及强大的网络管理功能，受到广大用户的青睐。

2. 数据库系统选择

目前，常用的数据库系统有Oracle、SQL Server、Sybase、Informix、Visual FoxPro 等。在选择数据库系统时，主要考虑其操作界面、数据的完整性和一致性、功能参数等。

(1) Visual FoxPro。Visual FoxPro(VFP)是微软公司的产品，它是一个面向对象的关系数据库管理系统。

VFP 使用了 Rushmore 技术，大大提高了查询检索速度；由于采用了结构复合索引文件，改变了单一的索引文件结构；在 VFP 中可以使用 SQL 命令，使得程序代码更少，并且能从一张或多张表中更快地查询到数据。

在数据操作方面，VFP 具有灵活多样的数据交换手段，支持众多的与其他应用程序进行数据交换的文件格式，如文本文件、电子表格等。这样，外界的数据可以很方便地加入到 VFP 中来，而 VFP 中表的数据也可以很容易地交付其他应用程序处理。

在程序设计方面，VFP 可以不必编程，或只要编写很少的程序代码，就能够很快地创建出可视化的应用程序。用 VFP 开发出来的应用程序有更高的可靠

性，还可以编译成为一个在 Windows 下独立运行的可视化应用程序。

在使用操作方面，VFP 具有功能完善的集成环境，操作和维护都很容易。在局域网中，它以文件服务器的工作方式为客户机所使用，所以网络传输量比较大。

在数据表方面，VFP 允许每个记录有多达 256 个字段，并同时使用 32767 个工作区。

VFP 适合于中、小型的数据处理项目，我国早期的数据处理系统都是基于 dBase、 FoxBase 和 FoxPro 开发的，因此目前仍有很多技术人员基于 VFP 系列开发数据处理系统。

(2) SQL Server、Access。SQL Server 是微软公司的产品，它是大型网络数据库管理系统，对硬件适应能力强、用户界面友好、性能可靠、使用方便，是电子商务、在线商店、大型管理信息系统首选的数据库管理系统。

SQL Server 是基于 C/S 模式的数据库管理系统，在 C/S 模式中处于服务器端，用于存储、提供、管理数据；在客户机端可以使用 VB、ASP(动态服务器网页)等可视化工具来开发。用 ASP 或.net 技术也很容易实现 B/S 的数据处理功能。

Access 数据库管理系统能支持多用户访问，Access 采用文件服务器的模式，会加大网络的流量，从而使访问的速度较慢，所以只能用于中、小型的数据处理系统。

(3) Oracle。Oracle 是标准 SQL 数据库语言的产品。它在数据管理、数据完整性检查、数据库查询性能及数据安全方面的功能强大，并且在保密机制、备份与恢复、空间管理、开放式链接及开发工具方面提供了与众不同的手段。例如，在数据安全机制方面，Oracle 采用表或记录加锁的方法来禁止同时写数据；使用扩大共享内存的方法来减少读写磁盘的次数，防止数据访问冲突；使用快照的方法进行备份，快照脱离原表，使得对于某些远程查询操作的数据可以在本地机上执行，从而减少了网络传输量。

SQL Server 及 Oracle 允许的数据库容量多达 1TB(1TB=1024GB)，而对于记录长度、一个记录中所能包含的字段个数等参数未加限制，仅受库容量大小的限制。

数据处理应用软件一般根据项目的要求进行定制开发。对于全国或全省性的数据处理项目都是由上级部门开发后下发，也有些用户根据上级部门要求上报的数据结构自行编制部分程序，方便自己的使用。当然，某些项目仍然要靠用户自己开发。个别很小的项目也可使用 Excel 或 Lotus-123 电子表格软件进行处理。

1.4 数据处理方式

数据处理方式是指计算机实现数据处理过程的方法。数据处理方式可分为单级数据处理和分级数据综合处理。单级数据处理又可分为批处理方式和联机实时处理方式两种；分级数据综合处理是根据一定的管理体制，自下而上进行数据汇总工作。

1.4.1 批处理方式和实时处理方式

1．批处理方式

批处理方式用于对数据处理的时间响应要求不是很高，数据处理点比较分散，无法实现联网或投资不允许等情况。批处理方式是定时将收集过来的数据输入计算机，并作出相应的处理。批处理方式投资少，稳定性好，但在数据汇总方面具有滞后性大等缺点。

批处理方式只要在时间许可的情况下，不需在计算机和人员方面进行大量的投入，有些系统只需要一人一机就能实现。若时间性要求很强，可采用局域网的技术，通过多人输入数据，达到数据处理的时间要求。

批处理方式的特点是费用较低，且可以较充分地使用计算机系统，一般用于以下几种情况。

(1) 固定周期的数据处理。

(2) 需要对大量的来自不同方面的数据进行综合处理。

(3) 需要将一段时间内积累的数据进行处理。

(4) 无法进行联机实时处理时。

2. 实时处理方式

某些数据处理系统要时刻关注其汇总结果，必须要采用实时处理系统。实时处理系统能随时反映数据处理系统的瞬间状态。实时处理系统是当数据一旦发生，就要立刻输入计算机，并作出相应的处理，如银行的存款和取款、大型且重要的选票统计。实现实时处理系统的关键是输入系统。银行、超市等行业使用数字输入设备和工作人员键盘输入相结合，某些系统采用网上即时输入数据或以涂卡的方式填写机读卡后送入读卡系统。实时处理系统必须要采用 Internet 技术，要保证网络畅通，系统稳定性要高，一旦系统产生瘫痪，将会严重影响工作。

联机实时处理方式一般用于以下几种情况。

(1) 需要迅速反应的数据处理。

(2) 负荷易产生波动的数据处理。

(3) 数据收集费用较高的数据处理。

数据处理方式不管采用批处理方式还是实时处理方式，一旦新的数据产生，都要及时录入计算机并及时处理，尽量减少由于数据输入的时间延迟而造成数据处理结果的滞后。

1.4.2 分级数据处理

由于某些数据处理系统牵涉的面广、数据量大，又要考虑时间，因此某些大型的数据处理要采用分级综合处理，如全国人口统计、全国职称外语考试等。分级数据处理可分为集中统一超级汇总处理和逐步分级综合处理。

1. 集中统一超级汇总处理

集中统一超级汇总处理是指将所有各基层收集过来的原始数据(如职称外语考试的报名表)都集中到最高一级数据处理中心进行数据录入、编辑、修改、汇总。这种处理方式的优点是数据的计算机处理工作点集中，数据的正确性和录入质量有可靠的保证；缺点是由于原始数据是以纸质方式来记录，给运输、验收等带来许多困难，若原始资料有问题，情况返回也很不方便。

当前，计算机已经非常普及，计算机技术人员也不缺乏，因此数据收集、录入均可放在基层完成。经录入的数据，经过编辑、校对后可传送到最高一级数据处理中心，最后进行统计汇总和分析工作。

2. 逐步分级综合处理

如果大量的基础数据都统一传输到某一台计算机进行最后的数据汇总，可能在数据存储、运算时间及软件平台上会难以实现。如我国人口普查，基本数据达 13 亿多条记录，又必须在指定的时间公布结果，用一台计算机运算很难达到理想的效果，甚至无法实现。因此，必须要做到各级统计部门遵照国家统计局的统一部署，统一软件，按计划严格执行，将分层的统计结果自下而上按级上报，最后得到最终的统计结果。

逐步分级综合处理是由基层进行数据的收集、录入、编辑校对，根据上级的要求进行统计汇总，然后将统计汇总结果上报上一级数据处理中心。上一级数据处理中心将所属的基层统计汇总数据进行第二次统计汇总，以此类推，最后由最高一级数据处理中心得出最终结果。

由于各个层次对数据统计汇总和分析指标不同，可以按逐级的要求设置报表。一般是下一级比上一级要求汇总内容少，输出的报表简单一些。因此，在汇总出一级报表时，不是简单地将下级汇总数据进行叠加，而必须采用综合方式，设计出综合数据汇总分析报表。如在成绩分析中，一个班只进行该班的平