# 软件世界
# SOFTWARE WORLD

（双月刊）

1984 年 9 月创刊

社　　长　周慕昌
主　　编　周锡龄
副 主 编　应　明　孙毓林
主　　办　机电部计算机与微电子发展研究中心
编辑出版　中国计算机报社《软件世界》编辑部
地　　址　(100846)北京万寿路 27 号
印　　刷　机电部电子情报所印刷厂
国内总发行　北京报刊发行局
订　　购　全国各地邮局
邮发代号　82-469
刊　　号　ISSN 1000-8926/CN11-2105
广告许可证号　京海工商广字 004 号
每期定价　1.50 元　全年定价　9 元
出版日期　1993 年 5 月 21 日


## MAIN CONTENTS



# 目　　录

1993 年第 3 期(总第 79 期)

# MIS 建设的困惑和出路

高复先

近些年来，我国管理信息系统(MIS)的建设遇到许多困难和问题，例如：开发效率低，维护投入大；急用项目无力开发，应用积压严重；分散开发的系统各自封闭，给其后的集成造成困难；硬件投资大，联网工作限于物理上的连通而缺乏数据的统一管理与共享，等等。MIS 理论与方法的引进，缺乏较全面的综合与分析，当遇到困难和问题时，往往想到要寻求一种"最好的"方法，好像有了这种"法宝"就可以摈弃其他方法而一举成功。比如，学了原型法就否定结构化方法；看到数据库的重要，就整年搞"数据库设计"；听说"面向对象"的方法是最新的，就急着搞"最先进的开发"。实际上，MIS 的理论像其他理论一样，都是从实践中总结出来的，这样的理论才能反过来指导实践。应该把总结自己的经验与学习他人的著述结合起来，了解一些近十年来国际上在这方面的总结与发展状况，探讨 MIS 建设的一些规律性东西，形成理性认识，不断实践，才能走出低谷，稳步地建设好自己的管理信息系统。

## MIS 建设需要有正确的方法论指导

詹姆斯·马丁(James Martin)于 80 年代初提出的信息工程(Information Engineering，IE)，标志着一种崭新的 MIS 开发方法论的诞生。信息工程比软件工程更为广泛，它包括了为建立基于数据库系统的计算机化企业管理所有相关的技术。信息工程的基本原理或前提是：

- 数据位于现代数据处理的中心；
- 数据是稳定的，处理是多变的；
- 最终用户必须真正参加开发工作。

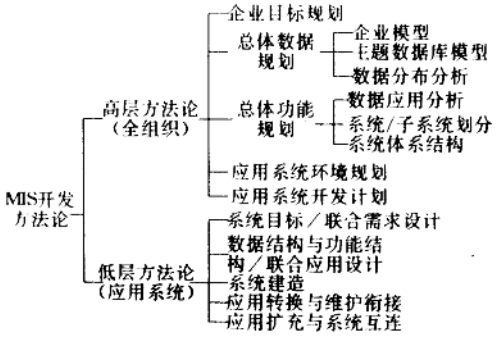
信息工程十分强调 MIS 建设的高层设计工作，即以总体数据规划为中心的总规划与总体设计。信息工程发展了关系数据库技术，提出了全组织数据环境建设的方法与步骤：数据文件(Data-Files)→应用数据库(Application Data Bases)→主题数据库(Subject Data Bases)→信息检索系统(Information Retrieval Systems)。信息工程吸取了以结构化技术为中心的软件工程的精华，克服了其不足之处；提倡组建先进的开发小组，掌握原型方法和四代语言，进行高质量、高效率的开发，但十分强调这些必须建立在企业模型、数据模型的深入研究和科学的信息系统体系结构的规划之上。

80 年代末、90 年代初，国际上激烈的企业、行业间的竞争，要求业务发展变化的步伐加快，计算机信息系统的更新必须能跟上业务的变化，不允许 MIS 阻碍业务的发展，于是快速应用开发(Rapid Application Development，RAD)应运而生。RAD 基础是：第四代语言和集成化计算机辅助软件工程(I-CASE)工具；最终用户全面介入开发过程，通过选拔与培训组成的包括最终用户人员参加的联合需求计划小组和联合应用设计小组，成为开发队伍的主体；富有成效的开发工作管理，开拓具有革新精神的管理方法；基于信息工程方法论的最佳生命周期——需求计划阶段，用户设计阶段，系统建造阶段，系统转换阶段。必须强调，RAD 是项目级的开发方法论，而不是全组织 MIS 建设的方法论。只有在信息工程的 MIS 体系结构的基础之上，RAD 才有用武之地；离开 MIS 体系结构和总体数据模型，是搞不好 RAD 的。

IBM 公司 1989 年 9 月宣布了应用开发周期(AD/Cycle)方法论，1991 年出版专著介绍。AD/Cycle 代表了应用软件自动建造、应用开发框架发展的重要里程碑。同当前国际上众多 CASE 工具和方法的杂乱无章情况相比，它使整个应用开发过程的规划、设计、实现与管理等发生根本性的变化。过去的 20 年，开发自动化已在多方面获得进展，但是，不可忽视的问题是，缺乏普遍性的标准和规范化。例如，用一个工具生成代码，但设计规格说明是由另一个工具完成的，这就需要复杂的界面处理，并影响到开发过程的明晰性。AD/Cycle 终于制定出一个总的框架(Framework)，组成综合的开发环境，建立起规范化过程并可在不同的工具间转换。在 AD/Cycle 框架中，应用开发是从企业目标、战略规划开始的，而不是从一个应用项目的需求开始。每一开发阶段的工作都通过共享资料，与下一开发阶段相衔接，即元库(Repository)和控制它的元库管理系统(Repository Manager)，集成化的开发工具在 AD 平台(Platform)上运行。值得注意的是，AD/Cycle 的企业模型建立，

高层的数据模型建立，是与信息工程方法论相一致的。

MIS 开发方法论是关于 MIS 建设的一整套具有指导意义的理论与方法保证体系，借鉴 80 年代发展起来的上述国际主流方法论，即信息工程方法论，结合从事 MIS 开发的体验，我们认为：MIS 开发方法论应明确提出由高低两层组成，高层方法论指导总体规划与高层设计，低层方法论指导应用项目的系统分析设计与实现。该方法论框架如下：




高层方法论属于系统工程范畴，其掌握与执行需要系统工程师——总体规划员、数据管理员、系统分析和设计员；低层方法论属于软件工程范畴，其掌握与执行需要软件工程师——数据库管理员、系统设计员和程序员。我们强调支持工具与方法的一体性，该框架也表示了支持工具的分层概念。

## MIS 建设的总体数据规划

诺兰(Nolan)关于 MIS 发展阶段的理论，认为 MIS 发展的客观规律性表现为六个阶段：初始阶段、扩展阶段、控制阶段、集成阶段、数据管理阶段和成熟阶段。信息工程的理论认为，从控制阶段到集成阶段是一个转折点，这时要进行总体数据规划，因为控制阶段和集成阶段的核心问题，是逐步建立多个应用系统的共用数据库。威廉·德雷尔(Willian Durell)关于数据管理的论述，指出必须对企业的数据资源进行全面的规划、分析、定义与使用控制，这是诺兰模型第五阶段的具体特征、工作内容和方法，使 MIS 的建设抓到关键，打下良好的基础。

管理信息系统有五个基本部分：人员、规程、数据库、计算机软件和计算机硬件系统。计算机辅助企业管理，需要建立起稳定的数据结构，尽管企业组织机构、业务过程和活动可能变化，但这种数据结构是基本不变的。因此，MIS 建设总体规划的基础与核心是总体数据规划。

MIS 发展的不同阶段，有不同的总体规划，用以推动、控制、协调 MIS 的发展。初级阶段要选好试点性项目开发，放手让有积极性的单位从简单的数据处理做起，鼓励自行开发，培养应用开发人才，激发使用兴趣，形成使用局面，逐步提出一些简单的数据管理规范。中级阶段的总体数据规划，要规划出 MIS 总体数据模型和功能模型，形成 MIS 体系结构，制定出以数据库建设为基础的新开发项目与改造扩充项目计划。高级阶段的总体规划，要求完成或完善全组织数据资源的战略规划，按数据管理标准规范，建立起稳定的全组织的数据模型，制定应用项目改造翻新的策略方法，使全组织的数据环境发展到第四类数据环境，以支持决策活动。

## 应用开发的新结构化方法

MIS 建设的应用项目开发即子系统的设计、实现与投入运行，应在总体规划指导下进行，为此，就应注意：与总体数据规划相衔接；以数据为中心，数据设计与处理设计同步；结合使用原型法，吸引最终用户参与开发；简化传统方法及其文档；使用辅助软件工具。

新结构化方法的主要步骤：概念设计——应用系统的概要设计或需求规范说明；逻辑设计——数据库设计和计算机化的处理过程设计；物理设计或实现设计——在确定的计算机硬软环境之上进行物理数据库和程序模块设计。这三个主要步骤都要设检查点，进行必要的反馈修正。试运行与系统的转换工作，交织着系统开发设计人员、用户代表、操作人员与系统维护人员多方面的交接、试用、变更等等一系列综合性活动。

需要再次强调的是，整个 MIS 建设中的“快速开发”主要的是指在稳定的概念设计与逻辑设计基础上的建造阶段，这一点不论是否采取了四代语言或辅助工具都是不变的。从已有的开发记录统计规律看，项目开发的实现设计的工作量仅在 30%左右，大量的工作还是在这之前进行的。因此，没有较为可靠的前三步设计，第四步的“快速”是毫无意义的。

## MIS 建设的数据管理

数据管理(Data Administration，简称 DA)，是数据处理发达国家近十年来才普遍重视的建设信息系统的基本技术。DA 技术是把组织内的数据和信息作为重要的资源，像对待能源、材料和资金一样，

进行全面的规划、科学的管理和有效的使用。行业和企业信息系统建设的集成化发展，必然要求数据管理标准化，事实上数据管理标准的制定与实施，是信息系统建设的基础，全企业的数据管理标准，应成为总体方案的重要组成部分。建立数据管理标准的意义在于，将数据管理标准纳入企业管理标准，从而保证管理工作中不可缺少的数据采集、存储、交换与使用的规范化。

数据管理标准包括：数据元素标准——数据元素的识别、定义和命名；信息分类编码标准——分类编码识别，制定编码规则和码表；用户视图标准——识别、定义、用户视图组成和标识主码；概念数据库标准和逻辑数据库标准。

数据字典用来存储上述标准的全部内容，以支持各个应用项目的开发，多个项目的互连，全系统的建设、运行、维护和数据信息的使用。

## MIS 集成化开发的辅助工具

系统集成是指先搞一个时期的分散开发，形成许多可以独立运行的"信息孤岛"，再通过多种技术和巨大的投资将它们互联起来成为"信息大陆"。其实，当提出 MIS 的建设目标时，就应该确定全组织范围的信息体系结构和集成化建设方针，开始集成化系统(Integrated System，IS)的建设。西方国家到 80 年代搞系统集成的时候，已经走过二、三十年信息系统建设的漫长道路。我们讲集成化开发，是在正确方法论指导和软件工具辅助下进行的，强调工具的实用性和集成性：

- 总体数据规划设计的高层建模工具，要便于业务人员理解和使用；
- 引进数据字典系统，加强数据标准化管理；
- 支持数据设计与应用设计同步的项目开发过程；
- 支持基于应用程序分类规划的程序生成；
- 简化设计规范，尽量采用计算机化文档。

我们自 1986 年研制中文微机数据字典系统以来，首先用于开发中数据实体与处理实体规范化描述，使开发组多人工作得到统一管理与协调。其后在总体数据规划实践中，我们研制使用了辅助工具。我们在十分重视 MIS 高层辅助工具研制的同时，也注意中低层辅助工具的研制与使用。在这些工作的基础上，进行系统化扩充，在微机环境下用汇编、C 语言和编译 dBASE 3-Plus 实现的成组工具如下：

1. 总体数据规划辅助工具 DPAT(Data-Planning Aided Tool)。建立数据管理基础，保证数据定义的一致性，规范化用户视图，辅助数据分析与数据库规划(概念数据库设计)、数据模型、功能模型和系统体系结构的建立。

2. 信息分类编码辅助工具 ICAT(Information-Catalogue Aided Tool)。信息分类编码对象的识别与记录管理，编码规则的编辑、修改与查询使用，参照/编码文件结构建立与维护，标准编码本的机内存储、快速查询与输出打印，企业/行业信息分类编码标准化管理。

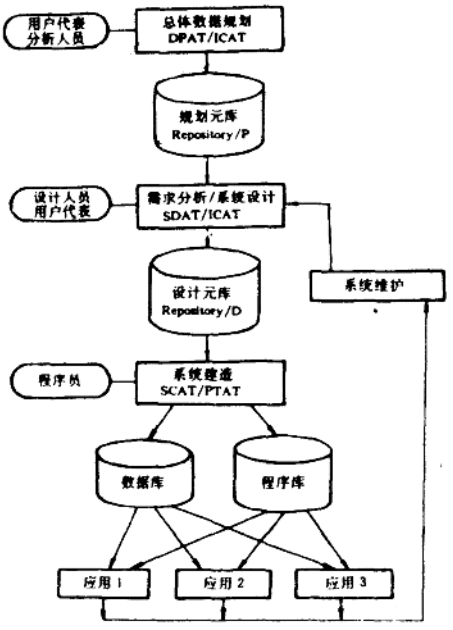
3. 结构化程序设计辅助工具 SDAT(Structured-Design Aided tool)。辅助新结构化设计方法的进程和主要文档工作，在自动转换规划元库为设计元库的基础上，提供数据实体和处理实体的规范描述与统一管理，使系统设计人员与建造人员有良好的界面，并为程序生成打下基础。

4. 系统建造工具 SCAT(System-Construction Aided Tool)。提供一组微机开发环境应用程序生成器，基于 SDAT 的设计元库，可以自动化或人机交互生成 dBASE/FoxBASE 数据结构及源程序。

5. 程序测试辅助工具 PTAT(Program-Testing Aided Tool)。主要用于辅助逻辑复杂的 dBASE/FoxBASE 程序调试，如复杂分支结构查错，多层循环结构查错，规范化程序清单列印等。

上述软件工具中，DPAT 属于高层辅助工具，具有逻辑独立性——面向高层概念设计，SDAT 属于中层的系统分析设计辅助工具，具有物理独立性——面向逻辑设计，在微机环境中运行具有较广泛的实用性。ICAT 可以结合这两种工具使用，也可以单独使用，特别在行业 MIS 开发中是很有效的。至于 SCAT 和 PTAT，我们在微机上运行并具体支持 dBASE/FoxBASE 的编程与测试，是因为中小型计算机系统不乏此类工具。

这套辅助工具之所以能支持集成开发，关键在于元库的设计。如图所示，总体数据规划的进程和结果都保存在规划元库中；当进行应用项目开发时，自动转化为设计元库中的信息而保证高低层方法的衔接，设计元库同样作为设计进程与结果的载体；系统维护与扩充则成为对设计元库信息的维护与扩充。五个工具既可连接使用，也可单独使用，而支持集成化开发的宗旨不会改变。针对大量微机应用系统的实际情况，如果能选用 DPAT、SDAT、ICAT、SCAT 和 PTAT 这套工具，无疑地会使整个 MIS 建设工作加快步伐，提高质量，同时也会随着应用系统的发展，为其后选用更加高级的辅助工具作好开发基础和技术素质上的准备。




MIS 支持工具与方法的一体性

## MIS 建设的组织领导和两类人员的合作

企事业单位 MIS 建设的领导者和组织者，在结合自己的经验进行学习和总结时，首先应该搞清这样一些基本观点：MIS 不是单纯的计算机技术系统，而是人机结合的社会系统，它的建设必须同企业的改革，同管理现代化进步结合起来；MIS 建设必须有正确的方法论指导，高层方法论更为重要；MIS 建设的基础性工作是信息需求分析与数据管理，在管理科学化发展过程中建设数据库环境；MIS 建设需要逐步升档的辅助工具，不能不分发展阶段地追求最理想化的工具，应遵循实用性原则，有一个适应与提高过程，使用工具更需要规范化，没有标准化就没有自动化；MIS 建设是群体性的长期协调活动，不是少数人的短期行为，像任何大型系统工程一样，MIS 建设要有领导，有组织，多种人员联合作战，因此是要讲规范标准，讲纪律性约束性的，不是个别人的随意自由创造。

最好的应用系统是最终用户真正参与开发的系统，最好的方法就是能吸引与组织最终用户参与开发的方法。MIS 建设的社会系统特征、群体行为、有效地使用工具和基础建设等一系列认识，都要落实到最终用户与技术人员的结合上来，两类人员在长期的过程中逐步提高业务技术素质，从而逐步完善 MIS 的信息组织与功能结构，共同成为 MIS 建设的主力。

MIS 建设作为社会系统工程，要处理好一系列关系，调动多方面的积极性，这也是 MIS 建设中组织领导的艺术。特别是领导的重视与参与关系，要使主要领导提高对 MIS 的认识，使领导班子对全组织 MIS 建设目标与步骤形成一致的意见，责成专人负责抓落实；领导层不是停留在一般号召式的重视，而是在确定目标、审核模型、制定方案、组建队伍等方面具体参与和领导。还有信息中心与其他业务部门的关系，部门 MIS 建设的骨干与其他业务人员的关系，习惯的方法与学习新技术的关系等，也要处理好。我们总结许多单位 MIS 建设的经验，看到其成功要素是：领导重视并亲身参与；有正确的理论指导；抓紧不断提高素质的开发应用队伍建设；总结选择实用的方法和辅助工具；扎实细致的基础性工作。

**编者的话：**

“技术研讨”是本刊的主要栏目之一。开设这个栏目的目的是为读者提供一个探讨交流软件技术的场所，更新知识、紧跟世界软件技术发展潮流的渠道。软件技术内涵丰富，范围广泛。有软件本身的技术；有软件开发的技术；有软件应用的技术；有针对不同处理对象、不同处理方式和不同计算机体系结构所产生的软件技术，等等。软件技术发展迅速，当前有不少热点，如面向对象的软件技术、分布处理和网络软件技术、并行处理软件技术、多媒体信息处理软件技术以及智能化软件技术等。软件技术中也存在着诸多难点，如如何解决提高软件生产率、软件功能和软件质量的问题，如何解决使人机界面更友善的问题，等等。本刊将就以上各个方面进行深入介绍、探讨。欢迎广大软件人员积极配合，踊跃投稿。稿件内容希望尽量做到新、实。所谓新，就是希望反映的是当前国内外最新的技术进展，最新的技术研究成果；所谓实，就是内容要具体、实用，能对工作有借鉴指导作用。

# 软件技术的前沿——遗传算法概要

陈　京

近年来，作为解复杂问题的方法，遗传算法受到人们广泛关注。这种方法的特点是受生物进化的启发，以它为模型来求出解。处理复杂问题时，常因陷入局部解而不能获得最好的解；如用遗传算法，则因同时使用许多遗传基因而可得优良的解，避免了上述危险。但这种算法刚刚提出，目前尚未建立起能解决实际问题的一般理论，只能针对具体情况来编制程序。然而它具有广阔的前景，对今后软件技术的发展将起重要作用。

## 一、遗传算法的基本概念

遗传算法(GA;Genetic Algorithms)是仿效生物物种进化原理的一种程序设计方法。生物通过自然淘汰和突然变异而进化，以适应环境。GA 也是使问题的解不断变化，以求出较好的解。

GA 最初由 Holland 在《Adaptation in Natural and Artificial Systems》(1975)一书中提出，但最近才引起人们关注。这是因为计算机硬件的进步，使得大规模的实验(模拟)成为可能。

### 1. GA 的基本原理

基本上可归纳为两点：一是把物种进化的原理用于求问题的解；二是只有最适合环境的物种才能保留下来，因而经反复求解后可以得最佳的解。

GA 基本上是生成和测试(Generate-and-Test)型算法。在这里使用三种遗传操作，即选择、交叉和突然变异。求出解的候补，作为基因型则由一维的染色体表示。各代是各个个体(解)的集合。各代中的个体数称为群体规模。

GA 的处理步骤如下：(1)生成初期群体；(2)在满足结束条件前反复进行：a)适应度评价，b)选择，c)交叉，d)突然变异。

首先，构成初期群体。通常，随机地生成所决定的个体数中的染色体。这时，个体数的决定和染色体长度、编码方法等，都是 GA 研究的中心课题。现在还只能由专业人员来做。

生成初期群体后，就要对各个个体进行适应度评价。这时所用的方法有多种，但基本上都是使较好的个体获得较高的适应度评价。

决定各个个体适应度后，基于它进行选择交配。其机理应使得适应度高的个体能留下更多子孙后代。因此，形成良好个体的遗传基因将在群体中广为分布。目前已提出若干种选择交配方法。选择哪一种方法，如何设定适应度同留下子孙数期望值的关系，都将改变淘汰压力。

决定了进行选择交配的个体对之后，可进行染色体交叉。现在也有若干种方法。基本上是采用双方染色体的各一部分，来生成子孙的染色体。通常，在某基因位置上复制同一基因位置上的双亲遗传基因。然后，加上突然变异。这是按一定概率，改变染色体一部分值的操作。

这些操作完成后，便形成了新一代的个体群。然后对这新的群，再进行适应度评价、选择交配、突然变异，形成更新的一代。如此反复进行下去，直到获得满意的解。

### 2. 图式(Schemata)定理

当染色体用一维的文字串表示时，就会产生具有意义的模式。这种模式便叫做图式。图式定理就是用以表示这样的模式能在多大程度上保留到下一代的定理。它是 GA 中非常重要的定理。

定义长用 $\delta(H)$ 表示，它表示图式最初固定部分和最后固定部分之间的距离。量数用 $o(H)$ 表示，它表示图式中决定"值"部分的数。例如，1＊＊＊1 这样的图式，定义长为 4 而量数为 2，而＊0＊11 这样的图式，定义长为 3 而量数也为 3。这里，1、0 通常是二进制数，星号＊表示不定的部分(wildcard)。

$m(H,t)$ 表示在 $t$ 代群体中存在的图式 H 个数。$f(H)$ 为包含图式 H 的个体平均适应度，而 $\bar{f}$ 则为群体中所有个体的平均适应度。这时，存在于第 $t+1$ 代图式 H 的个数期望值，可用下式表示：

$$m(H,t+1)=m(H,t)\frac{f(H)}{\bar{f}}$$

然而，由于图式可能因交叉和突然变异而破坏，因此需要在上式中加上考虑这些因素的项。某一图式因交叉而破坏的概率，可用 $pc[\delta(H)/(l-1)]$ 来表示，式中 $l$ 为染色体长度、$pc$ 为发生交叉的概率。因突然变异而破坏的概率，可用 $o(H)pm$ 表示，式中 $pm$ 为变异概率。这样，上式便修改成以下形式：

$$m(H,t+1)\geq m(H,t)\cdot\frac{f(H)}{\bar{f}}[1-pc\frac{\delta(H)}{l-1}-o(H)pm]$$

光从这一定理看，应该尽可能减少突然变异和交叉，才能在群体中增加适应的图式。然而谁都知道，如果没有突然变异和交叉就不会有进化（适应）。问题在于，图式定理并不是用于分析因突然变异和交叉生成的新图式的定理。图式定理只不过用来预测在现有群体中图式个数将作何变化而已。图式定理是目前GA中为数不多的定理，它被经常使用。

## 二、遗传的操作

### 1. 缩放(Scaling)

决定适应度时，没有必要把该值直接反映到选择时的概率中去，采用某一函数，只不过使适应度的差别放大或缩小而已。采用这样的函数便叫做缩放。其基本方法有三：(1)线性缩放($f'=af+b$)；(2)$\sigma$切断($f'=f-(\bar{f}-c\times\sigma)$)；(3)指数缩放($f'=f^k$)。以上各式中，f为原来适应度，f'为进行缩放后的新适应度，$\sigma$为群体的标准偏差。

### 2. 选择(Selection)交配

对某一个体如何进行交配，在给定选择淘汰压力时是非常重要的。现在有若干种交配方法，下面介绍其中具有代表性的。

**基本模型：适应度比例战略** 它又称为旋轮线模型或蒙特卡洛模型，是以同各个体适应度成比例的概率表示留下子孙可能性的模型。某一个体i，用各种选择方法选中的概率pselect可表示为

$$pselect=\frac{f_i}{\Sigma f_i}$$

可在0～1区间内产生随机数作为各个体的选择概率。选择概率大的个体参加交配的机会多，所以它的遗传基因在群体中分布很广。

**期望值战略** 在选择概率时，如果个体数目不足够多就会出现问题，因为随机数的波动有可能出现不能正确反映适应度的选择。期望值战略便是用以解决这一问题的方法。在期望值战略中，对各个体计算留下子孙的期望值。在选择各个个体时，从这一期望值减去0.5。这样，在最坏的情况下，即使与期望值相差0.5也能留下子孙。

**顺序战略** 按照适应度由大到小安排各个体顺序，对各顺序按照确定的概率而留下子孙。这时，选择概率取决于按适应度安排的顺序。所以它的问题在于，由于适应度和顺序的不同将引起给定选择概率的差别。

**保存优秀者战略** 这是把群体中适应度最高的个体，直接保留到下一代的方法。采用这一方法的优点是，这时最佳的解不会因交叉和突然变异而被破坏。但是，由于优秀个体的遗传基因可能在群体中迅速扩大，有可能陷入局部解(local minima)的危险。

### 3. 交叉(Crossover)

交叉是两种染色体相互交换的操作，如

A 1001|111→1001000

B 0011|000→0011111

便是在第4和第5遗传基因位置之间出现交叉位置，个体A染色体从第1至第4之间同个体B第5至最后的遗传基因，组成新的个体遗传基因；剩下的组成另一个新的个体遗传基因。

在这一例子中，交叉位置只有一个，称为单一交叉或单纯交叉。除此以外还有多个交叉和一致交叉。多个交叉时存在有多个交叉位置，如交叉位置为4和8，则个体A第1至第4、个体B第5至第8、个体A第9至最后，组成新的个体，剩下的组成另一个新的个体。一致(Uniform)交叉，交叉时使用掩模，掩模为0的位不交叉，掩模为1的位发生交叉。例如：

亲1 001111

亲2 111100

掩模 010101

子1 011110

子2 101101

### 4. 突然变异(Mutation)

突然变异是以一定概率改变遗传基因的操作。设置过大的变异概率，图式将全部破坏，因此会变成完全随机化。反之，如果完全没有变异，则无法在最初遗传基因组合以外的空间进行探索，因而求出的解的质量将受到限制。通常，突然变异系按设定的固定概率使各遗传基因发生变化，但也有动态地改变变异率的。适应变异便是这种方法之一。在适应变异中，由交叉结果生成的两个个体近似度，用加权平均距离测定，距离越近则用越高的变异率。这样，便可确保群体中遗传基因类型的多样性，以便向尽可能大的空间进行探索。

## 三、GA的应用

GA的应用现在还在探索中，但已经用于不少地方。其中最著名的是用以求巡回推销员的最优解。限于篇幅，下面只介绍一些有代表性的应用。

### 1. 在最优化问题方面的应用

GA的早期应用，以最优化问题为主。如用于煤

气管道的最优控制和巡回推销员问题。还可用于通信网络设计、铁路运输计划的优化，还被用于战斗机维修计划的优化。在工业上应用的有名例子，是美国通用电气公司将其用于高旁路比的涡轮机设计。下面介绍具体例子。

高旁路比内外函喷气发动机设计，是一个非常复杂的设计，具有$10^{387}$个解的集合。对这一问题使用GA同专家系统相互结合的手法，构成名为EnGENEous的系统。首先，生成规模为50个个体的初期群体。对这一群体中的各个个体(各个设计方案)用专家系统对其进行优化。然后，评价各设计方案，决定其适应度。通过遗传操作(选择、交叉、突然变异)生成下一代。对新的一代再用专家系统对其优化。这样的循环便是EnGENEous的基本循环。其特点在于，对各个体都用专家系统进行优化。这种方法被称为交错(Interdigitized)法。它拥有GA的总体取样，可以弥补不能进行局部探索的缺点。用交错法可以比单用GA法使解更快收敛而得出结论。通用电气公司的结果表明，基本设计，用人工要2人月以上，用GA法要17天，而用交叉法只要9天。

### 2. 在规则学习方面的应用

下面介绍用于战略获得。这也就是所谓逐次决策问题(Sequential Decision Tasks)，Pitts方法便是一例。这是应用SAMUEL(利用学习经验的战略获得法)系统，来获得躲避行动规则。所谓躲避行动是指战斗机进入敌方防空识别圈内时，如何进行躲避敌方导弹的飞行问题。在这里，把规则集合看作一个染色体，交叉是以规则集合为单位来进行的，交叉的结果将得到新的规则集合。

SAMUEL系统由仿真器、行动模块、学习模块所组成。行动模块使用竞争规则产生系统(CPS)来控制战斗机飞行。在这里使用了有128个节点的BBN Butterfly Machine，在各个节点上都装有CPS。因而能并行地对128个规则集合进行评价，得到各个规则集合的评价值。根据它，进行选择、交叉、突然变异。实验结果表明，使用SAMUEL进行规则学习的结果，在200代以后可获得能够躲避99%导弹的规则；而使用简单的规则组合，只能获得40%左右的躲避率。

### 3. 同神经网络结合

应用GA进行神经网络的学习和设计研究，也是一个重要的应用课题。在早期的研究中，采用了在染色体上表现网络的权重和节点的偏置这样的示教方法。但是这种方法将出现扩张性问题，使网络加大而收敛时间激增。

为此，改用了由L-system等文法规则生成网络构造的方法。这是一种高效率的方法。在这种方法中，作为遗传基因型获得不是网络构造的构造生成的文法规则。网络的构造可递归地用以产生文法。这种方法也叫文法编码法(Grammatical Encoding)。在以前的GA中，遗传基因型的解几乎都是用直接表现法(Direct Encoding)，遗传基因型和表现型成简单映射的对应关系；而在文法编码法中，在遗传基因型和表现型之间存在翻译器。这里，遗传基因型是指由遗传基因组合而成的模式，而表现型则指基于遗传基因型而形成的个体。

## 四、存在问题和今后研究方向

作为一种探索方法，GA的弱点在于它不是局部探索方法。为了克服这一缺点，提出了交错法和GA-BP法，但这并不是决定性的东西。

此外，解的表现方法即如何对染色体进行编码，尚无决定性的指导原则，现在还只能由专业人员凭自己的经验来决定。这方面需要进行理论分析。

尽管可以用评价函数，但除了一部分应用可以用评价函数明确定义外(也就是在这一领域可以确立解的评价方法)，评价函数的设定还都是专业人员的事。评价函数的好坏将直接对淘汰压力起影响，因而对解的质量和收敛速度有巨大影响。在最坏的情况下，由于使用不恰当的评价函数，也可能无法求出解。不过在GA时，尽管对评价函数不大清楚，如果能够进行模拟，那么也有可能进行评价。这方面的应用领域相当多。

至于今后研究的课题，首先就是要积累应用事例，获得一些真正重要的、具有象征意义的成功例子。的确，可以供GA研究的有趣材料很多，但也存在无法想像出其明确应用的危险。

其次，要发展新的GA词形变化(paradigm)法。现有的GA基本上是受生物进化论或新达尔文主义的启发而提出来的。但是，关于进化的学说，现在还有主张可能获得形质遗传的拉马克学说，以及木村资生的中立说，甚至还有Virus主张的种之间可以传递遗传信息的学说。这些学说至今还在争论中。这些学说都能给工程以有益的启发。

此外，有关共生进化(Co-Evolution)方面的研究还刚开始，可以期待获得进展。同时，分子生物学、遗传学上还有许多概念还没有用到GA上，这方面也是有极大潜力的。最近，最新分子生物学有许多新发现，因此有可能有助于提出一些崭新的概念。

# 一种管理异质混合数据库的面向对象方法

华中理工大学 冯 铃 冯玉才

**摘要** 本文提出了一种管理异质混合数据库的面向对象方法，讨论了该混合数据库系统的数据模型、组成结构以及各种异质数据库中信息的一致表示与集成方法。

## 一、引 言

管理异质混合数据库的系统是指对分散的按不同数据模型组织的信息从模型上进行集成构造一个具有多种风格的混合模型，而用户觉得是在一种共同的数据模型上访问混合数据库。由于面向对象数据模型的丰富性，它适合作为表示诸多数据模型的公共数据模型。我们这里用一个扩展的面向对象数据模型作为表示这些不同数据库模式的全局模式，例如。假定数据库 Product 由关系 DBS 管理，数据库 Employee 由层次 DBS 管理，数据库 Company 由网状 DBS 管理，我们可统一地按面向对象数据模型视图来集成管理、共享与存取这三个库中的信息。

将异质数据库进行集成，研制混合型 DBS 是对现有信息财富的继承与保留，通过一种有效的表示与转换手段，将各种已获得的按不同模式组织、存放与操作的数据集中管理，使得传统模式的数据库内容不随时间与应用的发展而被新模式的数据库系统(DBS)所淘汰是我们研制异质混合数据库最终要达到的目标。它对推进数据库(DB)领域的研究具有重大意义。异质混合 DBS 打破了传统数据库间信息彼此独立、用户不能共享的界限。它所具有的数据模型透明性使得用户不用关心各个具体库的模型与操作特性，就可以一种一致的方式，同时访问多个异质数据库中的内容。利用面向对象公共数据模型中的导出类概念，我们甚至可以定义同一库中、不同库间的信息关联，增加传统数据库不具备的语义表达能力，加强对现实世界的模拟；由于面向对象程序设计与数据模型有助于可扩充性，用此方法设计实现一个异质混合数据库系统，还可随时容纳附加的、新的类型的数据库，从而为现在与将来各种信息库的共享提供可能；此外，作为从传统数据库过渡到面向对象数据库的一种途径，管理异质混合数据库的系统也是很重要的。

为满足多个异质数据库的集成要求，在设计混合 DBS 的组成结构中，我们引进了两个抽象级映射：模式转换映射与模式集成映射，并采用了扩展的面向对象数据模型作为系统的全局模式。

## 二、异质混合数据库系统的数据模型

对多个异质数据库系统的集中管理，需要有一种能够统一表示各个库中信息的公共数据模型，使得用户对各库的访问均在此唯一的全局模式上进行，而不用区分其所需信息在具体库中的组织和操作方式。面向对象数据模型在日益广泛的应用领域中显示出了传统数据模型所不具备的强大生命力，其根据不同领域、不同对象分门别类地组织与操作数据的思想适合于各异质数据库的集成表达。这里我们采用面向对象数据模型作为异质混合 DBS 的全局模式，并对其进行必要的扩展。

在面向对象数据模型中，现实世界里的任何实体均作为对象，由系统赋予唯一的对象标识符，每个对象都把一个状态和一个行为封装在一起，其中对象的状态是该对象属性值的集合，而对象的行为是在对象状态上操作的方法(程序代码)的集合。共享同一属性值集合和方法集合的所有对象组合在一起，构成一个类，例如下面将讨论到的网状模型中的一个系，在面向对象的公共数据模型中就可用一个对象类来描述。根据一般化和特殊化的关系，我们可将某些性质上相似的类联系起来，使模式中的所有类组成一个有根的有向无环图或一个分层结构。一个类从其类分层结构中的直接或间接祖先那里继承所有的属性和方法，封装在一个对象中的状态和行为只能通过所显示的消息传送从外部存取和调用。

一般的面向对象数据模型支持类间的两种语义关系：特化其母类的子类所反映的概括(is-a)关系和对象通过属性的复合所反映的聚集(has-part-of)关系。为了适应异质混合数据库的应用环境，在此之上，我们又扩展了一种新的语义关系——对应(corresponding-to)关系，它反映了类与类之间的对应关

联，例如层次模型数据库中有一个一对多的从属关系（如图1所示）。

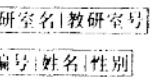



图1　层次模型数据库中的一个从属关系

一个教研室可有多名教授，在系统的公共数据模型中，这两个实体分别被模拟成两个类，通过教研室类中的对应关系，它的每一个对象实例与多个教授类中的实例对应起来。同样，通过此语义关联，我们还可表达网状模型数据库中一个系的首记录型与从记录型间的一对多关系。

除此之外，为了使用户通过集成环境，达到对各异质数据库的一致访问，我们还扩展了一般面向对象数据模型中的类的概念。通常，一个类的实例集与该类存放在一起，在异质混合 DBS 中我们称这样的类为外延类。另外，又增加了另一种内涵类，即类的实例集与该类的描述分开存放。内涵类又分外部类和导出类两种。外部类是对各异质数据库中基本信息单元的面向对象表示，例如：关系数据库中的一个关系、网状数据库中的一个系在混合 DBS 中分别表示成一个外部类，其实例，也即元组或系值存放在各自的基本库中，并未改变，它们只在需要时部分调入系统；而导出类是由基本类（外部类和外延类）导出，用基本类来描述，系统提供的语义关联，如归纳、聚集、特化等关系可视为导出类的构造子。对于既不能从外部类检索出、又不能被导出的信息，这时可通过用户定义的外延类及其存放在一起的实例得到。

## 二、异质混合数据库系统的组成结构

为了达到一致访问各不相同的数据库，而不用区分其数据模型的目的，我们在设计混合 DBS 时，引进了两个抽象级映射（如图2所示）。

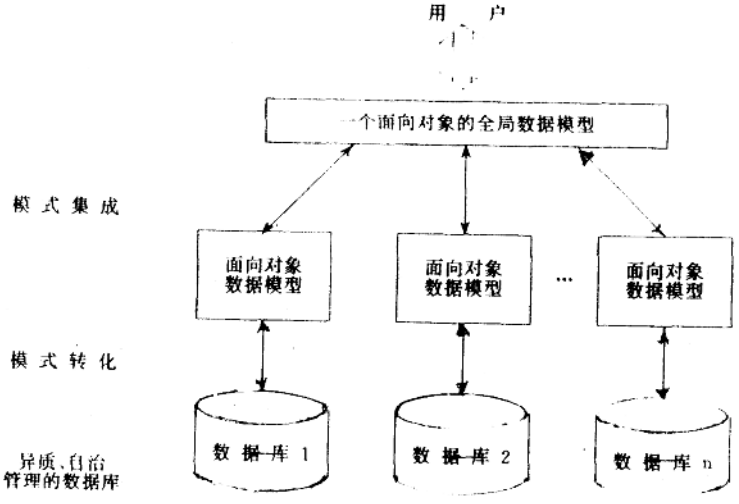



图2　异质混合DBS的组成结构

·模式转化映射负责将各个异质数据库中的信息内容转换成句法上一致的面向对象表示，经过该级转换，所有的数据库都用统一的面向对象模式来描述。例如，层次、网状或关系数据库中的任何实体都用面向对象模式中的类来描述，这些实体间原有的从属、对应关系等，构成相应类间的语义联系。这里各个库中的信息还未被集成。

·模式集成映射负责将由上级转化得到的具有相同模式的各个框架信息组合成一个统一的整体，用一个面向对象的全局模式来统一表达，并可根据用户的应用需求及语义集成要求建立各模式间的信息联系，增加原 DBS 所不能描述、处理的语义能力，如信息类间的概括、特化等，这反映到对导出类的定义。另外用户还可定义一些新的有用的类，它们与上述的外部类、导出类一起被集成管理起来。

下面两节，我们用具体例子说明如何实行这两级映射。

## 四、模式转化映射

为了将按不同数据模型组织的各个库中的信息转化成按一个统一的面向对象模式表达的形式，我们首先需要对各种异质数据库中表达信息的基本单位作相应的转化，而且这种转化必须是可逆的，因为用户在全局模式上的任何访问要求，最终是通过对最底层的基本数据库的操作得以实现的。该模式映射故而也应能向系统提供一种有效的手段，将面向对象模式描述的数据形式及操作转换成基本库中相应的数据及操作格式。我们用一个具体例子说明如何将各种数据模型中表达信息的基本单位转化成统一的面向对象表示。

假定有四个异质 DBS：DB Product 是关系模型，DB Company 是网状模型，DB Employee 是层次模型，DB Contract 是面向对象模型。关系模型中信息的基本单位是一个个关系，在面向对象模式中，我们将这一个个关系表示成类，例如有一个关系：

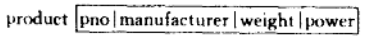



图 3 一个 product 关系框架

将它表示成一个类：

Define class product

type tuple (pno: integer, manufacturer: string, weight: real, power: integer)

method category:

method body:

该类的方法及方法体由用户按常规的面向对象模式中说明方法的格式加以定义。

网状模型中信息的基本单位是系，一个系由首记录型和属记录型组成，每一系中弧代表的是一对多关系，我们将每一个系表示成一个类，例如有一个系：

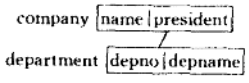



图 4 一个 company-department 系

因 company 与 department 均为记录型实体，故可用两个类分别描述，用类概念描述这个系为：

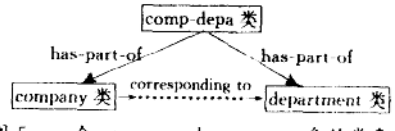



图 5 一个 company-department 系的类表示

……→代表类间的对应(corresponding to)关系，即一个 company 类的实例与一个或多个 department 类的实例相对应。

层次模型的图论表示为有向有序树，基本信息单位是树中一个个结点，我们用一个类来描述一个结点，在层次型中所包含的从属关系是一对多的关系，因此在一对母结点和子结点之间，对应于位于母结点的一个型，是位于子结点的另一个型的一个序列，如图 6 所示：

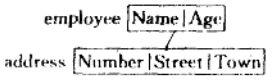



图 6 一个 employee-address 层次型片段

一名 employee 可以有多个住址，我们将 employee 与 address 分别用两个类来表示，它们具有对应(corresponding to)关系，以反映层次型实体间的从属关系，如图 7 所示：

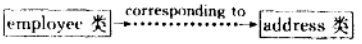



图 7 层次型片段的类表示

上述所有用于表示基本信息单位的类均是外部类，它们仅描述了类的信息框架和操作方法(由用户定义)，其实例仍是原按各种模式存放在多个数据库中的信息。用户通过发送消息，对某些类及实例对象的操作要求(即方法体中的操作代码)经由系统提供的相应转换机制，映射到原数据模型中对应的数据及操作语言，由基本数据库系统执行、实现。对于本是面向对象数据模型的单个数据库，则不需任何模式转化，且它包含的类都是外延类。但从用户的观点看，他对所有信息类的检索与操纵均具有一致的方式，即所有的类具有统一的对外接口。

经过模式转化映射，所有异质数据库表示成了多个同一的面向对象模式，下一节讨论如何将多个模式集成为一个全局模式。

## 五、模式集成映射

用一个全局的面向对象数据模型将多个面向对象模式集成管理起来，让用户觉得所有库中的信息均是按同一数据模型，存放在一个共同的面向对象数据库中，且对各库信息的访问也以此模型提供的消息方式完成，这就是本级映射所要达到的目的。

为解决异质 DBS 的集成问题，在面向对象全局模式中，我们区分两种类：系统类与应用类。系统类

由混合 DBS 本身建立、管理和维护，它不随应用的改变而改变。例如，有一系统类 Relclass，它描述了关系数据库中的关系与其公共模型中的类之间的所有映射转化方法，包括数据的映射方法（即将用户的操作对象类信息映射到关系数据库中的相应关系及元组）和操作的转化方法（即将一个用户用全局模式表达的操作要求转换成为关系数据库中的操作代数）。应用类是由用户定义或通过模式转化得到的类，应用类可作为系统类的子类，依据面向对象数据模型的继承机制，应用类具有系统类的全部功能。例如，若将上节中外部类，也即应用类 product 定义为 Relclass 的子类，则它可继承 Relclass 中特定的系统方法，从而能自行实行类与关系间的数据映射、操作转化等功能，达到无需用户干预，就能完成对混合数据库的访问，从这一类我们也可看出作为公共数据模型、管理异质混合数据库，面向对象视图是最佳方法。针对上节例子，除系统类 Relclass 外，我们还需 Hieclass 与 Netclass 两个系统类，各自完成层次模型中的实体，网状模型中的系与全局模式中对应的类间的映射与转化工作。

在模式集成一级，除了完成各外部类的集成外，用户还可定义新的有用的外延类，输入实例集。此外，为了加强混合 DBS 的语义表达能力，系统还提供了一个类构造子集合，将一个类构造子运用于类 $Y_1, Y_2, \cdots, Y_n$（$Y_i$ 既可是外部类，又可是外延类，$i \in [1,n]$）上，可得到一个新导出类，这里系统提供的任何一种语义关系均对应于一个二元类构造子。例如，根据一个特化（specialization）关系，可得到一个导出类 pro-premium：

```
DEFINE class pro-premium
superclass/specialization of : product
type tuple ( pno : integer, manufacturer : string,
             weight : real, premium : integer)
{if product.power>10 then premium:=1000 * power
 else premium:=0 }
```

类 pro-premium 是上节中外部类 product 的特化子类，它继承了 product 类的方法与状态，且又增加了自己的属性 premium，{ } 中给出了其属性值的导出方法。同样，根据概括（generalization）关系，我们可由类 $Y_1, Y_2, \cdots, Y_n$ 导出一新的归纳类。

与外部类相似，导出类也是内涵类，没有实例存储，其实例需通过导出方法与要求，从基本库或外延类中推导得到。外部类、导出类统称为应用类。至此，我们可用一个全局视图来描述上节各个面向对象模式集成映射后的结果（图 8 所示）。

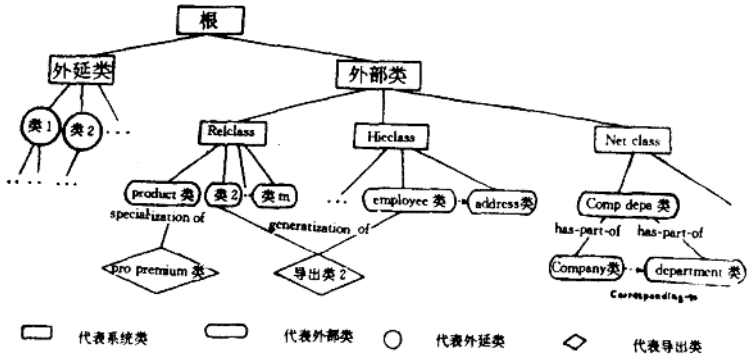



图 8　异质混合数据库的一个简单的全局视图

## 六、结　束　语

本文提出了一种管理异质混合数据库的面向对象方法，为了满足异质数据库系统的集成要求，加强其语义表达能力，我们对作为全局模式的面向对象数据模型进行了扩展，讨论了该混合数据库系统的组成结构，各异质数据库中信息的一致表示与集成问题。这里我们对异质、自治管理的数据库系统集成研究作了初步探讨，还有许多问题，如不同模式间信息的高效转化问题，用户访问要求的有效转述问题等，在今后的工作中还有待研究解决。

# 多层次检索管理信息系统的设计与实现

西南民族学院 杨宪泽 李宝群 秦文海

**摘要** 本文介绍在不影响已有管理信息系统的系统结构下增加的多层次检索方法。这些方法，以子模块形式设计，引入按常规方法设计的管理信息系统后，使其功能扩充、环境改善，具有实用意义。

## 一、引 言

大多数管理信息系统中，重要的一环是检索。就目前状况，常规管理信息系统设计有较成型的模式，各行业多采用。但其检索手段少、影响着管理信息系统的自动化程度和应发挥的效益；管理信息专家系统，人机接口很方便，检索的手段也较多，但技术还不十分成熟，系统开销大，研制周期长，只能在少数技术力量雄厚，经费充足的企事业单位试行，短期内不可能遍及各行业。基于上述情况，我们设计了多层次检索的三个方法。作为子模块，可引入常规管理信息系统，增加检索手段，提高自动化程度。

## 二、同 音 检 索

本节介绍在已有的管理信息系统基础上设计的同音检索子模块，解决中文关键词中误输入同音字导致检索失败的问题。

### 1. 方法思路

汉字机内码是机器内部表示汉字的代码，是中文管理信息系统体系结构设计的基础，也是同音检索实现的基础。汉字基本集标准 GB2312－80 包含一级汉字 3755 个。汉字按拼音字母顺序排列，同音字基本上在同一区中，少数跨越两区。每个汉字的机内码为两个字节，其高字节部分确定所在区号。我们构造的子模块其关键词内容以每一汉字所在区号进行索引，信息记录地址不变，见图 1。

使用时，若按常规查询误输同音字，导致检索失败，可进入这个子模块。子模块中进行关键词每一汉字的区号比较，两个关键词若每一汉字区号完全相同，则存储器中关键词对应的文献记录就是要检索的内容。这样，误输同音字导致检索失败的问题就得以解决或在很大程度上缓解。

### 2. 算法构造基点

同音检索的过程是，将要检索的关键词每一字符的高字节 ASCII 码与存储区内已建立的关键词每一字符高字节 ASCII 码一一比较，两者一致时存储区内关键词对应的文献记录为查询记录。同音字虽然机内码不相同，但高字节部分规定的区号是相同的。因此，子模块中首先建立关键词每一汉字高字节区号构成的索引。例如，关键词“电路节点”的区号索引为：21－34－29－21。

**关键词索引建立步骤：**

(1)初值 j＝1

(2)求关键词(字符串)长度：M←LEN(M$$_j$$)，其中 M$$_j$$ 为字符串。

(3)切分关键词成单一字符：

```
do i from 1 to M
K$j←MID$(M$j,i,1)
```

(4)将每个字符的区号(高字节部分)连接

```
do i from 1 to M step2
A$j←A$j+K$i
```

(5)j←j＋1，直至 j＝N，实施(2)～(4)。其中 M$$_1$$－M$$_N$$ 为管理信息系统内 N 个关键词。

(6)排序链接区号，并与原文献记录建立索引关系。

对于(6)，按 GB2312－80 规定，一级汉字出现在 16－55 区。如果按关键词第一区号排序，可采用分级技术。这里，关键词所有第一区号作为一级索引，通过简单计算即进入口。以后的关键词比较采用效率较高的二分检索法。

有少数汉字可能跨区，如宋键义和宋健义模块允许两种定义：43－28－50；43－29－50，它们均与原文献记录索引，见图 2。

此外，有可能出现区号完全相同的关键词，采用链接方式处理，检索结果将它们的文献记录均输出，由用户判断需要哪一个。

**算法主要点：**

(1)若常规检索失败，退出，以菜单方式询问用户是否要同音检索。

(2)进入同音检索子模块，待检索关键词求长度、切分、确定区号。

(3)关键词第一字符区号简单计算，进入相应区

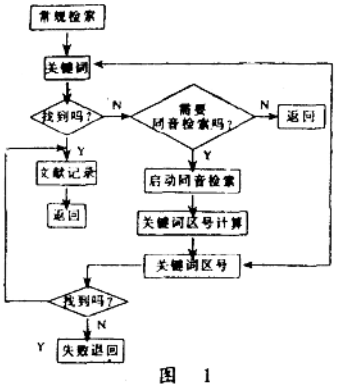



图　1

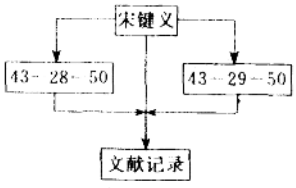



图　2

域。

(4)进行二分检索，第二字符区号用以确定二分范围。

(5)成功输出结果，失败退出。

### 3. 算法实现描述

A1：输入待检索关键词 N$。

A2：进入常规检索。找不到，询问用户是否要同音检索？要，进入同音检索子模块→A3；否，退出。

A3：求 N$ 长度，d←LEN(N$)。

A4：切分 N$ 为单一字符 K$₁，K$₂，…，K$ₐ

do i from 1 to d

K$ᵢ←MID$(N$，i，1)

A5：区号连接 B$=K$₁+K$₃+K$₅+…+K$ⱼ(j≤d)

A6：分级入口，从 K←ASC(K$₁)转相应程序段。

A7：二分检索，以 K←ASC(K$₂)确定二分范围。

A8：二分检索子程序运行(见有关书刊)。

A9：若找到相同区号，其索引的文献记录输出，检索成功。

A10：若找不到相同区号，检索失败，退出。

## 二、模糊分类与检索

如果人们给出的检索要求不明确，则认为要求是模糊的。例如，"请检索系统内与数据结构有关的文献"，这一要求就是模糊的，因为数据结构的定义目前尚未能统一，那么，究竟哪些关键词属于数据结构不能完全确定。

不明确的检索要求就应采用模糊检索方法。本节首先讨论如何在管理信息系统中使关键词实现模糊分类，即建立子模块；然后简述模糊检索的实施。

### 1. 分类算法构思

设管理信息系统中 n 篇文献由 $X_1, X_2, \cdots, X_n$ 组成。n 篇文献中含有 m 个不同的关键词 $K_1, K_2, \cdots, K_m$。这样，一篇文献 $X_r$ 可用 m 维向量来描述：

$X_r=(\varphi_{r1}\varphi_{r2}\cdots\varphi_{rm})$。其中

$$\varphi_{rj}=\begin{cases}1, X_r \text{ 中有关键词 } K_j\\ 0, X_r \text{ 中无关键词 } K_j\end{cases}$$

对于每一关键词 $K_1, K_2, \cdots, K_m$，事先标注它是属于某一类或多类(某一门课程或多门课程)。例如，作者发表一篇文章，给出三个关键词，只允许作者认定这篇文章属于某一门课程；那么这三个关键词均属这门课程；然而，在另一篇被认为属于另一门课程的文章中，有一个关键词与前述文章相同，那么这一关键词将属于两类。

统计 $K_j$ 在第 i 类中出现的概率 $P_{ij}$ 为

$$P_{ij}=\frac{L_{ij}}{L_j}(j=1,2,\cdots,m,\quad i=1,2,\cdots,d)$$

式中，$L_j$ 是 $K_j$ 在文献中出现的总次数；$L_{ij}$ 是 $K_j$ 认定属于 i 类的总次数，d 为分类数，隶属函数可由下式算出：

$$\mu_i(X_r)=\frac{\sum_{j=1}^{m}\varphi_{rj}P_{ij}}{\sum_{j=1}^{m}\varphi_{rj}}\left(\begin{matrix}i=1,2,\cdots,d,\\ r=1,2,\cdots,n\end{matrix}\right),$$

### 2. 分类算法描述

A1：确定分类数 d。每篇文献假定属于一类(作者填表认定)。

A2：r 从 1 至 N，输入每篇文献 $X_r$ 的关键词 $K_1, K_2, \cdots, K_m$(即建立索引)，有

(1)若 $K_j$ 所属文献为 i 类(i=1,2,…,d)，$L_{ij}=L_{ij}+1$($L_{ij}$初值赋 0)。

(2)若有 $K_j$ 相同(j=1,2,…,m)，$L_j←L_j+1$($L_j$初值赋 1)。

A3：r 从 1 至 N，$X_r$ 含有的关键词 $K_j$(j=1,2,…)对应 $\varphi_{rj}=1$(其余 $\varphi_{rj}$初值已赋 0)。

A4：[初值 j=1]i 从 1 至 d，计算

$$P_{ij}=\frac{L_{ij}}{L_j}.$$

A5：j←j+1，直至 j=m，重复 A4。

原书缺页

原书缺页

原书缺页