

浙江大学

博士学位论文

基于数字化的生物分类鉴定及知识集成研究

姓名：张小斌

申请学位级别：博士

专业：农业昆虫与害虫防治

指导教师：程家安;陈学新

20070430

中文摘要

21 世纪是信息时代,随着计算机与网络的广泛普及与应用,数字化信息越来越被人们推崇与重视,知识数字化、传播网络化已成为传统学科信息化发展的必然要求。生物分类是物种多样性研究与保护中重要的基础工作,尽管现有的纸质分类信息十分丰富,但由于分类信息数字化开发工具匮乏,分类专家直接数字化比例低,生物分类信息资源的数字化发展相对缓慢。为改善此局面,推进生物分类信息资源的数字化建设,论文依托现代信息技术,以检索表数字编码、智能编制和二次重构三大创新技术为核心研制了生物分类鉴定知识系统与物种多样性数据库开发工具,为检索表等生物分类数字化信息的制作、整理、发布、使用与推广提供一套完整的电子化解决方案,并在此基础上建立了中国昆虫鉴定分类系统 InsectX 与植物检疫性昆虫信息平台 W-QPM 等网络系统。

1. 检索表数字编码技术

基于规则的检索表数字化方式直接模拟专家思路,存在扩展性弱等诸多缺点。论文提出采用基于二维的特征分值数字矩阵保留检索表中对象与特征的匹配关系,并用 XML 结构化数据模板记录所有检索表信息。该方案数据元分离到位,数字矩阵十分适合计算机分析处理,深入挖掘扩展功能,它为生物分类信息数字开发与应用奠定了重要基础。

2. 检索表智能编制技术

为实现检索表直接数字化开发,提高设计效率与科学性,论文从检索表设计基本原则出发,设计了全局优化的新检索表编制算法,可基于专家建立的特征与对象关联矩阵产生备选检索表,通过比较检索表特征优化度、对象优化度或综合优化度,并参考检索表长度和特征总优先度,输出最佳检索表。以此为核心开发的二项检索表专业设计软件 KeyMaker 优化性能达到并高于同类 DELTA Key 软件,可代替分类专家智能设计出最佳的检索策略,减少手工编写检索表所花费的时间与精力。

3. 检索表二次重构技术

该技术是一种新颖的“表生表”重构技术,它以基础分类检索表为知识库,根据用户定制的重构对象,通过分类单元与鉴定特征间匹配关系的反向推理,自动搜寻并重组检索路线,

生成只含指定分类单元和必要特征的二次检索表,可大大简化依据传统分类检索表鉴定的繁琐过程,提高鉴定效率。目前 InsectX 已支持定制国内昆虫科及以上分类单元、蜜蜂科所有已知种在线进行二次检索表重构。

4. 生物分类鉴定知识系统开发工具

该工具是提供给分类专家、多样性工作者等从事分类信息数字化,提供网络信息服务的综合平台。用户只需收集、整理与入库分类资料,其他如网页制作、数据库集成和网站发布等技术性工作都由该工具自动完成。生成的系统内建系统分类树、辅助鉴定、二次检索表重构、物种鉴定训练、信息检索查询、分类专题定制等功能,完全能满足各种分类研究与学习的需要。

5. 物种多样性数据库开发工具

该工具收录分类地位、鉴别特征、地理分布、标本信息、物种图片等信息,并编译成易理解、可移植的基本信息库和物种图片库,用于第三方分类软件或系统的开发。以此开发构建的蜜蜂科多样性数据库已收录分类阶元 731 个,蜜蜂 569 种,检索表 128 个,是国内蜜蜂分类信息最全的多样性数据库。

6. 基于 Web 的中国昆虫鉴定分类系统 InsectX

InsectX 以昆虫系统分类树为框架,涉及分类阶元 1 873 个,检索表 984 个,物种图片 2 343 张,特征图片 284 张,包括名称、特征、分布、鉴定流程、术语释解、检索表等信息,是国内分类系统最完善、内容最全面的生物资源数据库之一。系统以提供网络检索表在线鉴定、二次检索表重构鉴定、物种直观鉴定训练等功能为特色,是昆虫分类鉴定与科普教学的重要电子工具。

7. 基于 Wap 的植物检疫性昆虫信息平台 W-QPM

W-QPM 是基于无线互联网的移动式专家系统,收录了 203 种植物检疫性昆虫信息,涵盖分类地位、检疫特征、来源产地、寄主植物、物种图片等,并提供这些物种快速的多途径检索功能。检疫人员在海港口岸等地可通过各种手机、PDA 等便携式移动终端实时访问,查询检疫信息,在线咨询专家,辅助决策鉴定。

关键词: 信息数字化 数字编码 分类规则 二次检索表 分类鉴定系统 生物多样性

Abstract

The 21-century is called information era, computer and network have become very common and digital information has been increasingly advocated in China. It's necessary for traditional disciplines to carry out digital information construction if they want to keep the step with the informatization trend worldwide. Biological taxonomy is in the case since biodiversity research and popularization require paper-based taxonomic bioinformation available online and shared globally. Due to lack of computer-aided tools and few experts' participation, the digitization of taxonomic bioinformation makes slow progress. Therefore, three innovative techniques for digitizing, generating and reproducing keys and two computer tools for building taxonomic identification systems and biodiversity databases, which could successfully provide a complete electronic solution for producing, integrating, publishing and sharing taxonomic bioinformation online, were developed by using information technology. InsectX and W-QPM were also developed as examples for their application based on this solution.

The major results of this study are summarized as follows:

1. The coding technique for key digitization

Due to direct simulation of expertise, the rule-based method of key digitization has many drawbacks, such as little expansibility, a 2D-matrix of character scores is used to record the matching relation between objects and characters. The 2D-matrix and other information of identification keys are saved in a high-structured XML module. This matrix-based method leads to rapid analysis and flexible application for computers and lays a foundation for further development and application of taxonomic bioinformation.

2. The division criteria for heuristic key generation

A new division criteria was proposed to generate dichotomous keys digitally and efficiently. It works on the matrix-based expertise of objects and characters, and chooses heuristically between possible ways of branching a key by using a global optimization which is primarily determined by character priority, taxon priority or compositive priority, and secondarily by key length and sum of character rank. As the new criteria runs as the core algorithm, KeyMaker, a new professional program for key generation which attains or surpasses DELTA Key in the capability of key-optimizing, could save considerable time and mental effort for entomologists in key construction.

3. The generation of secondary identification keys

Secondary identification keys(SIKs) only include user-specified objects and related characters, which simplifies the identification procedure with traditional keys and improves its efficiency. The generation of SIKs, known as the key-to-key technique, operates on backward reasoning based on available general identification keys and heuristic recombination of identifying pathways. Currently, InsectX has managed to generate SIKs of any insect taxa at family level or above and any known species of Braconidae in China at species level.

4. A tool for developing knowledge system to assist biological identification and taxonomic study

This tool is designed for taxonomists and biodiversity officers to make traditional taxonomic information digitized and available online. After users finish collecting and sorting out taxonomic materials, such hard work as making web pages, building databases and publishing websites will be done automatically. The system obtained is embedded with functions of learning evolutionary tree, assisting identifications, generating SIKs, training of species identification, searching and querying, then can meet the needs of taxonomic researches and studies entirely.

5. A tool for developing biodiversity databases

This tool collects biodiversity information (taxonomic hierarchy, discriminating characters, geographic distribution, sample description and species images), and compiles them into two comprehensive transplantable databases which are used as knowledge bases for developing other taxonomic software or systems. This tool was used to build the most complete diversity database of Braconidae in China, which consists of 731 taxa (including 569 species) and 128 keys.

6. InsectX, a web system for insect taxonomy in China

As one of the largest biodiversity databases in China, InsectX includes 984 keys and 1 873 taxa, which are described with names, characters, distribution, identifying procedure and terms explanation, and illustrated with images of 2 343 species and 284 characters. With all these information organized in the tree-form of systems, InsectX provides users with online identification using web-keys or SIKs and training of intuitionistic identification, which enables it to be an important electric tool for insect taxonomy and science education.

7. W-QPM, a mobile Wap-platform for plant quarantine pests

W-QPM is a mobile expert system based on Wap. It provides about 203 plant quarantine pests including information on taxonomic hierarchy, morphology, source or producing area, host plants and species images, and enables users to enhance capability of identification with multi-entry keys. Quarantine officers can log on it with mobile terminals like Wap-phones or PDAs anywhere at anytime, to view quarantine information, identify quarantine pests or inquire experts online.

Key words: information digitization, digital coding, division criteria, secondary identification keys, taxonomic identification system, biodiversity

第一部分

文献综述及本研究的目的是和意义

引言

随着计算机日益普及与发展,中国互联网事业发展日新月异。网络是四通八达的信息传播通道,网上发布的信息几秒间便可传递给成千上万的用户,加上它信息海量,更新快速,访问方便等特点,正吸引了越来越多人通过互联网了解资讯,学习知识。根据中国互联网络发展状况统计报告(CNNIC, 2006)分析,中国目前网民总人数为1亿2300万人,其中18-24岁网民占38.9%,学生网民占36.2%。根据年龄段与网民身份的比例来看,中国网民三成多是高中生和大学生。网民获取信息的主要途径调查中,通过网络的比例高达82.6%,通过书籍只占18.7%;网民普遍认为当前互联网对工作与学习有较大的帮助。可见,互联网已成为人们尤其青年网民获得信息、交流学习的重要渠道。

正是看到了互联网影响范围广大、发展前景广阔的优点,许多传统学科都迫切投入信息化的平台建设,提供网络化的信息服务,以推动传统学科的改造,加速提高学科的生存发展能力。生物多样性是现今生物领域中的热门学科,它的兴起源于人们对环境的日益关注,对自然资源保护和可持续使用的需求。分类鉴定是生物多样性研究中基础而重要的工作,要能鉴别物种、了解物种诸多生物学特性,才能再加以保护。随着环境污染与破坏的加剧,生物多样性保护的意识流正在全球快速蔓延,不断引发了人们对生物资源的关注,促进生物分类的广泛研究。为了便于人们了解物种多样性,学习鉴定物种,生物分类信息正渴望通过网络平台实现广泛的快速传播,生物多样性的研究与推广正逐步进入信息数字化和网络共享化的轨道。

然而生物种类繁多,信息庞杂,长期研究积累下来的分类知识大部分都记载在书中,只有其中一小部分通过计算机数字化进入了互联网,习惯了网上浏览学习的人们,还不得不要出入各大图书馆,翻阅书籍查寻分类信息。作为物种鉴定最重要的指导工具,检索表数字化程度更低,大量都沉睡在书本中,网上十分匮乏。尽管数字化期刊、电子图书馆已解决了部分生物分类信息的数字化问题,但这些资源收集有限,分布无章,尤其没有转化为方便有效的数字工具,还是难以获取使用。随着人们对生物多样性的关注与重视提高,这些矛盾正在不断显著加剧。缺乏相应通用软件与网络平台的技术支持,是导致检索表等分类信息资源不能尽快实现数字化并网上传播利用的主要原因。

因此,有必要将现代化信息技术有效地融合于生物多样性研究与知识普及过程,通过专门开发生物分类鉴定的数字化支持与管理工具,为分类专家提供发布与交流分类信息的网络平台,给一般用户开辟了解信息与物种鉴定的便捷通道,全面推动生物分类信息数字化开发和网络化服务的进程。

第一章 信息资源数字化建设

美国著名信息学家 Lancaster (1982) 曾预言未来 20 年后纸质图书馆将消亡, 将由纯电子介质信息替代, 人类社会将进入“无纸社会”。虽然目前这种发展趋势并没有像 Lancaster 预言的那样快, 但我们都已在身边真切感受到了数字图书馆、数字期刊、数字农业、数字媒体、数字家庭、数字电视等信息时代的产物。人类历史上还没有哪一个时代像今天一样, 现代信息技术特别是网络技术引发的全球信息化浪潮汹涌澎湃, 方兴未艾。

1 信息数字化的认识

面向信息时代, 数字化的理念也越来越为人们所关注。数字化是指把模拟信息转化为数字信息的过程, 现实世界中文字、图象、语音、动画、录像等各种可视信息都可以用 0 和 1 来数字化表示(Coates, 1992)。数字化是计算机的基础, 计算机能表示这些 0 和 1 组成的信息, 显示文字和图片, 播放音乐和影片, 还可以利用各种系统与工具软件对这些信息进行滤波、编码、加密、压缩等数字加工与处理。信息数字化的最基本理解是把其他载体类型(原始型、印刷型和模拟电子型)的信息变成计算机能识别和处理的信息, 数字化必然离不开各种计算机技术(杨晓农, 2004)。除了计算机, 数字化还与互联网结合, 以互联网的无限联接传输为平台基础, 达到数字化信息的大规模快速传播与共享利用。

2 数字化建设的意义

信息时代的构成不仅仅是众多的计算机和复杂的网络结构, 信息资源数字化的总量、信息资源揭示与利用的深度和管理水平, 已成为信息时代新的特征和衡量信息技术水平应用新的标准(林丹红, 2001)。国外发达国家信息资源数字化随网络建设同步发展, 文化与信息正在以前所未有的速度和规模向全球传播扩散(刘阳和丁银燕, 2002)。我国计算机普及始于 20 世纪 80 年代(沈被娜等, 2000), 网络建设启于 90 年代中叶(崔海东, 2005), 网上信息资源丰富程度不能与国外同日而语, 信息资源数字化已成为国家社会信息化发展的必然要求。而信息数字化的发展又促使科学实践、科学交流、科学思维、科学理论和知识结构等发生了深刻变革, 在信息层面上扩大了科学认识的活动空间, 打通了科学自主发展的瓶颈(于衍平, 1997)。因此, 加速信息资源的数字化建设, 不仅是国家发展战略竞争和社会信息化需求的必然, 而且是各领域学科实现科学知识数字化, 拓展学科自主发展生存空间的迫切之需。

3 数字化建设的方法

随着现代信息技术如计算机技术、缩微技术、光盘技术、多媒体技术、网络技术等高新技术群的快速发展，硬件应用的障碍大为减少，为信息资源数字化的建设和开发提供了实现的可能（林丹红，2001）。信息资源数字化建设包含两方面：

3.1 转化原有传统文献资源

对传统文献资源进行数字化加工处理，使之转化成为可以利用计算机识别、存取并进行网络传输利用的数字化信息资源。可以独立自主开发支持软件或系统工具建设，与其它信息机构合作开发建设，或者直接交给信息化专业公司加工定制（刘阳和丁银燕，2002）。

3.2 开发新的数字化信息资源

自主开发建设二、三次文献、数据库、电子书刊等数字化文献资源；再对这些新旧文献资源进行科学的加工、筛选、整合、重组、分类，建成新的数字化信息资源体系。或者建立导航网站，链接网上专业网站、搜索引擎、数据库服务中心、数字图书馆等，进行网上资源的导航组织、建立虚拟网上资源系统（刘阳和丁银燕，2002）。

4 数字化建设的原则

4.1 特色性原则

各领域学科在长期的发展建设中通过不断的研究探索和知识积累，已形成了自己的核心结构体系和专业资源特色。特色是信息资源数字化建设的生命。没有特色就没有竞争优势和发展潜力，就会失去生存价值。选择自身独有的，具有资源优势的专题和项目开发建设特色数字化资源系统、特色文献数据库，有利于形成学科专业优势，避免重复建设，实现资源优势互补、资源共享（刘阳和丁银燕，2002）。

4.2 标准化原则

数字化资源开发建设中，必须注意通用性原则问题，在开发建设中遵守网络传输协议、数据加工标准以及学科自身知识体系中的一些标准化规范，优先采用各种国际认可和流行的数据存储分发格式，确保数字化产品的通用性和标准化，便于网上传输、资源共享和全面普及（刘阳和丁银燕，2002）。

3 数字化建设的方法

随着现代信息技术如计算机技术、缩微技术、光盘技术、多媒体技术、网络技术等高新技术群的快速发展，硬件应用的障碍大为减少，为信息资源数字化的建设和开发提供了实现的可能（林丹红，2001）。信息资源数字化建设包含两方面：

3.1 转化原有传统文献资源

对传统文献资源进行数字化加工处理，使之转化成为可以利用计算机识别、存取并进行网络传输利用的数字化信息资源。可以独立自主开发支持软件或系统工具建设，与其它信息机构合作开发建设，或者直接交给信息化专业公司加工定制（刘阳和丁银燕，2002）。

3.2 开发新的数字化信息资源

自主开发建设二、三次文献、数据库、电子书刊等数字化文献资源；再对这些新旧文献资源进行科学的加工、筛选、整合、重组、分类，建成新的数字化信息资源体系。或者建立导航网站，链接网上专业网站、搜索引擎、数据库服务中心、数字图书馆等，进行网上资源的导航组织、建立虚拟网上资源系统（刘阳和丁银燕，2002）。

4 数字化建设的原则

4.1 特色性原则

各领域学科在长期的发展建设中通过不断的研究探索和知识积累，已形成了自己的核心结构体系和专业资源特色。特色是信息资源数字化建设的生命。没有特色就没有竞争优势和发展潜力，就会失去生存价值。选择自身独有的，具有资源优势的专题和项目开发建设特色数字化资源系统、特色文献数据库，有利于形成学科专业优势，避免重复建设，实现资源优势互补、资源共享（刘阳和丁银燕，2002）。

4.2 标准化原则

数字化资源开发建设中，必须注意通用性原则问题，在开发建设中遵守网络传输协议、数据加工标准以及学科自身知识体系中的一些标准化规范，优先采用各种国际认可和流行的数据存储分发格式，确保数字化产品的通用性和标准化，便于网上传输、资源共享和全面普及（刘阳和丁银燕，2002）。

4.3 安全性原则

网络是信息无限共享开放的空间, 数字化网络资源信息必须严肃对待安全保密问题。安全性包括防泄密与防数据丢失两个方面。防泄密、防病毒攻击或停电等意外事故造成数据丢失和系统破坏是数字化资源建设中的重大原则问题, 必须予以高度重视, 采取切实有效的安全措施和手段。如数据加密、限制使用范围、建立数据备份、制作只读光盘等(刘阳和丁银燕, 2002)。

5 文献资源数字化现状

以 20 世纪 60 年代美国国会图书馆正式发行书刊机读目录 LCMARC 为象征, 开始了馆藏文献资源电子化的实践。90 年代中期至今, 伴随着数字化概念的出现和信息处理技术的飞速发展, 文献资源进入了数字化发展阶段, 文献资源建设的重要性得到广泛的认同。全国高等教育文献资源保障系统计划启动, 中国国家图书馆的数字图书馆工程动工, 全国科技文献资源中心建设开始实施, 万方期刊数据群上网等(林丹红, 2001)。在这一发展过程中, 越来越多载体形式的文献逐渐被数字化, 不论是印刷文档、手写稿, 还是电子文档、音像文件等, 数据库建设朝着数字化、规模化方向发展。随着信息资源的飞速膨胀, 出现了数据仓库、数据集市、数据采集等新的信息技术概念和信息数字化处理方法, 逐步形成了大数据量存储和管理模式, 如清华同方光盘集团推出的机构知识仓库管理系统, 超星数字化公司推出的图文资源数字化 PDG 技术, 书生之家推出的全息数字化技术(杨晓农, 2004)等, 从各角度实践了各类文献数字化、信息化、标准化加工整理、编辑处理、数据存储和网络访问等。目前数字化文献保存和显示的主要类型有两种: 一是采用扫描录入方式将文献或图片资料按原貌逐页存储为图象文件, 如书生之家、超星数字图书馆; 二是以文本方式存储文献内容, 辅之以全文检索系统构成全文检索数据库, 如维普中文科技期刊全文数据库、中国学术期刊全文数据库。与此同时, 文献资源建设由个体向网络化方向发展, 着眼于全球性的资源布局和利用。

6 生物分类信息数字化兴起

信息化是全球化的基础和条件, 是人类社会当前最重要的历史潮流。传统分类学科为了适应信息时代发展的必然趋势, 提高学科的生存发展能力, 已积极投身传统学科的改造, 开展信息的数字化建设和服务。除了学科本身的发展需要, 生物分类信息走向数字化的动力还来自于人们对生物多样性认识与保护意识的不断加强。生物多样性兴起于 20 世纪 80 年代, 是指植物、动物和微生物的纷繁多样性及它们的遗传变异与它们所生存环境的总合, 是人类赖以生存和持续发展的物质基础(李宁和王姣, 2006)。我国生物多样性起始于 20 世纪 90 年代(迟德富等, 2006), 随着环境的污染和破坏造成生物多样性急剧下降, 人们对环境保

4.3 安全性原则

网络是信息无限共享开放的空间, 数字化网络资源信息必须严肃对待安全保密问题。安全性包括防泄密与防数据丢失两个方面。防泄密、防病毒攻击或停电等意外事故造成数据丢失和系统破坏是数字化资源建设中的重大原则问题, 必须予以高度重视, 采取切实有效的安全措施和手段。如数据加密、限制使用范围、建立数据备份, 制作只读光盘等 (刘阳和丁银燕, 2002)。

5 文献资源数字化现状

以 20 世纪 60 年代美国国会图书馆正式发行书刊机读目录 LCMARC 为象征, 开始了馆藏文献资源电子化的实践。90 年代中期至今, 伴随着数字化概念的出现和信息处理技术的飞速发展, 文献资源进入了数字化发展阶段, 文献资源建设的重要性得到广泛的认同。全国高等教育文献资源保障系统计划启动, 中国国家图书馆的数字图书馆工程动工, 全国科技文献资源中心建设开始实施, 万方期刊数据群上网等 (林丹红, 2001)。在这一发展过程中, 越来越多载体形式的文献逐渐被数字化, 不论是印刷文档、手写稿, 还是电子文档、音像文件等, 数据库建设朝着数字化、规模化方向发展。随着信息资源的飞速膨胀, 出现了数据仓库、数据集市、数据采集等新的信息技术概念和信息数字化处理方法, 逐步形成了大数据量存储和管理模式, 如清华同方光盘集团推出的机构知识仓库管理系统, 超星数字化公司推出的图文资源数字化 PDG 技术, 书生之家推出的全息数字化技术 (杨晓农, 2004) 等, 从各角度实践了各类文献数字化、信息化、标准化加工整理、编辑处理、数据存储和网络访问等。目前数字化文献保存和显示的主要类型有两种: 一是采用扫描录入方式将文献或图片资料按原貌逐页存储为图象文件, 如书生之家、超星数字图书馆; 二是以文本方式存储文献内容, 辅之以全文检索系统构成全文检索数据库, 如维普中文科技期刊全文数据库、中国学术期刊全文数据库。与此同时, 文献资源建设由个体向网络化方向发展, 着眼于全球性的资源布局和利用。

6 生物分类信息数字化兴起

信息化是全球化的基础和条件, 是人类社会当前最重要的历史潮流。传统分类学科为了适应信息时代发展的必然趋势, 提高学科的生存发展能力, 已积极投身传统学科的改造, 开展信息的数字化建设和服务。除了学科本身的发展需要, 生物分类信息走向数字化的动力还来自于人们对生物多样性认识与保护意识的不断加强。生物多样性兴起于 20 世纪 80 年代, 是指植物、动物和微生物的纷繁多样性及它们的遗传变异与它们所生存环境的总合, 是人类赖以生存和持续发展的物质基础 (李宁和王姣, 2006)。我国生物多样性起始于 20 世纪 90 年代 (迟德富等, 2006), 随着环境的污染和破坏造成生物多样性急剧下降, 人们对环境保

4.3 安全性原则

网络是信息无限共享开放的空间, 数字化网络资源信息必须严肃对待安全保密问题。安全性包括防泄密与防数据丢失两个方面。防泄密、防病毒攻击或停电等意外事故造成数据丢失和系统破坏是数字化资源建设中的重大原则问题, 必须予以高度重视, 采取切实有效的安全措施和手段。如数据加密、限制使用范围、建立数据备份, 制作只读光盘等 (刘阳和丁银燕, 2002)。

5 文献资源数字化现状

以 20 世纪 60 年代美国国会图书馆正式发行书刊机读目录 LCMARC 为象征, 开始了馆藏文献资源电子化的实践。90 年代中期至今, 伴随着数字化概念的出现和信息处理技术的飞速发展, 文献资源进入了数字化发展阶段, 文献资源建设的重要性得到广泛的认同。全国高等教育文献资源保障系统计划启动, 中国国家图书馆的数字图书馆工程动工, 全国科技文献资源中心建设开始实施, 万方期刊数据群上网等 (林丹红, 2001)。在这一发展过程中, 越来越多载体形式的文献逐渐被数字化, 不论是印刷文档、手写稿, 还是电子文档、音像文件等, 数据库建设朝着数字化、规模化方向发展。随着信息资源的飞速膨胀, 出现了数据仓库、数据集市、数据采集等新的信息技术概念和信息数字化处理方法, 逐步形成了大数据量存储和管理模式, 如清华同方光盘集团推出的机构知识仓库管理系统, 超星数字化公司推出的图文资源数字化 PDG 技术, 书生之家推出的全息数字化技术 (杨晓农, 2004) 等, 从各角度实践了各类文献数字化、信息化、标准化加工整理、编辑处理、数据存储和网络访问等。目前数字化文献保存和显示的主要类型有两种: 一是采用扫描录入方式将文献或图片资料按原貌逐页存储为图象文件, 如书生之家、超星数字图书馆; 二是以文本方式存储文献内容, 辅之以全文检索系统构成全文检索数据库, 如维普中文科技期刊全文数据库、中国学术期刊全文数据库。与此同时, 文献资源建设由个体向网络化方向发展, 着眼于全球性的资源布局和利用。

6 生物分类信息数字化兴起

信息化是全球化的基础和条件, 是人类社会当前最重要的历史潮流。传统分类学科为了适应信息时代发展的必然趋势, 提高学科的生存发展能力, 已积极投身传统学科的改造, 开展信息的数字化建设和服务。除了学科本身的发展需要, 生物分类信息走向数字化的动力还来自于人们对生物多样性认识与保护意识的不断加强。生物多样性兴起于 20 世纪 80 年代, 是指植物、动物和微生物的纷繁多样性及它们的遗传变异与它们所生存环境的总合, 是人类赖以生存和持续发展的物质基础 (李宁和王姣, 2006)。我国生物多样性起始于 20 世纪 90 年代 (迟德富等, 2006), 随着环境的污染和破坏造成生物多样性急剧下降, 人们对环境保

护和自然资源持续利用的意识在日益提高,但人们对生物多样性的认识比较肤浅,对现存物种种类和其功能多样性了解甚少,保护工作较难开展。因此学习鉴别物种、了解物种诸多的分类学知识,成为提高生物多样性认识水平的基础工作。目前,国家和地方各职能部门正积极扩大生物多样性知识的获取途径与传播范围,开展部门与学科的信息合作,促进生物分类知识的数字化收集整理,并通过网络发布在全社会范围内普及(Edwards, 2000; 纪力强, 2000)。正是在生物多样性与生物分类不断协同作用下,传统分类学科信息资源数字化建设的步伐越来越大。

7 生物分类信息数字化方法

按照生物分类研究所拥有的内容、性质、特点和作用等,分类信息的组成主要有文字信息、图象信息、声音信息、标本信息和检索信息等 5 个方面。按其特性或应用价值的差别,各个部分的数字化方法各有异同。

7.1 分类文字信息

系指生物分类领域中以文字为形式或载体而存有的分类信息,包括陈旧性文字信息与新鲜性文字信息。它们主要用以描述物种的各类名称、分类地位、形态特征、生物学、分布范围、标本记录、备注、参考文献等信息。不论新旧,它们都以各种分类学文献为渊源或载体而存在,如图书期刊、书籍著作及各种历史资料等。因此,这部分信息的数字化初期将以各种载体文献输入,通过光学扫描、文字识别、内容校正、版面还原等一整套工序完成从印刷版到电子版的转变;后期再利用数据加工和处理技术,实现这些电子信息的主题分类、自动标引、系统集成以及数据库存储等,便于深入的数据挖掘与分析处理。

7.2 分类图像信息

系指生物分类领域中以图像为载体而存有的分类信息,包括实物图、模式图、特征图和区域分布图等图像信息。分类图像信息不仅具有其它信息无可比拟的真实性,而且在信息的作用上又有其它信息没有的感召性和生动性(乔凤海, 2000)。图像一般通过数码拍摄或者光学扫描进行采集后保存为压缩统一格式的图像文件或存入数据库中,必要时加以图象剪裁、柔化、锐化、标注等后期处理。而模式图、区域分布图也可直接通过图像软件或者 GIS 系统数字化创建。

7.3 分类声音信息

系指生物分类领域中以音频为载体而存有的分类信息,包括鸣声(20-20KHz)、超声(>20KHz)。鸣声属于行为特征,在动物分类上应用至今尚不广泛,其潜力很大,特别是在

护和自然资源持续利用的意识在日益提高,但人们对生物多样性的认识比较肤浅,对现存物种种类和其功能多样性了解甚少,保护工作较难开展。因此学习鉴别物种、了解物种诸多的分类学知识,成为提高生物多样性认识水平的基础工作。目前,国家和地方各职能部门正积极扩大生物多样性知识的获取途径与传播范围,开展部门与学科的信息合作,促进生物分类知识的数字化收集整理,并通过网络发布在全社会范围内普及(Edwards, 2000; 纪力强, 2000)。正是在生物多样性与生物分类不断协同作用下,传统分类学科信息资源数字化建设的步伐越来越大。

7 生物分类信息数字化方法

按照生物分类研究所拥有的内容、性质、特点和作用等,分类信息的组成主要有文字信息、图象信息、声音信息、标本信息和检索信息等 5 个方面。按其特性或应用价值的差别,各个部分的数字化方法各有异同。

7.1 分类文字信息

系指生物分类领域中以文字为形式或载体而存有的分类信息,包括陈旧性文字信息与新鲜性文字信息。它们主要用以描述物种的各类名称、分类地位、形态特征、生物学、分布范围、标本记录、备注、参考文献等信息。不论新旧,它们都以各种分类学文献为渊源或载体而存在,如图书期刊、书籍著作及各种历史资料等。因此,这部分信息的数字化初期将以各种载体文献输入,通过光学扫描、文字识别、内容校正、版面还原等一整套工序完成从印刷版到电子版的转变;后期再利用数据加工和处理技术,实现这些电子信息的主题分类、自动标引、系统集成以及数据库存储等,便于深入的数据挖掘与分析处理。

7.2 分类图像信息

系指生物分类领域中以图像为载体而存有的分类信息,包括实物图、模式图、特征图和区域分布图等图像信息。分类图像信息不仅具有其它信息无可比拟的真实性,而且在信息的作用上又有其它信息没有的感召性和生动性(乔凤海, 2000)。图像一般通过数码拍摄或者光学扫描进行采集后保存为压缩统一格式的图像文件或存入数据库中,必要时加以图象剪裁、柔化、锐化、标注等后期处理。而模式图、区域分布图也可直接通过图像软件或者 GIS 系统数字化创建。

7.3 分类声音信息

系指生物分类领域中以音频为载体而存有的分类信息,包括鸣声(20-20KHz)、超声(>20KHz)。鸣声属于行为特征,在动物分类上应用至今尚不广泛,其潜力很大,特别是在

近缘种分类上更具有重要的意义(隋艳晖等, 2003)。声音数字卡采集可通过音频传感器(拾音器)配套数据采集卡录制到计算机中, 保存为波形文件(姚青等, 2001)。此后可用多媒体播放器回放, 音频编辑软件进行剪辑、修饰和降噪。

7.4 分类标本信息

系指在生物分类领域中以标本为载体而存有的分类信息, 包括生物形态学、解剖学、生理学、病理学、化石等标本信息。标本以最真实不过的特性, 向人们展示了其中存有的研究价值(乔凤海, 2000)。标本信息是生物多样性数据库的重要组成信息, 通过数码拍摄、3D 成像或者录像可记录标本全貌, 在计算机中浏览回放, 用于普及分类知识、教学及参观等。

7.5 分类检索信息

系指在生物分类领域中以检索表形式表达的分类信息, 主要指常用的二项式检索表。检索表以区分生物为目的编制, 从分目到分种每个生物分类水平都有相应的检索表, 分布广泛, 数量众多。这些信息大多都收录在各种分类学期刊、著作和指导物种鉴定的工具书中, 因此可把它们当作文字信息来数字化。不过与一般文字信息不同, 检索表具有统一的结构性, 把它变成电子版后需要经过一定的格式编码处理, 才能进入计算机数字化分析使用的流程。这要求专门制定电子检索表编码的统一标准, 既体现检索表的结构特色性, 又保证其通用性。

8 生物分类信息数字化资源

计算机的日益普及, 互联网的蓬勃发展以及人们对于生物多样性认识与保护的意识不断增强, 都促使生物分类信息逐步进入数字化、网络化的发展轨道。目前网上生物分类的数字化信息已较丰富, 以数字期刊、数字图书馆等收录的文献资源和生物多样性数据库、生物分类检索系统等收录的分类信息最为全面详尽。但其中数字文献以文字描述为主, 数量最多, 但随印刷文献发表在更新, 资料混杂不成系统(Norris, 2000)。而多样性数据库与检索系统采用系统化或专题化形式设计, 内容集中有序, 信息表现多样, 检索功能强大, 已成为较流行的生物分类数字化资源。

8.1 中国动物物种编目数据库

它是自 1992 年启动的中国科学院“八五”重大科研项目开始建设, 到目前为止已经收录了 25 000 余种(亚种)动物的基本信息, 由中科院动物研究所维护更新。该数据库包括动物的分类阶元、分类编号、原始文献、模式产地、同物异名、俗名、英文名、生境、海拔、分布范围等数据, 可根据动物的分类阶元, 如目名、科名、种名等, 查询动物的信息, 还可通过动物的分布区查询(纪力强, 2007)。同时, 中科院微生物研究所还提供了另外一个中

近缘种分类上更具有重要的意义(隋艳晖等, 2003)。声音数字卡采集可通过音频传感器(拾音器)配套数据采集卡录制到计算机中, 保存为波形文件(姚青等, 2001)。此后可用多媒体播放器回放, 音频编辑软件进行剪辑、修饰和降噪。

7.4 分类标本信息

系指在生物分类领域中以标本为载体而存有的分类信息, 包括生物形态学、解剖学、生理学、病理学、化石等标本信息。标本以最真实不过的特性, 向人们展示了其中存有的研究价值(乔凤海, 2000)。标本信息是生物多样性数据库的重要组成信息, 通过数码拍摄、3D 成像或者录像可记录标本全貌, 在计算机中浏览回放, 用于普及分类知识、教学及参观等。

7.5 分类检索信息

系指在生物分类领域中以检索表形式表达的分类信息, 主要指常用的二项式检索表。检索表以区分生物为目的编制, 从分目到分种每个生物分类水平都有相应的检索表, 分布广泛, 数量众多。这些信息大多都收录在各种分类学期刊、著作和指导物种鉴定的工具书中, 因此可把它们当作文字信息来数字化。不过与一般文字信息不同, 检索表具有统一的结构性, 把它变成电子版后需要经过一定的格式编码处理, 才能进入计算机数字化分析使用的流程。这要求专门制定电子检索表编码的统一标准, 既体现检索表的结构特色性, 又保证其通用性。

8 生物分类信息数字化资源

计算机的日益普及, 互联网的蓬勃发展以及人们对于生物多样性认识与保护的意识不断增强, 都促使生物分类信息逐步进入数字化、网络化的发展轨道。目前网上生物分类的数字化信息已较丰富, 以数字期刊、数字图书馆等收录的文献资源和生物多样性数据库、生物分类检索系统等收录的分类信息最为全面详尽。但其中数字文献以文字描述为主, 数量最多, 但随印刷文献发表在更新, 资料混杂不成系统(Norris, 2000)。而多样性数据库与检索系统采用系统化或专题化形式设计, 内容集中有序, 信息表现多样, 检索功能强大, 已成为较流行的生物分类数字化资源。

8.1 中国动物物种编目数据库

它是自 1992 年启动的中国科学院“八五”重大科研项目开始建设, 到目前为止已经收录了 25 000 余种(亚种)动物的基本信息, 由中科院动物研究所维护更新。该数据库包括动物的分类阶元、分类编号、原始文献、模式产地、同物异名、俗名、英文名、生境、海拔、分布范围等数据, 可根据动物的分类阶元, 如目名、科名、种名等, 查询动物的信息, 还可通过动物的分布区查询(纪力强, 2007)。同时, 中科院微生物研究所还提供了另外一个中

国动物物种编目数据库 (<http://www.bioinfo.cn/db05/BjdwSpecies.php>), 收录分布在我国哺乳类、鸟类、爬行类、两栖类、鱼类、无脊椎动物、昆虫等 19 000 种及亚种的信息, 内容与前者差不多, 但检索方式更多样。

8.2 动物物种多样性数据库

该数据库收录的物种信息较齐全, 包括拉丁名称、英文名称、特征、分布和生境, 所有物种从门到种依现行分类系统划分。最为可贵的是, 几乎每个物种都附有形象的图片说明, 并注明图片来源。它本身没有公布收录的物种数, 但从分类列表来看包括原生动植物到脊索动物共 25 个门 (EC 网络, 2007), 涉及面广, 内容丰富。该数据库中山大学生命科学学院出品维护。

8.3 生物数字标本区

该数据库是一个货真价实的图片数据库 (http://biomuseum.sysu.edu.cn/ASP/search/hexapod/hexapod_search.htm), 它以数字博物馆为着眼点, 用一幅幅清晰高质量图片展示了近 5 万个收藏标本, 其中昆虫 839 个, 其他动物 828 个, 植物 46 369 个, 化石 266 个。唯一不足的是, 每个标本只有名称及标本属性信息, 没有特征、分布等其他信息。不过该数据库提供的“标本局部放大”观察功能, 可让你看清标本几乎每个细节。它由中山大学数字博物馆制作出品, 该馆前身是在亚洲享有盛誉的原岭南大学自然博物采集所 (Biodata, 2007)。

8.4 Animal Diversity Web

Animal Diversity Web (<http://animaldiversity.ummz.umich.edu/site/index.html>) 是密歇根大学动物自然史、分布、分类和生物保护等内容组成的在线数据库, 由专业生物学家为其收录的物种及以上阶元准备了大量包含特征描述和一般生物学介绍的网页和图片。ADW 自比为面向全球的在线百科全书、科学学习工具和生动的博物馆, 某些物种还采用了“虚拟现实”的 3D 动画、声音、动物活动录象等形象展示 (Kaiser, 1999)。

8.5 Species 2000

Species 2000 & ITIS Catalogue of Life (<http://www.sp2000.org>) 每年收集 37 个分类数据库的信息形成一个固定的访问物种清单 (Annual Checklist), 并提供对 26 个在线分类数据库的即时访问 (Dynamic Checklist)。每条搜索结果中都含有信息来源的数据库链接。它以索引所有已知生物为目标, 计划到 2011 年收集并建成地球上所有已知物种的目录, 目前进度已近一半 (Norris, 2000; Bisby, 2000)。

8.6 Tree of Life

美国 1996 年 1 月开始的“生命之树网页工程 ToL” (<http://www.tolweb.org/tree>) 逐步建立起了整个生物界的分类系统, 它的目标是给地球上每一个物种或类群, 不管是活着的还是灭绝的, 提供文字描述、图片和相关网络资源导航的信息。这些信息按照系统发生关系以树形结构和超链接方式组织展示 (Pennisi, 2001)。ToL 目前已收纳 5 000 多个网页, 并继续由生物学家、教师、学生、科学爱好者等为不同的系统进化树节点上传资料, 生成更多的网页。

8.7 World Biodiversity Database

World Biodiversity Database (<http://nlbif.eti.uva.nl/bis/index.php>) 是一个正在成长中的分类数据库和信息系统, 它包括由 20 个生物项目构建的物种库, 内容除了常用分类信息、描述、图片、参考文献外, 还提供在线使用的检索表和交互式地理分布系统。该数据库由 ETI 公司著名的生物信息网络发布工具 Linnaeus II 制作。

8.8 等翅目昆虫分类系统

该系统隶属于浙江大学城市昆虫学研究中心, 由白蚁分类资料显示子系统、白蚁分类资料检索子系统、后台数据库在线管理子系统和白蚁种类鉴定子系统等 4 个功能模块组成。用户通过 Internet 网络, 可实现白蚁分类资料的查询与未知种类的鉴定 (徐晓国等, 2004), 其中种类鉴定子系统是基于二项检索表知识库实现的检索系统。

8.9 中国蝗总科分类查询及鉴定专家系统 ESCA

ESCA 是应我国蝗灾治理工作需要而建立的一种快捷高效的蝗虫分类、查询、鉴定工具。它以专家系统为模式设计, 建立了蝗总科形态学、分类学数据库, 并实现对该数据库智能化的高效自动查询, 并以问答式检索系统引导用户进行标本鉴定 (卢慧麓和黄原, 2003)。

表 1.1 对上述生物多样性数据库和检索系统进行了收录物种数与信息内容的对比, 供全面参考。同时网上还有专门的生物多样性与生物数据库索引或搜索引擎, 国外如密西根大学动物博物馆 Biocollection (NSF 资助) 网站 (<http://biocollections.org>) 致力于提供有关物种资源、分类权威文献、生物学家目录、DELTA 系统等软件以及大量生物多样性数据库的入口链接, 是网上查询生物多样性与生物数据库信息的重要索引。国内如物种数据库集成搜索引擎 (<http://www.biodata.cn>), 提供在近 100 个动植物、微生物多样性数据库中查询信息。

表 1.1 生物多样性数据库与分类检索系统收录信息比较

Table 1.1 Comparison of information collected in biodiversity databases and retrieval systems of biological taxonomy

多样性数据库	物种数	收录信息						
		基本 信息 ¹	特征	分布	生境	图片	检索 表	进化 树
中国动物物种编 目数据库	25 000	√		√	√			
动物物种多样性 数据库	-	√	√	√	√	√		
生物数字标本区	48 313	√				√		
Tree of life	-	√	√			√		√
Animal Diversity Web	-	√	√	√	√	√		
World Biodiversity Database	25 472	√	√	√	√	√	√	√
Sp2000 Annual Checklist 2006	884 552	√		√				√
等翅目分类系统	476	√	√	√	√		√	√
ESCA	855	√	√	√	√	√	√	√

9 生物分类信息数字化开发问题

我国数字化建设起步较晚，网上信息资源有限，相对于较完善的网络基础设施和人们日益提升的数字信息需求，资源数字化工作明显滞后（刘阳和丁银燕，2002）。随着数字化建设不断地深化开展，分类文献数字资源和生物多样性系统（数据库）正日趋丰富，但同生物分类研究的总体进度相比，信息资源的数字化发展仍显缓慢。这与分类信息数字化建设中出现的一些问题直接相关。

9.1 分类专家参与数字化比例低

从现有的生物分类数字化资源形成过程来看，分类专家一直扮演着分类知识原始生产者的角色，而对分类信息的数字化工作投入极其有限。以昆虫为例，国内互联网上大部分昆虫网站出自非专业人士或官方机构之手，内容侧重在对昆虫经济利用与观赏价值、基础知识的宣传介绍（吴焰玉和汪家社，2001），对专业系统的分类信息涉及粗略。真正掌握分类信息的研究专家，由于大部分时间专注于物种分类研究，缺少必要的信息软件技术支持，导致很

¹ 包括中文名、拉丁学名、分类地位等

表 1.1 生物多样性数据库与分类检索系统收录信息比较

Table 1.1 Comparison of information collected in biodiversity databases and retrieval systems of biological taxonomy

多样性数据库	物种数	收录信息						
		基本 信息 ¹	特征	分布	生境	图片	检索 表	进化 树
中国动物物种编 目数据库	25 000	√		√	√			
动物物种多样性 数据库	-	√	√	√	√	√		
生物数字标本区	48 313	√				√		
Tree of life	-	√	√			√		√
Animal Diversity Web	-	√	√	√	√	√		
World Biodiversity Database	25 472	√	√	√	√	√	√	√
Sp2000 Annual Checklist 2006	884 552	√		√				√
等翅目分类系统	476	√	√	√	√		√	√
ESCA	855	√	√	√	√	√	√	√

9 生物分类信息数字化开发问题

我国数字化建设起步较晚，网上信息资源有限，相对于较完善的网络基础设施和人们日益提升的数字信息需求，资源数字化工作明显滞后（刘阳和丁银燕，2002）。随着数字化建设不断地深化开展，分类文献数字资源和生物多样性系统（数据库）正日趋丰富，但同生物分类研究的总体进度相比，信息资源的数字化发展仍显缓慢。这与分类信息数字化建设中出现的一些问题直接相关。

9.1 分类专家参与数字化比例低

从现有的生物分类数字化资源形成过程来看，分类专家一直扮演着分类知识原始生产者的角色，而对分类信息的数字化工作投入极其有限。以昆虫为例，国内互联网上大部分昆虫网站出自非专业人士或官方机构之手，内容侧重在对昆虫经济利用与观赏价值、基础知识的宣传介绍（吴焰玉和汪家社，2001），对专业系统的分类信息涉及粗略。真正掌握分类信息的研究专家，由于大部分时间专注于物种分类研究，缺少必要的信息软件技术支持，导致很

¹ 包括中文名、拉丁学名、分类地位等

少能将分类成果整理成专业性的分类网站。从信息共建发展模式来看,分类专家可以主动依托具备数字信息开发能力的第三方机构或个人来完成分类信息的数字化工作,也可以直接利用专业的信息数字化开发平台,只需整理手头的分类资料便能实现网上发布。这样在占用较少时间的情况下,分类专家也能建立个人研究专题的专业分类网站,并将最新的分类信息转化为数字化资源。一旦分类专家参与数字化建设形成共识,必将大幅加快生物分类信息的数字化进程。

9.2 分类信息数字化开发工具匮乏

如今电子化办公已十分普及,许多第一手的分类成果可直接用于数字化建设。然而摆在分类专家面前的主要问题是如何将这些电子资料建成用于指导分类学习和鉴定决策需要的知识库,提供给更多人研究使用。这要求综合运用软件、数据库、多媒体和网络等技术才能实现分类知识的分步集成与网络发布。技术问题一直是制约信息资源共享的主要问题(杨晓农, 2004),分类专家掌握的信息技术有限,如果没有现成的技术支持工具可用,这些分类信息以印刷出版形式发表,造成第三方重复的数字转化劳动。因此,为分类专家量身打造分类知识集成与发布工具成为分类数字化建设的迫切要求。目前国外已有 ETI 等生物信息技术公司为系统分类学家和多样性研究者开发的多样性数据库与物种鉴定统一支持系统,并在积极推广中得到了成熟应用。而国内大多见到的是分类检索系统或者诊断专家系统的开发平台,能满足分类专家知识集成需要的数字化开发工具却不见踪影。这也难怪分类专家虽有数字化建设的热情,但仍难以付诸实际行动。

9.3 检索表数字资源急待深入开发利用

检索表是生物分类的产物,是生物分类信息重要的组成部分。从目前生物分类数字化资源组成看来,生物多样性数据库、检索系统收录的检索表较少,大量检索表集中在数字化文献中,甚至更多检索表还沉睡在书本中。对于仍处在印刷版的检索表,有待逐步完成电子化采集。对于电子版的检索表,已可以在网上传递共享,但使用起来与印刷版没什么不同。检索表的数字化水平并不局限于此,除了表面描述性信息,其内含的结构化鉴定思路也应提取出来进一步数字化处理,变成计算机也能理解利用的分类鉴定知识(Payne and Preece, 1980)。虽然检索系统普遍都有这方面的深入开发,但处理水平有限,不能满足检索表作为数字资源多方面作用的挖掘利用。从目前看来,电子检索表还需要建立一个统一的数字编码标准,以较高的数据集成模式记录检索表的文字信息和鉴定信息,以利于发挥检索表数字化后的潜在优势。

第二章 分类检索表概述

1 分类的意义

分类是认识客观事物最基本的方法。远在原始时代,人类在生产实践中,就需要辨别周围的事物,哪些是可吃的,哪些是不可吃的,哪些是有害的,哪些是无害的,就产生了初步的分类概念。千差万别的事物,只有通过分析对比与归纳才能分门别类。所谓分析对比与归纳,就是研究事物的特殊性与共同性的对立统一,没有共同性就没有特殊性,没有特殊性也就看不出共同性。分类就是建立在特殊性的分析与共同性的归纳之上(管致和,1999)。以认识昆虫为例,它们是世界上生物多样性最丰富的类群,全球预计昆虫有1000多万种(黄复生,1991),现已定名只约100多万种,也就是说绝大多数种类我们至今还不知道。物质是可以认识的,但我们要认识它,必须有正确的分类方法。从这个意义上来说,昆虫分类是认识昆虫的一种基础方法,是昆虫学其他所有分支学科的基础,如果不先对昆虫进行科学的分类,就无法以科学的方式研究昆虫,影响到其他研究结果的客观性、可比性和重复性,甚至给人类的生产和生活带来一定的损失。

随着生产的不断发展,人类对分类的要求也进一步提高,从个别的、表面的现象分类,进入到内在的、本质的分类。生物都由低等到高等进化过来,起源于共同的祖先,有着血缘的或近或远,或亲或疏的关系,有着进化上的间断性与连续性的对立统一。生物分类就要正确地反映出它们历史演化的过程,正确地反映这种潜在的系谱关系。因此,昆虫分类的任务,不仅仅是区别异同、鉴定名称,还要进而研究物种的渊源、自然位置,借以阐明自然界物种间以及类群间的亲缘关系,建立符合客观的分类系统,并以种群的观点,研究物种起源、分布中心、昆虫进化的过程和趋向,以及整个昆虫区系的形成、发展和演替,使我们能够更有效地控制有害昆虫和利用有益昆虫。事实证明,昆虫中亲缘关系愈接近的,它们的形态结构愈相近,对环境条件的要求愈相同,其发生发展的规律也愈相接近。例如鞘翅目昆虫都是咀嚼式口器,叶甲总科中的天牛科昆虫都是植食性的,它们的幼虫专钻蛀木材部分,而同一总科的豆象幼虫则专蛀食豆科的种子,叶甲则为害植物的叶,水叶甲专为害水生植物被水淹没的部分,而铁甲则潜入组织中取食叶肉。明确了它们的分类地位,也就等于了解它们的一部分生活规律(管致和,1999)。

2 检索表的作用

大千世界,物种丰富多样,用什么方法可以帮助我们认识各种各样的生物呢?现在所采用的方法主要有两种,一种是核对法,即将所采标本用标本室中专家已鉴定定名的标本进行

第二章 分类检索表概述

1 分类的意义

分类是认识客观事物最基本的方法。远在原始时代,人类在生产实践中,就需要辨别周围的事物,哪些是可吃的,哪些是不可吃的,哪些是有害的,哪些是无害的,就产生了初步的分类概念。千差万别的事物,只有通过分析对比与归纳才能分门别类。所谓分析对比与归纳,就是研究事物的特殊性与共同性的对立统一,没有共同性就没有特殊性,没有特殊性也就看不出共同性。分类就是建立在特殊性的分析与共同性的归纳之上(管致和,1999)。以认识昆虫为例,它们是世界上生物多样性最丰富的类群,全球预计昆虫有1000多万种(黄复生,1991),现已定名只约100多万种,也就是说绝大多数种类我们至今还不知道。物质是可以认识的,但我们要认识它,必须有正确的分类方法。从这个意义上来说,昆虫分类是认识昆虫的一种基础方法,是昆虫学其他所有分支学科的基础,如果不先对昆虫进行科学的分类,就无法以科学的方式研究昆虫,影响到其他研究结果的客观性、可比性和重复性,甚至给人类的生产和生活带来一定的损失。

随着生产的不断发展,人类对分类的要求也进一步提高,从个别的、表面的现象分类,进入到内在的、本质的分类。生物都由低等到高等进化过来,起源于共同的祖先,有着血缘的或近或远,或亲或疏的关系,有着进化上的间断性与连续性的对立统一。生物分类就要正确地反映出它们历史演化的过程,正确地反映这种潜在的系谱关系。因此,昆虫分类的任务,不仅仅是区别异同、鉴定名称,还要进而研究物种的渊源、自然位置,借以阐明自然界物种间以及类群间的亲缘关系,建立符合客观的分类系统,并以种群的观点,研究物种起源、分布中心、昆虫进化的过程和趋向,以及整个昆虫区系的形成、发展和演替,使我们能够更有效地控制有害昆虫和利用有益昆虫。事实证明,昆虫中亲缘关系愈接近的,它们的形态结构愈相近,对环境条件的要求愈相同,其发生发展的规律也愈相接近。例如鞘翅目昆虫都是咀嚼式口器,叶甲总科中的天牛科昆虫都是植食性的,它们的幼虫专钻蛀木材部分,而同一总科的豆象幼虫则专蛀食豆科的种子,叶甲则为害植物的叶,水叶甲专为害水生植物被水淹没的部分,而铁甲则潜入组织中取食叶肉。明确了它们的分类地位,也就等于了解它们的一部分生活规律(管致和,1999)。

2 检索表的作用

大千世界,物种丰富多样,用什么方法可以帮助我们认识各种各样的生物呢?现在所采用的方法主要有两种,一种是核对法,即将所采标本用标本室中专家已鉴定定名的标本进行

对照,特征相似者可初步确定可能就是该物种。另外一种方法是检索法,即在对标本进行全面的观察后,查阅各种工具书(如动植物志、图鉴、图说、图谱手册等)中的检索表对其进行检索鉴定,当检索出初步结果后,再与书中对该种详尽的特征描述进行核对,如其他特征是否一致、地理分布区域是否符合,以进一步确认鉴定结果(周云龙,2007)。

第一种方法要求核对前先要能辨认标本到科或属,进入标本室后才能方便地查阅核对。然而生物种类繁多,世界上没有一个标本室能容下数以万计的生物种类。尽管已有特定生物类群的标本室,但标本数量众多,鉴定效率难以提高,单靠比对标本特征,鉴定结果可信度也不高。相比之下,第二种方法要灵活方便,准确性也高。从检索表的设计来看,表中的特征一般都是从不同阶元提取出来的比较重要、突出、明显而稳定的特征,据此鉴定生物对象,速度又快,结果又好,即使检索过程中出了差错,也可从错判位置重新继续检索。从检索表的获取途径来看,由于它是文字工具,通过书籍、刊物等纸质载体即可传播使用,比去标本室核对特征要方便多。因此,检索法作为一种便捷高效的物种鉴定方法而一直沿用保留下来,为了能快速、方便地鉴定生物种类,无论哪种分类工具书,都会在书中编制大量检索表。可以认为,检索表已成为鉴定生物、认识生物不可缺少的工具,是认识它们的一把钥匙。

3 检索表的类型与结构

传统检索表根据其特征布局和排列形式,可分成包孕式、连续式和二项式三种。为了简明起见,一般以林奈所定的7目昆虫²为例加以说明(管致和,1999)。

3.1 包孕式

又名定距式、等距式、退格式、不齐头检索表,其优点是各不同单元的关系清晰醒目,缺点是相对性状相离很远,尤其在冗长的检索中,浪费篇幅,一般仅在包含种类数较少时应用,格式是:

- A. 有翅
 - B. 口器咀嚼式
 - C. 翅两对
 - D. 前翅膜质
 - E. 前翅不被鳞片
 - F. 雌腹部末端有蜚.....膜翅目
 - FF. 雌腹部末端无蜚.....脉翅目

²林奈7目的分类,现在已经不用,他的有吻目现在包括半翅目和同翅目,鞘翅目包括直翅目,无翅目包括很多无翅的目

对照,特征相似者可初步确定可能就是该物种。另外一种方法是检索法,即在对标本进行全面的观察后,查阅各种工具书(如动植物志、图鉴、图说、图谱手册等)中的检索表对其进行检索鉴定,当检索出初步结果后,再与书中对该种详尽的特征描述进行核对,如其他特征是否一致、地理分布区域是否符合,以进一步确认鉴定结果(周云龙,2007)。

第一种方法要求核对前先要能辨认标本到科或属,进入标本室后才能方便地查阅核对。然而生物种类繁多,世界上没有一个标本室能容下数以万计的生物种类。尽管已有特定生物类群的标本室,但标本数量众多,鉴定效率难以提高,单靠比对标本特征,鉴定结果可信度也不高。相比之下,第二种方法要灵活方便,准确性也高。从检索表的设计来看,表中的特征一般都是从不同阶元提取出来的比较重要、突出、明显而稳定的特征,据此鉴定生物对象,速度又快,结果又好,即使检索过程中出了差错,也可从错判位置重新继续检索。从检索表的获取途径来看,由于它是文字工具,通过书籍、刊物等纸质载体即可传播使用,比去标本室核对特征要方便多。因此,检索法作为一种便捷高效的物种鉴定方法而一直沿用保留下来,为了能快速、方便地鉴定生物种类,无论哪种分类工具书,都会在书中编制大量检索表。可以认为,检索表已成为鉴定生物、认识生物不可缺少的工具,是认识它们的一把钥匙。

3 检索表的类型与结构

传统检索表根据其特征布局和排列形式,可分成包孕式、连续式和二项式三种。为了简明起见,一般以林奈所定的7目昆虫²为例加以说明(管致和,1999)。

3.1 包孕式

又名定距式、等距式、退格式、不齐头检索表,其优点是各不同单元的关系清晰醒目,缺点是相对性状相离很远,尤其在冗长的检索中,浪费篇幅,一般仅在包含种类数较少时应用,格式是:

- A. 有翅
 - B. 口器咀嚼式
 - C. 翅两对
 - D. 前翅膜质
 - E. 前翅不被鳞片
 - F. 雌腹部末端有蜚.....膜翅目
 - FF. 雌腹部末端无蜚.....脉翅目

²林奈7目的分类,现在已经不用,他的有吻目现在包括半翅目和同翅目,鞘翅目包括直翅目,无翅目包括很多无翅的目

EE. 前翅密被鳞片	鳞翅目
DD. 前翅角质	鞘翅目
CC. 翅一对	双翅目
BB. 口器刺吸式	有吻目
AA. 无翅	无翅目

在这种检索表中，每一对相对特征前编有相关联的序号，并纵向相隔一定距离，且都书写在距书页左边同等距离的地方；每个分支的下边，又出现两个相对应的分支，再编写关联序号，书写在较先出现的一个分支序号向右退一个字格的地方，这样如此往复下去，直到编制的终点为止。

3.2 连续式

又名单项式，这种形式的检索表，具有包孕式检索表同样的优点，而篇幅比较节省，但相对性状还是相离很远，格式是：

1 (12) 有翅	
2 (11) 口器咀嚼式	
3 (10) 翅两对	
4 (9) 前翅膜质	
5 (8) 前翅不被鳞片	
6 (7) 雌腹部末端有蜚刺	膜翅目
7 (6) 雌腹部末端无蜚刺	脉翅目
8 (5) 前翅密被鳞片	鳞翅目
9 (4) 前翅角质	鞘翅目
10 (3) 翅一对	双翅目
11 (2) 口器刺吸式	有吻目
12 (1) 无翅	无翅目

在这种检索表中，每一条仅含一项，与其后所指示的特征相对应，所鉴定的对象若符合，就继续向下检索，若不符合，就检索其后括号中的序号，总条数为所含种类数 2 倍减 2。

3.3 二项式

又名齐头检索表，它是由图 2.1 这种结构重复出现组成，是目前最通用的形式。

- ① 特征序号：从 1 开始，随着特征对数增多而变大，其相对特征一般开头用“-”；
- ② 特征描述；
- ③ 分隔符：一般采用“-”或“•”符号，连续 3 次以上；

④ 跳转结果：指向特征序号或者鉴定结果。

其优点是每对性状互相靠近，便于比较，篇幅也节省，主要缺点是各单元的关系有时不明显。

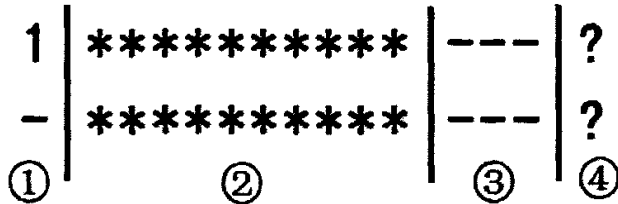


图 2.1 二项检索表的基本结构

Fig.2.1 Basic structure of dichotomous keys

在这种检索表中，每一条包含两项对应的特征，所鉴定的对象符合哪一项，就按哪一项所指示的条数继续向下检索，直至检索到其名称为止，总条数为所含种类数减 1（表 2.1）。

表 2.1 林奈 7 目昆虫二项检索表

Table 2.1 Dichotomous key to seven insect orders established by Linnaeus

1 无翅·····	无翅目
- 有翅·····	2
2 口器刺吸式·····	有吻目
- 口器咀嚼式·····	3
3 翅一对·····	双翅目
- 翅两对·····	4
4 前翅角质·····	鞘翅目
- 前翅膜质·····	5
5 翅不被鳞片·····	6
- 翅被鳞片·····	鳞翅目
6 雌腹部末端有蜚刺·····	膜翅目
- 雌腹部末端无蜚刺·····	脉翅目

近年来随着计算机技术尤其是编程技术、多媒体技术与网络技术的迅速普及，生物分类鉴定的方法得到了长足的进展，检索工具的作用不断增强，使用也更加方便，出现了一些新形式、新用法的检索表，如问答式检索表、图文式检索表、交互性多途径检索表（Edwards and Morse, 1995; Snow, 1999）、模式检索表（Pomar and Hidalgo, 1998; Natural History Museum, 2007）等。其中以交互性多途径检索表最有广泛影响，国外 Meka（Duncan and Meacham, 1986; Meacham, 2005）、XID Authoring System（Intelsys, 2001）、DELTA IntKey（Dallwitz et al., 2006b）、Lucid Professional（CBIT, 2007a）等都属于此类检索表工具。它突破了传统单途径检索必须拘

鉴于检索表已设计好的特征鉴定路线，用户可选择任何可用的特征开始鉴定，不同人可能以不同的特征比对顺序完成同一对象的鉴定，Snow (1999) 认为它是继分子系统发育分析后系统分类研究中另一重要创新。不过它对检索表编制的要求较高，设计者必须清楚了解对象与每个特征的匹配关系。除此之外，多途径检索表的制作与使用都需要专门的计算机软件，手工难以完成。

4 检索表的参考特征

为了方便分析与比较检索表的优劣，Calvo-Flores 等 (2006) 曾引入了一组与检索表直接相关的特征：

平均长度、最大长度、最小长度：平均长度是检索表鉴定物种平均步数的指示值。由于专家经常寻找能以最少步数完成鉴定的检索表，平均长度便成为一个重要的参考。

平均长度的变异度：它是检索表平衡度的衡量值。若该值较大，检索表各条鉴定路线的长度则差异较大。

叶节点数/阶元数比率：它是检索表分叉程度的指示值，叶节点数比待鉴定阶元数越多，该比值越大，检索表树型分叉越严重。

使用特征数/特征总数比率：检索表涵盖的特征不是鉴定每个阶元都用到的，有的阶元鉴定时可能只用到其中一个，有的则涉及多个甚至全部特征。该比值能反映检索表鉴定单个阶元的复杂度。

确认性特征数：检索表单个分叉节点可包括多个特征：第一个为主特征，其他为确认性特征。主特征与确认性特征分组作用相同，当主特征失效时，可继续用确认性特征鉴定。

终端节点与非终端节点数：终端节点即叶节点，非终端节点指该节点尚未变成叶节点，但已没有特征能继续划分。

实际适用性：非数量性特征，它需要收集广大用户的直观使用感受与反馈后加以评价分析才能得出。

5 检索表的优劣评价标准

检索表虽然结构较简单，但由于选用特征的不同，鉴定同一组对象可能有多个不同的检索表，即便使用相同的鉴别特征，因为特征的使用排序不同，也会导致出现各式各样的检索表。尽管这些检索表在分类功能设计上可能都是正确的，但它们实际上仍有优劣好差之分。那什么样的检索表才是设计好的呢？Metcalf (1954) 曾提出过一些检索表应具备的特点，是最早的雏形标准；后来 Dallwitz 等从分类学角度提出了目前较权威和普遍认同的评价标准 (Calvo-Flores, 2006)：

- i. 能把分类单元划分为相等的组群，这意味每次划分，组中的元素相等，即属于均分

鉴于检索表已设计好的特征鉴定路线，用户可选择任何可用的特征开始鉴定，不同人可能以不同的特征比对顺序完成同一对象的鉴定，Snow (1999) 认为它是继分子系统发育分析后系统分类研究中另一重要创新。不过它对检索表编制的要求较高，设计者必须清楚了解对象与每个特征的匹配关系。除此之外，多途径检索表的制作与使用都需要专门的计算机软件，手工难以完成。

4 检索表的参考特征

为了方便分析与比较检索表的优劣，Calvo-Flores 等 (2006) 曾引入了一组与检索表直接相关的特征：

平均长度、最大长度、最小长度：平均长度是检索表鉴定物种平均步数的指示值。由于专家经常寻找能以最少步数完成鉴定的检索表，平均长度便成为一个重要的参考。

平均长度的变异度：它是检索表平衡度的衡量值。若该值较大，检索表各条鉴定路线的长度则差异较大。

叶节点数/阶元数比率：它是检索表分叉程度的指示值，叶节点数比待鉴定阶元数越多，该比值越大，检索表树型分叉越严重。

使用特征数/特征总数比率：检索表涵盖的特征不是鉴定每个阶元都用到的，有的阶元鉴定时可能只用到其中一个，有的则涉及多个甚至全部特征。该比值能反映检索表鉴定单个阶元的复杂度。

确认性特征数：检索表单个分叉节点可包括多个特征：第一个为主特征，其他为确认性特征。主特征与确认性特征分组作用相同，当主特征失效时，可继续用确认性特征鉴定。

终端节点与非终端节点数：终端节点即叶节点，非终端节点指该节点尚未变成叶节点，但已没有特征能继续划分。

实际适用性：非数量性特征，它需要收集广大用户的直观使用感受与反馈后加以评价分析才能得出。

5 检索表的优劣评价标准

检索表虽然结构较简单，但由于选用特征的不同，鉴定同一组对象可能有多个不同的检索表，即便使用相同的鉴别特征，因为特征的使用排序不同，也会导致出现各式各样的检索表。尽管这些检索表在分类功能设计上可能都是正确的，但它们实际上仍有优劣好差之分。那什么样的检索表才是设计好的呢？Metcalf (1954) 曾提出过一些检索表应具备的特点，是最早的雏形标准；后来 Dallwitz 等从分类学角度提出了目前较权威和普遍认同的评价标准 (Calvo-Flores, 2006)：

- i. 能把分类单元划分为相等的组群，这意味每次划分，组中的元素相等，即属于均分

鉴于检索表已设计好的特征鉴定路线,用户可选择任何可用的特征开始鉴定,不同人可能以不同的特征比对顺序完成同一对象的鉴定, Snow (1999) 认为它是继分子系统发育分析后系统分类研究中另一重要创新。不过它对检索表编制的要求较高,设计者必须清楚了解对象与每个特征的匹配关系。除此之外,多途径检索表的制作与使用都需要专门的计算机软件,手工难以完成。

4 检索表的参考特征

为了方便分析与比较检索表的优劣, Calvo-Flores 等 (2006) 曾引入了一组与检索表直接相关的特征:

平均长度、最大长度、最小长度: 平均长度是检索表鉴定物种平均步数的指示值。由于专家经常寻找能以最少步数完成鉴定的检索表, 平均长度便成为一个重要的参考。

平均长度的变异度: 它是检索表平衡度的衡量值。若该值较大, 检索表各条鉴定路线的长度则差异较大。

叶节点数/阶元数比率: 它是检索表分叉程度的指示值, 叶节点数比待鉴定阶元数越多, 该比值越大, 检索表树型分叉越严重。

使用特征数/特征总数比率: 检索表涵盖的特征不是鉴定每个阶元都用到的, 有的阶元鉴定时可能只用到其中一个, 有的则涉及多个甚至全部特征。该比值能反映检索表鉴定单个阶元的复杂度。

确认性特征数: 检索表单个分叉节点可包括多个特征: 第一个为主特征, 其他为确认性特征。主特征与确认性特征分组作用相同, 当主特征失效时, 可继续用确认性特征鉴定。

终端节点与非终端节点数: 终端节点即叶节点, 非终端节点指该节点尚未变成叶节点, 但已没有特征能继续划分。

实际适用性: 非数量性特征, 它需要收集广大用户的直观使用感受与反馈后加以评价分析才能得出。

5 检索表的优劣评价标准

检索表虽然结构较简单, 但由于选用特征的不同, 鉴定同一组对象可能有多个不同的检索表, 即便使用相同的鉴别特征, 因为特征的使用排序不同, 也会导致出现各式各样的检索表。尽管这些检索表在分类功能设计上可能都是正确的, 但它们实际上仍有优劣好差之分。那什么样的检索表才是设计好的呢? Metcalf (1954) 曾提出过一些检索表应具备的特点, 是最早的雏形标准; 后来 Dallwitz 等从分类学角度提出了目前较权威和普遍认同的评价标准 (Calvo-Flores, 2006):

- i. 能把分类单元划分为相等的组群, 这意味每次划分, 组中的元素相等, 即属于均分

- ii. 划分标准选中的特征应稳定可靠, 种间多样性小
- iii. 能先鉴定出那些接触频率高或丰度高的物种, 或者说经常鉴定到的物种能最早从检索表中脱颖而出

Dallwitz 这一标准对检索表特征选用以及使用次序上提出了明确的目标, 这也正是检索表设计最核心的部分。一般来说, 好的检索表应选用最明显的外部特征, 而且要用绝对性状, 不要用“较大”、“较小”、“明些”、“暗些”等相对性状, 也不要重叠性状, 并用最简洁明确的文体表达出来以使用户了解(管致和, 1999)。这是对检索表质量最基本的要求。不过, 随着人们对生物多样性认识与保护意识增强, 接触并使用检索表的人越来越多, 检索表除了鉴别特征的选择、分类群系的划分次序、鉴别特征的语言表达等检索表本身因素外, 其使用方式与传播形式也逐渐成为影响检索表优劣的重要因素。20 世纪初基于计算机编程技术实现的问答式检索表(Goodall, 1968; 胡奇和马吉祥, 1990), 突破了检索表根据特征序号跳跃鉴定的传统使用方式, 方便了特征的调用, 这种人机互动的使用形式仍保留至今; 后来产生的图文分支式或链接式检索表(金瑞华等, 1996; Schäfer et al., 2000; 王心丽等, 2006), 给表中提到难以理解的鉴别特征配上形象直观的图片, 大大降低了检索表对于一般非专家用户的使用难度。近几年日益发达的网络化信息传播又推动形成了 PolyClave (University of Toronto Department of Botany, 1996)、X:ID (Marine Biological Laboratory, 2004)、ActKey (Brach and Song, 2005)、Phoenix Key (张小斌等, 2006a; CBIT, 2007b)、NaviKey (University of Bayreuth, Department of Mycology, 2007)、Lucid3 (CBIT, 2007c) 等网络检索表, 使足不出户便能快速访问并调用检索表工具变成了现实, 检索表资源的共享利用迈入了一个崭新阶段。

6 检索表开发工具

检索表开发工具是专门的计算机建表软件, 分成二项式和多途径两种类型。国外这类工具研究始于 20 世纪 70 年代, 目前较好的几款软件已进入商业化发展轨道(Dallwitz, 2007b), 本论文开发的 KeyMaker 即属于此类的中文工具。

6.1 DELTA 系统

DELTA 全称 Description Language for Taxonomy, 即分类描述语言, 它包含了一系列分类信息编码的规定, 以便于计算机加工处理。目前 DELTA 已被国际分类数据库工作组(TDWG)定为分类数据交互的标准(李健钧, 1996)。基于 DELTA 格式的分类信息可还原为自然语言描述, 生成一般及交互式检索表, 用于进化系统研究和检索系统构建。其中 DELTA Key 组件是建表工具, 可根据输入的 DELTA 格式编码的分类信息和配置文件, 自动计算选择合适特征再输出二项及多项式或退格式检索表; 而 DELTA IntKey 是同样基于 DELTA 编码文件的交互式检索表使用工具。

- ii. 划分标准选中的特征应稳定可靠, 种间多样性小
- iii. 能先鉴定出那些接触频率高或丰度高的物种, 或者说经常鉴定到的物种能最早从检索表中脱颖而出

Dallwitz 这一标准对检索表特征选用以及使用次序上提出了明确的目标, 这也正是检索表设计最核心的部分。一般来说, 好的检索表应选用最明显的外部特征, 而且要用绝对性状, 不要用“较大”、“较小”、“明些”、“暗些”等相对性状, 也不要重叠性状, 并用最简洁明确的文体表达出来以使用户了解(管致和, 1999)。这是对检索表质量最基本的要求。不过, 随着人们对生物多样性认识与保护意识增强, 接触并使用检索表的人越来越多, 检索表除了鉴别特征的选择、分类群系的划分次序、鉴别特征的语言表达等检索表本身因素外, 其使用方式与传播形式也逐渐成为影响检索表优劣的重要因素。20 世纪初基于计算机编程技术实现的问答式检索表(Goodall, 1968; 胡奇和马吉祥, 1990), 突破了检索表根据特征序号跳跃鉴定的传统使用方式, 方便了特征的调用, 这种人机互动的使用形式仍保留至今; 后来产生的图文分支式或链接式检索表(金瑞华等, 1996; Schäfer et al., 2000; 王心丽等, 2006), 给表中提到难以理解的鉴别特征配上形象直观的图片, 大大降低了检索表对于一般非专家用户的使用难度。近几年日益发达的网络化信息传播又推动形成了 PolyClave (University of Toronto Department of Botany, 1996)、X:ID (Marine Biological Laboratory, 2004)、ActKey (Brach and Song, 2005)、Phoenix Key (张小斌等, 2006a; CBIT, 2007b)、NaviKey (University of Bayreuth, Department of Mycology, 2007)、Lucid3 (CBIT, 2007c) 等网络检索表, 使足不出户便能快速访问并调用检索表工具变成了现实, 检索表资源的共享利用迈入了一个崭新阶段。

6 检索表开发工具

检索表开发工具是专门的计算机建表软件, 分成二项式和多途径两种类型。国外这类工具研究始于 20 世纪 70 年代, 目前较好的几款软件已进入商业化发展轨道(Dallwitz, 2007b), 本论文开发的 KeyMaker 即属于此类的中文工具。

6.1 DELTA 系统

DELTA 全称 Description Language for Taxonomy, 即分类描述语言, 它包含了一系列分类信息编码的规定, 以便于计算机加工处理。目前 DELTA 已被国际分类数据库工作组(TDWG)定为分类数据交互的标准(李健钧, 1996)。基于 DELTA 格式的分类信息可还原为自然语言描述, 生成一般及交互式检索表, 用于进化系统研究和检索系统构建。其中 DELTA Key 组件是建表工具, 可根据输入的 DELTA 格式编码的分类信息和配置文件, 自动计算选择合适特征再输出二项及多项式或退格式检索表; 而 DELTA IntKey 是同样基于 DELTA 编码文件的交互式检索表使用工具。

6.2 Lucid Professional

Lucid Professional是由澳大利亚昆士兰大学有害生物信息技术与推广中心(CPITT)精心研制开发的多途径检索表编制与使用工具,它支持特征与分类单元多种容错水平的匹配方式,并设计了“唯一性状”、“最佳特征”、“专家导航”等新颖的检索辅助功能,整体内容全面,功能强大,是迄今较为成熟和优秀的检索表开发工具(Snow, 1999; 孙冠英等, 2002; Shayler and Siver, 2006)。

6.3 Lucid Phoenix

Lucid Phoenix(CBIT, 2007b)是 Lucid 家族针对传统检索表设计的另一产品,包括 Builder、Importer、Player3 个组件,可编制新的一般通用检索表,导入现成扫描后的书面检索表,并给检索表整合超文本、图片等多媒体信息,通过跨平台 Java 技术直接嵌入浏览器在网上传播和交互使用(张小斌等, 2006a)。

6.4 PANKey

PANKey (Exeter, 2007)是一个基于 DELTA 分类标准的专业性生物诊断鉴定程序包,内含支持构建交互式检索表的 KCONI (Pankhurst, 1988)和一般通用检索表的 KEY3M3、支持在线鉴定的 ONLIN7、支序分析数据转换的 DELPAUP 以及特征比较、相似度分析等其他工具。检索表构建采用了 PANKey 算法,由于开发较早,需在 DOS 系统中使用(Jensen, 1990)。

6.5 XKey

XKey(Calvo-Flores et al., 2006)是一个基于决策树方法建树的分类检索表开发软件。它采用 XML 标准文档记录通用分类描述信息,通过引入机器学习的人工智能技术,使用户可在自动、半自动和交互模式下自由选择 4 种分类规则(熵 Entropy、信息增益率 The Gain Ratio、基尼多样性指数 Gini Diversity Index 和 Dallwitz 规则)分别建树,并根据输出检索表的评价参数选择最适合的检索表。XKey 是目前技术较新的检索表构建工具,软件界面为西班牙语。

7 检索系统开发工具

分类鉴定系统开发工具一般基于检索表鉴定模式设计,并以检索表作为知识库由计算机引导用户鉴定。由于检索系统使用普遍,此类工具国内外开发和应用均较成熟。

7.1 昆虫分类辅助鉴定多媒体专家系统通用平台 TaxoKeys

TaxoKeys 是根据昆虫分类学的特点,将昆虫分类的二项式检索表用数据库表示成系统知识库,利用计算机数据结构中二叉树结构的分枝结点搜索技术来实现其推理过程,进行昆虫分类的辅助鉴定(高灵旺等, 2003)。昆虫分类学家只要将建立的各种昆虫类群分类检索表

6.2 Lucid Professional

Lucid Professional是由澳大利亚昆士兰大学有害生物信息技术与推广中心(CPITT)精心研制开发的多途径检索表编制与使用工具,它支持特征与分类单元多种容错水平的匹配方式,并设计了“唯一性状”、“最佳特征”、“专家导航”等新颖的检索辅助功能,整体内容全面,功能强大,是迄今较为成熟和优秀的检索表开发工具(Snow, 1999; 孙冠英等, 2002; Shayler and Siver, 2006)。

6.3 Lucid Phoenix

Lucid Phoenix(CBIT, 2007b)是 Lucid 家族针对传统检索表设计的另一产品,包括 Builder、Importer、Player3 个组件,可编制新的一般通用检索表,导入现成扫描后的书面检索表,并给检索表整合超文本、图片等多媒体信息,通过跨平台 Java 技术直接嵌入浏览器在网上传播和交互使用(张小斌等, 2006a)。

6.4 PANKey

PANKey(Exeter, 2007)是一个基于 DELTA 分类标准的专业性生物诊断鉴定程序包,内含支持构建交互式检索表的 KCONI(Pankhurst, 1988)和一般通用检索表的 KEY3M3、支持在线鉴定的 ONLIN7、支序分析数据转换的 DELPAUP 以及特征比较、相似度分析等其他工具。检索表构建采用了 PANKey 算法,由于开发较早,需在 DOS 系统中使用(Jensen, 1990)。

6.5 XKey

XKey(Calvo-Flores et al., 2006)是一个基于决策树方法建树的分类检索表开发软件。它采用 XML 标准文档记录通用分类描述信息,通过引入机器学习的人工智能技术,使用户可在自动、半自动和交互模式下自由选择 4 种分类规则(熵 Entropy、信息增益率 The Gain Ratio、基尼多样性指数 Gini Diversity Index 和 Dallwitz 规则)分别建树,并根据输出检索表的评价参数选择最适合的检索表。XKey 是目前技术较新的检索表构建工具,软件界面为西班牙语。

7 检索系统开发工具

分类鉴定系统开发工具一般基于检索表鉴定模式设计,并以检索表作为知识库由计算机引导用户鉴定。由于检索系统使用普遍,此类工具国内外开发和应用均较成熟。

7.1 昆虫分类辅助鉴定多媒体专家系统通用平台 TaxoKeys

TaxoKeys 是根据昆虫分类学的特点,将昆虫分类的二项式检索表用数据库表示成系统知识库,利用计算机数据结构中二叉树结构的分枝结点搜索技术来实现其推理过程,进行昆虫分类的辅助鉴定(高灵旺等, 2003)。昆虫分类学家只要将建立的各种昆虫类群分类检索表

内容装入知识库,便可得到一个辅助鉴定的专家系统工具。目前 TaxoKeys 正在研究建立与 DELTA 系统之间的接口,以实现将 DELTA 建立的检索表导入 TaxoKeys,直接用于构建分类鉴定专家系统供一般用户使用。

7.2 二叉树型知识开发系统

与 TaxoKeys 相比,它是一个生物分类鉴定网络系统的开发工具。它由知识编辑器负责收集书本上或者专家头脑中的知识上传到 Web 服务器;由知识推理组件调用知识库进行推理,再将推理结果以网页的形式返回给用户(丘耘, 2006)。整个系统最大特色在于实行了可视化的知识编辑功能,即分类专家直接构造一棵形象化的二叉分类树来设计二叉检索路线,系统会自动将其转化为知识库。

7.3 LinnaeusII

LinnaeusII (Estep et al., 1989; ETI, 2007) 是 ETI 生物信息技术公司一款代表性产品,在多样性数据库构建与物种鉴定方面做得较完美。它由 4 大功能模块组成:分类数据库模块记录上种下分类单元的文字与多媒体信息,这是多样性数据库的基本内容;支持数据库模块记录相关的研究学者、参考文献、技术术语说明与索引等其他信息;鉴定模块提供了文本二项检索表、图片检索表和多途径检索表 IdentifyIt;地理信息模块以网格定位记录物种的分布信息,以实现基于网格的物种丰度查询与分布比较 (Schalk and Oosterbroek, 1996)。此外,该系统更新较快,是物种多样性研究的有利助手。

表 2.2 DELTA 等工具鉴定功能对比
Table 2.2 Comparison of identification with tools like DELTA et al.

工具名称	鉴定模式		分类信息扩增			在线 鉴定
	传统二项	多途径	特征	分类单元	GIS	
DELTA	✓	✓	✓	✓		
LinnaeusII	✓	✓	✓	✓	✓	✓
Lucid Professional		✓	✓	✓		✓
Lucid Phoenix	✓		✓	✓		✓
PANKey	✓	✓				✓
TaxoKeys	✓		✓			
XKey	✓					
二叉树型知识开发系统	✓		✓			✓

第三章 分类规则研究

人们在认识事物的过程中,发现它们具有各种各样的性状,其中有些是相同的,有些是不同的,分类即以这些共同性与特殊性作为事物分门别类的依据。然而,事物之间存在大量的异同性,分类不可能把它们都作为划分事物的依据,而是找出其中隐含的规则再据此进行判断。在动植物分类研究中,检索表便是这一规则的代名词。研究如何分类的问题,不仅是生物分类学家所面对的,也是人工智能中机器学习的研究范畴。机器学习也称为归纳推理,是通过学习训练数据集,发现模型的参数,并找出数据中隐含的规则,其中关联分析法、人工神经网络、决策树和遗传算法在数据挖掘中应用很广泛(朱建秋, 2007)。由于检索表是物种鉴定的决策工具,决策树研究对于检索表编制有着许多重要的参考意义。

1 决策树简介

决策树(Decision Tree)是一种树型结构的预测模型(图 3.1),其中树的非终端节点表示属性,叶节点表示所属的不同类别。根据训练数据集中数据的不同取值建立树的分支,形成决策树。决策树的主要功能是藉由分类已知的事例来建立一树状结构,并从中归纳出事例里的某些规律;而产生出来的决策树,也能利用来做样本外的预测(Tchen, 2007)。它与神经网络最大不同在于,决策制定的过程是可见的,可以解释结果是如何产生的。决策树一般产生直观、易理解的规则,而且分类不需太多计算时间,适于对记录分类或结果的预测。

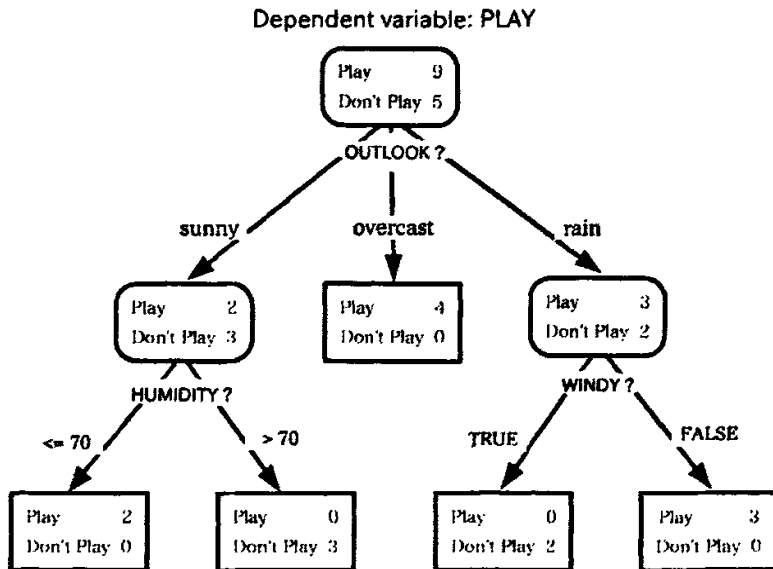


图 3.1 决策树范例(Quinlan, 1986)

Fig.3.1 Example of a decision tree (Quinlan, 1986)

第三章 分类规则研究

人们在认识事物的过程中,发现它们具有各种各样的性状,其中有些是相同的,有些是不同的,分类即以这些共同性与特殊性作为事物分门别类的依据。然而,事物之间存在大量的异同性,分类不可能把它们都作为划分事物的依据,而是找出其中隐含的规则再据此进行判断。在动植物分类研究中,检索表便是这一规则的代名词。研究如何分类的问题,不仅是生物分类学家所面对的,也是人工智能中机器学习的研究范畴。机器学习也称为归纳推理,是通过学习训练数据集,发现模型的参数,并找出数据中隐含的规则,其中关联分析法、人工神经网络、决策树和遗传算法在数据挖掘中应用很广泛(朱建秋, 2007)。由于检索表是物种鉴定的决策工具,决策树研究对于检索表编制有着许多重要的参考意义。

1 决策树简介

决策树(Decision Tree)是一种树型结构的预测模型(图 3.1),其中树的非终端节点表示属性,叶节点表示所属的不同类别。根据训练数据集中数据的不同取值建立树的分支,形成决策树。决策树的主要功能是藉由分类已知的事例来建立一树状结构,并从中归纳出事例里的某些规律;而产生出来的决策树,也能利用来做样本外的预测(Tchen, 2007)。它与神经网络最大不同在于,决策制定的过程是可见的,可以解释结果是如何产生的。决策树一般产生直观、易理解的规则,而且分类不需太多计算时间,适于对记录分类或结果的预测。

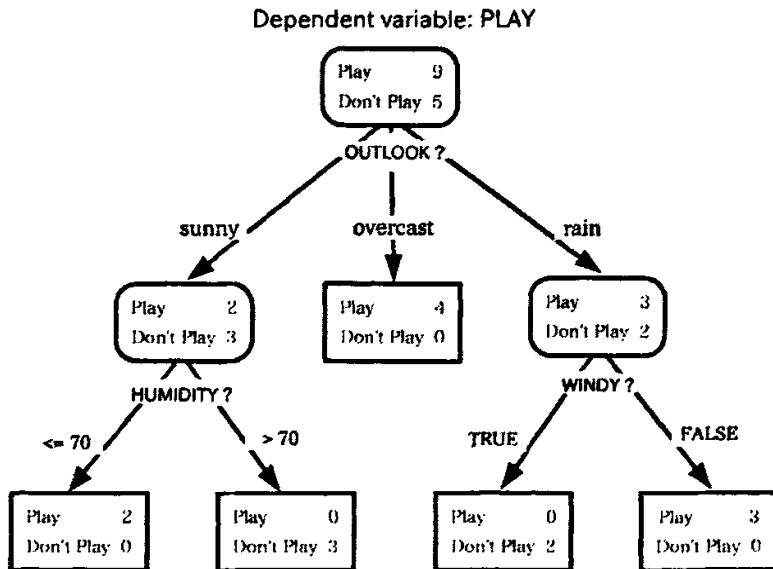


图 3.1 决策树范例(Quinlan, 1986)

Fig.3.1 Example of a decision tree (Quinlan, 1986)

决策树的产生是各个节点不断分支延伸的过程，由于在每一个节点上只用到一个自变量（即决策因子）来进行分支，因此决策树的每一次延伸都要选择依据哪一个自变量来分支以及如何决定分支的值。决策树的算法围绕此展开，它要求每次节点分支评估都要选择最佳的问题，以使分支后的每个节点，就所要预测目标而言，同构性或纯度越高越好，差异度越低越好。目前在决策树构建方法中，评估各种分支方式的常用标准算法有 3 种：信息熵、信息增益和基尼指数。

2 决策树的经典分类规则

2.1 信息熵

熵（Entropy）源至于信息论，1949 年由 Shannon C 与 Weaver W 提出，是关于不确定性的数学度量，用来表示某系统内所含有的信息量（Ke, 2007），定义如下：

若一份文件中用到 n 种字，且每一种字在该文件中的出现频率为 p_i ，则该文件的信息量或熵为：

$$-\sum_{i=1}^n p_i \times \log_2(p_i)$$

在决策树中，熵用来定义一个节点的纯度，纯度高者包含较少信息，熵也比较小。例 3.1：

30% 雌虫 ① 70% 雄虫 共 100 头

$$\begin{aligned} \text{Entropy} &= (-0.3 \times \log_2(0.3)) + (-0.7 \times \log_2(0.7)) \\ &= (0.3 \times 1.74) + (0.7 \times 0.514) \\ &= 0.881 \end{aligned}$$

50% 雌虫 ② 50% 雄虫 共 100 头

$$\begin{aligned} \text{Entropy} &= (-0.5 \times \log_2(0.5)) + (-0.5 \times \log_2(0.5)) \\ &= -\log_2(0.5) \\ &= 1 \end{aligned}$$

可见，第①批虫子的纯度要高于第②批。在选择节点分支时，则需根据加权熵(weighted entropy)来判断。加权熵是某节点分支后各节点的熵与其权重乘积的总和，例 3.2：

决策树的产生是各个节点不断分支延伸的过程，由于在每一个节点上只用到一个自变量（即决策因子）来进行分支，因此决策树的每一次延伸都要选择依据哪一个自变量来分支以及如何决定分支的值。决策树的算法围绕此展开，它要求每次节点分支评估都要选择最佳的问题，以使分支后的每个节点，就所要预测目标而言，同构性或纯度越高越好，差异度越低越好。目前在决策树构建方法中，评估各种分支方式的常用标准算法有 3 种：信息熵、信息增益和基尼指数。

2 决策树的经典分类规则

2.1 信息熵

熵（Entropy）源至于信息论，1949 年由 Shannon C 与 Weaver W 提出，是关于不确定性的数学度量，用来表示某系统内所含有信息量（Ke, 2007），定义如下：

若一份文件中用到 n 种字，且每一种字在该文件中的出现频率为 p_i ，则该文件的信息量或熵为：

$$-\sum_{i=1}^n p_i \times \log_2(p_i)$$

在决策树中，熵用来定义一个节点的纯度，纯度高者包含较少信息，熵也比较小。例 3.1：

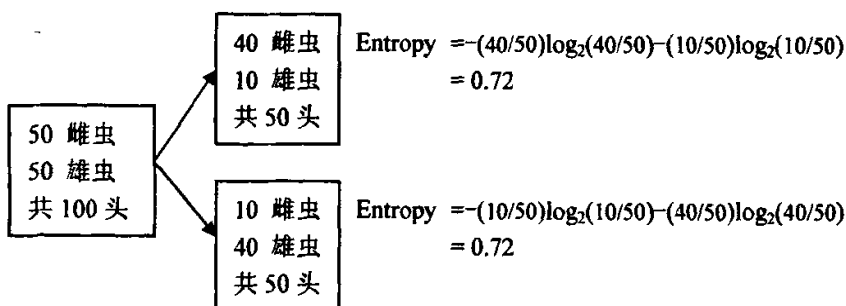
30% 雌虫
① 70% 雄虫
共 100 头

$$\begin{aligned} \text{Entropy} &= (-0.3 * \log_2(0.3)) + (-0.7 * \log_2(0.7)) \\ &= (0.3 * 1.74) + (0.7 * 0.514) \\ &= 0.881 \end{aligned}$$

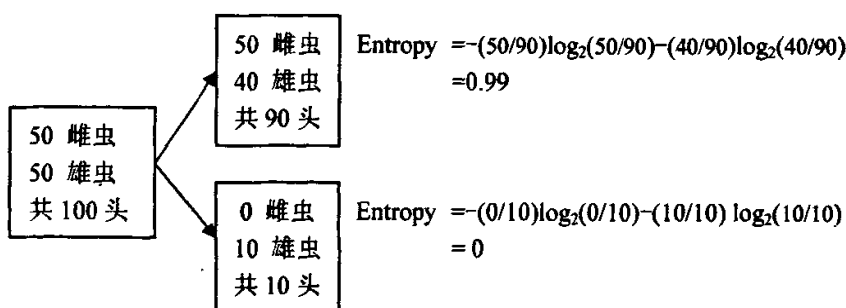
50% 雌虫
② 50% 雄虫
共 100 头

$$\begin{aligned} \text{Entropy} &= (-0.5 * \log_2(0.5)) + (-0.5 * \log_2(0.5)) \\ &= -\log_2(0.5) \\ &= 1 \end{aligned}$$

可见，第①批虫子的纯度要高于第②批。在选择节点分支时，则需根据加权熵(weighted entropy)来判断。加权熵是某节点分支后各节点的熵与其权重乘积的总和，例 3.2：



① Weighted Entropy $= (50/100)*0.72 + (50/100)*0.72 = 0.72$



② Weighted Entropy $= (90/100)*0.99 + (10/100)*0 = 0.89$

第①种节点分支的加权熵要低于第②种，因此决策树将选择按①分支方式往下延伸。

2.2 信息增益

信息增益 (Information Gain) 是由 Quinlan J R 1970 年代末期提出，并在 ID3^{*} (杨明和张载鸿，2002；孙细明和张晓鹏，2005) 开发工具中使用，它也是基于熵的一种分类算法，目的使对一个对象分类所需的期望测试数目达到最小，确保得到一棵简单的树 (Han and Kamber, 2001)。其定义为分支前节点的熵值与分支后各子节点加权熵总和平均之差，表示为：

$$G(A_j) = E(S) - E(A_i)$$

$$E(A_i) = \sum_{j=1}^n \left(\frac{m_j}{m} * E(C_j) \right)$$

其中 A_i = 属性 i

$E(A_i)$ = 以属性 i 分支后各子节点加权熵的总和平均

n = 属性 i 的可用值数

$E(C_i)$ = 属性值 i 对应子节点的熵值

m = 父节点笔数

$E(S)$ = 分支前节点熵值

m_j = 属性值 i 对应子节点笔数

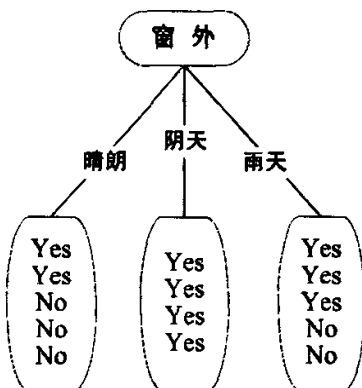
由于熵值越小，纯度越高，因此信息增益越大，分支效果越好。例 3.3：根据表 3.1 的训练集

数据开发决策树，以根据天气情况判定是否出去玩？

表 3.1 训练集数据(Quinlan, 1986)
Table 3.1 Data of training set (Quinlan, 1986)

不同天	窗外	温度	湿度	风力	出去玩
D1	晴朗	热	高	弱	NO
D2	晴朗	热	高	强	NO
D3	阴天	热	高	弱	YES
D4	雨天	温和	高	弱	YES
D5	雨天	凉	正常	弱	YES
D6	雨天	凉	正常	强	NO
D7	阴天	凉	正常	强	YES
D8	晴朗	温和	高	弱	NO
D9	晴朗	凉	正常	弱	YES
D10	雨天	温和	正常	弱	YES
D11	晴朗	温和	正常	强	YES
D12	阴天	温和	高	弱	YES
D13	阴天	热	正常	强	YES
D14	雨天	温和	高	强	NO

以“窗外”分支节点情况如下



计算用“窗外”分支节点时的信息增益：

$$\begin{aligned}
 E(S) &= \sum_{i=1}^n -p_i \log_2 p_i \\
 &= -p_{yes} \log_2 p_{yes} - p_{no} \log_2 p_{no} \\
 &= -\left(\frac{9}{14}\right) \log_2 \left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2 \left(\frac{5}{14}\right) \\
 &= 0.94
 \end{aligned}$$

$$E(A_i) = \sum_{j=1}^n \left(\frac{m_j}{m} * E(C_j) \right)$$

$$= (5/14)E(\text{晴朗}) + (4/14)E(\text{阴天}) + (5/14)E(\text{雨天})$$

$$= 0.694$$

故 Gain(窗外) = 0.94 - 0.694 = 0.246

同理 Gain(湿度) = 0.94 - 0.789 = 0.151

Gain(风力) = 0.94 - 0.892 = 0.048

Gain(温度) = 0.94 - 0.911 = 0.029

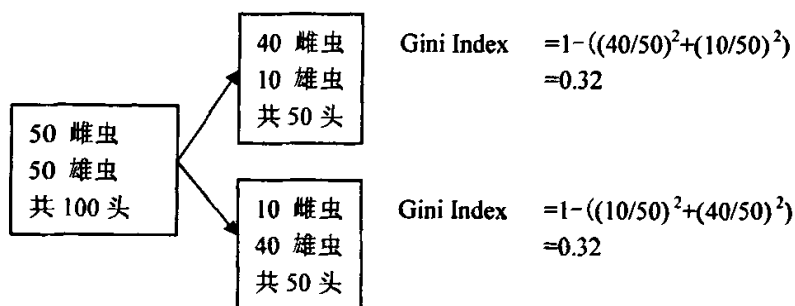
因此选择信息增益最大的“窗外”作为决策树分支的根节点。由于信息增益注重分支后子节点的纯度多于子节点的笔数，这种分支标准常常会导致产生许多分支后有很多很小子节点的决策树(Kononenko et al., 1984)，使决策树凌乱复杂。这一缺点在 Quinlan J R (1986) 后来提出的信息增益率 Gain Ratio(=Information Gain/分支的熵值)得到了解决。

2.3 基尼指数

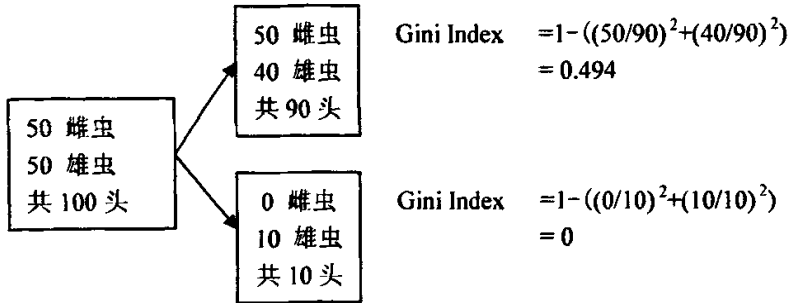
基尼指数(Gini Index)是一种不纯度分裂方法，由 Breiman 等人 (1984) 年提出，用来表示从同一个群落中随机取两个物种，它们属于同一物种的机率。在决策树中则表示某节点包含 n 种预测值，且每一种预测值在该节点中的出现频率为 p_i ，则该节点的基尼指数为

$$1 - \sum_{i=1}^n p_i^2$$

当每一种预测值在该节点中的出现频率都一样时，基尼指数最大，节点纯度最低；当整个节点只含有一种预测值时，基尼指数最小，节点纯度最高。在评价节点分支好坏时，则根据基尼度(Gini metric 或 Gini split)来比较。基尼度与前面的加权熵类似，是分支后各子节点的基尼指数与其权重乘积的总和(陈云樱等，2004)。例 3.4:



$$\textcircled{1} \text{ Gini Metric} = 50/100 * (0.32) + 50/100 * (0.32) = 0.32$$



② Gini Metric $= 90/100 * (0.494) + 10/100 * (0) = 0.445$

基尼度小的分支可产生纯度较高的节点，故决策树将选择按①分支方式往下延伸，此结果与例 3.2 用加权熵判断结果一致。

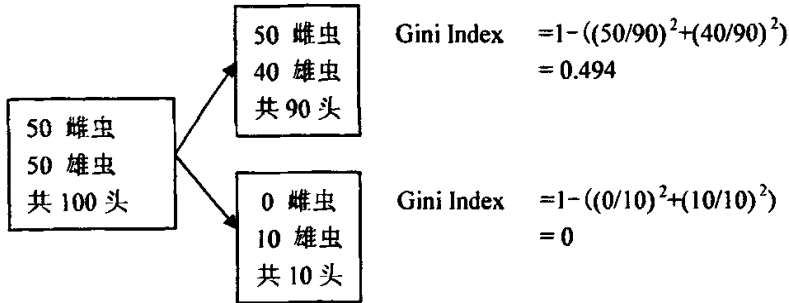
从以上决策树延伸方式的评估标准来看，尽管它们算法不同，但计算结果都一致倾向选择纯度较高的节点分支，这与事物分门别类后整体趋向一致（熵变小，纯度提高）是吻合的。检索表也可依照决策树的构建思路，根据已知类别的对象集合，运用数据挖掘技术从中发现分类规则，并萃取规则形成检索表（Calvo-Flores et al., 2006）。但这种做法需要改进，一方面由于进行决策树的规则归纳需要较多的训练数据才有代表性，另一方面检索表对象在数据集中都是唯一的，它们出现的机率均等，使用熵方法分支节点的纯度只与节点笔数相关，与节点组成成分无关，导致检索表倾向于是一个分支均衡的决策树模型。而从检索表优劣评价标准来看，构建时还应注重特征和对象这两个影响因素。或许可在评估标准中加入特征与对象的影响权重，以符合分类检索表设计的原则。

3 分类学中的分类规则

在分类学领域中，随着第二代电子计算机的发展，Möller(1962)、Hall(1970)、Pankhurst(1970a; Pankhurst and Walters, 1971)、Morse(1971)、Dallwitz(1974)、Payne and Preece (1980)、Payne and Thompson(1989)、Reynolds et al. (2003) 等都研究或介绍过检索表设计的分类算法。其中 Pankhurst 提出的 PANKey 算法曾在 Nature 上发表，知名度较大，现已开发为专业的商业软件 PANKEY(1971; 1991)，需付费使用；Dallwitz 开发的 DELTA Key 算法由于绑定 DELTA 分类描述系统免费提供，使用较广，是最经典的分类算法。

3.1 PANKey 算法

PANKey 算法是由 Pankhurst R J 1969 年提出，于 1970 在 Computer Journal 与 Nature 杂志上发表报道。Pankhurst(1970a, 1970b)认为检索表是从一个描述待鉴对象属性的矩阵产生的，尽管这个矩阵可以生成多个不同的检索表，但它们的适用性是不同的。Pankhurst key 程



② Gini Metric $= 90/100 * (0.494) + 10/100 * (0) = 0.445$

基尼度小的分支可产生纯度较高的节点，故决策树将选择按①分支方式往下延伸，此结果与例 3.2 用加权熵判断结果一致。

从以上决策树延伸方式的评估标准来看，尽管它们算法不同，但计算结果都一致倾向选择纯度较高的节点分支，这与事物分门别类后整体趋向一致（熵变小，纯度提高）是吻合的。检索表也可依照决策树的构建思路，根据已知类别的对象集合，运用数据挖掘技术从中发现分类规则，并萃取规则形成检索表（Calvo-Flores et al., 2006）。但这种做法需要改进，一方面由于进行决策树的规则归纳需要较多的训练数据才有代表性，另一方面检索表对象在数据集中都是唯一的，它们出现的机率均等，使用熵方法分支节点的纯度只与节点笔数相关，与节点组成成分无关，导致检索表倾向于是一个分支均衡的决策树模型。而从检索表优劣评价标准来看，构建时还应注重特征和对象这两个影响因素。或许可在评估标准中加入特征与对象的影响权重，以符合分类检索表设计的原则。

3 分类学中的分类规则

在分类学领域中，随着第二代电子计算机的发展，Möller(1962)、Hall(1970)、Pankhurst(1970a; Pankhurst and Walters, 1971)、Morse(1971)、Dallwitz(1974)、Payne and Preece (1980)、Payne and Thompson(1989)、Reynolds et al. (2003) 等都研究或介绍过检索表设计的分类算法。其中 Pankhurst 提出的 PANKey 算法曾在 Nature 上发表，知名度较大，现已开发为专业的商业软件 PANKEY(1971; 1991)，需付费使用；Dallwitz 开发的 DELTA Key 算法由于绑定 DELTA 分类描述系统免费提供，使用较广，是最经典的分类算法。

3.1 PANKey 算法

PANKey 算法是由 Pankhurst R J 1969 年提出，于 1970 在 Computer Journal 与 Nature 杂志上发表报道。Pankhurst(1970a, 1970b)认为检索表是从一个描述待鉴对象属性的矩阵产生的，尽管这个矩阵可以生成多个不同的检索表，但它们的适用性是不同的。Pankhurst key 程

序通过“寻树”方式探究所有可能的检索表形式，再从中选择一个最佳的检索表方案。这一步智能化的判断与选择要通过一个“比较函数F”实现。

$$F=F_1+F_2$$

$$F_1=(K-2)^2$$

$$F_2 = \sum_{i=1}^k |1 - \frac{n_i}{N} K|$$

N: 分类单元总数; K: 小组数; n_i : i 小组分类单元总数

PANKey 算法选择 F 值最小的检索表形式，当 $K=2$, $N=2$, $n_1=n_2=N/2$ 时, F 取最小值 0, 检索表达达到最理想状态。此后, F 值随着分组数 K 增加而变大, 同时若分组不均衡, 也会导致 F 值变大。如图 3.2, 阿拉伯数字为阶元, 罗马数字为鉴定需回答特征问题数, PANKey 算法选择 B 为输出结果。从树型结构来看, A 节点分布不均衡, 导致树型结构纵向发展, 层次增多, 完成所有分类单元鉴定需回答 9 个问题; B 节点分布平衡, 并比 A 少一层次, 完成所有分类单元鉴定需回答 8 个问题; 而随着节点数增加, 两者差异将更显著。因此, PANKey 算法会优先选择分组数少, 分组均衡的检索表, 这样使鉴定每个分类单元需回答的特征问题总数最少。

此外, PANKey 算法在枚举检索表时优先使用较高权重的特征, 冗余的特征 (对于相关的分类单元, 这些特征相同或者缺失) 则不再考虑。

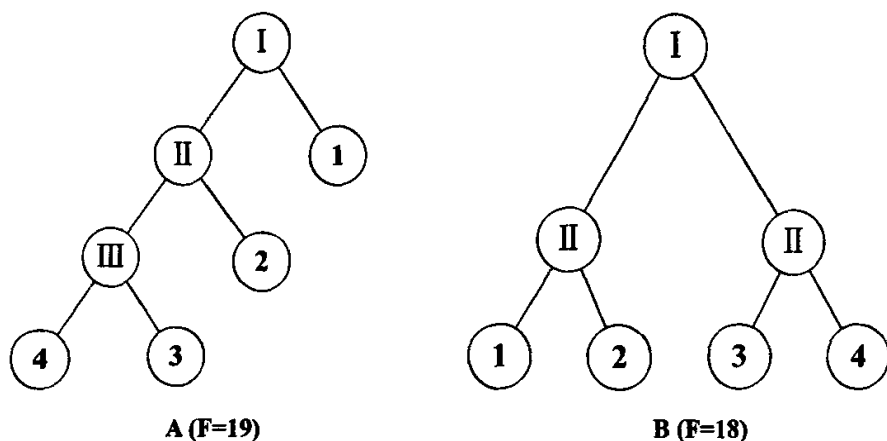


图 3.2 两种分支不同的检索表
Fig.3.2 Two keys with different branches

3.2 DELTA Key 算法

该算法是由 Dallwitz M J 1974 年为编写 DELTA 系统的建表组件 Key 而设计, 是分类学中较经典的分类标准。Dallwitz (2006a) 认为每个特征都有一定的使用成本, 这个成本可理解两层含义: 1.使用该特征鉴定产生错误结果的可能性或风险; 2.使用该特征鉴定的劳力 (如

观察、比对、判定的难易程度), 且特征成本具有加和性, 两个特征的使用成本是每个特征单独使用成本的总和。特征成本的评估由一个“比较函数 K (Dallwitz et al., 2006a)”来完成, 评估值最小的为最佳特征。因此, DELTA Key 创建检索表时先选择一个最佳特征, 把最初分类单元分成二个或以上的小组, 然后每个小组用剩余的最佳特征继续划分, 如此反复, 直到每个小组只包含一个分类单元或者没有再找到合适的特征。

$$K = c + c_{\min} [(\sum_{j=1}^s f_j \log_2 n_j) / (\sum_{j=1}^s f_j) + V]$$

$$V = (\frac{1-v}{v}) (\frac{n+8}{n \log_2 n}) (\sum_{j=1}^s n_j - n)$$

C : 特征成本 $C = Rbase^{5-r}$ (r : 特征可信度[0,10] $Rbase$: 基数[1,5])

C_{\min} : 考虑特征中的最小成本特征 S : 特征值数量

f_j : 第 j 个小组中分类单元频度总和

$f = Abase^{a-5}$ (a : 分类单元丰度[0,10] $Abase$: 基数[1,5])

n_j : 第 j 个小组中分类单元数量 n : 分类单元总数

V : 特征种间变异度或控制因子 v : 变异权重[0,1]

从比较函数看出, Dallwitz 引入了特征可信度、分类单元丰度、特征种间变异度等概念, 并认为高可信度、低种间变异度的特征应优先使用, 高丰度的分类单元应优先被鉴定出来。这些参数的共同作用, 使 DELTA Key 的建表思想更符合了分类检索表的设计标准和使用特点。有关 DELTA Key 建表效果和使用优缺点评价在第 5 章有专门介绍, 这里不再重复。

尽管国内每年都有大量新物种见诸报道, 但很少听闻分类学家使用 PANKEY、DELTA Key 等工具开发新检索表, 徐柱等 (1992) 曾讨论计算机产生中英文植物分类检索表, 但他采用的加权分类群计算公式直接取自 PANKey 的比较函数, 并非提出新的检索表分类算法。检索表的科学设计与编制在国内没有得到必要的重视, 仍是生物分类信息技术领域中的一个研究空白。

第四章 研究的目的意义和技术路线

1 研究的目的和意义

基于以上综述，为了适应信息资源数字化发展的时代潮流，实现生物分类信息的数字化开发与网络化共享利用，推动传统分类鉴定的数字化建设，论文综合运用计算机高级编程、数据库、网络平台和数字多媒体等技术，通过开发一款集检索表数字化编码、检索表智能设计与重构再生、生物分类检索系统自动生成和物种多样性数据库构建等多功能一体化的知识系统开发工具，为检索表等生物分类信息的制作、整理、发布、使用与推广提供一套完整的电子化解决方案。该工具的完成，将有助于提高检索表设计效率、科学性和利用率，降低生物分类检索系统开发的技术门槛，促进物种多样性信息的共享研究，推进国内生物分类研究与物种鉴定的电子化进程。

2 研究的技术路线

根据知识系统开发一般的设计流程，结合生物分类知识构架（图 4.1）的特点，本研究采用了如下技术路线（图 4.2）：

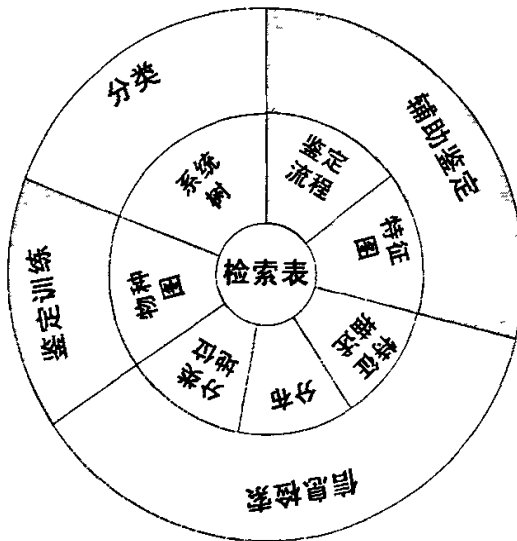


图 4.1 生物分类知识圆周结构图

Fig.4.1 Circumferential structure for biological taxonomic knowledge

第四章 研究的目的意义和技术路线

1 研究的目的和意义

基于以上综述，为了适应信息资源数字化发展的时代潮流，实现生物分类信息的数字化开发与网络化共享利用，推动传统分类鉴定的数字化建设，论文综合运用计算机高级编程、数据库、网络平台和数字多媒体等技术，通过开发一款集检索表数字化编码、检索表智能设计与重构再生、生物分类检索系统自动生成和物种多样性数据库构建等多功能一体化的知识系统开发工具，为检索表等生物分类信息的制作、整理、发布、使用与推广提供一套完整的电子化解决方案。该工具的完成，将有助于提高检索表设计效率、科学性和利用率，降低生物分类检索系统开发的技术门槛，促进物种多样性信息的共享研究，推进国内生物分类研究与物种鉴定的电子化进程。

2 研究的技术路线

根据知识系统开发一般的设计流程，结合生物分类知识构架（图 4.1）的特点，本研究采用了如下技术路线（图 4.2）：

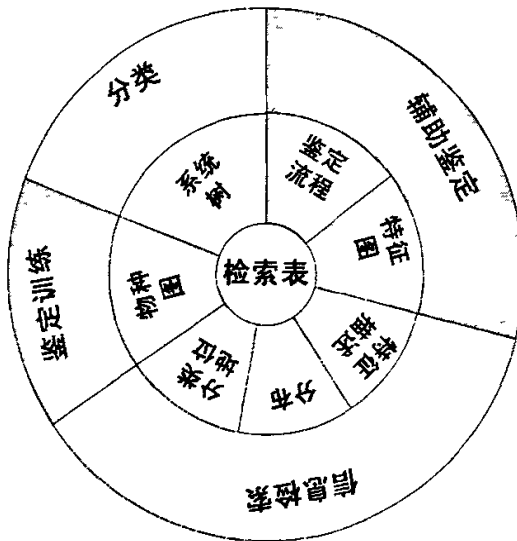


图 4.1 生物分类知识圆周结构图

Fig.4.1 Circumferential structure for biological taxonomic knowledge

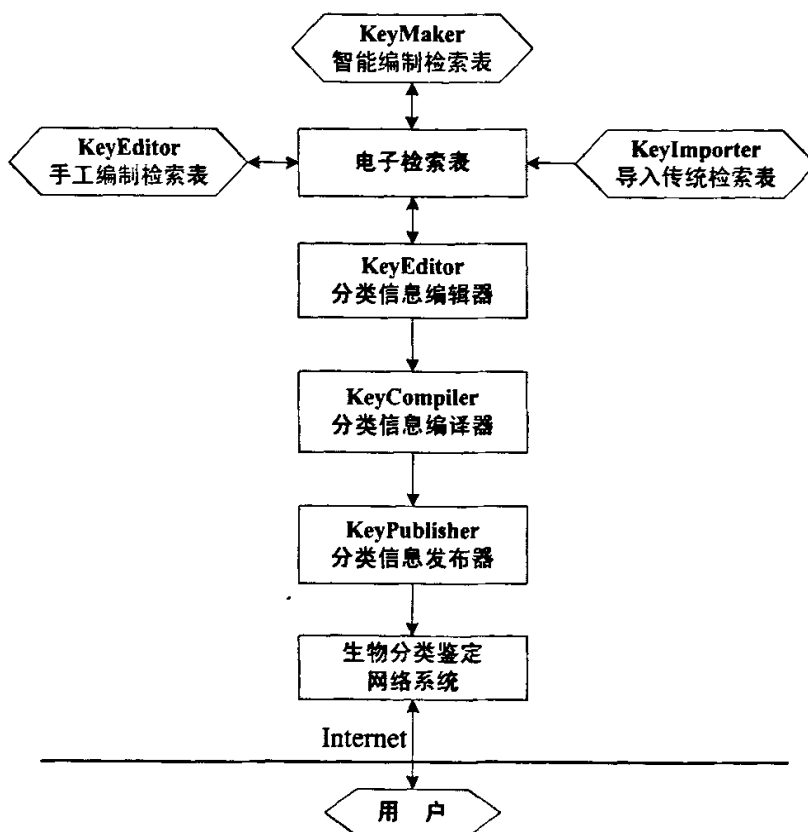


图 4.2 论文的总体设计和技术路线

Fig.4.2 Overall design and technical line of thesis

3 研究的创新点

3.1 检索表数字化编码方案设计完善，扩展性强

本研究采用数字编码矩阵保留检索表中对象与特征的匹配关系，并用 XML 结构化数据模板完整记录数字矩阵、对象与特征信息。该方案结构设计新颖，知识分离到位，信息可读性强，编辑修改方便，与一般的检索表数字化方式相比，更适合计算机分析处理，深入挖掘潜在知识，扩增多媒体信息，转换为其他使用灵活、鉴定方便的检索表形式。另外，还可直接用于检索表二次重构的开发。

3.2 检索表智能编制算法设计合理，优化水平高

该算法以检索表设计基本原则为出发点，根据专家建立的特征与对象关联矩阵，通过运

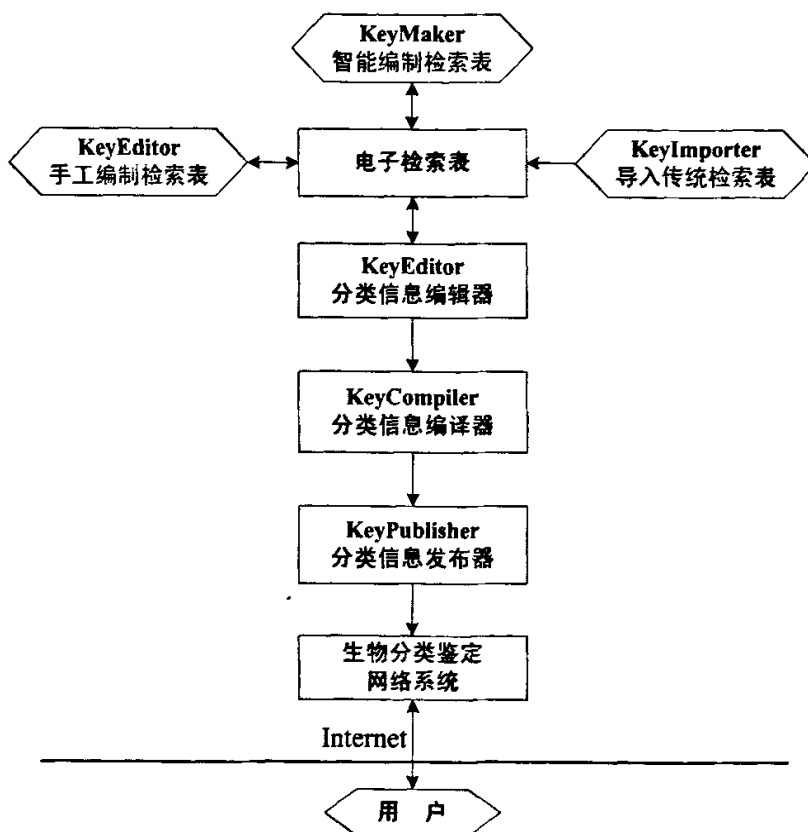


图 4.2 论文的总体设计和技术路线

Fig.4.2 Overall design and technical line of thesis

3 研究的创新点

3.1 检索表数字化编码方案设计完善，扩展性强

本研究采用数字编码矩阵保留检索表中对象与特征的匹配关系，并用 XML 结构化数据模板完整记录数字矩阵、对象与特征信息。该方案结构设计新颖，知识分离到位，信息可读性强，编辑修改方便，与一般的检索表数字化方式相比，更适合计算机分析处理，深入挖掘潜在知识，扩增多媒体信息，转换为其他使用灵活、鉴定方便的检索表形式。另外，还可直接用于检索表二次重构的开发。

3.2 检索表智能编制算法设计合理，优化水平高

该算法以检索表设计基本原则为出发点，根据专家建立的特征与对象关联矩阵，通过运

算以尽量逼近“特征优先序”和“对象优先序”为目标输出优化检索表。经比较验证,该算法科学可靠,优化设计达到期望水平,从而能代替分类专家智能设计出最佳的检索策略,减少手工编写检索表所花费的时间与精力,并提高检索表设计效率与合理性。以该算法为核心编写的检索表编辑大师 KeyMaker 是一款专业的二项检索表制作与智能优化工具,优化能力达到并高于同类 DELTA Key 软件。

3.3 检索表二次重构技术灵活性强,建表快速

二次重构是一种新颖的检索表自动构建技术,用户可不用求助分类专家,根据需要自行定制重构对象,便很快得到只含指定分类单元和必要特征的二次检索表,大大简化了依据传统分类检索表鉴定的复杂繁琐过程,在保证同等正确性的前提下提高鉴定工作效率。由于重构的基础是传统基础分类检索表,它也是一种高效灵活运用专家知识库的技术。而基于网络提供的二次重构服务,使更多人可以轻松地享受到该技术带来的好处。

3.4 网络检索表获取方便,辅助鉴定能力强

检索表在网上发布与使用,用户不必再四处奔走查询分类文献,即可在线访问使用这些电子检索表。而人性化设计的检索方式,如二叉跳转、动态交互、Phoenix 检索,使用户既方便快速地鉴定物种,又能随即查阅各类扩充信息,及时确认鉴定结果。这不仅能促进检索表资源的网络共享与交流,也有利于检索表快捷更新。现中国昆虫鉴定分类系统 InsectX 收录各类检索表共 984 个,已达到昆虫鉴定到科与茧蜂鉴定到种的要求。

3.5 生物分类检索系统开发便捷,功能强大

基于本工具开发各类生物群落的分类检索系统,用户只需在 KeyEditor 中完成分类资料的收集与整理,其他如网页制作、数据库集成和网站发布等技术性强、难度较高的工作都由 KeyCompiler 和 KeyPublisher 自动快速完成。生成的系统内建系统分类树、辅助鉴定、物种鉴定训练、信息检索查询、分类专题定制等功能,完全能满足各种分类研究与学习的需要。今后整个系统的维护与更新也十分方便。

3.6 物种多样性数据库构建方便,内容丰富

利用本工具开发的物种多样性数据库,将包含分类地位、系统关系、鉴别特征、地理分布、标本信息、物种图片等丰富内容。这些信息在 KeyEditor 中收集齐全后经 KeyCompiler 编译成结构化关系型的数据库和图片库,作为可移植的基础知识库,用于第三方分类软件或系统的开发,今后只需重新编译即可更新数据库。目前 InsectX 多样性数据库已收录昆虫分类阶元 1 873 个(含茧蜂 569 种),检索表 984 个,物种图 2 343 张,特征图 284 张,是国内分类系统最完善、内容最全面的生物资源数据库之一。

第二部分

生物分类鉴定知识系统开发

第五章 电子检索表的制作

随着计算机工具广泛被生物分类专家所熟悉使用,检索表的编制已逐渐从纸质书面向计算机无纸化过渡,但用计算机文本编辑器(如记事本、Office Word 等)代替纸张编写出来的检索表并不是真正意义上“电子检索表”,因为这样的检索表只有人能看懂,计算机仍无法理解其作用,更不用说让计算机辅助鉴定。所谓“电子检索表”应指分类专家头脑中或者现有传统检索表所蕴含的鉴定思路和分类数据经规则化处理后转换成计算机可判断识别并处理的电子信息。由于规则化处理方式不同,电子检索表的记录形式可能多种多样,但它们表现出的基本鉴定功能是不会变的。论文将以目前最通用的二项检索表为例介绍本研究设计的电子检索表数字化记录方式、计算机手工与智能编制电子检索表的原理与方法。

1 二项检索表的数字化技术

计算机拥有强大的运算推理能力,如果让它代替检索表引导用户鉴定,这样用户可以省力不少。但要实现这样的辅助鉴定,其先决条件是计算机已完全掌握了检索表中的分类信息和鉴定路线。

1.1. 检索表数字化方式

检索表数字化格式是检索表数字开发与应用的基础,国外由于生物分类信息技术研究起步较早,已出现了 Delta (Dallwitz, 1980)、Lucid Interchange Format (LIF) (TDWG-SDD, 2003; CBIT, 2007d)、NEXUS (Maddison et al., 1997)、XDELTA (Dodds L, 1999; Dallwitz, 2005) 等多个分类数据描述格式,它们设计各有千秋,应用各有侧重,对于分类信息自然描述、基于传统与交互检索表鉴定和支序进化研究都十分有利。国内在计算机应用开始盛行之初,分类工作者已尝试将检索表转换为计算机检索系统(蒋齐, 1991),把检索表中包含的鉴定推理变成计算机能获取利用的知识库,让计算机能根据用户的选择判断,引导用户完成鉴定。这种使用方式目前仍较流行,其知识库的组建方法普遍采用了“形式模拟”的数字化记录方式(陈乃中和沈佐锐, 2003;高灵旺等, 2003),即以检索表二叉结构为原型设计,将检索表各组成部分以“字段”的形式分别记录在数据库中(表 5.1),一般-1 表示已到达二叉树某一支的终止节点。

以这种结构记录在数据库中,数据的可读性很强,也很容易根据字段的逻辑关系重现二项检索表的最初模样。计算机编程也很容易实现二叉式分类推理,根据知识库中保存的分类信息,指导用户完成物种鉴定。

第五章 电子检索表的制作

随着计算机工具广泛被生物分类专家所熟悉使用,检索表的编制已逐渐从纸质书面向计算机无纸化过渡,但用计算机文本编辑器(如记事本、Office Word 等)代替纸张编写出来的检索表并不是真正意义上“电子检索表”,因为这样的检索表只有人能看懂,计算机仍无法理解其作用,更不用说让计算机辅助鉴定。所谓“电子检索表”应指分类专家头脑中或者现有传统检索表所蕴含的鉴定思路和分类数据经规则化处理后转换成计算机可判断识别并处理的电子信息。由于规则化处理方式不同,电子检索表的记录形式可能多种多样,但它们表现出的基本鉴定功能是不会变的。论文将以目前最通用的二项检索表为例介绍本研究设计的电子检索表数字化记录方式、计算机手工与智能编制电子检索表的原理与方法。

1 二项检索表的数字化技术

计算机拥有强大的运算推理能力,如果让它代替检索表引导用户鉴定,这样用户可以省力不少。但要实现这样的辅助鉴定,其先决条件是计算机已完全掌握了检索表中的分类信息和鉴定路线。

1.1. 检索表数字化方式

检索表数字化格式是检索表数字开发与应用的基础,国外由于生物分类信息技术研究起步较早,已出现了 Delta (Dallwitz, 1980)、Lucid Interchange Format (LIF) (TDWG-SDD, 2003; CBIT, 2007d)、NEXUS (Maddison et al., 1997)、XDELTA (Dodds L, 1999; Dallwitz, 2005) 等多个分类数据描述格式,它们设计各有千秋,应用各有侧重,对于分类信息自然描述、基于传统与交互检索表鉴定和支序进化研究都十分有利。国内在计算机应用开始盛行之初,分类工作者已尝试将检索表转换为计算机检索系统(蒋齐, 1991),把检索表中包含的鉴定推理变成计算机能获取利用的知识库,让计算机能根据用户的选择判断,引导用户完成鉴定。这种使用方式目前仍较流行,其知识库的组建方法普遍采用了“形式模拟”的数字化记录方式(陈乃中和沈佐锐, 2003;高灵旺等, 2003),即以检索表二叉结构为原型设计,将检索表各组成部分以“字段”的形式分别记录在数据库中(表 5.1),一般-1 表示已到达二叉树某一支的终止节点。

以这种结构记录在数据库中,数据的可读性很强,也很容易根据字段的逻辑关系重现二项检索表的最初模样。计算机编程也很容易实现二叉式分类推理,根据知识库中保存的分类信息,指导用户完成物种鉴定。

表 5.1 二项检索表数据库存储一般结构 (高灵旺等, 2003)

Table 5.1 General structure for storing dichotomous keys in database(Gao et al., 2003)

特征序号	特征描述	对应的下一级特征序号
1	特征 1	3
...
3	特征 3	7
...
7	分类单元 1	-1
...

然而这种数据存储方式,对于检索表所蕴含信息的深入挖掘以及功能扩展,并不十分有利。例如,根据表 2.1 检索表与表 5.3 分类单元与特征编号回答如下问题:

1. 分类单元 3 在鉴定过程中共使用了哪些特征?
2. 哪些分类单元在鉴定过程中使用了特征 3 进行判断?
3. 分类单元 3 与分类单元 7 最关键的特征区别是什么?

虽然计算机根据表 5.1 的检索表信息数字化存储结构进行推理也能获得答案,但从计算机数据挖掘与逻辑推理角度来分析,要回答问题 1:必须先定位分类单元 3 终止结点所在的特征序号,然后根据特征序号以及对应的下一级特征序号以反向推理的形式逐个回溯并记录所用过的特征,当到达特征序号 1 时,才找到分类单元 3 鉴定过程中所有用过的特征;而以此方法回答了所有检索表对象的问题 1 后,才能回答问题 2;问题 3 其实是寻找分类单元 3、7 之间的特征分歧点,只要比较一下这 2 个分类单元的特征使用次序,便能发现它们的特征分歧点。这种比较对于人来说轻而易举,但计算机需要一番较复杂的分析推算。从计算机程序设计角度看来,表 5.1 数据结构对于后期检索表扩展分析显得笨拙且效率低下。尽管有时可以针对解决问题 1、2、3 的需要,在数据库设计和数据填充时考虑简化后期计算机的分析推理过程,但这种“形式模拟”的数据结构对于信息数字化原则和论文研究目的来说,仍有许多问题存在。因此,有必要寻找一种既符合检索表二叉树结构,又能适宜计算机深入分析检索表信息的数字化记录的高效结构。

1.2. 特征分值数字编码

从检索表本质来看,它是由对象、特征及两者的匹配关系组成的,任何一个检索系统知识库也主要由这三者组成。其中匹配关系是核心,记录了对象所具备的特征、特征所对应的对象,等于这两者相互的映射矩阵。对匹配关系的数字化处理好差,也就直接决定了检索表数字化方式的优劣。表 5.1 所示的检索表数字化方式,直观记录检索表鉴定的推理路线,是

一种基于模拟专家思路或者规则的计算机表示方式。Dallwitz (1992; 2007a) 认为在分类鉴定系统中基于矩阵的表达方式要优于基于规则的方式。计算机擅长分析与处理有规则的编码信息, 没有对模拟过程进行有规律的数字转换 (即信息编码), 将导致计算机难以深入分析, 带来更多帮助。有鉴于此, 论文以检索表最重要的匹配关系为突破点, 抓住其映射矩阵的数学模型, 设计了一种新颖的特征分值数字编码方法(表 5.2)。

表 5.2 二项特征分值编码表
Table 5.2 Coding rule for binary-value character scores

鉴别特征	二项特征值	分类单元	特征分值
A	1	√	1
	2		
B	1		2
	2	√	
C	1	Null	0
	2	Null	

以此法对匹配关系进行编码处理后, 检索表的每个分类单元都获得一个由 0、1、2 组成的 N (N 代表鉴别特征总数) 长度的特征值。所有特征值并列一起, 便形成一个特征矩阵, 矩阵中每个非 0 的位点, 即反应了对象与特征的一次匹配关系。其实, 采用数字编码的思想已有各种先例, 如 DELTA 和 XDELTA 采用特征与特征值联位编码, NEXUS 同本论文采用特征位与特征值关联编码, LIF 采用特征值全组编码 (Dallwitz, 2005), 与它们相比, 特征分值编码虽不能支持多重特征匹配, 但对于处理二项检索表信息已绰绰有余, 相对于它们复杂的编码规范更易于理解应用。表 2.1 经特征分值编码后得到了表 5.3 的数学矩阵。

表 5.3 林奈 7 目昆虫特征分值表
Table 5.3 Character scores of seven insect orders established by Linnaeus

分类单元	特征编号					
	1	2	3	4	5	6
1. 无翅目	1	0	0	0	0	0
2. 有吻目	2	1	0	0	0	0
3. 双翅目	2	2	1	0	0	0
4. 鞘翅目	2	2	2	1	0	0
5. 鳞翅目	2	2	2	2	2	0
6. 膜翅目	2	2	2	2	1	1
7. 脉翅目	2	2	2	2	1	2

根据表 5.3, 我们再来回答前面提出的 3 个检索表问题:

1. 分类单元 3 在鉴定过程中共使用了哪些特征?

根据“双翅目”水平方向上非 0 特征分值所对应垂直方向上的特征即可得出为: 特征 1、2、3。

2. 哪些分类单元在鉴定过程中使用了特征 3 进行判断?

根据“特征编号 3”垂直方向上非 0 特征分值所对应水平方向上的分类单元即可得出为: 分类单元 3、4、5、6、7。

3. 分类单元 3 与分类单元 7 最关键的特征区别是什么?

通过比较两者完整的特征值可以发现, 在特征 3 之前, 它们的特征分值都是一样的, 在特征 3 时出现了取值的分歧, 因此答案是特征 3。

通过比较回答这 3 个问题的难度可以发现, 特征分值数字编码方法既简单又快速。对于计算机而言, 数字编码信息更容易接受处理, 因此它可以胜任分析更大更复杂的匹配矩阵。论文采用特征分值法编码检索表匹配信息给电子检索表修改、智能编制、网上使用和二次重构等都奠定了十分重要的开发基础。

1.3. 检索表数字化记录模板

特征分值数字编码方法已成功解决了检索表匹配关系的数字化问题, 而编码后产生的数字矩阵以及检索表对象、特征等文字信息, 论文选择了 XML (W3CHINA, 2007) 这一新型的文本数据库作为存储载体。作为可扩展标记语言, XML 与 HTML 有些相似, 都使用标记(文字由‘<’和‘>’括起)描述文档结构。XML 采用模块化的结构, 即文件由多个分离区块组成, 每个区块容纳不同的信息, 很容易被计算机识别与处理 (Maddison et al., 1997)。由于区块可自定义熟悉的名称来命名划分, XML 文件可读性也很好。基于文本的格式使它可以利用标准的文本编辑工具(如记事本、WORD)来读取和编辑。因此, 它已被普遍认同适用于描述结构化与半结构化的数据, 如电子表格、程序配置文件、数据库中所包含的信息(微软中国, 2007)。此外, XML 还有一个重要的优势, 它已作为信息交换的通用语言, 在不同编程平台、操作系统、开发人员之间的数据流通中发挥着十分关键的作用。正是看中了 XML 作为分类知识载体的各种优点, TDWG SDD (2003)1998 年以它为基础继 DELTA 后又推出了 SDD 子标准, 以适应信息时代对生物分类知识交流更高效更快捷的要求; Dodds (1999) 开发了 DELTA 标准的 XML 版本即 XDELTA, 极大地方便了 DELTA 信息的解析和扩展。为了简明起见, 论文根据二项检索表的基本结构设计了表 5.4 的检索表数字化记录 XML 模板。

表 2.1 经特征分值法编码后再和对象、特征等信息一起填入表 5.4 的 XML 模板, 便得到了林奈 7 目昆虫分类检索表完整的数字化记录文件(表 5.5)。最后, 保存该 XML 文件并修改文件扩展名为 .dam。至此, 一个完整的 DAM 电子检索表便诞生了。

表 5.4 检索表数字化记录 XML 模板
Table 5.4 XML-based key descriptive model

```
<key parent_key=value>
  <Characters>
    <Character1>
      <Node1 Value=Text image=text />
      <Node2 Value=Text image=text />
    </Character1>
    <Character2>
      <Node1 Value=Text image=text />
      <Node2 Value=Text image=text />
    </Character2>
  </Characters>
  <taxa>
    <taxon1 Score=numbers image=text />
    <taxon2 Score=numbers image=text />
  </taxa>
</key>
```

表 5.5 林奈 7 目昆虫检索表数字化记录文件
Table 5.5 Key descriptive document for seven insect orders established by Linnaeus

```
<key parent_key=无>
  <Characters>
    <翅>
      <Node1 Value=无翅/>
      <Node2 Value=有翅/>
    </翅>
    <口器>
      <Node1 Value=口器刺吸式/>
      <Node2 Value=口器咀嚼式/>
    </口器>
    <翅数>
      <Node1 Value=翅一对/>
      <Node2 Value=翅两对/>
    </翅数>
```

```

<前翅>
  <Node1 Value=前翅角质/>
  <Node2 Value=前翅膜质/>
</前翅>
<鳞片>
  <Node1 Value=翅不被鳞片/>
  <Node2 Value=翅被鳞片/>
</鳞片>
<蜚刺>
  <Node1 Value=雌腹部末端有蜚刺/>
  <Node2 Value=雌腹部末端无蜚刺/>
</蜚刺>
</Characters>
<taxa>
  <无翅目 Score=100000/>
  <有翅目 Score=210000/>
  <双翅目 Score=221000/>
  <鞘翅目 Score=222100/>
  <鳞翅目 Score=222220/>
  <膜翅目 Score=222211/>
  <脉翅目 Score=222212/>
</taxa>
</key>

```

1.4. 电子检索表二项形式还原方法

从表 5.5 可以看出, 电子检索表文件结构与原来形式截然不同。考虑到发表分类成果、出版分类著作等实际需求, 电子检索表仍需要能随时转换回到传统通用的二项表形式。其实, 论文在设计特征分值法时, 已经考虑了这种应用需求, 并预备了电子检索表二项形式的还原方法。从表 5.5 可发现, 特征的记录顺序与原检索表特征的排列顺序是相同的, 只要找到每条特征描述所指向的下一级特征序号或对象, 还原自然就可以实现。同时对照表 2.1 和表 5.3 还可以发现:

当对象在某特征完成鉴定时, 其后的特征分值均为 0。反之, 如果某对象特征分值在某特征位以后都是 0, 那该位所对应的特征描述应指向该对象。例如, “双翅目”特征值为 221000, 从第 3 位以后特征分值都为 0, 那特征 3 第 1 分值 (即翅一对) 在还原检索表中指向 “双翅目”。依此类推, 可以找出所有直接指向对象的特征描述。

如果某特征位后面是非 0，或者隔几个 0 后出现一次非 0，那它后面第一个非 0 特征位所对应的特征编号便是当前特征描述所对应的下一级特征序号。仍以“双翅目”为例，第 1 位特征分值是 2，其后紧跟着为 2 的特征分值，故可判断，特征 1 第 2 分值特征描述（即有翅）指向特征序号 2。同理，可找出所有特征描述所指向的下一级特征序号。

这一还原方法虽然比起表 5.1 的检索表数字化记录方法要复杂一点，但对于计算机而言，两者处理的速度和效果相差无几。Delphi 编程实现该还原过程如下：

```

Procedure export_dichotomouskey(txtfile:string) {
var temp:textfile;
begin
  assignfile(temp,txtfile);
  rewrite(temp);//新建输出文件

  //分析特征矩阵
  j:=dom.getElementsByTagName('characters').item[0].childNodes.length;
  setlength(tig,j,2); //节点数组
  with dom.getElementsByTagName('taxa').item[0] do
    for i:=0 to childNodes.length-1 do begin
      fengz:=childNodes[i].attributes[3].text;//读取特征值
      for h:=1 to j do//从第 1 个特征分值开始,寻找非 0 的特征分值
        for m:=1 to 2 do
          if (strtoint(copy(fengz,h,1))=m) and (tig[h-1,m-1]='') then begin
            for l:=h+1 to j do
              if strtoint(copy(fengz,l,1))>0 then begin
                tig[h-1,m-1]:=inttostr(l);//记录下一个特征序号
                break;
              end;
            if l>j then begin//如果后面全部是零，记录改鉴定到分类单元
              tig[h-1,m-1]:=childNodes[i].nodeName+childNodes[i].attributes[0].text;
              break;
            end;
          end;
        end;
      end;
    end;

  //输出二项检索表
  with dom.getElementsByTagName('characters').item[0] do
    for i:=0 to j-1 do begin

```

```
keyid:=inttostr(i+1);
pretig1:=keyid; //pretig 是特征序号
pretig2:='-';
writeln(temp,txt_res(pretig1+putmain(childNodes[i].childNodes[0].text,tig[i,0])));
writeln(temp,"");
writeln(temp,txt_res(pretig2+putmain(childNodes[i].childNodes[1].text,tig[i,1])));
writeln(temp,"");
end;
//写入保存文件
flush(temp);
closefile(temp);
showmessage('文本二项检索表导出完毕!');
end;
}
```

2 检索表智能编制技术

检索表是重要的分类分析工具，学习分类的人都要掌握检索表的制作。手工编制检索表的一般过程如图 5.1。

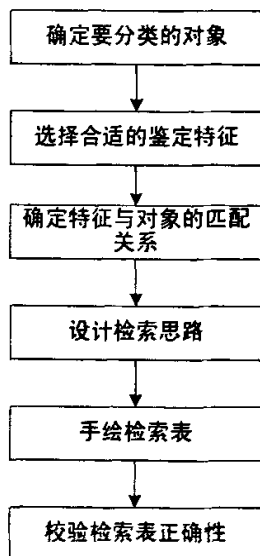


图 5.1 手工编制检索表的一般过程

Fig.5.1 General procedure in manual construction of keys

手工编制检索表的大部分时间与精力都花费在后三个环节，尤其是第四个环节，由于它

```
keyid:=inttostr(i+1);
pretig1:=keyid; //pretig 是特征序号
pretig2:='-';
writeln(temp,txt_res(pretig1+putmain(childNodes[i].childNodes[0].text,tig[i,0])));
writeln(temp,"");
writeln(temp,txt_res(pretig2+putmain(childNodes[i].childNodes[1].text,tig[i,1])));
writeln(temp,"");
end;
//写入保存文件
flush(temp);
closefile(temp);
showmessage('文本二项检索表导出完毕!');
end;
}
```

2 检索表智能编制技术

检索表是重要的分类分析工具，学习分类的人都要掌握检索表的制作。手工编制检索表的一般过程如图 5.1。

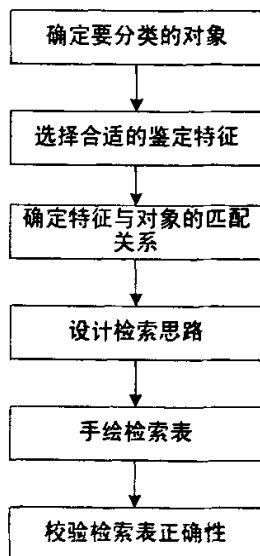


图 5.1 手工编制检索表的一般过程

Fig.5.1 General procedure in manual construction of keys

手工编制检索表的大部分时间与精力都花费在后三个环节，尤其是第四个环节，由于它

密切关系检索表成形后的鉴定效能,往往需要设计者反复思考与比较。而智能编制检索表的好处正在于它能取代后面三个环节的工作,减轻设计者的劳动负担,并通过内置的检索表优化技术,提高检索表设计的科学性与实用性。

计算机没有自主设计与创造的能力,但它能按照人为设定好的程式或规则,对输入信息进行运算、比较、判断等操作后输出结果。整个过程计算机处理速度极快,可以不断重复而“乐此不疲”,在很多层面上可以代替人脑的工作,这也是计算机被广泛用于科学研究的重要原因。而研究编制检索表的计算机工具,除了实现无纸化的检索表编制外,更重要的是发挥计算机的超级运算能力,达到一些人难以实现的期望目标,并提高检索表使用效能和设计效率。目前国内还没有此方面的专业工具,国外此类工具 PANKey(Pankhurst, 1970a)、DELTA Key(Dallwitz, 1974)等开发较早,存在设计算法不能达到检索表的最佳优化效果、二项检索表编制支持不好、中文兼容效果差、操作方式不符合分类专家使用习惯等弊端。因此,论文旨在根据二项检索表的设计特点,严格参考检索表的优劣评价标准制定优化目标,设计优化算法,以真正实现检索表的智能编制与使用效能优化,并适合分类专家使用。

2.1. 三个度的优化标准

检索表设计的最基本标准一能快速而正确地指导物种鉴定,论文在此将它具体分化为“三个度”的优化标准:

2.1.1. 特征优化度

特征优化度(Character Priority 简称 CP)定义为检索特征调用先后顺序与专家设计思路中的特征优先序一致性程度,其结果越接近 100%,检索表设计越佳。一般专家在设计检索表时会考虑,越重要的特征,应越优先使用,对于物种的划分结果则越可靠;重要性表现在特征性状是否稳定,是否多样性变异少(Calvo-Flores et al., 2006)(多样性变异指特征存在的可能性状,如触角有棒状、线状、具芒状等;翅有膜翅、鞘翅、鳞翅等),是否可方便观察。随着生长季节的变化,物种表现出现的特征也在变化,某些特征可能消失或变得不明显而难以观察,因而在设定特征优先序,即特征相对重要性排序时也应考虑这些因素。

2.1.2. 对象优化度

对象优化度(Taxon Priority 简称 TP)它与特征优化度作用相似,是对象鉴定完成的先后顺序与专家设计思路中的对象优先序一致性程度。一般来说,越高丰度的分类单元,越应优先能被鉴定(丰度是指某生态环境中,分类单元出现的频率,也可理解为常见性,物种丰度存在地区与时间区间上的差异)。越常见的物种,越容易被观察发现,人们遇到该物种的概率越高,应优先被断定或排除。另一方面,从分类单元与农业生产、经济生活等方面的利弊关系程度来考虑,也可以制定出一个对象优先序,用于指导检索表的优化编制。

2.1.3. 综合优化度

在检索表智能编制应用时发现, 由于受特征与对象匹配矩阵的制约, 不少检索表特征序达到最佳优化时, 对象序优化水平较低; 反之, 对象序达到最佳优化时, 特征序优化水平又较低。这种优化“一边倒”的现象, 使我们很难只根据特征优化度或者对象优化度来选择最好的优化效果。为了解决这一问题, 论文又继续引入另一优化标准—综合优化度 (Compositive Priority), 并定义为同时兼顾考虑前面两种效应后得出的复合结果。该值较高时, 其特征优化度和对象优化度水平均衡, 不会相差很大, 从而达到平衡优化的效果。正如人随着季节交替、天气冷暖变化着装一样, 检索表可以根据特征优先序与对象优先序的变化作出综合调整, 以获得当时当地环境下最适合使用的检索策略。

2.2. 三个度的优化算法

2.2.1. 特征优化度算法

特征优化度是衡量检索表设计好坏的重要指标之一, 只有当所有对象的检索特征调用次序符合专家事先设定的特征优先序时, 检索表的 CP 值才能达到 100%, 即完全优化。

特征优化度计算原理: 根据检索表中每个对象的特征鉴定路线, 计算单一对象的特征优化度 SCP, 随后将每个对象的 SCP 进行加和叠加, 获得 CP 值。

特征优化度计算公式(1)(2)(3):

$$CP = \frac{\sum_{n=1}^t SCP(V_n)}{t} * 100\% \quad (1)$$

V_n : 第 n 个对象完成鉴定所用的特征数; t : 检索表对象数

$$SCP(v) = \frac{\sum_{c=1}^{v-1} \sum_{h=c+1}^v P(c, h)}{\sum (v-1)} \quad (v > 1) \quad (2)$$

当 $v=1$ 时, $SCP=1$; P : 单个特征的相对优先度

$$P(c, h) = \begin{cases} 1 & (CR_c \leq CR_h) \\ 0 & (CR_c > CR_h) \end{cases} \quad (3)$$

CR_h : 第 h 个特征的优先级

例 5.1: 某检索表中有 3 个特征, 特征优先级分别为 $C_1=1$; $C_2=2$; $C_3=3$, 有 4 个对象, 对象优先级分别为 $T_1=1$; $T_2=2$; $T_3=T_4=3$, 每个对象鉴定时特征使用次序分别为 $S_1=C_1$; $S_2=C_1C_2$; $S_3=S_4=C_1C_2C_3$, 则

$$SCP(1)=1;$$

$$SCP(2)=\sum_{c=1}^1 \sum_{h=c+1}^2 P(c,h)=1;$$

$$SCP(3)=\frac{\sum_{c=1}^2 \sum_{h=c+1}^3 P(c,h)}{\sum 2}=1;$$

$$SCP(4)=SCP(3)=1;$$

$$CP = \frac{SCP(1)+SCP(2)+SCP(3)+SCP(4)}{4}=100\%$$

即该检索表的特征调用达到最佳优化水平。

2.2.2. 对象优化度算法

对象优化度也是衡量检索表设计好坏的重要指标。从检索表使用过程来看, 最高优先级的对象应该使用最少的特征判断, 便能完成鉴定。因此对象优化度应与对象完成鉴定所使用的特征数量相关。如果对象鉴定特征使用数在大小排序上与专家事先设定的对象优先序一致, 检索表的 TP 值可达到 100%, 即完全优化。

对象优化度计算原理: 根据检索表中每个对象的特征鉴定路线, 计算每个对象的对象优化度 STP, 随后将每个对象的 STP 进行加和叠加, 获得 TP 值。

对象优化度计算公式(4) (5) (6):

$$TP = \frac{\sum_{n=1}^t STP(n)}{t} * 100\% \quad (4)$$

n: 第 n 个对象; t: 检索表对象数

$$STP(c) = \frac{\sum_{h=c+1}^t P(c,h)}{t-c} \quad (c < t) \quad (5)$$

当 $c=t$ 时, $STP=1$; P: 单个对象的相对优先度

$$P(c, h) = \begin{cases} 1 & (TF_c \leq TF_h) \\ 0 & (TF_c > TF_h) \end{cases} \quad (6)$$

TF_h : 第 h 个对象完成鉴定所使用的特征数量

同上例 5.1 计算检索表的对象优化度:

$$TF_1=1; \quad TF_2=2; \quad TF_3=TF_4=3;$$

$$STP(1) = \frac{\sum_{h=2}^4 P(1, h)}{3} = 1;$$

$$STP(2) = \frac{\sum_{h=3}^4 P(2, h)}{2} = 1;$$

$$STP(3) = \sum_{h=4}^4 P(3, h) = 1;$$

$$STP(4) = STP(3) = 1;$$

$$TP = \frac{STP(1) + STP(2) + STP(3) + STP(4)}{4} = 100\%$$

即该检索表的对象鉴定也达到最佳优化水平。

2.2.3. 综合优化度算法

论文将上述检索表各个分类单元的两项检验指标进行几何加权作为权衡双向优化水平的综合指标 CTP。尽管“一边倒”的检索表不可能达到完全优化的水平，但通过考核综合优化度，仍可以找到较佳的平衡优化结果。

$$CTP = \frac{\sum_{n=1}^t (\sqrt{SCP(V_n) * STP(n)})}{t} * 100\%$$

显然上例 5.1, $CTP=100\%$, 这种情况是检索表设计中喜闻乐见的。

2.3. 检索表智能编程序

分类专家在设计检索表时, 不需要事先勾勒出检索表的鉴定流程或策略, 只需筛选出一些有效的鉴别特征, 与分类单元建立关联, 并设定检索表优化筛选时的参考标准 (特征优先

序与对象优先序), 由智能编制算法自动产生备选的检索表。备选的结果可能多种, 但分类专家可根据每个结果的优化程度、检索表长度、特征总优先级的高低排序, 选择最适用的检索表。默认情况下, 算法以优化程度最高的结果为最佳检索表。

2.3.1. 建立对象与特征匹配矩阵

检索表研究实体是特征、对象以及两者的匹配对应联系, 检索表研究的目的可认为是针对各种不同的使用环境和用户对象, 提供一种最符合实际需求、快而准的鉴定工具。为了实现计算机代替人脑设计检索表, 减轻工作难度, 提高工作效率, 目前大部分检索表计算机开发程序, 包括植物分类领域中知名的 DELTA Key 和 Lucid 多途径分类检索软件, 都不约而同地选用了“二维矩阵”的数学模型来实现特征与对象匹配关系的保存与再现。

由于检索表智能编制技术设计时特征取值一般都是三到多项, 不再局限于二项形式, 故原来的二项特征分值数字编码方法不能适用, 但将它进行分值扩展, 改造成多项特征分值便可解决问题。经过表 5.6 编码后, 每个对象仍可获得一个特征值, 下面 XML 文档中 score 值已记录了对象与特征的匹配关系, 其中大于 9 的特征分值转换成十六进制保存。

<Objects>

<甜瓜绢野螟 sname="" score="191121767F7" />

<小蔗螟 sname="" score="192112767B7" />

<橘蛀果点翅螟 sname="" score="181111567A7" />

<南美玉米苗斑螟 sname="" score="191111567F7" />

<柚叶螟 sname="" score="191112667F7" />

<西南玉米秆草螟 sname="" score="152112767F7" />

<南方玉米秆草螟 sname="" score="19211276797" />

<松异带蛾 sname="" score="392320773F6" />

<麦绢蛾 sname="" score="190321277F7" />

<美国白蛾 sname="" score="391321047F7" />

<苹果蠹蛾 sname="" score="28131277787" />

</Objects>

2.3.2. 对象逐层分组

检索表运作本质就是用不同的特征分组对象, 直到目标组中只有一个对象, 即完成鉴定。这种分组模式在二项检索表与多途径检索表中都存在, 唯一不同的是, 二项检索表是二分分组, 即一次一个特征把对象分成 2 组, 而多途径检索表由于特征取值是多项性的, 一次一个特征能把对象分成若干个组, 这意味着检索表每次的走向都是多重选择。检索表智能编制技术所要处理的分组模式既是二分组又是多途径的, 二项表现在它在每个分组水平上都期待用一个特征把对象分成 2 个组, 多途径则表现在同一分组水平可能有多个特征能将对象二分组。这种“二向性”的分组模式使检索表的构建因出现多层次多途径而变得十分复杂, 但它却为

优化技术提供了产生备选结果的物质基础。下面用“特征的对象值”与“逐层分组”加以说明：

表 5.6 多项特征分值转化表
Table 5.6 Converting rule for multi-value character scores

鉴别特征	特征值	分类单元	特征分值
A	1		2
	2	√	
		
B	n		n
	1		
	2		
C		0
	n	√	
	1	不需核对	
	2		
		
	n		

特征的对象值是指某一个特征所有对象的取值，如果在匹配矩阵中，以对象为横坐标，特征为纵坐标，那对象的特征值就是与 Y 轴平行的竖线，特征的对象值就是与 X 轴平行的横线。从特征的对象值可以清楚地判断哪些特征可以将哪些对象组进行二分组，它对于匹配矩阵的逐层分析十分有益。如表 5.7 特征与对象匹配矩阵所示，横向分值为特征的对象值，纵向分值为对象的特征值，在第 1 层水平上，只有特征 I 才能将对象(ABCDE)二分组，第 2 层水平亦同，而在第 3 层水平上，III 与 IV 特征都能将对象(ABC)二分组，从而扩展出 2 个二叉分支，并继续往下分组。按照这样的逐层分组模式下推，获得图 5.2 的分组节点树，从第一层沿着特征节点下寻，可以得到 2 条对象分组途径。但这只是个简单的逐层分组模型，在实际鉴定问题中，单层中很容易出现两到多个分叉节点，随着对象分组的不断深入，可能的分组路径会成倍数增加，分组节点树也将变得庞大而复杂。

表 5.7 特征与对象匹配矩阵
Table 5.7 Matrix value between character and object

特 征	对 象				
	A	B	C	D	E
I	1	1	1	1	2
II	3	3	3	2	1
III	2	4	4	1	3
IV	2	2	1	3	3

为了在计算机中保留并重现多组节点树，论文引入了“二叉树”数据结构，并将其扩展设计为“五叉单节点（图 5.3）”，以满足后续流程中数据记载与分析的需要。其中父节点、左右节点为主分叉，用于记录节点间的层次关系，左右对象节点用于记录分组后的对象成员。经验证，该结构能有效地为后期快速萃取备用检索表奠定基础。

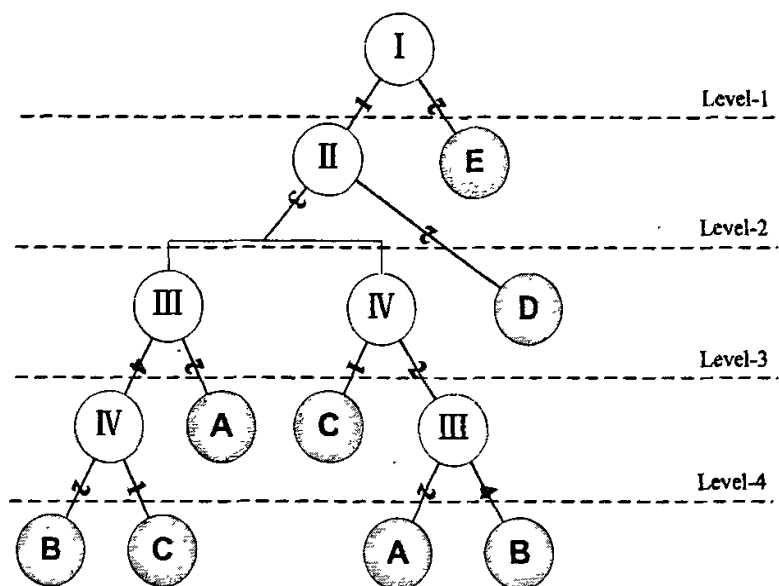


图 5.2 分组节点树

Fig.5.2 Tree-like grouping structure

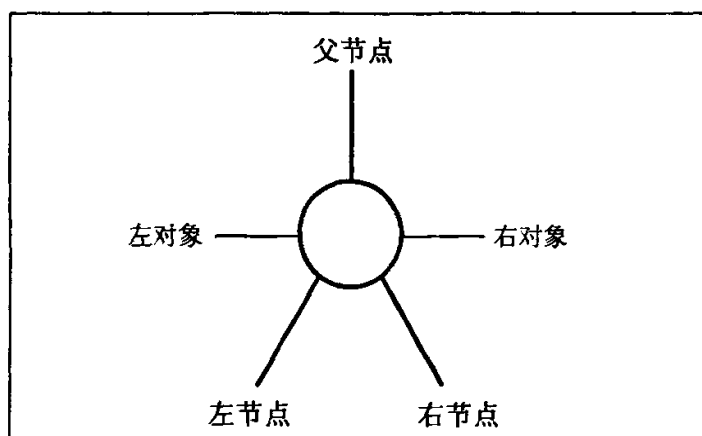


图 5.3 五叉单节点

Fig.5.3 A single node with five branches

2.3.3. 筛选备用检索表

通过对象与特征匹配矩阵逐层分组分析后，矩阵中所蕴含的对象分组路径已基本水落石

出。从节点设计上来说,一条对象分组路径是由一组无序排列的五叉单节点组成,它经匹配矩阵还原后对应一个可备用的检索表。但某些分组路径在逐层分析过程中缺少有效特征(左右节点标记为-1)而未能继续下延,则被当作无效路径排除。

表 5.8 KD 试验备选特征

Table 5.8 Optional characters for KD test

特征编号	二项特征值
1	翅一对
	翅两对
2	前翅退化为伪平衡棒
	前翅正常
3	第一腹节前伸与胸部愈合,第二节呈腰状
	腹部无此现象
4	前翅骨化为翅鞘,无翅脉
	前翅膜质或革质
5	翅密被鳞片
	翅不被鳞片
6	前后翅大小相似,翅多横脉成网状
	前后翅大小不一或二翅,翅少横脉不成网状
7	口器从头的前方生出
	口器从头的后方生出
8	口器咀嚼式、虹吸式或舐吸式
	口器刺吸式
9	翅狭长,有细长缘毛
	翅无细长的缘毛
10	跗节末端有泡状囊,无爪
	跗节末端无泡状囊,有爪
11	下颚及下唇之内、外片分离
	下颚及下唇之内、外片有愈合

2.3.4. 输出优化检索表

基于对象与特征匹配矩阵生成的检索表可能有几种到几千种的表现形式,但这些结果的特征优化度、对象优化度和综合优化度都不尽相同,按照这些评价指标从高到低进行排序,便可从中发现符合优化目标的检索表。但在某些情况下“特征序最优”的检索表不一定是“对象序最优”的检索表,反之亦然,这是因为对象与特征匹配矩阵不能同时满足“特征优先序”与“对象优先序”的优化准则。在检索表编辑大师 KeyMaker 软件中,有三种优化方案供选择:特征序优先、对象序优先、特征与对象兼顾,默认情况下推荐第三种即综合优化度最高的检索表作为最优结果输出。

输出的检索表自动保存为 DAM 电子检索表格式，以便于文献发表、检索系统知识库构建或网上传播与使用。

2.4. KeyMaker 与 DELTA Key 建表试验

KeyMaker 是论文基于上述检索表智能编制技术开发的建表工具，它可根据分类专家建立的对象与特征关联矩阵，以“三序标准”进行优化运算后输出检索表，供专家选择采用。DELTA 是植物分类学描述语言的国际标准格式，DELTA Key 是其系统组件之一，专门利用 DELTA 格式记录的对象与特征匹配关系及其他指示信息生成二叉或多叉的检索表。本研究将通过两者的建表试验（简称 KD 试验），比较它们输出的检索表结果，分析两者的优缺点，并进一步验证智能编制算法的可行性。由于 PANKey 为商业付费软件，其设计原理与 KeyMaker 和 DELTA Key 相差较大，论文未作比较讨论。

2.4.1. KD 试验建表基础材料

建表对象为 8 个常见的农业昆虫目：鳞翅目、鞘翅目、双翅目、膜翅目、同翅目、半翅目、直翅目、脉翅目，可用特征见表 5.8。

对象与特征匹配关系经数字编码处理后得表 5.9。

KD 试验比较过程：提供上述相同的对象、特征和匹配关系，分别输入到它们各自的信息采集环境中，以相同的分类优化思想进行参数设置，最后对软件及输出结果进行优化程度、检索表长度、特征总优先度、软件操作性、稳定性等方面的评价。

表 5.9 KD 试验特征分值矩阵

Table 5.9 Character-object matrix for KD test												
对象	特征值											
半翅目	2	2	2	2	2	2	1	2	2	2	2	2
同翅目	2	2	2	2	2	2	2	2	2	2	2	2
脉翅目	2	2	2	2	2	1	0	1	2	2	2	2
鳞翅目	2	2	2	2	1	2	0	1	2	2	2	2
鞘翅目	2	2	2	1	2	0	0	1	2	2	2	2
直翅目	2	2	2	2	2	2	0	1	2	2	1	1
双翅目	1	2	2	2	2	2	0	1	2	2	2	2
膜翅目	2	2	1	2	2	2	0	1	2	2	2	2

2.4.2. KeyMaker 设置参数及结果

对象优化序为：鳞翅目、鞘翅目、双翅目、膜翅目、同翅目、半翅目、直翅目、脉翅目；

特征优化序为：翅数:翅鳞片:翅缘毛、前翅:前翅质地、腹部:翅脉、口器位置:口器类型、跗节:下颚（冒号相连表示处于同级）；

输出参数：枚举结果无限制；层节点数无限制；允许重复使用特征；综合优化方案
输出检索表：

表 5.10 KeyMaker 输出结果-检索表 A

Table 5.10 Key A generated by KeyMaker

1 翅密被鳞片	鳞翅目
- 翅不被鳞片	2
2 翅两对	3
- 翅一对	双翅目
3 前翅骨化为翅鞘, 无翅脉	鞘翅目
- 前翅膜质或革质	4
4 第一腹节前伸与胸部愈合, 第二节呈腰状	膜翅目
- 腹部无此现象	5
5 口器刺吸式	6
- 口器咀嚼式、虹吸式或舐吸式	7
6 口器从头的后方生出	同翅目
- 口器从头的前方生出	半翅目
7 下颚及下唇之内、外片分离	直翅目
- 下唇及下颚之内、外片有愈合	脉翅目

表 5.11 KeyMaker 输出结果-检索表 B

Table 5.11 Key B generated by KeyMaker

1 翅密被鳞片	鳞翅目
- 翅不被鳞片	2
2 翅两对	3
- 翅一对	双翅目
3 前翅骨化为翅鞘, 无翅脉	鞘翅目
- 前翅膜质或革质	4
4 第一腹节前伸与胸部愈合, 第二节呈腰状	膜翅目
- 腹部无此现象	5
5 口器刺吸式	6
- 口器咀嚼式、虹吸式或舐吸式	7
6 口器从头的后方生出	同翅目
- 口器从头的前方生出	半翅目
7 前后翅大小不一或二翅, 翅少横脉不成网状	直翅目
- 前后翅大小相似, 翅多横脉成网状	脉翅目

2.4.3. DELTA Key 指示参数及结果

DELTA 下载地址: <http://delta-intkey.com/win32/delt32.exe>

CHARACTER RELIABILITIES: 1,10 2,9 3,8 4,9 5,10 6,8 7,7 8,7 9,10 10,6 11,6

ITEM ABUNDANCE: 1,4 2,5 3,6 4,3 5,10 6,9 7,7 8,8

Rbase = 1.40 Abase = 2.00 Reuse = 1.01 Varywt = .80

输出检索表(因 DELTA KeyEditor 中文名支持不佳, 部分对象名前需加字母才能输入):

表 5.12 DELTA Key 输出的检索表
Table 5.12 Key generated by DELTA Key

1. 翅密被鳞片.....	l鳞翅目
翅不被鳞片.....	2
2(1). 前翅骨化为翅鞘,无翅脉.....	q鞘翅目
前翅膜质或革质.....	3
3(2). 翅一对.....	s双翅目
翅两对.....	4
4(3). 第一腹节前伸与胸部愈合,第二节呈腰状.....	o膜翅目
腹部无此现象.....	5
5(4). 口器咀嚼式.....	6
口器刺吸式.....	7
6(5). 前后翅大小相似,翅多横脉成网状;下唇及下颚之内外片有愈合.....	m脉翅目
前后翅大小不一或二翅,翅少横脉不成网状;下颚及下唇之内外片分离.....直翅目
7(5). 口器从头的前方生出.....	b半翅目
口器从头的后方生出.....	t同翅目

2.4.4. 结果分析

从表 5.13 看出, KeyMaker 输出的 2 个检索表虽然优化程度都不错,但最后一个特征的差异(“翅脉”与“下颚”的特征优先序相差 2 级),使得它们的优化结果不尽相同。由于表 A 的综合优化率高于表 B,且两者的特征总优先度相差较小,表 A 应优先选择使用。不过,两者的综合优化度仅差 0.85%,若考虑到特征的总可靠性,用户也可以考虑使用表 B。一般来说,KeyMaker 输出的结果以优化度为主参考条件,检索表长度和特征总优先度为辅参考条件来判断选择。DELTA Key 输出的结果虽然对象优化度比 KeyMaker 略高,但由于特征优化度相差较大,在检索表长度和特征总优先度达到相同水平下,总体优化水平即综合优化度仍没有 KeyMaker 高。

表 5.13 KeyMaker 与 DELTA Key 建表结果比较

Table 5.13 Comparison of KeyMaker and DELTA Key in producing keys

对比项目	KeyMaker		DELTA Key
	表 A	表 B	
综合优化度%	98.91	98.06	94.04
对象优化度%	97.92	97.92	100
特征优化度%	100	98.33	88.75
检索表长度	34	34	34
特征总优先度	20	18	18
软件可操作性	简便		较复杂
软件稳定性	稳定		常自动错误关闭

从两者的制表原理来分析, DELTA Key 采用分组优化即贪心算法 (Reynolds et al., 2003) 的思路, 在每次对象分组时选择算法认为最佳的特征使用; KeyMaker 属于全局优化的思路, 选择整体优化水平高的检索表, 相当于一次性决定每次对象分组使用的特征; 优化思路的差异将导致 KeyMaker 的综合优化效果一般要比 DELTA Key 好, KD 补充试验的对比结果 (表 5.14) 也证实了这一分析结论。此外, KeyMaker 的易用性和稳定性也都优于 DELTA Key。

表 5.14 KD 补充试验结果比较³

Table 5.14 Comparison of results from additional KD test

对比项目	KeyMaker	DELTA Key
综合优化度%	73.48	62.85
对象优化度%	70.60	55.18
特征优化度%	78.33	83.33
检索表长度	32	32
特征总优先度	34	34

2.4.5. 优缺点总结

KeyMaker 是中文开发工具, 自然在中文支持与使用设计上比 DELTA Key 更适合中国人。DELTA Key 没有中文界面, 对大多数分类学者使用操作并不影响, 但在 Win2000 与 WinXP 环境下试用后发现存在如下显著问题:

³ KeyMaker 特征序: 翅数、前翅、腹部、前翅质地、翅鳞片、翅脉、口器位置、口器类型、翅缘毛、跗节、下颚; 对象序: 半翅目、同翅目、脉翅目、鳞翅目、鞘翅目、直翅目、双翅目、膜翅目。DELTA CHARACTER RELIABILITIES 1.10 2.9.5 3.9 4.8.5 5.8 6.7.5 7.7 8.6.5 9.6 10.5.5 11.5; ITEM ABUNDANCE 1.5 2.10 3.9 4.8 5.7 6.6 7.3 8.4, 其他同上 KD 试验

1. 无法输入超过 6 个汉字的对象名称
2. 明明不同的中文对象名被警告“命名不唯一”的错误，需加字母补救
3. 特征值有时输入超过 6 个汉字后确定，引起程序出错关闭
4. 操作过程中程序会因其他不明原因而自动结束，导致前面的操作结果全部丢失

DELTA 虽然颇具盛名，但这些问题实在令人生畏。相比之下，Lucid 家族软件这方面要好得多。试用还发现，特征条目中有 3 个或 3 个以上特征值参选时，DELTA Key 一般生成多叉检索表，因此用 DELTA Key 编制二项检索表应限制特征取值为 2 条，否则不能适用。而 KeyMaker 专为二项检索表制作设计，不存在此问题。

除此之外，DELTA 要求用户给每个对象或特征手工输入一个 0—10 之间的实数表示对象丰度或特征可信度的相互关系，值越高越有利，大小表示期望差异。这种设置操作不仅麻烦容易出错，而且修改或增加参数时容易造成“牵一发而动全身”的影响。KeyMaker 由于采用序的设置概念，通过可视化的排序即可完成相同操作，快捷方便。

不过，DELTA Key 支持特征取值的多重选择，并在特征使用有剩余时尝试把这些特征作为确认性特征放在主特征（分组特征）的后面，如表 5.12 中把“下颚”特征放在了“翅脉”特征的后面，以保证检索顺利进行。KeyMaker 虽然同一特征也允许有多个取值，但每个对象的特征值仍必须唯一，不能出现多重取值。这一限制对于描述某些多样性变异大而较难以统一的特征较为不利，但这并不意味特征多重取值不适合二项检索表的构建，只是引入这一匹配机制后，检索表中同一对象可能出现多次，即拥有多条不同的检索路线。这种情况在二项检索表设计中一般尽量避免。

3 使用方法

3.1. 计算机手工编制电子检索表

计算机手工编制与智能编制电子检索表的差异见表 5.15。手工编制时在 KeyEditor 中输入必要的特征、二项值和分类单元，然后开始建立特征与对象的匹配关系；这是一对多的映射关系，即一个特征为多个对象共有，或说一个对象拥有多个不同的特征。根据这种逻辑关系，KeyMaker 与 KeyEditor 都设计了双向关联的功能。

表 5.15 计算机手工与智能编制检索表差异
Table 5.15 Difference of key construction in manual and heuristic way

不同点	手工编制	智能编制
鉴定策略设计	专家	计算机
使用特征	固定	智能选择
特征值	二项	二项或多项
编辑器	KeyEditor	KeyMaker

1. 无法输入超过 6 个汉字的对象名称
2. 明明不同的中文对象名被警告“命名不唯一”的错误，需加字母补救
3. 特征值有时输入超过 6 个汉字后确定，引起程序出错关闭
4. 操作过程中程序会因其他不明原因而自动结束，导致前面的操作结果全部丢失

DELTA 虽然颇具盛名，但这些问题实在令人生畏。相比之下，Lucid 家族软件这方面要好得多。试用还发现，特征条目中有 3 个或 3 个以上特征值参选时，DELTA Key 一般生成多叉检索表，因此用 DELTA Key 编制二项检索表应限制特征取值为 2 条，否则不能适用。而 KeyMaker 专为二项检索表制作设计，不存在此问题。

除此之外，DELTA 要求用户给每个对象或特征手工输入一个 0—10 之间的实数表示对象丰度或特征可信度的相互关系，值越高越有利，大小表示期望差异。这种设置操作不仅麻烦容易出错，而且修改或增加参数时容易造成“牵一发而动全身”的影响。KeyMaker 由于采用序的设置概念，通过可视化的排序即可完成相同操作，快捷方便。

不过，DELTA Key 支持特征取值的多重选择，并在特征使用有剩余时尝试把这些特征作为确认性特征放在主特征（分组特征）的后面，如表 5.12 中把“下颚”特征放在了“翅脉”特征的后面，以保证检索顺利进行。KeyMaker 虽然同一特征也允许有多个取值，但每个对象的特征值仍必须唯一，不能出现多重取值。这一限制对于描述某些多样性变异大而较难以统一的特征较为不利，但这并不意味特征多重取值不适合二项检索表的构建，只是引入这一匹配机制后，检索表中同一对象可能出现多次，即拥有多条不同的检索路线。这种情况在二项检索表设计中一般尽量避免。

3 使用方法

3.1. 计算机手工编制电子检索表

计算机手工编制与智能编制电子检索表的差异见表 5.15。手工编制时在 KeyEditor 中输入必要的特征、二项值和分类单元，然后开始建立特征与对象的匹配关系；这是一对多的映射关系，即一个特征为多个对象共有，或说一个对象拥有多个不同的特征。根据这种逻辑关系，KeyMaker 与 KeyEditor 都设计了双向关联的功能。

表 5.15 计算机手工与智能编制检索表差异
Table 5.15 Difference of key construction in manual and heuristic way

不同点	手工编制	智能编制
鉴定策略设计	专家	计算机
使用特征	固定	智能选择
特征值	二项	二项或多项
编辑器	KeyEditor	KeyMaker

特征关联：点击工具面板上的“联特征”，特征面板上的特征性状前都会出现选择框(图 5.4)。此时选中阶元面板上的任一对象，与该对象已关联的特征都会被打勾选中。点击特征性状选择框，可将两者匹配或解除匹配。再次点击“联特征”，关闭特征关联。

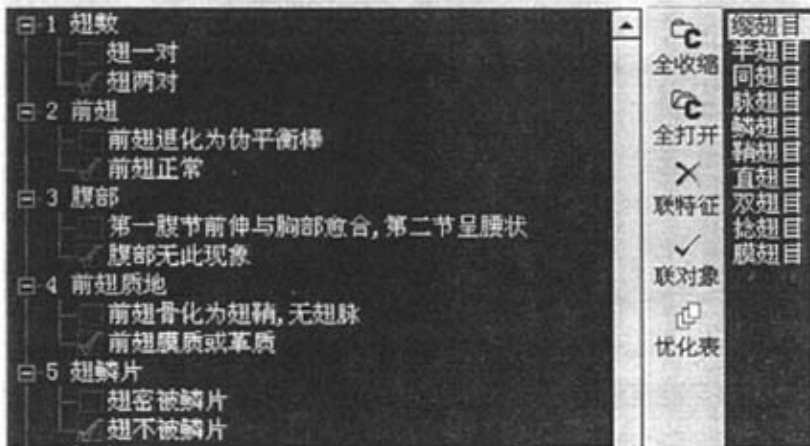


图 5.4 特征关联操作情景

Fig.5.4 Situation of associating with characters

阶元关联：点击工具面板上的“联对象”，阶元面板上对象前面都会出现选择框(图 5.5)。选中特征面板上的任一特征形状后，与此特征已匹配的对象选择框都会被打勾选中。此时选中对象，可将两者匹配或解除匹配。再次点击“联对象”，结束阶元关联。

此外，若 2 个对象前面特征都相同，只在最后一个特征出现性状分歧时，用户可在完成第一个对象的特征关联后，在第二对象上右击菜单中选择“复制关联”，在弹出的窗口中选中第一个对象进行关联复制，即可把第一个对象的特征匹配信息复制给第二个对象。然后再修改一下最后一个不同特征的关联，便很快完成了第二个对象的特征关联。灵活使用“复制关联”功能，可起到“循环利用，事半功倍”的效果。

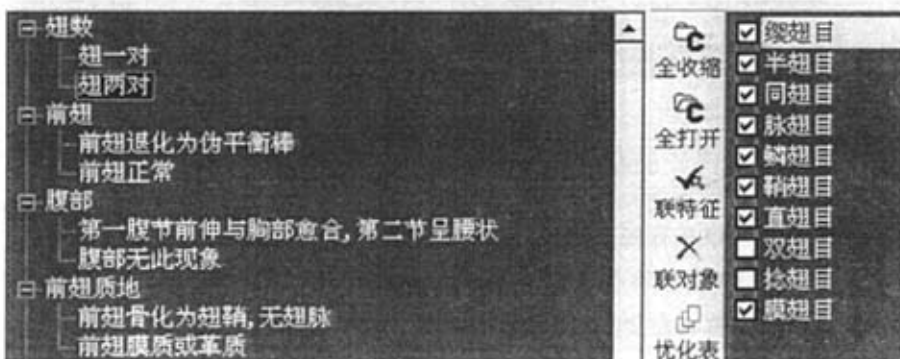


图 5.5 对象关联操作情景

Fig.5.5 Situation of associating with objects

3.2. 导入/导出传统二项检索表

一直以来,分类学家编制的检索表都通过刊物专著等以书面形式发布或者汇总,许多检索表都沉睡在书本中,不能被更多的人广泛使用。KeyImporter 正是为了解决这部分检索表资源的数字化问题而专门设计的工具。从 KeyEditor “文件” 菜单中调出 KeyImporter, 它能把文本二项检索表转换成 DAM 电子检索表。书上的检索表扫描后用 OCR 软件进行识别,再导入 KeyImporter 中。一旦检索表加载后,导入模块会自动分析,用不同颜色标出检索表的各组成部分(红色:错误;灰色:特征序号或序号前缀;黑色:特征描述;绿色:分隔符;紫色:跳转序号或拉丁学名;蓝色:中文名),并将找到的格式错误显示在右边的错误提示窗口中。点击错误提示,导入模块会自动选中包含错误的行。用户及时修正所有提示错误后,再按“分析”重新检查检索表。如果没有发现任何错误,即可导入 KeyEditor 中。

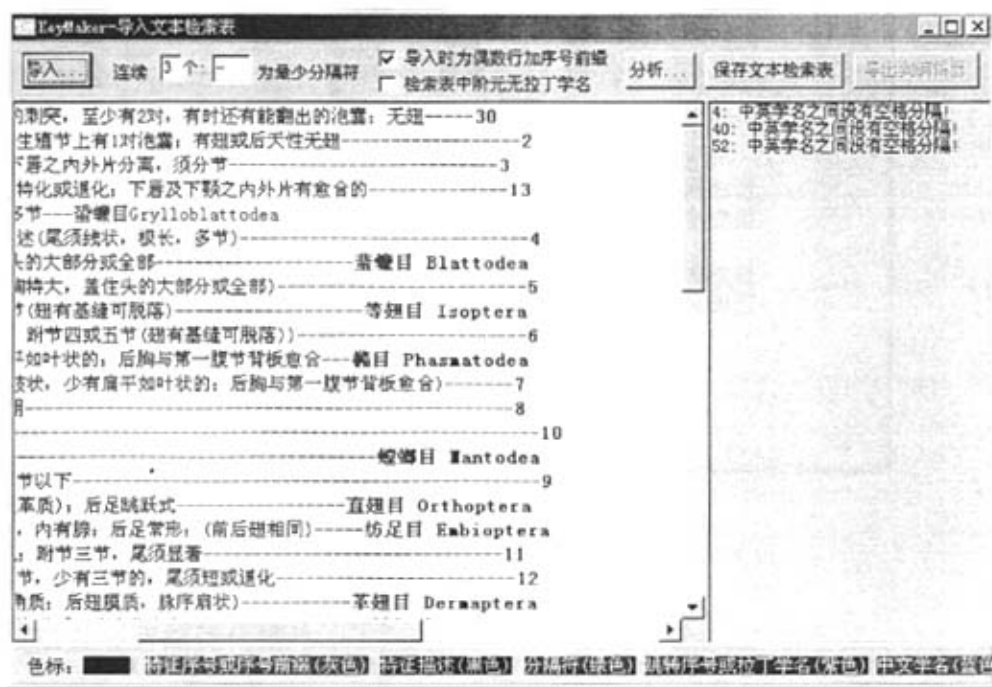


图 5.6 导入模块分析检索表

Fig.5.6 Key analysis in import module

与 KeyImporter 相反, KeyEditor 支持导出电子文本格式或者 WORD 格式的二项检索表。打开“导出”菜单, 点击“文本二项检索表”或者“WORD 排版检索表”, 选择保存位置后即可完成。导出的检索表会自动对齐排版, WORD 输出格式更适合在投稿论文或出版书籍中发表, 省去了手工排版设计检索表的时间与精力, 一键搞定。

3.3. 检索表鉴定策略校验

通常分类专家在设计检索表时，会先大致勾勒出一条鉴定路线，即用某一特征把对象分为2组，再用另一特征继续划分，直到所有对象都独立为一组。在检索表完成后，分类专家会希望能形象化地查看一下检索表整体的鉴定思路，以核对是否与最初的设计相一致。

KeyEditor 支持查看检索表“树型拓扑结构”的功能。用户在完成特征与对象关联后，点击工具面板上的“树拓扑”，便可清晰地看到分类单元在特征划分后向右收缩进显示，越往右，越迟完成鉴定（图 5.7）。点击+、-符号可显示或隐藏节点内容。

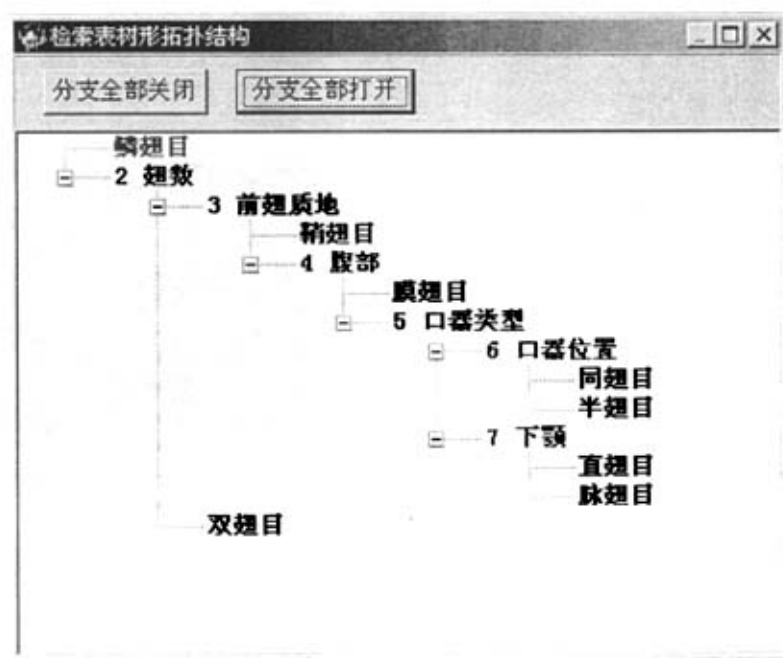


图 5.7 检索表（表 5.10）结构拓扑树

Fig.5.7 Topological structure of the key in Table 5.10

3.4. 检索表智能优化操作

检索表智能优化是基于较完整的特征与对象匹配矩阵实现的，因此在完成关联工作后，才能点击工具面板中“优化表”进入检索表智能优化窗口（图 5.8）。首先设定特征序和对象序，通过“上下移动”可调整排序，“上下合离”可使多个特征或对象处于相同序水平。然后点击“优化预处理”，运算所需时间视特征数量、对象数量以及两者关系矩阵的复杂度而定，一般几秒之内完成，并将结果按指定优化方案得到的优化度排序输出，位于顶端的结果为最佳优化检索表。此时双击该条结果，可在 KeyEditor 中查看检索表结构，并继续附加其他分类信息。如果计算机内存较小，大量的枚举运算导致内存溢出，可调小“层节点数”或“枚举结果”重新预处理。有时可能输出多个优化度相同的结果，用户可调整优化方案或者

依据优化度、检索表长度、特征总优先度的综合判断，择优选用。

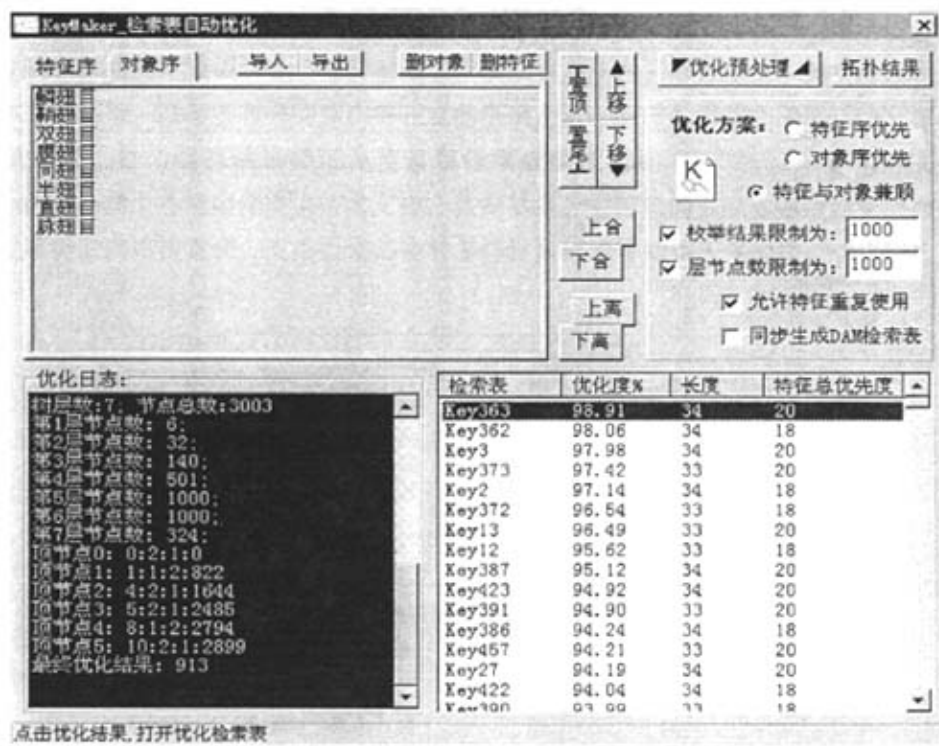


图 5.8 检索表智能优化操作界面

Fig.5.8 Operating interface for heuristic optimization of keys

第六章 分类知识库构建

分类知识由基本分类信息、分类检索表和辅助鉴定信息三部分组成，其中检索表是整个知识体系的基础，几乎所有分类知识都在它的框架上进行有序的添加和关联。论文在构建分类知识库时，分为信息编辑整理与信息编译集成两个流程(图 6.1)，并由 KeyEditor 和 KeyCompiler 两个功能组件提供技术支持，获得基本分类信息的多媒体网页、支持在线鉴定与特征解析的网络检索表、支持查询的物种多样性数据库等知识表现形式。

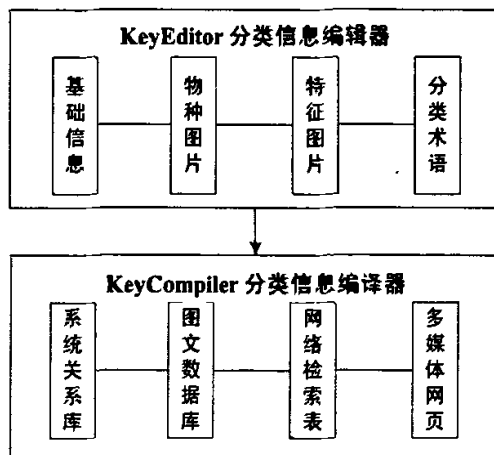


图 6.1 分类知识库构建流程

Fig.6.1 Building procedure of taxonomic knowledge base

1 技术方法

1.1. 物种多样性数据库的构建

建立物种多样性数据库是论文开发的重要功能之一，它由两个基础库组成，分别以表的形式保存在 Microsoft Access 数据库中。

1.1.1. 基本信息库

该库负责记录生物分类阶元的基础信息，设计上由 KeyEditor 负责收集这些基础信息，再 KeyCompiler 整理集成数据库，提供给第三方软件、网站等后期开发使用。在所有字段中，“path”标记了从总纲到该分类单元完整的系统关系，是计算机比较物种分类地位、定位分类单元 XML 源文档的重要索引；“upclass”标记的是分类单元的上级阶元，它是为方便建立

第六章 分类知识库构建

分类知识由基本分类信息、分类检索表和辅助鉴定信息三部分组成，其中检索表是整个知识体系的基础，几乎所有分类知识都在它的框架上进行有序的添加和关联。论文在构建分类知识库时，分为信息编辑整理与信息编译集成两个流程(图 6.1)，并由 KeyEditor 和 KeyCompiler 两个功能组件提供技术支持，获得基本分类信息的多媒体网页、支持在线鉴定与特征解析的网络检索表、支持查询的物种多样性数据库等知识表现形式。

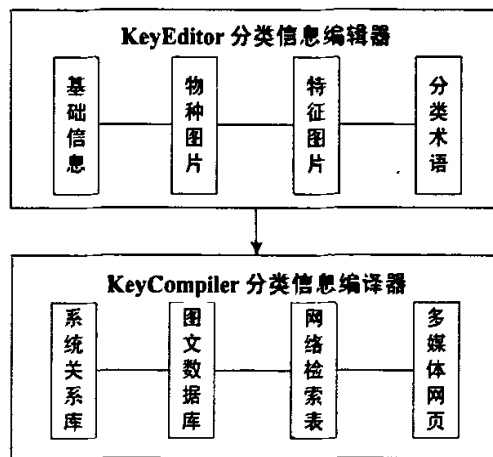


图 6.1 分类知识库构建流程

Fig.6.1 Building procedure of taxonomic knowledge base

1 技术方法

1.1. 物种多样性数据库的构建

建立物种多样性数据库是论文开发的重要功能之一，它由两个基础库组成，分别以表的形式保存在 Microsoft Access 数据库中。

1.1.1. 基本信息库

该库负责记录生物分类阶元的基础信息，设计上由 KeyEditor 负责收集这些基础信息，再 KeyCompiler 整理集成数据库，提供给第三方软件、网站等后期开发使用。在所有字段中，“path”标记了从总纲到该分类单元完整的系统关系，是计算机比较物种分类地位、定位分类单元 XML 源文档的重要索引；“upclass”标记的是分类单元的上级阶元，它是为方便建立

系统分类树而专门设立的。其他字段含义如表 6.1 所述。

表 6.1 基本信息数据库结构
Table 6.1 Database structure for basic information

字段名	含义	实例
id	编号	443
unit	分类单元	科
enname	拉丁名	Cicadidae
chname	中文名	蝉科
path	保存路径	Hexapoda\Insecta\Homoptera\Cicadoidea
upclass	上级阶元	蝉总科(Cicadoidea)
intro	特征	略
distr	分布	略

1.1.2. 物种图片库

该库负责收集由开发者通过网络搜索、书籍扫描或实物拍摄等途径得到的大量物种图片,是提供丰富多样性信息的支持库,同时也是后续开发物种直观识别训练功能的资源库。该库结构如表 6.2,其中 origin 字段专门注明图片引用的 URL 地址或文献出处,以备核对和版权声明。

表 6.2 物种图片数据库结构
Table 6.2 Database structure for species picture

字段名	含义	实例
id	编号	622
enname	拉丁名	Glyptobasis brunnea
chname	中文名	褐端长角蛉
origin	引用地址	www.mdsesd.com.tw/~kinmatsu/animal/neuroptera-1.html
path	保存地址	Hexapoda\Insecta\Neuroptera\Ascalaphidae\2.jpg

1.2. 辅助鉴定信息库的构建

相对于多样性信息库,辅助鉴定信息库侧重于解决分类检索表在使用过程中可能遇到的问题,通过两个基础库与检索表的定位结合,发挥其功用。

1.2.1. 特征图库

为了便于用户了解文字描述中一些抽象或较难懂的特征,论文设计了特征图库的开发功

能,通过示意图形象说明这些特征。特征图库没有采用数据库的管理模式,而直接用文件目录来保存,即以特征部位命名文件夹,将对应的特征图片放入其中,如把各种触角形状的图片存入“antenna”文件夹中。针对某些分类特征描述较复杂,还需建立专门的信息索引文件,详细注明这些文件和文件夹的意义。KeyEditor 则会专门解读这些索引文件,并动态建立特征图目录树(图 6.7-④)。

1.2.2. 分类术语

术语是表达各个专业的特殊概念,由于通用范围有限(百度百科,2007c),行外人看来都抽象难懂。分类检索表中特征用语讲究简洁扼要,经常会出现一些术语,成为一般用户使用检索表的“拦路虎”。为了帮助他们克服这一困难,论文收集了 100 条分类术语,进行了通俗易懂的注解;KeyEditor 也为分类专家提供了专门的术语知识库管理工具,进行添加、修改操作。术语数据库采用表 6.3 的数据结构,并启用数据库密码保护,防止第三方直接篡改,保证术语解释的权威性。

表 6.3 分类术语数据库结构
Table 6.3 Database structure for taxonomic phrase

字段	含义	实例
id	编号	35
voc	中文术语	蛻
eng	英文术语	exuvia
txt	解释	昆虫脱皮时脱下的那层旧表皮
img	示意图片	exuvia.jpg

1.3. 网络检索表的设计

生物分类信息的数字化,其中一个很重要的工作就是实现分类检索表的电子化和在线使用。基于已完成的检索表数字化方案,论文根据传统、问答等普遍适用的检索形式,满足不同用户的使用习惯,设计了 3 种网上可直接交互使用的检索表。

1.3.1. 二叉跳转

该检索方式是传统二项检索表的再现,适合习惯书面二项检索表的用户。由于它基于 Web 设计,在一般二项检索表的外观下补充了两个实用功能:首先,特征序号与跳转索引间建立了书签链接,点击每个索引序号可自动跳到下一个鉴定特征,鉴定到结果时,则自动调入下一级检索表继续鉴定或打开相关分类信息页面;其次,特征描述中嵌入隐性解析机制,即与特征图库和分类术语库关联,点击这些关联信息,得到对应的图片或文字解释说明。二项检索表的产生过程与基于特征分值数字编码方法的二项形式还原方法一致。

1.3.2. 动态交互

与二叉跳转相比,该检索方式更加智能方便,其鉴定思路:先给出每步鉴定需核对的鉴定部位、鉴别特征及示意图片,用户在两项特征中选择其一,检索表会自动排除特征不符的分类单元,动态显示此特征的鉴别能力;点击剩下的分类单元可直接切入下级阶元的检索;选择“继续鉴定”进入下一组特征的判断,选择“后退”可在特征误判时逐步返回上级特征重新开始;如果中间误判太多,可直接点击“重新开始”回到鉴定起点。如此往复,直至弹出唯一的鉴定结果。由于整个鉴定过程都在基于 JavaScript 的通用推理模块 JCDM 中进行,只要导入不同的检索表数字化文件(扩展名为.dam),便可实现不同分类单元的动态交互鉴定。JCDM 模块如下:

```
<script>
var dom=new ActiveXObject("Microsoft.XMLDOM");//加载 XML 文档
var pos=0;//鉴定位置
var kes="";
var score="";//用户的选择
var count=0;//鉴定结果数量
var op1=new Array();//定义数组,分割特征和图片
var op2=new Array();
var ch1=new Array();
var ch2=new Array();
function tezback(){略} //后退功能
function showtez(){略} //显示特征选择
function showresult1(option){略} //选择特征后显示结果,但选择结果不记入 score
function showresult2(option){ //继续鉴定时选择结果记入 score
count=0;
score=score+option;
family.backward.disabled=false;//恢复后退
for(n=0;n<nodeEle.childNodes.length;n++){
kes=nodeEle.childNodes(n).attributes(3).text;
if (score==kes.substring(0,score.length)){//取前一段特征值进行比较
count=count+1;
if (count==1){
for(i=pos+1;i<kes.length;i++){
if (kes.charAt(i)!='0'){
pos=i;//下一个鉴定特征位
break;}

```

```

else
    {score=score+'0';}
}
}
}
dom.async=false;
dom.load('../Hexapoda.dam');//导入不同的检索表数字化文件
nodeEle=dom.getElementsByTagName("分类界元").item(0);
starting.innerHTML=nodeEle.attributes(0).text;
showtez();
</script>

```

1.3.3. Phoenix 检索

Phoenix 是澳大利亚昆士兰大学生物信息技术中心专为网上演示和发布基于计算机的二项检索表而编写的计算机工具(CBIT, 2007b)。它采用跨平台 JAVA 编程技术, 输出的 Phoenix 网络检索表整合文本、超文本、图片等多媒体信息, 可以 ActiveX 插件嵌入浏览器网上传播和在线使用。Phoenix 检索思路与上述动态交互相似, 但它额外扩增了分类单元过滤器及特征问题“跳过忽略”的功能(张小斌等, 2006a), 让用户检索更加轻松方便。此外, Phoenix 也采用了开放的 XML 结构文档, 因此论文可在搞清 Phoenix 数据保存结构的基础上, 无需利用 Phoenix Builder 逐个编制, 通过检索表数字文件(.dam)的结构转换便能生成 Phoenix 检索表, 相当于 KeyEditor 也能用来开发 Phoenix 检索表。Phoenix 检索表的 XML 结构见表 6.4。

表 6.4 六足总纲分纲 Phoenix 检索表
Table 6.4 Phoenix key to Hexapoda at Class level

```

<?xml version="1.0" encoding="Unicode" ?>
<PhoenixKey title="六足总纲" CreatedIn="Lucid Phoenix Builder. File Format © Copyright
  Centre for Biological Information Technology, The University of Queensland.">
  <Identity id="i0" name="原尾纲(Protura)" icon="" url="Protura\main.asp" />
  <Identity id="i1" name="弹尾纲(Collembola)" icon="" url="Collembola\main.asp" />
  <Identity id="i2" name="双尾纲(Diplura)" icon="" url="Diplura\main.asp" />
  <Identity id="i3" name="昆虫纲(Insecta)" icon="" url="Insecta\main.asp" />
  <Steps firstStepID="s0">
    <Step id="s0" text="腹部-触角-眼-前足">
      <Lead stepid="s0" leadid="s0l0" goto="i0" icon="../character/abdomen\1.jpg">
        <Text>腹部 12 节, 无触角, 无复眼和单眼, 前足很长, 代替触角的功用</Text>
      </Lead>
    </Step>
  </Steps>

```



```

<Lead stepid="s0" leadid="s0l1" goto="s1" icon="./character/abdomen\1.jpg">
  <Text>腹部 11 节以下；没有此综合特征</Text>
</Lead>
</Step>
<Step id="s1" text="腹部-腹管-跳器-触角">
  <Lead stepid="s1" leadid="s1l0" goto="i1" icon="./character/abdomen\1.jpg">
    <Text>腹部 6 节以下；腹部下面有腹管及跳器等附属器官；触角 3-6 节</Text>
  </Lead>
  <Lead stepid="s1" leadid="s1l1" goto="s2" icon="./character/abdomen\1.jpg">
    <Text>腹部 6 节以上；无此综合特征</Text>
  </Lead>
</Step>
<Step id="s2" text="口器-眼-中尾丝">
  <Lead stepid="s2" leadid="s2l0" goto="i2" icon="">
    <Text>口器缩在头壳内；没有眼，没有中尾丝</Text>
  </Lead>
  <Lead stepid="s2" leadid="s2l1" goto="i3" icon="">
    <Text>口器生在头壳外；有眼，腹部末端部分有尾须或中尾丝</Text>
  </Lead>
</Step>
</Steps>
</PhoenixKey>

```

1.4. 信息页面批量生成

KeyCompiler 在发布信息时将产生大量的信息页面，包括阶元汇总、单阶元介绍、网络检索表、分类单元鉴定流程等。这些网页都由预先准备的信息模板派生而成，看起来整齐划一。

1.4.1. 模板的作用

模板的概念源于印刷业，是创建统一副本的底板。办公软件 Office PowerPoint 就引入了模板功能，提供样式文稿的格式、配色方案、母版样式及产生特效的字体样式等（百度百科，2007a）。用模板设计的幻灯片内容不同，但整体的显示风格和谐一致。模板在网页设计中也使用甚广，事先规定好网页平面的框架样式，只需按照预留位置填入相应的文字图片信息，便可快速生成各类内容不同的网页。

1.4.2. 分类单元信息模板

图 6.2 是论文中为单阶元信息介绍设计的网页模板，其中类似“<#.>”的信息元便是预留的信息插入位点。在各个阶元信息页面生成时，KeyCompiler 会自动用实际内容替换这些信息元，并保存为不同的文件。<#system>信息元将被页面生成时间代替，以表明信息的更新时间。

<#title>				
学 名:	<#sname>	俗 名:	<#cname>	查看鉴定流程
介 绍:	<#intro>			
分 布:	<#distribution>			
鉴 定 部 位:	<#location>			
<#images>				
<#system>				

图 6.2 单阶元信息介绍模板
Fig.6.2 Template for introduction of one taxon

1.5. 鉴定流程的自动生成

鉴定流程是分类单元完成鉴定时所核对特征的集合列表，它完全得益于特征分值法的扩展应用。由于检索表完成数字编码后，每个分类单元都有一个特征值，把其中的非 0 编码进行特征还原，并按次序串联起来，就是一个鉴定流程。如果按照分类系统关系把多个特征值合并，编码还原后便是一个更长的鉴定流程。根据用户需要，它可以从目到属，也可以从纲到种。在“中国昆虫分类鉴定系统 InsectX (张小斌等, 2006b)”中，KeyCompiler 会为每个分类单元生成它回溯到六足总纲的鉴定流程（图 6.3），以完整显示分类单元最具鉴别力的特征。如果某些分类单元存在多条检索途径，鉴定流程中以数字索引标出各条分叉的流程路线。

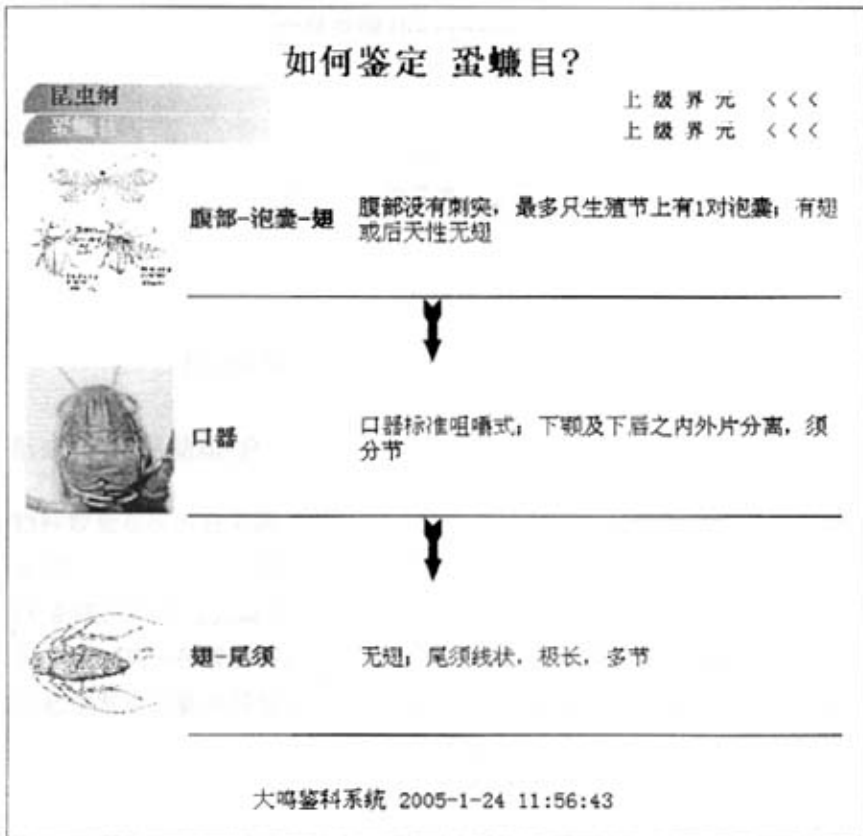


图 6.3 蜚蠊目鉴定流程图
Fig.6.3 Diagnostic flow of Grylloblattodea

1.6. 系统分类树的实现

系统分类树是生物分类单元之间系统关系的直观显现，是生物多样性数据库和生物分类系统的重要组成部分。KeyEditor 在新建电子检索表时即要求按照分类地位存放，如膜翅目 Hymenoptera 电子检索表应保存在昆虫纲 Insecta 的文件夹中。经过这样的预安排，整个文件目录默认就隐含了分类系统关系，利用 KeyCompiler 重新遍历整个文件目录，便能生成分类单元的关系数据库（表 6.5）。

以上级阶元 Hexapoda 进行数据库查询，可得到六足总纲的首级分类树（图 6.4）；再用下级阶元继续嵌套查询，只要分类单元关系数据库足够完整，可显示整棵系统分类树。

六足总纲(Hexapoda)

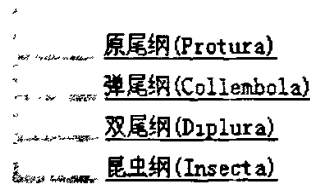


图 6.4 六足总纲分纲系统树
Fig.6.4 Systematic tree of Hexapoda at Class level

1.7. 分布式信息更新维护

昆虫物种数量高居所有生物类群之首，但目前只发现了其中的 10%多（黄复生，1991），今后相当长的时间内，随着新的分类阶元不断产生，整个昆虫分类系统处于不断更新的状态中，InsectX 系统的维护与更新工作必不可少。由于分类工作的不断深入细化，绝大多数昆虫分类专家只专注于某一种群的研究，导致分类资料广泛分散，统一整理难度较大。此外，生物多样性信息学发展也要求开发有助于信息交换使用和跨平台集成，并能相嵌形成完整知识构架的系统（Frondorf and Waggoner, 1996; Bisby, 2000）。为此，论文专门设计了“分布式信息维护模块（图 6.5）”，各地分类专家可利用此模块自行构建某一类群的分类树，添加相关文字图片信息，再由维护中心统一收集这些单一种群的分类数据，依据分类关系进行系统的衔接与整合，在保证系统资料可靠性的基础上，完成分类信息的集成更新。用户在浏览时可注意每个页面下方的系统产生时间，了解该内容的更新情况。

表 6.5 分类单元关系数据库结构示例
Table 6.5 Example of relational database for insect taxonomy

分类阶元	中文名	拉丁学名	获取路径	上级阶元
总纲	六足总纲	Hexapoda	\	
纲	原尾纲	Protura	Hexapoda\Protura	Hexapoda
纲	双尾纲	Diplura	Hexapoda\Diplura	Hexapoda
纲	弹尾纲	Collembola	Hexapoda\Collembola	Hexapoda
纲	昆虫纲	Insecta	Hexapoda\Insecta	Hexapoda
...
种

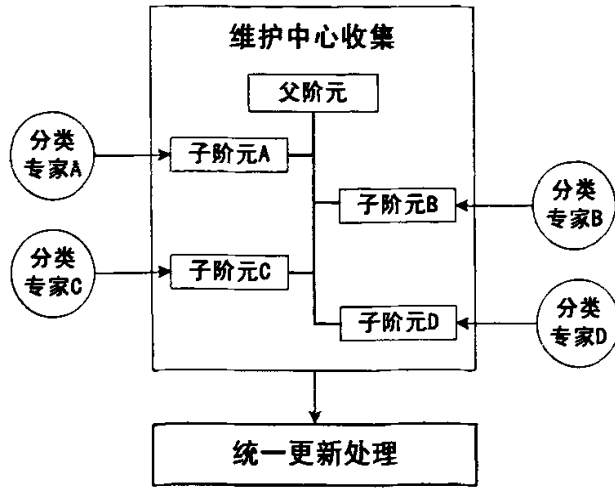


图 6.5 分布式信息维护流程

Fig.6.5 Maintenance procedure for distributed information

2 使用介绍

2.1. 分类单元基础信息的附加

给分类单元附加学名、俗名、特征、分布等文字信息是构建物种多样性数据库的主要工作，论文以检索表体系为框架，有序地组织输入上述多样性信息。在 KeyEditor 中编辑电子检索表时，双击分类单元打开“基础信息维护窗口（图 6.6）”补全信息。其中名称信息随检索表直接存入 XML 文档，特征分布信息保存在置于分类单元拉丁名目录的 info 文本文件中。

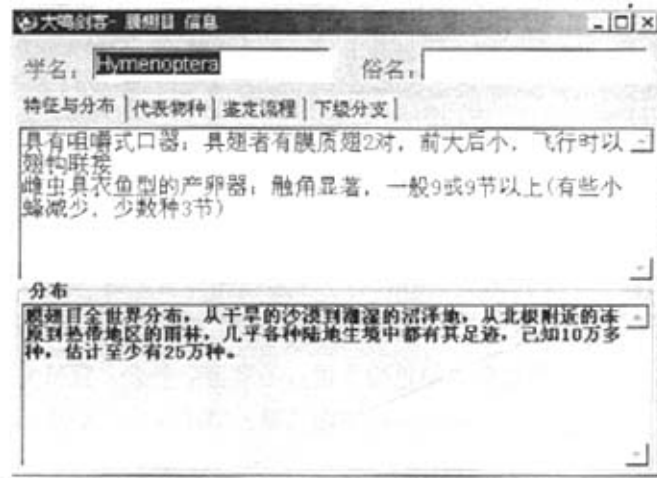


图 6.6 基础信息输入界面

Fig.6.6 Entry interface for basic information

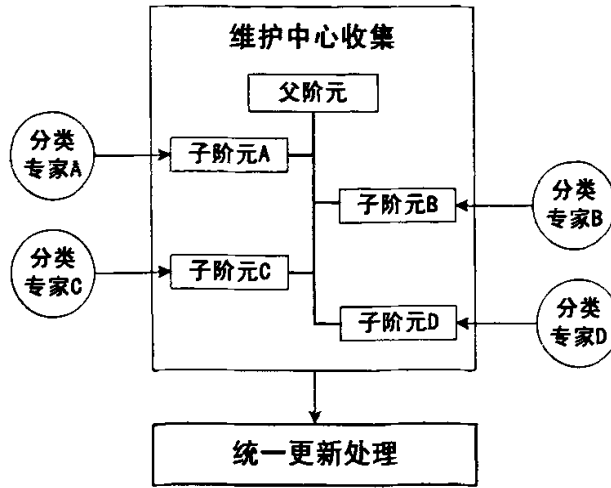


图 6.5 分布式信息维护流程

Fig.6.5 Maintenance procedure for distributed information

2 使用介绍

2.1. 分类单元基础信息的附加

给分类单元附加学名、俗名、特征、分布等文字信息是构建物种多样性数据库的主要工作，论文以检索表体系为框架，有序地组织输入上述多样性信息。在 KeyEditor 中编辑电子检索表时，双击分类单元打开“基础信息维护窗口（图 6.6）”补全信息。其中名称信息随检索表直接存入 XML 文档，特征分布信息保存在置于分类单元拉丁名目录的 info 文本文件中。

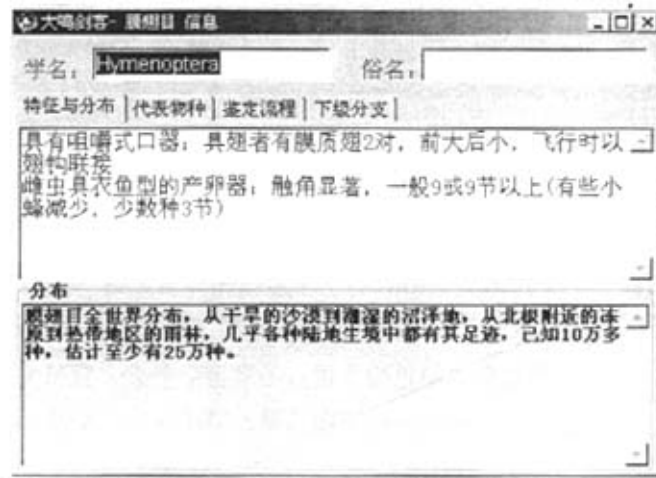


图 6.6 基础信息输入界面

Fig.6.6 Entry interface for basic information

2.2. 特征与示意图图片的关联

特征图关联是手动操作过程，目的是实现文字描述与特征图的一一对应，使检索表在编译时能自动嵌入这些图片的链接信息。KeyEditor 中选择任一特征描述后，在右键菜单中点击“特征图片”打开特征图关联界面（图 6.7）。如果特征描述是由分号隔开的多个不同特征组成，该特征描述会被自动拆成多行，以分别关联不同的特征图。



图 6.7 特征关联界面

(① 检索表特征 ② 特征列表 ③ 工具栏 ④ 特征图导航 ⑤ 特征图)

Fig.6.7 Interface for associating character description with pictures

(① key characters ② character list ③ toolbar ④ picture guider ⑤ character pictures)

2.3. 分类术语词库的管理

术语词库管理在于不断丰富或修正这本注解各种分类术语的专业词典，增强它对辅助鉴定的作用。KeyEditor 中从“解释”菜单打开“专业词库”管理界面（图 6.8）。已入库的专业词汇以“汉字首字拼音”为序分组显示，每个都可添加或修改中文名、英文名、具体含义及示意图等信息。供选的示意图默认都存放在 phase\images 目录下。

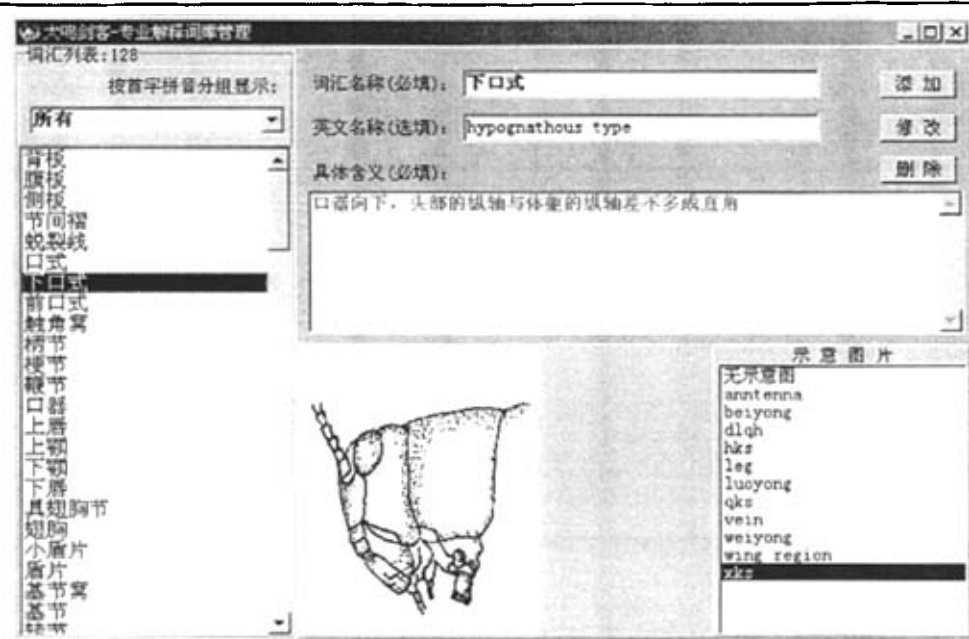


图 6.8 分类术语管理界面

Fig.6.8 Management interface for taxonomic terminology

2.4. 系统内容的更新

生物分类系统初步完成后,随着分类单元的增加、特征图库与术语词库的补充完善或者信息模板的修饰改造,系统需要定期更新知识内容,否则容易出现陈旧老化的问题。论文考虑到了这一重要的信息维护环节,针对内容更新的不同作用范围,开发了全面与局部两种灵活的更新方式:

全面更新由 KeyCompiler (图 6.9) 执行操作,作用范围是指定分类阶元下所有电子检索表包含的分类单元,由用户选择“系统目录”设定;更新内容可包括所有分类信息介绍页面与网络检索表,由用户多选决定。全面更新多在信息模板重新设计或术语词库大幅改动后使用,若因涉及较大物种类群导致更新占用较长时间时,可考虑分子类群完成。



图 6.9 KeyCompiler 操作界面

Fig.6.9 Operating interface of KeyCompiler

局部更新由 KeyEditor 编辑时完成，作用范围是当前操作电子检索表中的分类单元。用户修改过知识内容后时选择“编辑”菜单下的“编译阶元信息”或“编译检索表”命令更新网页或网络检索表。局部更新快而灵活，是日常维护的常用操作。

第七章 网络分类与检索

网络的兴起与发达,使人们足不出户便能接触到丰富的资讯信息。分类知识实现数字化,除了便于计算机分析处理、二次开发,更重要的目的在于能利用网络的优势促使知识的快速传播、广泛使用与共享研究。论文在前面已成功解决了检索表电子化设计编制、现有二项检索表电子化改造、物种多样性数据库构建、分类知识系统化集成等问题,所有通过编辑和编译的分类信息最终将由 KeyPublisher 向互联网推送,支持分类知识库的在线浏览、学习与利用。KeyPublisher 是通用的分类知识网络发布平台,只要装入预编译的知识信息,便可以“生物分类鉴定网络系统”的身份在互联网上公开发布(图 7.1),提供分类检索的信息服务。

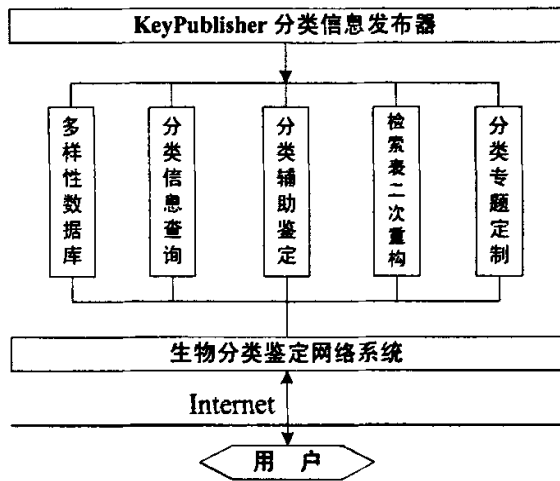


图 7.1 分类知识发布流程

Fig.7.1 Publishing flow of taxonomic knowledge base

1 检索表二次重构技术

检索表二次重构是一种新颖的生物检索表构建技术。该技术利用丰富的基础分类检索表资源,通过分类单元与鉴定特征间匹配关系的反向推理(沈爱华等,2006),自动生成一种由指定对象和必要特征组成的二次检索表(Secondary Identification Key,简称 SIK)。简而言之,它是一种“表生表”的创建技术,目前 KeyPublisher 已集成了论文开发的昆虫分类检索表重构工具 SIKey,并支持定制科及以上分类单元、蜜蜂科所有已知种进行二次重构。

1.1. 二次检索表产生背景

检索表在生物分类中有着广泛的应用。为了指导人们识别鉴定物种,分类专家建立了一

第七章 网络分类与检索

网络的兴起与发达,使人们足不出户便能接触到丰富的资讯信息。分类知识实现数字化,除了便于计算机分析处理、二次开发,更重要的目的在于能利用网络的优势促使知识的快速传播、广泛使用与共享研究。论文在前面已成功解决了检索表电子化设计编制、现有二项检索表电子化改造、物种多样性数据库构建、分类知识系统化集成等问题,所有通过编辑和编译的分类信息最终将由 KeyPublisher 向互联网推送,支持分类知识库的在线浏览、学习与利用。KeyPublisher 是通用的分类知识网络发布平台,只要装入预编译的知识信息,便可以“生物分类鉴定网络系统”的身份在互联网上公开发布(图 7.1),提供分类检索的信息服务。

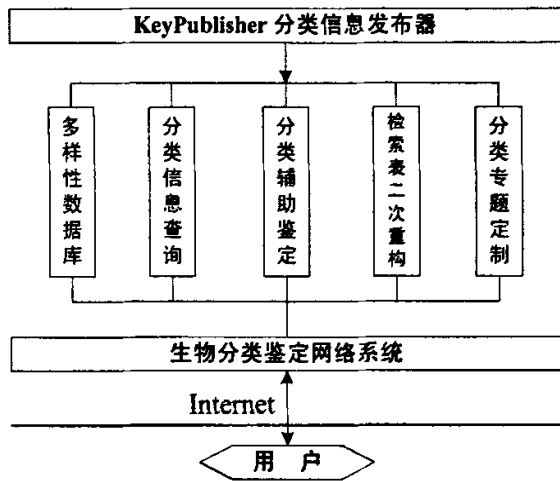


图 7.1 分类知识发布流程

Fig.7.1 Publishing flow of taxonomic knowledge base

1 检索表二次重构技术

检索表二次重构是一种新颖的生物检索表构建技术。该技术利用丰富的基础分类检索表资源,通过分类单元与鉴定特征间匹配关系的反向推理(沈爱华等,2006),自动生成一种由指定对象和必要特征组成的二次检索表(Secondary Identification Key,简称 SIK)。简而言之,它是一种“表生表”的创建技术,目前 KeyPublisher 已集成了论文开发的昆虫分类检索表重构工具 SIKey,并支持定制科及以上分类单元、蜜蜂科所有已知种进行二次重构。

1.1. 二次检索表产生背景

检索表在生物分类中有着广泛的应用。为了指导人们识别鉴定物种,分类专家建立了一

整套较完善的生物分类体系，从纲到种，检索范围覆盖每一种已发现的物种，信息量大，检索表多。由于检索表主要从形态特征相似性的角度来划分物种，并尽量体现物种间的亲缘关系，当应用于某些与农业生产或社会生活有密切利害关系的物种鉴定时，涉及分类单元多，物种间关系复杂，导致检索表信息冗长复杂，使用起来难度较大。比如，白蚁是重要的社会性昆虫，我国已知种类 474 种，其中危害人类生产生活的常见种主要有十几个(黄复生等, 2000)。如果利用现有的基础分类检索表鉴定这十几种白蚁，所需查阅的检索表不下十几个，所需核对的鉴别特征不下几十个，鉴定难度不是一般非昆虫分类专家所能承受的。为了解决这一问题，分类专家会针对这十几种白蚁，重新选择合适特征，另外建立专用的分类检索表，以降低鉴定的使用难度，提高鉴定工作效率。这种为了弥补分类研究与实际应用间差距而定制检索表的现象(Watson and Milne, 1972)，在水稻、棉花、蔬菜、果树、仓库以及检疫等害虫鉴定中也常见到。一旦鉴定对象发生变化，它们必须要求有关分类专家重新构建新的专用检索表，以满足实际应用的需要。为了适应这方面日益增多的需求，并减少分类专家为编制专用检索表花费的额外时间与精力，论文开始了检索表二次重构的研究。

1.2. 二次重构原理和算法

从分类系统整体来看，任何两个分类单元，尽管它们处于不同的目、科或属等分类地位上，但在二项检索表中必有一个分叉点(crossing node)即一个特征的差异，能将它们区分开来，而分叉点以上的特征是它们所共有的。因此，只要基于一定数量的检索表，给定任何两个分类单元，都可以通过遍历(tree-tracing)检索表，找出它们的分叉点，并由此构建它们两者的鉴定检索表。同理，给定任何三个分类单元，根据它们在分类系统中的亲缘关系(图 7.2-α)，将其中两个关系较近的先划为一组，建立两者鉴定用的子检索表(sub key)；再将它们分叉点以上的共同特征看作一个新的分类单元，与剩余的分类单元比较，建立另一个子检索表；最后将两个检索表依照系统关系进行衔接，便可得到它们三者的鉴定检索表(图 7.2-β)。根据这样的组建思路，即使给出任意个分类单元，只要基于它们在现有分类系统的分叉点和亲缘关系，都可以构建出一个鉴别特征明确，信息长度较小的新检索表。图 7.3 是实现检索表二次重构的基本算法。

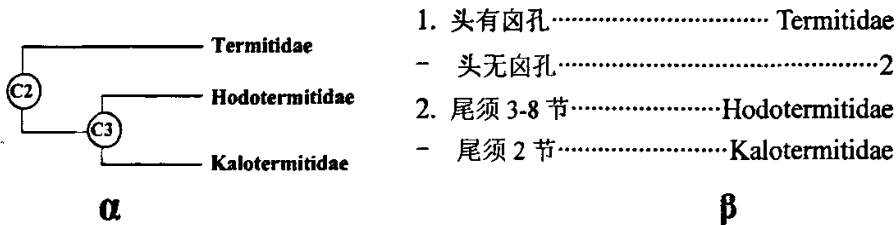


图 7.2 三个分类单元重构演示

(a: 三者亲缘关系; β: 子检索表)

Fig.7.2 Demonstration of reproducing a sub key of three taxa

(a: relationship with three taxa; β: sub key)

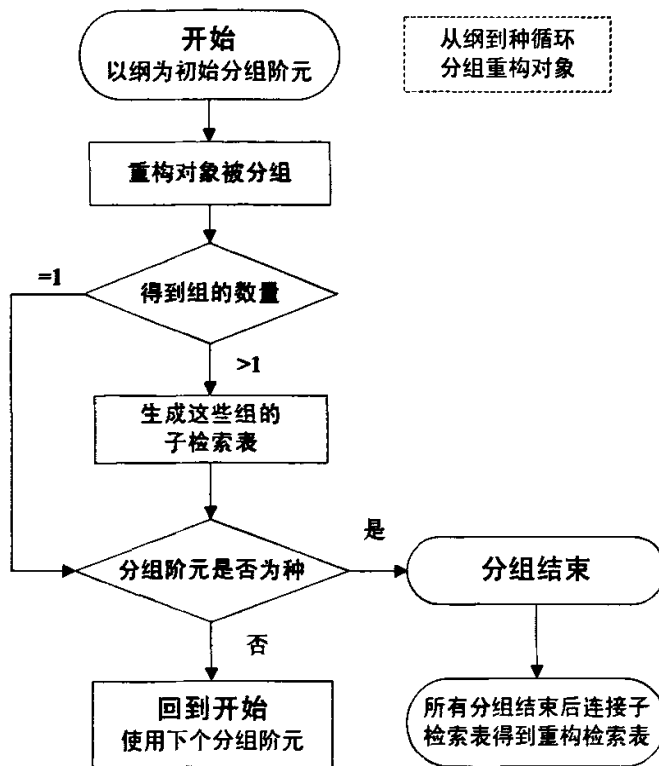


图 7.3 二次检索表重构算法

Fig.7.3 Algorithm for generation of Secondary Identification Key

1.3. 二次重构技术的优点和意义

二次重构技术从基础分类检索表中剔除了无关的鉴定对象，保留了与鉴定对象有关的特征信息，起到了化繁为简的作用。从表 7.1 对比结果可看出，利用二次检索表鉴定，涉及用到的检索表只有一个，需要核对的特征比利用分类检索表少得多，使用变得简单快捷；从理论上讲，二次检索表是完全基于基础分类检索表资料库，通过反向逻辑推理与比较方法得到的，其鉴定准确性与传统分类检索表保持一致，而经人工实际推理检验也证明了其可靠性。二次重构技术对于用户自定义物种的鉴定要求特别有效，上述的白蚁鉴定问题以现存的权威白蚁分种检索表为知识基础进行二次重构，即可获得用于常见危害白蚁种类鉴定的二次检索表（表 7.4）。它不需要分类专家的重复劳动，便可从纷繁冗长的传统分类检索表中提取出关键特征信息，即使出于地理分布或者环境条件的差异而造成鉴定对象有所增减，都可以随时再进行重构，得到使用方便而鉴定准确的二次检索表。应用二次重构技术，还可轻松地获得其他一些有直接应用价值的二次检索表，如常见棉花害虫、仓库害虫、果树害虫、检验害虫、名贵蝴蝶等检索表，它为广泛高效利用分类学家的专家知识发挥了重要的作用。

表 7.1 使用二次检索表 SIK 和分类检索表鉴定中国常见危害白蚁对比结果
Table 7.1 Comparison of diagnosis of common termite pests in China using SIK and conventional keys

物种	使用检索表 (个)		核对特征 (个)	
	SIK	分类检索表	SIK	分类检索表
<i>C. declivis</i>	1	4	5	17
<i>C. domesticus</i>	1	4	5	15
<i>R. flaviceps</i>	1	5	6	26
<i>R. chinensis</i>	1	5	6	24
<i>C. formosanus</i>	1	4	3	35
<i>O. formosanus</i>	1	4	4	39
<i>M. barneyi</i>	1	4	4	41
平均值	1	4.3	4.7	28.1

1.4. 二次重构的计算机实现方法

在实际应用中, 二次检索表主要用于种水平上的重构。而在传统分类鉴定中, 要将一个未知对象鉴定到种, 通常要查阅分目、分科、分属再到分种的检索表, 其间涉及的亲缘关系之复杂, 分类特征之多, 人工重构十分困难。即使在一个检索表内作分类单元的重构, 有时由于分类单元众多, 检索信息复杂, 也使人工重构相当麻烦。如果把重构的设计思路编写成计算机程序, 无论检索表多么庞大, 重构对象数量多少, 计算机都能轻而易举地迅速完成。因此, 二次重构技术必然要依托计算机技术来实现二次检索表的自动化生成。下面以中国常见危害白蚁为重构对象, 介绍论文已实现应用的计算机二次重构方法:

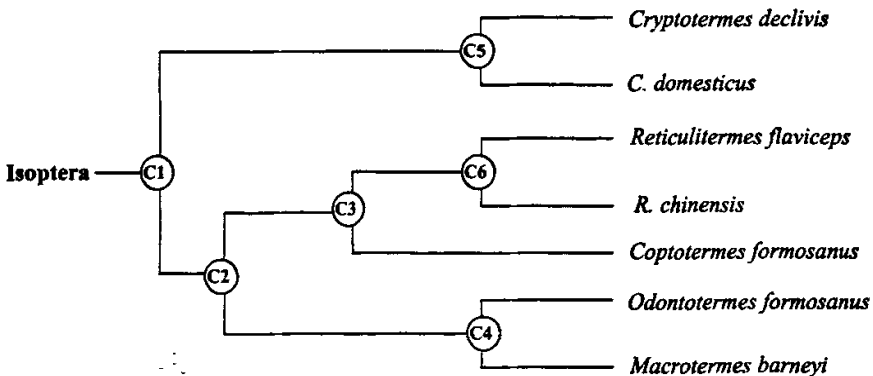


图 7.4 中国常见危害白蚁亲缘关系树型图

Fig.7.4 Tree-like relationship of common termite pests in China

1.4.1. 重构执行方式

由于重构单元的分类地位不同，它们往往分布在各个不同的检索表中，重构时一般会出现三种情况：水平重构、垂直重构和交叉重构。当重构对象都处于同一个分类检索表中时，它们相互间的区别特征通过相对特征的回溯比较即可获得，这种水平重构适合于提取同级分类单元的特征差异；而在垂直重构中，分类单元都位于不同的检索表中，它们的区别特征不能直接通过检索表比较提取，而必须先寻找重构单元分类地位上的相似点，用它们上级阶元的特征差异作为它们的重构特征，类似于将垂直重构转换为水平重构处理。例如，常见水稻害虫黑尾叶蝉属于同翅目叶蝉科，稻象属于鞘翅目象甲科，它们两者的区别特征应该用昆虫分目检索表中同翅目与鞘翅目的区别特征来代替；交叉重构是以上两种重构的混合形式，也是二次检索表重构中最常见的形式，只有正确了解各重构单元在生物分类系统的亲缘关系，才能尽快发现它们之间的特征分叉点，顺利地完成交叉重构。

1.4.2. 对象亲缘关系分析

分清重构对象在分类系统中的亲缘关系并据此进行分组，是构建子检索表的基础。该过程一般从纲、目、科、属、种等基本分类阶元依次开始，相同分类地位的对象归为一组，再转入下个分类阶元继续划分，直到各组中只有一个对象。图 7.4 是国内七种常见危害白蚁亲缘关系的树型图，它们分组后的结果如表 7.2。

表 7.2 中国常见危害白蚁亲缘关系表
Table 7.2 Taxonomic relationship of common termite pests in China

白蚁种类	科	属	亚属	种
铲头堆砂白蚁 <i>Cryptotermes declivis</i> 截头堆砂白蚁 <i>C. domesticus</i>	木白蚁科 Kalotermitidae	堆砂白蚁属 <i>Cryptotermes</i>	—	C5
黄胸散白蚁 <i>Reticulitermes flaviceps</i> 黑胸散白蚁 <i>R. chinensis</i> 台湾乳白蚁 <i>Coptotermes formosanus</i>	C2 鼻白蚁科 Rhinotermitidae	C3 散白蚁属 <i>Reticulitermes</i>		C6
黑翅土白蚁 <i>Odontotermes formosanus</i> 黄翅大白蚁 <i>Macrotermes barneyi</i>	白蚁科 Termitidae	C4 土白蚁属 <i>Odontotermes</i> 大白蚁属 <i>Macrotermes</i>		

1.4.3. 提取特征分叉点

利用计算机实现检索表二次重构，最关键在于分析处理基础分类检索表中对象与特征的匹配关系，以从中找出任两个对象的特征分叉点。根据论文介绍的检索表数字化技术，利用特征分值数字编码方法对表 7.2 中每个同组同级地位的重构单元进行特征编码，转化后得到表 7.3 的检索表数字编码信息。它们完整地记录了重构对象与所有分类特征的匹配关系，不过当中许多对于重构对象鉴别是没有作用的，如每个分类特征的数字串（竖行）中，没有 1 与 2 编码同时出现，说明两个重构对象该特征相同或者缺乏该分类特征的信息，那么这一数字串便是无效数据，可从此级数字信息中删除。同理依次对其他分类特征进行分析筛选后，剩下的便都是对区别重构对象有贡献的有效鉴别特征，也就是特征分叉点。

表 7.3 中国常见危害白蚁特征编码信息
(方框中的有效编码代表分叉点特征)

Table 7.3 Character coding to common termite pests in China
(valid coding or crossing nodes are bold-faced in a frame)

白蚁种类	科	属	亚属	种
<i>C. declivis</i>	2	1200		1202202
<i>C. domesticus</i>				1201000
<i>R. flaviceps</i>	220	12	200222	20
<i>R. chinensis</i>			11	
<i>C. formosanus</i>			200100	
<i>O. formosanus</i>	22020	11220000	201100000000000000000000	
<i>M. barneyi</i>		11220000	200000000000000000000000	

1.4.4. 组合子检索表

此步是对传统二项检索表的重建过程，相当于编码信息的翻译表达。由于每个特征分叉点对应一个子检索表（图 7.2），根据检索表数字化时保留的特征编码信息，逐个反向转化分叉点为对应的特征描述信息，并按照二项检索表的书写格式输出，即可得到所有需要的子检索表。最后依照所有分叉点的树型结构，以树叶到树根的顺序衔接组合子检索表，得到最终的重构结果（表 7.4）。

基于矩阵数字编码的检索表信息，计算机完成二次重构非常方便，原来分类学家需要几个小时才能设计完成的专用检索表，计算机在几秒之内便能输出相同作用的二次检索表。这也显示了检索表数字化带来的极大便利。

表 7.4 SIKey 重构得到的中国常见危害白蚁二次检索表
Table 7.4 SIK to common termite pests in China generated by SIKey

1 头无凶孔	5
- 头有凶孔	2
2 前胸背板扁平, 无前叶 (前缘不翘起)	3
- 前胸背板马鞍形, 有前叶 (前缘翘起)	4
3 凶孔在头前, 有 1 短管	台湾乳白蚁(<i>Coptotermes formosanus</i>)
- 凶孔在正常位置, 无短管	6
4 二型或三型	黄翅大白蚁(<i>Macrotermes barneyi</i>)
- 仅单型	黑翅土白蚁(<i>Odontotermes formosanus</i>)
5 触角窝上方的额角突很小, 下方颊角突发达头堆砂白蚁(<i>Cryptotermes domesticus</i>)	
- 触角窝上方的额角突和下方颊角突同样发达铲头堆砂白蚁(<i>Cryptotermes declivis</i>)	
6 额平或微隆	黑胸散白蚁(<i>Reticulitermes chinensis</i>)
- 额峰明显, 或强隆	黄胸散白蚁(<i>Reticulitermes flaviceps</i>)

1.5. 二次重构的基础开发

作为一种新颖的生物检索表创建技术, 二次重构技术生成的专用检索表使用之便, 检索之快, 是包含了冗长信息的基础分类检索表不能比拟的。然而, 实现二次重构的重要能源来自自己完成数字化编码的基础分类检索表。如果要能随心所欲地重构任何所有已知的昆虫物种, 首当其冲应完成现有昆虫所有分种基础检索表的数字化工作。但昆虫数量众多, 分类系统复杂, 完成这一工作, 将十分庞大和艰巨。目前, 本研究已利用前期开发完成的 KeyEditor 和 KeyImporter 完成了昆虫科及以上分类阶元、蜚蠊科所有分类检索表的数字化工作。访问由 KeyPublisher 发布的“中国昆虫分类鉴定系统 InsectX”, 输入若干科名或蜚蠊种名, SIKey (图 7.5) 会自动完成这些指定对象的二次重构, 并在线输出相应的专用检索表。考虑到真正有实际用途的一般是种级甚至是亚种级的二次重构, 编码并转化科级以下分类阶元的检索表信息势在必行, 但它需要得到国内各路昆虫分类专家的支持与合作, 才能奠定二次检索表种级重构所必需的数字化基础。

2 其他技术

2.1. 分类信息解析机制

昆虫分类鉴定是一项专业性较强的工作, 为了使一般用户能理解一些较难的特征描述与专业术语, 降低鉴定难度, 本研究中使用了两种的分类信息解析机制:

表 7.4 SIKey 重构得到的中国常见危害白蚁二次检索表
Table 7.4 SIK to common termite pests in China generated by SIKey

1 头无凶孔	5
- 头有凶孔	2
2 前胸背板扁平, 无前叶 (前缘不翘起)	3
- 前胸背板马鞍形, 有前叶 (前缘翘起)	4
3 凶孔在头前, 有 1 短管	台湾乳白蚁(<i>Coptotermes formosanus</i>)
- 凶孔在正常位置, 无短管	6
4 二型或三型	黄翅大白蚁(<i>Macrotermes barneyi</i>)
- 仅单型	黑翅土白蚁(<i>Odontotermes formosanus</i>)
5 触角窝上方的额角突很小, 下方颊角突发达头堆砂白蚁(<i>Cryptotermes domesticus</i>)	
- 触角窝上方的额角突和下方颊角突同样发达铲头堆砂白蚁(<i>Cryptotermes declivis</i>)	
6 额平或微隆	黑胸散白蚁(<i>Reticulitermes chinensis</i>)
- 额峰明显, 或强隆	黄胸散白蚁(<i>Reticulitermes flaviceps</i>)

1.5. 二次重构的基础开发

作为一种新颖的生物检索表创建技术, 二次重构技术生成的专用检索表使用之便, 检索之快, 是包含了冗长信息的基础分类检索表不能比拟的。然而, 实现二次重构的重要能源来自自己完成数字化编码的基础分类检索表。如果要能随心所欲地重构任何所有已知的昆虫物种, 首当其冲应完成现有昆虫所有分种基础检索表的数字化工作。但昆虫数量众多, 分类系统复杂, 完成这一工作, 将十分庞大和艰巨。目前, 本研究已利用前期开发完成的 KeyEditor 和 KeyImporter 完成了昆虫科及以上分类阶元、蜚蠊科所有分类检索表的数字化工作。访问由 KeyPublisher 发布的“中国昆虫分类鉴定系统 InsectX”, 输入若干科名或蜚蠊种名, SIKey (图 7.5) 会自动完成这些指定对象的二次重构, 并在线输出相应的专用检索表。考虑到真正有实际用途的一般是种级甚至是亚种级的二次重构, 编码并转化科级以下分类阶元的检索表信息势在必行, 但它需要得到国内各路昆虫分类专家的支持与合作, 才能奠定二次检索表种级重构所必需的数字化基础。

2 其他技术

2.1. 分类信息解析机制

昆虫分类鉴定是一项专业性较强的工作, 为了使一般用户能理解一些较难的特征描述与专业术语, 降低鉴定难度, 本研究中使用了两两种的分类信息解析机制:

2.1.1. 显性解析

帮助理解的文字或图片直接显示在相关信息的旁边上下,如书中的附注、配图一样,读者一眼便能明白这些信息的含义。这种显性解析方式简单明了,广泛用于报刊杂志等平面媒体。本研究由 KeyEditor 维护的特征图片都与对应的文字描述配套显示在鉴定流程、动态交互检索和 Phoenix 检索的网页中。

2.1.2. 隐性解析

隐性信息顾名思义不会直接出现在页面中,但都带有一定的信息隐藏标记,读者在觉得需要深入了解时,通过一定操作便能打开浏览,成为显性信息,这种解析方式是网络化信息普及后才出现的产物。与显性解析相比,它不占用平面空间,易于版面的美观设计和信息层次的有序组织。例如,网页有别于常见的平面媒体,是一种可交互、可扩展的知识媒介,在它表面之外的第二层便是隐性解析发挥的空间。隐性解析在网页中有两种实现方式:超文本和隐藏信息。

超文本(HyperText)是网页的灵魂,是用超链接方法将各种不同空间的文字信息组织在一起的网状文本(百度百科,2007b)。读者点击超文本可打开并阅读与此相关的更多信息页面。由于超文本在显示风格上与普通文本不同,读者很容易识别这些信息的链接,加上它设置方便,调用灵活,已成为信息隐性解析最常用的方式。由 KeyEditor 维护的分类术语知识便以超文本链接的形式穿插在分类单元的特征分布介绍中,用户点击高亮显示的文字,便可查看相关的解释信息。这部分的信息关联是由 KeyCompiler 在生成网页时自动完成。

隐藏信息是指鼠标在图片或者超链接上稍作停留时,以提示(Hint)方式出现在鼠标右下方的信息。它们在 HTML 编码中以 alt=“信息”或者 title=“信息”嵌入在、<a>、<td>等标签中。这种信息解析方式多用于图注、链接内容提示、信息跟踪显示等,读者一旦发现这些信息,便会留下较深刻的印象。

2.2. 多媒体信息版权的保护

系统分类资料是所有生物分类专家的研究成果,享有不容侵犯的知识产权,但这些信息在网上公开发布后,如不加以一定的防护措施,很容易被第三者通过各种电子手段窃取、下载与非法传播。针对网络开放所可能造成的负面影响,论文在设计分类信息网络发布平台 KeyPublisher 时,除了进行必要的版权声明,还提出了多方面的系统控制与页面保护措施,尽可能防止电子分类信息被随意复制滥用。

2.2.1. 登记注明 URL 出处

互联网是极广阔的资源共享平台,在引用他人文字作品时应主动注明 URL 出处,并将网站地址列入参考文献中,这是对原创作者基本的尊重,也便于根据作者要求及时调整或删除这部分信息。网上收集下载一些特征清晰的物种图片时,可在图片保存目录下用一个“说

明”文本文件按图片序号记录引用 URL，如 1: <http://www.jfps.tpc.edu.tw/~yoyo/New23/9207/94.jpg>，KeyCompiler 会将这些信息编译收录在物种图片数据库中，KeyPublisher 在调用这些图片时也会提示其引用地址。

2.2.2. 用户访问权限控制

帐号认证机制可有效监管并控制用户允许访问的内容，防止未授权用户的进入。KeyPublisher 设置了一个用户数据库，通过“permitrange”字段限制用户的访问范围。例如，“permitrange”内容设为 Hexapoda 时用户可以浏览六足总纲及下属所有分类阶元的信息，若设为 Neuroptera，用户则只能查看脉翅目及下属分类阶元的信息，其他目页面被禁止进入。为了全面应用这种访问权限控制，KeyCompiler 在自动编译生成各种信息页面时，在页面头部加入一句代码，如 `<!-- #Include File=../authentication.asp -->`，再将该页面保存为 asp 动态网页。这样服务器在调用信息给用户时，才能执行帐号权限审查功能。

2.2.3. 强制禁止网页复制下载

用户在网上看到有价值的文字或图片时，一般通过三种方法将这些信息保存在本地电脑中，而 KeyCompiler 针对它们都设计了禁用措施，可有效防止文字与图片信息的电子复制传播。

(1) 在选中文字或图片上右击菜单中选择“复制”时，KeyCompiler 使用下面这段 JavaScript 代码禁用文字拖选、禁止右键菜单：

```
<script>
With (document){
oncontextmenu=function(){return false;};
ondragstart=function(){return false;};
onselectstart=function(){return false;};
}
</script>
```

(2) 通过“查看源文件”复制文字信息时，KeyCompiler 将所有重要信息都放在帧(frame)网页中，用户看到的只是父帧网页的源代码；即使用户知道子帧网页的 URL 地址，但由于 KeyCompiler 在每个网页编译时都加入“子帧锁定”代码：

```
<script>if (window == top) top.location.replace('/index.asp');</script>
```

普通用户永远无法单独查看子帧网页，也无法在顶层直接引用该网页。

(3) 通过“文件另存为”下载信息是获取网页所有信息最省力的办法，为了禁止这种信息保留方式，KeyCompiler 在每个网页编译中都加入“干扰代码”：

```
<noscript><iframe src=*.html></iframe></noscript>
```

用户利用正常方式“另存为网页.全部”时，都以“无法保存网页”而失败告终。即使用户改用“另外为网页.仅 HTML”时，得到的网页也只有父帧内容。

3 使用介绍

3.1. 检索表二次重构

检索表二次重构的最大优点是以用户自由选择的物种为目标对象生成鉴定检索表，这些物种可能分布范围相同、危害对象相同或者经济利害相似，甚至是出于个人兴趣或研究需要而圈定，二次检索表都可直接用于这些目的的物种鉴定，而无需求助于纷繁复杂的基础分类检索表。下面以寄生农业害虫稻纵卷叶螟的茧蜂为对象，介绍利用 KeyPublisher 提供的在线重构工具 SIKey 生成二次检索表的过程：

(1) 找出寄生农业害虫稻纵卷叶螟的茧蜂种类。登录“中国昆虫分类鉴定系统 InsectX”网站，在“信息检索”功能模块选中“分布全文”，再以“稻纵卷叶螟”为关键词搜索，可得到 6 个寄生稻纵卷叶螟的茧蜂。


(2) 添加重构对象。点击搜索结果旁边的图标，将该茧蜂加入重构对象组，待所有对象添加完毕后出现图 7.5 的窗口。其中“初始重构”可清空所有重构对象，“添加单元”可输入种名手工添加重构对象。



图 7.5 检索表二次重构工具 SIKey 界面

Fig.7.5 Web interface of SIKey for producing SIKs

(3) 完成二次检索表重构。点击“开始重构”，SIKey 会在新窗口中显示重构得到的新检索表（表 7.5）。整个操作简便快速，轻松完成。

表 7.5 寄生稻纵卷叶螟的茧蜂检索表

Table 7.5 Key to Braconids reared from *Cnaphalocrocis medinalis*

- 1 端跗节(与基跗节比较)宽大;前足跗节第 2-4 节多少变短,通常与基跗节等长或更短;前翅 3-M 脉大部分通常骨化;若缺翅,则头前口式,脸上具一个多少发达突起.....2
- 端跗节正常;前足跗节第 2-4 节正常,通常长于基跗节;前翅 3-M 脉通常大部分不骨化;若缺翅,则触角着生部位正常,不长在脸部的突起上或头部不呈前口式.....3
- 2 腹部第 2-3 背板背方大部分膜状,几乎均比它们的侧板骨化程度低,而且并胸腹节中纵脊短或缺;梗节与柄节约等长,而且(或)胸腹侧脊缺;腹部第 1 背板侧方平坦,通常宽。全世界分布.....稻纵卷叶螟索翅茧蜂(*Hormius Nees moniliatus*)
- 腹部第 2-3 背板与其侧板骨化程度相同或更强,若骨化程度低,则并胸腹节中纵脊长;梗节明显短于柄节,或若相对较长,则胸腹侧脊存在;腹部第 1 背板多少均匀凸起,其侧方部分窄或缺.....眼蝶脊茧蜂(*Aleiodes Wesmæli coxalis* (Spinola))
- 3 前翅 SR1 脉部分或全部不骨化,导致缘室端部开放;腹部常短;后翅扼叶可能大;无缺翅型或短翅型.....5
- 前翅 SR1 脉全部骨化,管状,达翅缘,因而缘室端部关闭;腹部通常长;后翅扼叶通常小;有短翅或缺翅型.....4
- 4 各足第 2 转节前侧(亚)端部具梳状的钉状刺,偶尔后足第 2 转节无钉状刺;腹部着生于并胸腹节的位置稍在后足基节上方;后头脊缺;中胸盾片中叶多少比侧叶凸出。全世界分布.....纵卷叶螟长体茧蜂(*Macrocentrus Curtis cnaphalocrocis* He & Lou)
- 足第 2 转节无钉状刺;腹部着生位置至少部分在后足基节之间;若稍在后足基节上方,则后头脊存在;中胸盾片中叶与侧叶同样凸出.....白跗赛茧蜂(*Zelet albiditarsus*)
- 5 盾纵沟和腹板侧沟光滑;翅烟褐色,具 2 条透明带;并胸腹节中室在前方 1/5 处具中纵脊.....横带折脉茧蜂(*Cardiochiles Nees philippensis*)
- 盾纵沟和腹板侧沟多少具弱刻纹;翅烟褐色,有时色浅,但全翅较一致,不成带状斑;并胸腹节中室在前方无明显的中.....纵脊纵卷叶螟黑折脉茧蜂(*Cardiochiles Nees fuscipennis*)

3.2. 物种直观鉴定训练

由 KeyEditor 建立的物种图片库,除了随特征介绍显示在单阶元信息页面中,也用于物种直观鉴定的学习训练。所谓“直观鉴定”是指一眼看到物种,不需要深入核对特征,便能识别或大概知道其分类地位。专家对自己所研究的分类群落十分熟悉,具有较高的直观鉴定能力,而这种能力正是通过平时多看多观察培养形成的。登录“中国昆虫分类鉴定系统

InsectX”网站后进入“物种图库”模块，选择准备学习的分类单元，如要学习“鳞翅目”昆虫的直观鉴定，即在“鳞翅目”前打勾；点击“学习”开启在线学习窗口(图 7.6)，该窗口中每 5 秒会自动更换显示一张物种图片(图顶注有“科名: 种名”，图尾注有图片引用出处)，让用户不断反复地观察，加深物种名称的视觉印象。如果用户觉得学有所成，可关闭学习窗口，点击“练习”打开在线练习界面。此时窗口中同样显示物种图片，但在图顶等待用户输入物种的科名或种名，对用户输入判断后再返回正确结果。通过“学习—练习”的训练模式，可有效增强用户直观鉴定的判断能力。

3.3. 分类信息检索查询

信息检索是物种多样性数据库最常用的功能，KeyPublisher 为其发布的生物分类系统提供了丰富多样的信息查询方式，以满足用户快速定位物种分类或生物学信息的需要。进入“信息检索”模块，可直接按拉丁学名字母索引浏览所有收录的分类单元，或者按“中文名/拉丁学名+分类级别”、“特征/分布全文+阶元范围”进行组合式查询。在搜索结果(图 7.7)中点击图标即可浏览分类信息或在线使用网络检索表，物种图片库也可支持中文名或拉丁学名的单项查询，全方位快速提供物种的多样性信息。



图 7.6 InsectX 物种图片鉴定学习窗口

Fig. 7.6 Training window of species identification based on pictures in InsectX

输入关键词: 每页显示结果

☒ 中文名 ☐ 拉丁学名 ☐ 图片库 分类级别:

☐ 特征全文 ☐ 分布全文 阶元范围: (如输入Braconidae, 即检索至蜂科所有阶元)

ID	中文名	拉丁学名	介绍	鉴定流程	汇总	二叉检索	交互检索	Phoenix检索
1	<input checked="" type="checkbox"/> 古蠹目	Eosentomata						
2	<input checked="" type="checkbox"/> 古蠹科	Eosentomidae						
3	<input checked="" type="checkbox"/> 长角蠹科	Entomobryidae						
4	<input checked="" type="checkbox"/> 光蠹科	Epylampridae						
5	<input checked="" type="checkbox"/> 壳白蚁属	Enhantermes						
6	<input checked="" type="checkbox"/> 瘦头蠹科	Eupusidae						
7	<input checked="" type="checkbox"/> 姬姬科	Eusepteriidae						
8	<input checked="" type="checkbox"/> 姬姬科	Eumastacoidae						
9	<input checked="" type="checkbox"/> 姬姬科	Eumastacidae						

图 7.7 信息检索主界面
Fig.7.7 Main page of information search engine

3.4. 分类研究专题的定制

分类研究专题是为较高级生物分类研究者提供的专题信息页面，涵盖相关生物分类领域最详细、更新最快的研究成果与进展内容，信息全面集中，专业性强。分类专家可以此作为个人主页平台向网络发布。定制分类研究专题，需首先提供分类群落完整的系统结构、基础分类检索表和物种信息，由 KeyEditor 完成专题信息库的基本构建，并融入整个生物分类系统中；同时再提供研究专家、形态特征、生物习性、地理分布等群落整体方面的介绍，形成专题内容的全面概貌；最后由 KeyCompiler 集成所有分类信息形成内容丰富的物种多样性数据库，由 KeyPublisher 对外发布时嵌入英名、汉名索引和中文名、拉丁学名检索等功能，提供给专题作者唯一的访问 URL，完成专题定制。

第三部分 总讨论

第八章 结论与展望

1 总结

论文以支持和推动生物分类信息资源的数字化建设与网络化发展为目的,借助信息技术的平台,利用网络技术、数据库技术、计算机编程技术、多媒体技术、人工智能技术、专家系统工具等现代信息技术,从研究并解决分类检索表的数字化编码问题起步,开发检索表智能设计与编制技术和二次检索表自动重构再生技术,通过建立 KeyEditor、KeyMaker、KeyImporter、KeyCompiler 和 KeyPublisher 等软件支持工具,实现生物分类检索系统定制开发与物种多样性数据库知识集成的功能,为检索表等生物分类信息的制作、整理、发布、使用与推广提供一套完整的电子化解决方案,并提高检索表的设计效率、科学性和利用率,降低生物分类鉴定系统开发的技术门槛,促进物种多样性信息的共享利用,加速国内生物分类研究与物种鉴定的电子化进程。

论文主要完成了以下技术或工具的设计开发,并建立了相关的生物分类鉴定系统:

1.1. 检索表数字化编码技术

实现生物分类信息数字化,解决检索表统一编码问题是关键。论文在分析了基于规则(类似模拟专家思路)的检索表数字化记录方式的缺点后,提出了采用基于二维的特征分值数字矩阵保留检索表中对象与特征的匹配关系,并用 XML 结构化数据模板完整记录该数字矩阵、对象与特征信息。该模板数据元分离到位,信息可读性强,编辑修改方便,更适合计算机分析处理,深入挖掘扩展功能。它为论文后期开发检索表智能编制与二次重构技术奠定了重要的数据基础。

1.2. 检索表智能编制技术

直接开发数字化检索表是生物分类信息资源数字化建设的主要途径之一。论文以检索表设计基本原则为出发点,在研究分类学中现有的检索表分类算法基础上设计了新的检索表智能编制技术,其原理是根据专家建立的特征与对象关联矩阵,通过“三个度”的智能优化运算输出最佳检索表。经比较验证,该技术科学可靠,优化算法达到期望水准,可代替分类专家智能设计出最佳的检索策略,减少手工编写检索表所花费的时间与精力,并提高检索表设计的效率与合理性。目前基于该核心技术开发的专业二项检索表制作工具 KeyMaker 优化性能达到并高于同类 DELTA Key 软件。

第八章 结论与展望

1 总结

论文以支持和推动生物分类信息资源的数字化建设与网络化发展为目的,借助信息技术的平台,利用网络技术、数据库技术、计算机编程技术、多媒体技术、人工智能技术、专家系统工具等现代信息技术,从研究并解决分类检索表的数字化编码问题起步,开发检索表智能设计与编制技术和二次检索表自动重构再生技术,通过建立 KeyEditor、KeyMaker、KeyImporter、KeyCompiler 和 KeyPublisher 等软件支持工具,实现生物分类检索系统定制开发与物种多样性数据库知识集成的功能,为检索表等生物分类信息的制作、整理、发布、使用与推广提供一套完整的电子化解决方案,并提高检索表的设计效率、科学性和利用率,降低生物分类鉴定系统开发的技术门槛,促进物种多样性信息的共享利用,加速国内生物分类研究与物种鉴定的电子化进程。

论文主要完成了以下技术或工具的设计开发,并建立了相关的生物分类鉴定系统:

1.1. 检索表数字化编码技术

实现生物分类信息数字化,解决检索表统一编码问题是关键。论文在分析了基于规则(类似模拟专家思路)的检索表数字化记录方式的缺点后,提出了采用基于二维的特征分值数字矩阵保留检索表中对象与特征的匹配关系,并用 XML 结构化数据模板完整记录该数字矩阵、对象与特征信息。该模板数据元分离到位,信息可读性强,编辑修改方便,更适合计算机分析处理,深入挖掘扩展功能。它为论文后期开发检索表智能编制与二次重构技术奠定了重要的数据基础。

1.2. 检索表智能编制技术

直接开发数字化检索表是生物分类信息资源数字化建设的主要途径之一。论文以检索表设计基本原则为出发点,在研究分类学中现有的检索表分类算法基础上设计了新的检索表智能编制技术,其原理是根据专家建立的特征与对象关联矩阵,通过“三个度”的智能优化运算输出最佳检索表。经比较验证,该技术科学可靠,优化算法达到期望水准,可代替分类专家智能设计出最佳的检索策略,减少手工编写检索表所花费的时间与精力,并提高检索表设计的效率与合理性。目前基于该核心技术开发的专业二项检索表制作工具 KeyMaker 优化性能达到并高于同类 DELTA Key 软件。

1.3. 检索表二次重构技术

论文开发的检索表二次重构技术是基于检索表数字化编码方案实现的检索表自动创建技术,其原理是利用丰富的基础分类检索表数据,通过分类单元与鉴定特征间匹配关系的反向推理,自动搜寻并重组检索路线,可理解为“表生表”的过程。该技术角度新颖,功能实用,用户不用求助分类专家,根据需要自行定制重构对象,便很快得到只含指定分类单元和必要特征的二次检索表,大大简化了依据传统分类检索表鉴定的复杂繁琐过程,在保证同等正确性的前提下提高鉴定工作效率。目前 KeyPublisher 已内置检索表重构工具 SIKey,基于 InsectX 现有的多样性数据库便可支持定制昆虫科及以上分类单元、蜜蜂科所有已知种进行二次检索表重构。

1.4. 生物分类鉴定知识系统开发工具

由 KeyEditor、KeyCompiler 和 KeyPublisher 组成的生物分类鉴定知识系统开发工具,是提供给分类专家、多样性工作者等实现分类信息数字化,提供网络信息服务的重要平台。用户只需在 KeyEditor 中完成分类资料的收集与整理,其他如网页制作、数据库集成和网站发布等技术性强、难度较高的工作都由 KeyCompiler 和 KeyPublisher 自动快速完成。生成的系统内建系统分类树、辅助鉴定、二次检索表重构、物种直观鉴定训练、信息检索查询等功能,完全能满足各种分类研究与学习的需要。今后整个系统的维护与更新也十分方便。

1.5. 物种多样性数据库开发工具

由 KeyEditor 和 KeyCompiler 组成的物种多样性数据库开发工具,专门收录生物类群的分类地位、系统关系、鉴别特征、地理分布、标本信息、物种图片等丰富信息。这些信息在 KeyEditor 中收集齐全后经 KeyCompiler 编译成结构化关系型的基本信息库和物种图片库,作为可移植的基础知识库,用于第三方分类软件或系统的开发,今后只需重新编译便可更新数据库。目前由该工具开发的蜜蜂科多样性数据库已收录分类阶元 731 个,蜜蜂 569 种,检索表 128 个,是国内蜜蜂分类信息最全的多样性数据库。

1.6. 基于 Web 的中国昆虫鉴定分类系统 InsectX

基于上述技术与工具建立的中国昆虫鉴定分类系统 InsectX,是国内首个全面介绍昆虫物种多样性的网站,以郑乐怡和归鸿(1999)介绍的昆虫分类系统为信息框架,从六足总纲到科涉及分类阶元 1 873 个(现蜜蜂科已到种,含 569 个),检索表 984 个,物种图片 2 343 张,特征图片 284 张,覆盖名称、特征、分布、鉴定流程、术语释解、检索表等信息,是国内分类系统最完善、内容最全面的生物资源数据库之一。系统还提供二叉跳转、动态交互、Phoenix 等多种网络检索表在线使用、自定义物种群的二次检索表重构再生、分类信息检索查询、物

种直观鉴定训练、分类专家个人网站定制等功能,是昆虫分类鉴定与科普教学的重要电子工具(张小斌等,2006b)。

1.7. 基于 Wap 的植物检疫性昆虫信息平台 W-QPM

论文开发的植物检疫性昆虫信息平台 W-QPM,是基于无线互联网的移动式专家系统,收录了 203 种植物检疫性昆虫(孙冠英,2003)信息,涵盖分类地位、检疫特征、来源产地、寄主植物、物种图片等,并提供这些物种快速的多途径检索功能。检疫人员在海港口岸等地可通过各种手机、PDA 等便携式移动终端实时访问,查询检疫信息,在线咨询专家,辅助决策鉴定。W-QPM 是 3G 网络时代来临之前,无线互联网环境下开发生物分类鉴定系统的一次有益尝试(张小斌和程家安,2006)。

2 讨论

2.1. 加强跨学科领域的知识技术交流与合作

生物专家与计算机工程师是属于两个完全不同的学科领域,但生物专家编制各类电子化系统,离不开计算机技术的支持;计算机工程师设计好的生物应用软件,也离不开生物背景知识的指导。虽然目前网络发达,信息流通方便,但双方的知识视野和思维结构差异较大,若各置一域“自给自足”地单干,或只是表面性地简单合作,仍容易导致生物专家永远不会用更先进更合适的计算机技术编写应用系统,也可能一番尝试后发现技术行不通,或者实际人为工作量太大难以实现而无果告终;计算机工程师开发的生物软件功能也不能真正地贴近专业或实际的要求,造成人力与投资的浪费。双方只有从共识点出发,加强知识与技术的交流,促使彼此的研究与开发都能双赢获益,才能真正促进交叉学科的发展与成果创新。

2.2. 加大对优秀生物软件的宣传与推广力度

随着计算机技术在各学科领域的不断应用,产生了许多生物应用软件。许多当时较优秀的软件因缺乏必要的业界认同和研发经费支持,停滞改进或销声匿迹,造成了开发与应用的严重断层。这说明软件本身的宣传与应用推广力度不够,缺乏必要的用户群才日渐失去价值。其实生物软件与整个软件业一样,形成良好的产业链循环机制才能长久发展(李广友,2006)。软件开发完成之后,应积极提倡在所属领域的知名刊物上发表宣传,并在相关学术交流会议上主动推荐介绍,打响产品的知名度和影响力;或者直接交给更有推广能力与经验的第三方单位负责软件的宣传,再由他们收集使用建议返回改进。生物软件只有通过各种渠道的推广,把技术尽快转化为生产力,并占据一定的应用市场,才能持续稳定地发展,增强与国外同类软件的竞争力。而多样性数据库的构建,更需要一个权威或者有广泛影响力的单位领头策划

种直观鉴定训练、分类专家个人网站定制等功能,是昆虫分类鉴定与科普教学的重要电子工具(张小斌等,2006b)。

1.7. 基于 Wap 的植物检疫性昆虫信息平台 W-QPM

论文开发的植物检疫性昆虫信息平台 W-QPM,是基于无线互联网的移动式专家系统,收录了 203 种植物检疫性昆虫(孙冠英,2003)信息,涵盖分类地位、检疫特征、来源产地、寄主植物、物种图片等,并提供这些物种快速的多途径检索功能。检疫人员在海港口岸等地可通过各种手机、PDA 等便携式移动终端实时访问,查询检疫信息,在线咨询专家,辅助决策鉴定。W-QPM 是 3G 网络时代来临之前,无线互联网环境下开发生物分类鉴定系统的一次有益尝试(张小斌和程家安,2006)。

2 讨论

2.1. 加强跨学科领域的知识技术交流与合作

生物专家与计算机工程师是属于两个完全不同的学科领域,但生物专家编制各类电子化系统,离不开计算机技术的支持;计算机工程师设计好的生物应用软件,也离不开生物背景知识的指导。虽然目前网络发达,信息流通方便,但双方的知识视野和思维结构差异较大,若各置一域“自给自足”地单干,或只是表面性地简单合作,仍容易导致生物专家永远不会用更先进更合适的计算机技术编写应用系统,也可能一番尝试后发现技术行不通,或者实际人为工作量太大难以实现而无果告终;计算机工程师开发的生物软件功能也不能真正地贴近专业或实际的要求,造成人力与投资的浪费。双方只有从共识点出发,加强知识与技术的交流,促使彼此的研究与开发都能双赢获益,才能真正促进交叉学科的发展与成果创新。

2.2. 加大对优秀生物软件的宣传与推广力度

随着计算机技术在各学科领域的不断应用,产生了许多生物应用软件。许多当时较优秀的软件因缺乏必要的业界认同和研发经费支持,停滞改进或销声匿迹,造成了开发与应用的严重断层。这说明软件本身的宣传与应用推广力度不够,缺乏必要的用户群才日渐失去价值。其实生物软件与整个软件业一样,形成良好的产业链循环机制才能长久发展(李广友,2006)。软件开发完成之后,应积极提倡在所属领域的知名刊物上发表宣传,并在相关学术交流会议上主动推荐介绍,打响产品的知名度和影响力;或者直接交给更有推广能力与经验的第三方单位负责软件的宣传,再由他们收集使用建议返回改进。生物软件只有通过各种渠道的推广,把技术尽快转化为生产力,并占据一定的应用市场,才能持续稳定地发展,增强与国外同类软件的竞争力。而多样性数据库的构建,更需要一个权威或者有广泛影响力的单位领头策划

并有序地开展,才能带来数据库发布后的广泛关注和应用,避免重复性建设。

2.3. 国内分类软件开发注重方法论的学习与应用

软件应用的介入给传统生物分类注入了新活力,20世纪70年代以来国外涌现了不少优秀的分类软件,如 DELTA、PANKEY、Lucid、Linnaeus II 等,它们推动了检索表的电子设计、交互式与多途径检索的广泛使用。而国内这方面研究大多只在检索系统开发的层面,二叉推理检索技术的应用十分普遍,但国外软件中文支持差等问题依旧令人诟病。出现这一落差,原因可能在于我们容易停留在计算机技术本身的应用,而没有注重对其他基础学科方法论的吸收和借鉴,而外国人只是把计算机视为电子工具,能及时地将人工智能、数理统计等其他领域中的新方法与新技术应用在检索表开发与鉴定上,由此产生了新思路和新软件(Payne and Preece, 1980)。因此,我们在观察洞悉国外生物软件发展方向的同时,自身也应加强分类方法与理论的学习,实现软件创新。

2.4. 提倡直接利用现有的电子化分类信息

生物分类发展已有较长历史,积累了一大批宝贵的分类学文献著作。由于纸质、印刷质量、中文识别率等客观因素影响,InsectX 在前期整理昆虫科类数字化信息时,花费了大量时间在文字扫描和识别校对上。而后期制作蜜蜂科分类子系统时,直接从专家手中拿到了第一手的电子材料,节省了大量机械式的重复劳动,加快了开发进度。由此可见分类学家直接的信息支持对于生物分类系统开发的重要性,在系统初期设计定义时,应主动与有关专家接洽索取电子版本,或者查寻各类数字图书馆,充分使用现成的电子材料,尽量避免重复建设劳动。

3 今后进一步研究

3.1. 增强分类软件与系统的功能

论文完成的生物分类鉴定知识系统开发工具曾在中国科学院动物所公开演示介绍,收到了一些有指导意义的反馈建议。如利用地理信息系统技术动态显示物种的地理分布图,提供形象直观的认识;物种图库扩增“模式标本”图片,提高特征核对与新种判断的可信度;多样性数据库增加参考文献的字段,注明分类信息的引用来源,有案可考;支持输出 EXCEL 等数据格式的分布信息,可用于物种区系分析等用途。这些都有待今后在系统逐步完善与改进中得以实现。

并有序地开展,才能带来数据库发布后的广泛关注和应用,避免重复性建设。

2.3. 国内分类软件开发注重方法论的学习与应用

软件应用的介入给传统生物分类注入了新活力,20世纪70年代以来国外涌现了不少优秀的分类软件,如 DELTA、PANKEY、Lucid、Linnaeus II 等,它们推动了检索表的电子设计、交互式与多途径检索的广泛使用。而国内这方面研究大多只在检索系统开发的层面,二叉推理检索技术的应用十分普遍,但国外软件中文支持差等问题依旧令人诟病。出现这一落差,原因可能在于我们容易停留在计算机技术本身的应用,而没有注重对其他基础学科方法论的吸收和借鉴,而外国人只是把计算机视为电子工具,能及时地将人工智能、数理统计等其他领域中的新方法与新技术应用在检索表开发与鉴定上,由此产生了新思路和新软件(Payne and Preece, 1980)。因此,我们在观察洞悉国外生物软件发展方向的同时,自身也应加强分类方法与理论的学习,实现软件创新。

2.4. 提倡直接利用现有的电子化分类信息

生物分类发展已有较长历史,积累了一大批宝贵的分类学文献著作。由于纸质、印刷质量、中文识别率等客观因素影响,InsectX 在前期整理昆虫科类数字化信息时,花费了大量时间在文字扫描和识别校对上。而后期制作蜜蜂科分类子系统时,直接从专家手中拿到了第一手的电子材料,节省了大量机械式的重复劳动,加快了开发进度。由此可见分类学家直接的信息支持对于生物分类系统开发的重要性,在系统初期设计定义时,应主动与有关专家接洽索取电子版本,或者查寻各类数字图书馆,充分使用现成的电子材料,尽量避免重复建设劳动。

3 今后进一步研究

3.1. 增强分类软件与系统的功能

论文完成的生物分类鉴定知识系统开发工具曾在中国科学院动物所公开演示介绍,收到了一些有指导意义的反馈建议。如利用地理信息系统技术动态显示物种的地理分布图,提供形象直观的认识;物种图库扩增“模式标本”图片,提高特征核对与新种判断的可信度;多样性数据库增加参考文献的字段,注明分类信息的引用来源,有案可考;支持输出 EXCEL 等数据格式的分布信息,可用于物种区系分析等用途。这些都有待今后在系统逐步完善与改进中得以实现。

3.2. 进一步完善智能优化技术

论文设计的检索表智能优化技术,已完全能胜任电子检索表的自动编制,输出的结果也达到了较高的优化水平。但在实测中已发现,该技术还有两个重要的性能提升空间。首先,当特征与对象二项矩阵维数较大时,优化算法可能需要较长时间枚举备选检索表,甚至引起内存溢出。虽然软件本身已能通过限制层节点数和枚举结果避免此种情况的发生,但实际上它以牺牲部分备选结果(视参数设置而定)为代价,并不十分可取。假如编程时采用多线程运算技术加速检索表的枚举,并采用硬盘缓存代替内存负载,可以较好地缓解上述问题。不过这需要对编程策略大幅改动,有待今后深入研究。其次,目前对象的每个特征只能单个取值,这种限制对于性状多样性变化大的物种,特征匹配会有所不便,解决办法是改进特征分枝法以支持多特征值的存储。这已牵涉检索表电子化的设计原理,并将引起后续设计的一连串变动。不过这是最行之有效的办法,值得一试。

3.3. 建设二次重构的基础分类库

二次检索表检索快捷,使用方便,但要随心所欲地得到任何已知物种的二次检索表,前提是以事先编码转化好的基础分类数据作为重构的基础。就昆虫而言,数量众多,分类系统复杂,要完成现有昆虫分种基础检索表的信息编码,将是十分庞大和艰巨的任务。考虑到真正有实际鉴定用途的一般是种级甚至是亚种级的二次重构,编码并转化这些分类单元的检索表信息仍势在必行。因此,本论文呼吁国内各路昆虫分类专家的支持与合作,共同加入昆虫基础分类检索表电子信息化的工作,以最终奠定二次检索表种级重构所必需的基础知识库。届时,它将成为基于各种应用需求而自定义物种鉴定的指导工具,具有重要的利用价值和深远意义。

3.4. 扩充网页模板,增添界面美观

网页是网络系统中最主要的信息载体,KeyCompiler 在编译知识库时,统一采用预先设计的模板,填入数据库中存储的分类信息,生成大量网页。虽然这些网页整齐划一,但由于出自同一模板风格,界面整体上单一欠雅。按照自助式网站设计惯例,一般会设计多套不同主题、色彩各异的模板,以便在网页批量制作或更新时灵活更换使用,增强网页的变化性与美感。一个网站只有内容丰富实用,界面和谐雅致,才能吸引更多人的目光,在网上流行开来。因此 KeyCompiler 今后也应增设多套简洁雅观的网页模板,实现上述“换肤”功能。

3.5. 推进本工具的国际化发展和应用

由于国外分类软件与系统的开发已走在前列,其制定的分类标准、编码规范已被国际广泛认同接受,如 DELTA 已被定为分类学描述语言编码的国际标准。为了与 DELTA 格式兼容,

与国际软件接轨，本工具今后将致力开发 DELTA 数据接口（高灵旺等，2003），支持 KeyImporter 导入、KeyEditor 输出 DELTA 格式的分类文件，以便能吸收丰富的 DELTA 分类信息为己所用，并利用其他软件分析处理分类信息。此外，本工具还将设计英文版或支持多国语言显示，为软件国际化推广应用奠定基础。

参考文献

- Bisby A F. The quiet revolution: biodiversity informatics and the internet [J]. *Science*, New Series, 2000, 289(5488): 2309-2312.
- Brach A R, Song H. ActKey: a Web-based interactive identification key program [J]. *Taxon*, 2005, 54(4): 1041-1046.
- Breiman L, Friedman J H, Olshen R A, et al. Classification and regression trees [R]. Monterey, California: Wadsworth International Group, 1984.
- Calvo-Flores M D, Contraras W F, Gibaja Galindo E L, et al. XKey: A tool for the generation of identification keys [J]. *Expert Systems with Applications*, 2006, 30: 337-351.
- CBIT. LucidCentral: Lucid Professional. <http://www.lucidcentral.com/Lucid2> [EB/OL], 2007a-4-5.
- CBIT. LucidCentral: Lucid Phoenix. <http://www.lucidcentral.org/phoenix> [EB/OL], 2007b-3-30.
- CBIT. LucidCentral: Lucid 3. <http://www.lucidcentral.com/Lucid3> [EB/OL], 2007c-4-27.
- CBIT. LucidCentral: Lucid Translator. http://www.lucidcentral.org/lucid3/lucid_translator.htm [EB/OL], 2007d-4-28.
- Chao-Hsuan Ke. Machine learning-decision tree. <http://bioinfo.ec.kuas.edu.tw/news/file/sample.ppt> [EB/OL], 2007-4-5.
- Coates V T. The future of information technology [J]. *Annals of the American Academy of Political and Social Science*, 1992, 522: 45-56.
- Dallwitz M J, Paine T A, Zurcher E J. How the program selects characters. http://www.delta-intkey.com/www/uguide.htm#_5.4_How_the [EB/OL], 2006a-10-9.
- Dallwitz M J, Paine T A, Zurcher E J. The interactive identification program Intkey. http://www.delta-intkey.com/www/uguide.htm#_7._The_Interactive [EB/OL], 2006b-10-9.
- Dallwitz M J. A comparison of matrix-based taxonomic identification systems with rule-based systems [C]. *Proceedings of IFAC Workshop on Expert Systems in Agriculture*. Beijing: International Academic Publishers, 1992: 215-218.
- Dallwitz M J. A comparison of matrix-based taxonomic identification systems with rule-based systems. <http://delta-intkey.com/www/expertid.htm> [EB/OL], 2007a-4-6.
- Dallwitz M J. A flexible computer program for generating identification keys [J]. *Systematic Zoology*, 1974, 23: 50-57.

- Dallwitz M J. A general system for coding taxonomic descriptions [J]. *Taxon*, 1980, 29: 41-46.
- Dallwitz M J. Data requirements for natural-language descriptions and identification. <http://delta-intkey.com/www/descdata.htm> [EB/OL], 2005-9-13.
- Dallwitz M J. Programs for interactive identification and information Retrieval. <http://delta-intkey.com/www/idprogs.htm> [EB/OL], 2007b-4-7.
- Dodds L. XDELTA—Deriving an XML based format for taxonomic information. <http://www.ldodds.com/delta> [EB/OL], 1999-10-22.
- Duncan T, Meacham C A. Multiple-entry-keys for the identification of angiosperm families using a microcomputer [J]. *Taxon*, 1986, 35: 492-494.
- Edwards J L, Lane M A, Nielsen E S. Interoperability of biodiversity databases biodiversity information on every desktop [J]. *Science*, New Series, 2000, 289(5488): 2312-2314.
- Edwards M, Morse D R. The potential for computer-aided identification in biodiversity research [J]. *Tree*, 1995, 10(4): 153-158.
- Estep K W, Hasle A, Omli L, et al. Linnaeus: interactive taxonomy using the Macintosh computer and hypercard [J]. *BioScience*, 1989, 39(9): 635-638.
- ETI. LinnaeusII. <http://www.eti.uva.nl/products/linnaeus.php> [EB/OL], 2007-4-5.
- Exeter. PANKey, programs for identification. <http://www.exetersoftware.com/cat/pankey/pankey.html> [EB/OL], 2007-4-5.
- Fronsdorf A, Waggoner G. Systematics information as a central component in the National Biological Information Infrastructure [J]. *Annals of the Missouri Botanical Garden*, 1996, 83: 546-550.
- Goodall D W. Identification by computer [J]. *BioScience*, 1968, 18: 485-488.
- Hall A V. A computer-based system for forming identification [J]. *Taxon*, 1970, 19(1): 12-18.
- Intelsys Inc. XID Authoring System. <http://www.xidservices.com> [EB/OL], 2001.
- Jensen R. PANKEY [J]. *The Quarterly Review of Biology*, 1990, 65(4): 538-539.
- Kaiser J. NetWatch: Animals, animals [J]. *Science*, New Series, 1999, 285(5434): 1635.
- Kononenko I, Bratko I, Roskar E. Experiments in automatic learning of medical diagnostic rules [R]. Ljubljana: Jozef Stefan Institute, 1984.
- Lancaster F W. Libraries and librarians in the age of electronics [M]. Arlington, VA: Information Resource Press, 1982.
- Maddison D R, Swofford D L, Maddison W P. NEXUS: An extensible file format for systematic information [J]. *Systematic Biology*, 1997, 46(4): 590-621.
- Marine Biological Laboratory. X: ID. <http://uio.mbl.edu/services/key.html> [EB/OL], 2004-6.

- Meacham C A. Meka. <http://ucjeps.berkeley.edu/meacham/meka> [EB/OL], 2005-8-22.
- Metcalf Z P. The construction of keys [J]. *Systematic Zoology*, 1954, 3: 38-45.
- Möller F. Quantitative methods in the systematics of actinomycetales. IV. The theory and application of a probabilistic identification key [J]. *Giornale Di Microbiologia*, 1962, 10: 29-47.
- Morse L E. Specimen identification and key construction with time-sharing computers [J]. *Taxon*, 1971, 20(2/3): 269-282.
- Natural History Museum. British bumblebee identification guide. <http://www.nhm.ac.uk/nature-online/life/insects-spiders/bumblebee-id/british-bumblebee-identification-guide.html> [EB/OL], 2007-4-3.
- Norris S. A year for biodiversity [J]. *BioScience*, 2000, 50(2): 103-107.
- Pankhurst R J, Walters S M. Key generation by computer, in data processing in biology and geology, for systematics association [M]. Academic Press, 1971: 189-203.
- Pankhurst R J. A computer program for generating diagnostic keys [J]. *The Computer Journal*, 1970a, 13(2): 145-151.
- Pankhurst R J. An interactive program for the construction of identification keys [J]. 1988, *Taxon*, 37(3): 747-755.
- Pankhurst R J. Botanical keys generated by computer [J]. *Watsonia*, 1971, 8: 357-368.
- Pankhurst R J. Key generation by computer [J]. *Nature*, 1970b, 227: 1269-1270.
- Pankhurst R J. Practical taxonomic computing [M]. Cambridge University Press, 1991.
- Payne R W, Preece D A. Identification keys and diagnostic tables: a review [J]. *Journal of the Royal Statistical Society. Series A (General)*, 1980, 143(3): 253-292.
- Payne R W, Thompson C J. A study of criteria for constructing identification keys containing tests with unequal costs [J]. *Computational Statistics Quarterly*, 1989, 1: 43-52.
- Pennisi E. Preparing the ground for a modern 'Tree of Life' [J]. *Science, New Series*, 2001, 293(5537): 1979-1980.
- Pomar J, Hidalgo I. An intelligent multimedia system for identification of weed seedlings [J]. *Computers and Electronics in Agriculture*, 1998, 19: 249-264.
- Quinlan J R. Induction of decision trees [J]. *Machine Learning*, 1986, 1: 81-106.
- Reynolds A P, Dicks J L, Roberts I N, et al. Algorithms for identification key generation and optimization with application to yeast identification [J]. *Lecture Notes in Computer Science*, 2003, 2611: 107-118.
- Schäfer K, Goergen G, Borgemeister C. An illustrated identification key to four different species of adult *Dinoderus* (Coleoptera: Bostrichidae), commonly attacking dried cassava chips in West

- Africa [J]. *Journal of Stored Products Research*, 2000, 36: 245-252.
- Schalk P H, Oosterbroek P. Interactive knowledge systems: meeting the demand for disseminating up-to-date biological information [J]. *Biodiversity Letters*, 1996, 3(4/5): 119-123.
- Shayler H A, Siver P A. Key to freshwater algae: a web-based tool to enhance understanding of microscopic biodiversity [J]. *Journal of Science Education and Technology*, 2006, 15(3): 298-303.
- Snow N. Lucid Professional for windows: contemporary identification tools [J]. *Systematic Biology*, 1999, 48(4): 828-830.
- TDWG-SDD. SDD part 0: Introduction and primer to the SDD standard. <http://160.45.63.11/Projects/TDWG-SDD/Primer/index.htm> [EB/OL], 2003-12-31.
- University of Bayreuth, Department of Mycology. NaviKey Home. <http://www.navikey.net> [EB/OL], 2007-3-15.
- University of Toronto Department of Botany. PolyClave. <http://prod.library.utoronto.ca:8090/polyclave> [EB/OL], 1996-10-10.
- Watson L, Milne P. A flexible system for automatic generation of special purpose dichotomous keys, and its application to Australian grass genera [J]. *Australian Journal of Botany*, 1972, 20(3): 331-352.
- Biodata. 昆虫数字标本库. <http://www.biodata.cn/site/show/昆虫数字标本> [EB/OL], 2007-4-6.
- CNNIC. 中国互联网络发展状况统计报告. <http://www.cnnic.cn/uploadfiles/doc/2006/7/19/103601.doc> [EB/OL], 2006-7-19.
- EC 网络. 动物物种多样性数据库. <http://life.zsu.edu.cn/animal/index.php> [EB/OL], 2007-4-6.
- Han J, Kamber M. 数据挖掘概念与技术 [M]. 北京: 机械工业出版社, 2001: 188-196.
- Tchen. 数据挖掘之五: 分类规则法. <http://www.tribo.nfu.edu.tw/~tchen/DataMining2/ch5.ppt> [EB/OL], 2007-4-5.
- W3CHINA. XML. <http://www.xml.org.cn/index.html> [EB/OL], 2007-3-29.
- 百度百科. PPT 模版与母版的作用和区别. <http://zhidao.baidu.com/question/18163561.html> [EB/OL], 2007a-3-30.
- 百度百科. 超文本. <http://baike.baidu.com/view/156868.htm> [EB/OL], 2007b-3-30.
- 百度百科. 术语. <http://baike.baidu.com/view/168249.htm> [EB/OL], 2007c-3-30.
- 陈乃中, 沈佐锐. 一种计算机昆虫检索系统的制作方法 [J]. *植物检疫*, 2003, 17(1): 20-21.
- 陈云樱, 吴积钦, 徐可佳. 决策树中基于基尼指数的属性分裂方法 [J]. *微机发展*, 2004, 14(5): 66-68.
- 迟德富, 孙凡, 严善春, 等. 保护生物学. <http://jpkc.nefu.edu.cn/bhswx/分栏/相关教材/保护生物>

- 学.doc [EB/OL], 2006-5-19.
- 崔海东. 互联网建设十年回顾 [J]. 电信工程技术与标准化, 2005, 9: 1-5.
- 高灵旺, 沈佐锐, 刘志琦, 等. 基于二叉分类推理的昆虫分类辅助鉴定多媒体专家系统通用平台 TaxoKeys 的设计与开发 [J]. 昆虫学报, 2003, 46(5): 644-648.
- 管致和. 昆虫学通论上册 [M]. 北京: 中国农业出版社, 1999: 128-133.
- 胡奇, 马吉祥. 用计算机进行昆虫分类检索研究初探 [J]. 昆虫知识, 1990, 27(1): 40-44.
- 黄复生, 朱世模, 平正明, 等. 中国动物志·昆虫纲第十七卷·等翅目 [M]. 北京: 科学出版社, 2000, 159-889.
- 黄复生. 昆虫种类数量的变化 [J]. 昆虫知识, 1991, 28(6): 374.
- 纪力强. 生物多样性信息系统建设的现状及 CBIS 简介 [J]. 生物多样性, 2000, 8(1): 41-49.
- 纪力强. 中国动物物种编目数据库. <http://zd1.brim.ac.cn/speciessrch.asp> [EB/OL], 2007-4-6.
- 蒋齐. 用 PC-1500 计算机进行昆虫分类检索表的编写及应用 [J]. 昆虫知识, 1991, 28(2): 118-119.
- 金瑞华, 王心丽, 张家娴. 苹果蠹蛾及其近似种成虫彩色图解式检索表 [J]. 植物保护, 1996, 22(1): 49.
- 李广友. 软件平台化推动中国软件产业链发展 [J]. 程序员, 2006, 10: 38.
- 李健钧. 处理植物分类学描述语言的国际标准—DELTA 系统 [J]. 植物分类学报, 1996, 34(4): 447-452.
- 李宁, 王姣. 全球生物多样性的减少与对策 [J]. 国土与自然资源研究, 2006, 4: 73-74.
- 林丹红. 信息资源数字化建设的设想与探索—谈中医药文献资源建设 [J]. 情报探索, 2001, 80(4): 1-4.
- 刘阳, 丁银燕. 论图书馆信息资源数字化建设 [J]. 图书馆工作与研究, 2002, 107(2): 22-24.
- 卢慧甍, 黄原. 中国蝗总科分类、查询及鉴定专家系统 (ESCA) 设计与实现 [J]. 动物分类学报, 2003, 28(3): 428-433.
- 乔凤海. 医学信息的组成与分类. <http://www.chis.com.cn/新世纪/医院/第二篇> 医学信息的组成与分类.htm [EB/OL], 2000-9.
- 丘耘. 知识系统开发工具用户操作手册 [EB]. 北京: 中国农业科学院农业信息研究所多媒体技术研究室, 2006.
- 沈爱华, 唐启义, 程家安. 基于二叉分类检索表正、反向推理的研究及应用 [J]. 浙江大学学报 (农业与生命科学版), 2006, 32(5): 541-545.
- 沈被娜, 刘祖照, 姚晓冬. 计算机软件技术基础 [M]. 北京: 清华大学出版社, 2000: 6-8.
- 隋艳晖, 徐洪富, 孙淑君. 昆虫发声行为的研究现状 [J]. 山东农业大学学报 (自然科学版),

- 2003, 34(3): 443-446.
- 孙冠英, 陈学新, 程家安. Lucid: 多途径的分类检索和诊断专家系统 [J]. 动物分类学报, 2002, 27(4): 871-875.
- 孙冠英. 基于网络的进出境植物检疫信息管理和辅助决策系统 [D]. 杭州: 浙江大学, 2003
- 孙细明, 张晓鹏. 基于信息增益的决策树算法实现 [J]. 计算机与数字工程, 2005, 33(11): 94-95, 121.
- 王心丽, 万霞, 鲍荣, 等. 蚁蛉亚科电子图文链接式分族检索表 [J]. 昆虫知识, 2004, 41(2): 174-176.
- 微软中国. 了解 XML. <http://www.microsoft.com/china/MSDN/library/archives/library/dnXML/html/understXML.asp> [EB/OL], 2007-3-29.
- 吴焰玉, 汪家社. 简体中文昆虫学国际互联网站介绍与评述 [J]. 昆虫知识, 2001, 38(3): 236-237.
- 徐晓国, 莫建初, 程家安. 基于 Web 的等翅目昆虫分类系统的设计与开发 [J]. 昆虫分类学报, 2004, 26(2): 86-90.
- 徐柱, Dallwitz M J, Watson L. 计算机产生中英文植物分类检索表 [J]. 中国草地, 1992, 1: 53-57.
- 杨明, 张载鸿. 决策树学习算法 ID3 的研究 [J]. 微机发展, 2002: 5.
- 杨晓农. 我国文献信息数字化技术的发展 [J]. 中国信息导报. 2004, 5: 32-33.
- 姚青, 赖凤香, 傅强, 等. 多功能昆虫鸣声信号采集和分析系统及其在褐飞虱鸣声研究中的应用 [J]. 中国水稻科学, 2001, 18(2): 171-175.
- 于衍平. 信息数字化对科学认识的拓展 [J]. 自然辩证法研究, 1997, 13(12): 29-33.
- 张小斌, 陈学新, 程家安. Lucid Phoenix: 交互式多媒体网络检索表工具 [J]. 昆虫分类学报, 2006a, 28(3): 231-234.
- 张小斌, 陈学新, 程家安. 基于 Web 的中国昆虫科级鉴别分类系统 InsectID 的设计与开发 [J]. 昆虫分类学报, 2006b, 28(1): 63-68.
- 张小斌, 程家安. 移动网络环境下的植物检疫性昆虫信息平台 W-QPM [J]. 浙江大学学报(农业与生命科学版), 2006, 32(4): 406-409.
- 郑乐怡, 归鸿. 昆虫分类 [M]. 南京: 南京师范大学出版社, 1999: 1-1070.
- 周云龙. 如何鉴定生物对象. <http://www.nwnu.edu.cn/sky/jpkc/zhifu/shixizhidao/5.php> [EB/OL], 2007-4-5.
- 朱建秋. 数据挖掘的背景知识. <http://javainsight.bokee.com> [EB/OL], 2007-4-5.

致谢

本论文是在导师程家安教授和陈学新教授的悉心指导下完成。从论文的规划、设计、实施以至撰写、修改的各方面,无不凝聚着导师大量的精力和心血。基于数字化的生物分类信息开发与应用是顺应全球数字化信息建设潮流开展的研究课题,在整个完成过程中,程老师一直给予我莫大的支持与鼓励。他安排我赴澳参与检索软件合作开发,赴京宣传推广论文成果,这些宝贵而难得的机会令我大开视野,增进阅历。基于检索表的成功开发得益于陈老师给予我的设计灵感和研究思路,他对基础分类研究的执着追求和敬业精神激发了我对课题的研究热情,尤其在访澳期间一路的关怀照顾,令我对澳洲一行留下了毕生难忘的印象。五年来,两位导师不仅在学习和工作上言传身教,支持帮助,而且在生活上给予无微不至的关怀。在此谨向他们表示由衷的敬意和万分的感谢!

同时,本所的莫建初教授和蒋明星教授对我平时的生活学习给予贴心的关怀与照顾,唐启义研究员对分类软件的统计算法和程序编制给予许多宝贵的建议和帮助,胡萃教授、何俊华教授、刘树生教授、叶恭银教授、娄永根教授、祝增荣教授、徐志宏教授、施祖华教授、张传溪教授等给予诸多方面的热情指导与关心;马永芳、马云、芮开宁、吴晓晶、袁熙贤等老师多年来对我助管和网管兼职工作给予热心的支持与协作。在此向他们表示诚挚的感谢!

此外,已毕业的孙冠英、徐晓国、滕立、杨天赐、陈华才、林欣大、张素芳、姜永厚、施婉君、张时妙、何黄英、王伟、王六彩、孙传恒、李俊敏、姚刚、金铃等,在研的潘程远、程梦林、方军、邓天福、王争艳、叶敏、吴琼、伍代胥等,大学同学施卫兵、朱旭华、蒋跃平、朱洁、方丽,研究生室友张小全、王福民、刘汾阳、张国忠、刘翔、叶健等在五年的学习与生活过程中给予许多成面的帮助照顾、支持合作,在此也向他们表示衷心的感谢!

感谢我的父母等家人一直以来对我的关心、爱护、理解和支持!

感谢李拓、江盛鸿、陈科、廖白璐等四年来曾支持过音乐滑板校园网站的所有合作伙伴和校友们!

感谢所有与我一起分享快乐和忧愁的所有未提及姓名的朋友们和同学们!

感谢参与本论文评审和答辩的所有老师和同学们!

张小斌

2007年4月于杭州.浙江大学华家池校区

博士期间发表的学术论文和研究成果

- Zhang X B, Chen X X, Cheng J A. SIKey: A tool to generate secondary identification keys for targeted diagnosis. *Expert system with application*, 2006 (In press).
- 张小斌, 程家安. 移动网络环境下的植物检疫性昆虫信息平台 W-QPM. 浙江大学学报. 农业与生命科学版, 2006, 32(4): 406-409.
- 张小斌, 陈学新, 程家安. Lucid Phoenix: 交互式多媒体网络检索表工具. 昆虫分类学报, 2006, 28(3): 231-234.
- 张小斌, 陈学新, 程家安. 基于Web的中国昆虫科级鉴别分类系统 InsectID的设计与开发. 昆虫分类学报, 2006, 28(1): 63-68.
- 张小斌, 陈学新, 程家安. 为何海洋中的昆虫种类如此稀少? 昆虫知识, 2005, 42(4): 471-475.
- 大鸣剑客 InsectID 昆虫辅助分类鉴定软件 (软件著作权 2006SR17549)
- 中国昆虫鉴定分类系统 InsectX: <http://insectx.3322.org>
- 植物检疫性昆虫信息平台 W-QPM: <http://qpm.3322.org>
- 植物检疫信息管理与辅助决策网络系统: <http://qpm.8866.org>

作者: [张小斌](#)
学位授予单位: [浙江大学](#)

相似文献(2条)

1. 学位论文 [邵洪民](#) [MSK调制技术在数字化列车自动控制系统中的应用研究](#) 2003

该论文研究的重点是探讨MSK信号调制技术应用于数字编码轨道电路的可行性,提出并完成具体实现方案.论文中介绍并分析了国外具有代表性的数字化列车自动控制系统;研究了列控安全信息数字化传输的关键技术问题;结合不同数字调制方式的比选,数学模型分析、以及软件仿真等相关工作,在提出国产数字化列车自动控制系统整体设想的基础上,重点阐述了基于最小频移键控(MSK)调制实现列控信息数字化传输的技术解决方案,并且分别采用可编程定时计数器(PTC)和直接数字频率合成技术(DDS)加以实现.对于两个不同的设计方案,均完成了核心模块的硬、软件的设计、关键技术分析、功能验证以及方案间的对比分析.

2. 学位论文 [李延鹏](#) [战伤伤员信息数字化研究](#) 2006

以往对战伤伤员状态(伤部、伤类、伤势等)、运动特征(分类、去向等)以及与战伤伤员有关的卫勤信息(救治措施、后送工具等)的描述多是定性的,难以对伤员状态定量评价,影响了战时卫勤信息资源的开发和利用.本研究基于信息论和数字编码技术,将战伤伤员及其相关信息数字化,使伤员状态定量评价成为现实,不仅为实现伤员的精确化医疗后送提供了可能,同时,亦为战时卫勤C4I建设提供了理论基础,因此,战伤伤员信息数字化研究是战时卫勤信息化的重要基础性研究课题。

战伤伤员信息描述的是战伤伤员状态及其运动变化的属性.战伤、战伤伤员以及伤员医疗后送产生的信息,构成了战伤伤员信息系统.基于信息系统是构建战时卫勤C4I和救治机构伤病员管理系统,其前提是伤员信息数字化.本研究基于信息论,运用系统科学的方法,从战伤信息源头上,对战伤、战伤伤员的发生及其相关信息进行了系统性研究;分析了战伤伤员信息系统,构建了信息系统模型;基于通讯领域数字化的概念、技术和方法,定义了伤员信息数字化的概念、方法和内容;提出了数字化伤员概念,构建了数字化伤员模型.特别是针对高技术局部战争参战人员心理精神创伤具有不同于一般战伤的特点,本研究又重点对战时心理损伤伤员及其信息数字化编码进行了研究。伤员数字化实质就是对伤员及其医疗后送要素信息的数字化,包括反映战伤状态的伤部、伤类、伤势信息和战时伤病员分类、救治和后送等信息的数字化,其基本方法是运用数字编码技术对上述信息进行编码.在此基础上对数字化伤员概念进行了探索,认为数字化伤员是基于伤员信息要素的数字化,采用计算机对伤员信息进行描述、控制与管理,通过数字通讯技术联网和后勤指挥一体化系统,实现伤员信息实时化感知、网络化传输、智能化管理的数字化信息虚拟伤员.该理论的提出为战时卫勤信息化建设提供了伤员救护、后送等信息管理理论上的参考。针对高技术战争,参战人员心理损伤,提出应将心理损伤纳入战伤范围,并对其信息进行数字化编码。

本文链接: http://d.g.wanfangdata.com.cn/Thesis_Y1124655.aspx

下载时间: 2010年5月2日