引　言

伴随着信息时代的来临，科学数据已成为信息化技术快速发展的一个关键。据统计，人类社会最近 30 年所积累的科学数据总量已经超过了人类 5000 年发展历史所积累的数据量总和。尤其是美国，它所拥有的数据库总量占全世界总量的一半以上。20 世纪 70 年代，随着计算机技术的飞速发展，美国科学数据库的开发如雨后春笋般地发展起来，而发展最快的领域就是地球科学。1990 年，美国航空航天局建立了“分布式最活跃数据档案中心群”，它标志着美国国家层面上对科学数据库建设和共享工作的高度重视。该数据库中更是包括诸如地物光谱数据库、陆地资源卫星遥感数据库等，它们是国家空间信息基础设施建设的重要组成部分，其主要服务对象就是政府决策、国家重大科研项目、国家各个科研群体以及企业部门，在国家的科技发展、国民经济建设和国防建设中发挥了巨大的作用。

遥感已成为地球空间信息获取的主要技术手段，而地物波谱研究是遥感理论与应用技术发展的基础，正如在童庆禧院士和田国良研究员合著的“中国典型地物波谱及其特征分析”一书中所写的那样，“遥感技术的基本任务,就是在以电磁波辐射为表现形式并在传递媒介复杂多变的环境信息中,通过各种有效手段来收集、处理、分析和提取所需要的特征，达到认识、识别研究对象的存在、状况和动态的目的。这就要求深入研究环境要素的电磁波辐射信息的发生、变化、转换和传输规律，提高遥感技术的作用。有人认为：遥感技术就是一个收集、分配、处理和分析电磁波辐射信息的系统，这种观点从物理和技术的角度来看，无疑是十分正确的”。

高光谱分辨率（简称为高光谱）遥感或成像光谱遥感技术的发展是过去二十多年中人类在对地观测方面所取得重大技术突破之一，是当前遥感的前沿技术（陈述彭等，1998）。它融合了成像技术和光谱技术，其核心特点是图谱合一，即能够获取目标的连续、窄波段的图像数据的技术。高光谱技术自从诞生以来，就以其丰富的光谱维信息显著区别于传统的遥感技术，在地质、植被生态、土壤，以及城市应用等方面的研究中取得了引人注目的成果，已经成为当前遥感的一个重要发展方向，高光谱分辨率遥感信息的分析处理集中于光谱维上进行图像信息的展开和定量分析。通过高光谱成像所获取的地球表面图像包含了丰富的空间、辐射和光谱三重信息，因此，自从 70 年代末，美国喷气推进实验室（JPL）在美国宇航局（NASA）支持下首先对成像光谱仪进行概念设计和研究（Goetz,1992）以来，地物目标识别和属性探测以及高光谱成像仪的开发一直是遥感的一个重要方法和技术发展方向。航空成像光谱仪已经发展成熟,现在运行的有美国的 WIS、AVIRIS,加拿大 CASI、SFSI,澳大利亚的 HYMAP（Cudahy et al,2000;Farrand,2000;Martini et al,2000;Peters et al,2000）和我国的 PHI、OMIS 等（Tong Qingxi,1998）。在航天领域中,除人们所熟知的美国对地观测系统（EOS）计划中的中分辨率成像光谱仪（MODIS）和欧洲空间局的中分辨率成像光谱仪（MERIS）之外，搭载 220 通道高光谱仪 Hyperion 的 EO-1 卫星已经升空，而另一颗载有高光谱遥感器的 Orbview4 也

即将发射。可以预料，在二十一世纪初，空间高光谱成像卫星将成为对地观测中的一项重要前沿技术，在地球资源开发利用及地球环境监测中发挥越来越重要的作用，高光谱遥感实用化的时代即将到来，开展高光谱遥感的基础和应用研究在我国已具有重要的意义和迫切的需求。

图 1.1 为高光谱遥感技术的示意图。我们把高光谱遥感获取如图所示的二维空间、一维光谱的数据立方体称为图像立方体。利用高光谱遥感图像立方体，一方面可以通过目标的空间几何形状信息对目标特征进行分析、识别和定位，另一方面，可以通过目标的波谱特性来确认目标或揭示目标更丰富的本质属性。

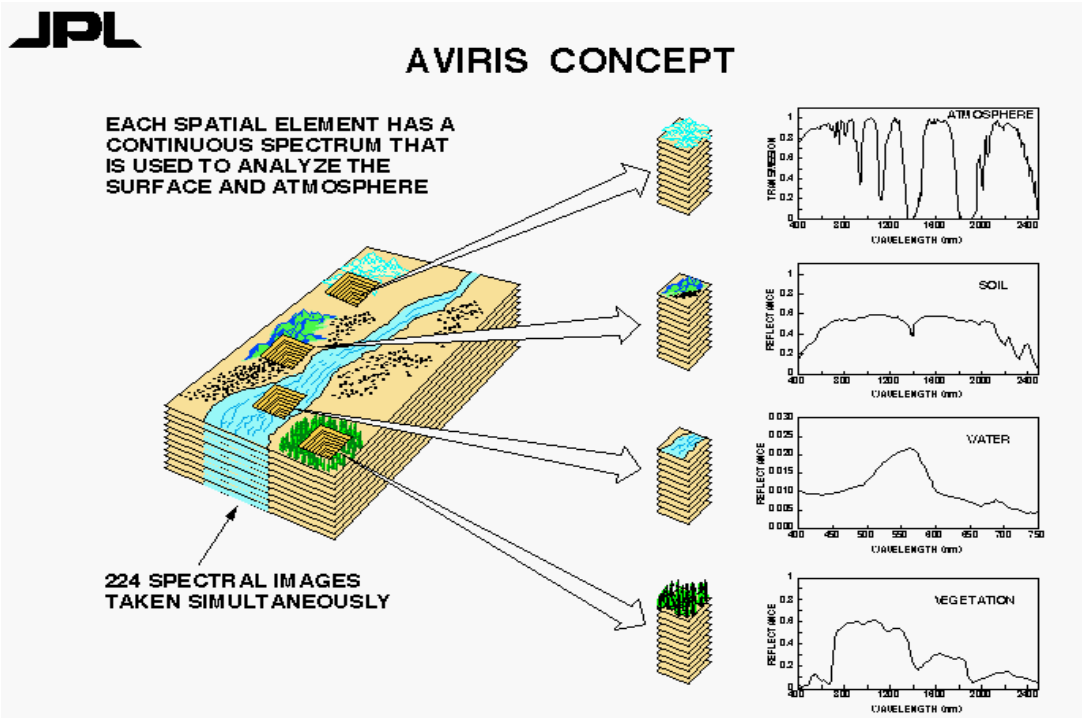


图 1－1　高光谱遥感技术示意图

当前定量遥感已成为新世纪遥感科学发展的最显著特征，定量遥感研究更离不开波谱数据库的有力支持。现如今国内的遥感应用水平参差不齐，遥感基础研究条件建设严重不足，对于高光谱这样的新技术更是如此。同时在计算机方面，数据库技术已经很好的解决了数据存储、查询、更新、维护等基本需要，如何能把两者很好的结合起来，尤其是在高光谱的图谱合一的特性上有所发展，是一个很重要的课题。在数据库中可以将前人辛苦的工作成果，处理、分析过的标准样本数据存储起来，这对于后人相应方面的工作是很大的贡献。在高光谱应用的直接方向上，图像处理、分析、分类等工作，有一个样本丰富、功能强大的数据库用于比对、借鉴、分析、整合数据等等，都将是一个很有力的支持。在数据存储的基础上，如果可以结合数据处理，加以网络化，将对推动高光谱技术的发展，密切国内、国际同行的

交流、沟通与合作起到重要的作用。近年来，数据仓库技术日益成熟，可以针对不同的应用需求所关注的问题，重新整合数据库数据，进行综合和分割，并采用多维查询模式，从而实现数据挖掘（Data Mining, 简记为 DM)。这项技术已经大量应用在商业领域，在 GIS 领域中也已有应用，它在分析海量数据的内在联系，进行决策支持方面取得了显著的成果。如果我们可以借鉴这类先进的技术，努力探索数据挖掘技术应用在遥感数据方面，特别是高光谱数据，以希望深入分析试验中所取得的生物参量、光谱、图像之间潜在的联系，必将会进一步推动高光谱技术的发展和应用。

尤其是经过若干年的发展，高光谱遥感已经积累了很多的数据和经验，如何把他们整合在一起，将实验性的工作成果向应用转化，已经是当前很重要的课题之一，也已经有很多实际研究涉及到了这方面的工作，比如：863 项目中的地物标准波谱库、精准农业高光谱数据库等等。同时，从应用实践的角度出发，人们还希望能够把对高光谱数据的各种分析功能与数据库的存储、整理特点相结合，这样就可以把原有的典型的标准地物的信息和新采集的地物信息加以比对、分析、精练，从而迅速、准确的确定目标的类别、性状等等，这样就可以形成一套完整、实用的高光谱应用系统。比如：农业中加入高光谱信息与作物理化参量的回归分析；岩矿应用中比对测量光谱与标准光谱从而识别矿物成分等等。这就对我们构建数据库提出了更高的要求，不但要实现数据库的基本功能，还要考虑附加的分析功能，尤其是系统的可扩展性，因为需求也许会不断地提出，良好的扩展性将为系统未来的发展打下坚实的基础。

同时在这个信息通讯日益发达的时代，如何能利用目前先进的网络技术，让我们的成果能够尽量为更多的遥感专业人员服务，甚至将来为野外等实地工作服务也是很有意义的课题之一。当前的网络技术已经在向着无线技术方向不断前进，并且有了很重大的突破，蓝牙、WI-FI、迅驰这一系列新鲜的名词不断的出现在无线网络应用的领域里。这就意味着网络可以摆脱网线的羁绊在更加自由的天空里驰骋，这对于遥感领域也是一个重要的消息！试想如果我们可以建成一个功能强大、数据丰富的网络数据库，那么在野外工作的高光谱遥感工作者，只要简单拥有一台具有无线上网功能的笔记本电脑，就可以现场比对采得的样本和数据库中的标准样本，从而迅速的得出结论，也可以以先进的数码科技，将新的数据、图片等及时添加进数据库供在其他野外实地工作和实验室中的科技人员使用。可见网络化对于数据库，尤其是高光谱遥感数据库应用有着“无限”的意义，正是基于这种考虑，本系统把数据库系统的网络化作为本系统的一个重要方向。

综上所述，本次研究主要以现有的高光谱数据为前提和基础，力求建立一个基于网络的高光谱数据库系统，该系统除了具有数据库基础的数据存储、查询、添加、删除等功能，还应针对需求，具备一些简单的数据分析功能，同时力求针对高光谱数据的特点，做一些能够在网络媒体上体现高光谱数据－－图谱合一特点的研究。同时，在当前数据库的新热点－－数据仓库和数据挖掘的概念上，本文针对高光谱应用的现状进行了初步的探讨，提出了一些

有益的探索。

本篇论文共分七个章节，第一章从传统的光谱数据库介绍开始，针对高光谱数据的特点和数据库实践，提出了创新性的高光谱数据库的概念，它点明了整篇论文的理论核心，说明了本系统的理论价值和创新性。从第二章开始从全系统需求分析、设计、开发的角度，围绕高光谱数据库这个概念，从五个方面、共分五个章节进行论述。第二章的重点是高光谱数据库的需求，这是整个系统的出发点和源头。第三章、第四章、第五章分别从基于网络的光谱数据库子系统、基于网络的数据分析子系统、基于网络的光谱图像子系统等三个子系统全面介绍了整个高光谱数据库系统的设计方案与建设实施，并提出了在高光谱数据库系统中的数据存储规范以及ORACLE数据平台下的数据存储模式，并进行了比较和分析。这对于以后类似的高光谱数据库系统的构建具有重要的借鉴意义。其中在第三章里，整体所采用的网络服务平台和前台界面也做了详细的介绍。第六章结合当前数据库发展的最新热点数据仓库和数据挖掘，对在高光谱数据方面的应用前景，进行了前瞻性的研究。高光谱数据数据量大，关系紧密，如果能够成功的应用数据仓库和数据挖掘，将会有很光明的应用前景。第七章是结论，总结了全文的创新点与特色，同时对系统的发展做了进一步的展望。

综上所述，本篇论文紧密围绕“网络环境下的高光谱数据库构建及其应用实践”这一主题，分七个章节逐步展开论述，论文结构如图 1 所示。

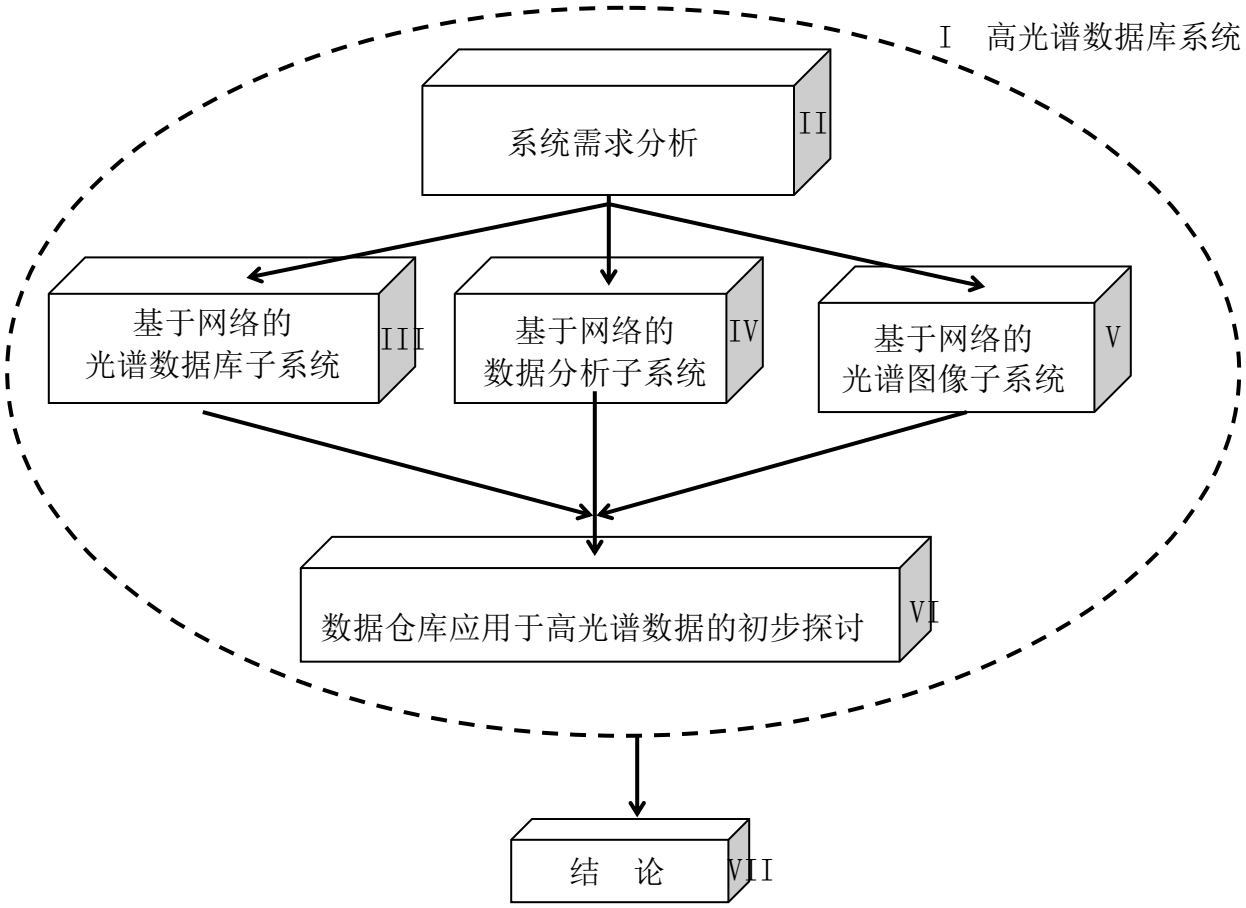


图 1 论文组织结构

第一章 高光谱数据库系统概念与逻辑结构

目前，在遥感技术领域，国内国外都已经建成了一些光谱数据库，但其存储内容多为单条光谱曲线的集合。如引言中提到的，高光谱数据的特点就是图像与光谱的紧密结合，这与一般的光谱数据还有很大区别，在不断深入的用于存储高光谱数据的数据库开发实践过程中，本文提出一个全新的概念－－高光谱数据库。

1.1 光谱数据库发展现状

目前已经有一些研究机构在长期工作的基础上，建立了一些光谱的数据库。借鉴他们的成果，吸取他们的长处对于我们开展这项研究会有很大的帮助

1.1.1 目前已有的一些国外光谱数据库

在高光谱遥感方面，国外比国内起步要早一些，发展也要快一些，国外已经建立了几个光谱方面的数据库，主要有：

1、IGCP—264 光谱库：由美国 IGCP—264 项目于 1990 年收集建立，包括由 5 种光谱仪测量所得到的 5 个光谱库，它们是：

（1）科罗拉多大学空间对地研究中心（SCES）采用改制的 Beckman5270 双光路反射光谱仪测量的光谱。光谱分辨率为 3.8nm，重采样成 1nm 分辨率。

（2）SCES 采用 GER 公司 SIRIS 便携式野外光谱仪测量的光谱。SIRIS 是单光路三个光栅的光谱仪，第一个光栅波长范围为 350nm—1080nm；第二个为 1080nm—2500nm；第三个为 1800nm—2500nm。

（3）SCES 采用 PIMA Ⅱ野外光谱仪在实验室条件下测得的光谱，光谱分辨率约为 2.5nm。

（4）布朗大学采用 Relab 光谱仪测量的光谱。光谱分辨率为 2—13nm，在 400—2500nm 范围内重采样成 5nm。

（5）USGS 丹佛光谱实验室采用计算机控制的 Beckman 光谱仪测量的光谱。光谱分辨率在可见光范围为 0.2nm，在近红外为 0.5nm。

2、John Hopkins 大学光谱库：采用 Beckman 和 FTIR 光谱仪测量。测量对象包括各种火成岩，变形岩，沉积岩，雪，土壤，水体，矿物，植被以及人工目标等多类物质（Korb，1996；Salisbury，1994）。

3、JPL 的 ASTER 光谱库：采用 Beckman5240 光谱仪测量。包括 160 种矿物岩石在 125—500 微米，45—125 微米，小于 45 微米三种微粒尺度下的光谱，以研究微粒尺度与光谱之间的关系。光谱波段宽度在 400—800nm 之间为 1—4nm，800—2200nm 之间小于 20nm，2200—2500nm 之间为 20—40nm（Grove，1992）。

另外，在许多遥感商用软件中也包括高光谱数据库模块。如在 ENVI 软件中拥有波谱库管理、编辑及分析模块，它包含了美国地质调查局的 USGS 光谱库，喷气推进实验室的 JPL 标准物质成份波谱库，John Hopking 大学 2-25μm 热红外及植被波谱库，用户可查看、建立、重

采样标准波谱库和自己的波谱库，从而使用户可进行物质成份、热红外分析和植被分析。在 PCI 软件的高光谱分析（Hyperspectral Data Analysis）模块中也提供了基于 USGS 光谱库发展的高光谱地物库，并支持用户有限光谱通道的光谱库，即可由用户自行组合成有限光谱通道（如 10－20 个）的光谱曲线库。它同时提供用户各种光谱分析能力，自动地物判识（根据光谱特点）等功能，用户可用上述工具对高光谱影像进行辅助的或半自动的地物判识，或结合 PCI 软件的多光谱分析（Multispectral Analysis）和神经元网络分类模块及其他影像解译方法进行地物判识。在 ERDAS 软件的高光谱工具模块（Hyperspectral tools）中，也包括 JPL，USGS，以及用户自定义的光谱库。

1.1.2 基于 FOXPRO 的 HIPAS 数据库

中科院遥感所高光谱室是国内在这一领域开展工作比较早的一个单位，在积累了多年的数据和实践经验之后，为了更好的整理已有的工作成果，高光谱室以微软的 VISUAL FOXPRO 为平台建立了一套挂靠于图像分析系统 HIPAS 的相对独立的高光谱数据库系统。里面收集了土壤、岩矿、水体、植被和人工物体五大类目标，包涵了对象的光谱数据、测量的地学属性数据，例如：风速、云量、太阳高度角等、测量的仪器参数等等。本课题最初就是在此基础上提出的，为了适应数据的不断增长以及更大范围应用的需求，所以希望能从普通的桌面数据库平台 FOXPRO 向大型的数据库平台 ORACLE 发展。所以这里我先介绍一下基于 FOXPRO 的高光谱数据库。

图 2－1 基于 FOXPRO 的高光谱数据库数据输入界面

该系统主要实现了数据库基本的查询检索、添加、删除，修改等功能，可以为对象的光谱画出曲线，并且可以实现去除包络线，以突现特征波段。下面几幅图是该系统最主要的界面，其中反映出了查询出来的数据对象的各种属性以及其光谱曲线。可以看到该系统已经有了最初步的分析功能，如：去除包络线显示特征波段，计算太阳高度角等。

图 2－2 基于 FOXPRO 的高光谱数据库查询结果界面

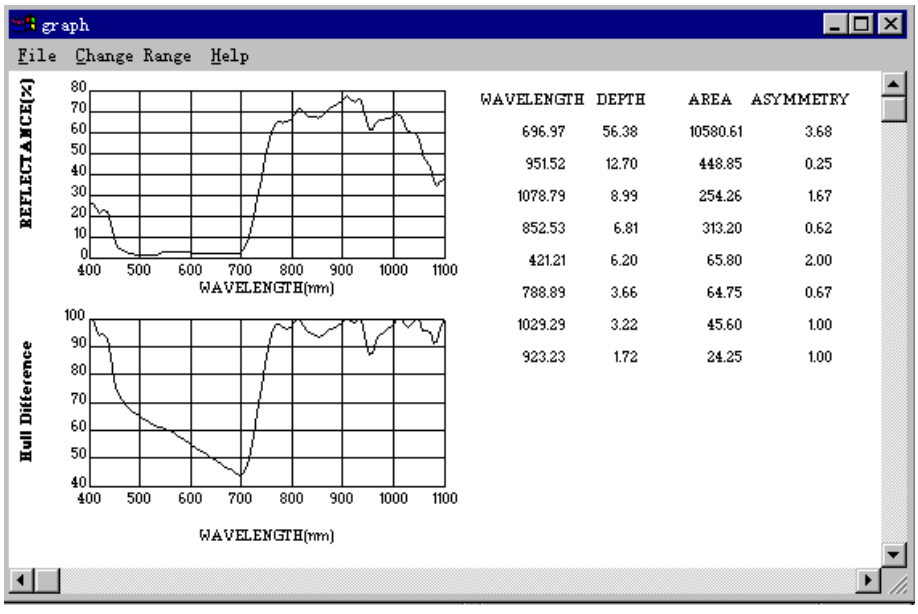


图 2－3 光谱曲线及特征波段界面

可见，在光谱数据库方面，前人已经做了很多的工作，可以基本上满足对光谱曲线的查询等操作。这为本文的成果打下了良好的基础。

1.2 高光谱数据库概念的提出

引言中已经提到，高光谱数据有着其与众不同的特点，所以决定了在数据库开发方面也有其与众不同的特点，也就诞生其高光谱数据库这个全新的概念。***高光谱数据库系统是专门面向高光谱数据，体现图谱合一特性，综合了光谱数据库、光谱分析功能和数据挖掘功能于一体的专用数据库系统。***它与通常意义上的光谱数据库最大的不同在于它不仅存储室内或野外光谱辐射计获取的目标光谱数据，它还可以存储以图像立方体形式存在的高光谱图像光谱数据，并根据用户需要，可以从标准图像数据块中提取所要求的任意象元级上的光谱曲线。高光谱数据库在以下两个方面具有非常现实的意义：一是面向近地面的地面成像光谱辐射计，二是从航空或航天层次上开展的地面光谱遥感成像测量。

（1）目前，国际上主要的光谱辐射计均采用点测量模式，只获得测量对象的光谱曲线。然而任何对地观测都有尺度效应的问题，在遥感器视场角不变的情况下，不同的遥感飞行高度对应不同的象元地面大小。野外光谱测量所对应的地面视场范围与光谱图像象元的大小相对应是图像光谱定标质量好坏的一个关键因素，而目前野外光谱辐射计在均难以做到地面视场的精确设定。

在目前成像光谱遥感科学与技术发展中，无论是航空还是航天成像光谱仪，其图像反映的是像元尺度上的光谱信息。而现有地面光谱光谱辐射计又只适合点状地物的光谱测量，两种仪器之间在空间尺度上衔接的不够紧密，因此在象元组分光谱分析、混合象元分结等问题上缺乏创新性和精确的测量手段。

有鉴于此，地面成像光谱辐射计的研制已经提上议事日程。它不仅保留了地面光谱测量的特点，也具有精确设定视场范围和高地面分解力的优势，尤其适合于混合光谱研究和可变尺度下的象元光谱组构分析。

但是目前的光谱数据库还没有考虑到应对这种集图像与光谱为一体的地面成像光谱辐射计数据，高光谱数据库概念的提出和系统建设将可有效解决这个问题。

（2)另外，高光谱数据库在军事侦察中也具有十分重要的意义。美国海军研究办公室(ONR)和海军研究实验室制定了一个高光谱遥感技术发展计划(HRST)即 Hyperspectral Remote Sensing Technology Program。其核心就是要研制和发射美国海军的地球制图观测卫星 NEMO，以作为海军的大型高光谱数据的基准平台。它所带的 COIS 高光谱成像仪可以获取地面分辨率 60 或 30 米、210 波段的高光谱图像立方体数据。由于无法到别的国家进行地面光谱测量工作，通过这一计划，美国军方就通过光谱成像方式来建立世界范围内 50 个热点地区的高光谱数据集。因此，相应的高光谱数据库的建设是十分必要的。

从以上两点可以看出，面对这种新型光谱测量仪器以及光谱测量方式的出现，高光谱数据库都具有十分重要的意义。

1.3 高光谱数据库系统（逻辑结构）的基本组成

一个完整的高光谱数据库系统应该有如下的组成部分：核心部分是高光谱图像样本库系统；为了保证数据的多样性，以及和地面光谱的匹配，基于已有技术的光谱数据库辅助系统是必要的；为了扩展数据库的功能，在存储、搜索等基本功能上丰富数据库的应用，建立一个高光谱数据分析系统会有很大的帮助；在整个数据库系统之上，面向高光谱海量数据集的带有数据挖掘功能的数据仓库将为高光谱数据的应用提供更广阔的空间；最后由前台界面系统与用户交互，得到其关心的条件并返回其需要的结果。下面将各系统功能进行详细描述：

（1） **高光谱图像样本库系统**： 这是高光谱数据库系统的核心所在，它主要负责高光谱样本图像光谱数据以及其对应的属性数据的存储、查询、浏览、添加、修改、删除等基本操作，其中存储的可以是地面成像光谱辐射计或者航空、航天高光谱成像仪采集的图像光谱数据，其兼容性保证各种光谱波段设置的高光谱成像仪采集的样本均可以存储。同时，出于应用角度的考虑，在库中存放的应该是经过纠正、处理的标准图像光谱数据，同时其相应的属性中不但可以包括地学、测量方面的属性数据，还可以面向需求扩展到经纬度信息乃至分类信息。

（2） **光谱数据库辅助系统**： 高光谱数据库系统是来源于已有的光谱数据库系统之上的，而且现有的很多高光谱航空、航天成像仪还是需要地面调查，取得地面光谱的配合，所以建议一个光谱数据库辅助子系统无论从技术角度还是应用角度都是很有必要的。目前国内光谱数据库系统技术也尚不成熟，建立一套综合了光谱、图片、其他属性数据，和高光谱数据库处于同一个大型数据库平台下的光谱数据库系统既可以为高光谱数据库起到必要的辅助作用，也可以进一步提高我国光谱数据库的发展水平。如果可以将其网络化，则将极大的扩展系统的应用范围，后面提到的附加的数据分析系统也将进一步促进光谱数据库的应用。

（3） **高光谱数据分析系统**： 为了满足科研的需要，作为一个成熟的数据库应用系统，光有数据库基本的管理功能是不够的。尤其是高光谱数据中富含了很多信息，相应的处理模型、方法也已经作了很多的研究，将一些常见的处理方法软件化，为全系统搭建一套高光谱数据分析子系统将极大的促进高光谱数据库系统的应用范围，给高光谱研究应用带来极大的便利。

（4） **带有数据挖掘功能的数据仓库**： 一景 1024×1024 分辨率的高光谱图像数据就将含有 1M 条光谱曲线，如果每条光谱曲线有 100 个波段，就将是 100M 的数据量，明显可以看出高光谱数据是典型的海量数据。同时，高光谱图像光谱数据具有很强空间几何信息，相互之间具有复杂的关系，多景同类图像之间更会有复杂的关系，这就给利用数据仓库整合数据，再进行数据挖掘提供了良好的基础。数据挖掘、数据仓库正是当前数据库发展的前沿和热点，其目标就在于数据积累的基础上，深入探索数据间的关系与规律，从而得到进一步有价值得信息和结论。可以

预见，将数据挖掘、数据仓库应用于高光谱数据必将有光明的前景。

（5） **前台界面系统**：它的目标很明确就是和用户交互，得到其意见并反馈其需要的信息。这也是系统中最富于变化的一部分，从形式上讲它可以采取网络化或者单机运行两种方式，各有优势，但网络化必将是发展的趋势；从开发而言则有许多常见的工具可以选用；从样式上，则随着开发的投入可以不断改进。

由高光谱图像样本库系统、光谱数据库辅助系统、高光谱数据分析系统以及带有数据挖掘功能的数据仓库等就构成了一个面向高光谱数据应用的完整的高光谱数据库系统。从这些子系统的设计可以看出，高光谱数据库系统这个崭新的概念必将为高光谱应用开创一个崭新的局面。

1.4 高光谱数据库系统的开发及应用过程

高光谱数据库系统的开发与应用可以分为下列几个步骤：

（1） 系统需求分析：高光谱数据库系统可能会因为应用的领域不同在个别子系统中有不同的设计和变化，充分的进行系统需求分析可以明确用户的需要，从而进行相应的设计，为开发工作的实施打下坚实的基础。

（2） 高光谱数据结构设计： 前面提到了，高光谱数据数据量非常大，而且直接和图像紧密联系，再加上其他的属性数据，因此，在数据库中的数据结构设计直接影响到整个高光谱数据库系统的性能。在国内，光谱数据库的建设也不成熟，如何在大型数据库平台上合理的存储光谱数据也是一个新的课题。

（3） 数据库前台界面设计：随着不同的用户需求，这是变化比较大的一个部分。数据结构设计好将使得整个高光谱数据库系统可以在各种前台界面下给出结果，返回信息。但总体上讲，无论采用什么样的软件、语言，无非是单机版和网络化两种方案，其各自有各自的用途，但毫无疑问，网络化可以促进数据的交流、共享与应用，必将是发展的方向。

（4） 数据分析系统设计：根据不同应用领域用户的需求，如：岩矿、农业、城市等等，他们关注的波段、目标、结果并不一致，所以这个部分将是全系统中变化最大的。

（5） 数据获取：数据获取可以分为两个部分，数据采集＋数据整理。数据采集是指用地面成像光谱辐射计或者航空、航天高光谱成像仪采集的图像光谱数据，以及用地面光谱辐射计采集的地面点光谱。数据整理则包括基本的数据处理工作，包括结合 GPS 系统的定标、定位、几何纠正、辐射纠正等等，这样做的目的就是使高光谱数据库中的数据标准化为使用做好准备。

（6） 数据入库并使用：最后就是将前面整理好的数据按照数据库设定的数据结构，分门别类的存入数据库中。如果能够网络化，则可以实现远程访问，乃至工作现场的访问。同时，其附加的数据分析子系统也可以在应用过程中起到重

要的辅助作用。

本文后面几章将按照上述思路具体介绍一个高光谱数据库原型系统的建立，并在前台界面实现网络化，这样为今后全系统的建立和完善奠定了坚实的基础。

1.5 本章小结

本章作为全文的第一章，点明了全文的核心概念，高光谱数据库系统是专门面向高光谱数据，体现高光谱图谱合一特性，综合了光谱数据库、光谱分析功能和数据挖掘功能于一体的专用数据库系统。从现有的国内外光谱数据库发展现状讲起，提出全新的高光谱数据库系统的概念，然后对全系统的组成、概貌以及开发、应用的过程都做了比较详细的介绍。后面各章都将围绕如何建立一个高光谱数据库原型系统而展开。最后本节将以关系图 2 来表示整个带有创新性的高光谱数据库系统的概念。

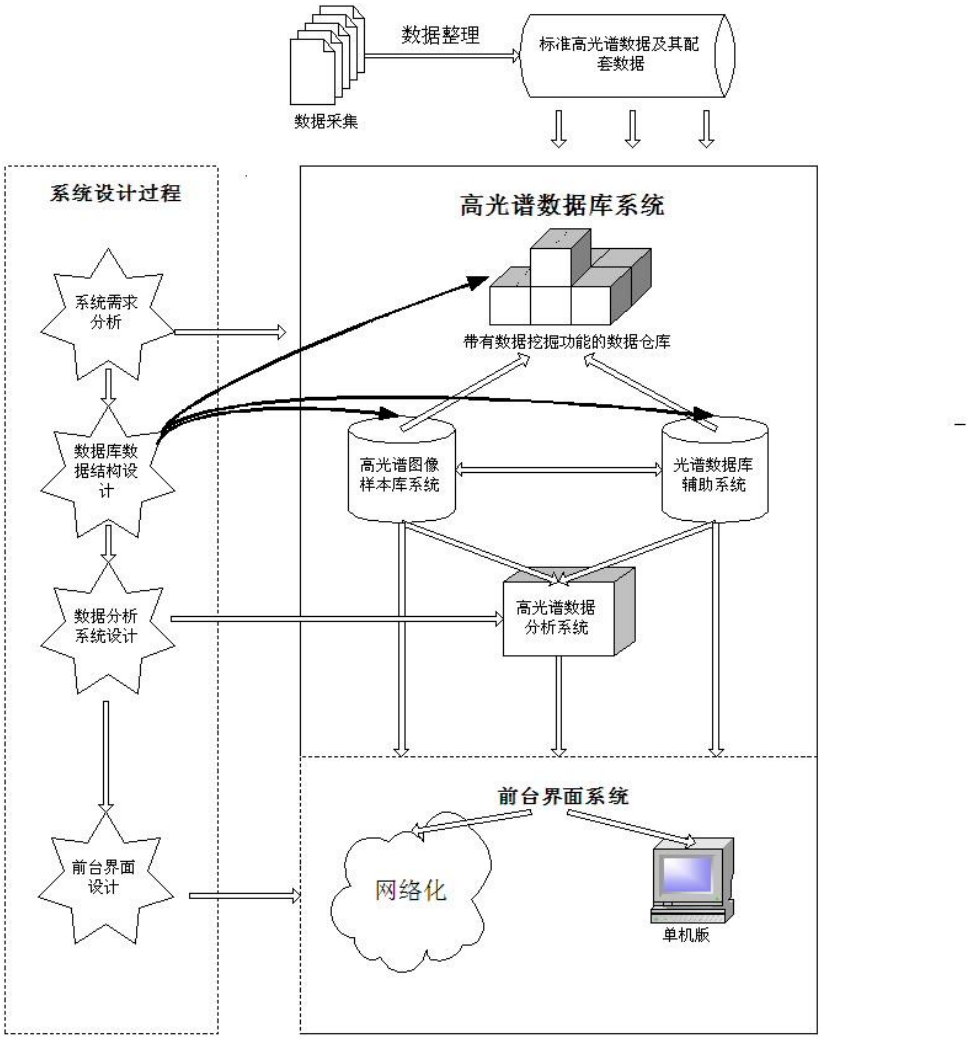


图 2 高光谱数据库系统概念示意图

第二章 系统需求、特点与整体结构

2.1 系统需求分析

正如引言所提到的，高光谱的应用范围非常广泛，不同的应用可能会有不同的需求，而作为一个工程项目而言，具有明确的需求对于整个设计思想的形成、功能模块的设计、系统的开发乃至类、函数和变量的设计都是十分重要的。本次研究除了为高光谱数据库系统建立原型之外，主要是依托于由北京农业信息技术研究中心承担的北京市自然科学基金资助的重点项目－－面向精准农业的作物光谱信息系统研究。该研究的一个主要目标就是将提供一个具有易维护，易操作，便移植，便扩展功能的面向农业的作物光谱数据库系统，为用户提供查询，检索，分析处理，共享与交流等服务；为成像光谱遥感定量化理论研究和应用服务。该项目有比较清楚明确的需求已经整理好的数据集，这正好可以作为高光谱数据库系统中光谱数据库辅助系统来设计。具体而言主要需求有：

（1）图像样本库模块：作为高光谱数据库系统发展的核心研究内容，希望能够在数据平台上建立一个能够反映高光谱图谱合一特点的图像样本库。这项功能在 envi 等高光谱数据处理软件中已经得到实现，但是在数据库平台下，乃至网络环境中还鲜有实现。

（2）高光谱数据存储和应用的研究：本系统的一个关键是设计良好的数据结构，使得光谱信息与图像信息能够很好的相结合。同时，在大型数据库平台上建立光谱数据库在国内也是前沿的课题，能够建立实用性的数据规范，对于光谱以及高光谱数据库系统都是很重要的。

（3）网络环境下的发布：为了使高光谱数据库系统能够得到最大的发展，为之建立网络平台无疑将是最有利也是最有用的解决方案。这可以使得系统在可能的情况下为尽可能多的研究人员、用户服务。对于光谱数据库这也是一个崭新的课题。

（4）为光谱数据建立数据库：以小麦、玉米为主的作物全生育期的生长动态光谱及其相关环境信息（如测量时间、地点、天气条件、测量仪器等）为主要数据源，选用 ORACLE 开发环境，采用面向对象的编程方法，最终形成一个易维护，易操作，便移植，便扩展的光谱数据库。

（5）拓展原有的光谱数据库系统：将数据库平台从 FOXPRO 移植到 ORACLE 上，为系统扩容，以及向更大规模发展打下基础；保持数据库的基本功能，如：查询、添加、修改、删除等等；提供更加容易扩展的接口规范，以利于更广范围的应用；为了便于数据共享与交流，希望能在网络的平台上发布；进一步丰富其应用性的数据处理功能等等

（6）光谱分析功能模块：为了研究以小麦和玉米为主的作物全生育期光谱特征及其在不同水肥等环境因子作用下的变化规律，深入探讨光谱诊断机理，本数据库将具有一定的光谱分析功能，比如光谱重采样；去包络线；有限特征参量提取，包括：

红边、红谷、绿峰、黄边、NDVI 等；光谱和生化参量线性回归统计分析；导数光谱等等；也是高光谱数据库系统分析模块的一个原型系统。

（7）系统的扩展性：首先需要考虑的是不同高光谱仪器数据的存储，其次这个系统还应该考虑将来可能会有更多的分析功能不断的加入，而且可能会有新的数据模型加入，所以在系统设计时，考虑到系统的延展性，可扩展性。

2.2 系统特点

从上面的讨论和系统需求分析中，我们可以看出，本研究需要考虑的方面很多，并且具有其与以往数据库系统不同的特点，完成整个系统的设计与建设，具体来讲主要需要面对如下几个主要的问题:

（1） FOXPRO 是属于桌面型的中小型数据库，ORACLE 则是很不相同的大型的数据库平台，数据管理系统，数据结构等等都复杂许多，由此决定，从后台的数据结构、数据管理系统到前台的界面系统都不可能照搬 FOXPRO 下的设计，需要重新进行整理、规划。

（2）高光谱数据集中最重要的是光谱数据，而光谱数据又有与众不同的很多特点，这些直接决定了设计的复杂性。

A.一个对象对应几十波段乃至上千波段不等的数据，而且每个波段的数据都可能需要单独提出，以用作分析或者计算，这就为数据结构的设计提出了问题。

B.高光谱仪器林林总总，会有不同的波段设置，不同的对象很可能采取不同仪器测量，也就是一个库中可能有多个波段数的光谱数据。作为一个兼容性好的数据库需要考虑到这些情况，要能适应各种数据。

C.高光谱数据的图谱合一的特性，需要将变化性很强的光谱数据与图像数据很好的结合在一起，这也需要对数据结构进行精心的设计。

在考虑到上述因素的同时，数据库尤其是基于网络环境的数据库，其查询时的读取速度也是特别需要注意的问题。

（3）在网络条件下进行数据库设计与在普通的软件环境下设计数据库又有着很多不同的因素，如：速度、读取方式等等。

（4）这个研究项目中，高光谱数据的其他属性也比较复杂，种类繁多而且几乎都是查询的检索关键字，如何能够兼顾数据结构的简洁性与适度冗余性，从而兼顾速度、安全性和存储空间也是一个需要慎重的问题。

（5）在为网络环境设计的大型数据库 ORACLE 中，图片的存储也有多种可以考虑的方式，现在常见的是以文件的方式存放于操作系统之中，但如何能够针对高光谱数据的特点，选择一种合适的方式是本项目一个需要考虑的问题。

（6）本系统除了需要实现数据库的基本功能，而且要外加很多分析功能模块设计，这在系统整体结构的设计上提出了新的要求。

由此可见，这项研究工作具有很重要的意义，同时在设计、开发、操作时也具有相当的难度。

2.3 系统总体结构

针对各项需求，全系统需要从图像样本库子系统，光谱数据库子系统，网络服务子系统以及前台发布子系统来进行整体结构设计，同时需要专门考虑数据分析处理系统。本文将分别就这几个系统一一详述，这节将把总体的技术路线以及设计思想进行一个概括性的描述。

2.3.1 全系统逻辑结构设计

2.3.1.1 三层网络服务模式

首先本系统整体上采用了网络服务模式是由数据库应用特点出发的。一方面，数据共享，数据互连正在成为各行各业努力达到的一个信息化目标，高光谱数据自然也不例外，而且随着有线、无线网络的不断发展，数据库网络化更加有利于数据库系统在实际应用中所发挥的作用；同时，作为一个大型的数据库系统，不可能要求每一个用户都去购买一套价格昂贵的ORACLE 数据平台，而且就数据库系统的维护、更新、升级等等也需要专业人员操作，那么如何能够让一个数据库系统能够有更多的用户使用，网络解决方案就是一条根本出路。它忽略了物理距离，只要授予用户权限，用户就可以在任何支持网络的地方访问、使用数据库系统，这也是多数数据库系统要以网络形式开发的一个主要原因。

在网络服务刚刚兴起的阶段是比较流行的是两层的网络服务模式，即 client/server 模式。两层体系结构在实际应用中已暴露出一些问题。如：客户机直接（或通过存储过程）访问数据库，所有客户机均访问数据库，不利于安全控制，难以防止黑客的恶意攻击。同时，网络流量很大，易形成网络瓶颈。还会造成数据库访问瓶颈及数据库连接数过多，影响数据库的响应速度，降低系统性能。另外，两层应用体系结构还有维护、扩展方面的问题。

当前流行的网络服务结构是三层网络服务结构，由于分布式技术不断发展，在一些大型企业管理系统中，三层结构逐渐取代了两层结构。三层结构是在分布式技术成熟之后建立起来的，它的基本思想是将用户界面同企业逻辑分离，把信息系统按功能划分为用户层、中间层和数据层三大块，分别放置在相同或不同的硬件平台上。

（1）用户层：也称表示层，是信息系统的用户接口部分，即人机界面，是用户与系统间交互信息的窗口，主要功能是指导操作人员使用界面，输入数据、输出结果。它并不拥有企业逻辑，或只拥有部分不涉及企业核心机密的应用逻辑。

（2）中间层：也称用户逻辑层，是应用的主体，包括了系统中核心的和易变的企业逻辑（规划、运作方法、管理模式等），它的功能是接收输入，处理后返回结果。

（3）数据层：数据层即数据库平台，主要由数据库管理系统（DBMS）负责管理，包括对数据库的读写和维护等等，能够迅速执行大量数据的更新和检索。

三层结构与过去的两层结构相比有如下优点：

（1）进程管理：通过对服务进程的管理，使得在正常情况下，能用尽量少的服务进程处理尽量多的请求，减少进程的启动/终止次数。在峰值情况下，控制服务进程的总数，使得服务器在设定的负载下工作，不被压跨。总之，通过中间件对服务进程的有效管理，可以使系统在额定的功率下稳定工作，当请求服务的数量超过了服务器的处理速度时，中间件会把请求排队进行缓冲。

（2）保持和复用数据库连接：服务进程访问数据库都要和数据库建立连接，如打开和关闭数据库等。中间件通过采用长驻服务进程的手段，使得与数据库的连接被保持和复用，从而大大减少与数据库连接的次数和时间。

（3）优化了系统结构：将系统分为三层（或多层），业务逻辑放在应用服务层，软件的维护集中在应用服务层，客户端的维护就相对简单多了，有利于软件维护及系统管理。

（4）提高了应用系统的安全性：将客户端与数据库隔离起来，客户端无权限直接访问数据库，有利于安全管理，可有效防止恶意攻击。还可以利用中间件的安全管理特性进一步加强权限控制管理。

（5）卓越的扩展能力：若要提高系统性能、处理速度，可增加应用服务器，分担一部分应用服务工作即可，而原来的应用服务器几乎可以不动。

（6）减少网络数据流量和提高数据库响应速度：两层应用体系结构中客户机直接（或通过存储过程）访问数据库，会造成数据库访问瓶颈及网络瓶颈，从而降低了整个系统的性能。三层应用体系结构中，应用服务层的引入有效地解决了网络瓶颈和数据库连接数过多引起数据库性能下降的问题。应用服务层往往有多台服务器，可有效地解决客户机访问服务层瓶颈。应用服务器与数据库服务器（物理距离很近）可方便地采用宽带网连接，不会产生与数据库服务层网络瓶颈。

（7）提高系统性能：三层应用体系结构能更好地调整应用体系，还可利用中间件的特点来选择路由、平衡负载，提高整个系统的性能。

总的来说，三层应用体系结构使系统的性能、安全性、扩展性有了很大的提高，也方便了系统的维护和管理。

2.3.1.2 本系统的逻辑结构设计

作为一个网络环境下的高光谱数据库系统，借鉴当前流行的网络服务模式，本研究也将体系结构设计为基于一个三层的分布式环境（图 2－1）。这个三层系统模型，即客户层、应用逻辑层（含 WEB 服务器和各类应用服务器群）、数据库层（数据库服务器群）。在这三个逻辑层中的应用单元通过一组符合业界标准的协议、服务和软件连接器连接起来。这样，各级用户实际上处于理论上的平等地位，实际的区别取决于操作用户端软件的人员、单位的授权差异，这种差异由系统的安全控制部分管理和引导。

本网络数据库系统的功能结构划分：

（1）客户层：基于通用浏览器的各类用户与系统的人机界面，负责人机交互，接受用户提交的各种信息和要求，并以文字、图片或者曲线的方式反映给用户查询、分析的结果。

（2）应用逻辑层：一般包括应用服务器和网络服务器。是各类数据对外服务系统，负责应对用户端对数据中心的请求，与数据库层通信，提供数据和功能服务，对数据进行分析、处理，以及提供相关网络调节功能。

（3）数据库层：基于 ORACLE 平台，负责存储用户有权可访问利用的数据，它负责用户对直接数据库访问操作，这些访问操作由数据服务中心的服务系统根据用户访问的需求而制定。需要考虑数据的安全、一致性、易维护管理、易更新等指标。

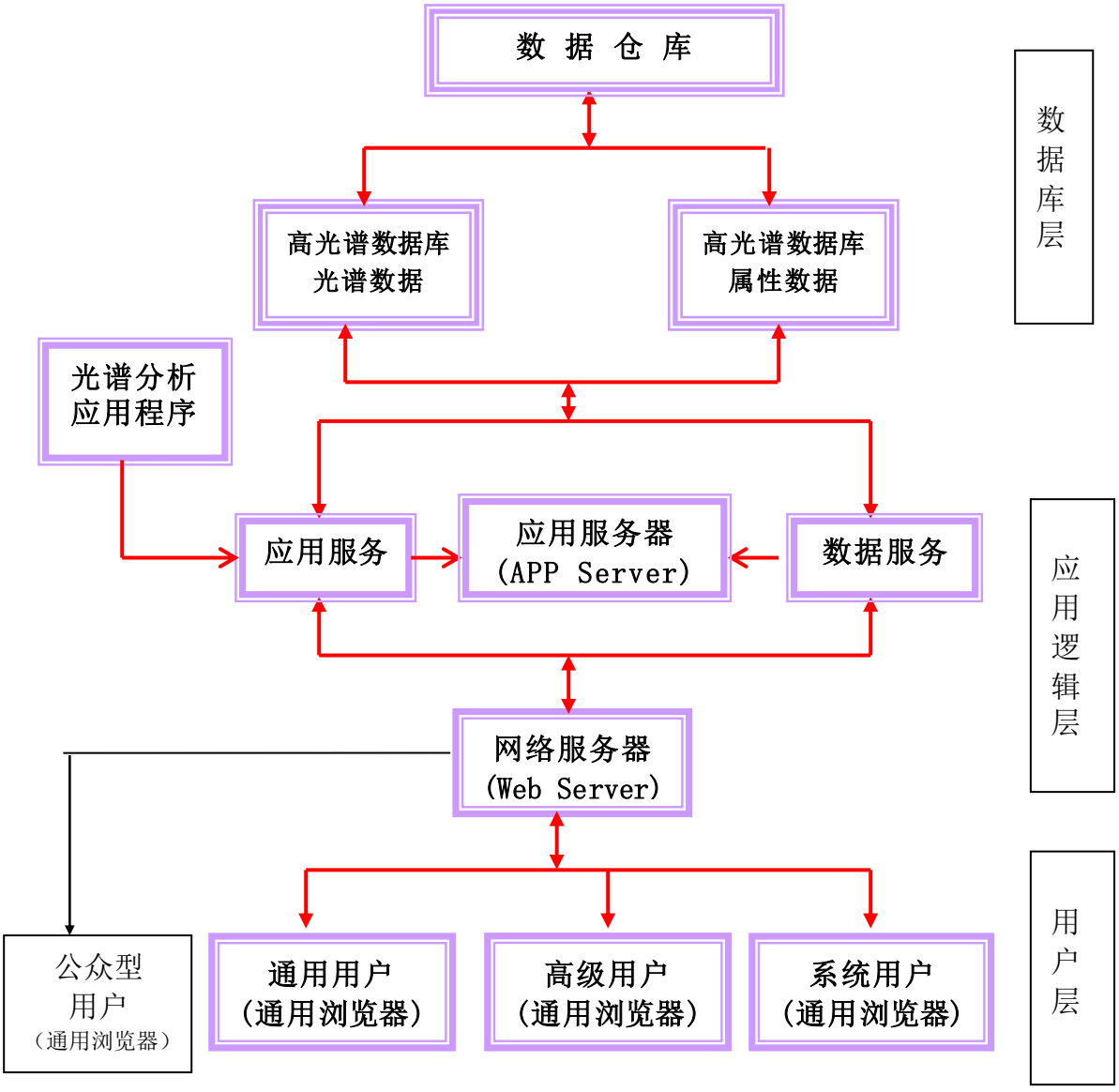


图 2-4 系统体系结构

2.3.2 系统物理结构设计

作为一个完整的数据库系统，硬件的合理配置对于最大限度的发挥系统的能力，保证整个系统的运行有着至关重要的作用。硬件的疏忽很可能会导致系统访问的缓慢、资源配置的不合理、瓶颈的出现等等干扰系统正常运行的问题。作为我们一个规模不大的应用系统，无法和真正的商用系统相比，所以这些设计可能并不是全部都能够实现，但是作为一个完整的系统设计这个方面是不可或缺的。其整体设计见图 2－5。

在开发实现中，数据库、网管工作站、应用服务器、网络服务器均集中与一台计算机上，防火墙则是集成在路由服务器中。

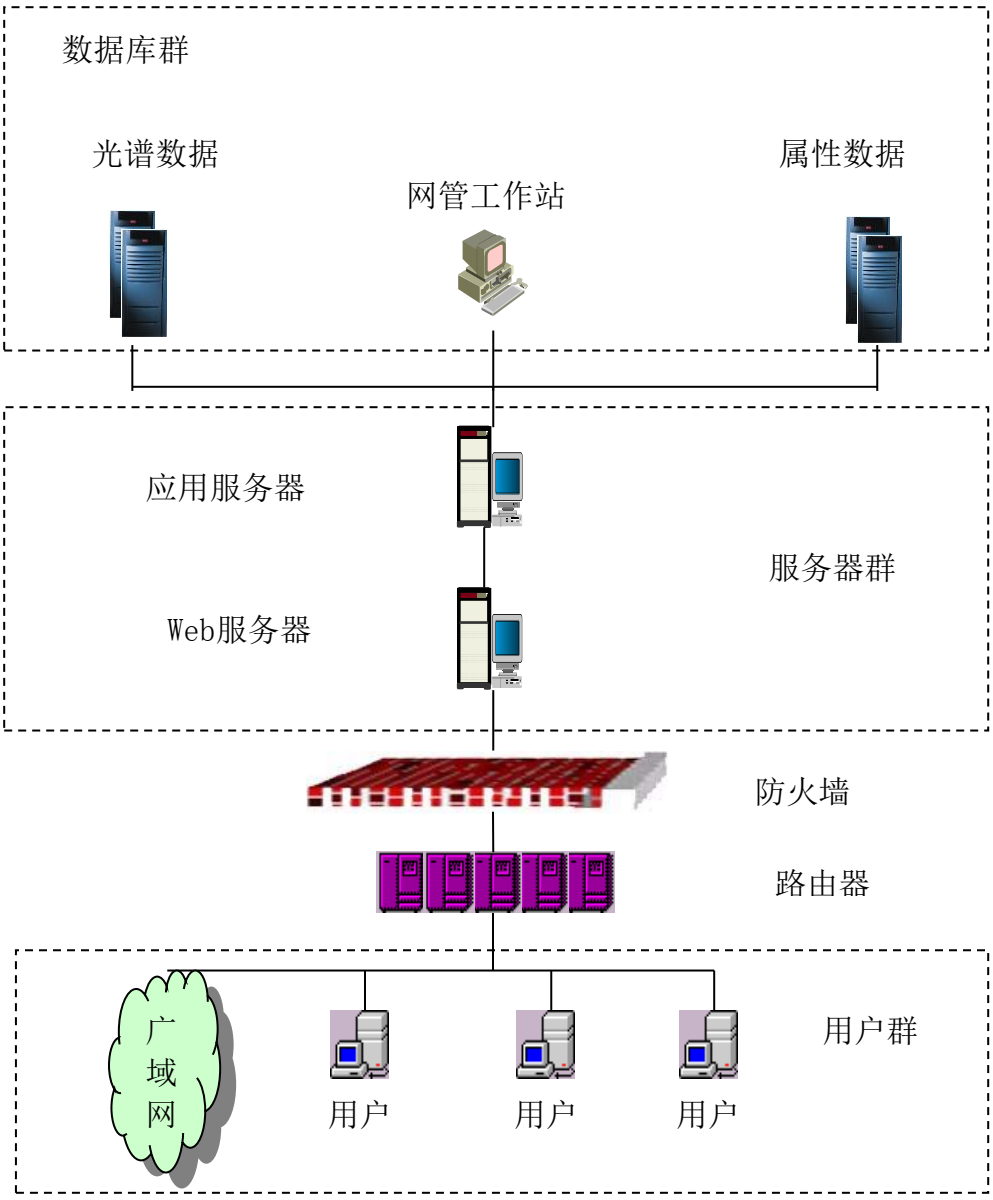


图2－5　高光谱数据库系统物理结构图

2.3.3 开发环境

基于上述设计思想，针对已有需求的特点，本系统立足于已有的环境，选用了一些比较常见的开发组建和开发语言，这对于降低开发难度，便于系统未来扩展，增强易维护性等等都是必要的。基本的开发环境是在 WINDOWS XP SP1+SP2 的环境下，后台数据库平台已经由需求选定在国际上通用的大型数据库开发环境 ORACLE 开发，该系统具有易于管理，可靠性好，便于扩展等优势，采用面向对象的编程方法，本系统采用的 ORACLE9i(9.0.0.1)。数据库访问则是由基本的 SQL 语言以及 ORACLE 中的 PL/SQL 语言开发完成。前台界面中的静态部分采用 html 语言，动态部分以及中间层数据分析模块的开发则采用 SUN 公司出品的 jsp 和 javabean 网络开发语言和组件，这套开发工具基本采用 JAVA 语言的语法，也是采用面向对象的编程方法，具有多平台适用性，方便灵活，简便易学，封装性好等特点，特别适合于网络开发。中间层选用了 Apache＋Tomcat 这一网络服务组件作为网络服务器和应用服务器，这是当前网络服务模式中非常流行的一个配置方式，具有全免费，配置灵活等特点，是 SUN 公司官方网站的推荐配置。整个系统最终结果具有易维护，易操作，便移植，便扩展功能等特点

2.4 本章小结

本章开始从总结系统需求：建立图像样本库模块；高光谱数据存储和应用的研究；网络环境下的发布；为光谱数据建立数据库；拓展原有的光谱数据库系统；建立光谱分析功能模块；设计系统的扩展性等入手， 结合本研究的具体内容，深入分析了高光谱数据库系统特点与难点：数据库 FOXPRO 平台向 ORACLE 平台迁移；高光谱数据的多变性、采集平台的多样性、图谱合一特性；网络环境下数据库结构设计；其他辅助属性的复杂性；图片数据存储方式以及分析功能模块设计等等，从而为全系统的开发做好了准备。同时，本章还介绍了全系统的开发环境以及为网络化应用而设计的，分为用户层、中间层和数据库层的三层整体系统结构。

第三章 基于网络的光谱数据库子系统

3.1 已有数据整理工作

在此研究的开始阶段，整理了原 FOXPRO 数据库的大量数据，而且重新整理、合并了其结构，虽然由于没有提出更具体的需求，最后实现时也没有这部分数据，但是这项工作为今后在此项目的基础上，整合、装入这些数据会起到很重要的作用。原库数据结构将 6 个部分的数据一共以 47 个表来存储，这样查询时可能会比较频繁的进行多表链接，容易降低读取效率，所以我将之整合为 8 个表，主要是将很多只表征一个字段属性的表合并入主表之中。因为库中的数据一旦输入，则主要是用作查询、检索，与会频繁修改的商业数据不同，所以合并在一起有利于提高效率。其关系图如图 3－1 所示。图中，文本框中的都是原库中的表名，线上为被合并表中的的主键，线下为合并表中的连接字段，线上没有标识的则是最后整理剩下的表。

合并之后的数据主要存储于 envirome，Instrume，Manmade，Rock，Rock_ana，Soil，Vegetati，Water 等 8 个表中，各自对应于通用属性及光谱数据、仪器相关属性数据、人工目标属性数据、岩矿属性数据、岩矿化学成分属性数据、土壤属性数据、植被属性数据、水体属性数据。这样就比较合理的兼顾了查询效率、存储空间等问题，为今后入库做好了准备。

3.2 数据来源

本系统光谱数据库子系统采用的数据源，主要来自于北京市自然科学基金资助重点项目北京市自然科学基金“面向精准农业的作物光谱数据库研究”的支持。其中的数据是 2002 年中科院遥感所与北京农业科学院合作在小汤山精准农业示范基地地面测得的，使用的仪器是美国生产的 ASD－500 高光谱仪，波段范围从 350nm－2500nm。这套数据体现了上面提到的高光谱数据的普遍特点，每个样本都有其高光谱数据，相应的配套属性数据，其中包括：测量属性数据，如：云量、风速等天气资料数据，仪器名称、测量时间等测量相关数据，叶面积指数、含氮量、含水量等理化参量数据，以及每个样本相应的图片数据。

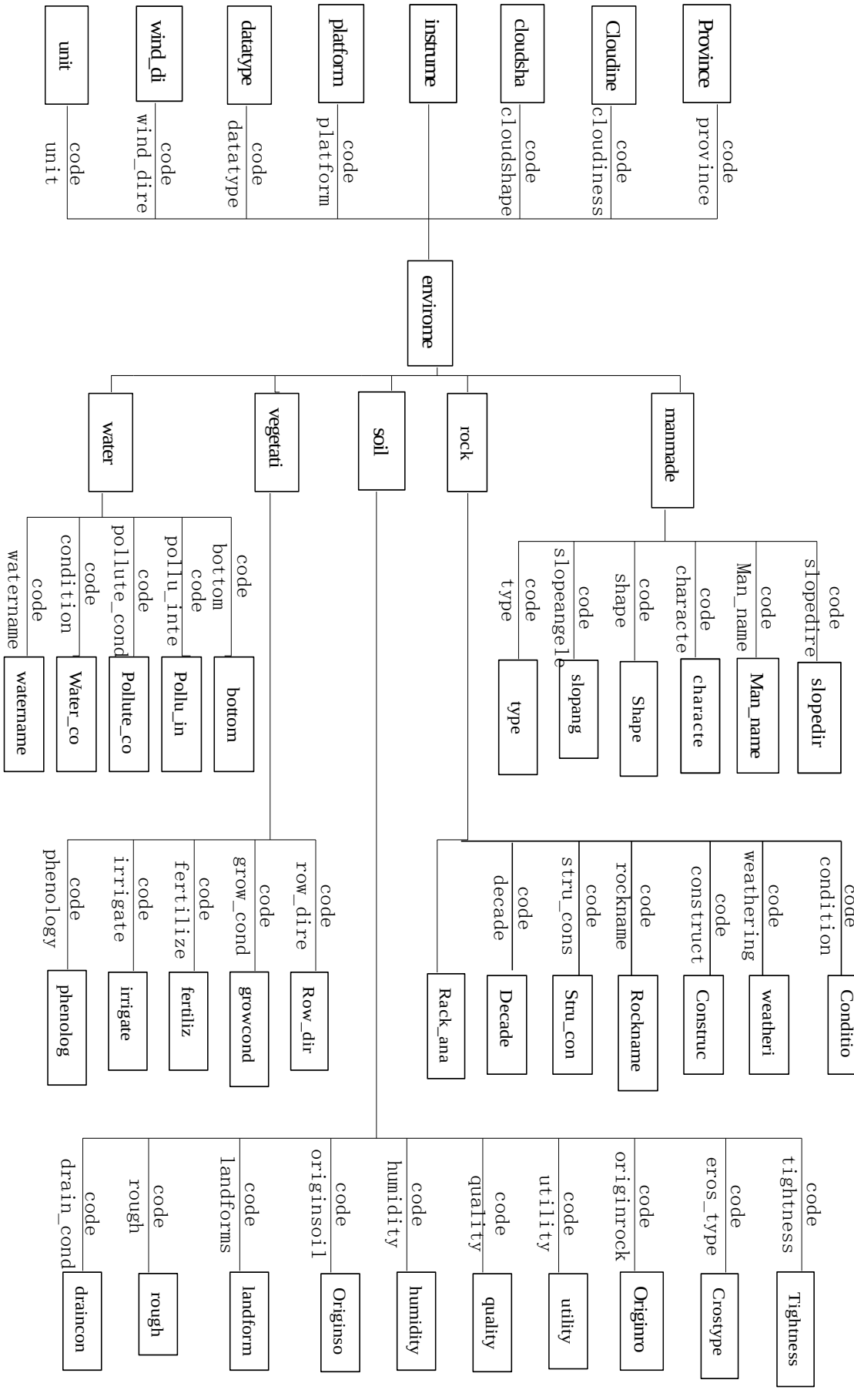


图3—1 FOXPRO高光谱数据库数据结构整理一览

3.3 数据结构设计

数据结构设计主要指数据在数据库中的存储模式，这是整个系统的根本所在，数据集中在这里，整个系统能否顺利的运行，运行的是否通畅很大程度上取决于这部分实现的好坏。这部分主要是由当前很为流行的大型数据库平台 ORACLE 来实现。

3.3.1 相关技术背景

ORACLE 数据库平台是有美国甲骨文公司（ORACLE）开发的，该公司自 1977 年诞生第一个数据产品，到现在已经发展到了 ORACLE9i，11i，其在数据库市场的份额一直都居于前列。这是整个系统的一个核心，主要负责数据的存放、整理、保护，结合这个应用需求而言，他的优势在于：

（1）ORACLE 是一个成熟的、大型的数据库应用开发软件，具有良好的可用性、易管理性、易维护性、易移植性和可靠的安全性。

（2）ORACLE 在管理上可以赋予不同用户不同的角色和权限，可以动态、静态多种方式进行数据库的备份，从而最大限度的保证了数据库的安全性。

（3）ORACLE 突破传统的关系数据库理论，引入了面向对象技术。借助对象的封装、继承和多态性，ORACLE 能以最便于应用的模式存放结构化和非结构化数据。这使得系统能很容易地用多维格式来存放数据。

（4）ORACLE 8i 是第一个专门为 Internet 的开发和发布而设计的数据平台,在网络方面的性能上比较出众，尤其是和同样在网络方面具有优势的 java 语言相结合，为本系统将来多用户、远程访问、大范围应用提供了技术保障。我们这里更是采用了更进一步的 ORACLE9i。

（5）ORACLE9i 进一步强化了 ORACLE 对网络功能的支持，在原有单一的数据库平台基础上，集成了 ORACLE 自己开发的网络应用服务器，即：IAS，其中包括自带了 Apache Web Server，并增强了对 JAVA 开发语言的支持， 例如：对 JAVA、XML、SQL 的一体化支持，更加完善的 J2EE 平台等等。

（6）ORACLE 具有强大的分布式数据处理能力和数据集成能力，这样为数据库系统在大型化之后，还能够保持良好的系统特性，比如：可靠的稳定性、访问的快速性等提供了有力的保障。

（7）作为一个大型数据库，ORACLE 为数据仓库的数据集成提供了有立的支持，提供了完善的数据仓库解决方案，也就是在存贮了大量数据的后期的综合分析工作方面也有不错的设计，这为本高光谱数据库系统未来的发展，发掘已有数据间潜在地联系提供了技术支持。

其主要结构如下图，数据库平台是全系统的心脏，是数据存放的核心，是各种诉求的最

终目的地，其安装、配置、性能调整本身也是一项复杂的工作。

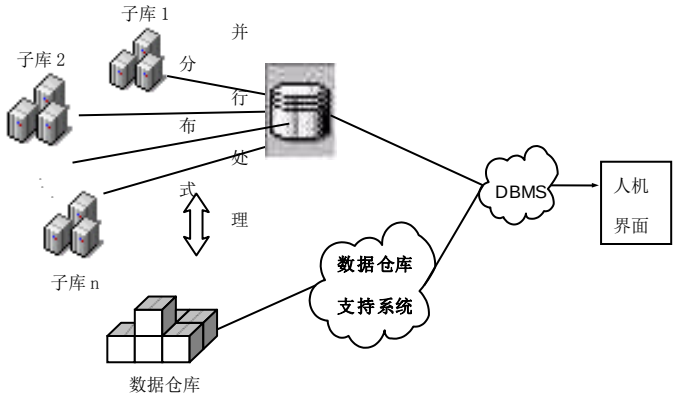


图 3－2　数据库结构图

3.3.2 数据结构设计流程

这里主要指的是数据库中存储结构的逻辑设计，其内容包括了对应于概念级的概念模式，即数据库管理系统要处理的数据库全局逻辑结构，也包括了对应于用户级的外模式。整个设计过程可分为 5 个阶段，如下图 3-3 所示。

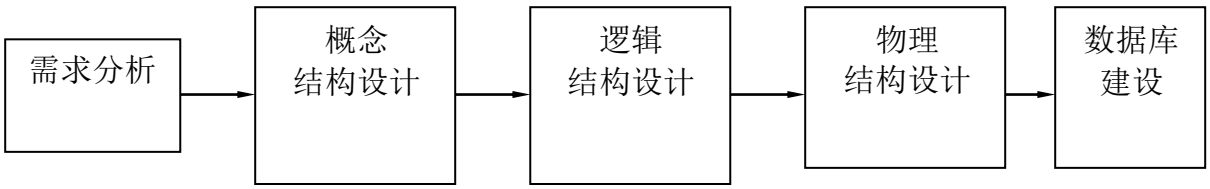


图3－3 高光谱数据管理子系统设计流程

根据上述流程，整个过程可以细化为：

（1）需求分析，上文已有，不再赘述；

（2）根据制定的测量标准和规范，对应于各个测量参量，设计概念模式，分析数据间关系，得到核心的数据作为主键，确定连接关系；

（3）结合关系数据库设计规范，考虑到数据存储的冗余性、数据查询的效率，进行合理的逻辑结构设计，规划到数据表这一层；

（4）结合硬件基础和数据库存放原理，进行合理的物理结构设计，以保证数据能够得到最有效的存放，并有一定的冗余度以保证数据的安全；

（5）实施数据库建设。

其中逻辑结构的设计是最为核心的部分，他决定了最基本的数据库存储单位－－表的结

构，这也就决定了整个系统中数据的输入、输出，它将根本性影响到整个系统的性能。

3.3.3 逻辑结构设计

逻辑结构设计主要是指数据在 ORACLE 平台下的数据库中存储的逻辑结构，具体就是数据表和数据表空间的设计。数据表是数据存储的基本单位，各种数据根据自身的特点以及互相之间的联系，按照 ORACLE 规定的一定的数据类型，设定表的字段设置，确定主键，确定连接方式。如果有需要，例如：有图片数据等数据量比较大的字段，则需要分配好数据表空间，以保证查询时不必反复的读取这一庞大的空间，从而提高查询效率。

本研究在提出针对需求设计表结构的同时，力图根据高光谱数据的特点，确定一个高光谱数据库系统中基本的数据存储规范，从而为今后开发高光谱数据库系统打下基础。针对高光谱数据的普遍特点是光谱数据＋图片数据＋各种属性数据，本研究提出的规范就是：应该采用：**光谱数据表组＋属性数据表组**的存储方式。每个表组中表的数量根据数据量的大小而定，两者通过能够唯一确定对象样本的字段进行连接，即：主关键字。光谱数据表组中的表主要存放对象的光谱数据，根据 ORACLE 设定的数据类型，又有几种存储方式，此处和图像样本库中分别采用了两种方式，最后将加以比较。而属性数据表组中的表则主要存放对象的各种其他属性数据，包括：测量属性数据、地学属性数据、其特征属性数据以及图片属性数据。图片属性数据一般需要单独开辟表空间存储，因为一般情况下，图片数据相对于普通属性数据属于需要占据较大存储空间的数据，分别存储有利于查询是提高效率，节省时间。

结合上述存储规范，本系统数据设计如下：

首先全部数据存放于两个表，三个表空间中，一个表是光谱数据表，表名为 wheatspectrum，主要存放光谱数据，见表 3－1；一个表是其他属性数据表，表名为 Wheatnature，见表 3－2，两者通过统一的联合主关键字连接，即 Sampleno+Sampledata 两个字段，因为该数据源是有时间和样本两者来确定唯一性的。三个存储表空间分别对应于两张表和存储图片数据的 picture 字段。

在存储光谱数据时，这里采用了展开存储的方式，即每个波段一条记录，目前一共有的是一天 48 个样本的光谱，每个样本有 350nm－2500nm，总共 2151 个波段的数据，也就是说目前光谱数据表中总共有 1×48×2151＝103248 条数据。这种存储方式的特点是：每个波段设定为一条记录，所以读取速度快，易于查找，结构简单，扩展性强，不受仪器变化，即波段总数变化的影响，对于后面数据分析、处理时提取数据十分有利。这种存储方式的缺点有：记录条数众多，如果是 100 天的数据，则会有一千万条数据，但是对于 ORACLE 这种大型数据库来说，仍然可以满足需要处理的需要；Sampleno+Sampledata 两个字段做的联合主关键字存储的冗余度比较大，占据了一定的存储空间，但是对于当前日益便宜的存储空间而言，是可以接受的，同时这种冗余可以通过在数据表上加索引（index），进行分区等数据库性能调整操作来进一步的提高查询检索的效率。

字段	字段类型	长度及小数位	显示名称	备注
Datano	Number	10	编号	
Sampleno	Varchar2	16	样本名称	
Wavelength	Number	4	波段	
Wavedata	Number	11，7	值	
Sampledate	Number	8	采样日期	

表 3－1 光谱数据表 wheatspectrum 结构

字段	字段类型	长度及小数位	名称	备注
Datano	Number	8	编号	
Sampleno	Varchar2	16	样本名称	
Kind	Varchar2	8	品种	
Season	Number	1	季相	0:冬小麦（春玉米）；1：春小麦（夏玉米）
Operation	Varchar2	16	处理名称	
Crown	Number	1	冠层/叶片	0：冠层；1：叶片(玉米与小麦相同)
Sampledate	Number	8	采样日期	
Stages	Varchar2	16	生育期	
Leafage	Number	4，2	叶龄指数	
Leafcolor	Varchar2	8	叶色	
plantheight	Number	4	株高	单位：mm
Cover	Number	3，2	覆盖度	百分比
Stemdiameter	Number	3，1	茎粗	单位：mm
Rowspace	Number	3	行距	单位：mm
LAD	Number	3	平均叶倾角	－90－－＋90
Leafdirec	Number	3	叶向值	－90－－＋90
Growthstatus	Varchar2	128	生长状况	非检索字段
Measuredate	Number	4	观测时间	前两位表示小时，后两位表示分
Instrument	Varchar2	16	测试仪器	
FOV	Number	3，1	视场角	
Height	Nmber	4	离地高度	单位：mm

Obangle	Number	2	观测角度	°
Windspeed	Number	4，2	风速	m/s
Winddire	Varchar2	4	风向	
Cloudine	Number	4，2	云量	%
Visibility	Number	2	能见度	Km
Sunangel	Number	2	太阳角	°
LAI	number	4，2	叶面积指数	
SLA	Number	6，3	比叶重	单位:ug/cm2
Ltemperature	number	4，2	叶片温度	°C
LWC	number	4，2	叶片含水量	%
RWC	number	4，2	叶片相对含水量	%
PWC	number	4，2	植株含水量	%
lchla	number	4，2	叶片叶绿素 a	%
lchlb	number	4，2	叶片叶绿素 b	%
lchlab	number	4，2	叶片叶绿素 ab	%
ltn	number	4，2	叶片全氮	%
lprotein	number	4，2	叶片蛋白质	%
lamylum	number	4，2	叶片淀粉	%
lcellulose	number	4，2	叶片纤维素	%
llignin	number	4，2	叶片木质素	%
lanthocyanin	number	4，2	叶片花青素	%
lxanthin	number	4，2	叶黄素	%
Photospeed	number	5，3	光合速率	
Risespeed	number	5，3	蒸腾速率	
schla	number	4，2	茎鞘叶绿素 a	%
schlb	number	4，2	茎鞘叶绿素 b	%
schlab	number	4，2	茎鞘叶绿素 ab	%
stn	number	4，2	茎鞘全氮	%
sprotein	number	4，2	茎鞘蛋白质	%
samylum	number	4，2	茎鞘淀粉	%
scellulose	number	4，2	茎鞘纤维素	%
slignin	number	4，2	茎鞘木质素	%
picture	Blob		照片	

表 3—2　属性数据表 Wheatnature 结构

3.3.4 管理系统设计

数据管理工作主要由 ORACLE 系统的数据库管理员，即 DBA 来完成，在数据库级别的数据管理、维护、备份，用户的身份认证、授权等采用 ORACLE 平台的管理工具即可，所以初级开发无需专门开发管理员界面，对于不是很复杂的系统这样有利于系统的易维护性。

这里需要特别提到的是，在大规模录入数据时，采用了 ORACLE 的批量数据录入工具－SQL*Loader。ORACLE 提供了多种批量录入数据的方式，如从 ODBC 设置连接，使用 SQL*Loader 工具等等，经过实际测试，SQL*Loader 更为方便，该工具允许多种输入文件，格式，数据类型；接受 SQL 函数；同时运行加载多个表；产生主键；从不同数据源加载；错误报告；使用高性能的“直接路径”加载等等,本系统初期均是将数据整理成文本文件,然后以 SQL*Loader 工具顺利导入。

3.4 网络服务系统设计

网络服务系统主要指中间层中的网络服务器和应用服务器组件，在整个系统中这部分起着纽带的作用，连接了前台的用户服务界面和后台核心的数据存储平台，还负责对网络协议的支持以发布到 internet 之上。下一章数据分析系统的程序也都挂靠在这一部分，由应用服务器负责和整个系统接驳。当前很多中小网站采用的是比较流行的 Apache＋Tomcat 组件配置，尤其是 Tomcat4.0 可以直接单独使用，本系统这部分采用的就是 Tomcat4.0

3.4.1 相关技术背景

Tomcat 是 jakarta 项目中的一个重要的子项目，其被 JavaWorld 杂志的编辑选为 2001 年度最具创新的 java 产品(Most Innovative Java Product)，同时它又是 sun 公司官方推荐的 servlet 和 jsp 容器(具体可以见 http://java.sun.com/products/jsp/tomcat/)，因此其越来越多的受到软件公司和开发人员的喜爱。servlet 和 jsp 的最新规范都可以在 tomcat 的新版本中得到实现。tomcat 随着 java 的流行，作为一个开源的 servlet 容器，其在 web 上的应用也越来越广，应用前景越来越广。tomcat 最新版本为 4.0.1，这个版本用了一个新的 servlet 容器 Catalina，完整的实现了 servlet2.3 和 jsp1.2 规范。安装该组件之前，系统必须安装了 java 的开发工具 jdk1.2 以上的版本。Tomcat4.0 包含三个主要的部分，包括：

（1）Catalina - 一个符合 Servlet API 规范 2.3 的 Servlet Container

（2）Jasper - 一个符合 JSP 规范 1.2 的 JSP 编译器和运行环境

（3）Webapps - Tomcat 中包含的一些例子和用于测试的 web 例程，以及相关文档。

3.4.2 服务平台比较

当前流行的网络服务平台主要有如下几种：Apache+Tomcat；微软的 PWS 和 IIS；以及 IBM 的 Webserver+weblogical 等。这三种各自有他们的特点：

（1）Apache+Tomcat：这套组件在上文已有说明，Apache 在解释静态页面上比较出色，以至于 ORACLE9i 之后，也将其作为自己的一个组件固化在产品之中，Tomcat 则在动态页面的解释上更为出色，4.0 以后更有兼容 apache 的趋势，也就是只要一个 Tomcat 就可以支持网络服务了。这套组件优势在于所有全部是免费的，而且对 java 系列的网络开发有着良好的支持，整个组件短小精悍；缺点在于相对来说可能需要比较复杂的配置，而且免费则缺乏相关售后服务。

（2）Microsoft PWS & IIS: 其全称分别是 Personal Web Server 和 Internet Information Server，是微软开发的网络应用服务器组件，其作为一个组件分别包含在 windows98 和 windows NT 系列操作系统之中，所以实质其实是一致的。它主要是针对微软的网络开发语言 ASP 所设置的，ASP 一直都是网络开发中非常流行的一种开发语言，在微软发布了.NET 技术之后，更有更进一步的趋势。这套组件的特点主要是由于他嵌入在大家常见的 windows 操作系统之中，所以安装、操作都相对方便，用户也比较熟悉，缺点则是安全性相对较差。

（3）IBM 的 Webserver+weblogical：这套系统是 IBM 开发的大型网络服务组件，这套系统的最大特点就是真正的商用服务系统，所以有着技术成熟、体系结构完善和良好的售后服务等特点，是商务应用系统的首选方案。缺点就是价格不菲。

市场中还有诸如 ORACLE 开发的 IAS 等网络服务组件，这里就不一一详述了，而我之所以选择了 Apache+Tomcat，主要则是由于本系统的前台和应用分析模块都是以 java 语言系列 JSP、javabean 等来开发的，而且系统规模相对较小，投入也不可能太多，所以选择了对 jsp 等支持良好而又免费的 Apache+Tomcat 这套网络服务组件。

3.4.3 系统配置及与数据库的连接

作为整个系统的纽带，网络应用服务系统起到的是承上启下的作用，承上是指和前台页面进行交互，接受用户提交的各种信息，这部分主要是由系统内部的一系列协议控制完成的；启下则是与后台的数据库平台进行通信，提交从前台接受信息而形成的指令，并从数据库中获取相应信息的过程。

网络服务系统和数据库连接在这里主要有两种方式：

一种是采用 ORACLE 自身携带的类库中的固定类，在连接程序中主要以 jdbc:ORACLE 的方式直接连接 java 的网络数据库连接驱动程序 JDBC 和 ORACLE 数据库；另一种则是以微软的数

据库连接驱动程序 ODBC 作为中间连接层，连接 JDBC 和 ORACLE,在连接程序中的表现形式主要为 jdbc:odbc:ORACLE。

显然，前者的连接更为直接，所以运行速度也会相对较快，但是由于本系统的开发还是基于 WINDOWS XP 平台，所以配置 jdbc 需要 java 类库的支持，相对比较复杂，同时为日后的扩展和移交等工作造成一定麻烦；后者则是以 windows 平台自带的 odbc 数据库驱动程序作为中间层，在设置上具有简便易行的特点，虽然可能速度相对稍慢，但是绝对不是系统瓶颈，亦不会对全系统性能造成根本影响，同时极大的方便了系统的扩展性、易维护易操作性，所以选用了后者。下图是底层连接的示意图。

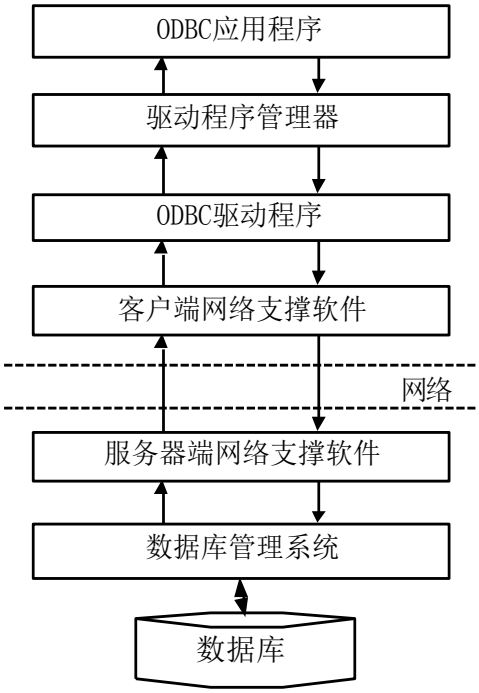


图 3-4 通过 ODBC 与远程数据库交互的框图

3.4.4 网络服务系统基本设计

本系统主要是通过网页等常见的网络形式，根据用户提出的需求，将数据库服务获得的内容发布，并向用户提供对象直观的属性数据和光谱曲线图，以及分析结果等信息。随着本系统的不断发展，将来还会不断添加各类与地物波谱相关的计算、显示的软件工具以及附加的直接使用的分析功能，同时还应具有调整网络性能等网络服务功能。

服务系统的运行性能设计主要考虑稳定、安全、容错、维护、效率等方面的内容：

（1）稳定性：在软硬件环境配置正确的条件下，系统能够长时间稳定运行，不宕机；

（2）安全性；能够进行严格的用户身份检查和数据保密性检查，不同权限的用户享有

对不同数据的浏览、查询等操作能力，禁止越权访问；

（3）容错性：对用户给出的各种查询条件出现错误时，应给用户明确的错误提示，并指导用户完成操作；

（4）系统维护简便，升级平稳，无须专业技术人员指导，维护人员即可根据使用说明完成系统的维护及升级操作；

（6）系统采用优化的查询及数据处理技术，减少用户请求的响应时间。

（7）在一定时间之内以电子邮件、硬复制、网页等形式响应各类用户的需求。

3.5 前台发布界面设计

前台界面是本系统的“脸面”，主要负责和用户交换信息，从用户处接收需求信息，然后提交的中间的应用逻辑层，同时得到中间层从数据库层返回的数据，并以数据、表格、图片、曲线乃至下载文件等方式显示给用户。

3.5.1 相关技术背景

在前台中，核心技术是 html 语言和 java 语言系列的 jsp、java bean，这是本数据库的另一个核心，作为整个数据库系统的前端界面（GUI），使得用户能够和数据库进行良好、友善的交流，同时支持作为日后国家级的数据库的网络使用能力。JAVA 语言与 C、C++语言是当前软件开发人员最为常用的两种语言，具有很多优秀的特点:

（1）Java 是标准的面向对象的编程语言，符合当前的编程语言发展趋势，具有良好的结构化编程方法，程序易读性强，具有多线程处理能力，无论在网络还是非网络的开发都可以胜任。

（2）Java 具有良好的可移植性，利用自己独创的 java 虚拟机技术（JAVA VM）可以无视平台的差异，更在前阶段与微软公司的官司中取得了胜利，强制在 windows 系统中加入 java 虚拟机，为 java 的发展起到了重要的作用，对于国家优先发展的 unix/linux 系统更是有着天生良好的兼容性。

（3）Java 在网络方面具有特天独厚的优势，基于 java 技术的 jsp，javabeans，java servelet，java applet 等专门技术，同时也提供了很多免费的功能组件和插件，为大型数据库的远程多用户访问提供了坚实的技术基础。

（4）Java 具有良好的易学易用性，避开了 c＋＋中最为艰深的指针概念和内存管理，自动收集、分配资源，封装性好，组件丰富，jsp、javabean 等网络开发组件的语法与 java 几乎一致，无需单独学习，从而为快速、广泛的培训编程、使用、维护人员，提供了非常便利的条件。

（5）Html：超文本标记语言，自从万维网（WWW）蓬勃发展以来，一直是网络开发的最基本技术，简单易学，可以被浏览器简单的判读，是最常见的 internet 技术。XML，扩展标识语言发展，是其未来的发展方向。

选择 JAVA 语言系列的一个主要考虑因素就是 ORACLE 和 java 的紧密关系。ORACLE 的开发一直都对 java 语言给予了极大的支持，其第一个测试版本也是以 java 语言测试的，基于对网络方面的考虑，在 9i 之后更是在自己的数据库产品中加入了对 java、xml、sql 等的一体化支持，其不但采用了 apache 作为其网络服务器，其自己开发的应用服务器组件 ORACLE IAS 更是 100％的 J2EE 架构。基于以上这些因素，以及 java 在网络开发方面的优势，本系统采用了 SUN 公司的 JAVA 语言系列进行开发。

3.5.2 界面设计与开发

在界面设计和开发中，本系统充分调研了当前国内外相关类型数据库的界面方式，并结合了自身需求以及高光谱数据的各种特点，力求达到让用户感觉友好，充分考虑到用户使用的便捷性，并且要具有相当的容错性及错误提示，来保证整个数据库系统能够方便、顺畅的运行。

界面设计的第一步是要根据需求，确定页面所要提供的功能模块与选项，其次按照功能模块以及用户的需求设定连接关系，即网络站点结构，然后进一步细化需求，确定输入、输出的方式，最后确定每一界面的具体形式。根据这些设计，就可以逐步的设定类、变量等以 html 和 jsp 等技术进行开发。

3.5.2.1 界面功能结构

界面功能结构设计实际上就是将需求模块化，将同样的功能聚类，比如：样本光谱的显示和图像样本库中图谱合一的光谱显示就可以统一为一类功能，这样可以为前台开发理清思路，同时为其进行适当的页面规划，为后来设计站点结构和页面打好基础。在本系统中，主要的模块有如下几个方面：数据库的维护操作、属性数据的查询和浏览、光谱数据的显示以及光谱分析功能。如图 3－5 所示。

3.5.2.2 站点结构图

根据整理好的功能结构模块，结合用户的需求，考虑到使用情况，下一步就是设计整个网络前台界面的站点结构，也就是整个网络系统的连接关系。如下图 3－6 所示。

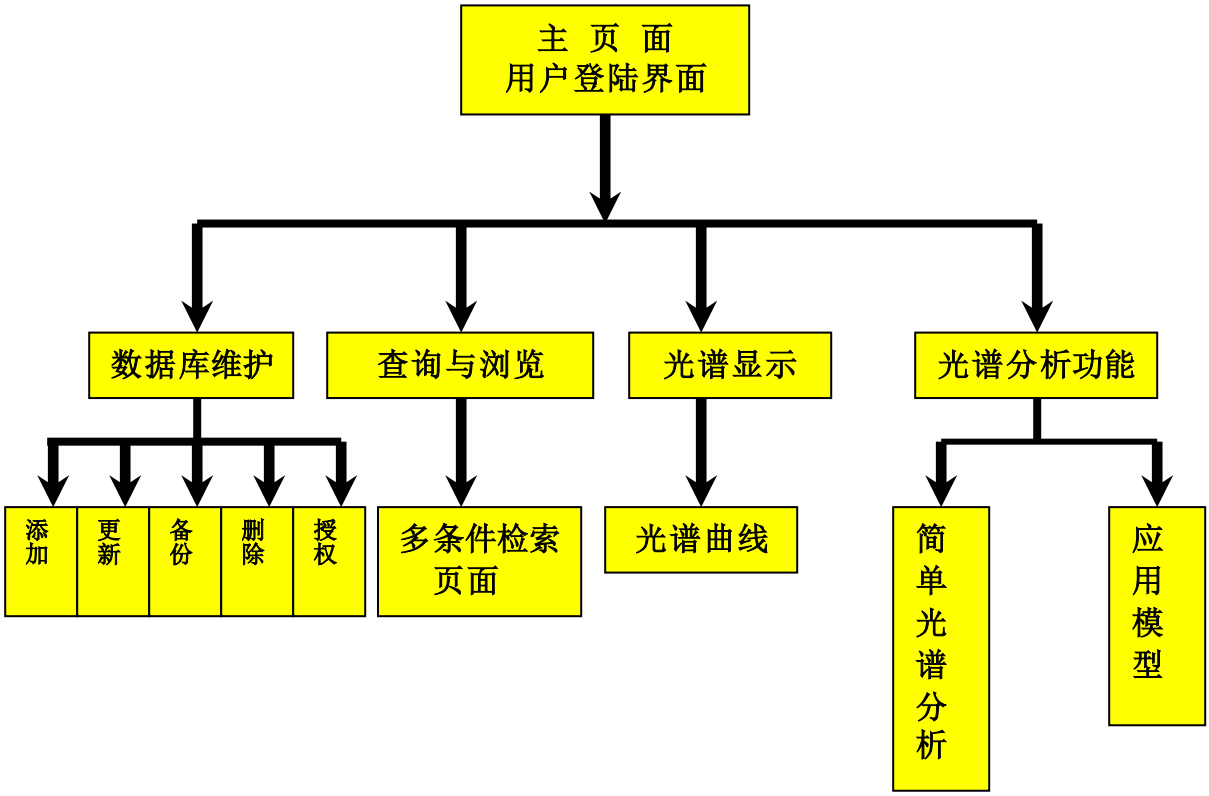


图 3－5　功能界面结构图

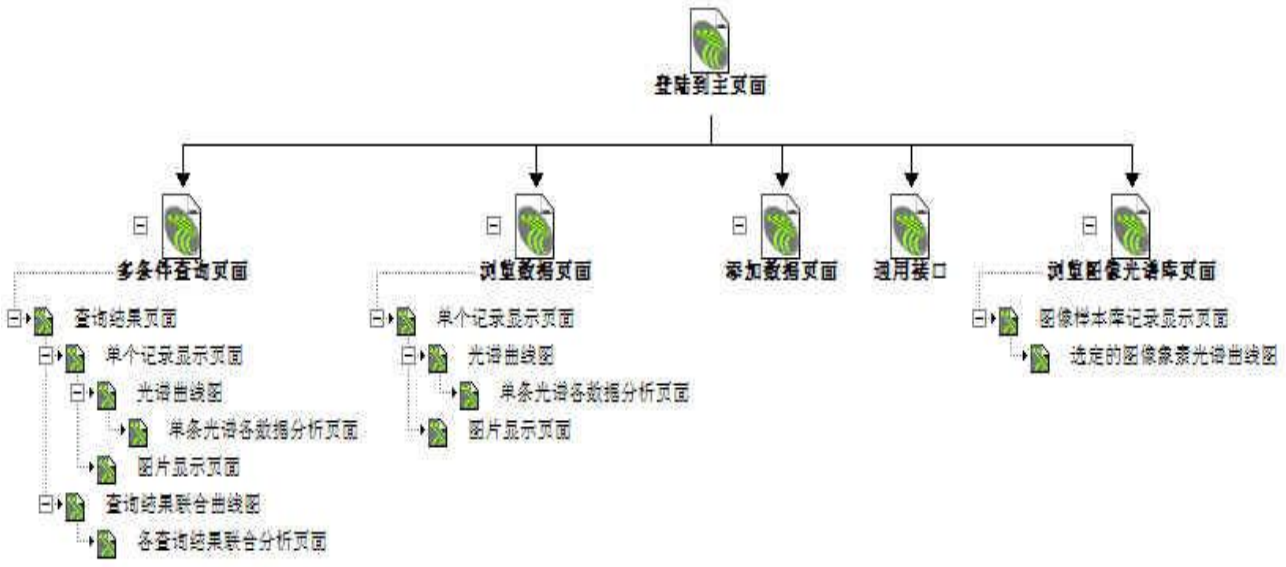


图 3－6 系统网络站点结构图

3.5.2.3 界面样式

当前网络已经不再是很新鲜的东西，各种网站也层出不穷，因此很多时候借鉴已有的成果可以降低开发的难度，同时提高界面的友好性。在选择界面样式的时候，本系统参考了国内外一些网络数据库的界面样式。比较典型的几个例子有：

（1）国外的 science@direct 文献数据库：（www.info.sciencedirect.com）

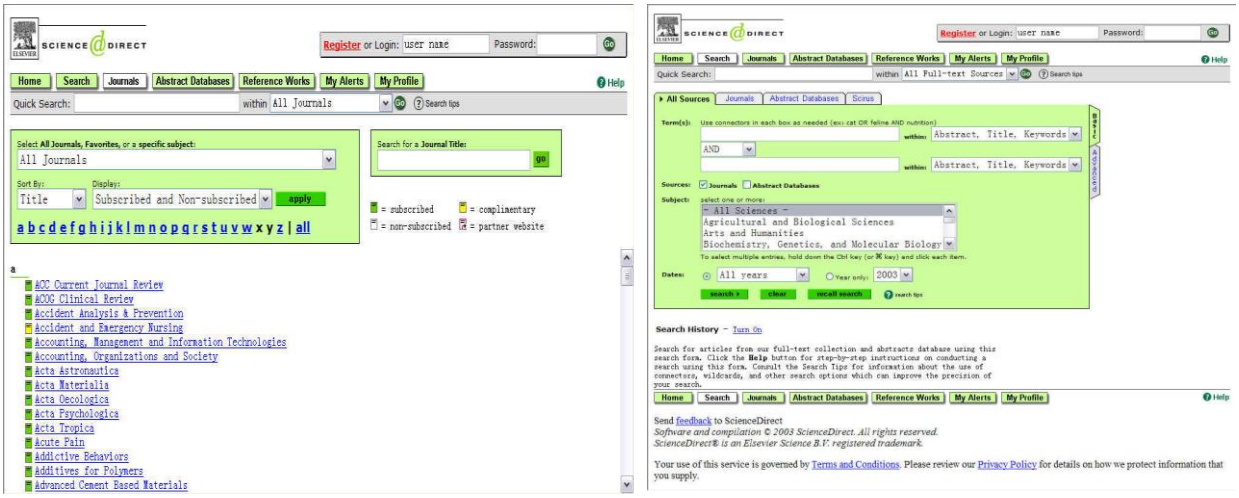


图 3—6 science@direct 文献数据库界面

（2）中国科技书刊数据库：（http://159.226.100.28）

图 3—7 中国科技书刊数据库界面

（3）中国水生维管束植物图谱数据库：（http://159.226.162.2）

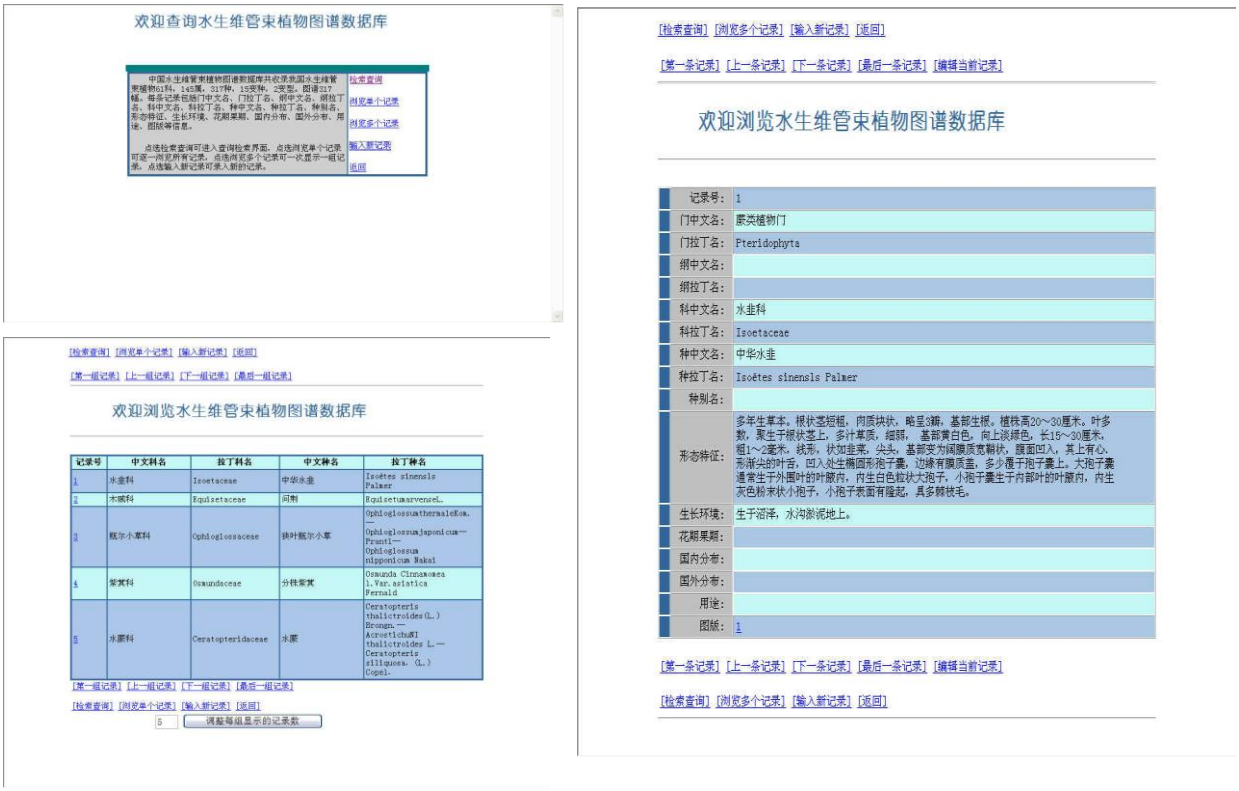


图 3—8　中国水生维管束植物图谱数据库界面

从这三个数据库的样式可以看出，2、3 两个样式比较简洁，界面比较友好，2 属于文献数据库，3 则属于返回数据的数据库，所以两者的数据性质有区别，这也导致两者在技术选择上的不同，2 使用了 html 中的 frame 技术，而 3 主要应用的是 html 中 table 技术。一般来说 table 较 frame 有读取速度快，格式简洁，不易出错等优点。

同时考虑到本系统特点，本系统属于需要返回数据的数据库，尤其是数据的属性繁多，使用 frame 则会出现浏览数据时需要频繁翻页的情况，再考虑到 table 的优势，选用了 table 作为主要技术，考虑到本系统属性过于繁多，与 3 的特点不是很符合，所以又借鉴了中科院遥感卫星地面站开发的“遥感卫星图像检索数据库”的界面样式（cbs.rsgs.ac.cn）：

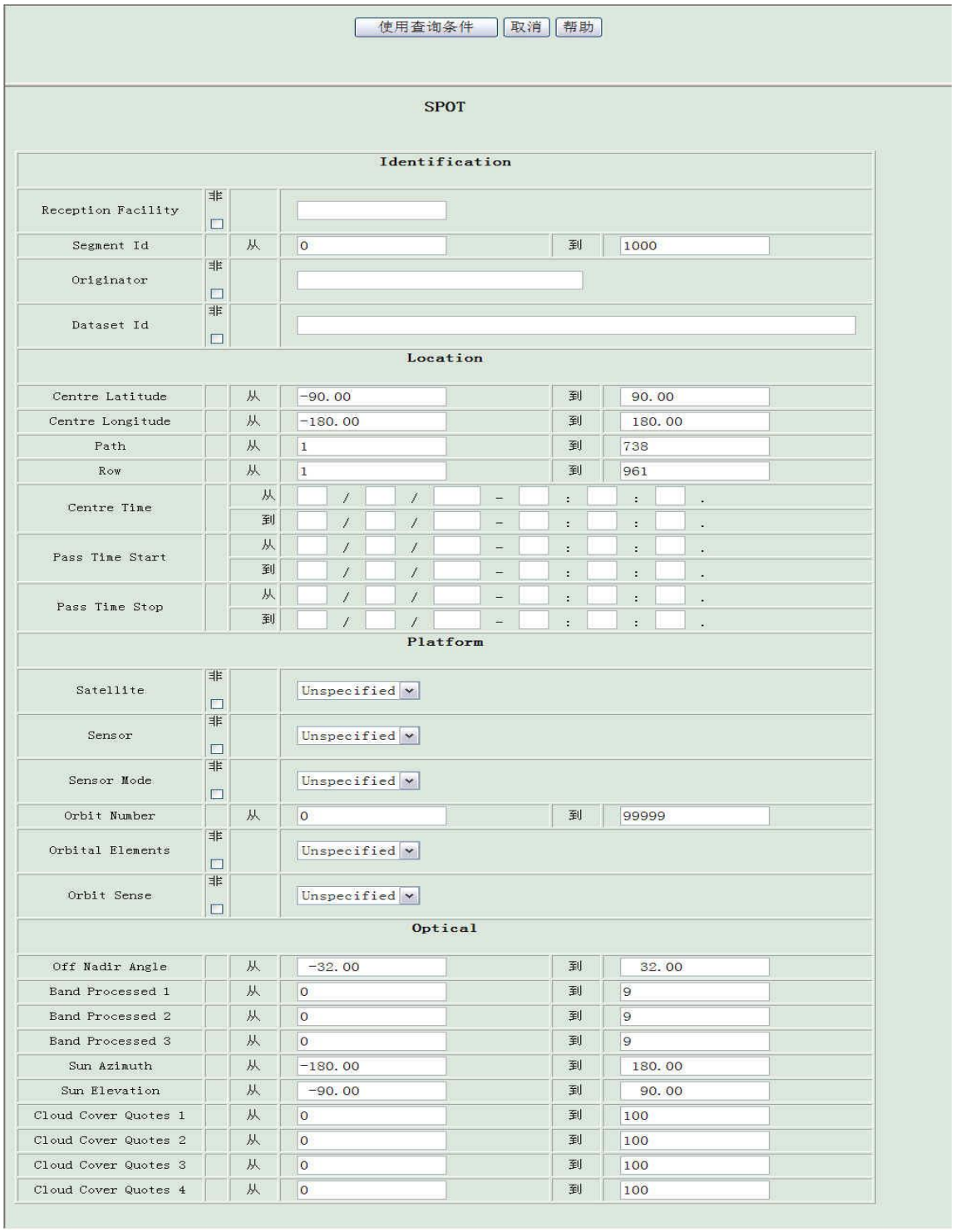


图 3－9　　遥感卫星图像检索数据库界面

该界面与本系统最大的共同点就是需要罗列的查询条件以及返回的数据繁多，这为本系统的建立提供了良好的参考。所以，综合以上各个界面的优点，以及本系统数据的特点，最后本系统页面的主要样式为：

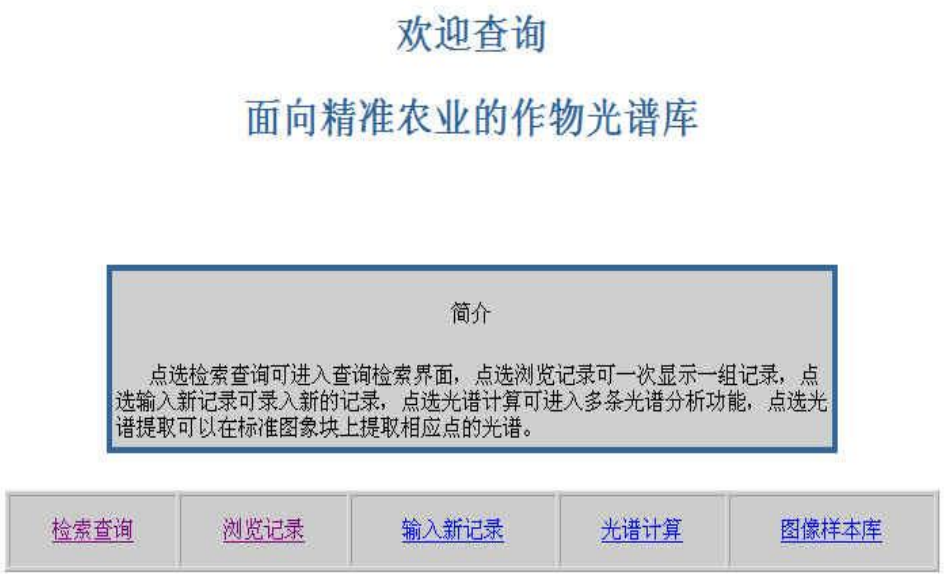


图 3－10 主页面

图 3－11 多条记录浏览及查询结果页面

图 3－12
查询及单条记录显示页面

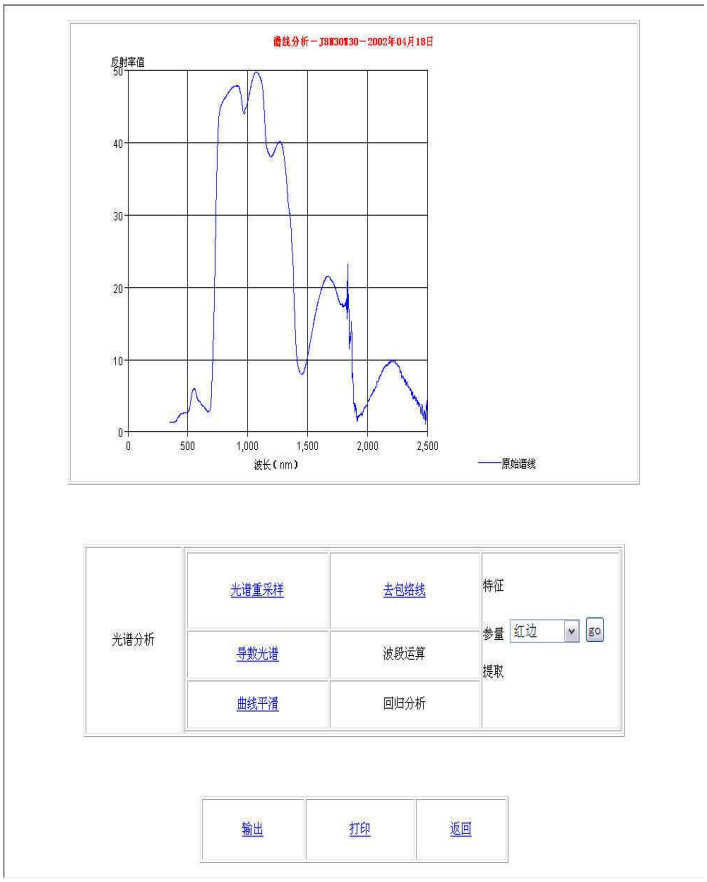


图 3－13 样本光谱显示及分析页面

3.5.3 属性、图片与光谱的发布

整个前台各种数字以及文字属性的发布都是由 JSP、Javabean 通过中间层向数据库执行命令得到的。图片的发布是以二进制数据流的方式从库中提取，并同样以二进制数据流的形式用 JSP 中固定的函数向浏览器发布的。

光谱曲线的发布，首先根据用户提供的信息或查询或浏览从属性数据表中得到主关键字，然后以此主关键字在光谱数据表中找到该样本的全波段波段值和反射率值，以两个数组的方式提取出来，分别作光谱曲线的 X 轴和 Y 轴。曲线的描绘有两种方式：一种是生成图片后，

存放于客户端，然后同样以二进制数据流的形式传到客户端；一种则是由 java applet 根据数组数据直接生成曲线图。本系统采用了后者的方式，并且选用了一个 java 的曲线图描绘插件 Kavachart，通过将两个数组数据以及其他一系列固定的参数传过去，该插件即可直接将曲线图显示在客户端。

还有就是一些其他操作的选项，比如翻页、添加、删除、修改等等，也都是通过在 html 中获得相应的标记参数，并通过传递参数执行数据库的相应命令。

3.6 本章小结

本章主要介绍了本系统的光谱数据库子系统的设计、思考与实现，同时介绍了全系统的网络中间层和前台界面系统的设计与实现，提出了高光谱数据在关系数据库中的存储规范：**光谱数据表组＋属性数据表组**。光谱数据库子系统实现了和农业信息中心合作项目需求的最主要部分，也就是数据库相关功能部分，包括普通属性数据、图片数据、光谱数据等的查询、检索、添加、删除、修改等等功能。它通过后台的数据库平台，中间的网络应用服务平台以及前台开发的界面平台有机的结合在一起，形成一个有机的整体，从而实现了一个网络化数据库系统的良好运行，也将国内的光谱数据库系统的水平向前迈进了一步。同时，作为高光谱数据库系统的一个子系统，其数据库组织结构完全可以满足全系统的要求，解决了第一章提到的光谱数据库子系统的建立问题。本章还将全系统的网络服务结构和前台界面设计的思路与实现做了比较详细的描述，以后将不再赘述。

第四章 基于网络的数据分析子系统

前一章讲述了整个光谱数据子系统的设计与实现，实现了数据库的基本功能。但是本项目还涉及到了数据处理、数据分析等基于数据库之上的更高级的需求。本章将针对这一需求，讲述基于网络，以及建设与数据管理系统之上的数据分析子系统。

4.1 分析功能简介

本系统在基本的数据库功能需求之外，还增添了数据分析的需求，也就是需要根据农学数据和高光谱数据的特点，对数据进行运算，得到一些对农学有价值的参量，并且对光谱做一定的变换以突出某些光谱波段的特征。具体讲需要的分析功能与变换主要有，光谱重采样；去包络线；有限特征参量提取，包括：红边、红谷、绿峰、黄边、NDVI 等；光谱和生化参量线性回归统计分析；导数光谱等等。同时本系统还将与其他的光谱模拟及反演的模型相连接，所以需要提供一定的模型接口。

根据上面提到的分析需求，本节将一一详述各项分析功能的原理和应用。下列各公式中未标注的： λi 为波段 i 波长值，ri 为波段 i 的反射率值。

4.1.1 光谱重采样

光谱重采样主要是在任意的波段间隔中，变换波段间隔重新采样，这样可以放大某部分区间，也可以模拟某个传感器的数据。其对应的处理程序文件为 resample.jsp

需要输入的数据有： 起始波段：r1 结束波段：r2 间隔：b

计算公式为：

$$r_i' = \frac{\sum_{i-\frac{b}{2}}^{i+\frac{b}{2}} r_j}{n+1}$$

ri’：重采样后的 λi 反射率值 ri：重采样前的 λi 反射率值 n：b（b 为偶）：b＋1（b 为奇）

4.1.2 曲线平滑

曲线平滑的目的主要是去除光谱曲线中的毛刺和随机噪声，保留主要的光谱数据，是曲线更加平滑，其对应的处理程序文件为 calculate.jsp

其计算公式为：

$$r_i' = \frac{r_{i-1} + 2r_i + r_{i+1}}{4}$$

4.1.3 NDVI

归一化差异植被指数，这个参数广泛应用于农学的应用，主要表征了植被地覆盖度和绿度。NDVI 的时间序列可以表征农作物长势的时间变化以及物候历。其对应的处理程序文件为 calndvi.jsp

其计算公式为：

$$NDVI = \frac{\lambda_{800} - \lambda_{680}}{\lambda_{800} + \lambda_{680}}$$

4.1.4 导数光谱

主要是突出植被和目标物体的信息，消除背景信息。也可以用于确定下面的红边等信息其对应的处理程序文件为 derivative.jsp

其计算公式为：

$$R'(\lambda) = \frac{R(\lambda_{i+1}) - R(\lambda_{i-1})}{\lambda_{i+1} - \lambda_{i-1}}$$

其中：R’代表反射率的一阶导数光谱，R 代表反射率，λ代表波长，i 代表光谱通道

4.1.5 农业光谱特征

这些特征或是波段位置，或是某计算值，都表征了很多农作物的生理、生化参量的变化规律，对于估算生化参量等有着重要的意义。

1、红边、红谷、红边面积：

其对应的处理程序文件为 red.jsp

其计算公式为：

红边： 在 680－750nm 之间的光谱数据求导，其中最大值为红边斜率 Srg，对应的 λi 为红边位置。

$$d_i = \frac{r_{i+1} - r_{i-1}}{\lambda_{i+1} - \lambda_{i-1}} \quad \text{Srg=max(di)}$$

红谷： 在 550-680nm 之间光谱反射率值极小点对应的 λi 为红谷位置

红边面积：对 580-700nm 间的光谱反射率值做去包络线处理，在求得的去除包络线之后的曲线所与 X 轴，即波段值轴，围成的面积 S。

$$S = \sum \frac{1}{2}(r_{i+1} + r_i) * 1$$

2、 黄边：在 550－650nm 间光谱反射率的一阶导数的极大值为黄边，其对应的 λi 为黄边位置。其对应的处理程序文件为 yellow.jsp。

其计算公式与红边位置计算公式相同。

3、绿峰：在 500-600nm 间光谱反射率的极大值，其对应的 λi 为绿峰位置。其对应的处理程序文件为 green.jsp。

4.1.6 包络线消除算法

其对应的处理程序文件为 antidconnium.jsp

一般来说，由于地物组成复杂，每个图像像元点对应的地物并不纯粹，它的光谱通常是多种物质光谱的合成，因此直接从光谱曲线上提取光谱特征不便于计算，还需对光谱曲线进行进一步的处理以突出光谱的吸收和反射特征。为此，我们引入了包络线消除算法。

光谱曲线的包络线从直观上来看，相当于光谱曲线的“外壳”，如图 2 所示。因为实际的光谱曲线由离散的样点组成，所以我们用连续的折线段来近似光谱曲线的包络线。

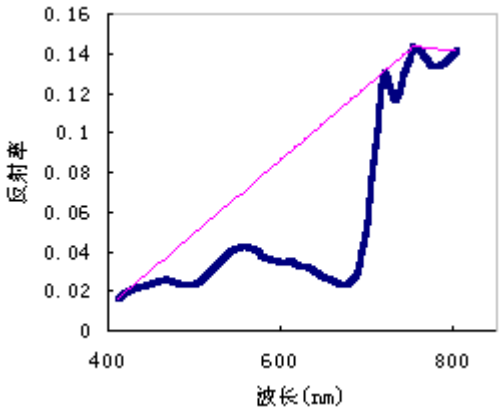


图 4－1　光谱曲线及其包络线

求光谱曲线包络线的算法描述如下：

设有反射率曲线样点数组：r（i），i=0，1，…，k-1；

波长数组：w（i），i=0，1，…，k-1；

1、i=0，将 r（i），w（i），加入到包络线节点表中；

2、求新的包络节点。如 i=k-1 则结束，否则 j=i+1；

3、连接 i，j；检查（i，j）直线与反射率曲线的交点，如果 j=k-1，则结束，将 w（j），r（j）加入到包络线节点表中，否则：

（1）m=j+1；

（2）若 m=k-1 则完成检查，j 是包络线上的点，将 w（j），r（j）加入到包络线节点表中，i=j，转到 2；

（3）否则，求 i，j 与 w（m）的交点 r1（m）。

（4）如果 r（m）<r1（m），则 j 不是包络线上的点，j：=j+1，转到（3）；如果 r（m）≤r1（m），则 i，j 与光谱曲线最多有一交点，m：=m+1，转到(2)。

4、得到包络线节点表后，将相邻的节点用直线段依次相连，求出 w（i），i=0，1，…，k-1 所对应的折线段上的点的函数值 h（i），i=0，1，…，k-1；从而得到该光谱曲线的包络线。

显然有：

$$h(i) \geq r(i)$$

求出包络线后对光谱曲线进行包络线消除：

$$r'(i) = r(i) / h(i),\ i=0, 1, \cdots, k-1;$$

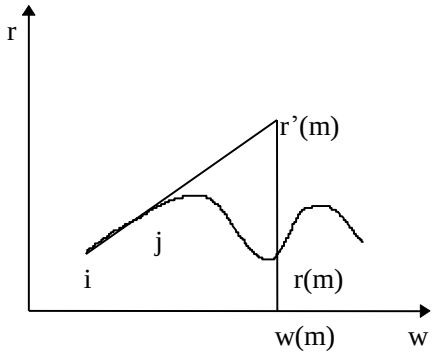


图 4－2 包络线去除算法示意图

如图 4－3 所示，左边为原光谱曲线，右边为包络线消除后的光谱曲线。进行包络线消除后将反射率归一化到 0∽1，光谱的吸收和反射特征也归一到一个一致的光谱背景上，并且得到了很大的增强，因此可以更加有效的和其他光谱曲线进行光谱特征数值的比较，进行光谱的匹配分析。

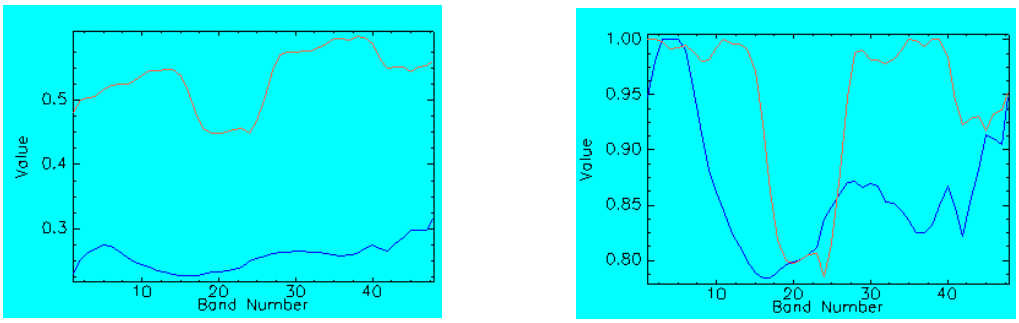


图 4－3 包络线消除前后的光谱曲线图

4.1.7 线性回归分析

在本系统中，不但有光谱数据，有光谱数据对应的各种属性数据、理化参量，还有新计算出来的各种参量，如：NDVI 等。在农业应用中，根据这些光谱及其计算得到的光谱参量与叶面积指数（LAI）等农业理化参量进行相关性分析是有很重要的意义的。最简单易行的就是进行线性回归分析。其对应的处理程序文件为 analyze.jsp。

问题可以描述为：有数列 x（n），y（n），欲求其线性关系，即 y＝ax+b，求 a，b。本系统

中则是以 NDVI 以及各波段数据作为 x（n），以叶面积指数（LAI），叶片含氮量等作为 y（n），根据用户的选择求其线性关系。

其计算公式为：

$$a=\frac{\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y})}{\sum_{t=1}^{n}(x_t-\bar{x})^2} \qquad b=\bar{y}-\bar{x}a$$

4.2 子系统设计与实现

数据分析子系统如前面总体结构所讲，主要是以 JSP 语言将这些分析功能以程序的形式实现，挂靠在中间的网络应用层，从前台界面中得到用户的输入，向后台数据库平台查询得到需要数据，然后在经过运算得到结果。简单数据就直接也窗口的形式返回，曲线分析则同样通过数组传到 kavachart 插件，以曲线图的方式反馈给用户。

以叶面积指数（LAI）＝3.5 为查询条件得到的数据，并进行各个分析功能，得到的结果如下列各图所示。

图 4－4　查询结果列表

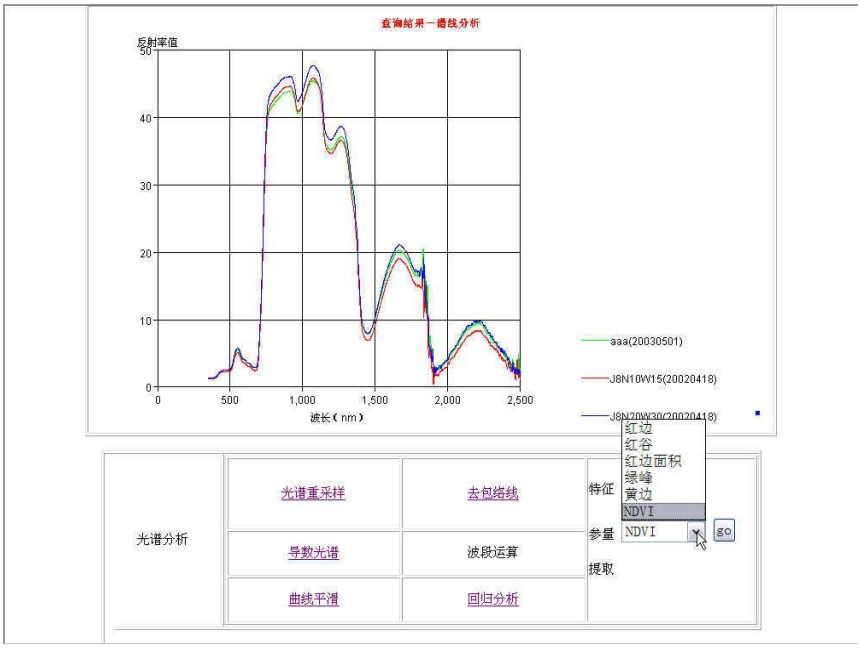


图 4—5 查询结果光谱曲线图

图 4—6 重采样输入界面

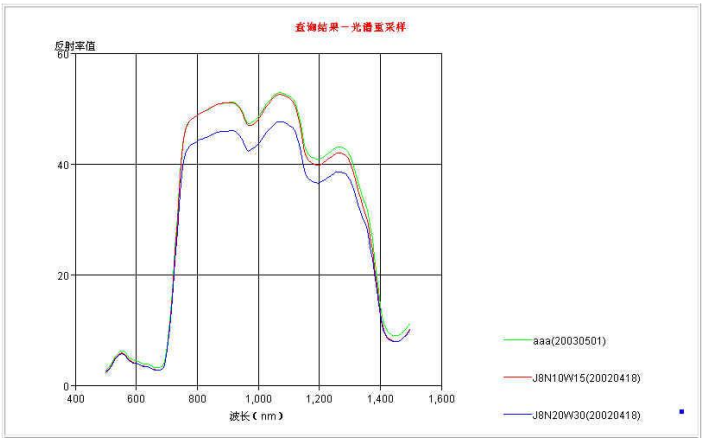


图 4—7 重采样结果曲线图

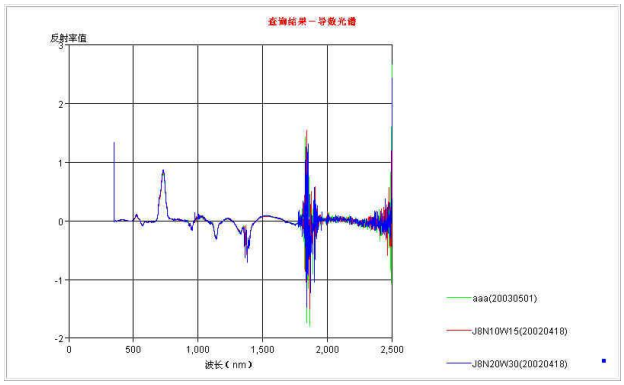


图 4—8 导数光谱曲线图

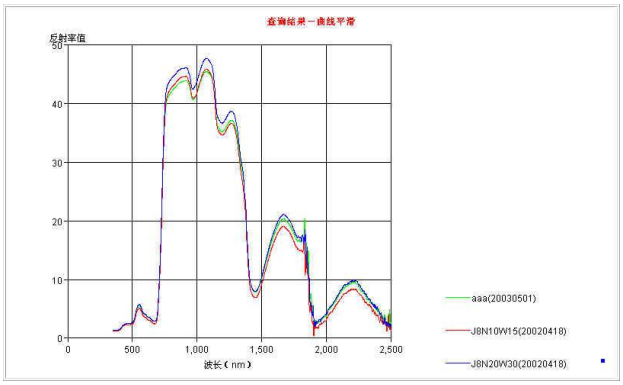
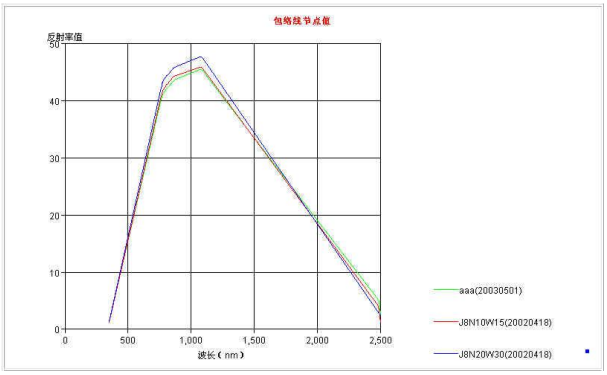


图 4—9 曲线平滑光谱曲线图



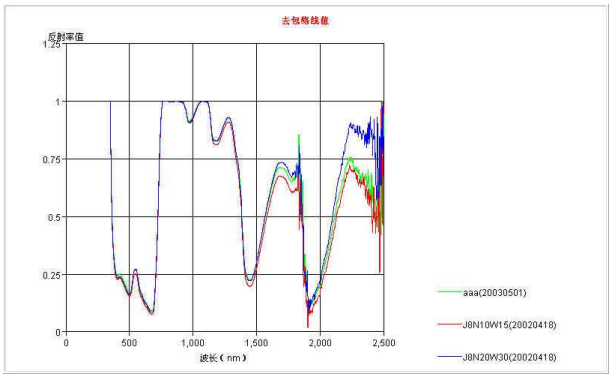
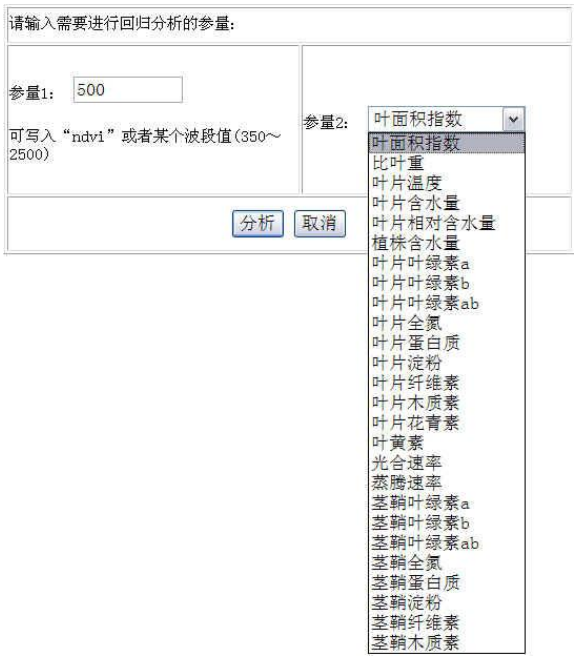


图 4—10 包络线节点及去除包络线后光谱曲线图



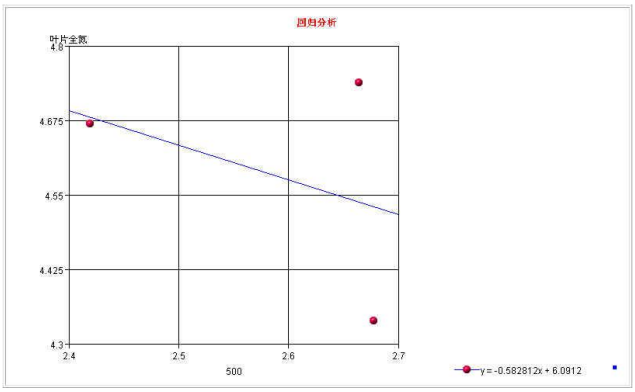


图 4—11 回归分析曲线图(取 500nm 波段反射率值和叶片全氮含量回归分析)

红谷

采样编号	采样日期	红谷位置	红谷反射率值(nm)
J8N20W30	2002年04月18日	554	5.8256837
J8N10W15	2002年04月18日	554	5.1646179
aaa	2003年05月01日	555	5.5650423

红边

采样编号	采样日期	红边位置	红边斜率
J8N20W30	2002年04月18日	730	0.872615
J8N10W15	2002年04月18日	730	0.852714
aaa	2003年05月01日	730	0.827824

红边面积

采样编号	采样日期	红边面积
J8N20W30	2002年04月18日	82.282958
J8N10W15	2002年04月18日	81.941623
aaa	2003年05月01日	84.29605

NDVI

采样编号	采样日期	反射率(680nm)	反射率(800nm)	ndvi
J8N20W30	2002年04月18日	2.9347	44.2372	0.8756
J8N10W15	2002年04月18日	2.5164	42.6606	0.8886
aaa	2003年05月01日	2.9548	41.9802	0.8685

黄边

采样编号	采样日期	黄边位置	黄边斜率
J8N20W30	2002年04月18日	550	0.026186
J8N10W15	2002年04月18日	550	0.023293
aaa	2003年05月01日	550	0.025534

绿峰

采样编号	采样日期	绿峰位置	绿峰反射率值(nm)
J8N20W30	2002年04月18日	554	5.825684
J8N10W15	2002年04月18日	554	5.164618
aaa	2003年05月01日	555	5.565042

图 4－12　农业光谱特征参量的计算

4.3 光谱分析功能应用实践

以上这些光谱分析功能都是在很多理论研究的基础上提出的，在实践中也已经有过比较成熟的应用，其不仅对于目标领域，即农业方面是有着很重要意义的；将来将这些功能移植到其他高光谱应用领域，对于光谱识别等等也可以发挥重要的作用。下面仅举一两例加以说明。

4.3.1 导数光谱对于突出特征光谱信息的作用

导数光谱的突出作用就是凸现其光谱上的特征，消除背景噪声。在植被及农作物光谱中，730nm 附近会有一个反射率值陡升的部分，一般称之为红边，在上面农业光谱特征参量的计算上也能看出。这是一个判读光谱为植被光谱的一个典型判据。而导数光谱就可以在相应的区间里凸现出这一特征。如图，红色圆圈所标注的就是原始数据中的红边部分，以及相应的导数光谱中的特征值。这样，如果我们在其他应用中遇到未知的实物光谱曲线，通过对其 500－1000nm 波段之间做导数光谱，我们就可以判断其到底是不是属于植被。

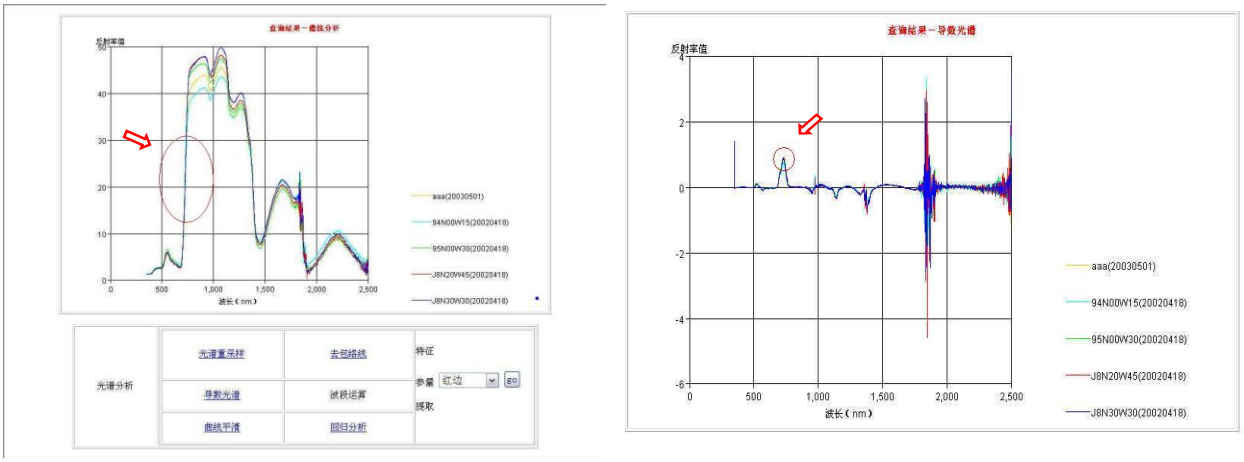


图 4－13 红边在原始光谱曲线和导数光谱曲线中的位置

4.3.2 通过回归分析反演农作物生化参量以及生化参量填图

运用多元回归、线性、指对数回归等分析方法，将农作物的生化参量，典型的如：叶面积指数、叶片全氮含量、含水量等等，于光谱参量数据，如：NDVI、特定波段的数值等进行分析，可以得到他们之间的数学关系。利用这些得到的关系，可以进行农作物生化参量反演、填图、长势评估、估产等等具有很强实际意义的工作。在这方面，刘良云博士的博士后出战报告《高光谱遥感在精准农业中的应用研究》做了比较详细的分析与说明，这里仅就已经实现软件化的简单线性回归做简要分析。

假设有需求是：根据指定条件（叶片全氮含量＝4.34）检索得到的作物样本光谱数据 ndvi 及生化参量 LAI 的数值，进行线性回归分析，并希望由此可以通过测得的 ndvi 反演区域内相应的 LAI。

其处理思路为：先根据查询条件得到所需样本，然后在相关分析的选项中选择回归分析，即可得到相应的关系为： $y=23.889503x-17.3761$ 由于其数据非常规范，其相关系数经计算为：0.9671。这样，新测的光谱值可以先通过计算 ndvi 的选项得到 ndvi 值，然后代入公式即可。其过程如下图所示：

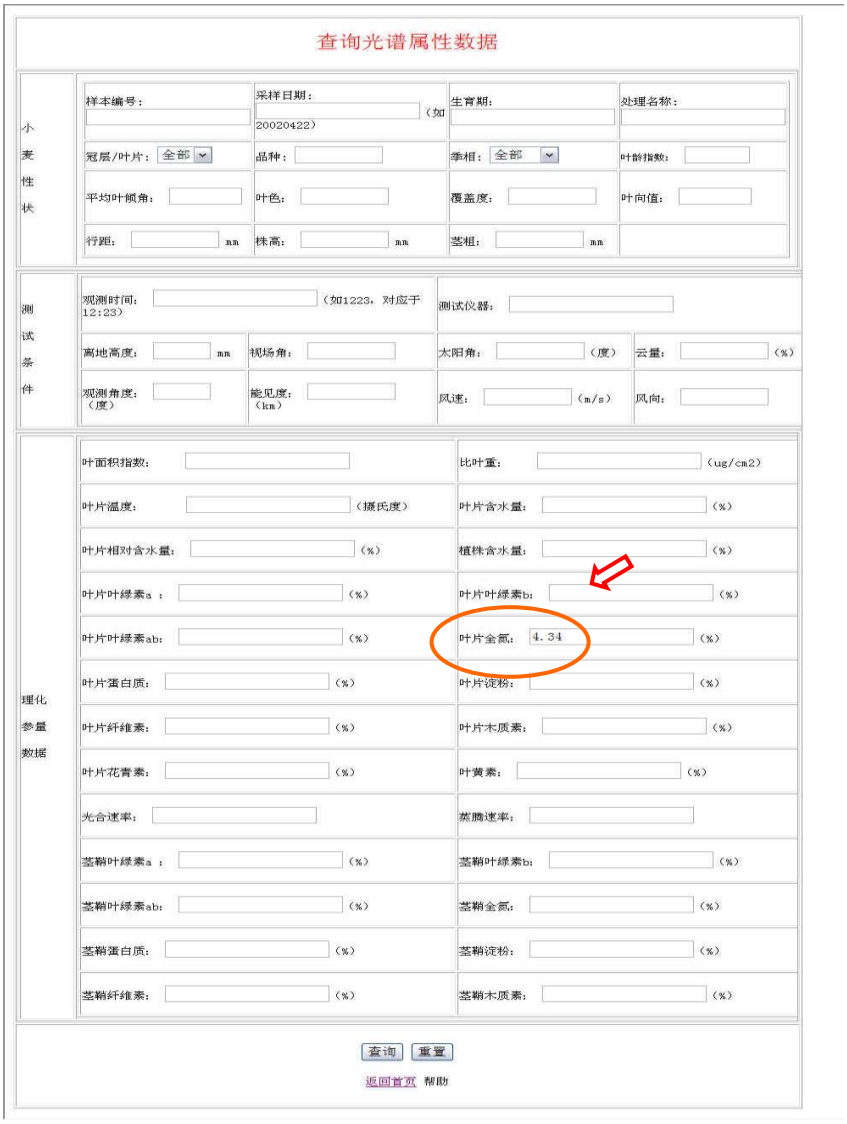


图 4－14　查询条件页面

图 4－15
按叶片全氮＝4.34 查询结果页面

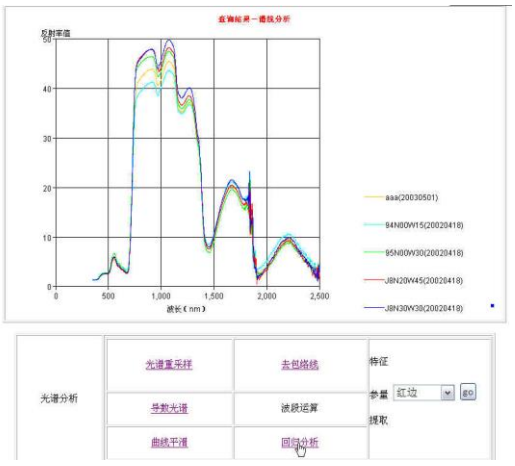


图 4－16
按叶片全氮＝4.34 查询结果曲线图

图 4－17
选取 ndvi 和 LAI 回归分析

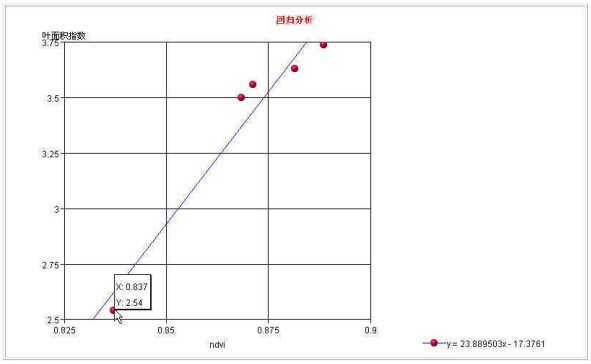


图 4－18 直线方程及直线图

通过这两个小例子可以看出，本系统所附带的光谱数据分析子系统可以满足基本的数据分析的需要，加上网络化的实现，极大的方便了用户的使用。

4.4 本章小结

本章讲述了高光谱数据库系统的数据分析子系统的原理、设计与实现。如前面所述，这一部分是全系统中变化最大的一个部分，目前系统中仅针对农业应用的需求开发了相应的数据分析、处理工具。但是，首先，如光谱重采样、导数光谱、光谱平滑、去包络线等功能是通用型的高光谱处理功能；其次，对于光谱数据库系统来说，添加了数据分析功能等于如虎添翼，为国内光谱数据库系统的发展又添上了新的一笔；再次本子系统的实现，确定了在高光谱数据库系统中数据分析模块的技术结构与设计实现模式，将之挂靠于网络中间层，为将来的扩展、变化奠定了技术基础。最后还以导数光谱和回归分析两个例子具体的说明了数据分析模块对于数据库应用的重要性。

第五章 基于网络的光谱图像子系统

图谱合一是高光谱数据最突出的特点，图像光谱库系统也是高光谱数据库系统的核心所在，尤其是在网络上实现，起点比较高，在国内、国际上目前还属于比较新鲜的尝试，所以这个子系统作为整个系统中最有特色的一部分，为高光谱的发展还是做了有益的探索。同时，该子系统联系了图像、光谱，因此对数据库结构有了进一步的要求，本章里将结合前面提到过的高光谱数据库系统的存储规范，对几种存储方式给以归纳、总结、比较，这项工作将为今后高光谱数据库系统在主流数据库平台 ORACLE 上的存储以及网络化实现都具有极其重要的意义。

5.1 系统功能简介

高光谱数据图谱合一的特点具体的说就是在扫描成像时，对于图像中每一点（pixel）都有其相应的一个全波段的光谱数据，所以如果以几何的坐标为第一维和第二维，光谱的坐标为第三维，这里就可以形成一个图像的立方体。这里尝试把每一点对应的第三维展开在一个二维的图谱曲线上。这一点已经有一些高光谱图像处理软件实现了，比如比较常见的 ENVI，下图为 ENVI 中展开高光谱数据的一个例子。其中左图为取 651nm 为红波段，553nm 为绿波段，455nm 为蓝波段合成的彩色图片，右下为图上某点坐标，右上为该点的光谱曲线图。本系统就是希望能够在网络数据库环境下模拟实现这一功能。

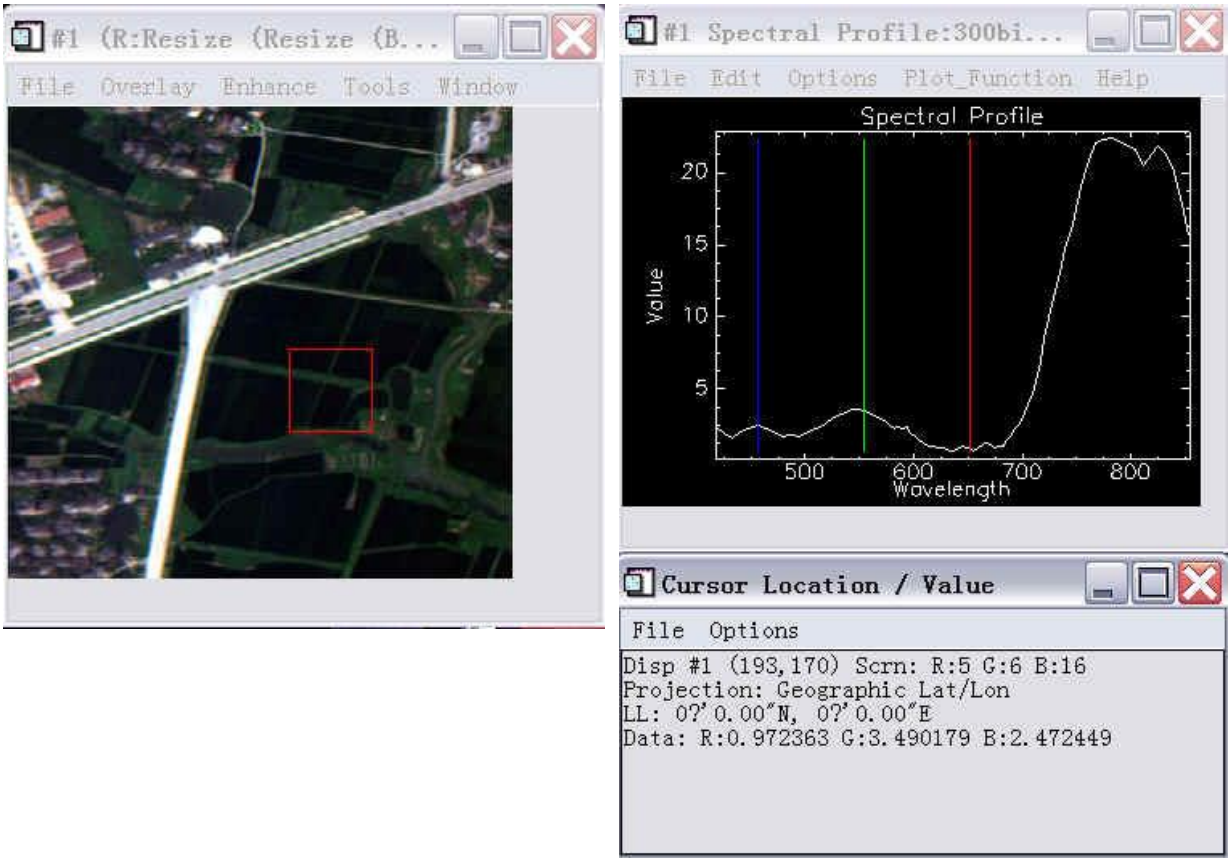


图 5－1 ENVI 中的图谱合一功能

5.2 光谱图像样本数据标准

在本研究项目的图像样本库中，需要每点都有相应光谱的高光谱图像数据，一般这种图像数据都是航拍甚至是卫星图像，该项目中主要为地面测量数据，无法提供，重新选取了航拍图像作为数据源。这里采用了的是常用图像处理软件 ENVI 的图像文件格式：*.img。该格式有三种存储模式：BSQ、BIL、BIP。简言之，BSQ 为从第一波段到最后一个波段的所有象素点的光谱反射率值以二进制的方式顺序排列；BIL 则是按照从第一行象素点到最后一行象素点的所有波段光谱反射率值以二进制的方式顺序排列；BIP 则是按照从第一象素点到最后一个象素点所有波段光谱反射率值以二进制的方式顺序排列。其具体格式标准见下图。

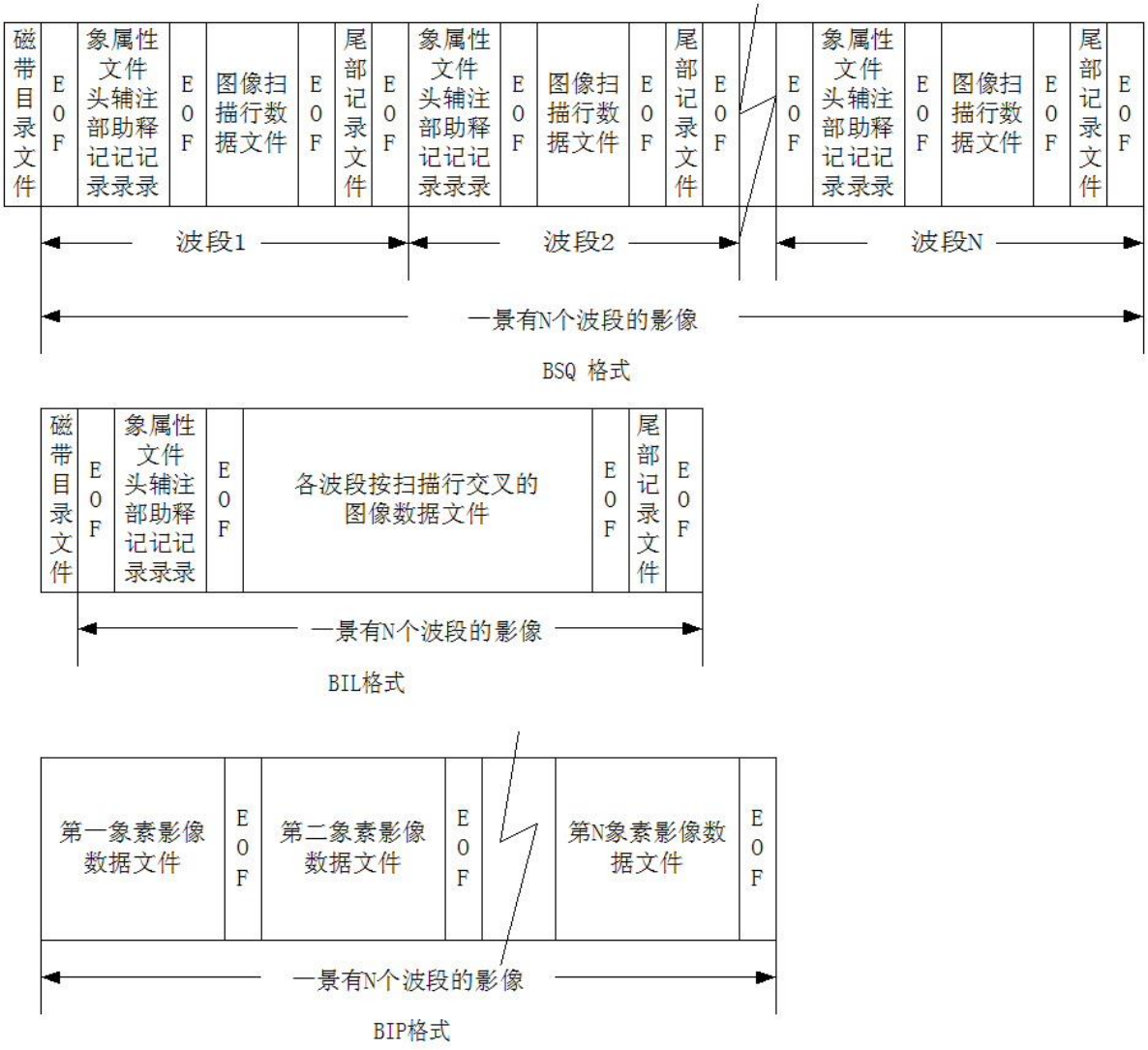


图 5－2　ENVI 中*.img 图像文件格式的三种存储模式图解

可见，选取 BIP 的格式对于实现我们的目的最为有利，这一点后面还会具体提到。同时，为了达到标准化的目的，本光谱图像库中的数据，拟采用 300×300pixel 作为统一标准。此次的数据选用了由我国上海技术物理所研制的高光谱推扫式成像仪，即 PHI，得到的航拍数据，该仪器共有 80 个波段。以 1999 年 9 月在江苏常州的实验飞行图像数据中从中截取了一块 300×300pixel 的图像样本块，并存储其相应的高光谱数据。

5.3 光谱图像库子系统设计与实现

这个子系统虽然和前面的光谱数据库子系统整合在一起而且也是存储的高光谱遥感数据，但是由于其图像和光谱的联合性，造成了其数据结构的不同。本节在讲述此子系统设计方案的同时还将对比几种 ORACLE 平台的各种大型数据的存储方式，众所周知，一个系统的数据结构的好坏将决定该系统的性能，所以这项工作对于今后系统的发展和进步，以及相关的设计都有着举足轻重的意义。

5.3.1 逻辑结构的设计

同样此子系统的数据结构设计也采用了前面提到的数据存储规范，即：普通属性表组＋光谱数据表组。具体而言，imagespectrum 表为光谱数据表，isnatures 表为光谱属性表，其中 imageno 相当于主关键字。可以看到本子系统的光谱数据表与光谱数据库子系统的光谱数据表在结构上有了明显的差异。在此表中应用了两种大对象数据类型 clob 和 blob，即文本型大对象和二进制型大对象，这在前面存储图像数据时也用到过，后面还将具体对比讲述。Wlno 是为了系统能够有比较好的扩展性，这又回到刚开始时提到的本系统的难点所在，因为本系统中的光谱图像可以是任意的高光谱仪器得到的，所以其波段数也是任意的，wlno 就是用于解决这一问题。Wavelength 以文本方式存储了相应仪器的波段值，每行一个波段，然后再以 clob 的形式存储于数据库中。反射率值则是一个整体的形式出现的，而不是像前面光谱数据库子系统那样分成每一波段存储，后面将进行解释和对比。该字段将存储整个的 BIP 模式的*.img 数据文件。具体的表结构见下表：

字段	字段类型	长度及小数位	显示名称	备注
Imageno	Varchar2	16	图像编号	
Wlno	Number	4	波段个数	
Wavelength	Clob		波段	
Wavedata	Blob		反射率值	

表 5－1 光谱数据表 imagespectrum 结构

字段	字段类型	长度及小数位	名称	备注
Datano	Number	8	编号	
Imageno	Varchar2	16	图像编号	
imagedate	Number	8	日期	
Instrument	Varchar2	16	测试仪器	
Height	Nmber	4	离地高度	单位：Km
Area	Varchar2	16	图片地区	
Windspeed	Number	4，2	风速	m/s
Winddire	Varchar2	4	风向	
Cloudine	Number	4，2	云量	%
Visibility	Number	2	能见度	Km
Sunangel	Number	2	太阳角	°
picture	Blob		照片	

表 5－2 属性数据表 Isnatures 结构

5.3.2 光谱图像库子系统的实现

在数据结构中提到了，光谱数据是采用的 BIP 模式的*.img 数据，即在光谱数据中是按照每点的所有波段（从低向高）的数据顺序排列的。在入库时需要经过 java 的程序改写，因为实验发现 java 读取二进制数据时会与 c 语言的格式不同。图片则是按照红波段 R＝651nm，绿波段 G=553nm，蓝波段 B＝455nm 取值，在 ENVI 中合成为*.jpg 格式的彩色图像然后截取了 300×300 象素的区域，以 blob 的字段类型存入数据库中。运行时，先显示出该图片的一般属性并用 java applet 显示该图片。然后接收鼠标点击的位置（x，y），然后按照 k＝300×x＋y 的公式算出该象素在光谱数据文件中的象素个数。在数据库的光谱数据字段中，以 wlno×k 得到该象素的起始读数的位置，向后读取 wlno 个二进制数据放入 y 数组，则该 pixel 的反射率数值就已经得到了。再以读行数据的方式，从 wavelength 字段中读取波段值存入 x 数组，以 x，y 两个数组传入 kavachart 插件返回给用户即可。

其对应于第一节中 ENVI 图像的同样采样点的光谱曲线如下图 5－3。需要注意的一点是，在 ENVI 中，左上角的原点坐标为（1，1），而在本系统中左上角的原点坐标为（0，0），所以可以看到在原图中的（193，170）在本系统中则为（192，169）。

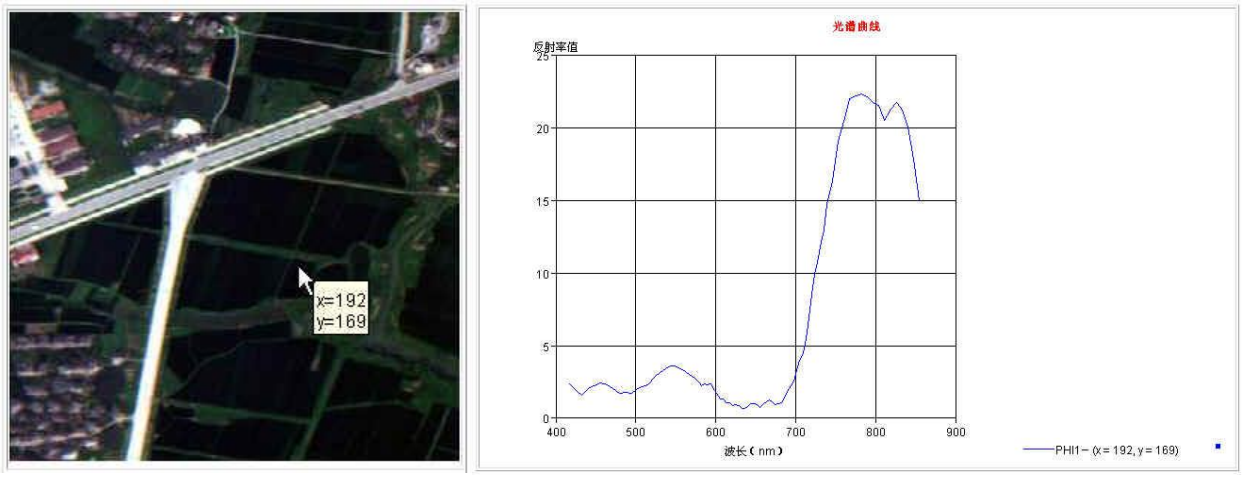


图 5－3 光谱图像子系统对应的采样点光谱曲线

5.3.3 高光谱数据存储结构的比较和总结

上文提到了，在整个系统中，两个子系统：光谱数据库子系统和图像光谱子系统虽然使用了同样的设计规范，但是其核心表－－光谱数据表的数据结构并不相同。这是由他们各自的特点决定的，下文将对这几种存储结构进行对比研究。

（1）波段独立顺列式（简称：顺列式）

这个模式即是光谱数据库子系统采用的模式，具体而言就是在光谱数据表中，以每一个波段及其相应的反射率值作为一个记录，相应的两个主要字段均以 number(m,n)作为字段类型。这种模式的优点是存储时波段相互独立，存储、查询、处理等都非常迅速，也便于分别提取，特别有利于需要波段值单独操作的应用，而且必然需要冗余字段作为定位，所以可以对冗余的定位字段进行各种数据库性能调整操作，如：加索引（index）、建立分区等等，以利于提高效率。当然这种存储方式的缺点也是很明显的，就是记录数相对较多，冗余字段会浪费一定的存储空间。

（2）波段集中整合式（简称：整合式）

这个模式就是光谱图像库子系统采用的模式，具体而言就是在光谱数据表中，以每一个样本为一条记录，无论是波段值还是反射率值均以一个类似文件的方式存储，相应的两个字段分别以 clob、blob 作为字段类型。这种模式的明显好处是结构性更好，更为直观，容易理解，存储上也比较节约空间，但是这样做在任何后续的读取、处理的过程中，均需要以单独的程序对该字段进行操作，比如定位、跳跃等等，这就影响了整个应用的速度。即便以添加、读取、修改等基本数据库操作，大对象类型数据仍然需要专门的包来操作，这就增加了开发的难度。

（3）表单位式

这可以说是最容易理解的一种方式。由于高光谱遥感数据量大。类型繁多，所以一种类型的数据一个表可以说是最简单最容易理解的方式。既可以以某种高光谱仪器为单位来建表，也可以以对象为单位见表等等，总之共同点就是会有一些比较固定的数据维，或者是波段数固定，或者是光谱数固定，这样就可以根据该维来设定该表的结构。这样实施，数据存储量不会冗余，浪费空间；查询等操作的效率也会比较快；开发亦比较容易，但是最致命的缺点就是扩展性差，因为作为数据库开发人员不应该允许用户随着数据量的增大经常进行建表的工作。所以这种方式仅限于一些特殊要求或者数据量比较固定的情况下，本系统中没有加以涉及。

以上是三种比较典型的高光谱数据的存储模式，各有特点，用途也不同，本系统实现了其中的两种。下面我将对比说明前两种格式。

在光谱数据库子系统中，我之所以采用波段独立顺列式，主要就是考虑到首先，该样本所使用的ASD-500具有2000多个波段，这在高光谱遥感数据中亦属于波段数比较多的数据了。一天的样本对应了100000条数据对于ORACLE这种大型数据库而言，还是可以接受的，而且关键在于需求中对于画光谱曲线，以及回归分析等等都是可能需要单独处理各波段数据的，所以综合考虑处理性能和存储空间，采用了第一种方式。

在图像光谱子系统中，也曾经设计过使用第一种存储方式来存储与图像对应的高光谱遥感数据。其设计思路如下：假设为光谱数据表为Imagespe，其结构如下表5－3。以象素编号作为冗余字段，即第imageno图片的第pixelno点的各个波段及其值。这里只有80波段，目前只有一副图片，300×300pixel，80 波段/pixel，计算下来总共 1×90000×80＝720 万条记录。这样如果本子系统录入100幅图片，记录数就会暴增到7，200，000×100＝7.2×108条记录，即7亿2千万条记录。如果该库中存入我国的OMIS高光谱成像仪，128个波段，乃至更多的比如：USGS的HYDICE,有210个波段，那么记录数就会上10亿条量级了，这即便对于ORACLE这样大型的数据库依然是一个很可怕的数字。所以在图像光谱子系统中，牺牲了一部分处理效率，以及增大的开发难度，采用了第二种波段集中整合式来存储光谱数据。

字段	字段类型	长度及小数位	显示名称	备注
Datano	Number	10	编号	
Imageno	Varchar2	16	图像编号	
Wavelength	Number	6，2	波段	
Wavedata	Number	11，7	值	
Pixelno	Number	8	象素编号	

表5－3　假设采用波段独立顺列式的光谱数据表Imagespe结构

5.3.4 图像数据存储结构的设计与比较

在光谱图像子系统中，光谱数据是和图像数据紧密联系在一起的，所以图像数据如何以合理的数据结构存储也是一个重要的课题。一般来说网络化的 ORACLE 数据库系统中图像数据有三种存储方式：在数据库中以 blob 字段类型存储；在数据库中以 bfile 字段类型存储；独立于数据库以图像文件的方式存储。

（1）在数据库中以 blob 字段类型存储

blob 字段是 ORACLE 数据库提供的一种存储大对象数据的数据类型，他以二进制的形式存储数据，所以常被用于存储图片等数据。使用 blob 字段存储数据主要特点是使得图片数据与整个数据库的数据成为一个整体，这样在库内容的迁移、升级、恢复等等方面与其他数据都将统一为一个整体，十分有利于数据库的管理和操作。其缺点是读取时速度略慢，不如在库外读取时效率更高。

（2）在数据库中以 bfile 字段类型存储

bfile 字段类型实际上是 ORACLE 数据库中一种文件指针，该字段类型仅仅将图像数据在系统中的存放位置存储在数据库中，所以该图像文件整体还是游离于数据库之外的。这就失去了刚才提到的 blob 字段的优势。

（3）图像文件形式存储：

这也是当前网站比较流行的一种存储方式，图像文件仅仅以一个文件夹的形式放置于操作系统中，并进行各种操作，这样最主要的优势就是操作效率高，速度快。但是，作为本系统是一个科研型的系统，其数据有着密不可分的关系，是一个有机的整体，因此这种方式并不适宜。

综上所述，本系统在各个子系统的应用都采用了 ORACLE 数据库中的 blob 字段类型作为存储图像数据的基本数据类型。

5.4 光谱图像库应用实践

在本光谱图像子系统中，在开发时仅用了一个样本作为例子。高光谱图谱合一的特点有着其特殊的优点，下面将举一个例子说明。

下面两副图均是 PHI 获取得，按照通常的图片所示，取红、绿、兰，即 651，553，455nm 三个波段合成的，一幅是 1999 年于中国常州获得，一幅是日本南牧获得。可以看出，在红框区域内，常州地区是水稻作物，南牧地区暂时不知道，两者看上去区别不大，则图像光谱的第一个优势显现出来，可以即时的从图像中提取光谱进行对比、分析。合成的真彩色图片以及相应地区采样点的光谱曲线请见图 5－4、图 5－5。

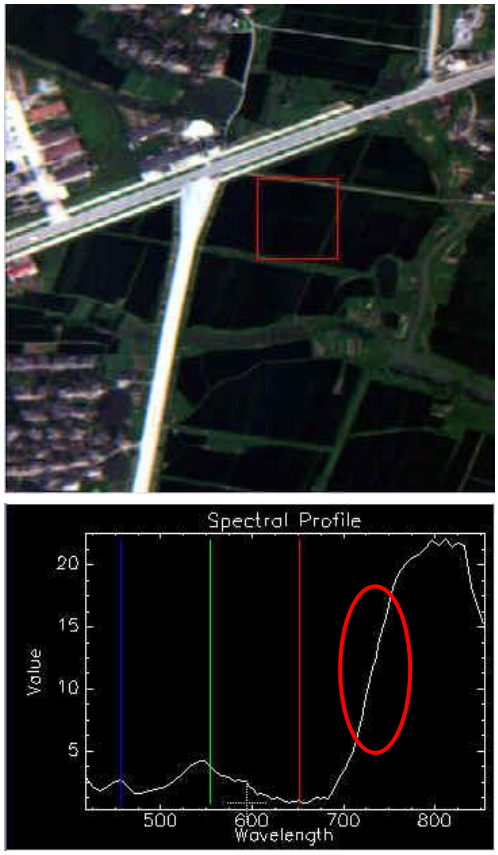


图 5－4 中国常州图像光谱

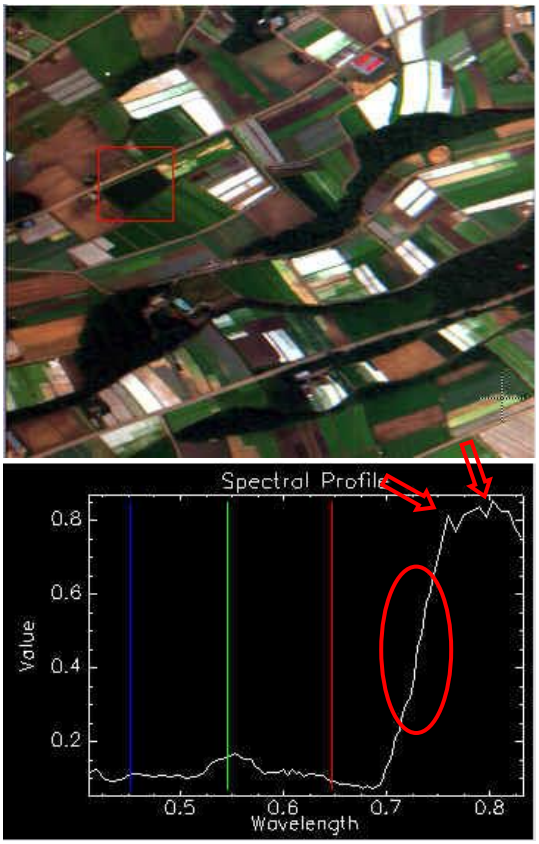


图 5－5 日本南牧图像光谱

可以看出两者的光谱曲线形状十分类似，而且如红色椭圆中标注的，两者在 730nm 附近均有陡升，所以可见后者和前者应该同属于植被类型。但后者在 700nm 以前的波形较为平缓，尤其是在 800nm 附近有两个小波谷，如红色箭头所示，与前者明显不同，所以应该不是水稻。将两者做去除包络线分析可以得到图 5－6、图 5－7：

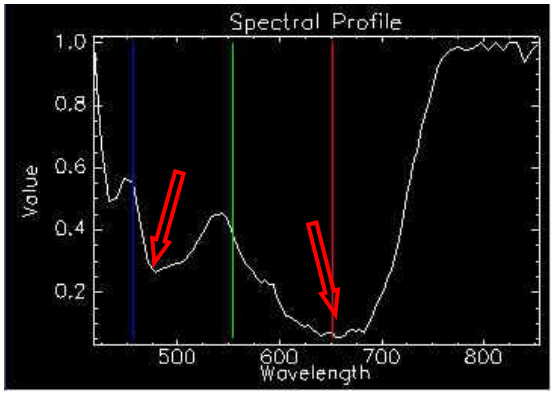


图 5－6 图 5－4 光谱去包络线结果

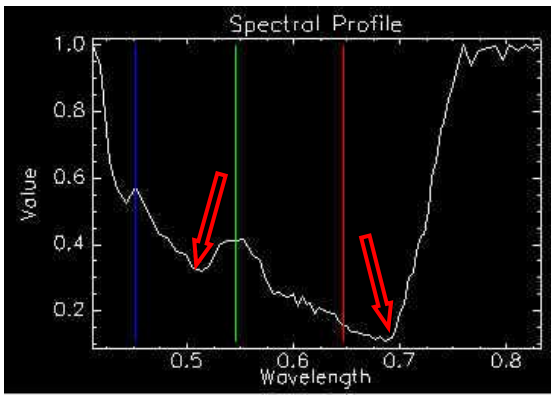


图 5－7 图 5－5 光谱去包络线结果

去除包络线以后可以看出，在红色箭头所示的地方。两个吸收谷的位置并不相同。500nm 附近的吸收谷，常州地区的在 500nm 之前，而南牧地区的在 500nm 之后；700nm 附近的吸收

谷常州地区的在 650nm 附近，而南牧地区则很接近 700nm。可见两者还是不同的，后者不应该是水稻。经地面调查得知，该区域是种植的是架豆作物。

由此可见，在高光谱数据库系统中如果能够取得比较标准的地物波谱图像块，并将其以这种网络的形式发布，那么对于将来应用于实地野外勘测中地物的判别、成分的分析等等都将发挥重要的作用。

5.5 本章小结

本章介绍了高光谱数据库系统中的核心－－基于网络的光谱图像子系统。该系统与普通的光谱数据库系统最大的不同在于，它存储的是图像光谱数据，可以从图像中直接获得光谱，更加直观，在数据结构上更加紧凑，原来近 0.1M 个对象的光谱曲线，现在一幅 300×300 象素的图像光谱数据就可以存储。为此，本章再次强调了第三章中提出的高光谱数据库系统的数据结构存储规范：**光谱数据表组＋属性数据表组；** 同时，提出了在 ORACLE 数据平台下的三种高光谱数据存储模式：**波段独立顺列式、波段集中整合式以及表单位式，**并将他们进行了对比分析，扼要说明了各自的优缺点以及选择条件，为今后高光谱数据库系统的发展奠定了基本的技术基础。本章还具体讲述了基于网络的光谱图像子系统的实现过程和结果，最后用一个小例子介绍了光谱图像子系统的应用实践。

第六章 数据仓库应用于高光谱数据的初步探讨

随着数据库技术的迅速发展和管理系统的广泛应用，人们积累的数据越来越多。数据的背后隐藏着许多重要信息，人们希望能够对其进行更高层次的分析，而不只是简单的查询,以便更充分地利用这些数据。目前的数据库系统可以高效地实现海量数据的录入、修改、统计、查询等功能，但无法发现数据中存在的关系和规则，无法理解数据中包含的信息,无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。人们迫切需要把这些看似分散的数据，提炼成一条条有价值的信息。数据挖掘技术能自动分析数据，对它们进行归纳性推理和联想，寻找数据间内在的某些关联，从中找出一些专家们不易察觉的极有用的信息，发掘出潜在的、对信息预测和决策行为起着十分重要作用的模式，从而达到作出正确决策的目的。

6.1 几个与数据挖掘技术相关联的概念

6.1.1 数据挖掘技术的产生和发展

数据挖掘(Dta Mining)，又称数据库中的知识发现(Knowledge Discovery in Database，KDD)，是指从大型数据库或数据仓库中提取隐含的、未知的、非平凡的及有潜在应用价值的信息或模式，它是数据库研究中的一个很有应用价值的新领域，融合了数据库、人工智能、机器学习、统计学等多个领域的理论和技术。

数据挖掘技术是人们长期对数据库技术进行研究和开发的结果。起初各种商业数据是存储在计算机的数据库中的，然后发展到可对数据库进行查询和访问，进而发展到对数据库的即时遍历。数据挖掘使数据库技术进入了一个更高级的阶段，它不仅能对过去的数据进行查询和遍历，并且能够找出过去数据之间的潜在联系，从而促进信息的传递。表 6－1 给出了数据进化的四个阶段。

进化阶段	时间段	技术支持	生产厂家	产品特点
数据搜集	60 年代	计算机、磁带等	IBM，CDC	提供静态历史数据
数据访问	80 年代	关系数据库、结构化查询语言 SQL	ORACLE 、 Sybase 、 Informix、IBM、Microsoft	在纪录中动态历史数据信息
数据仓库	90 年代	联机分析处理、多维数据库	Pilot、Comshare、Arbor、Cognos、Microstrategy	在各层次提供回溯的动态的历史数据
数据挖掘	正在流行	高级算法、多处理系统、海量算法	Pilot、Lockheed、IBM、SGI、其他初创公司	可提供预测性信息

表 6－1 数据挖掘进化时间表

6.1.2 数据仓库

业界公认的数据仓库概念创始人 W.H.Inmon 在《建立数据仓库》一书中对数据仓库的定义是：数据仓库就是面向主题的、集成的、不可更新的(稳定性)、随时间不断变化（不同时间）的数据集合，用以支持经营管理中的决策制定过程，是存储数据的一种组织形式。
数据仓库中的数据面向主题，与传统数据库面向应用相对应。主题是一个在较高层次上将数据归类的标准，每一个主题对应一个宏观的分析领域：数据仓库的集成特性是指在数据进入数据仓库之前，必须经过数据加工和集成，这是建立数据仓库的关键步骤，首先要统一原始数据中的矛盾之处，还要将原始数据结构做一个从面向应用向面向主题的转变；数据仓库的稳定性是指数据仓库反映的是历史数据的内，而不是日常事务处理产生的数据，数据经加工和集成进入数据仓库后是极少或根本不修改的；数据仓库是不同时间的数据集合，它要求数据仓库中的数据保存时限能满足进行决策分析的需要，而且数据仓库中的数据都要标明该数据的历史时期。

主题是指用户使用数据仓库进行决策时所关心的重点方面；面向主题是指数据仓库内的信息是按主题进行组织的，为按主题进行决策的过程提供信息；集成是指数据仓库中的信息不是从各个业务处理系统中简单抽取出来的，是经过系统加工、汇总和整理，保证数据仓库内的信息是关于整个企业的一致的全局信息；稳定是指一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓库中一般有大量的插入和查询操作，但修改和删除操作很少；包含历史数据是指数据仓库内的信息并不只是关于企业当时或某一时点的信息，而是系统记录了企业从过去某一时点到目前的各个阶段的信息，通过这些信息可以对企业的发展历程和未来趋势作出定量分析和预测。把信息加以整理归纳，并及时提供给相应的管理决策人员，是数据仓库的根本任务。

数据仓库主要有三方面的作用：首先，数据仓库提供了标准的报表和图表功能，其中的数据来源于不同的多个事务处理系统，因此，数据仓库的报表和图表是关于整个企业集成信息的报表和图表；其次，数据仓库支持多维分析，多维分析是通过把一个实体的多项重要的属性定义为多个维度，使得用户能方便地汇总数据集，简化了数据的分析处理逻辑，并能对不同维度值的数据进行比较，而维度则表示了对信息的不同理解角度。应用多维分析可以在一个查询中对不同阶段的数据进行纵向或横向比较，这在决策过程中非常有用；第三，数据仓库是数据挖掘技术的关键基础，数据挖掘技术要在已有数据中识别数据的模式，以帮助用户理解现有的信息，并在已有信息的基础上，对未来的状况作出预测。在数据仓库的基础上进行数据挖掘，就可以针对整个企业的状况和未来发展作出较完整、合理、准确的分析和预测。

数据仓库是我们进行数据挖掘的良好基础。当我们进行数据挖掘是时，挖掘对象如果是数据仓库，那么这些工作往往在生成数据仓库时已经准备妥当。如果挖掘的数据对象存储在

数据库系统中，则往往不适合直接在这些数据上面进行知识挖掘，需要做数据准备工作，一般包括数据的选择（选择相关的数据）、净化（消除噪音、冗余数据）、推测（推算缺失数据）、转换（离散值数据与连续值数据之间的相互转换，数据值的分组分类，数据项之间的计算组合等）、数据缩减（减少数据量）。

数据仓库的市场巨大，数据仓库产品很多，其中比较有代表性的产品有：Business Objects 和 Sybase、Platinum Technology 等的解决方案。

6.1.3 OLAP

OLAP （On-Line Analytical Processing）代表联机分析处理，他是一种用于对大容量数据归总与分析的技术，是一种用户接口，而不是数据存储的概念。他支持多维性、课钻取性、可旋转性以及多视图等功能。

OLAP 和 DM 虽然都是数据库(数据仓库)的分析工具，但其应用范围和侧重点是不同的。OLAP 的在线性体现在与用户的交互和快速响应，多维性则体现在它建立在多维视图的基础上。用户积极参与分析过程，动态地提出分析要求、选择分析算法，对数据进行由浅及深的分析。而 DM 与 OLAP 不同，主要体现在它分析数据的深入和分析过程的自动化，自动化是说，其分析过程不需要用户的参与，这是它的优点，也正是它的不足，因为在实际中，用户也希望参与到挖掘中来，如只想对数据的某一子集进行挖掘，以及对不同抽取、集成水平的数据进行挖掘，还有想根据自己的需要动态选择挖掘算法等等。

由此可见，OLAP 与 DM 各有所长，如果能将二者结合起来，发展一种建立在 OLAP 和数据仓库基础上的新的挖掘技术，将更能适应实际的需要。这种结合的产物就是 OLAM(Online Analytical Mining 或 OLAP Mining)。

6.2 数据挖掘的分析方法与常用技术

数据挖掘的核心模块技术历经了数十年的发展，其中包括数理统计、人工智能、机器学习。今天，这些成熟的技术，加上高性能的关系数据库引擎以及广泛的数据集成，让数据挖掘技术在当前的数据仓库环境中进入了实用的阶段。

数据挖掘利用的技术越多，得出的结果精确性就越高。原因很简单，对于某一种技术不适用的问题，其它方法即可能奏效，这主要取决于问题的类型以及数据的类型和规模。数据挖掘方法有多种，其中比较典型的有关联分析、序列模式分析、分类分析、聚类分析等。

（1）关联分析

关联分析，即利用关联规则进行数据挖掘。在数据挖掘研究领域，对于关联分析的研究开展得比较深入，人们提出了多种关联规则的挖掘算法，如 APRIORI、STEM、AIS、DHP 等算法。关联分析的目的是挖掘隐藏在数据间的相互关系，它能发现数据库中形如"90%的顾客在一次购买活动中购买商品 A 的同时购买商品 B"之类的知识。

（2）序列模式分析

序列模式分析和关联分析相似，其目的也是为了挖掘数据之间的联系，但序列模式分析的侧重点在于分析数据间的前后序列关系。它能发现数据库中形如"在某一段时间内，顾客购买商品 A，接着购买商品 B，而后购买商品 C，即序列 A→B→C 出现的频度较高"之类的知识，序列模式分析描述的问题是：在给定交易序列数据库中，每个序列是按照交易时间排列的一组交易集，挖掘序列函数作用在这个交易序列数据库上，返回该数据库中出现的高频序列。在进行序列模式分析时，同样也需要由用户输入最小置信度 C 和最小支持度 S。

（3）分类分析

设有一个数据库和一组具有不同特征的类别(标记)，该数据库中的每一个记录都赋予一个类别的标记，这样的数据库称为示例数据库或训练集。分类分析就是通过分析示例数据库中的数据，为每个类别做出准确的描述或建立分析模型或挖掘出分类规则，然后用这个分类规则对其它数据库中的记录进行分类。举一个简单的例子，信用卡公司的数据库中保存着各持卡人的记录，公司根据信誉程度，已将持卡人记录分成三类：良好、一般、较差，并且类别标记已赋给了各个记录。分类分析就是分析该数据库的记录数据，对每个信誉等级做出准确描述或挖掘分类规则，如"信誉良好的客户是指那些年收入在 5 万元以上，年龄在 40～50 岁之间的人士"，然后根据分类规则对其它相同属性的数据库记录进行分类。目前已有多种分类分析模型得到应用，其中几种典型模型是线性回归模型、决策树模型、基本规则模型和神经网络模型。

（4）聚类分析

与分类分析不同，聚类分析输入的是一组未分类记录，并且这些记录应分成几类事先也不知道。聚类分析就是通过分析数据库中的记录数据，根据一定的分类规则，合理地划分记录集合，确定每个记录所在类别。它所采用的分类规则是由聚类分析工具决定的。聚类分析的方法很多，其中包括系统聚类法、分解法、加入法、动态聚类法、模糊聚类法、运筹方法等。采用不同的聚类方法，对于相同的记录集合可能有不同的划分结果。

聚类分析和分类分析是一个互逆的过程。例如在最初的分析中，分析人员根据以往的经验将要分析的数据进行标定，划分类别，然后用分类分析方法分析该数据集合，挖掘出每个类别的分类规则；接着用这些分类规则重新对这个集合(抛弃原来的划分结果)进行划分，以获得更好的分类结果。这样分析人员可以循环使用这两种分析方法直至得到满意的结果。

数据挖掘中最常用的技术有：

人工神经网络：仿照生理神经网络结构的非线形预测模型，通过学习进行模式识别。

决策树：代表着决策集的树形结构。

遗传算法：基于进化理论，并采用遗传结合、遗传变异、以及自然选择等设计方法的优化技术。

近邻算法：将数据集合中每一个记录进行分类的方法。

规则推导：从统计意义上对数据中的"如果-那么"规则进行寻找和推导。

神经网络：聚集，偏差分析…

遗传算法

模糊逻辑

约略集（rough set）

概念学习（concept learning）

简单的基于规则的推理

数学发现算法 AM

规则发现算法 Meta-Dendral

经验发现算法 BACON

类比发现算法 Phineas

采用上述技术的某些专门的分析工具已经发展了大约十年的历史，不过这些工具所面对的数据量通常较小。而现在这些技术已经被直接集成到许多大型的工业标准的数据仓库和联机分析系统中去了。

6.3 数据挖掘技术的体系结构、过程及设计

6.3.1 数据挖掘技术的体系结构

数据挖掘工具可以是独立于数据仓库（和数据库）以外的，它们需要独立地输入输出数据，以及进行相对独立的数据分析。为了最大限度地发挥数据挖掘工具的潜力，它们可以紧密地和数据仓库集成起来。这样，在人们对参数和分析深度进行变化的时候，高集成度就能大大地简化数据挖掘过程。集成后的数据挖掘体系有自己的特点。应用数据挖掘技术，如上所述，较为理想的起点就是从一个数据仓库开始。例如：这个数据仓库里面可以保存着所有客户的合同信息，并且还应有相应的市场竞争对手的相关数据。这样的数据库可以是各种市场上的数据库：ORACLE 、Sybase、Redbrick、或者其他等等，并且可以针对其中的数据进行速度上和灵活性上的优化。

数据挖掘系统的出现代表着常规决策支持系统的基础结构的转变。不象查询和报表语言仅仅是将数据查询结果反馈给最终用户那样，数据挖掘高级分析服务器把用户的模型直接应用于其数据仓库之上，并且反馈给用户一个相关信息的分析结果。

基于数据仓库的数据挖掘技术过程：

基于数据仓库的数据挖掘过程都要有数据准备、执行挖掘算法和表达结果等几个阶段.数据挖掘过程细分为以下几个步骤:

（1）理解和定义问题

（2）数据的搜集和抽取

（3）数据净化

（4）数据引擎

（5）算法引擎

（6）运行数据挖掘算法

（7）评估结果

（8）重新精化数据和问题

（9）使用结果

上述的九个步骤在数据挖掘过程中要反复多次。其中，每一个步骤都是必不可少的，下面分别讨论各个步骤：

（1）理解和定义问题：理解和定义问题是解决任何事情的必经步骤，在数据挖掘过程中，它要花费很多的时间。数据挖掘不是简单的把数据挖掘算法应用到数据库上而得到一些结果。如果没有很好的理解问题，得到的结果将没有任何用处。一个问题有多种解决办法，但有些是行得通，有些是行不通的。即使是行得通的办法，也要考虑其执行效率等方面的问题。

（2）数据的搜集和抽取：一旦问题定义完毕，就要进行相关数据的搜集。大多数情况下，相关数据是从已存在的数据库或数据仓库中提取的。通常，数据挖掘算法不能直接在任何一个随意的数据库中工作。我们需要从相关的数据库中提取数据，并将它们存储为数据挖掘算法可以识别的格式。在数据挖掘算法中，一般采用标准数据库查询语言 SQL，或自行设计 DMQL。因为挖掘算法的大部分时间都花费在对数据库的访问上，所以通过数据库管理系统的查询引擎，可以大大提高数据挖掘过程的速度。目前，数据挖掘算法通常是基于一个抽取出来的二维关系表。对于用户所提出的发现任务，确定感兴趣的属性域，进行各种数据汇集的操作。利用抽样技术对数据库中符合条件的元组进行抽样。

（3）数据净化和数据理解：搜集完相关数据后，接下来就要处理数据库。这有两方面的原因：首先，数据分析者要理解数据库的内涵，而不是仅停留在知道数据库中有哪些字段。其次，在数据搜集的过程中(通常是由几个库抽取出信息组成一个新的数据库)，不可避免的存在着一些错误。如：字段值输入错误；字段名称发生错误；字段内容不详；对于同一字段的同一内容的不同表达方式，也可能会造成算法对数据含义理解的不确切性。净化带噪音的数据是一个复杂、牵扯到多方面的

过程。数据的净化过程包括：检查拼写错误、去掉重复的记录、补上不完全的记录、解决不一致的记录、用测试查询来验证数据、根据验证结果反复迭代上述步骤。数据净化的目标是保证所表达数据的一致性(Consistently)，确保数据的参照完整性和数据的精确性。

（4）数据引擎：前面所涉及的步骤都是在谈论如何产生一个挖掘的基础，即一个从原始的数据库到一个挖掘数据库的过程。这个挖掘数据库由所有要在数据挖掘过程中使用到的信息组成。但是，还存在着两个问题：

A. 在原始数据库中包含了许多可以忽略掉的属性。如何选择原始数据库中包含的所有属性的子集。

B. 另外，挖掘数据库中包含的数据信息量有可能远远超过我们所要求的在有限时间内所能处理的信息量，因此，我们必须从中找出样本数据库。到此为止，上述步骤均为整个过程的数据准备阶段，工作量很大，而且也是较难深入的部分。

（5）算法规划：在选择了挖掘数据库后，有很多的数据挖掘算法，但我们需要知道选择哪种算法和怎样应用它。算法的选择直接影响着所挖掘模式的质量。另外，即使选定了某一种算法，这个算法中参数的改变也会影响所产生的模式。在许多时候，有效的数据挖掘算法也可能不能直接用来解决问题，还需做一些辅助的工作来修改算法。这可能因为数据挖掘系统中的工具集不全，或者还没有一个解决某种特定问题的合适算法。

（6）运行数据挖掘算法：如何运行数据挖掘算法是数据挖掘分析者和相关领域专家最关心的阶段。因为只有这个阶段才能给出人们所关心的东西。这个阶段称之为真正意义上的数据挖掘。所有的数据挖掘算法都要事先提出一些标准来度量产生的模式，并在搜寻所有模式的过程中，使用这些标准来决定保留什么，丢弃什么，哪些模式需要继续挖掘。目前，通常利用一些简单的统计属性作为评估标准，如支持度(Support)、置信度(Confidence)和感兴趣度(Interesting)等。对预测型模式好坏的判断比较容易。由于可预测型模式是预测某一属性的值，而这个属性的值又存在于训练集合中，所以一般来说，通过把预测的值与存在于训练集中的那个属性的实际输出值相比较，计算模式的误差程度，从而做出对模式的评估。相比较，对信息型模式的评估较难。

（7）结果的初步评估：他是用来评估可预测型模式好坏的方法，依赖于所要解决的问题，所以仅仅给出某种模式的精确度是没有用的。最重要的是，使用模式模拟实际的行为并给出使用它的结果报告。但是由于数据挖掘所找到的模式可能只是某一段时间内的较短暂的规律，所以即使我们选用了各种评判方法，如数学的或其他的非客观性的方法，它也只是一种估测。真正的检测只能在实际的应用中进行。

（8）重新精化数据和问题。，如果再实际的检验中，挖掘结果或者模式难以令人满意，就要重新进行新一轮的数据挖掘过程。通常，数据挖掘的过程是由粗略到细致，由简单到复杂的过程。经过几次反复精化之后，如果模式的执行情况足够好，就可以进入到使用结果的阶段了。

（9）使用结果：在前面讨论了数据挖掘的许多准备工作及论证所挖掘出的模式的有效性。一旦当到达数据挖掘的最后一步，我们就可以应用基于所发现模式的决策了。

其中，前四个步骤为整个数据挖掘过程的数据准备阶段。工作量直大，占了全部的约60%，而且在目前，也较难有深入研究。五六两个步骤是数据挖掘的核心，也是挖掘者和专家最为关心的阶段。目前，在数据挖掘算法方面，已经取得了并且正在不断取得进展。尤其是在Apriori 算法（属于关联方法）及其派生算法基础上，进行了大量的试验与研究。如新的数据挖掘结构 FP－tree 和相应的构造算法（15），以及包含负属性的关联规则采掘算法（16）等等。

一个典型的基于数据仓库的数据挖掘工具为ARMiner（16）（Association Rules Miner），，主要包括数据预处理、数据采掘、数据评价三部分。其中：

（1）数据预处理模块：将来自关系数据库、多维数据库、数据仓库或者文件系统的数据进行转化，对于大数据集，可以通过数据采样减少处理的数据量，然后利用数据清理等手段清除脏数据，将数据整合成能被采掘算法利用的数据，最后存入数据采掘库。同时可以利用概念层次树(运用领域知识)对原始数据进行必要的抽象，使得采掘模块能够处理数据各个抽象层次，而不是仅对细节数据进行采掘。

（2）数据采掘模块：在友好的导航界面（Wizard）引导下，使用合适的算法(包含负属性的关联规则采掘算法)对数据采掘库中的数据进行采掘，它可以使用索引、并行或删减分支等技术提高运行效率，并把采掘结果输出给数据评价模块。

（3）数据评价模块：将数据采掘模块存储在采掘库中的结果以可视化的形式表示出来。对于关联规则采用网格和图形两种表示方式。对于采掘结果，用户可以进行评价结果的信任程度，并将其存储在数据采掘库中，供以后采掘使用。

在具体的采掘过程中，用户可以循环调用以上模块，直至获得满意的决策信息为止。

6.3.2 数据挖掘系统设计

数据挖掘系统，由上面讨论的工具模型，预计主要包括数据预处理、数据采掘、数据评价三部分，，可用于 B/W/S（浏览器/Web 服务器/数据库服务器）三层结构。将核心算法存放在服务器端，客户端通过 API 函数调用各种功能。

（1）数据预处理模块：将来自关系数据库或者文件系统的数据进行转化，对于大数据集，可以通过数据采样减少处理的数据量，然后利用数据清理等手段清除脏数据，

将数据整合成能被采掘算法利用的数据，最后存入数据采掘库。

（2）数据采掘模块：使用合适的算法对数据采掘库中的数据进行采掘，它可以使用索引、并行或删减分支等技术提高运行效率，并把采掘结果输出给数据评价模块。

（3）数据评价与评价模块：将数据采掘模块存储在采掘库中的结果表示出来，最好采用可视化的表示方法。对于采掘结果，用户可以进行评价结果的信任程度，并将其存储在数据采掘库中，供以后采掘使用。在具体的采掘过程中，用户可以循环调用以上模块，直至获得满意的决策信息为止。以上三个主要的模块中，数据采掘模块是数据挖掘技术的核心技术，包括数理统计、人工智能、机器学习。这些技术，加上高性能的关系数据库引擎以及广泛的数据集成，让数据挖掘技术在当前的数据库环境中进入了实用的阶段。

数据挖掘利用的技术越多，得出的结果精确性就越高。因为对于某一种技术不适用的问题，其它方法即可能奏效，这主要取决于问题的类型以及数据的类型和规模。数据挖掘方法有多种，其中比较典型的有关联分析、序列模式分析、分类分析、聚类分析等。

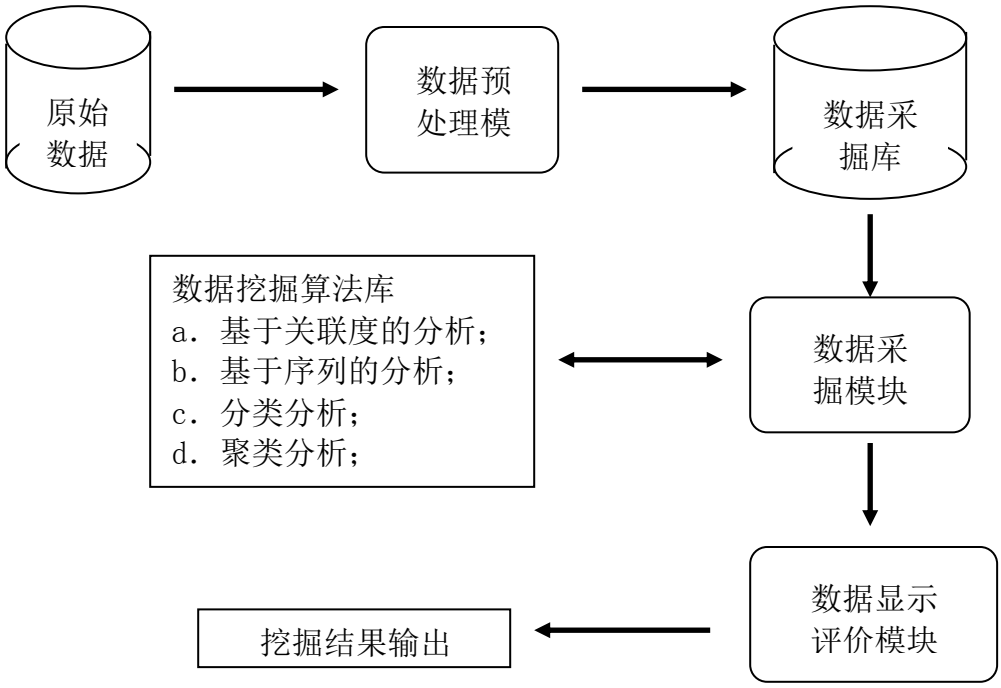


图 6－1 挖掘系统设计简图

6.4 结合遥感高光谱数据应用展望

当前，遥感数据，尤其是高光谱数据的特点就是数据量大，信息量丰富，但是目前很多数据分析处理工作，还是要靠手工来进行，而且这种分析工作难免挂一漏万。更重要的是，很多来自航拍，乃至卫星拍摄的数据还远没有得到充分的利用。所以在这个领域引入数据挖掘和数据仓库技术将对遥感数据处理技术，无论从数据的利用率还是从分析过程的自动化，都会有一个很大的提高。

6.4.1 小汤山精准农业项目数据分析

2001 年，中科院遥感所高光谱室参加了北京农业信息中心主持的小汤山精准农业示范项目，其中完成了 350~2500nm 反射率光谱与生化参量的相关分析；利用地面光谱信息反演小麦生化参量回归分析，如红边、红谷及其它吸收特征与生化参量分析，完成了光谱特征反演生化参量的模型分析与论证；完成了 4 月 11 日和 26 日两次 OMIS 高光谱数据反演生化参量研究工作，进行了多种生化参量的填图。下面附图是其中一些分析结果：

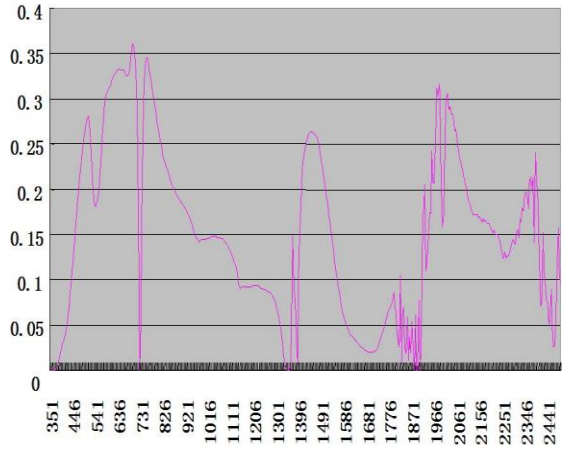


图 6－2 全氮含量与光谱反射率的相关性

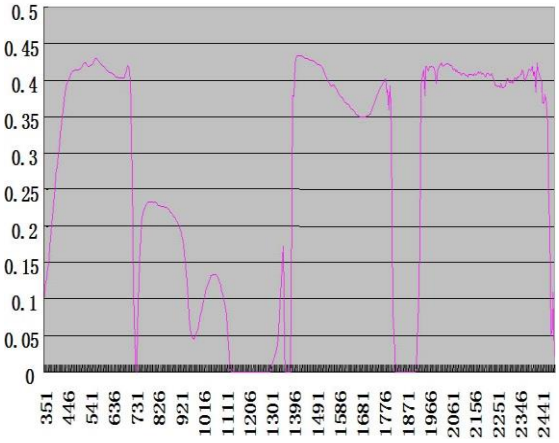


图 6－3
可溶性糖含量与光谱反射率的相关性 R2 曲线

还有叶绿素浓度，叶面积指数等等生化参量与光谱反射率之间相关性分析。通过线性回归算法，可以得到相关系数曲线：

全氮含量与光谱反射率的相关性 R2 曲线：

Total Nitrogen = -0.2216σ + 11.863　　σ：红边宽度　单位：nm

可溶性糖含量与光谱反射率的相关性 R2 曲线：

Sugar = -1.5436 λp + 1114.5　　λp：红边位置　单位：nm

可见，目前的分析工作完全是可以由数据挖掘技术来替代的。我们可以遵循上述的数据挖掘过程，首先将光谱信息和生化参量等放入库中，然后根据我们的需要设定主题，比如：全氮含量与光谱反射率的相关关系，再通过我们得到的方程，设定我们希望的规则，并产生方法，比如：线性回归，然后由数据挖掘的过程进行数据筛选，选取合适的数据，进行挖掘。我们可以根据得到的相关系数，选取几个合适的波段，采用多元回归的方法，反演生化参量。我们还可以多选取几个生物参量，利用分析关联度，以及聚类等方法，同时发掘他们之间以及和光谱数据之间的关系，这也正是人力分析不便的地方。利用得到的结论，我们可以结合编程和数据库内的信息，实现生化参量填图的自动化生成，也可以对精细农业的田间管理等起到指导作用。

6.4.2 沧州南大港蝗虫灾害数据分析

近年来，蝗虫灾害屡有发生，给我国农业发展造成了很大影响。国内正在加紧对蝗灾的研究，以期能够有效的检测、减轻灾害乃至遏制蝗灾的发生。高光谱作为监测蝗灾的有效手段之一，近年来发展很快。高光谱遥感的图谱合一的特点，特别是它获取的地物精细光谱（光谱分辨率可达 10-2λ），使得它在地表物质的识别以及相关信息提取方面的独具优势，同时结合植被的时间动态特征，将更加有利于适时准确地对地物分类、农作物品质、病虫害等性状进行监测。尤其是近 2—3 年来高光谱卫星 （如 MODIS，Hyperion）的成功发射，预示着高光谱遥感作为对地观测中的一项重要前沿技术，将在研究地球资源、环境监测中发挥越来越重要的作用。

有很多环境因素蝗灾的发生，气候的因素，包括合适气温与土壤温度，降水量的多少，湖水的涨落，都有这重要的作用。同时气候的因素也影响着植物的生长，间接使蝗虫的营养和它的生殖能力产生变化；土壤的物理化学性质对飞蝗发生地的形成和消灭也是重要的，其中特别是含水量与含盐量的升降；人为因素也是影响蝗虫消长的极其重要的环境因素，直接防治、兴修水利、垦荒、精耕细作都是防治蝗虫的有利途径。所以我们可以通过地面光谱数据来分析研究地表土壤，植被的状况，来达到监测蝗灾程度，以及发现蝗虫滋生地等。

根据南大港的数据的初步分析结果（见下图）

0611---　光谱测量日期　　1—12：编号为 1—12 的不同地块

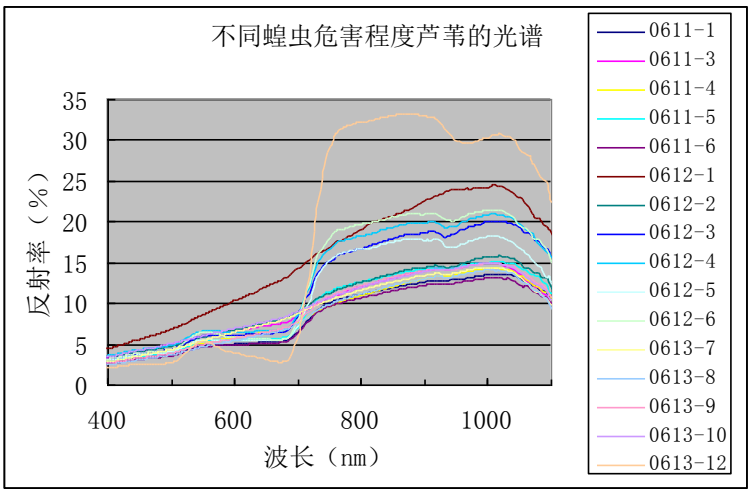


图 6—4　南大港受灾作物光谱曲线图

可见，其中 0613—12 反映出的植物的光谱最为完整，也说明其受损最少；其他不同程度的遭到蝗虫的破坏，光谱完全可以反映出植物受损的状况。我们可以根据实际情况设定植物的受损程度，并根据程度的轻重进行分级，然后以植物最为敏感的红边光谱和受损程度为我

们关心的数据集，进行关联度分析，以后就可以根据所得结果通过光谱直接判断蝗灾的严重程度。

同时，由于我们已经可以得到蝗虫的生长受环境制约的生态数学模型，例如：

下列简式说明(R……代表雨量，T……代表温度，S……代表日照　角标数字代表月份)：

长江下游地区：T5/21+80/R4 下 5 上>2→夏蝗发生，T5/21+80/R4 下 5 上+240/R7.8>3→秋蝗发生，

淮河流域：T5.9/(2 × 21)+S4 下 5 上 /42+100/R7>3 → 秋蝗发生，R7.8/300+100/R9.10+T9.10/(2×21)>3→来年夏蝗发生。

黄河流域：

南部：100/R5.6+T5.9/(2×21)>2→秋蝗发生，R8/180+T9/21+100/R5.6>3→来年夏蝗发生。

北部: 60/R5.6+T5.6/(2×20)>2→秋蝗发生，R8/180+T9/21+S5.6/42>3→来年夏蝗发生。

海河流域：60/R5.6+S6/21+T5.9/(2×20)>3→当年秋蝗来年夏蝗发生。

可以根据这个模型来开发相应的数据挖掘的算法和规则，然后以从实地获得的测量光谱中反演出的土壤水分等生化参量，以及相关环境资料为研究数据集，根据我们开发的挖掘算法进行挖掘，就可以预测出蝗虫可能大量繁殖的地区，从而着手加以防治。

由上述例子可以看出，对于高光谱这种海量遥感数据完全可以通过数据仓库进行整合，通过数据挖掘算法进行挖掘，从而得到更深层次的规律和信息。这方面的应用必将有着光明的前景。

6.5 本章小结

当前数据库应用和发展的热点就在数据仓库和数据挖掘概念上，它解决了数据量爆炸而知识贫乏的问题，特别适用于海量数据的处理。高光谱数据正好符合数据量大，数据间关系复杂的特点，因此在高光谱数据库系统上建立起数据仓库将会极大的促进高光谱的发展和应用。本章从介绍数据仓库、数据挖掘、在线联机分析等基本概念入手，进一步解释了数据挖掘中常见的技术和分析方法，并且对于如何建立一个数据挖掘系统，从设计过程和体系结构等方面进行了描述。最后结合已有的高光谱数据，以小汤山精准农业研究和沧州南大港蝗虫灾害预报研究为例展望了数据挖掘和数据仓库在高光谱中的应用前景。

第七章　　结　论

本文的最终目的是提出高光谱数据库的概念，并且建立基于网络的高光谱数据库原型系统。围绕这一主线，总结、提高了前人的工作成果，针对高光谱数据在 ORACLE 平台上的存储的各种模式进行了研究和讨论，通过对整个系统的难点和重点进行分析、设计与编程实现，提出了在大型数据库平台 ORACLE 下与其他属性、图像结合的高光谱数据的存储规范，并且对比了几种存储模式的优缺点，并最终实现了两种模式。发展了已有的光谱数据库系统，同时在基本的数据管理功能之上添加了数据分析的功能，并且在国内首次在网络上实现了高光谱数据的图谱合一的特点。最后，本文对于将最新的数据仓库和数据挖掘应用于高光谱遥感数据进行了初步的探讨。

7.1 主要工作及结论

（1）在总结前人工作以及数据库开发实践的基础上，本文提出了高光谱数据库的概念，确定了高光谱数据库系统的构成。***高光谱数据库系统是专门面向高光谱数据，体现图谱合一特性，综合了光谱数据库、光谱分析功能和数据挖掘功能于一体的专用数据库系统。***高光谱数据库系统的主要组成部分有：高光谱图像样本库系统；光谱数据库辅助系统；高光谱数据分析系统；带有数据挖掘功能的数据仓库以及前台界面系统。在其中将会存储经过整理的高光谱图像光谱样本数据，再以网络化前台反映给用户，整个系统将为高光谱科学的发展和应用起到重要的作用。

（2）本文为高光谱数据库系统中的图像光谱数据以及普通光谱数据在关系数据库中的存储设计了存储规范：**光谱数据表组＋属性数据表组**。为 ORACLE 数据库平台下高光谱数据库系统中高光谱数据的存储设计了三种存储模式：**波段独立顺列式、波段集中整合式以及表单位式**。同时，在数据库开发实践中加以应用，进行了对比分析，提出了选用的条件。

（3）根据本文中提到的高光谱数据库系统的系统组成建立了高光谱数据库原型系统。

（4）以北京农业信息中心 2002 年小汤山的数据为数据样本，建立了一套基于大型数据库平台 ORACLE 的集基本属性、图像数据、光谱数据为一体的面向精准农业的高光谱数据库系统。将国内光谱数据库系统又向前迈进了一步。

（5）将整个高光谱数据库系统以网络平台加以发布，实现了高光谱数据的远程互连、访问以及数据共享，促进了高光谱遥感技术的信息化、普及化建设。

（6）在实现了数据库管理的各项基本功能后，针对农业应用的特殊需求，在系统中加入了光谱数据以及属性数据的分析和处理，为数据能够更加有效的得到利用奠定了坚实的基础。

（7）针对高光谱遥感的海量数据，对于数据仓库和数据挖掘在高光谱遥感数据中的应用进行初步的探索与讨论。

7.2 论文的特色与创新点

7.2.1 论文特色

本文以工程项目为背景，在解决实际问题的同时注意归纳总结，参考文献，注重科学研究，一方面以科研的精神和思路使得自己工程的进展、分析、设计更加具有结构性和合理性，一方面以工程建设的思路和方法在科研方面提出了具有广泛意义的规范和模式，这样不但在工程上做到了思路清楚，研发合理，同时也是自己的学术水平有了进一步的提高。

7.2.2 创新点

（1） 在国际上，本文总结了前人的工作成果，结合自己的研究实践，首次提出了高光谱数据库的概念，陈述了高光谱数据库系统组成以及实现的途径，并建立了原型系统。

（2） 本文首次为高光谱数据在关系数据库中的存储提出了存储规范。数据以数据库平台进行管理已经成为大家的共识，提出一种存储规范，对于今后高光谱遥感数据的整理和统一有着重要的作用。

（3） 本文首次提出了在 ORACLE 数据库环境下高光谱数据几种存储模式，对其各自的优缺点进行对比分析，并对于其中两种加以实现。ORACLE 是当前比较流行的大型数据库软件，应用比较广泛，这为今后高光谱数据在数据存储、共享方面的开发打下了一定的基础。

（4） 首次将高光谱数据库系统以网络化的形式加以发布，既是将光谱数据库进行网络发布，也是第一次将高光谱数据的图谱合一的特点在网络环境下实现，这不但给新提出的高光谱数据库系统建立了范本，同时为促进高光谱研究的发展，加强高光谱数据的交流和共享，更直观、更信息化的普及高光谱遥感的概念起到重要的作用。

（5）本文首次针对高光谱遥数据在数据仓库这一数据库最新的发展领域进行了初步的探讨，尤其是高光谱数据也是海量数据，如果能够在存储的基础上对其进行进一步的数据挖掘，必将更加充分的利用已有的数据，发现更深层次的联系和规律。

7.3 全系统的发展和展望

由于时间和条件限制，高光谱数据库系统距离一个完善、成熟的应用系统还有着不小的差距，主要存在以下几个方面可以进一步完善与提高：

（1）本系统只是初步建立起了数据结构，对于提高整个系统的实际应用性能考虑尚嫌不足，有许多数据库的性能调整操作没有实施。

（2）本系统更注重于决定整个性能的数据结构的设计与实现，在应用层面上对高光谱数据多样性的特点考虑尚嫌不足。

（3）在查询方面可以向区间查询发展，即查询时可以写入任意区间，这在实现上有一

定的难度，但是可以极大的方便用户的使用。

（4）在数据分析功能上，只是实现了一些最基本的光谱分析功能，还可以添加更多的功能为应用服务，同时面向其他方面，如岩矿等方面的应用亦会有不同的应用需求可以进一步扩展。

（5）在光谱图像子系统中，由于采用了第二种数据集中整合式光谱数据存储模式，只实现了基本的图谱合一的功能，没有时间开发其他光谱分析功能，也有待日后完善。

（6）光谱图像子系统只是建立了原型系统，实现了主要功能，很多数据管理的功能还有待进一步完善。

作为一个独立运行的网络应用系统，本系统但肯定还有不少问题有待完善与提高，但本文以及本系统的实现已经为网络化的高光谱数据库系统开发建立了良好的规范和实现实例，并且完成了最初的基本的框架，可以顺畅的应用和运行。在此基础上，高光谱数据库系统这一概念必将得到进一步的发展，为整个高光谱技术的应用与实现发挥重要的作用。

参 考 文 献

1．Michael J.Corey 等著 陈跃等译， ORACLE8 数据仓库分析、构建使用指南 新华书店北京发行所 2001 年 1 月第一版

2．杨绍方著， 编程使用技术与案例 清华大学出版社 2001 年 11 月第一版

3．Ben Chang 等著 高波等译， ORACLE Xml 开发手册 新华书店北京发行所 2001 年 1 月第一版

4．王克宏主编， JAVA 语言编程技术 清华大学出版社 1997 年 5 月第一版

5．孙林等，一个新的数据挖掘模型与算法. 计算机应用研究 2001(2) 43－44

6．汤语松、刘相风等， 数据挖掘系统设计. 系统工程理论与实践 2000(9)

7. 周斌等， 一个例子. 计算机工程 2000(6)

8. 谢夏丹， web 上的数据挖掘技术和工具设计. 计算机工程与应用 2001(6) 86-88

9. 韩嘉伟等， web 挖掘研究 计算机研究与发展 2001(4)406-409

10. 邹淘， www 上的信息挖掘技术与实现 计算机研究与发展 2001(8)

11. 邹炎昆， 大型空间数据仓库除探 测绘通报 2000(8)

12. 陈才扣， 基于 web 的时间模式挖掘 计算机应用与研究 2000(7)

13. 邹淘， 基于 www 的文本信息挖掘 情报学报 1999 年第 18 卷第 4 期

14. 许振航，基于 xml 的 web 数据挖掘 应用技术 2001(1)

15. 陈才扣， 挖掘基于 web 的访问路径模式 小型微型计算机系统 2001(1)

16. 梁旭， 一种新的高效关联规则数据挖掘算法 大连铁道学院学报 2001(1)

17. 张雷， OLAM 以及基于 web 的 OLAM 计算机工程与应用 2000(9)

18. 陆丽娜， web 日志挖掘中的数据预处理研究 计算机工程 2000(4)

19. 宋擒豹， Web 页面和客户群体的模糊聚类算法 小型微型计算机系统 2001(2)

20. 申瑞民， 个性化数字服务模型 微电子学与计算机 2001(1)

21．宗 锋，Tomcat 全攻略 西北大学计算机系硕士 2001 年 12 月

22．白继伟，基于去包络线算法的高光谱图像分类方法，硕士论文，2002 年 6 月

23．吴长山，多时相的高光谱数据对农作物的生物物理参量的估算模型研究，硕士论文，1999 年 6 月

24. Robot， Apache Tomcat 4.0 的新特性 www.chinajavaworld.net

25. 吴学启， 信息系统三层结构及其在某配送业态中的应用 云南工业大学信电学院 98 级研究生

26. 王志华 杨斌， 中间件 TUXEDO 在电信计费营帐系统中的应用 深圳市现代计算机有限公司

27．周永銮， 数据仓库技术简介 2001/07/21

28．马世骏， 《马世骏文集》 中国环境科学出版社，1995，p66－67

29. 浦瑞良，宫鹏， 高光谱遥感及其应用 高等教育出版社 2000.8
30. 刘良云， 高光谱遥感在精准农业中的应用研究 博士后出站论文 2002.12
31. 陈述彭，童庆禧，郭华东， 遥感信息机理研究 科学出版社 1998.7
32. Jonathan Gennnick, Carol McCullough-Dieter, Gerrit-Jan Linker 著 赵艳勤，刘冠英等译， ORACLE8i DBA 宝典 电子工业出版社 2000.8
33. Karl Avedal, Danny Ayers, Timothy Briggs 等著，黎文 袁德利 吴焱等译， JSP 编程指南 电子工业出版社 2001.4
34. 清华大学计算中心培训部， ORACLE8i 数据库课程讲义 2001.11
35. John Zukowski 著， 邱仲潘等译， JAVA2 从入门到精通 电子工业出版社 1999.4
36. Philip Heller, Simon Roberts 著，邱仲潘等译 JAVA2 高级开发指南 电子工业出版社 1999.6
37. Michael Abbey, Michael J.Corey, Lan Abramson 著，乐嘉锦，王兰成等译， ORACLE8i 初学者指南 机械工业出版社 2001.8
38. 丁钺 编著， ORACLE8/8i 数据库系统管理 人民邮电出版社 2001.3
39. Scott Urman 著， 陈维军 王蕾等译， ORACLE9i PL/SQL 程序设计 机械工业出版社 2002.8
40. Marlene Theriault, Rachel Carmichael, James Viscusi 著， 王兰成等译， ORACLE 数据库管理员基础教程 机械工业出版社 2000.10
41. Robert J.Muller 著， 王华驹，李连，曲宁，郭天杰等译， ORACLE Developer 使用指南， 机械工业出版社， 2000.6
42. 袁志发，周静芋， 多元统计分析，科学出版社，2002.10
43. 严士健，刘秀芳， 概率论与数理统计， 高等教育出版社， 1993.4
44. 张兵 ，时空信息辅助下的高光谱数据挖掘，博士论文，2002.12

课题研究和文章情况

硕士期间参加课题研究项目：

（1） 参加了北京市基金项目“面向精准农业的作物高光谱数据库”，负责报告撰写，数据库系统设计与建设

（2） 参加了国家 863－13“我国典型地物标准波谱库”项目，参与了报告撰写和数据库建设工作

（3） 2002 年与北京农林科学院合作的北京小汤山精准农业示范项目，参与了地面调查工作

（4） 参与并组织了马来西亚遥感专业人员来华培训；负责外宾接待，技术支持，网络环境建构及软硬件调试，维护

（5） 参加了清华大学计算中心举办的 ORACLE 培训班的学习，获得了结业证书

硕士期间撰写的文章

（1）The preliminary research in using the technology of data mining to analyze the remote sensing data(已接受,SPIE, 2003)

（2）Designing a Hyperspectral Database System based on the web(待发表,SPIE, 2003)

（3） 高光谱数据在数据库中的高效存储研究（已接收，《遥感学报》）

致　　谢

完成本文以后，回想三年来的学习与研究经历，不禁感慨万分。三年中我经历了风风雨雨，既有努力付出的喜悦，也有品尝挫折的痛苦，正是得到了许多老师的谆谆教诲和同学的大力支持和无私的帮助，我才能够顺利完成学业。

首先要衷心感谢我的导师童庆禧院士、郑兰芬女士和张兵研究员的精心指导。童老师和郑老师亲自为我指定了研究方向，使我能够在一个全新的领域学习和研究，张老师则在很多具体研究时遇到的问题解决上以及论文撰写上给了关键性的指导和帮助。他们在百忙中总是关注着我的学习和研究进展，并不断地从思想上、方法上为我排疑解难。童老师幽默睿智的谈吐和广博的见识以及对问题入木三分的见地无不令我深深地折服，郑老师对我的日常学习、生活与工作给予的细致入微的关怀更令我深深地感动，张老师认真负责的工作态度，扎实的工作作风，丰富的工作经验和专业知识令我十分钦佩。

其次要衷心感谢的是高光谱室这个小集体：其中赵永超博士出色的学识，和蔼、热心的态度使我受益非浅，终身难忘；张霞博士以其丰富的工作经验和乐于助人的精神在我的课题和论文方面给了我极大的帮助；李兴博士在高光谱方面广泛的积累，开阔的思路以及真挚的友情对我有很大的启发和激励；周丽萍女士做了很多细致、艰苦的工作，和胡兴堂、耿修瑞、陈正超、刘良云、刘团结、吴传庆、白继伟、刘伟东、关燕宁等同学和同事的交流也让我学到了很多东西。三年来，大家的友情、关怀与帮助使我能克服各种困难完成硕士论文，使我有了无比的收获。

还要衷心感谢的是我在清华的同学和老师们，罗萌、潘澈、高山、刘为、李炜、孟虎、闻立杰以及计算中心的老师们，正是由于他们的帮助，我才能够解决计算机技术上一个又一个的问题，使整个系统得以最终完成，顺利完成我整个硕士的工作。我的母校清华大学以及我的自动化系给了我精神上莫大的安慰和鼓舞，在此深表感谢！

本所图像室的朱海青和安老师在数据库方面也与我有过很多交流；工程中心的芮小平博士也在 java 方面给了我很多的帮助，在这里一并表示感谢。

衷心感谢研究生部主任余琦老师、吴晓清老师在学习、生活以及工作方面给我的关心和帮助。

衷心感谢我的女友，她的关心和鼓励让我能够在最感到挫折和泄气的时候重新鼓舞和振奋。

许多老师和同学在我的学习过程中都给予过真挚的关心，感激之情，拳拳在心。

最后一句和大家共勉：

无论我们失去多少，我们得到的只会更多

^_^