基于数据挖掘的 ARP 数据分析系统①

王凤霞，卢景秀，杜义华

(中国科学院计算机网络信息中心 ARP 中心，北京 100184)

摘 要：中国科学院资源规划项目(Academia Resource Planning，简称 ARP 项目)，是实现中国科学院资源规划的信息系统工程。该项目从中国科学院院所两级治理结构出发，以科技计划与执行管理为核心，综合运用创新的管理理念和先进的信息技术，对全院人力、资金、科研基础条件等资源配置及相关管理流程进行优化与整合。本文通过对当前数据挖掘技术的研究，利用数据挖掘技术对 ARP 系统大量数据进行了分析，获取其中有价值的信息和知识，方便院领导、管理人员及科研人员提供信息服务和决策支持，充分发挥 ARP 效益。

关键词：中国科学院资源规划项目；数据挖掘；管理决策支持系统

ARP Data Analysis System Based on Data Mining

WANG Feng-Xia, LU Jing-Xiu, DU Yi-Hua

(ARP Center, Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Academia Resource Planning project is an information system project to achieve resource planning of Chinese Academy of Sciences. ARP project serves two-level management organizations including CAS and its institutes, integrating and optimizing the business management flow of scientific research programs, human resources, comprehensive finance and scientific research conditions. This paper attempts to analyze large amounts of ARP data based on currently data mining techniques, obtaining valuable information and knowledge to provide information services and decision support to facilitate leaders, managers and researchers, maximizing the effectiveness of ARP system.

Key words: ARP; data Mining; DSS

1 引言

1.1 研究背景

ARP 数据分析系统，是面向中科院综合管理的需要，通过数据交换平台，从各业务系统中获得相关数据源，集中存储于信息资源数据库，在此基础上，建立可靠的、易用的数据展示平台，并通过对这些数据的分析、组织，结合对展示界面的分析、设计，实现对综合性数据的可视化利用，进而为各级领导、管理和科研人员提供信息服务和决策支持。

随着 ARP 系统数据量的逐年增加以及运行方式更加灵活多样，目前 ARP 系统面临着业务量的大量增加与资源日趋紧张的矛盾，同时科学院的一些机构也在不断地改革变化，这些都给 ARP 数据分析系统带来了前所未有的发展和挑战。

1.2 当前现状与趋势

当前情况下，科学院各级领导层都希望从整体的、宏观的角度了解形势，优化资源配置，提高资源利用率。因此，建立一个集数据汇聚、传输、处理、存储、利用为一体的高效的 ARP 信息资源管理与服务平台则显得十分必要。该系统的各项功能除了能满足院、所两级查询和综合统计外，还可以通过数据挖掘等技术开发历史数据，提取隐含在其中的事先未知的、潜在的、有价值的信息，为各级领导、管理部门、科研人员提供不同层面的信息服务和决策支持。

① 收稿时间:2011-07-27;收到修改稿时间:2011-09-08

2　数据挖掘分析技术

2.1 管理需要数据挖掘分析

基于数据仓库的数据挖掘技术，其任务是发现数据仓库中尚未被发现的知识。对于那些决策者明确了解的信息，可以用查询工具直接获取，而另外一些隐藏在大量数据中的关系、趋势等信息就需要数据仓库技术。数据仓库技术可从数据仓库中找出大量真正有价值的信息和知识，可以更好地对科学院的发展历程和未来趋势做出定量的分析和预测。为各 ARP 的管理决策者提供更科学的决策基础，从而有效地提高数据质量，有针对性地加强科学管理。

根据目前 ARP 数据系统的特点，首先需要在较高层次上将不同信息系统中的数据梳理、归类，并进行分析利用的抽象，即建立数据仓库，在数据仓库的基础上进行联机分析处理和数据挖掘，为科学决策提供依据支持。

2.2 数据挖掘分析的任务

数据挖掘的任务是发现知识，主要包括以下几类知识的发现：广义型的知识，反映同类事务共性的知识；特征型知识，反映事物各方面特征的知识；差异性知识，反映不同事物之间属性差别的知识；关联型知识，反映事物之间依赖或关联的知识；预测性知识，根据历史和当前的数据推测未来的数据；偏离型知识，揭示事物偏离常规现象[1]。

3　ARP数据分析系统技术分析

3.1 ARP 数据分析系统架构

根据科学院的特点，我们在“十一五”期间组织建设了集数据汇聚、传输、清洗、存储、分析、挖掘及展示功能为一体的高效的 ARP 信息资源管理与服务平台，该平台针对全局（院所两级平台）信息资源，构建了科研管理资源目录，建立了共享发布平台，按业务需求组织资源目录，纳入结构化和非结构化资源，提供了统一的应用方式，实现资源信息各级平台内的共享。通过该平台可以支持满足不同需求的个性看板配置，并提供动态、多维信息分析服务功能以及统一的报表管理功能和强大的信息搜索功能。其总体架构图如下所示：

3.2 数据挖掘方法在系统中的应用

数据仓库的结果体现在知识的发现上，面对科学院管理的需要，如何从众多的挖掘技术中精心选择出有效的技术和方法，是研究和开发 ARP 数据分析系统的首要问题。ARP 系统中主要用的数据挖掘技术有以下几种：

图 1　ARP 数据分析系统总体架构图

(1) 人工神经网络：用于分类、聚类、特征挖掘、预测和模式识别。人工神经网络从结构上模仿生物神经网络，通过简化、归纳、提炼总结出来的一类并行处理网络。利用其非线性映射的思想和并行处理的方法，用神经网络本身的结构来表达输入和输出的关联知识。

(2) 概率论和数理统计：侧重于应用研究随机现象本身的规律性来考虑资料的收集、整理、分析，从而找出相应随机变量的分布律或数字特征，尽可能做出较合理精确的推断，包括决策树推断、规则推断、最近邻方法、聚类方法、关联规则等。

(3) 关联规则方法：用于对关系数据库发现有价值的关联模式，或对半结构化数据（如文档数据）进行关联规则挖掘。关联分析可以分为两种，关联规则和时序分析。关联规则即在当前记录的各个特征间寻找内在的联系。时序分析即在历史数据中寻找具有时间上相关的记录间的规律性。

3.3 ARP 数据分析系统主要功能

ARP 数据分析系统依据信息价值链全生命周期管理的理念，针对 ARP 系统院所两级体系架构，从信息的获取、交换与传输，信息的存储与管理，信息的处理发布以及信息共享服务等方面进行设计与开发。功能模块包括了搭建数据指标体系、信息交换与传输、信息采集与管理、信息处理与发布、信息共享与服务、辅助决策动态展示以及系统管理等内容。

1、构建数据指标体系

为规范数据质量提供的明确的数据范围，同时为

了给后续数据挖掘分析提供坚实、准确的数据保障，ARP 数据分析系统首先对数据指标体系进行了梳理，针对 ARP 系统及三级用户群的特点，数据指标体系工作包含以下三部分内容：

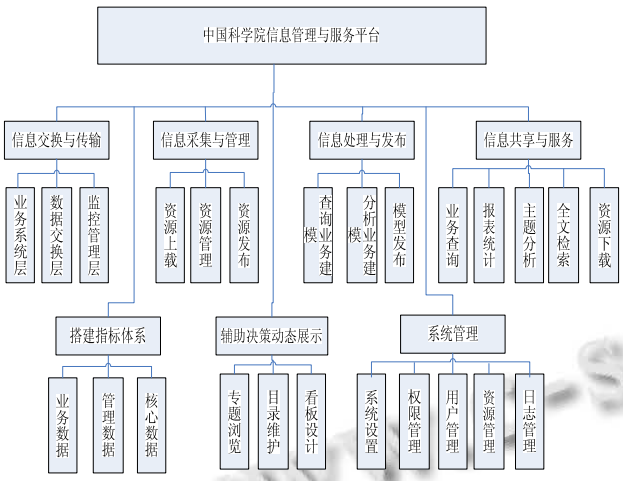

图 2 ARP 数据分析系统功能模块

(1) 业务基础数据指标：满足各业务系统日常管理与应用的数据指标。主要包含院所两级系统共 10 个模块的业务数据实体。

(2) 综合管理数据指标：满足国家、院、所综合管理部门的数据指标，来源于业务基础数据指标。主要包括交换到信息资源中心的 10 个模块的业务数据实体。

(3) 核心数据指标：满足院宏观管理决策层需要的数据指标，来源于基础数据指标和综合管理数据指标。主要包括信息资源服务用于院宏观管理决策的相关数据指标。

2、信息交换与传输

数据交换与传输平台通过集成中间件来构架数据交换的基础中间件平台，建立完全分布式的数据交换体系。数据交换平台的总体架构分为三层：业务系统层、数据交换层及监控管理层。

业务系统层是信息资源交换与共享的数据源头，提供了信息交互与共享的原始数据；

数据交换层的交换传输是通过不同种类的适配器和数据交换服务总线的方式，将各业务子系统封装为松耦合的服务接口，对服务按照业务需求进行编排形成数据交换流程。使得资源提供者提供的资源能够通过数据交换平台，根据业务同步规则在安全通信的前提下，定时批量的将共享的资源传递给资源使用者。

监控管理层提供了基础信息库管理与维护、日志管理及统计分析、用户及权限管理、数据备份以及对整个交换平台的运行情况监控等管理监控功能，采用基于 B/S 架构的管理控制中心能够实现移动管理和远程维护。

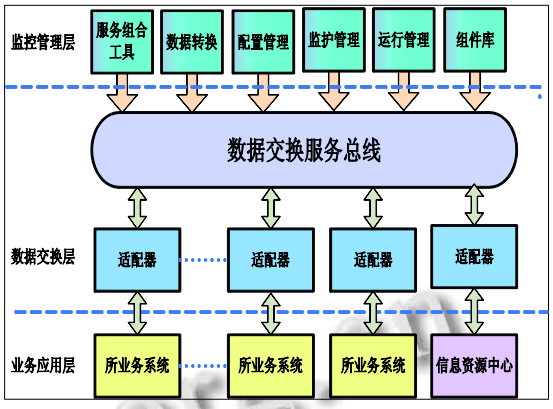

图 3 ARP 数据交换平台架构图

3、信息存储与管理

基于内容管理，实现信息上载、发布、浏览、下载功能，可以为部门间业务互动及数据共享提供支持。

(1) 资源上载：各类信息由各自职责部门职责岗位业务人员上载；

(2) 资源发布：信息上载成功后，可由相关权限人员进行发布操作，从而决定该信息的授众，进而控制信息的安全性；进行发布操作时，提供面向角色和面向用户的两种发布操作；为简化发布操作，结合资源目录的操作权限，对未进行发布操作的信息进行默认发布操作，即各信息继承其所属资源目录的分配权限。

(3) 资源的浏览与下载：在授权范围内查询浏览业务信息，并且可对权限范围内的信息进行下载操作。

4、信息处理与发布

(1) 通过定义查询业务数据集，利用平台提供的内置查询解析引擎，解释并执行预定义的查询模型。在此基础上可以进一步进行元数据的维护，定义业务数据集的输出项及显示形式以及定义业务数据集之间的联系，从而提供数据挖掘功能。

(2) 利用数据仓库建模，把现有数据库以二维关系表存储的数据转化为具有多维特征的数据集，处理过程中我们利用 OWB 工具把二维关系表转化为星型架构的多维分析表存储于数据仓库中。多维立方体中的事实表简称为立方。而在系统底层数据仓库中包括了涉及科学院科研管理活动中的人事、课题、经费、产出物、资产等五个立方体以及他们的公共维度表和私有维度表。

在建立好维度和立方之间的映射关系后，经过 ETL 过程，最终把数据通过建立好的映射关系存储于 OWB 建立的资料档案库中，进行实际的物理存储，提

升数据分析人员进行数据分析时的响应速度。

5、信息共享与服务

在信息资源服务上通过查询业务建模的定义与发布，为用户提供常用的查询服务，服务内容有：

(1) 业务查询

通过查询业务建模模块的定义与维护，无缝集成院所两级管理数据库、业务数据库、主题库丰富的业务数据，提供界面标准、操作标准、流程标准的业务查询功能，实现一站式管理数据的综合查询。

(2) 报表统计

根据实际的业务需求，在日常管理中、不同部门会有多种统计报表的需求，通过报表组件提供强大的中国式报表统计功能，并且提供集成功能，使用户在不必关心报表平台的操作细节即可方便的浏览各类统计报表

(3) 主题分析

根据实际的管理需求，对关系较强、较复杂的一个数据集合，从多个视角、不同层次、不同组合模式为信息管理与服务平台提供基于人力资源管理、科研条件管理、科研项目管理、综合财务管理业务等各类智能分析应用。

(4) 全文检索

通过用户输入的关键字，在资源目录下进行全局检索，同时查询得到匹配关键字的文档资源、图片资源、报表信息、综合分析等相关结构化信息和非结构化信息。

(5) 资源下载

信息管理与服务平台的终端用户，在授权范围内查询浏览业务信息，并且可对权限范围内的信息进行下载操作。

6、辅助决策动态信息展示

在利用业界先进工具搭建服务平台时，一方面受制于工具/产品特性（以技术为导向），用户在应用时步骤繁琐、操作复杂；另一方面由于信息资源中心服务的目标对象范围庞大，各自需求及习惯差异较大，导致所提供的信息服务缺少统一性、标准性、友好度。如何将信息服务由技术导向转变为业务导向已成为信息化建设的当务之急。因此，针对这个问题，在平台建设的过程中，设立了辅助决策动态信息展示系统的专项开发，进一步分析科研业务的内在规律，总结关键决策点规律，充分发挥数据、信息和图表对业务管理、决策支持的效力，提升服务质量。

辅助决策动态信息展示系统的应用架构为一个核心——管理元素（资源元素）；两类基础——资源管理、系统管理；三种模式——浏览模式、设计模式及管理模式。

(1) 一个核心

管理元素：是本系统的核心，每个管理动态看板由多种基础元素组合而成，例如：描述性文本、分析结论性文本、效果图、统计表等，其管理和应用贯穿于系统三种模式之间。

(2) 两类基础

系统管理：提供用户管理、权限管理等保证系统日常运行的管理功能；

资源管理：维护管理元素基本信息（包括类型、名称等），同时维护相关展示信息及数据信息。

(3) 三种模式

浏览模式：基础应用模式，提供查询、播放、下载、导出等常规功能。

设计模式：设计看板模式，基于管理元素提供看板的编辑功能，即通过管理元素组合编辑管理看板，提供所见即所得的设计界面。

管理模式：系统管理模式，提供管理元素的维护功能，包括维护元素的展示方式和业务数据集合，提供资源管理、系统管理相关功能。

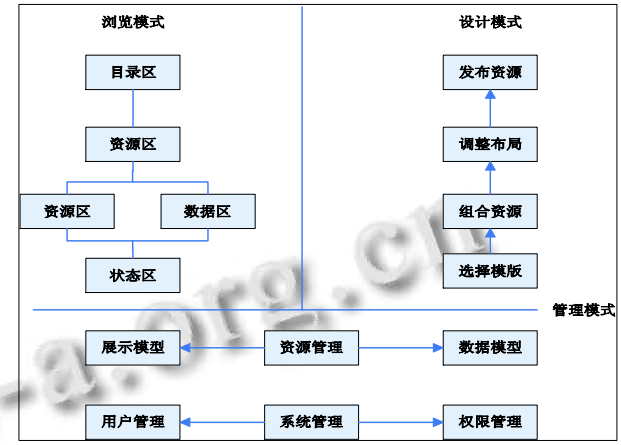

图 4 辅助决策动态信息展示系统应用架构

4 总结和展望

数据仓库、数据挖掘和决策支持系统都是新兴前沿科学，数据挖掘技术为管理决策提供了一种有效的、可行的解决方案。随着 ARP 系统的不断深入应用，ARP 系统的作用将逐渐从服务于日常科研管理活动向为领导决策提供支持服务功能方面进行转变。

参考文献

1 张震.数据挖掘技术分析及其在高校管理决策中的应用.远程教育,2005.6.