基于数据仓库的土壤环境监测综合挖掘模型构架

郑向群[1,2], 赵 政[1※], 刘东生[3]

（1．天津大学计算机科学与技术学院，天津 300072； 2．农业部环境保护科研监测所，天津 300191；
3．农业部规划设计研究院，北京 100026）

摘 要：为解决土壤环境监测全流程优化问题，提出了监测流程综合挖掘方法，采用数据仓库和工作流挖掘技术，构建了一个土壤环境监测综合挖掘模型构架。首先，采用雪花结构建立了数据模型构架，为流程挖掘提供数据底层；然后建立了监测流程综合挖掘模型构架，给出了监测单元挖掘、监测点位挖掘和监测数据挖掘，以及综合挖掘的模型函数和参数，以局部优化和全局优化的思路实现了土壤环境监测全流程优化。研究结果对于提高土壤环境监测效率，拓展 LIMS 实验室信息管理系统功能，构建土壤环境监测数据仓库和专家系统提供指导。

关键词：土壤环境监测；数据挖掘；数据仓库；模型构架
中图分类号：TP311.13；X51 **文献标识码**：A **文章编号**：1002-6819(2008)-8-0162-07

郑向群, 赵 政, 刘东生. 基于数据仓库的土壤环境监测综合挖掘模型构架[J]. 农业工程学报，2008，24(8)：162－168.
Zheng Xiangqun, Zhao Zheng, Liu Dongsheng. Integrated mining model framework for soil environmental monitoring data based on data warehouse[J]. Transactions of the CSAE, 2008, 24(8)：162－168.(in Chinese with English abstract)

0 引 言

随着中国土壤环境监测体系的逐步完善，监测数据快速积累和增长，给传统的数据分析方法带来巨大的挑战，如何利用海量的监测数据为宏观决策服务成为环境监测领域的重要课题。数据挖掘（datamining，DM）与知识发现（knowledge discovery from databases，KDD）技术是人工智能、机器学习与数据库技术相结合的产物，其主要用于从商业数据库和数据仓库中，即通过提取有用知识来支持高层管理部门的决策[1]。但它不是简单地检索和查询信息，而是从数据集合中发现隐含的、先前不知道的潜在有用知识[2]。目前，数据挖掘和知识发现已经由计算机领域扩展到其他行业，在人工智能、医学、市场营销、土地管理、农业管理等诸多领域已经获得诸多应用[3]。空间 OLAP 的出现又使数据挖掘拓展到空间数据领域[4]。但在环境保护领域，国内开展这方面研究还不多见，而且基本处于经典挖掘方法的实例应用方面。应用较多的挖掘方法主要有数理统计、神经网络、决策树、模糊聚类、地统计等等，本文作者在 2003 年就对这方面的研究做了比较全面的描述[5]。统计分析技术是目前应用最广、最成熟的挖掘方法，大量土壤监测数据分析研究成果均基于统计分析完成。其他计算机挖掘算法应用于土壤环境质量研究则相对较少，例如，中科院南京土壤所的檀满枝等人利用模糊 C-均值聚类法对土壤进行了重金属污染空间分布的预测等[6]。而相对较多的是在土壤面源污染上的一些应用，例如，西安理工大学的李家科等人利用支持矢量机技术预测面源污染负荷[7]，南京大学的钟晓兰等人利用地统计技术分析长江三角洲地区的土壤重金属空间分异特征等[8]，同时也有部分学者开始利用多种数据挖掘模型进行面源污染的空间模拟研究[9]。综合上述各类专家学者的研究成果，我们发现目前环境监测数据挖掘技术仅仅只限于某个挖掘算法在监测过程某个环节的具体应用,而如何对整个监测流程进行优化，实现环境监测综合挖掘的研究尚未见报道。

本文将数据挖掘技术引进环境监测评价领域，提出一种土壤监测数据流综合挖掘模型构架，以数据仓库为基础建立数据模型，以 LIMS 实验室管理信息系统（Lab Information Management System, LIMS）为平台，以全流程挖掘为基础建立综合挖掘模型。先对监测流程的单个环节进行分步挖掘，实现局部最优，然后将每个过程的挖掘结果作为对象进行集成优化，达到全局最优，从而实现提升监测活动整体代表性的最终目的。本文讨论的土壤环境监测综合挖掘模型实质上是一种综合评价模型，是对土壤环境监测实行全流程优化控制，保证最终评价结果的科学合理。研究结果将极大地拓展土壤环境监测数据的外延和以宏观决策服务为目标的数据分析方法，对有效发挥监测数据的作用，为中国社会经济的持续发展提供基础支撑具有重要意义。

1 土壤环境监测工作流分析

土壤环境监测是一个大数据量的工作流系统，是由监测单元划分——布点采样——样品测试——数据处理/评价——检测报告等结点组成[10]。具体流程如下：首先，依据监测对象和监测区域的特点，将监测对象性状相近、监测环境状态基本类似的区域归并为一个单元，进行监

收稿日期：2008-03-14 修订日期：2008-07-26
作者简介：郑向群（1974－），男，湖北英山人，副研究员，博士研究生，主要研究方向：农业环境信息，数据挖掘，地理信息系统，遥感技术。天津农业部环境保护科研监测所，300191。Email: zhengxiangqun@cae.org.cn
※通讯作者：赵 政（1948－），男，教授，博士生导师，主要研究方向：数据库、地理信息系统、网络、CIMS。天津 天津大学计算机科学与技术学院，300072。Email: zhengzh@tju.edu.cn

测单元划分，然后，在各监测单元内布点采样，要求各采样点必须能代表该监测单元的环境质量特点，样品采集完成之后，进行上机测试，得到样品的环境质量数据，进行数据处理和评价，得到监测单元的环境质量状况，最后依据数据评价结果，综合监测区域的调查及采样信息，提出土壤环境条件决策意见，形成检测报告，整个土壤环境监测活动完成。

显而易见，土壤环境监测业务工作流是一个具有时间复杂度和空间复杂度的工作流系统，具备以下两个特点：第一，工作流相关数据格式多样、数据量大，有文本、数据、遥感图象、GIS 数据，以及知识数据等，涉及土壤环境背景数据、调查数据、现场测试数据、实验室检测数据等，目前，已经有 LIMS 实验室管理信息系统可以对数据的生成、流转以及存储进行管理[11]，但如何有效的进行数据管理，提高数据利用效率，仍然需要有挖掘方法来支持解决。第二，在整个土壤环境监测工作流中，单元划分、采样布点、上机检测、分析评价等活动结点都会接受前驱结点的资源和数据，受到前驱结点的误差影响，一旦处理不当，就会在信息传递的过程中产生信息扭曲现象，即所谓的牛鞭效应（Bullwhip Effect）[12]，导致误差逐层放大而不可控，整个监测业务流程就会失去应用价值和意义。

土壤环境监测业务的目的是为了获得具有代表性的数据，为土壤环境治理提供决策依据。因此，如何提高监测工作的代表性是我们必须考虑的问题，传统的方案是以质量保证和质量控制技术为主导，在布点采样和上机检测等单维数据空间中采取措施，简化了监测业务的海量背景信息，没考虑监测区域的历史、地理等多维数据空间的交叉影响，选取的采样区域、点位、样品和数据只能在一维扁平空间中具有代表性，而在时间、空间跨度上无法反映监测区域的真实情况；同时，传统质量保证方法只能实现整个监测业务链的单环节局部最优，无法实现监测业务的全流程优化，因此，牛鞭效应导致的误差放大问题无法得到有效解决，而且，一旦某个环节出现差错，即使其他环节实现了局部最优化，整个监测业务流程也将缺乏代表性。

因此，如何综合考虑各种监测活动影响因素，充分利用监测工作流相关数据，实现监测流程的全局优化，提高土壤环境监测工作流运行效率，一直是广大环境监测技术人员长期关注的问题。

2　土壤环境监测综合挖掘模型构架

本文提出的综合挖掘模型构架是以土壤环境监测业务标准工作流为基础，以数据仓库技术解决业务工作流大数据量的管理和利用问题，以分步挖掘和综合挖掘技术解决监测流程总体优化问题。整个模型构架包括两部分：数据模型构架、流程挖掘模型构架。首先，采用雪花结构建立了数据模型构架，为流程挖掘提供数据底层；然后建立了监测流程挖掘模型构架，给出了监测单元挖掘、监测点位挖掘和监测数据挖掘、以及综合挖掘的模型函数和参数，以局部优化和全局优化的思路实现了土壤环境监测全流程优化。

2.1　数据模型构架

设计数据模型构架的目的是为挖掘活动提供数据底层，构建初始分析空间。环境监测数据量巨大，存储格式多样，要想充分有效地利用海量监测数据进行挖掘，最关键的是必须设计出一套合理的数据结构。本文构建的综合挖掘模型构架是基于数据仓库的，但由于数据仓库是为一个非常大的群体服务的，所以对于任何一个需求集合而言，性能和便捷性都不是最优[13]。因此，最终选择数据仓库中形成的数据集市运用到应用层面。

2.1.1　数据仓库建模方法

整个数据模型构架建模方法是：先从不同介质、不同结构的数据源中获取数据，包括资料收集、现场调查、实验分析、预测评估等活动中取得的文本、数据、遥感图象、GIS 数据，以及评价规则等，取得数据后要经过数据变换和集成，目的是消除数据的属性特征差异、空间特征差异、时间特征差异、数据精度差异、数据整体表现的差异等[14]，然后将他们按照一定的粒度和尺度进行规范化纠正，使得各种数据类型能够在定义域空间中叠加。然后进行抽取、整理等操作，取得规范标准的数据集合装载进数据仓库，形成数据集市，最后提供给流程挖掘模型进行数据挖掘。具体数据建模流程如图 1。

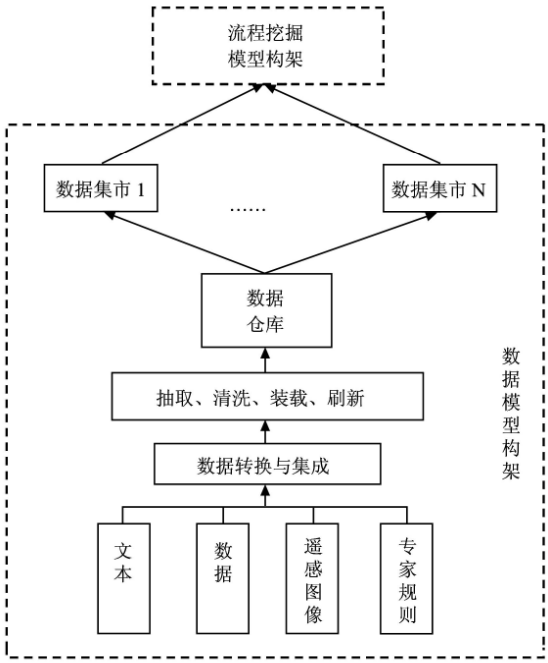

图 1　数据仓库建模流程图

Fig.1　Flowchart of building the model of data warehouse

2.1.2　数据模型结构

本文使用最常用的 E-R 模型方法作为概念模型组建数据仓库和形成数据集市，由于系统的数据繁杂，为了提高挖掘效果，在模型中采用了雪花结构（Snow Flake Schema）[15]。诚然，在数据挖掘过程中，星型结构是最常见的模型范例，但是，在土壤环境监测数据挖掘模型中，维表大而复杂，可能拥有上百万条数据和各种各样的属性，为了改善查询性能，本文对维表进一步层次化，

在星型结构的基础上拓展成雪花结构，在维表上连接对事实表进行详细描述的详细类别表，从而缩小事实表，最大限度的减少数据存储量以及联合较小的维表来改善查询性能[16]。雪花模型增加了用户必须处理的表数量，减少了数据仓库结构的直观性，但这种方式可以处理多对多关系的结构，使系统进一步专业化和实用化，而且，一旦需求已知，我们可以将数据集市转化成一个最优的星型连接结构，从而达到数据仓库的高性能和易使用性。同时，本文在数据仓库建模过程中引进知识库建立维表，拓展了后期流程挖掘优化的向量空间。

“检验结果”由于涉及较多污染要素分析结果，将是一个被大量载入数据的实体，是一个核心属性，而“监测单元”、“点位”等实体的数据量相对单一。因此可以将“检验结果”确定为事实表，“监测单元”、“点位”等确定为维表。然后对这些维表进行扩展，将“监测单元”、“点位”设计成小的事实表，将它们的各种属性分散到小维表中。事实表中通常有多个外关键字(FK)，外关键字是连接事实表和维度表的桥梁，它连接到相关维度表的主关键字（PK）上。使用这种连接方法得到的数据仓库引擎只需使用一次事实表的索引，就可以求得两类表间的任意 N 种连接结果[17]。“专家规则”组成知识库，对“检测结果”有着很强的支持度，因此将其单独建成“检测结果”的一个维表。系统的雪花结构模型如图 2 所示。

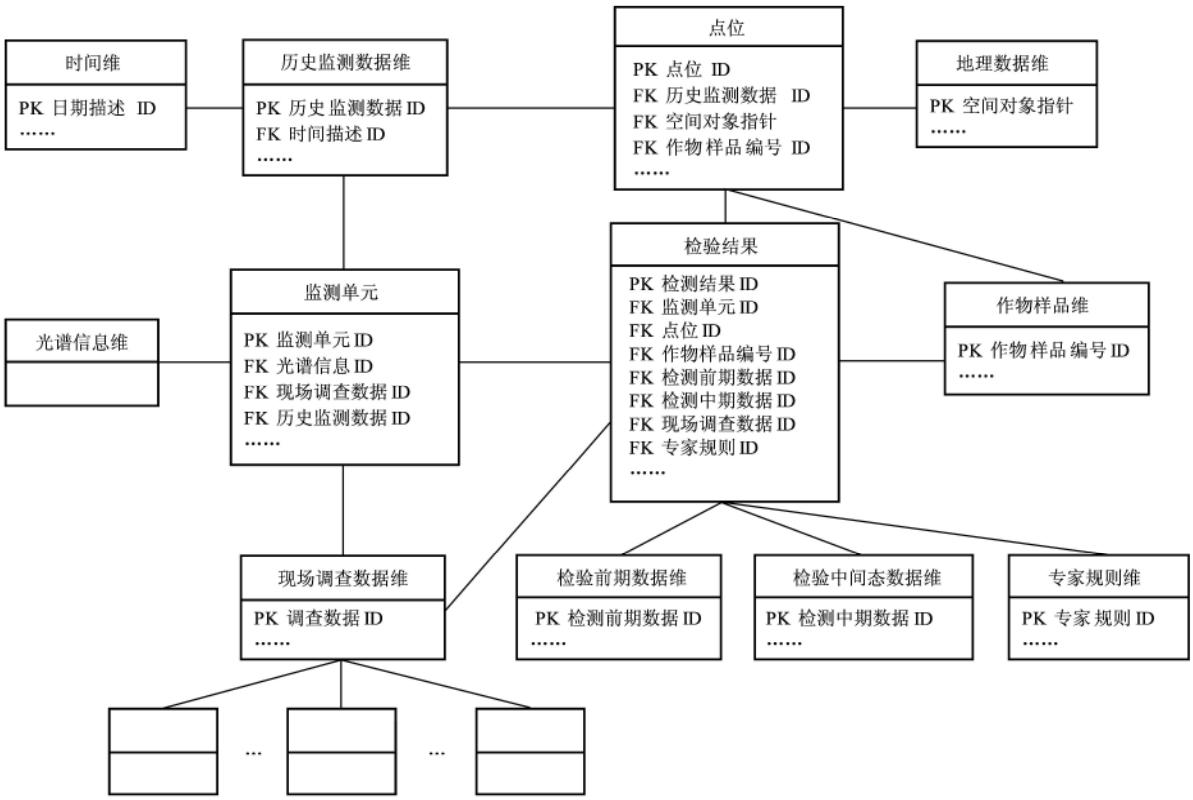

图 2 土壤环境监测数据雪花模型简图

Fig.2 Diagram of the snow flake model for soil environmental monitoring data

事实表：

1）检测结果事实表，保存检测结果相关信息。其结构包括（PK 检测结果 ID，FK 监测单元 ID，FK 点位 ID，FK 作物样品编号 ID，FK 检测前期数据维 ID，FK 检测中期数据维 ID，FK 现场调查数据维 ID，FK 专家规则 ID，检测结果数据向量）。

2）监测单元事实表，保存监测单元相关信息。其结构包括（PK 监测单元 ID，FK 光谱信息 ID，FK 现场调查数据 ID，FK 历史监测数据 ID，监测单元面积，……）。

3）点位事实表，保存监测点位相关信息。其结构包括（PK 点位 ID，FK 历史监测数据 ID，FK 空间对象指针，FK 作物样品维 ID，经纬度，……）。

维度表：

1）历史监测数据维表，保存历史监测数据信息。其结构包括（PK 历史监测数据 ID，FK 时间描述 ID，历史监测结果，……）。

2）现场调查数据维表，保存监测现场调查数据信息。其结构包括（PK 调查数据 ID，时间 ID，农业生产土地利用状况向量，区域土壤地力状况向量，土壤环境污染状况向量，土壤生态环境状况向量，土壤环境背景向量，植被分区向量，作物长势向量，土地利用规划向量，……）。该维表可以依据调查数据量的多少再进行分层，依据各向量进行分类，继续向外扩展成其他类别表，例如农业生产土地利用状况维表、区域土壤地力状况维表、土壤环境污染状况维表、……等，在图 2 中以空白表表示。

3）作物样品维表，保存作物样品相关信息。其结构包括（PK 作物样品编号，作物长势向量，污染程度向量，……）。该维表可以根据实际情况由现场调查数据维表衍生，也可以单独建立，这样使得数据结构更加直观，而且能够提高查询速度。

4）检验前期数据维表，保存样品检测前期的相关信息。其结构包括（PK 检测前期数据 ID，样品类别，温度，湿度，含水量，……）。

5）检测中间态数据维表，保存样品检测中产生的数据，可以通过实验室信息管理系统（LIMS：Lab Information Management System）平台进行提取。其结构包括（PK 检测中期数据 ID，平行检测结果，盲样检测结果，……）。

6）光谱信息维表，保存监测地点的遥感数据信息。主要由遥感影象的各波段数据组成，可以利用元数据库进行继续扩展。

7）地理数据维表，保存监测地点的 GIS 空间数据和属性数据。其结构包括（PK 空间对象指针，空间对象名称，临近空间对象的拓扑关系，……），也可以利用元数据库进行继续扩展。

8）时间维表，保存监测活动时间信息，也是每个业务主题必须用到的，因为每个业务主题都是时间序列的。其结构为（PK 日期描述 ID，年，季度，月，日，……）。

9）专家规则维表，保存在监测活动中，不同判定规则组成的知识。其结构为（PK 专家规则 ID，规则空间，规则，……）。

数据仓库建立起来后，依据各挖掘步骤的需要形成各种不同的数据集市。流程挖掘模型根据输入空间的要求选取相应数据集市作为底层数据源，构建初始空间，驱动挖掘活动。

2.2　流程挖掘模型构架

本文提出的土壤环境监测的流程挖掘构架是一个分步挖掘和总体挖掘的综合过程。先进行监测单元挖掘、点位挖掘和监测数据挖掘，然后建立综合挖掘算法，通过反馈信号相互修正各阶段的挖掘结果，得到最优挖掘结果。在挖掘过程中，所有数据均来自上述数据仓库形成的数据集市。为了挖掘方便，本文将样品测试、数据处理/评价、检测报告三者综合考虑，一起并入监测数据挖掘，在后面的论述中可以看见这样处理的优点。由于每个具体挖掘过程可以采取不同的经典挖掘算法来实现，因此本文不就算法问题进行深入探讨，只是给出挖掘模型构建方法，为建立土壤环境监测数据综合挖掘专家系统提供指导。可以选取 LIMS 实验室信息管理系统作为挖掘平台， LIMS 已经基本实现了监测活动的数据流管理，同时建立了基本的数据库结构体系，便于数据仓库的建立和流程挖掘模型的技术实现。同时，本文建立的土壤环境监测综合挖掘模型构架，同样可以扩展 LIMS 系统，可以将用于普通监测流程数据管理的实验室信息管理系统拓展为具有人工智能的专家决策系统。整个挖掘模型可以用图 3 表示。

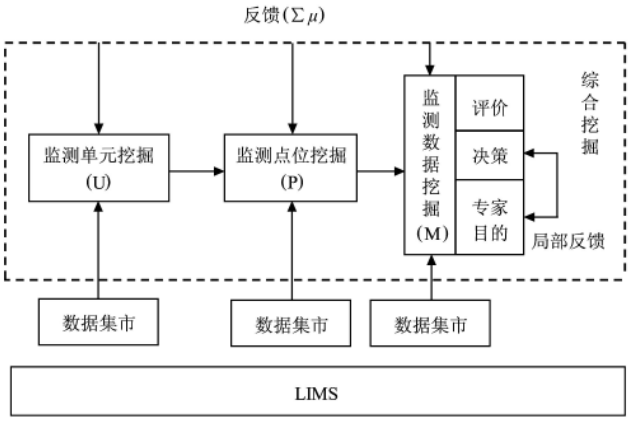

图 3　流程挖掘模型构架图

Fig.3　Framework of process mining model

用数学模型可以表示为：

$$I_{opt}=f(U, P, M, \Sigma\mu)$$

其中，U 是监测单元挖掘；P 是点位挖掘；M 是监测结果挖掘；$\Sigma\mu$ 是各挖掘过程的反馈信号合集，$\sum\mu=\mu_0\cup\mu_U\cup\mu_P\cup\mu_M$，$\mu_0$ 是自身反馈，μ_U 是监测单元挖掘反馈，μ_P 是点位挖掘反馈，μ_M 是监测数据挖掘反馈；f 是综合挖掘算法。具体含义是：监测单元挖掘 U、点位挖掘 P 和监测结果挖掘 M 分别取得最佳监测单元划分结果、最优点位和获得最具代表性监测结果,然后以反馈信号对三者进行修正优化，进行综合挖掘之后取得线性平衡，得到全局最优 I_{opt}。反馈信号可以来自 U、P、M 等分步挖掘过程，也可由 I_{opt} 自身反馈生成，相比而言，由 I_{opt} 生成的自身反馈更加重要，可以促使各分步挖掘过程尽快趋于平衡优化。很明显，从土壤环境监测目的出发，生成 I_{opt} 是为了促使 M 生成最佳决策，因此，M 是整个挖掘活动的重点。

下面就各阶段的挖掘模型 U、P、M 等给出构建方法。

2.2.1　监测单元挖掘

所谓监测单元挖掘，也就是从大面积耕地中，把土壤性能相似、污染程度相近、种植制度基本相同的土地进行归类，作为一个监测单元，这样可以减少监测工作量，提高监测结果的代表性。传统的监测单元划分方法只是依靠经验和现场考察结果进行划分，往往造成监测单元划分错误，监测结果产生偏差。本系统提出以监测区域遥感影象、现场调查数据、历史监测数据为基础进行相似度聚类，通过后续挖掘过程反馈信号修正之后自动识别特征值相近的空间为监测单元，具体模型如下：

$$U=f(\Sigma(r, g, m), \mu_P, \mu_0)$$

其中：r 为监测区域的遥感光谱信息数据；g 为监测区域的现场调查数据向量，g={农业生产土地利用状况∪区域土壤地力状况∪土壤环境污染状况∪土壤生态环境状况∪土壤环境背景资料∪植被分区∪土地利用规划}；m 为历史监测数据向量。显而易见，所有挖掘数据输入均能在数据仓库的维表中直接抽取生成。μ_P 为点位反馈信号，由点位挖掘过程触发一个行为（Action），依据挖掘结果生成一个反馈，参入当前挖掘的修正；μ_0 是综合挖掘算法生成的反馈，对 U 在初始空间内进行修正；$f(\ )$

为挖掘算法，通常采用自组织映射 SOM（Self-Orgernization Map）对 r、g、m 三层数据叠加后进行相似度识别，然后通过反馈信号对模型进行修正，包括修正算法的偏置量、神经元数量等等，以求得到最优的监测单元划分。具体挖掘流程如图 4。

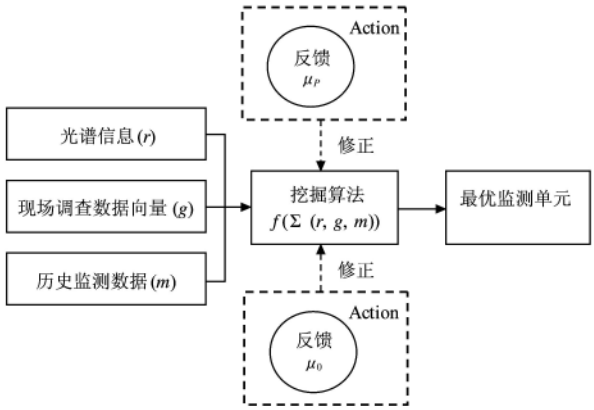

图 4 监测单元挖掘简图

Fig.4 Diagram of monitoring cell mining

2.2.2 点位挖掘

土壤环境监测过程中，点位的选择必须考虑是否能够代表监测单元的污染性状，是否能代表单元内农作物的污染和生长性状等。本文提出的点位挖掘模型如下：

$$P=f(analysis(k, m), v, \mu_M, \mu_0)$$

其中：k 为地理数据向量矩阵；m 如前所述，是历史监测数据向量；v 为作物样品特性向量，具体包括作物长势和污染状况等，从作物样品维表中抽取；μ_M 为反馈信号，由监测数据挖掘过程触发一个行为，依据挖掘结果生成一个反馈，参入当前挖掘的修正；μ_0 如前所述；$f(\ )$ 为挖掘算法，通常可以利用神经网络算法实现。

该算法先由地理数据向量和土壤历史监测数据分析出土壤污染分布的各向异性，然后叠加作物特性向量作为挖掘算法的输入，可以得出土壤污染和作物污染之间的关系，将作物生长受土壤污染影响最严重或最敏感的区域作为监测样点输出，得到最优点位。在求解最优点位之前，可以先从数据集市中选取历史点位数据作为训练集，训练生成最优点位挖掘算法模型，然后再进行点位预测挖掘。具体挖掘流程如图 5。

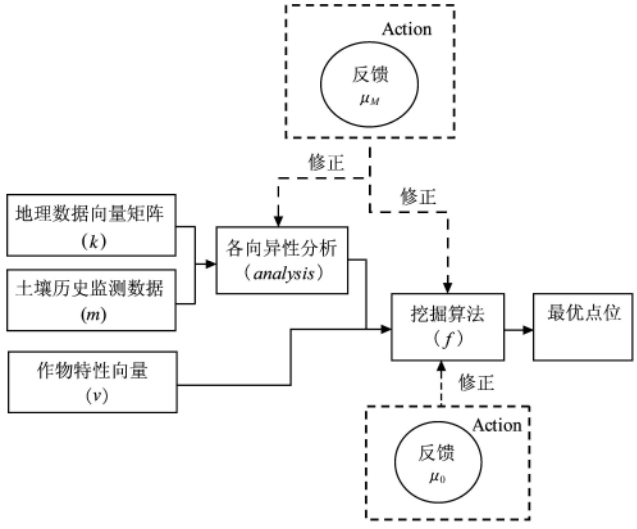

图 5 点位挖掘简图

Fig.5 Diagram of point mining

2.2.3 监测数据挖掘

在综合挖掘模型构架中，监测数据挖掘是最关键的一步，目前大量土壤监测数据分析与评价工作基本都处于这个阶段。由于监测数据挖掘受检测分析处理前、中、后多要素影响，因此，本文将样品测试、数据处理/评价、检测报告等过程综合考虑，统一抽取参数，提出挖掘模型如下：

$$M=f(s(o, i, d), b, v, \mu_0, \mu_1, a)$$

其中：o 为样品检测前期数据向量，例如样品数量、种类、温度、湿度等影响结果分析的状态数据；i 为检测中间态数据向量，包括样品在检测过程中的质量保证和质量控制数据，虽然不参加结果分析，但影响到数据分析的趋势和精度；d 为检测结果向量；s 为数据整理算法，包括数据归一化、标准化等过程，目的是消除脏数据和消除数据各异性干扰；b 为土壤背景值向量，可以从现场调查数据维表中抽取；v 为作物特性（长势、污染状况）向量，可以从作物样品维表中抽取；μ_0 如前所述；μ_1 为局部反馈，由自身挖掘算法触发生成，可以修正专家目，并受专家规则影响，反过来修正挖掘算法；a 为专家规则，反映检测报告的目的和用途，可以从专家规则维表中抽取；$f(\ )$为挖掘算法。

引进了专家规则作为一个输入变量，是为了控制数据评价结果导向，根据监测结果形成决策。很明显，该模型构架已经不再是简单的数据挖掘过程，而是一个决策系统，$f(\ )$实质上就是一个决策机，可以和专家规则 a 形成局部反馈，从而得到决策知识，一旦挖掘结果不符合监测目的需要，可以及时修改专家规则，或者依据专家规则及时更改挖掘算法，使得挖掘结果能够生成正确的决策。具体挖掘过程如图 6。

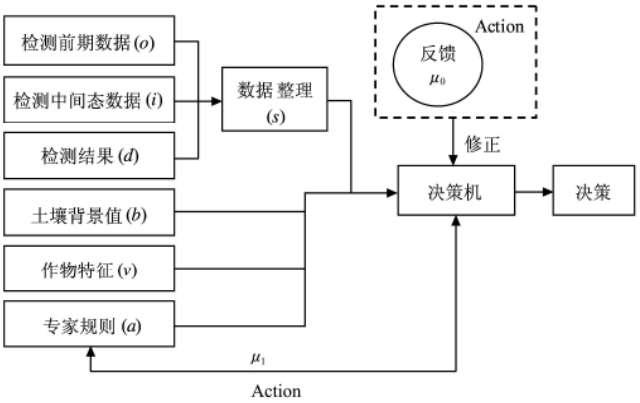

图 6 监测数据挖掘简图

Fig.6 Diagram of monitoring data mining

显而易见，该监测数据挖掘模型与传统的监测数据分析方法有明显的区别。首先，将检测前、中、后的数据状态，土壤背景历史、以及土壤对作物的影响均纳入挖掘范围，这样可以全面考虑监测活动对监测结果的影

响，避免了以往只考虑最终检测结果的片面性。例如，如果两个土壤样品的温度、湿度、生物活性等原始状态不一样，在烘干之后上机分析得到的数据即使一样，也不能说明这两个土壤样品对作物的污染性状一样。其次，将专家规则作为参变量引入挖掘算法，使数据挖掘（或者传统意义上的监测评价）过程变成一个决策系统，从而增加了挖掘模型的使用效率，扩展了模型的适用范围。但是，该模型的缺点也恰恰在此，由于考虑因素较多，造成挖掘难度大，到目前为止尚未完全实例化。通常应用的模型基本上都是忽略样品处理前、中的 o、m 变量，主要考虑样品检测结果 d;回避专家目的 a 和局部反馈 μ_1，仅以得到监测知识为目的。于是得到下述三种形式的模型构架简化。

简化一：仅以样品检测结果 d 为挖掘对象，得到：

$$M = f(s(d))$$

很明显，这就是目前通用的依据《土壤环境质量标准》的评价方法[18]；

简化二：不但考虑土壤样品检测结果 d，还考虑土壤背景历史影响 b，得到：

$$M = f(s(d), b)$$

这就是目前应用较多的土壤背景值评价方法[18]，环境监测学术界认为该方法优于《土壤环境质量标准》评价方法；

简化三：把土壤样品检测结果 d 和作物特性向量 v 一并考虑，得到：

$$M = f(s(d), v)$$

这就是目前的研究热点——作物有效态评价方法[19]，该简化模型基于“土壤——农作物”环境质量进行挖掘，能够较好的反映土壤污染对农作物的影响，是一种公认的比较科学、具有较强应用价值的土壤污染评价方法。

从上述三种简化模型可以看出，随着挖掘参数的增多，挖掘结果会更精确，更能反映监测结果的真实性。

3 结　论

土壤环境监测数据综合挖掘模型是将数据仓库和数据挖掘技术应用于土壤环境监测优化管理流程而建立起来的定量化模型，具有适用性广、指导性强等特点。但就整个土壤环境监测领域来说，数据挖掘建模方法的研究和应用起步较晚，目前仅仅只限于某个挖掘算法在监测过程某个环节的具体应用，而如何对整个监测流程进行整体优化，实现环境监测综合挖掘的研究尚未见报道，更没形成一套系统的、成体系的研究与开发方法。本研究针对土壤环境监测数据挖掘模型开发框架的不足，将数据仓库和数据挖掘技术应用于土壤环境监测的全过程。首先，采用雪花结构建立了数据模型构架，为流程挖掘提供数据底层；同时，还建立了监测流程综合挖掘模型构架，给出了监测单元挖掘、监测点位挖掘和监测数据挖掘，以及综合挖掘的模型函数和参数，以局部优化和全局优化的思路实现了土壤环境监测全流程优化。

基于数据仓库建立的土壤环境监测综合挖掘构架是国内首次提出的研究成果，具有较强的开创性和探索性。和传统数据库相比，雪花结构的数据仓库模型较好的解决了多源数据的集成使用问题，使得数据调用和驱动过程标准化和规范化，极大提高了数据使用效率；挖掘模型构架提高了全流程挖掘的容错能力和置信度，和传统挖掘方式相比，具有较高的挖掘效率。研究结果对于提高土壤环境监测效率，拓展 LIMS 实验室信息管理系统功能，构建土壤环境监测数据仓库和专家系统提供了指导，将极大地拓展土壤环境监测数据的外延和以宏观决策服务为目标的数据分析方法，对有效发挥监测数据的作用，为中国社会经济的持续发展提供基础支撑具有重要意义。

随着数据挖掘技术的不断发展和对土壤环境监测建模的更深入研究，今后的工作应继续在该挖掘模型的基础上，结合 LIMS 实验室信息管理系统，就如何构建土壤环境监测计算机综合评价和挖掘专家系统开展进一步研究。由于综合挖掘模型的定义域空间过大，数据观察集的约简和模型求精将是后续研究的技术难点，有必要进行深入研究和探讨。

［参　考　文　献］

[1] 杨　光．数据仓库及联机分析处理技术[J]．计算机工程与科学，2000，2(1)：76－80．

[2] Han Jiawei, Kamber. Data Mining: Concepts and Technique[M]. Morgan Kaufmann，2000：23－56，77－106．

[3] Yost M. Data warehousing and decision support at the national agriculture statistics service[J]. Social Science Computer Renew，2000，18(4)：434－441．

[4] Miller H J, Han Jiawei. Geographic Data Mining and Knowledge Discovery[M]. Taylor and Francis，2001：74－109．

[5] 郑向群．农业环境信息数据分析中数据挖掘技术的应用[J]．农业环境与发展，2003，1：35－37．

[6] 檀满枝，陈　杰，郑海龙，等．模糊 C 一均值聚类法在土壤重金属污染空间预测中的应用[J]．环境科学学报，2006，26(12)：2086－2092．

[7] 李家科，李怀恩，赵　静．支持向量机在非点源污染负荷预测中的应用[J]．西安建筑科技大学学报，2006，38(6)：756－760．

[8] 钟晓兰，周生路，李江涛，等．长江三角洲地区土壤重金属污染的空间变异特征——以江苏省太仓市为例[J]．土壤学报，2007，44(1)：33－40．

[9] 胡远安，程声通，贾海峰．非点源污染模拟的空间分割优化[J]．清华大学学报：自然科学版，2005，45(3)：367－370．

[10] HJ/T 166 -2004, 土壤环境监测技术规范[S] ．

[11] Mcdowall R D. Laboratory Information Management System, Concepts, Integration and Implementation. W ilmslow, Cheshire, Engiand: Sigma Press l988．

[12] Chen F, Drezner Z, Ryan J K, et a1. Quantifying the bullwhip effect in a simple supply chain: the impact of forecasting, leadtimes and information[J]. Management Science，2000，46(3)：436－443．

[13] Inmon W. Building the Data Warehouse. 2nd Edition, John[M]. Wiley & Sons Inc. 1993．

[14] 邹逸江．空间数据仓库的概略设计[J]．测绘科学，2002，27(3)：13－15．

[15] Sid Adelman, Larissa Terpeluk Moss. 数据仓库项目管理[M]. 北京：清华大学出版社，2003.
[16] 曾叶虹，奚建清. 如何有效使用雪花模型[J]. 广东工业大学学报，2002，19(2)：24－27.
[17] 周志艳，罗锡文. 农作物生产管理信息数据仓库维度建模初探[J]. 农业工程学报，2005，21(11)：112－115.
[18] NY/T 395-2000, 农田土壤环境质量监测技术规范[S].
[19] 刘凤枝，师荣光，徐亚平，等. 农产品产地土壤环境质量适宜性评价研究[J]. 农业环境科学学报，2007，26(1)：6－14.

Integrated mining model framework for soil environmental monitoring data based on data warehouse

Zheng Xiangqun[1,2], Zhao Zheng[2※], Liu Dongsheng[3]

(1. *Department of Computer Science and Technique, Tianjin University, Tianjin* 300072, *China*;
2. *Agro-Environmental Protection Institute, Ministry of agriculture, Tianjin* 300191, *China*;
3. *Chinese Academy of Agricultural Engineering, Beijing* 100026, *China*)

Abstract: In order to optimize the soil environmental monitoring process, an integrated mining model framework was presented on the basis of data warehouse and workflow mining technologies for the first time. This integrated mining model is made up of a data model and an integrated mining model. The data model is the data source for the integrated mining which is built by using Snow Flake Schema. The integrated mining model consists of local mining and global mining. That is to say, firstly, monitoring cell mining, point mining and monitoring data mining are fulfilled respectively, then global mining model is run to opimize the monitoring process. The paper brings forward how to build up these models and the framework, and what variables should be delivered. This study provides a guideline for enhancing work efficiency of soil environmental monitoring, for expanding the functions of Lab Information System(LIMS), and for building the soil environmental monitoring expert system in the future.

Key words: soil environmental monitoring; data mining; data warehouse; model framework