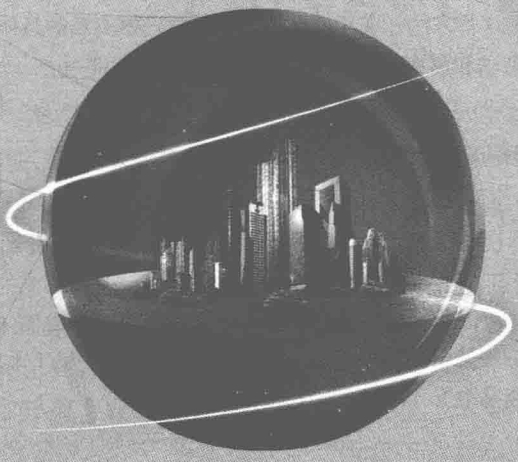
■环境科学新技术应用丛书■

环境保护档案数据挖掘理论与实践

潘鹏 诸云强◎著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

■环境科学新技术应用丛书■



环境保护档案数据挖掘理论与实践

潘鹏 诸云强◎著

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

内 容 简 介

在介绍相关研究背景与意义、国内外研究现状的基础上，本书阐述基于本体的环境保护档案数据挖掘的理论方法，包括环境保护本体原型设计及其构建方法、基于本体的环境保护档案文本信息抽取方法及表格信息抽取方法、基于本体的环境保护档案多层空间关联规则挖掘方法等，并说明依据这些方法开展环境保护档案数据挖掘实践的具体情况，最后对环境保护档案数据挖掘作出总结和展望。

图书在版编目（CIP）数据

环境保护档案数据挖掘理论与实践 / 潘鹏，诸云强著. —北京：电子工业出版社，2016.8
（环境科学新技术应用丛书）

ISBN 978-7-121-28761-9

Ⅰ. ①环… Ⅱ. ①潘… ②诸… Ⅲ. ①环境保护—档案资料—数据处理—研究 Ⅳ. ①X

中国版本图书馆 CIP 数据核字（2016）第 096437 号

策划编辑：董亚峰
责任编辑：郝黎明
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：720×1 000　1/16　印张：12.75　字数：285.6 千字
版　　次：2016 年 8 月第 1 版
印　　次：2016 年 8 月第 1 次印刷
定　　价：48.00 元

PREFACE 前言

在环境污染和生态破坏形势日益严峻的情况下，开展环境保护活动显得尤为重要，而环境保护活动无疑需要大量环境保护相关信息作为支撑，环境保护档案作为环境保护部门在环境保护活动中直接形成的有价值的各种形式的历史记录，其中含有大量能够为环境管理和环境保护工作提供强有力支撑的有用信息，充分利用这些信息将会对环境保护工作的开展起到无可估量的促进和支撑作用。

环境保护工作开展多年以来，虽已经积累产生了大量的环境保护档案数据，但目前对这些数据的利用还停留在以全文检索和查阅浏览为主的初级阶段，环境保护档案数据资源还没有得到深度的开发，在环境保护档案中真正有用的关键数据没有被单独抽取出来集中管理，隐含在关键数据背后的深层次有用信息还没有被挖掘出来加以利用，环境保护档案在环境保护活动中应有的价值没能得到充分的发挥。然而，要抽取环境保护档案中的关键数据并挖掘关键数据中隐含的信息，需要相应的数据挖掘理论技术方法作为指导。

数据挖掘于 20 世纪被提出，用来发现隐藏在大量数据中的有用知识，以解决“数据丰富，信息贫乏”的问题。传统的数据挖掘主要用于发现大型事务数据库中人们事先不知道的、潜在有用的知识，但随着研究的深入，目前数据挖掘的处理对象已从事务数据拓展到文本数据、空间数据、多媒体数据等其他类型的数据，各类数据挖掘已经得到研究者们的广泛关注并在资源环境、信息安全、图书情报和银行证券等重要领域中得到了成功的应用。环境保护档案主要以文本档案为主，可以借鉴现有的文本数据挖掘方法开展关键数据的挖掘工作，但即便如此，文本数据的高度非结构性和一般文本信息挖掘方法的领域依赖性以及环境保护信息本身的高度复杂性，使得常规的文本数据挖掘方法在处理环境保护文本档案数据时的适用性不强，换言之，环境保护档案数据挖掘目前缺少具有针对性的数据挖掘理论技术方法。

本体又称为本体论，在西方哲学中也被称为存在论，是指研究世界本原的学说，20 世纪末，哲学本体论被人工智能和信息科学等引入到各自领域的研究中，用来作为抽象概念以及概念之间关系的规范性描述。本体作为一种能够提

供领域共同知识的交流工具，目前已经在地理科学、农业科学、生物医药、电子商务等科学和领域中得到了广泛的研究与应用。从本质上来讲，本体是一种描述概念世界的模型，是一种实现知识表达与组织管理的有效工具，本体的这一本质特征也使得其非常适合于表达复杂的环境保护知识。

在以上背景下，本书结合本体的相关理论和技术，介绍环境保护档案数据挖掘的理论方法，并以火电建设项目环境影响报告的数据挖掘为例开展相关实践，可以促进环保档案的开发利用及相关领域的教学与研究工作。

首先，从理论角度来讲，提出环境保护本体，可以为后续环境保护领域的知识表示与组织、信息集成与共享利用等研究工作提供参考；介绍基于本体的环境保护档案数据挖掘方法，可以丰富和完善数据挖掘理论体系，扩大数据挖掘理论的应用范围。其次，从技术角度来讲，利用现有的本体和数据挖掘开发工具，构建环境保护档案数据挖掘原型系统并开展挖掘实践，可以解决环境保护档案数据挖掘中的关键技术问题，能够为今后开展环境保护档案数据挖掘提供技术支撑。最后，从应用角度来讲，介绍环境保护档案数据挖掘的理论方法并开展相应实践，对于促进环境保护档案中关键信息的快速获取，提升环境保护档案在环境保护中的信息支撑作用，进而更好地保护好人类共同生存和发展的环境具有重要现实意义。

本书内容组织如下。

第 1 章：绪论。概述环境保护档案的内容、种类和特点，介绍环境保护档案的重要作用及其数据挖掘的迫切需求，总结本体和数据挖掘的研究现状，提出环境保护档案数据挖掘的研究内容。

第 2 章：环境保护本体。阐述环境保护本体的定义和结构框架，指出环境保护本体构建的方法，包括应遵循的原则和步骤，应采用的本体语言和构建工具等。

第 3 章：环境保护档案文本信息抽取方法。总结文本信息抽取中的几个关键问题，归纳文本信息抽取的三类主要方法，指出其各自的优势与问题，重点介绍一种基于本体和隐马尔可夫模型的自由文本信息抽取方法和一种基于本体和语义相似度的表格信息抽取方法，并以火电建设项目环境影响报告为例，分别说明这两种方法的具体应用情景。

第 4 章：环境保护档案空间关联规则挖掘方法。介绍空间关联规则挖掘的基本理论，分析环境保护档案中的空间信息并指出其空间维和属性维的概念层次关系，阐述基于本体的环境保护档案多层空间关联规则挖掘方法，详细描述方法的思路和具体实现步骤，最后以火电建设项目环境影响报告为例，介绍该方法的具体应用情景。

第 5 章：环境保护档案数据挖掘实践。以火电建设项目环境影响报告为例，

设计构建环境保护本体并加以构建实现，设计环境保护档案数据挖掘系统的总体架构、功能体系、开发环境，解决系统实现涉及的键技术问题，编码实现系统原型，并开展环境保护档案数据挖掘实践。

第 6 章：环境保护档案数据挖掘总结与展望。总结环境保护档案数据挖掘理论研究与实践的主要成果和创新点，并做研究展望。

本书的编写和出版得到了以下项目的资助，它们是环保公益性行业科研专项项目（200909110）、国家自然科学基金项目（41371381）、科技基础性工作专项项目（2013FY110900）、国家地球系统科学数据共享平台（2005DKA32300）和江苏省地理信息资源开发与利用协同创新中心建设项目，在此特别表示感谢！

作　者

CONTENTS 目录

第 1 章 绪论……1

1.1 环境保护档案概述……1

1.1.1 环境保护档案的内容……1

1.1.2 环境保护档案的种类……3

1.1.3 环境保护档案的特点……4

1.2 环保档案的重要作用及其数据挖掘的需求……5

1.2.1 环境保护档案的重要作用……5

1.2.2 环境保护档案数据挖掘的迫切需求……7

1.3 本体、文本信息抽取及空间数据挖掘研究现状……9

1.3.1 本体研究现状……9

1.3.2 文本信息抽取研究现状……13

1.3.3 空间数据挖掘研究现状……15

1.4 数据挖掘与相近领域的关系……18

1.5 环境保护档案数据挖掘研究的主要内容……20

1.5.1 环境保护本体研究……20

1.5.2 环境保护档案的信息抽取方法研究……21

1.5.3 环境保护信息的空间数据挖掘方法探讨……22

1.5.4 环境保护档案数据挖掘原型系统构建与应用实践……23

第 2 章 环境保护本体……24

2.1 环境保护本体的定义……24

2.1.1 本体的定义与分类……25

2.1.2 环境保护本体的定义……30
2.2 环境保护本体的结构……32
2.2.1 环境保护本体的逻辑构成……32
2.2.2 环境保护本体的概念框架……35
2.3 环境保护本体的构建方法……39
2.3.1 环境保护本体的构建原则……39
2.3.2 环境保护本体的构建过程……41
2.3.3 环境保护本体的描述语言……46
2.3.4 环境保护本体的构建工具……50

第 3 章 环境保护档案文本信息抽取方法……53

3.1 信息抽取的关键问题……53
3.1.1 信息抽取的主要任务……53
3.1.2 文本的表示模型……56
3.1.3 语义单元的粒度……58
3.1.4 中文文本的自动分词……59
3.2 信息抽取方法分析……61
3.2.1 基于自然语言处理的信息抽取方法……61
3.2.2 基于规则方式的信息抽取方法……62
3.2.3 基于统计学习的信息抽取方法……63
3.3 基于本体和隐马尔可夫模型的自由文本信息抽取方法……65
3.3.1 隐马尔可夫模型……65
3.3.2 基于本体和隐马尔可夫模型的自由文本信息抽取思路……68
3.3.3 基于本体和隐马尔可夫模型的自由文本信息抽取实现方法……69
3.3.4 应用情景分析……73
3.4 基于本体和语义相似度的表格信息抽取方法……76
3.4.1 语义相似度及其计算方法……76
3.4.2 基于本体和语义相似度的表格信息抽取思路……80
3.4.3 基于本体和语义相似度的表格信息抽取实现方法……83
3.4.4 应用情景分析……85

第 4 章 环境保护档案空间关联规则挖掘方法……89

4.1 空间关联规则挖掘……90
4.1.1 空间关联规则及其分类……90
4.1.2 空间关联挖掘的过程模型……93
4.1.3 空间关联规则挖掘的算法……94
4.2 环境保护档案的空间信息及其概念层次关系……97
4.2.1 环境保护档案的空间信息及特点……97
4.2.2 环境保护空间信息中的空间关系及其描述模型……99
4.2.3 环境保护档案空间信息的概念层次关系……103
4.3 基于本体的环境保护档案多层空间关联规则挖掘方法……106
4.3.1 基于本体的环保档案多层空间关联规则挖掘思路……106
4.3.2 基于本体的环境保护档案多层空间关联规则挖掘实现步骤……108
4.4 环境保护档案多层空间关联规则挖掘应用情景分析……113
4.4.1 火电厂与配套设施的多层距离关联规则挖掘……113
4.4.2 火电厂与居民点空气污染物浓度的多层方位关联规则挖掘……115

第 5 章 环境保护档案数据挖掘实践……118

5.1 环境保护本体构建……118
5.1.1 火电行业建设项目环境影响评价本体设计……119
5.1.2 火电行业建设项目环境影响评价本体的实现……121
5.2 火电建设项目环境影响报告数据挖掘原型系统设计……127
5.2.1 系统总体架构……127
5.2.2 系统功能体系……129
5.2.3 系统开发环境……131
5.3 火电建设项目环境影响报告数据挖掘原型系统关键技术实现……132
5.3.1 基于 VSTO 的文本档案数据处理技术实现……132
5.3.2 基于本体和 ICTCLAS 的中文文本分词技术实现……140
5.3.3 基于 Jena 的本体解析与推理技术实现……145
5.4 环境保护档案数据挖掘及效果分析……151
5.4.1 环境保护档案数据挖掘的数据范围……151

5.4.2 环境保护档案数据挖掘的结果展示……152
5.4.3 环境保护档案数据挖掘的效果分析……154

第 6 章 环境保护档案数据挖掘总结与展望……156

6.1 环境保护档案数据挖掘总结……156
6.2 环境保护档案数据挖掘展望……159

附录 火电行业建设项目环境影响评价本体核心概念 OWL 描述……161

参考文献……178

第1章

绪　论

环境保护档案中含有大量能够为环境管理和环境保护工作提供强有力支撑的有用信息，充分利用这些信息将会对环境保护工作的开展起到无可估量的促进和支撑作用，但获取这些信息需要借助具有针对性的数据挖掘技术方法。本章概述环境保护档案的内容、种类和特点，阐明环境保护档案的重要作用及其对数据挖掘的迫切需求，总结本体和数据挖掘的研究现状，提出环境保护档案数据挖掘的主要研究内容。

1.1　环境保护档案概述

1.1.1　环境保护档案的内容

环境信息是环境管理、环境科学、环境技术、环境保护产业等与环境保护相关的数据、指令和信息等，以及其相关动态变化信息〔《环境信息分类与代码》，HJ/T 416—2007〕。由于环境保护档案是

环境信息记录的真实载体，因此，环境保护档案中包含了大量原始的、规范化的且具有重要价值的环境信息和其他信息。同时，为了能够方便管理和利用，环境保护档案中除包含环境信息和其他信息外，还包含环境保护档案的说明性信息，即环境保护档案的元数据信息。因此，环境保护档案信息由两大部分信息构成：一部分是环境保护档案实体信息，另一部分是环境保护档案元数据信息，具体内容如表 1-1 所示。

表 1-1　环境保护档案信息分类与内容

环境保护档案信息构成	环境保护档案信息分类与内容	
环境保护档案实体信息	环境质量信息	环境功能区划、环境质量数据、环境质量报告等
	生态环境信息	自然生态信息、农村生态信息、生物多样性信息，生物安全信息等
	污染源信息	工业、农业、生活、交通运输、施工工地污染源信息，集中式污染治理设施信息，污染物信息等
	环境保护业务信息	环境保护管理制度，污染防治信息，生态环境保护与修复信息，环境监测信息，环境监察信息，环境专业人才管理认证信息，公众参与信息，环境保护宣传、培训信息等
	环境保护科技及其管理信息	环境保护科技信息，环境保护科技管理信息，环境保护认证管理信息等
	环境保护产业信息	环境保护产品信息、环境保护产业项目，环境保护产业组织，环境保护工程设计，清洁生成，循环经济等信息
	环境保护管理信息	环境保护机构、人事信息，日常政务信息，资产管理信息，文档管理信息，会议管理信息等
	环境保护政策法规标准信息	环境保护政策、法规和标准信息
	其他环境保护信息	自然环境信息、社会经济信息等

续表

环境保护档案信息构成	环境保护档案信息分类与内容	
环境保护档案元数据信息	文件实体信息	文件题名、文件分类、文件主题、文件日期、文件语种、文件种类等信息
	责任者实体信息	责任者层级、责任者标识、责任者描述、责任者权限、责任者行为历史等信息
	业务实体信息	业务层级、业务标识、业务法规依据、业务描述、业务权限、业务处理过程等信息
	关系实体信息	关系实体标识、关系实体类型、相关实体标识、相关实体类型、关系定义、关系时间等信息
	长期保存实体信息	签名信息、锁定签名信息、编码等信息

1.1.2　环境保护档案的种类

按照不同的划分方法，环境保护档案可以划分为不同的类型。

（1）依据人类环境保护活动的分工和环境保护档案记述的内容，环境保护档案可以划分为环境管理档案、环境监测档案、环境污染及其防治档案、自然保护档案、环境科学研究档案、环境工程与基本建设档案、设备与仪器档案、环境标准与计量档案和其他各类辅助档案（《中国档案分类法环境保护档案分类表》，1994）。

环境管理档案主要包括：环境保护法规档案，环境保护政策档案，环境保护管理制度档案，环境监理档案，环境行政处罚、复议和诉讼档案，环境规划、计划档案。

环境监测档案主要包括：环境信息管理档案，环境宣传与教育档案，监测质量保证档案，监测数据管理、环境质量报告管理档案，污染物排放状况档案。

环境污染及其防治档案主要包括：监测网络系统管理档案，污染

源调查档案，环境污染及其防治档案，污染事故档案，城市环境综合整治档案，清洁生产档案。

自然保护档案主要包括：自然资源保护档案，农村生态建设档案，乡镇环境保护档案，自然保护区档案，生物多样性保护档案，风景名胜区档案，自然生态区档案。

环境科学研究档案主要包括：自然保护研究档案，环境质量研究档案，环境保护产品开发研究档案，污染防治技术研究档案，环境污染及生态破坏研究档案，全球环境问题研究档案，环境管理研究档案，基础研究档案。

环境工程与基本建设档案包括：环境工程档案，民用建筑档案，环境科研、监测专用设施档案。

设备与仪器档案包括：设备档案，仪器档案，设备、仪器质量监督管理档案。

环境标准与计量档案包括：环境标准，计量。

其他各类辅助档案主要包括：世界各国和地区表，中国地区表等。

（2）依据环境保护档案的承载介质来划分，环境保护档案可以分为纸质档案、录音档案、录像档案、照片档案、实物档案及其他档案等。

1.1.3 环境保护档案的特点

环境保护档案具有来源上的广泛性、使用上的现实性、在产生过程中的原型性、形成中的系统性四大特点（吴文超，2000）。

1. 来源上的广泛性

环境保护工作涉及非常广泛的领域，涉及自然科学和社会科学等

各种学科，因此，环境保护档案的来源也十分广泛，包括工业、农业、交通、能源、生态、城乡建设等各个方面，同时其来源上的广泛性也决定了其内容上具有丰富性。

2. 使用上的现实性

环境保护档案完整、准确和系统地记录了人类在环境管理、污染防治、工程建设和项目建设中的真实活动，因此，环境保护档案在使用上具有现实性，这也决定了其在实际使用中的价值。

3. 在产生过程中的原型性

环境保护档案是人类在进行环境保护工作时直接形成的原始记录，是对人类的生存环境的客观的、直接的描述，不是人为事后编造的，因此，环境保护档案具有产生过程中的原型性。

4. 形成中的系统性

环境保护档案是在实际工作中自然和陆续形成的，记录了人类从事环境活动的各项运动规律，且是按照一定的科学程序积累而形成的，因此具有形成中的系统性。

1.2 环保档案的重要作用及其数据挖掘的需求

1.2.1 环境保护档案的重要作用

目前，我国同时面临着环境污染和生态破坏两个方面的环境问题。

环境污染主要集中在城镇地区，主要表现有工业环境污染、农业化学物质污染等。例如，2010 年“全国工业固体废物产生量月为 240943.5 万吨”（《中国环境状况公报》，2010）；生态破坏主要发生在广大农村地区，主要表现有水土流失、草原退化、耕地减少等，例如，截至 2010 年，我国“现有水土流失面积 356.92 万 km^2，占国土总面积的 37.2%”（《中国环境状况公报》，2010）。

自然因素和人为因素是造成以上环境问题的两大因素。自然因素包括地震和海啸等自然灾害、环境因素引起的地方病等；人为因素包括人类生产导致的各种环境污染和生态破坏、开发利用自然资源造成的生态环境质量恶化、自然资源枯竭等。自然因素为不可抗因素，从近年来自然灾害和流行疾病频发的趋势来分析，在未来自然因素极有可能还将对环境造成更加严重的影响；而因人口不断膨胀和社会经济发展的需要，人为因素在短时间内也不能完全消除，在未来也势必会使得环境问题更加突出。

环境保护档案是环境保护部门在环境保护活动中直接形成的有价值的各种形式的历史记录，它与人类生存环境的关系十分密切，在各种自然因素和人为因素已经并且还将持续给我国带来严重环境问题的形势下，环境保护档案在处理目前已经出现以及今后还将继续出现的环境问题的过程中具有重要作用，这主要体现在决策支持、参考依据和法律凭证三个方面（吴文超，2000），具体如下。

1. 决策支持

系统准确、真实可靠的环境保护档案，能为施行环境影响评价、制定国家环境保护规划、确定环境保护标准和法规以及开展国际环境问题谈判等提供有力支持。

2. 参考依据

环境保护档案中包括了环境保护工作者的丰富知识和经验教训，包含了大量有用的环境信息数据，能够较好地为环境科研和环境治理等工作提供重要参考依据。

3. 法律凭证

环境保护档案客观记录了环境问题参与者在环境污染事故发生中的现状，具有真实性和可靠性并具有法律效率，可以作为环境保护主管部门依法处置环境污染事故和纠纷等问题的重要凭证。

1.2.2 环境保护档案数据挖掘的迫切需求

虽然环境保护档案中包含有大量可以为环境保护决策、环境科研、环境治理等环境保护工作提供重要依据或指导作用的环境现象和环境规律信息，但要从大量的环境保护档案中获取这些信息，还需要借助环境保护档案数据挖掘的手段，而目前环境保护档案的数据挖掘还缺乏相应的理论技术方法。

一方面，随着国家对污染防治和生态保护问题的重视，以及多年来一系列的环境政策、法规的制定与实施，环境保护相关部门积累产生了海量的环境保护档案，其中含有大量的环境保护信息，这些信息能够为环境管理与保护工作提供强有力的支撑。以建设项目环境影响评价档案为例，我国实施环境影响评价制度以来，每年审批的建设项目约 30 多万个，每个建设项目都有一套完整的环境影响评价档案，其中尤其典型的是建设项目环境影响报告。环境影响报告是对建设项目进行环境影响评价后产生的系统性、客观性、规范性成果文件，科学