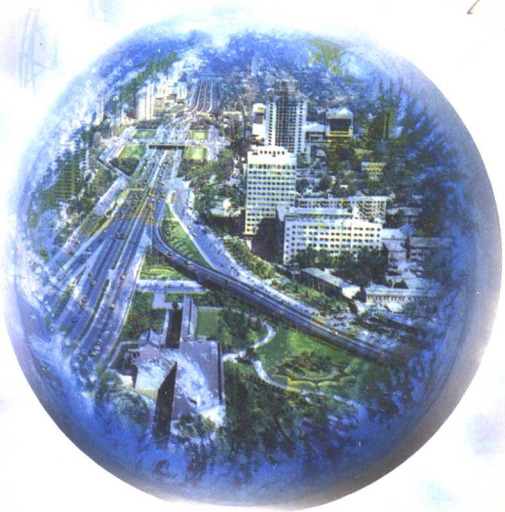
城市空间数据挖掘方法与应用

李新运 著



山东大学出版社

城市空间数据挖掘方法与应用

李新运 著

山东大学出版社

图书在版编目(CIP)数据

城市空间数据挖掘方法与应用/李新运著.
济南:山东大学出版社,2005.9
ISBN 7-5607-3079-5

Ⅰ.城…
Ⅱ.李…
Ⅲ.城市—地理信息系统—研究—山东省
Ⅳ.P208

中国版本图书馆 CIP 数据核字(2005)第 105586 号

山东大学出版社出版
(山东省济南市山大南路 27 号　邮政编码:250100)
山东省新华书店经销
山东旅科印务有限公司印刷
850×1168 毫米　1/32　8.25 印张　209 千字
2005 年 9 月第 1 版　2005 年 9 月第 1 次印刷
印数:1000 册
定价:25.80 元

内容简介

本书以城市发展决策中的空间信息需求为目标，以空间信息科学、城市地理学和区域经济学为指导，在已有空间数据挖掘研究的基础上，对城市空间数据挖掘理论、方法、技术和应用进行了较深入和系统地研究。主要内容包括：城市空间数据挖掘体系、空间基础计算模型、城市空间分布（静态）数据挖掘、城市空间动态预测、城市空间与时序关联规则提取、城市群数据挖掘等方面，提出了一个总体框架体系和若干新的挖掘方法，改进了一些已有的空间数据挖掘方法，进行了大量应用实验研究，初步建立了一个城市空间数据挖掘实验系统。

本书可作为地理、测绘、土地管理、城市规划、区域经济、信息管理等专业高年级大学生和研究生的教学参考书，也可供相关专业领域的科研人员、工程技术人员和管理工作者参考。

序

地理信息系统是对空间数据进行自动化管理、智能化分析与可视化输出的技术系统。目前，对空间数据的管理与可视化技术已取得明显进展，但对空间数据的智能化分析却面临许多困难。这主要是因为空间数据库是对空间实体的抽象和表达，而空间实体具有属性特征、空间分布、动态变化等特点，其数据组织形式要比关系数据库复杂得多。而区域经济、社会、环境的发展又迫切需要各种深层的空间信息，如空间动态变化信息、空间优化配置信息、空间分布差异信息等。在这种背景下，空间数据挖掘(Spatial Data Mining)技术应运而生。空间数据挖掘是提取隐含在空间数据库中的知识的过程，它至少具有四个特点：一是面对的数据是海量的，二是获取的知识是事先未知的，三是获取的知识是有用的，四是挖掘的方法具有智能性和自动性。20 世纪 90 年代后期以来，空间数据挖掘逐渐成为地理信息科学研究的热点领域，众多专家和学者在此不懈耕耘，已出现了一批研究成果。

李新运同志在 2001 年 9 月至 2004 年 6 月攻读博士学位期间，对空间数据挖掘的理论、方法、技术和应用进行了较深入和系统地研究，完成了博士论文《城市空间数据挖掘方法与应用研究》，现在经整理后即将出版，作为导师，我们感到由衷的高兴和欣慰。

本书是一部高水平的理论与实践相结合的著作，其创新性是多方面的：

第一，在基础理论方面：在总结已有研究工作的基础上，提出了一个空间数据挖掘框架体系和城市空间数据挖掘的任务体系；提出了位置—属性一体化的实体信息模型，并给出了三种空间距离测度，可以作为空间计算的基础准则；通过对空间权重矩阵进行拓展，提出了空间实体关联矩阵和空间状态关联矩阵的概念，并给出了建立方法，为空间数据挖掘提供了新的基础工具。

第二，在城市空间分布（静态）数据挖掘方面：采用空间—属性一体化概念模型，把空间坐标、空间关系和属性特征纳入到统一的空间计算模型，分别对城市土地适宜性评价和城市功能区划分中的空间聚类方法进行了研究，并提出了一种分类图层的平滑算法；针对城市土地的空间优化配置，提出了一种空间遗传算法（SGA），该算法中的选择、交叉、变异算子都是在空间上进行的；对多要素离散空间场之间的相关性测度，定义了一种基于信息熵的规范的相关指数，并给出了计算方法。

第三，在城市空间动态挖掘方面：对离散状态属性预测和模拟，建立了一种具有操作性的细胞自动机预测方法，即从历史空间数据中自动提取局部状态转换规则，在预测和模拟计算阶段采用随机试验的方法确定未来时间的单元状态，更符合实际；对连续状态属性预测，提出了一种空间关系与属性特征一体化的空间自回归分析方法，可以用于空间单元网络的连续属性的群体预测；对城市空间扩展预测，根据区域扩散的思想提出了一种点源射线扩散的预测方法和计算模型。

第四，在城市空间关联知识挖掘方面：对静态关联规则，根据粗糙集理论归纳出了基于数据约简和等价类划分的两种空间关联规则提取方法；对动态关联规则，提出了一种时序信息表的生成方法，可广泛用于时序关联规则的挖掘。但是这些规则提取方法与

空间关系计算是分离的，即规则提取方法本身不是空间计算方法。针对较复杂的空间关联知识，研究了根据空间关联矩阵挖掘空间实体关联知识和空间状态关联知识的方法，这些方法是基于空间计算的。

第五，在城市群空间数据挖掘方面：提出了坐标—属性一体化的城市群分布轴线挖掘思路和参数估计方法，包括直线、抛物线和一般二次曲线；提出了几种新的空间离散度指数及计算方法，包括加权平均间距、空间标准差、空间基尼系数等。从理论上证明了普通 Voronoi 图在描述城市群吸引范围中的缺陷，提出了一种更为合理的属性加权的曲边 Voronoi 图模型，并对生成方法进行了初步探讨，但还不成熟；根据城市规模—等级定则（Zipf 定则）、坐标分离策略和遗传算法（GA），分别提出了两种城市群重心移动预测的方法；最后对城市群空间引力场和潜能场的可视化方法进行了研究。

第六，在技术开发方面：通过编制 50 个空间数据预处理和挖掘计算程序，并与 GIS 平台和其他数据分析软件集成，初步建立了城市空间数据挖掘实验系统 USDMS，结合济南市和山东省城市群进行了大量应用实验研究，获取一大批城市空间数据挖掘结果。

当然，本书还存在一些不足，一些空间挖掘算法的研究还有待深入，算法的技术开发与系统集成还有待提高。

空间数据挖掘涉及地理学、测绘学、数学、计算机技术、信息科学、经济学等学科，要求研究者具有多样化的知识结构和无私的奉献精神。希望本书的出版能为地理信息科学的发展出一臂之力，为区域经济社会发展决策提供信息技术支撑。

林宗坚　李成名

2005 年 7 月

目　录

第1章 绪 论

城市空间数据挖掘(urban spatial data mining, USDM)既指空间数据挖掘(spatial data mining, SDM)理论、方法和技术在城市地域的应用,又包括针对城市地域的特点新发展的空间数据挖掘理论、方法和技术。从经济角度看,城市地域是区域经济社会发展的极核和辐射源,经济社会资源高度集中,经济社会活动非常活跃,城市景观变化剧烈;从信息角度看,城市地域内基础地理数据和专题要素资料相对丰富,同时城市发展决策对空间信息定量化、自动化、可视化的要求越来越迫切。所以,城市地理信息系统(urban geographic information system, UGIS)建设和城市空间数据挖掘理论、方法、技术与应用研究是地理信息科学中最为活跃的领域之一。

1.1 空间数据挖掘概述

1.1.1 空间数据挖掘的内涵

基于数据库的知识发现或称从数据库中发现知识(knowledge discovery in database, KDD),是指从大量数据中提取有效

的、新颖的、潜在有用的和可被理解的模式的非平凡过程[1, 37]。其中：

——数据是一个有关事实 F 的集合，用于描述事物的基本信息，一般来说 KDD 面对的数据都是海量的，也是准确无误的；

——模式是语言 L 中的表达式 E，E 所描述的数据是集合 F 中的一个子集 F_E，F_E 表明数据集合 F 中的数据具有特性 E，作为一个模式，E 比枚举数据子集 F_E 简便；

——有效性是指获取的模式必须具有一定的置信度；新颖性是指发现的模式应是事先未知的；

——潜在有用是指获取的模式现在或将来可以被实际应用。可被理解是指模式应简洁明了，便于使用；

——非平凡性是指 KDD 具有智能性和自动性，是传统的数据分析方法和技术的提升和延伸。

一般认为，KDD 是一个过程，它主要由以下步骤组成[1]，如图 1-1 所示。

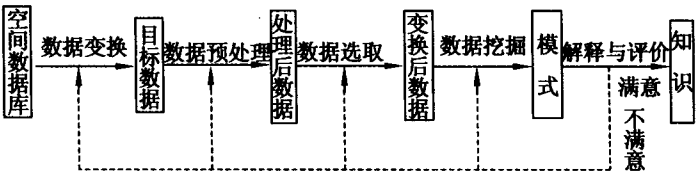


图 1-1　典型的 KDD 流程

——数据选取：确定目标数据，根据挖掘任务从原始数据库中检索相关数据，主要是对原始数据库进行有关操作；

——数据预处理：主要任务是检查数据的完整性和一致性，包括滤除冗余数据，填充缺失数据，消除不一致性，拟制噪声干扰等；

——数据变换：对经过预处理的数据进行再处理，把数据变换成适合挖掘的形式，如简要统计、相对指标计算、投影变换等；

——数据挖掘(data mining, DM):选择智能化方法,从数据中提取模式,并以一定的方式把发现的模式表示出来。DM是KDD的核心步骤;

——模式评估:对提取的数据模式进行解释,根据某种测度度量模式的有趣性,判定真正有趣的模式即知识。

DM是KDD过程中的关键步骤,承担着提取模式和表达模式的任务,包含很多计算方法和复杂的技术。所以学术界有时也把DM与KDD混用。

SDM是DM的一个重要分支,USDM则是SDM的一个主要方向,SDM或USDM面对的都是空间数据库(spatial database, SDB)。由于空间数据库储存了空间实体的位置特征、几何特征和属性特征这3种基础数据,空间实体之间又具有空间拓扑、空间距离、空间方位这3种关系,所以空间数据的分析与计算要比关系数据库(RDB)复杂。这意味着SDM的方法要比普通DM复杂得多,技术的实现也更困难。所以USDM研究既要继承DM和SDM的理论、方法和技术,又要发展针对城市空间信息管理特点的新的方法和技术。

1.1.2 空间数据挖掘的任务

SDM的任务是提取SDB中隐含的信息和知识。Jiawei Han等把数据挖掘功能分为6大类:描述、关联分析、分类和预测、聚类分析、孤立点分析及演变分析[1]。M. H. Dunham则把数据挖掘的任务分为8个方面,其中预测性挖掘(分类、回归分析、时间序列分析、预测)利用从历史资料中得到的结果对数据做出预测,而描述性挖掘(聚类、概括、关联规则、序贯分析)则识别出数据中的模式和关系[144]。事实上,由于空间数据的复杂性和多样性,SDM的任务也是多方面的,我们归纳为以下8类:

(1)空间数据概括。空间数据概括(Summarization)也称为空

间数据特征化(Characterization)或泛化(Generalization),主要是抽取空间数据的总体信息和一般特征,如某地区中小城市的基本特征归纳等。其逆操作是特化(Specialization)。

(2)空间特征参数计算。主要提取单个图层内空间实体的分布特征参数或多个图层之间的相关特征参数。如空间要素或实体的分数维、离散度、自相关系数等的计算,空间要素之间的互相关系数计算等。

(3)空间实体分布计算。对空间实体集合的抽象和归纳。如城市群分布重心、分布轴线的计算。

(4)空间分类与聚类。把空间实体集合划分成若干个组群。空间分类是根据指标数据把空间实体划分到既定的类别中去,而空间聚类则根据类内最相近、类间最相异的原则把空间实体分组。空间分类需要训练样本,空间聚类则是非监督的。

(5)空间回归分析。主要提取空间实体之间或实体属性之间的静态的定量模型。如提取地区经济水平与中心城市科技教育投入之间的定量关系,区域经济水平和人口素质之间的定量关系等。空间回归分析一般需要多个空间样本,回归模型既可以是线性的也可以是非线性的,模型的参数一般采用最小二乘法进行估计。

(6)时间序列关联分析。主要分析多个空间实体之间的时变关联性或单个实体属性之间的时变关联性,可采用统计学中的相关系数、灰色系统理论中的关联度或信息论中的熵进行定量表达。

(7)空间动态预测。空间单元(实体)网络是在相互作用中发展变化的,空间动态模型是对空间过程的动态预测和模拟。包括连续空间场预测(如行政区网络的人口数量增长预测等)、离散空间场预测(如城市土地利用类型变化预测等)和空间实体运动预测(如城市空间扩展预测等)。大家熟知的细胞自动机(cellular automata, CA)就是对离散空间场序列(空间离散、状态离散、时间离散)在全局法则(每一个细胞的状态按同一确定的法则演化)和