计算机科学学术文库·数据库

空间数据挖掘及其相关问题研究

张志兵 著

Research of Spatial Data Mining and Related Issues

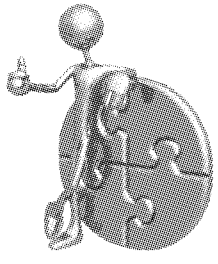
清华大学出版社
http://www.tup.com

计算机科学学术文库·数据库

空间数据挖掘及其相关问题研究

Research of Spatial Data Mining and Related Issues

张志兵 著



华中科技大学出版社
http://www.hustp.com
中国·武汉

内 容 提 要

本书围绕空间数据挖掘的相关技术进行了卓有成效的研究。首先，研究了数据聚类有关问题；接着，提出了一个改进的支持大的数据集和任意形状聚类、且具有良好的抗噪性能和能满足高维数据要求的算法；然后，分析了与空间数据挖掘和分析相关的空间索引及查询技术；最后，设计了一个融合神经网络、模糊集和遗传算法的空间数据挖掘系统。

本书可作为人工智能、模式识别、空间数据库、统计学、空间信息系统和网络等学科相关专业学生的教材及参考资料。

图书在版编目(CIP)数据

空间数据挖掘及其相关问题研究/张志兵　著. —武汉：华中科技大学出版社，2011.10
ISBN 978-7-5609-7081-3

Ⅰ.空…　Ⅱ.张…　Ⅲ.数据收集-计算机应用-地理信息系统　Ⅳ.P208-39

中国版本图书馆 CIP 数据核字(2011)第 091037 号

空间数据挖掘及其相关问题研究　　张志兵　著

责任编辑：王汉江
封面设计：潘　群
责任校对：祝　菲
责任监印：周治超
出版发行：华中科技大学出版社(中国·武汉)
　　武昌喻家山　　邮编：430074　　电话：(027)87557437
录　　排：华中科技大学惠友文印中心
印　　刷：湖北新华印务有限公司
开　　本：710mm×1000mm　1/16
印　　张：7.5
字　　数：146 千字
版　　次：2011 年 10 月第 1 版第 1 次印刷
定　　价：18.00 元

本书若有印装质量问题，请向出版社营销中心调换
全国免费服务热线：400-6679-118　竭诚为您服务

前　言

在当今数字化的时代，数据的积累正在呈爆炸性的增长，例如，商业企业、科研机构或政府部门都积累了海量的数据资料。据估计，人们日常接触的数据80%都与空间有关，人们收集到的空间数据远远超过人脑的分析能力，导致了空间数据这种丰富的数据资源不能很好地为决策者提供便利的服务。面对浩渺无际的空间数据海洋，人们在呼唤一种新的技术，一种能够帮助人们从繁杂的空间数据中去伪存真、去粗存精的技术，于是空间数据挖掘技术应运而生了。空间数据挖掘就是从空间数据库中抽取隐含的、以前未知、潜在有用的知识的过程，这就要求通常的数据挖掘技术和空间数据库的集成。

决策者们已经不满足于明确显现在数据表层的检索、查询，也希望能够抛弃无意义的信息，深入到数据深层，而只对感兴趣的数据进行更高层次的分析和利用。可是，在过量的空间数据面前，空间知识显得相当贫乏。克服空间数据灾难的重要途径之一，是以从空间数据中挖掘得到的知识指导数据利用。因此，“空间数据过量而知识贫乏”已经成为空间信息学的瓶颈。空间数据挖掘是一个很有发展前景的领域，其应用涉及人们生活的各个方面，如广泛应用于GIS、地理经济、远程遥感、基于内容的图像检索、交通控制、城市规划、环境研究。正是由于空间数据的规模巨大、数据类型和存取方法复杂，所以探索高效的空间数据挖掘是一个富有挑战性的难题。

空间数据库挖掘和知识发现是空间数据获取技术、空间数据库技术、计算机技术、网络技术和管理决策支持技术等发展到一定阶段的产物，是多学科相互交融和相互促进的新兴边缘学科，汇集了机器学习、人工智能、模式识别、空间数据库、统计学、空间信息系统和网络等各学科的技术成果。只有研究新的理论、技术和方法，才能从日益丰富的空间数据库中自动或半自动地挖掘隐藏在其中的、事先未知的、可信的、新颖的、有效的、能被人理解的空间知识。不能单纯利用某种方法就企图将数据中隐含的最大价值挖掘出来，需要将多种方法融合，综合研究和应用。

为了有效地从大量的空间数据中挖掘隐藏在其中的知识，本书围绕空间数据挖掘的相关关键技术进行了大量卓有成效的研究。

(1) 详细讨论了聚类算法的相关问题，介绍了多种常用的空间数据聚类算法，分析了这些算法的效率及其优缺点。

(2) 在详细分析DENCLUE算法的基础上，提出了改进的DENCLUE算法，

改进的算法在不损失太多精度的情况下能把某些网格当成一个数据点来计算，并给出了估计误差，从而提高算法的效率。

(3) 仔细研究了 DENCLUE 算法中的参数问题，并给出了基于密度熵的 DENCLUE 算法参数的优化估计，从而得到较好的聚类效果。

(4) 通过对 R-树的特点和基于 R-树的空间查询及空间连接算法分析，提出了改进的空间查询和改进的空间连接算法，改进的算法能有效地减少不必要的空间 overlap 谓词的测试次数。

(5) 通过对现有的 R-树代价模型的分析，给出了数据非均匀分布时的代价模型，从而使得代价模型的精度有所提高。

(6) 在分析 R-树最近邻查询算法的基础上，给出了两个 R-树 k 近邻查询的剪枝规则，进而提出了一个新的 R-树 k 近邻查询算法。

(7) 给出了一个融合神经网络、模糊集和遗传算法的数据挖掘系统 NFGDM，这个系统把多种软计算工具相融合，有效地解决了空间数据挖掘中的分类和规则提取。

张志兵

2010 年 12 月

目　　录

第1章 绪 论

1.1 研究背景、目的和意义

计算机信息处理技术的进步、信息技术的高速发展给人类社会带来了巨大的变化和影响，数据成为最重要的战略资源。由于技术的进步，人们能以更快速、更容易、更廉价的方式获取和储存数据，数据库应用的规模、范围和深度不断扩大，数千万个数据库被用于商业管理、政府办公、科学研究和工程开发等，并且这一势头仍将持续发展下去，使得数据及其信息量呈指数形式增长。根据粗略估计，早在20世纪80年代，全球信息量每隔20个月就要增加一倍。而进入20世纪90年代，全世界所拥有的数据库及其所存储的数据规模增长更快。一个中等规模企业每天要产生100 MB以上来自生产经营等多方面的商业数据。以美国宇航局的数据库为例，每天从卫星下载的数据量为TB量级，而为了研究的需要，这些数据要保存多达7年之久。20世纪90年代互联网的发展与普及，以及随之而来的企业内部网、企业外部网及虚拟私有网的产生和应用，使整个世界互联形成一个小小的地球村，人们可以跨越时空，在网上交换信息和协作工作。这样展现在人们面前的已不是局限于本部门、本单位和本行业的庞大数据库，而是浩瀚无垠的信息海洋。由此便出现了一个新的挑战：在这个信息爆炸的时代，信息过量几乎成为人人需要面对的难题。由于海量数据的复杂性和数据处理的时效性妨碍了人们对数据的使用，人们陷入了“数据丰富，但知识缺乏”的窘境，数据库急剧增长与人们对数据库处理和理解的困难之间形成了强烈的反差。据估计，目前一个大型企业数据库中的数据，只有7%得到了很好的应用。在这些大量数据的背后隐藏了很多具有决策意义的信息，那么如何及时得到这些有用的知识呢？如何才能不被信息的汪洋大海所淹没，提高信息利用率呢？面对这一严峻挑战，数据挖掘和知识发现(data mining and knowledge discovery，DMKD)技术应运而生，并得以蓬勃发展，越来越显示出其强大的生命力。

数据挖掘(data mining)就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的但又是潜在有用或是感兴趣的信息和知识的过程。还有很多和这一术语相近似的术语，如从数据库中发现知识

(knowledge discovery in database, KDD)、数据分析(data analysis)、数据融合(data fusion)及决策支持(decision supporting)等。就像从矿山中采矿一样,人们可以把原始数据看做是形成知识的源泉。原始数据可以是结构化的,如关系数据库中的数据,也可以是半结构化的,如文本、图形、图像数据,甚至是分布在网络上的异构数据。发现知识的方法可以是数学的,也可以是非数学的;可以是演绎的,也可以是归纳的。发现的知识可以用于信息管理、查询优化、决策支持、过程控制等,还可以用于数据自身的维护。因此,数据挖掘是一门很广义的交叉学科,它汇聚了不同领域的研究者,尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员。它是人工智能、机器学习技术发展的结果,其目的是为理解与应用数据库提供自动化、智能化的手段。无论是商业企业、科研机构还是政府部门,在过去若干时间里都积累了海量数据,目前这些数据仍然保持着强烈的增长势头。如此大量的数据向传统的信息处理和知识抽取方法提出了巨大的挑战。数据挖掘或数据库知识发现作为一个新的研究领域和新的技术正方兴未艾,它用于从大型数据库中发现有趣的、隐含的、以前不知道的知识。数据挖掘研究及其推广应用对于解决信息爆炸、辅助人类高层次决策,以及促进信息产业和知识经济的发展具有重大意义。特别要指出的是,数据挖掘技术从一开始就是面向应用的。它不仅是面向特定数据库的简单检索查询调用,而且要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理,以指导实际问题的求解,期望发现事件间的相互关联,甚至利用已有的数据对未来的活动进行预测。同时需要指出的是,这里所说的知识发现,不是要求发现放之四海而皆准的真理,也不是要去发现崭新的自然科学定理和纯数学公式,更不是什么定理证明。所有发现的知识都是相对的,是有特定前提和约束条件的,而且是面向特定领域的,同时还要能够易于被用户理解。由于数据挖掘如此有价值,数据挖掘将变得更加重要,以至于企业将不会丢失与客户之间有关的任何信息。如果你不在这方面做些什么,那么你将失去你的客户。如果数据挖掘能够对改善商务过程起到明显作用,那么它就是一种能够赢得竞争的武器。

1.1.1 研究背景

在当今数字化的时代,数据的积累正在呈爆炸性的增长,例如,商业企业、科研机构或政府部门都积累了海量的数据资料。据估计,人们日常接触的数据80%都与空间有关。空间数据在数量、时效性和复杂性等方面激剧增长,收集到的数据远远超过人脑分析的能力,导致了空间数据这种丰富的数据资源不能很好地为决策者提供便利的服务。面对浩渺无际的数据海洋,人们在呼唤一种新的技术,一种能够帮助人们从繁杂的数据中去伪存真、去粗存精的技术,于是空间数据挖掘技术应

运而生了。决策者们已经不满足于明确显现在数据表层的检索、查询，也希望能够抛弃无意义的信息，深入到数据深层，而只对感兴趣的数据进行更高层次的分析和利用。可是，在过量的空间数据面前，空间知识显得相当贫乏。克服空间数据灾难的重要途径之一，是以从空间数据中挖掘得到的知识指导数据利用。因此，空间数据过量而知识贫乏已经成为空间信息学的瓶颈。

空间数据库挖掘就是抽取隐藏在空间数据库中的知识、空间关系或者其他感兴趣的模式，这就要求通常的数据挖掘技术和空间数据库的集成。空间数据库与常规的关系数据库有许多不同：空间数据库有很强的局部相关性，带有拓扑和距离信息；其数据通常是经过精心组织，采用多维索引作为其存取方法，并且常常要求其有空间推理、计算几何能力及空间知识表达技术的。这就使得空间数据库的挖掘需求不同于常规的关系数据库的挖掘需求，特别是空间自相关性(spatial-auto-correlation)这一概念，即相似的对象趋向于在地理空间上聚集，它是空间数据挖掘所特有的概念，这就使得以对象间相互独立的统计学为基础的、常规的关系数据库挖掘方法不再适用。空间数据挖掘是一个很有发展前景的领域，其应用涉及人们生活的各个方面，如广泛应用于地理信息系统(geographical information system，GIS)、地理经济、远程遥感、基于内容的图像检索、交通控制、城市规划、环境研究。正是由于空间数据的规模巨大、数据类型和存取方法复杂，所以探索高效的空间数据挖掘是一个富有挑战性的难题。

空间数据库挖掘和知识发现是空间数据获取技术、空间数据库技术、计算机技术、网络技术和管理决策支持技术等发展到一定阶段的产物，是多学科相互交融和相互促进的新兴边缘学科，汇集了机器学习、人工智能、模式识别、空间数据库、统计学、空间信息系统和网络等各学科技术的成果。只有研究新的理论、技术和方法，才能从日益丰富的空间数据库中自动或半自动地挖掘隐藏在其中的事先未知的、可信的、新颖的、有效的、能被人理解的空间知识，才能利用所发现的空间知识指导遥感信息解译的自动化和智能化等。不能单纯利用某种方法就企图将数据中隐含的最大价值挖掘出来，需要采取多种方法融合，综合研究和应用。正是在这个大背景下，本书对空间数据挖掘的相关技术进行了大量卓有成效的研究。

大量空间数据从遥感、GIS、多媒体系统、医学和卫星图像等多种应用中收集出来。空间数据的产生速度在加快。空间数据是人们用于认识自然和改造自然的重要数据，如常见的气温数据。由于雷达、红外线、光电、卫星、多光谱扫描仪、数码相机、成像光谱仪、全球定位系统 GPS(global positioning system)、全站仪、天文望远镜、电视摄像仪、电子显微成像仪、CT 成像仪等各种宏观与微观传感器或设备的使用，以及常规的野外测量、人口普查、土地资源调查、地图扫描、统计图表等空间数据获取手段的更新和提高，在计算机、网络、GPS、遥感 RS(remote sensing)和

GIS 等技术在空间数据的应用和发展中，空间数据的数量、大小和复杂性及其传输的速度都在飞快地增长。而且，空间数据的膨胀速度也极大地超出了常规的事务型数据。例如，以高空间、高光谱、高动态为标志的新型卫星传感器不仅波段数量多、光谱分辨率高、数据传输速率高、周期短，而且数据量特别大，一般情况下数据的容量均在 GB 量级以上。空间数据基础设施的建设速度的加快，也积累了大量的电子地图数据库、规划道路网络数据库、工程地质信息数据库、用地现状信息数据库、总体规划信息数据库、控制性详细规划数据库、市政红线数据库、建筑红线与用地红线数据库、地籍数据库及土地利用和基本农田保护规划数据库等空间基础数据。这些空间数据极大地满足了人类研究地球资源和环境的潜在需求，拓宽了可供利用的信息源。可是，过量的空间数据不仅难以消化、难以辨识真假和难以保证安全，而且形式不一，难以统一处理。因此，空间数据的数量激剧增长、时效性日益增强和复杂性迅速增加远远超过人脑的数据分析能力（见图 1.1），导致了空间数据灾难。人们对空间数据的要求越来越高，已经不满足于明确显现在数据表层的检索、查询，也希望能够抛弃无意义的信息，深入到数据深层，而只对感兴趣的数据进行更高层次的分析和利用。可是，在过量的空间数据面前，人们的空间知识显得相当贫乏。克服空间数据灾难的重要途径之一，是以空间数据中挖掘得到的知识指导数据利用。

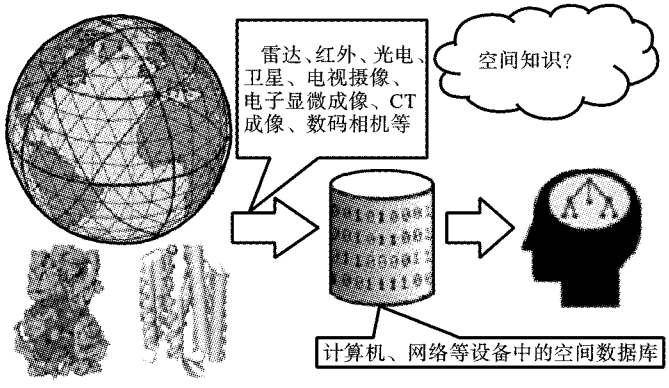


图 1.1　空间数据的生产和传输能力远大于数据分析能力

人类一开始就致力于在数据中利用信息搜寻感兴趣的有用模式，以改善自己的工作和生活。例如，渔民在鱼类的觅食活动中、农民从作物的生长中、政治家从选民的意见中、气象专家在天气的变化中等寻找规律。在研究人类智能的过程中，产生了人工智能，人工智能又分为符号主义、联结主义和行为主义三大学派。以 Robinson 的归结原理为基础的符号主义方法，把符号作为认知基元，通过符号操

作来实现智能行为，其代表是 Lisp 和 Prolog 语言。它们着重问题求解中的启发式搜索和推理过程，在模拟逻辑思维时取得成功，如自动定理证明和专家系统。但是，因为问题求解和逻辑推理的本质仅仅是演绎，所以归结原理不可能成为所有数学分支的证明基础。在信息时代，人工智能也不大可能像物理学那样，建立起一种最基本、简洁、普遍适用的理论或公理系统，用以解决人类智能的五彩缤纷的现实问题。这便形成了对专家系统的挑战。为了利用专家系统完成知识的自动获取，在 20 世纪 60 年代末出现了用计算机模拟人类学习的机器学习（machine learning，ML），借助机器学习，专家系统可以完成知识的自动获取。1980 年，在美国召开了第一届国际机器学习研讨会；1984 年，《机器学习》杂志问世。在研究应用机器学习与数据库技术的过程中，数据库管理系统一般被用来存储数据，而机器学习则用来分析数据。以 Hopfield 为代表的联结主义方法，认为人的思维基元是神经元，把智能理解为相互联结的神经元竞争与协作的结果，其中最典型的是人工神经网络，其中，反向传播网络模型（BP）和 Hopfield 网络模型更为突出。它着重结构模拟，研究神经元特征、神经元网络拓扑、学习规则、网络的非线性动力学性质和自适应的协同行为。可是，能不能从大量的个案中自动形成知识库？在把构造专家系统过程中把知识工程师的事情交给计算机去做？以 Brooks 为代表的行为主义方法，认为反馈是控制论的基石，没有反馈就没有智能，典型成果是机器人和智能控制。它处理不确定性的最基本手段，是根据目标与实际行为之间的彼误差来消除此误差的控制策略。可是，控制论的机器人是“感知-行为”模式，智能行为体现在系统与环境的交互之中，功能、结构和智能行为不可分割。可见，人工智能、机器学习和数据库技术都不能独立地将数据中隐含的最大价值挖掘出来，需要结合三者，综合研究和应用。

1.1.2 研究目的和意义

空间数据挖掘系统的应用相当广泛，涉及人们生活的各个方面，下面我们列举一些空间数据挖掘应用比较多的领域。

（1）气象领域：预测气温气压的异常，天气预报。

（2）城市规划：市政设施的选址。

（3）环境监测：挖掘需要重点监测的地区。

（4）石油、天然气勘探：预测油气储层和油气产能。

（5）经济研究：预测地区经济。

（6）人口种族研究：挖掘人口流动和迁徙。

（7）政府和防卫：评估军事战略，预测资源的消耗。

（8）交通研究：挖掘乘客旅行路线的选择规律等。

(9) GIS 专题挖掘:挖掘不同用户、不同地区的用水规律,从而指导水资源的合理利用,还用于预测森林火灾、地质灾害等。

(10) 电力与能源:电站、电网建设及合理的电力调度。

由于空间数据挖掘有广阔的应用前景,对于国内知名的数据库管理系统——达梦数据库来说,紧跟国际先进技术研制和开发出自主知识产权的空间数据库挖掘系统也就显得十分必要了。据我们所知,目前空间数据库挖掘的研究和应用还处在起步阶段,商用化的产品几乎没有。

本课题正是基于上述原因提出的,其研究的主要目的是跟踪国际先进技术,为拥有完全自主知识产权的达梦数据库管理系统研制通用的空间数据挖掘工具打下基础,因此,本课题的主要意义在于研制出自主知识产权的空间数据挖掘工具,丰富我们国产数据库管理系统的功能。可以预见,随着数据库应用的广泛深入,空间数据挖掘也将会广泛地用于经济建设的各个方面。

1.2 相关研究现状

目前数据挖掘已经成为数据库和信息领域的前沿和研究热点,引起了许多优秀学者和专家的关注,也是商业软件竞争的制高点之一。作为一种新的学术研究领域,与数据挖掘领域研究联系比较密切的领域有数据库、统计分析、机器学习、人工智能、可视化及社会科学,等等。不同领域的学者从自身领域的特点对数据挖掘给出了不同的解释和定义。目前比较公认的一种定义是由 U. M. Fayyad、G. Piatetsky-Shapito 等人提出的[1,2],即数据挖掘就是从数据库中抽取隐含的、以前未知、潜在有用的知识。从数据挖掘的定义可以看出,数据挖掘与数据库中的知识发现(KDD)的含义是相似的。实际上,很多研究者认为数据挖掘和 KDD 在概念上是等价的,只是称呼不一样而已,数据库领域的研究者习惯称之为数据挖掘,而人工智能领域的研究者则称之为知识发现。不过也有些研究者认为数据挖掘是 KDD 过程中的一个主要步骤。随着 DM、KDD 研究逐步走向深入,数据挖掘和知识发现的研究已经形成三大技术支柱:数据库、人工智能和数理统计。因此,KDD 学术会议曾经由这三个学科的权威人物同时出任主席。目前数据挖掘与知识发现 DMKD 学术会议的主要研究内容包括基础理论、发现算法、数据仓库、可视化技术、定性定量互换模型、知识表示方法、发现知识的维护和再利用、半结构化和非结构化数据中的知识发现以及网上数据挖掘等。

1.2.1 数据挖掘的产生与发展

有关KDD的学术会议可以追溯到1989年的人工智能联合会议(IJCAI)的数据库中的知识发现专题(KDD-89)。随着研究者的广泛关注,该专题在1995年发展成国际上著名的学术会议ACM SIGKDD,每年由ACM组织召开,其他比较有名的学术会议有PAKDD、PKDD、DS(Discovery Science)、SIAM-Data Ming(SDM)、SPIE-Data Ming、IEEE-Data Ming(ICDM)、DaWaK,等等。由G. Piatetsky-Shapito维护的网站http://www.kdnuggets.com(Knowledge Discovery Nuggets[3])上提供了很多数据挖掘方面的书籍、软件和学术会议信息,现在该网站已成为数据挖掘研究者交流的一个重要场所。

空间数据库挖掘和知识发现已经渗透到数据挖掘和知识发现、地球空间信息学等相关学科的学术活动中。由美国人工智能协会(AAAI)主办的国际DMKD学术会议,规模由二三十人发展到六七百人,论文比例快速增长;Data Mining、Advanced Spatial Databases、Very Large Databases、International Symposium on Digital Earth、ACM、IFIS和SIGMOD等国际学术会议定期举行。在这些国际学术会议中,空间数据挖掘和知识发现从无到有,已成为关注热点。此外,目前各种规模的国际GIS学术会议及国际摄影测量与遥感学会(ISPRS)都把它作为重要的研究主题,目前,"Data Ming and Knowledge Discovery"学术杂志已经被SCI(Science Citation Index)全部收录,难度指数跃居信息领域的前列,空间数据挖掘和知识发现为该学术期刊的重要研究内容。在IEEE Transactions on Knowledge and Data Engineering、Int. J. of Very Large Databases、Int. J. of Geographical Information Science、Applied Intelligence、Computational Intelligence、Intelligent Data Analysis、Int. J. of Intelligent Information Systems、Journal of Intelligent Systems、Machine Learning、Knowledge and Information Systems、Lecture Notes in Computer Science和Lecture Notes in Artificial Intelligence等国际学术期刊或专著中,也出现了空间数据挖掘的研究成果。Kluwer Publication、Springer-Verlag、Academic Press、WIT Press等著名的国际出版公司也开始出版发行空间数据挖掘的学术期刊、专著或论文集。

目前,国外有许多研究机构、公司和院校从事数据挖掘工具的研究与开发。这些工具主要采用决策树、神经网络、聚类、遗传算法、贝叶斯信任网络、统计分析等方法。许多数据挖掘系统已经成功应用于零售业、银行业、市场营销、电信业、保险业、医疗部门等领域。世界上比较有影响的典型数据挖掘系统有SAS公司的Enterprise Miner、IBM公司的Intelligent Miner、SGI公司的SetMiner、SPSS公司的Clementine、Sybase公司的Warehouse Studio、RuleQuest Research公司的

See5，还有 CoverStory、EXPLORA、Knowledge Discovery Workbench、DBMiner、Geominer、Quest 等。网站 http://www.datamininglab.com 提供了许多数据挖掘系统和工具的性能测试报告。

随着国外知识发现的兴起，我国也很快跟上了国际步伐。1993 年国家自然科学基金首次支持该领域的研究项目。目前，国内的许多科研单位和高等院校竞相开展数据挖掘与知识发现的基础理论及其应用研究。其中，北京大学从事了数据立方体代数的研究，复旦大学、浙江大学、中国科技大学、中科院数学研究所、吉林大学等开展了对关联规则开采算法的优化和改造，南京大学、东南大学、中科院计算所和上海交通大学等探讨了非结构化数据的知识发现及 Web 数据开采，中科院自动化所和中科院软件所开展了多媒体数据挖掘和海量数据挖掘的研究，中科院计算所还深入研究了文本挖掘和检索技术，中国科学技术大学研究了具有时态约束的关联规则的发现及序贯模式的发现，四川联合大学探讨了周期规律的采掘，武汉大学对地球空间数据挖掘做了一些有益的工作，其中华中科技大学数据库研究所对浓缩数据立方及其梯度挖掘进行了深入的研究，等等。对于空间数据挖掘，在国内较少见到这方面的研究报导，国家对空间数据挖掘和知识发现也给予了极大的重视。国家自然科学基金、国家高技术研究发展计划(863)、国家重大基础研究规划(973)和教育部博士点基金，以及军事国防等纵向、横向的科研或应用基金，都相继把空间数据挖掘和知识发现列为资助项目范围，尤其是国家自然科学基金还将其列入信息学部重点项目。

1.2.2 空间数据挖掘的研究内容

空间数据挖掘(spatial data mining，SDM)，或称“从空间数据库中发现知识”(knowledge discovery from spatial databases)，是指从空间数据库中提取用户感兴趣的空间模式与特征、空间与非空间数据的普遍关系及其他一些隐含在数据库中的普遍的数据特征，它是对 KDD 技术在空间数据库方面应用的延伸。

空间数据库是一类重要的、特殊的数据库，地理信息系统(GIS)是空间数据库发展的主体，另外还有图像数据库、CAD 数据库等。GIS 中含有大量的空间和属性数据，有着比一般关系数据库和事务数据库更加丰富和复杂的语义信息，隐藏着丰富的知识。空间数据挖掘和知识发现技术的应用，一方面可使 GIS 查询和分析技术提高到发现知识的新阶段，另一方面从中发现的知识可构成知识库用于建立智能化的 GIS 系统，空间数据挖掘和知识发现技术的引入，将使系统具有自动学习的功能，能使系统自动获取知识，从而使其真正成为智能的空间信息系统。

那么空间数据挖掘系统是怎样从空间数据库中挖掘出有用的知识的？空间数据挖掘是一个多处理阶段，其一般的模型是由 U. M. Fayyad 等人提出的[1,2]，如图

1.2 所示。空间数据挖掘过程包括四个主要的过程：数据预处理、数据挖掘、结果解释和评价、知识表示。整个过程是一个不断循环和反复的过程，因而可以对所挖掘出来的知识不断求精和深化，并且使这些知识易于理解。

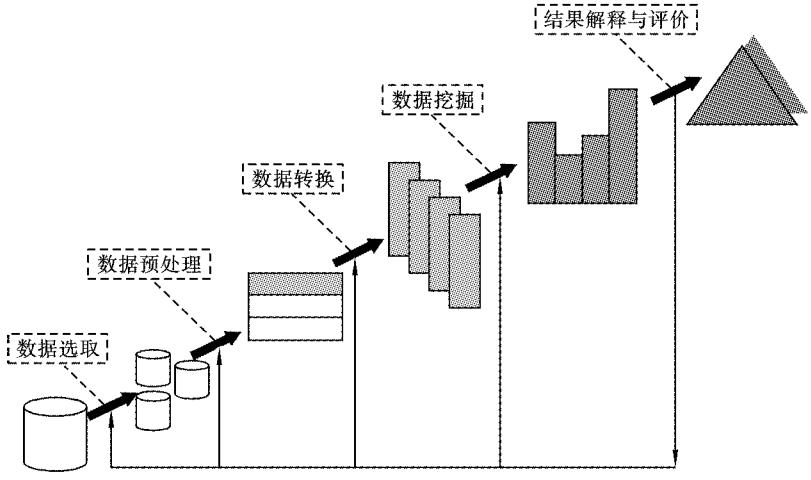


图 1.2 空间数据挖掘过程

1. 数据预处理

具体来说，数据预处理包括以下几个方面。

(1) 数据清理：消除噪声或不一致数据。

(2) 数据集成：对不同数据源中的数据进行合成，包括数据完整性和一致性的检查及噪音数据的过滤和不完整信息的填补，等等。

(3) 数据选择：根据挖掘任务从合成的数据库中选择性地提取与数据挖掘有关的数据。数据选择的目的是缩小处理的范围，提高数据挖掘的质量。

(4) 数据预处理：根据数据挖掘算法的要求对数据选择后再进行一些投影、选择等数据库操作，以便于挖掘算法处理。

2. 数据挖掘

数据挖掘包括以下几个方面。

(1) 确定数据挖掘的目标，确定挖掘的知识类型。

(2) 根据挖掘的知识类型选择合适的挖掘算法。

(3) 运用选定的挖掘算法从数据库中抽取所需的知识，并且用一种知识表示形式(如产生式规则、模式等)来表示所挖掘出来的知识。

3. 结果解释和评价

结果解释和评价包括以下几个方面。