学校编码：10384　　分类号___密级____
学　　号：200431065　　UDC____

厦门大学

硕　士　学　位　论　文

基于 GIS 的空间同位规则挖掘算法的实现及应用研究

Research on Realization and Application of Spatial Co-location Rule Mining Algorithm Based on GIS

张希雯

指导教师姓名：米　　红　　教授
专　业　名　称：模式识别与智能系统
论文提交日期：2007　年　6　月
论文答辩时间：2007　年　7　月
学位授予日期：2007　年　7　月

答辩委员会主席：_____

评　　阅　　人：_____

2007 年 6 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（√ ）

（请在以上相应括号内打“ √”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

摘 要

空间数据挖掘(Spatial Data Mining，简称 SDM)是数据挖掘的一个重要分支，它对于理解空间数据，寻找空间数据之间、空间与非空间数据之间内在关系，以简洁方式表达空间数据规律起着重要作用。空间数据挖掘面向的是空间数据库，空间数据库是一类重要的、特殊的数据库。地理信息系统(Geographic Information System，简称 GIS)是空间数据库的载体，GIS 数据库中含有大量的空间和属性数据。因此，利用 GIS 作为开发空间数据挖掘工具的平台，能够使空间数据的整合利用更加方便以及知识的表达更加直观。

空间关联规则是空间数据挖掘所要发现的一种重要知识。Tobler 的第一地理规则描述了这样一种空间依赖性：“所有的事物都是有联系的，一个地方发生的事件总是与它附近发生的事件有关联，并且相距近的事物之间的联系一般比相距远的事物之间的联系要紧密。”如果能从这些数据中找出其规律性或相互联系，就可以反过来推断客观世界的情况。这就是空间关联规则挖掘的任务。

一般的空间关联规则研究是基于传统的关联规则，然而这些方法在处理空间关系时是不适用的。同位规则问题的提出，很好的解决了挖掘正确有效的空间关联规则的需要。

本论文分为五章。第一章是概述，对空间数据挖掘、GIS 以及空间关联规则等概念和理论框架进行简要的阐述。第二章介绍空间数据的相关理论以及 GIS 与空间数据挖掘的集成模式。第三章是空间同位规则挖掘算法的研究。介绍了用于处理布尔型空间数据的事件中心模型和拓展到处理空间多维分类数据的同位规则挖掘算法，对后者进行了改进，有效的控制了生成的候选同位模式的规模，减少数据库扫描次数，从而提高了算法的效率。第四章是改进算法在 GIS 平台上的实现，并以北京市大兴区 2005 年的经济普查数据为例，分析工业、餐饮业和学校三者间的空间分布规律，得出相关结论。第五章是对本文的工作、创新点以及下一步的研究方向进行总结。

关键词：空间数据挖掘；GIS；空间同位规则

ABSTRACT

Spatial Data Mining (SDM) is a very important branch of Data Mining. It has great efforts on understanding spatial data; find the intrinsic correlations among spatial data, and between spatial data and non-spatial data, and expressing the rules of spatial data concisely, which allow the extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases.

SDM face to spatial databases which are important and special. Geographic Information System (GIS) is the carrier of spatial databases with a mass of spatial data and attribute data. Therefore, using GIS as the framework of SDM tool can make the use of spatial data more conveniently and express knowledge more intuitively.

Tobler's first law of geography describes the spatial correlations as: Everything is related to everything else but nearby things are more related than distant things. The discovery of spatial association rules is a descriptive mining task aiming at the detection of associations between reference objects and some task-relevant objects, the former being the main subject of the description while the latter being spatial objects that are relevant for the task at hand and spatially related to the former. If we can find the rules or mutual associations in these data, we can conclude the external world in reverse. This is the task of Spatial Association Rule Mining.

General studies on spatial association rules are based on conventional association rule algorithms, which treat spatial databases as usual data sets. Co-location rule algorithms meet the demand of mining spatial association rules effectively and exactly.

This paper consists of five chapters. Chapter 1 outlines the basic concepts, theories and applications of SDM, GIS and spatial association rules. Chapter 2 introduces basic concepts on spatial data and integration mode of GIS and SDM. Chapter 3 studies on spatial co-location rule algorithms. First we introduce the Event Center Model which deals with boolean spatial data, and the spatial co-location rule algorithm which deals with spatial multi- dimensional classified data. Then we

propose an improvement on the latter by controlling the size of candidate co-location patterns effectively and reducing the scan time of database, in order to enhance efficiency of the algorithm. Chapter 4, take economy census statistical data in Daxing district, Beijing, 2005 as an example, analyze the spatial distribution rules of industry, catering industry and school, and obtain some conclusions using the improved algorithm. Finally, in Chapter 5, summarize the content, innovation and limitations of this paper.

Key Words: Spatial Data Mining; Geographic Information System; Spatial Co-location Rule

目 录

CONTENTS

第一章 绪论

1.1 空间数据挖掘综述

1.1.1 知识发现与数据挖掘

随着信息应用的普及和数据库技术的成熟，人类累积的数据量正在呈指数级增长，全世界每天存入的数据量超过万兆字符。面临浩如烟海的数据，人们呼唤从数据的汪洋大海中去芜求精、去伪求真。因此，“从数据库中发现知识”（Knowledge Discovery in Database，KDD）及其核心技术——数据挖掘（Data Mining，DM）应运而生。

KDD 被定义为从数据中发现隐含的、先前不知道的、潜在有用信息的非平凡过程，它是数据和数据库急剧增长远远超过人们对数据处理和理解能力的背景下产生的，也是数据库、人工智能、统计和可视化等多学科与技术交叉、渗透、融合发展形成的交叉学科[1]。

还有很多含义相同或相近的术语，如数据分析（Data Exploring）、数据融合（Data Fusion）、数据抽取（Data Extraction）、知识提取（Knowledge Discovery）、数据捕捞（Data Dredging）等，尽管提法不尽相同，但其本质是一样。因此，为统一认识，在 1996 年出版的总结该领域进展的权威论义集《知识发现与数据挖掘研究进展》[2]中，Fayyad 等人重新给出 KDD 和数据挖掘的定义，将二者加以区分：KDD 是从数据中辨别有效的、新颖的、潜在有用的、最终可理解模式的过程，涉及的范围比较广；数据挖掘是 KDD 中通过特定的算法在可接受的计算效率限制内生成特定模式的一个步骤，是 KDD 的一部分。换句话说，KDD 是一个包括数据选择、数据预处理、数据变换、数据挖掘、模式评价等步骤，最终得到知识的全过程，而数据挖掘只是其中一个关键步骤。KDD 的全过程描述如图 1.1 所示。

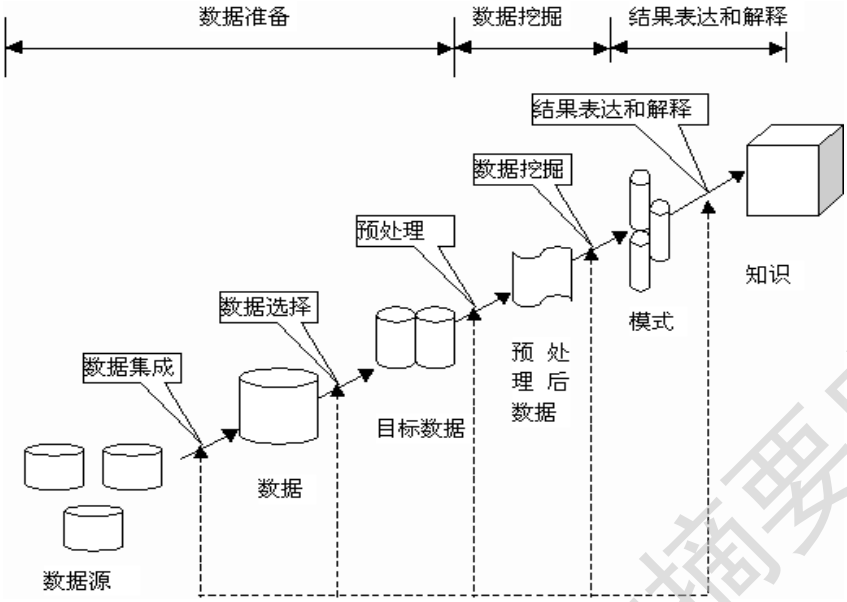

图 1.1　知识发现全过程示意图

IEEE 的 Knowledge and Data Engineering 会刊在 1993 年出版有关知识发现的技术专刊，所发表的 5 篇论文代表当时 KDD 研究的最新成果和动态，较全面地论述 KDD 系统方法论、发现结果的评价、KDD 系统设计的逻辑方法，集中讨论鉴于数据库的动态性冗余、高噪声和小确定性、KDD 系统与其它传统的机器学习、数据库技术、专家系统、人工神经网络、数理统计分析系统的联系和区别，以及相应的基本对策。并行计算、计算机网络和信息工程等其它领域的国际学会、期刊也把数据挖掘和知识发现列为专题和专刊讨论，成为当前计算机科学界的一大研究热点。1997 年 1 月，国际上第一本 KDD 杂志——数据挖掘和知识发现(Data Mining and Knowledge Discovery)创刊。1998 年成立 TKDD 的组织——ACMS ICKDD。2000 年，J.Han 教授出版第一部数据挖掘的专著——《数据挖掘：概念与技术》，该书系统介绍数据库技术的发展和数据挖掘应用的重要性，数据仓库 OLAP（联机分析处理）技术，数据预处理技术，数据挖掘技术，先进的数据库系统中的数据挖掘方法，数据挖掘的应用和一些具有挑战性的研究问题。随后，许多关于数据挖掘的书籍纷纷面世。

在国内，从 1993 年开始，一些基金和企业也开始资助数据挖掘和知识发现的研究。目前，国内许多高等院校和科研单位已经开展数据挖掘的基础理论及其应用研究，并取得很大的成果。