文章编号:1001 - 9081(2005)03 - 0710 - 03

基于 J2EE 的空间数据挖掘系统设计与实现

涂建东[1,2],陈崇成[1,2],黄洪宇[1,2],张群洪[1,2]
(1. 福州大学 数据挖掘与信息共享教育部重点实验室,福建 福州 350002;
2. 福州大学 福建省空间信息工程研究中心,福建 福州 350002)
(chencc@fzu.edu.cn)

摘 要:在分析空间数据挖掘特点的基础上,充分集成空间数据仓库技术、空间数据挖掘技术以及空间信息表达等技术,设计了一个基于 J2EE 的空间数据挖掘原型。重点介绍该原型系统的功能框架与体系结构、空间关联规则挖掘模块、挖掘结果的可视化表达模块的设计和实现办法。最后给出系统以某市土地利用现状数据集为例的空间关联规则挖掘结果界面。结果表明该系统可较好地满足可靠性、扩展性、可用性等业务需要。

关键词:空间数据挖掘;空间关联规则;J2EE;可视化
中图分类号: TP392 **文献标识码**:A

Design and implementation of J2EE-based spatial data mining

TU Jian-dong[1,2], CHEN Chong-cheng[1,2], HUANG Hong-yu[1,2], ZHANG Qun-hong[1,2]
(1. *Key Laboratory of Data Mining and Information Sharing of Ministry of Education, Fuzhou University, Fuzhou Fujian* 350002, *China*;
2. *Spatial Information Research Center of Fujian Province, Fuzhou University, Fuzhou Fujian* 350002, *China*)

Abstract: Based on the characters of spatial data, a spatial data mining prototype system with J2EE was designed, which integrated spatial data warehouse, spatial data mining, and spatial visualization techniques. The major functional modules of the prototype were described: spatial data management and interactive module, spatial data mining module, and visualization module. Finally, implementation of the prototype was demonstrated using the land use data of a certain city, and the typical user interfaces were illustrated. It is available to meet the need of reliability, expansibility and usability.

Key words: spatial data mining; spatial association rules; J2EE; visualization

0 引言

数据挖掘(DM)技术已经成为解决"数据爆炸、知识贫乏"问题的有效手段,各类数据挖掘方法与信息提取手段层出不穷,研究和应用领域也从最初的关系数据和事务数据挖掘拓展到(地理)空间数据挖掘(SDM)。面向地学和环境等领域应用的空间数据挖掘,作为数据挖掘的一种类型有其共性,也有特殊的方面,应该说空间数据挖掘受空间数据特性和应用范围的影响表现为更为复杂烦琐,提取和发现的知识类型也更为多样和丰富。除了能提取地理实体几何特征知识外,空间数据挖掘还能发现空间分布、空间关联、空间层次、空间分区、空间演变等空间知识[1]。

目前,国内外都开展了空间数据挖掘与知识发现方面的研究,但大多集中在挖掘算法研究上。加 Han Jiawei 教授领导的小组,较早对此进行系统全面的研究[2],并在 MapInfo GIS 平台上建立了空间数据挖掘的扩展模块 GeoMiner[3]。为克服传统的统计分析和数据挖掘技术消耗资源大、必须离线操作、不允许用户与结果交互、无法实时改变各种参数等弱点,Lu 等开发了两个基于网络的可视化和挖掘原型系统(Mapcube 和 Mapview),用于汇总分析交通和人口普查数据的时空模式和趋势[4]。May 等基于 EJB 多层体系结构,开发了空间数据挖掘与交互式可视化的集成平台[5]。国内武汉大学的李德仁院士最早关注空间数据挖掘问题,并作了开拓性工作,随后李德仁院士及邸凯昌等人提出了状态空间理论和云理论并将其运用到空间数据挖掘中来[6]。孙连英提出基于超图模型的空间数据挖掘模型[7],袁红春等人在 MapInfo GIS 平台上用 VC++ 开发出原型 GisMiner[8]。本文在分析空间数据挖掘特点的基础上,设计了一个基于 J2EE 的空间数据挖掘原型系统,并实现了它在土地利用现状空间分布相邻优势的规则挖掘中的应用。

1 空间数据挖掘系统设计

1.1 系统体系结构和功能框架

基于 J2EE 的空间数据挖掘原型系统(SircGeoMiner)的设计目标,就是要建立一个分布式的空间数据挖掘平台,集成空间数据库、数据挖掘模型和数据挖掘知识表达等功能。系统以存储在底层数据库(或数据仓库)中的空间数据作为数据挖掘及分析的对象,应用 GIS 空间分析原理和方法提取诸如空间拓扑关系等空间分布和关系信息,通过具体数据挖掘模型对其进行进一步的分析、处理、转化、挖掘结果的图形化表达,以探询空间信息内在的、通过传统 GIS 空间分析功能无法分析得到的抽象规则。

为了满足系统需求,该原型系统采用基于 EJB 三层结构的 B/S 模式,后台空间数据库存储要进行挖掘的空间数据表

收稿日期:2004 - 08 - 17;**修订日期**:2004 - 10 - 27 **基金项目**:国家 863 计划项目(AA633010)
作者简介:涂建东(1980 -) 女,广东人,硕士研究生,主要研究方向:空间数据挖掘、数据可视化; 陈崇成(1968 -),男,福建人,副教授,博士,主要研究方向:资源与环境信息工程、空间信息集成技术、城市与环境遥感.

以及数据挖掘系统的数据参考,并以空间数据引擎 SDE 为应用连接器。在 Tomcat 应用服务器上配置相应的数据挖掘 EJB,把数据挖掘客户端的 JAR 文件和相应的 JNLP 描述文件存放在 Tomcat 服务器上,在多个客户机通过访问对应的 Web 页面,激发 Java Web Start,下载并运行数据挖掘客户端。系统总体功能框架由三大部分模块构成(图 1),即基于 WebGIS 的空间数据管理和人机交互模块、空间数据挖掘模块(SDM)以及图形用户界面。其中 SDM 模块是 SircGeoMiner 系统的核心模块,是用 ArcSDE 来完成空间数据挖掘中大量空间信息的抽取过程,而 GIS 组件则用于原始图层及挖掘结果的显示,从而实现了 GIS 技术与空间数据挖掘系统的集成。

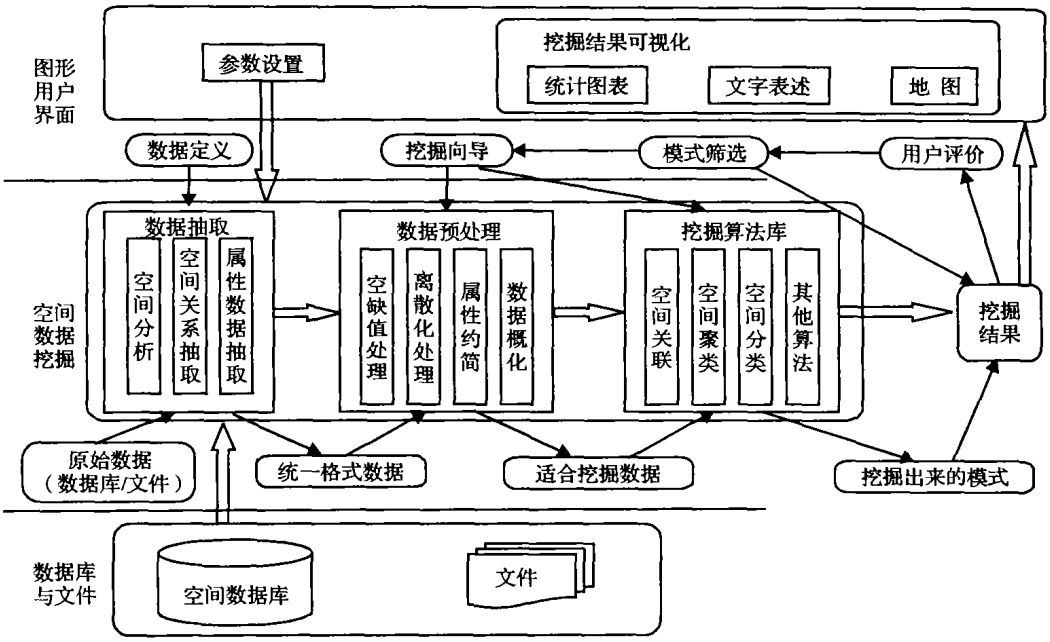


图 1　系统总体功能框架图

1.2　空间数据挖掘模块设计

本系统采用 JSP 技术开发动态网页,空间数据挖掘模型以 JavaBean 的形式嵌入到网页中。下面以初步实现的空间关联规则数据挖掘算法对空间数据挖掘模块设计作描述。

空间关联规则挖掘主要需经历下面两个步骤:

1) 数据预处理

原始空间数据中存在噪音数据或空缺值,需要对其进行一定处理,否则可能对挖掘结果产生重要负面影响,甚至可能造成算法失效。

接着对数据进行离散化或抽象化处理,连续属性离散化在数据挖掘中是一个很重要的问题。本系统实现了两种离散化方法,一种是无监督的学习方法:等频算法;另一种是有监督的学习方法:VDM 算法。由于挖掘算法本身需要考虑决策类,所以 VDM 算法比等频算法要好得多。

2) 空间关联规则挖掘算法

在分析传统关联规则数据挖掘算法基础上,结合空间数据特点,同时借鉴其他研究者的经验,本模块采用如下空间关联规则挖掘算法,具体算法流程如图 2 所示。

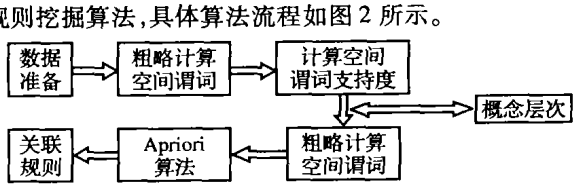


图 2　空间关联规则挖掘算法流程示意图

第一步:通过执行空间查询和空间分析,将所有与目标相关的空间对象、目标空间对象的参照集 S、与挖掘目标任务相关的空间关系的集合收集到数据库 Task_relevant_DB 中。

第二步:在一个粗略层次上执行高效的谓词计算。计算目标空间对象的最小限定矩形(MBR)的交,抽取 MBRs 间距离落在预设阈值之内的对象,并将描述对象间空间关系的谓词存储在数据库 Landuse_predicate_DB 中,其属性值是单个值或一组值。

第三步:为 Landuse_predicate_DB 中的每个谓词计算支持度,并过滤支持度低于最小支持度阈值的对象,从而形成数据库 Frequent_landuse_predicate_DB。

第四步:在 Frequent_landuse_predicate_DB 上执行精确空间计算,即采用 MBR 技术对经过第三步剪枝后的空间谓词关系进行检查,滤去与实际不相符合的空间谓词关系,形成新的拓扑关系数据表,并计算这些谓词的支持度,滤去支持度小的项目形成数据库 Fine_predicate_DB。

第五步:采用概念层次树(图 3 所示)对第四步形成的新的拓扑关系表进行概化后形成新的拓扑关系数据表,并采用 Apriori 算法在 Fine_predicate_DB 上抽取强空间关联规则并提取出关联规则.其具体算法步骤如下:

```
precedure find_frequent_predicates_and_mine_rules( DB );
for( l: = 1; L[l, 1] ≠ Ø and l < max_level; l ++) do begin
  /* L[1, 1]是在概念树 l 层上高频 1-谓词集表 */
  L[l, 1] : = get_frequent_1_predicate_sets( DB, l );
  /* 获取概念树 l 层上高频 1-谓词集表 L[l, 1] */
  for( k: = 2; L[l, k - 1] ≠ Ø; k ++) do begin
    Pk : = get_candidate_set( L[l, k - 1]);
    /* 从第 l 层的高频(k - 1)- 谓词集表获取 Pk */
    for each objects in S do begin
                                  /* 参照集 S 中的每一对象 s */
      Ps : = get_subsets( Pk, s );
                          /* 从 Pk 中获取满足 s 的候选集 Ps */
      for each candidate p ∈ Ps do p. support ++;
       /* 计算 Ps 中每个候选项的支持度 */
      end;
      L[l, k] : = { p ∈ Pk | p. support ≥ minsup[l]};
      /* 由 Ps 中每个候选项的支持度大于最小支持度阈值的项
      组成概念树 l 层上高频 k- 谓词集表 L[l, k] */
```

```
    Output: = generate_association_rules( L[l, k]);
    /* 从 L[l, k] 中导出强关联规则 */
  end;
 end;
end;
```

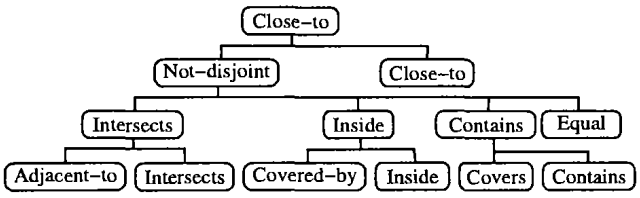


图3 空间拓扑关系的概念层次树示例

在上述步骤中,第2行表示从最顶层开始逐层挖掘关联规则,直至高频1-谓词集为空或抵达最低概念层。第3行表示对每一层 l 计算高频 l-谓词集并将其放入表 $L[l,1]$ 中。第4行到第10行采用Apriori算法逐步计算在概念树 l 层上的高频 k-谓词集 $L[l,k]$,直至在 l 层上的高频 $(k-1)$-谓词集 $L[l,k-1]$ 为空。第11行表示从概念树的第 l 层到高频 k-谓词表 $L[l,k]$ 中导出强关联规则。

1.3 可视化表达模块设计

原型系统可视化的目标,是将挖掘出来的晦涩的规则转化成易于理解、易于识别的三维图形和符号化表达方式,并通过Web方式实现远程访问,支持多用户同时交互浏览。在空间关联规则可视化中,至少需涉及如下五组参数:1)常用空间或非空间谓词集;2)候选空间或非空间谓词集;3)常用空间或非空间谓词集与候选空间或非空间谓词集之间的关联;4)关联的可信度;5)关联的支持度。

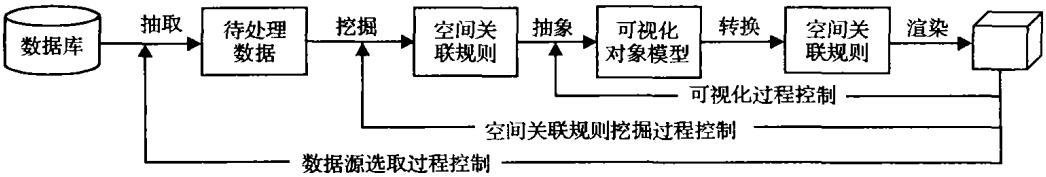


图4 空间关联规则结果的可视化流程图

结合空间关联规则可视化的特点及目标,笔者提出(图4所示)空间关联规则可视化的一般流程。

可视化过程具体采用以下几种类型方法:

1) 数据建模

关联规则可视化的建模方式可大致分为两大类:二维图形表示和三维图形表示。由于空间关联规则经常挖掘出涉及多个空间(或非空间)谓词组合的复杂规则,所以考虑采用三维图形方式,以方便复杂空间规则的表达。

具体模型拟用不同颜色的三维球体分别代表常用空间(或非空间)谓词集和候选空间(或非空间)谓词集;具体常用空间(或非空间)谓词集和候选空间(或非空间)谓词集的名称采用二维字体标注;谓词之间的关联用圆柱体连接表示,关联方向用圆锥体锥尖指向表示。

2) 三维交互设计

空间关联规则挖掘中对于同一空间(或非空间)谓词集,用户若选取不同的最小支持度和最小可信度作为域值控制,挖掘结果可能会有很大差别。如果域值较低,往往会出现挖掘出大量规则。这时,就需要设置不同的鼠标事件实现对场景的控制,例如,平移、旋转、放大、缩小等,以方便用户从各种角度观察可视化结果。

3) 关联强度的展示

关联强度内容较多,一次性全部展示将影响整体的可视化效果,所以,采用通过鼠标拾取谓词或关联,弹出提示标签的方式,来展示与所点取实体对应的关联强度的具体内容。

2 系统实现与应用

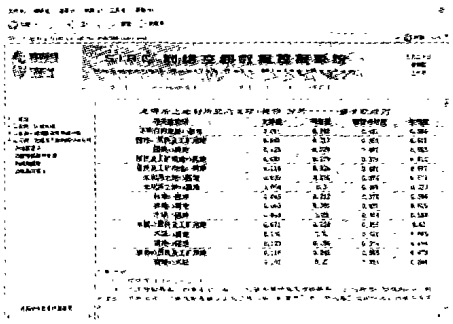
2.1 系统实现

空间数据挖掘系统原型SircGeoMiner,在实现上采用当前流行的B/S模式进行开发,服务层以Apache TOMCAT Server 5.0为基础,使用J2EE技术构建服务器应用程序来实现网络业务逻辑功能。数据层则选用SQL Server 2000和其扩展模块Analysis Server数据库管理系统全面支持系统的数据操作。表现层的用户界面采用HTML语言和JavaScript语言编写。

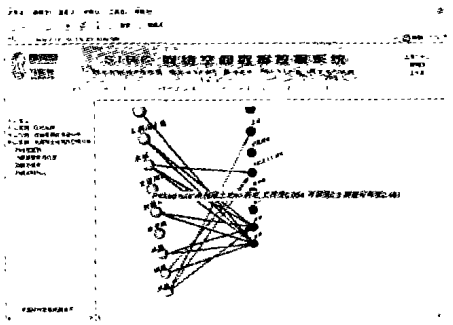
采用J2EE技术系统进行底层开发的具体步骤包括:组件模块模型的设计;组件模型的部署、组件的具体化、产生代码、编译、链接等过程。

本系统采用JSP/Servlet实现用户界面,中间核心部分由EJB组件完成,EJB的Session组件实现业务逻辑,EJB的Entity组件实现数据存储。开发EJB组件模块首先定义三个类:Bean类本身,Bean的本地(Home)和远程(Remote)接口类。然后进行配置、组装。最终,实现J2EE技术、空间数据仓库技术、空间数据挖掘技术以及空间信息表达等技术的无缝集成,并构成一个可扩展的软件平台。

2.2 应用示例



(a)



(b)

图5 某市土地利用数据集空间关联规则挖掘结果

笔者将空间数据挖掘系统应用于某市土地利用类型空间分布相邻优势分析中。某市土地利用现状数据集收集了该市九种

(下转第716页)

个安全尾段组也将被包括在内[1]。安全性的安全尾段组只将包含一个 UST 段。一旦 EDIFACT 结构被密码化后,其他的 EDIFACT 安全服务将不被提供。

以下是该系统加密安全性使用原则:

1) 多安全服务

如果同一时间超过一个安全服务被请求,除了安全性要被执行外,依照在 ISO 9735-5 中定义的规则,在 EDIFACT 结构发送方加密之前应用。接受方在解密之后将执行相关的审核确认操作。

2) 安全性

提供的 EDIFACT 结构的安全性必须与在 ISO 10181-5 中定义的原则相一致。安全性的安全服务必须在安全头段组中被详细描述,运算法则将在段组 1 中的一个 USA 段中被确定[4]。这个 USA 段也可能包含建立在通讯双方如安全创建人和安全接受者之间的主要关系必需的数据。

担当安全创建人的一方将 EDIFACT 结构加密,从它的头段(数据交换,数据组,报文或数据包)结束之后,到第一次记录它的段尾(数据交换,数据组,报文或数据包)之前,并把结果视为密码化的数据。在接受密码化的数据之上,通讯双方扮演安全接受者,将解码密码化的数据复原到最初的 EDIFACT 结构,去掉头段和尾段。

3) 内部表示和过滤函数

加密技术处理的结果表面上是一组随机的字符串。它的实现在确定的有限制能力的互联网络上可能会有一定的困难。为了要避免这个问题,比特串可以被映射到可逆的依靠一个过滤函数设定的特殊字符集之上[5]。

使用一个过滤函数要扩张密码化的数据大小。不同的过滤函数有不同的扩充因素。一些可能允许被过滤的文本包含目标字符集的字符,包括服务性字符,例如段结束,然而其他过滤函数可能过滤掉这些服务性字符[6]。

由 USD 和 USU 段传送的"八位组的数据长度"数据元中数据的长度,将说明密码化的(压缩并过滤)数据的长度。这将用来找出密码化数据的末端。用到的过滤函数将被简要地在加密安全头段组中 USH 中的 0505(过滤函数,编码)进行说明。

4) 在加密之前的压缩技术的使用

编写密码的计算开销直接与要加密的数据的大小有关系,它可能对加密之前进行的压缩数据有帮助。大多数的压缩技术对密码化的文本是不会有效的,因此如果压缩是需要的,它将在加密之前被应用。

总之,当压缩用在安全性服务的时候,安全头段组可能包含数据在加密技术之前被压缩的指示、识别压缩的运算法则和可选择的参数等信息。在如此的一个情形中,在密码化的数据解密之后,数据必须在 EDIFACT 结构恢复之前解压缩。

5) 处理操作的次序

当处理 EDIFACT 结构提供安全性时,操作依下列各项将被运行:压缩 EDIFACT 结构(可选择的),计算被压缩的数据的完整性的数值;将已压缩和完整性保护的 EDIFACT 结构加密;过滤(已压缩和完整性保护)密码化的数据(可选择的)。

当处理密码化的 EDIFACT 结构复原到最初的 EDIFACT 结构的时候,操作将依下列各项运行:反过滤已被过滤的密码化的数据(如果过滤了);解码密码化的数据;对被压缩的数据(如果有完整性的值)校验完整性的值和伸展(也就是解压缩)解码了的数据,复原最初的 EDIFACT 结构(如果是被压缩的)。

参考文献:

[1] ISO9735-1990, Electronic data interchange for administrator, commerce and transport-Application level syntax rules(Syntax version number:4) [S], 1990.

[2] STEVENS A, WALNUM C. Standard C++ Bible[M]. 北京:电子工业出版社, 2001.

[3] 刘尊全. 刘氏高强度公开加密算法设计原理与装置(第2版)[M]. 北京:清华大学出版社, 1998.

[4] SCHNEIER B. 应用密码学[M]. 北京:机械工业出版社, 2000.

[5] 胡予濮,张玉清,肖国镇. 对称密码学[M]. 北京:机械工业出版社, 2002.

[6] 胡志远. 口令破解于加密技术[M]. 北京:机械工业出版社, 2003.

(上接第 712 页)

土地利用类型的空间分布数据,该数据集存储在 ArcGIS 的空间数据引擎 SDE 中。数据集经空间关联规则挖掘后,得出一系列土地地类周边相邻区域较可能出现的土地地类情况。

图 5 为对该土地利用现状数据集实施空间关联规则挖掘的结果。图 5(a)得出某种土地类型的周边地块出现的地类有一定规律。如当目标地类是居民及工矿用地时,其周边区域地类常为耕地,而当目标地类是耕地时,周边的地类为水域的概率较高等。图 5(b)为空间关联规则挖掘可视化结果,其中"Picked rule:未利用土地⇒耕地,支持度 0.054,可信度 0.3,期望可信度 0.481"为用户拾取查询结果,即未利用土地周边出现耕地的支持度为 5.4%,可信度为 30%。

参考文献:

[1] 李德仁,关泽群. 空间信息系统的集成与实现[M]. 武汉:武汉测绘科技大学出版社, 2000. 85-188.

[2] KOPERSKI K, ADHIKARY J, HAN J. Spatial Data Mining: Progress and Challenge[A]. Proceedings of SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96)[C]. Montreal, Canada, 1996.

[3] HAN J, KOPERSKI K, STEFANOVIC N. GeoMiner: A System Prototype for Spatial Data Ming[A]. Proceedings of ACM SIGMOD International Conference on Management of Data[C], 1997. 553-556.

[4] LU C-T, KOU Y-F, WANG H-J, *et al*. Two Web-based Spatial Data Visualization and Mining Systems: Mapcube & Mapview[A]. Proceedings of International Workshop on Next Generation Geospatial Information[C]. Cambridge, 2003.

[5] MAY M, SAVINOV A. An Integrated Platform for Spatial Data Mining and Interactive Visual Analysis[A]. Proceedings of Data Mining 2002, the 3rd International Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields[C]. Bologna, Italy, 2002.

[6] 邸凯昌. 空间数据发掘和知识发现[M]. 武汉:武汉大学出版社, 2002.

[7] 孙连英,彭苏萍,张德政. 基于超图模型的空间数据挖掘[J]. 计算机工程与应用, 2002, 38(11):31-33.

[8] 袁红春,熊范纶,杭小树,等. 一个适用于地理信息系统的数据挖掘工具——GisMiner[J]. 中国科学技术大学学报, 2002, 32(2):217-221.