

基于 SVG 的空间关联规则挖掘

李 慧, 李 岩, 王兴芳

(华南师范大学 计算机学院, 广东 广州 510631)

摘要:随着网络技术的飞速发展,SVG 成为矢量图形发布的新一代标准,越来越多的 SVG 文档涌现出来。SVG 文档中隐藏着大量有趣的空间信息,因而如何从 SVG 文档中发现有趣的空间信息成为数据挖掘领域中值得研究的问题。讨论空间关联规则的挖掘,采用多维多层交叉关联规则挖掘技术,综合利用 SVG 文档中的空间信息和非空间信息进行挖掘,可以较好地 SVG 文档中挖掘隐藏的空间关联规则。

关键词:SVG 文档;矢量图形;空间数据挖掘;空间关联规则挖掘

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2007)10-0116-04

Spatial Association Rule Mining Based on SVG

LI Hui, LI Yan, WANG Xing-fang

(Computer School of South China Normal University, Guangzhou 510631, China)

Abstract: With the popularization and application of the computer network, SVG has been the new publishing standard for vector graphic. More and more SVG documents appear. There are large interest spatial information hidden in SVG document. Many researcher focus their research on finding hidden spatial information in SVG documents. In this paper, apply cross multilevel association rule discovery technique to find spatial association rules from a combination of spatial and non-spatial information of SVG document.

Key words: SVG document; vector graphic; spatial data mining; spatial association rule mining

1 概念介绍

1.1 SVG 文档

随着 Internet 的迅速发展,图形图像信息也需要通过网络实现共享。目前浏览器显示的大都是栅格形式的图像,栅格文件具有很多限制性,这些问题矢量图形可以很好地解决。W3C 推出的 SVG(Scalable Vector Graphics, 可伸缩矢量图形),是一种基于 XML 的图形标准,因它具有纯文本、开放、动态、可缩放和平台无关等特性,已成为进行空间矢量数据发布的生力军。SVG 基于 XML 来描述二维矢量图形,对于基本图形元素和非几何特征都可以很好表达。每一个 SVG 文档含有一个根元素<SVG>,需表达的整个内容处于根元素<SVG></SVG>之间,空间对象以图层组织。图层 Layer 可采用<G>表示,属于同一图层的几何元素可以是任何基本几何元素及其组合,包含于同一组<G></G>之间。点集、线集、面集、复杂几何体类等也可采用<G>表示,<G>元素的 ID 用于标识不同类型。基本的图形元素:点 Point、线 Line

String、面 Polygon,可采用<circle>、<rect>或<path>等表达。

几何图形的非几何特征表达在 SVG 中一般采用两种方式:内嵌法和外联法。内嵌法一般是作为元素(如 circle, path 等元素)的属性进行表达^[1];外联法是指非几何特征数据存储在服务器的数据库中,并且通过 ID 与 SVG 文档中的相关地进行连接。

SVG 文档的结构可以认为是树状结构,图 1 是简单的 SVG 文档,文档包含了两个图层,分别是行政区域图层和水库分布图层。

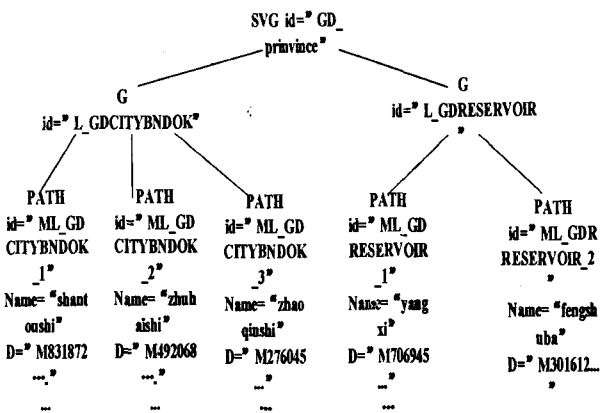


图 1 SVG 文档示例

收稿日期:2006-12-29

作者简介:李 慧(1980-),女,山西人,博士,研究方向为空间信息;李 岩,硕士,教授,研究方向为空间信息。

1.2 空间拓扑关系

几何对象间的空间关系通常为三类:空间拓扑关系、度量空间关系和顺序空间关系。空间拓扑关系对于空间数据挖掘非常重要。通过空间关系进行归纳和分类,得出了 5 种基本空间拓扑关系^[2]:相离关系(Disjoint)、相接关系(Touch)、相交关系(Cross)、包含于关系(Within)、部分覆盖关系(Overlap)。这些空间拓扑关系之间具有层次性,图 2 是空间拓扑关系层次图。图中非相离关系是相离关系的补集。相邻关系是相离关系的子集,如果两个空间对象相离,并且两者距离小于给定值,那么它们间的拓扑关系是相邻。因为在实际应用中,只有两个相离几何对象间距离在一定范围内,才具有实际讨论意义。

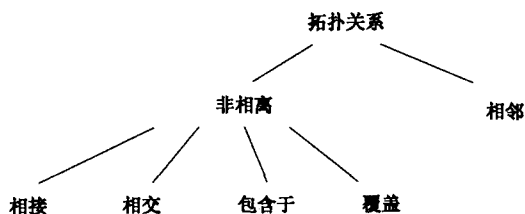


图 2 空间拓扑关系概念层次图

1.3 关联规则

设 $I = \{i_1, i_2, \dots, i_m\}$ 是项的集合。记 D 为交易的 T 集合,并且 $T \subseteq I$ 。如果项集 $X \subseteq I \wedge X \subseteq T$,那么称交易 T 包含 X 。支持度是交易集中包含 X 和 Y 的交易数与所有交易数之比,即 $\text{sup}(X \Rightarrow Y) = \text{sup}(X \cup Y)$ 。可信度是包含 X 和 Y 的交易数与包含 X 的交易数之比,即: $\text{conf}(X \Rightarrow Y) = \text{sup}(X \cup Y) / \text{sup}(X)$ 。

关联规则是形如 $X \Rightarrow Y$ 的蕴涵式,这里 $X \subset I, Y \subset I$,并且 $X \cap Y = \emptyset$ 。同时满足最小支持度阈值和最小置信度阈值的规则成为强规则。

一直以来关联规则挖掘都是数据挖掘研究热点,它与空间领域结合产生了空间关联规则挖掘。目前空间关联规则挖掘研究主要是针对栅格图像^[3-5],在矢量数据上进行挖掘也有一定的研究^[6]。但基于结构化的 SVG 文档进行空间关联规则挖掘还很少研究,文中探讨如何基于 SVG 文档进行空间关联规则挖掘。

2 空间关联规则挖掘

2.1 数据预处理

SVG 是结构化文档,元素中存在层次性。空间对象间的拓扑关系具有层次性。非空间属性也可以定义其概念层次结构,因而适合在 SVG 文档上进行多维多层关联规则挖掘。但是 SVG 文档并不是为了空间挖掘而建立的,不能直接用于多维多层关联规则挖掘,需要将其转换成适合挖掘的形式。数据预处理中的基本

思路是从 SVG 文档中选取挖掘任务所需几何对象,根据其空间、非空间信息,构造事务数据库。

拓扑关系是挖掘所需主要空间信息,但 SVG 文档中通常没有显示地记录几何对象间的拓扑关系,需要调用相应的空间分析算子进行空间分析,构造空间拓扑关系表^[7],如表 1 所示。构造规则: $I_{\max}^{\text{db}} = \{(O_1, O_2, T) \mid O_1, O_2 \in \text{SDoc}, O_1 \text{distance} = \text{dis}t O_2 \wedge \text{dist} \leq \max \wedge O_1 T O_2\}$ 。SDoc 是 SVG 文档, max、distance 均为数值数据, T 为拓扑关系。

表 1 空间拓扑关系表

TID	O ₁	O ₂	Topology
1	ML-GDCITYBNDOK-1	ML-GDRESERVOIR-5	Within
2	ML-GDCITYBNDOK-5	ML-GDCITYBNDOK-6	Touch
3

许多挖掘任务,不仅仅在空间信息上挖掘,还需要与非空间信息相结合。根据不同挖掘任务,从 SVG 文档中提取空间对象的非空间特征,在空间拓扑关系表中提取对象间的拓扑关系,构造关系表,如表 2 所示。

表 2 地区城市人口与道路拓扑关系表

TID	O ₁	O ₂	Topology	Population
1	ML-GDCITYBNDOK-1	ML-GDRoad-5	Disjoint	100
2	ML-GDCITYBNDOK-2	ML-GDRoad-6	Crosses	130
3

进行多维多层数据挖掘前,数据还需要先转换成单维单层结构。根据 SVG 文档结构、拓扑层次结构、非空间特征概念层次结构对关系表中所有项编码。每个项所对应编码第 1 位用于区分项的类型。取值 1 表示是空间对象,2 表示是拓扑关系,3 表示是非空间属性。编码其他位取值由其所对应层次结构决定。空间对象的编码可借助 DOM(Document Object Model)操作,采用深度遍历来实现。算法如下:

输入:SVG 文档,路径 P 为空字符串

输出:空间对象编码 $\langle L, P \rangle$

步骤:

1)调用 GetSVGDocument()函数,得到 SVG DOM 文档对象。

2)调用 GetElementByid()函数,得到结点 L 。判断结点 L :

(1)如果 L 是叶结点,访问标志改为 True,输出 $\langle L, P \rangle$,返回父结点。

(2)如果结点 L 是非叶结点:

a. 如果 L 的访问标志为 False,输出 $\langle L, P \rangle$,记录访问标志改为 True,字符 1 作为尾字符加入路径 P ,访问 L 的第一个子结点;

b. 如果 L 的访问标志为 True, L 有未访问子结点,

路径 P 的最后字符值加 1,访问 L 的下一个子结点;
c.如果 L 的访问标志为 True,L 所有子结点已被访问,去掉 P 的最后一个字符,更改后的 P 的最后一个字符值加 1,返回父结点。

3)依次循环,直到 SVG 文档所有对象被访问。
拓扑关系和非空间属性编码相对简单,根据预先定义的层次结构就可以得到相应编码(如图 3,4 所示)。图 3 是拓扑关系编码图,拓扑关系所对应的编码是从根结点到本结点路径上所有编码的组合,例如相接关系的编码为“211”,相邻关系编码为“22”。图 4 是非空间属性人口、收入概念层次结构编码图。表 3 是表 2 运用上述编码规则后形成的编码表。

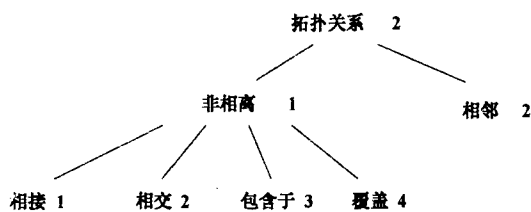


图 3 空间拓扑概念层次编码图

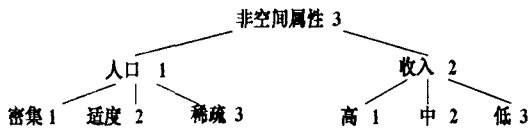


图 4 非空间属性:人口、收入概念层次编码图

表 3 地区城市人口与道路拓扑关系编码表

TID	O ₁	O ₂	Topology	Population
1	111	131	211	311
2	112	131	22 *	312
3	112	132	212	312
...

2.2 空间关联规则生成

算法如下所示:

输入:编码表 Ctab、最大层次数 Max_l、各层最小支持度 Minsup[Max_l]、最小置信度 Mincon[Max_l]

输出:强关联规则

步骤:

1)扫描编码表 Ctab,计算 1 层频繁 1 项目集到 Max_l 层频繁 1 项目集。

2)循环调用如下过程产生 L 层频繁 2 项目集到 L 层频繁 k 项目集:

a.由频繁 L 层(k-1)项目集产生 L 层候选 k 项目集。

b.循环计算 L 层候选 k 项目集中各项目的支持度,得到 L 层频繁 k 项目集。

3)根据最小置信度,得到对应的强关联规则,检查冗余,剔除冗余规则。

算法具体分析如下:

步骤 1:频繁 1 项集生成中不同层赋予不同的最小支持度,较低层使用递减的最小支持度,避免丢掉出现在较低抽象层中有意义的关联规则。

步骤 2:L 层 k 项目候选集 C_k 生成,需要 L 层(k-1)项目集进行自身连接生成 L 层 k 项目候选集,还要使用 1 层 1 项目集、2 层 1 项目集...L-1 层 1 项目集与 L 层(k-1)项目集连接生成 L 层 k 项目候选集。

步骤 3:各层规定不同的最小置信度,由各层得到的频繁项生成强关联规则。由于项之间的“祖先”关系,如果规则的祖先,它的支持度和置信度都接近于“期望”值,那这个规则是冗余的,应该剔除^[8]。

3 应用分析

当前社会经济、人口迅速发展,生态环境越来越受到关注。图 5 是某区域的 SVG 矢量图,图中包括行政区域、长年河、水库、林地、草地等图层。通过挖掘空间关联规则,可帮助分析此区域的环境特征。例如:分析此区域城市与水域、植被等分布的关系,可运用上述空间挖掘思路进行空间挖掘。由于此挖掘任务未涉及非空间属性,只需选取此区域中几何对象城市、水域、森林、草地以及其拓扑关系。城市、水域、植被等图层处于 SVG 文档层次关系第二层,拓扑层次关系总共分为三层。这里列出由第二、三层挖掘后的部分频繁模式。第二层最小支持度 56%,第三层最小支持度 40%:

is_a(A, town), is_a(B, river), not_disjoint(A, B),
is_a(C, forest), not_disjoint(A, C), is_a(D, grass), ot_disjoint(A, D), close_to(A, E), cross(A, B), within(A, C), within(A, D)

根据置信度,可以得到强关联规则,如

is_a(A, town)(is_a(B, river)(cross(A, B)

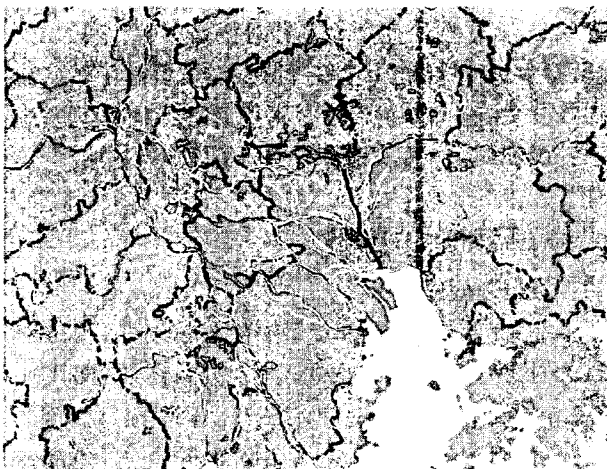


图 5 某地区 SVG 矢量图

以上信息表明在此区域城市大都分布有水源、各

种植被。但是还不能提供进一步详细信息,如城市水源是否充足,植被覆盖是否茂盛等。因而在之前空间信息基础上,加入非空间属性:城市面积、人均水量、人均草地面积、人均森林面积等,并且建立非空间属性层次图进一步进行挖掘,得到一组强关联规则。如规则: $is_a(A, town)(is_a(C, grass)(cross(A, C)(ave_river(Low))$,表明大多数城市都覆盖有草地,人均草地占有量低。

目前各个地区都处于发展当中,影响区域发展的因素很多,这些因素与发展的相关性通常都是隐性的,需要通过相关数据进行分析。如图 6 是某区域的 SVG 矢量图,图中有城市边界、公路两个图层,以及土地利用等属性数据。

在图 6 数据基础上分析城市的公路分布与土地各项利用变化之间的关系,进行空间挖掘。预处理中利用文中所述方法得到地区城市与道路拓扑关系编码表,此表中包含非空间属性,分别是公路分布密度、耕地、城乡工矿居民用地和未利用地的使用变化,其中公路分布密度利用地区城市与道路拓扑关系可求得。图 7 可视化地显示了各区域的公路分布和土地利用变化,各个区域根据公路分布密度的不同赋予了不同颜色。

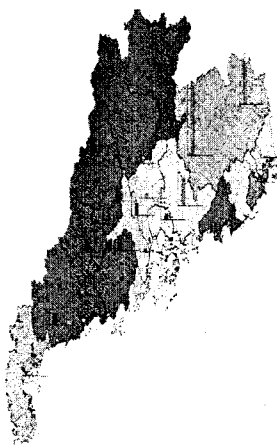
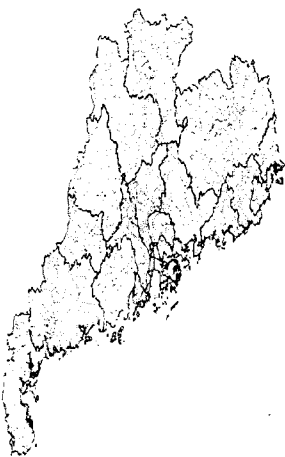


图 6 某地区 SVG 矢量图 图 7 可视化分析结果

最后设置最小支持度、置信度进行空间关联规则的生成,得到一组强关联规则,这些规则所显示的信息

可作为区域的发展策略的参考依据,帮助区域达到更好的和谐发展。

4 总 结

探讨了如何在 SVG 文档上挖掘空间关联规则。为了得到有趣信息,综合利用空间信息和非空间信息,基于挖掘任务进行多维多层交叉挖掘,并且应用于实际分析,扩展了 SVG 的研究与应用。但是由于基于 SVG 文档的挖掘还处于起步阶段,所做工作存在很多不足,如关联规则挖掘中没有考虑增量挖掘的情况;对于各层最小支持度和置信度依赖于经验进行手工设置;冗余规则的检测原则过于简单,没有考虑实际情况。这些问题都有待于进一步解决。

参考文献:

- [1] 徐云和,谢刚生,程朋根,等.基于 SVG 的空间数据的可视化[J].计算机应用研究,2005,22(2):46-48.
- [2] 邹 伦.地理信息系统——原理、方法和应用[M].北京:科学出版社,2001:65-66.
- [3] Kangkachit T, Waiyamai K. A business-oriented spatial association rule mining system prototype(Bosarm)[C]//In: Proceeding of Information and Computer Engineering Postgraduate Workshop. Thailand:[s. n.],2002.
- [4] Maleba D, Lisi F A. An ILP method for spatial association rule mining[C]//In: Working Notes of the First Workshop on Multi-relational Data Mining. Freiburg, Germany:[s. n.], 2001:18-29.
- [5] Sabhananda W, Waiyamai K. Data Mining: A Novel Approach for Multi-level association Rules Mining in Large Databases [C]//In: The Fifth National Computer Science and Engineering Conference. Chiang-Mai, Thailand:[s. n.],2001.
- [6] 库向阳,许五弟,薛惠峰.矢量空间数据库中关联规则的挖掘算法研究[J].计算机应用,2004,24(8):47-49.
- [7] Ester M, Kriegel H P, Sander J. Spatial Data Mining: A Database Approach[C]//in: Proc. 5th Int. Symp. on Large Spatial Databases. Berlin, Germany:[s. n.],1997:47-66.
- [8] Han Jiawei, Kamber M. 数据挖掘——概念与技术[M].北京:高等教育出版社,2005:236-237.

(上接第 115 页)

服以往 BP 学习过程中的缺点,使这一应用更有效。

参考文献:

- [1] 李孝安.神经网络与神经计算机导论[M].西安:西北工业大学出版社,1994.
- [2] 杨红涛,潘 昊.拆分组装方法与神经网络的结合[J].武

汉理工大学学报:信息版,2003(5):13-16.

- [3] 杨红涛,潘 昊.用 GDR-GA、拆分组装法训练神经网络[J].武汉理工大学学报:信息版,2004(1):30-34.
- [4] 陈方泽.用 EGA-GDR 训练神经网络[J].化工学报,1996(4):421-426.
- [5] 金菊良,杨晓华,丁 晶.标准遗传算法的改进方案——加速遗传算法[J].系统工程理论与实践,2001(4):8-13.