

基于邻接关系的空间数据挖掘技术的研究

张楠^{1,2}, 王子牛^{1,2,3}, 刘念^{1,2}

(1. 贵州大学 多媒体与信息安全实验室, 贵州 贵阳 550003;

2. 贵州大学 信息工程学院, 贵州 贵阳 550003;

3. 贵州大学 信息化管理中心, 贵州 贵阳 550003)

摘要:随着现代科学技术的迅速发展,复杂多变的空间数据日益膨胀,远远超出人们的解译能力,迫切地需要数据挖掘和知识发现为其提供知识。文中从空间数据挖掘的基本概念出发,详细阐述了空间数据的特点、空间邻接关系及其相关操作,并针对空间邻接关系给出了几种典型的空间数据挖掘方法。

关键词:空间数据挖掘; 地理信息系统; 邻接关系

中图分类号: TP311.13

文献标识码: A

文章编号: 1673-629X(2007)04-0154-04

Research of Spatial Data Mining Technique Based on Neighborhood Relation

ZHANG Nan^{1,2}, WANG Zi-niu^{1,2,3}, LIU Nian^{1,2}

(1. Multimedia and Information Security Lab., Guizhou Univ., Guiyang 550003, China;

2. College of Information Engineering, Guizhou Univ., Guiyang 550003, China;

3. Information Management Center, Guizhou Univ., Guiyang 550003, China)

Abstract: With the application and development of modern science and technique, the tremendous amounts of spatial and non-spatial data have been stored in large spatial database (SDB). These are far beyond the human ability to interpret and analyse of them, which is badly in need of spatial data mining (SDM) to provide knowledge. From the basic conceptions of SDM, introduces characteristics of spatial data, spatial neighborhood relation and its operation. Also discusses typical spatial data mining methods based on spatial neighborhood relation.

Key words: spatial data mining; geography information system; neighborhood relation

0 引言

空间数据库含有空间数据和非空间数据,空间数据主要是地表在地理信息系统中的二维投影,非空间数据则是除空间数据以外的一切数据。随着对地观测、获取设备的迅速发展,空间数据资源日益丰富。然而,数据资源中蕴含的知识远远没有得到充分的挖掘和利用,导致“空间数据爆炸,但空间知识贫乏”;同时,要求用户详细分析这些数据并提取感兴趣的知识或特征是不现实的。因此,从空间数据库中自动地挖掘知识,寻找数据库中不明确的、隐含的知识,空间关系或其它模式,即空间数据挖掘技术就显得越来越重要。

空间数据挖掘(Spatial Data Mining, 简称 SDM),是指从空间数据库中抽取隐含的知识、空间关系或非显式地存储在空间数据库中的有意义的特征或模式^[1,2]。这种技术可用于发现空间数据与非空间数据间的关系、构建空间知识库、优化查询、重组空间数据库和获取简明的总体特征等方面,它在 GIS、遥感、图像数据库、医疗影像处理、机器人导航等领域具有广阔的应用前景。

1 空间数据的特点

空间对象具有空间位置和距离属性,并且距离邻近的对象之间存在一定的相互作用,因此空间数据之间的关系更为复杂(不仅多了拓扑关系、方位关系,而且度量关系还与空间位置和对象间的距离有关),与其它类型的数据之间存在明显的差异。

空间数据具有如下复杂性特点^[2~5]:

(1)海量的数据。海量数据常使一些算法因难度

收稿日期:2006-07-03

基金项目:国家教育部“春晖计划”科研项目基金(Z2004152016);贵州大学“211工程”重点建设项目基金(2005)

作者简介:张楠(1979-),男,山东烟台人,硕士研究生,研究方向为 Rough Set 理论、数据挖掘;王子牛,教授,研究方向为数据挖掘、软件工程及网络技术。

或计算量过大而无法实施,因而空间数据挖掘的任务之一就是创建新的计算策略并发展新的高效算法,克服海量数据造成的技术困难。

(2)空间属性之间的非线性关系。它是空间系统复杂性的重要标志,反映了系统内部作用的复杂机制,是空间数据挖掘的主要任务之一。

(3)空间数据的尺度特征。空间数据在不同观察层次遵循的规律以及体现出的特征不尽相同。尺度特征是空间数据复杂性的又一表现形式,利用该性质可以探究空间信息在泛化和细化过程中所反映出的特征渐变规律。

(4)空间维数的增高。空间对象的属性增加极为迅速,如在遥感领域,由于感知器技术的飞速发展,波段的数目由几个增加到几十甚至上百个,如何从几十甚至几百维空间中挖掘数据、发现知识成为研究中的又一热点。

(5)空间信息的模糊性。模糊性几乎存在于各种类型的空间信息中,如空间位置的模糊性、空间相关性的模糊性以及模糊的属性值等。

(6)空间数据的缺失。数据缺失现象是由于某种不可抗拒的外力使数据无法获取或发生丢失。如何对丢失数据进行恢复并估计数据的固有分布参数,成为解决数据复杂性的难点之一。

2 空间邻接关系以及相关操作

2.1 空间邻接关系

空间关系^[4,6,7]是空间实体之间由于空间位置和形状的不同而造成的相互之间的各种联系。一般分为三种类型:拓扑关系(Topological Relation)、尺度关系(包括距离关系(Distance Relation)和方向关系(Orientation Relation)、顺序关系(依观察者的位置而定)。

下面主要讨论平面坐标下两个物体之间的二元关系(Binary Relation)。

(1)拓扑关系。两物体之间的拓扑关系具有不因参照物的拓扑变换(如放缩、旋转)而改变的特点,可通过9个相交矩阵模型定义。

A、B是两个空间实体,可以分别为一个点(P)、一条线(L)或一个多边形表示的面(R),符号 $A^0, A^-, \Gamma A$ 分别表示A的内部、外部和边界。A、B的 3×3 相交矩阵如下,用它可判别A、B之间的拓扑关系,包括: A disjoint B, A meet B, A overlap B, A equal B, A cover B, A covered - by B, A contain B, A inside B。

$$\begin{Bmatrix} \Gamma A \cap \Gamma B & \Gamma A \cap B^0 & \Gamma A \cap B^- \\ A^0 \cap \Gamma B & A^0 \cap B^0 & A^0 \cap B^- \\ A^- \cap \Gamma B & A^- \cap B^0 & A^- \cap B^- \end{Bmatrix}$$

(2)距离关系。距离关系 A distance δ_c B 是由两个物体A、B之间的距离 $\text{dist}(A, B)$ 与某个给定的常量 c 经过算术比较(<、>、=)确定的,即:若距离关系 A distance δ_c B 成立,当且仅当 $\text{dist}(A, B) \delta_c$ 。

(3)方向关系。是指A、B在空间分布上的相对方位关系,用 $B R A$ 表示,R是方向关系,A是源物体,B是目标物体。

方向关系的确定依赖于所考虑的构成物体点的个数。这里假定用源物体A的某个具有代表意义的点 $\text{Rep}(A)$ 为中心来代表A,对目标物体B则考虑其所有的点,则有:

$B \text{ north } A$ 成立,当且仅当 $\forall b \in B, b_y \geq \text{Rep}(A)_y$ 。同理可以定义 south, west, east。

$B \text{ northeast } A$ 成立,当且仅当 $\forall b \in B, b_x \geq \text{Rep}(A)_x$,且 $b_y \geq \text{Rep}(A)_y$ 。同理可以定义 southeast, southwest, northwest。

$B \text{ any-direction } A$ 恒成立。

图1、图2分别给出了空间距离关系和空间方位关系的几种示例。

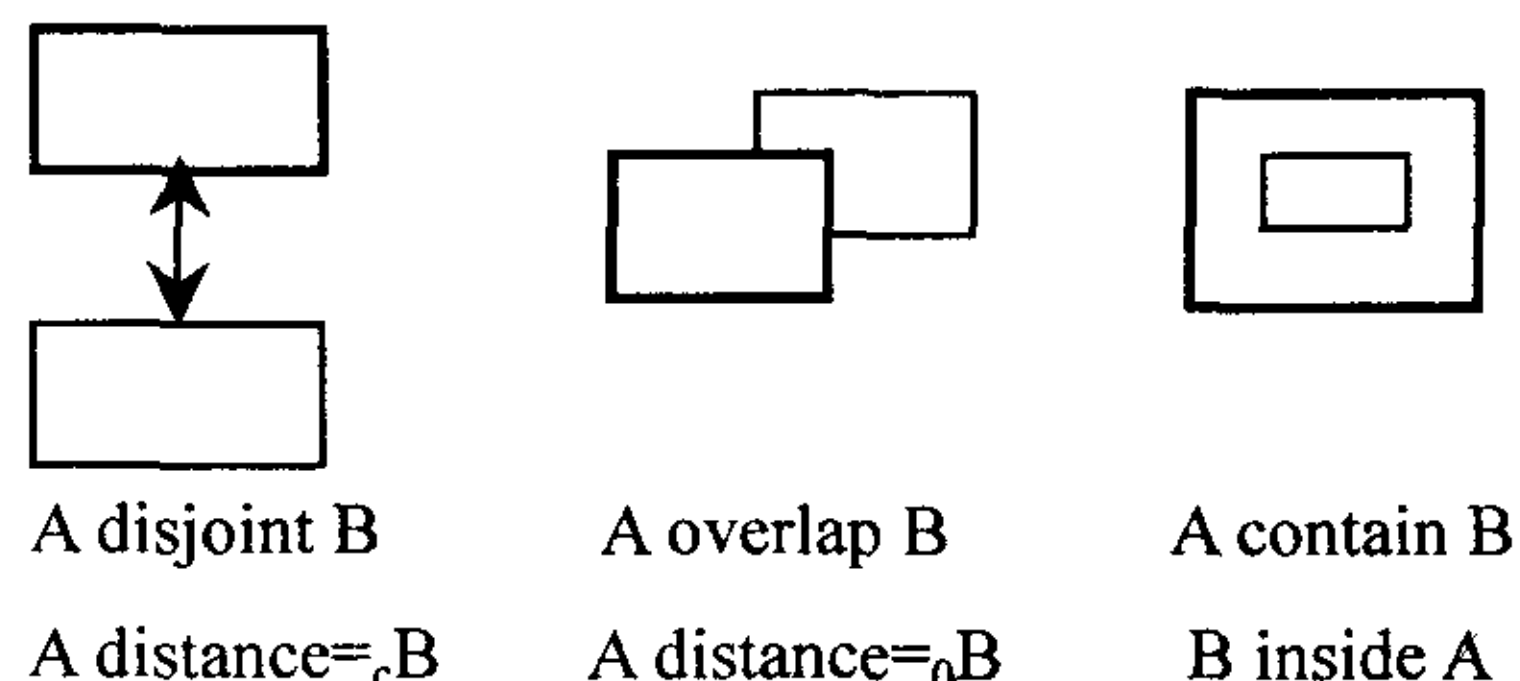


图1 空间物体间的距离关系

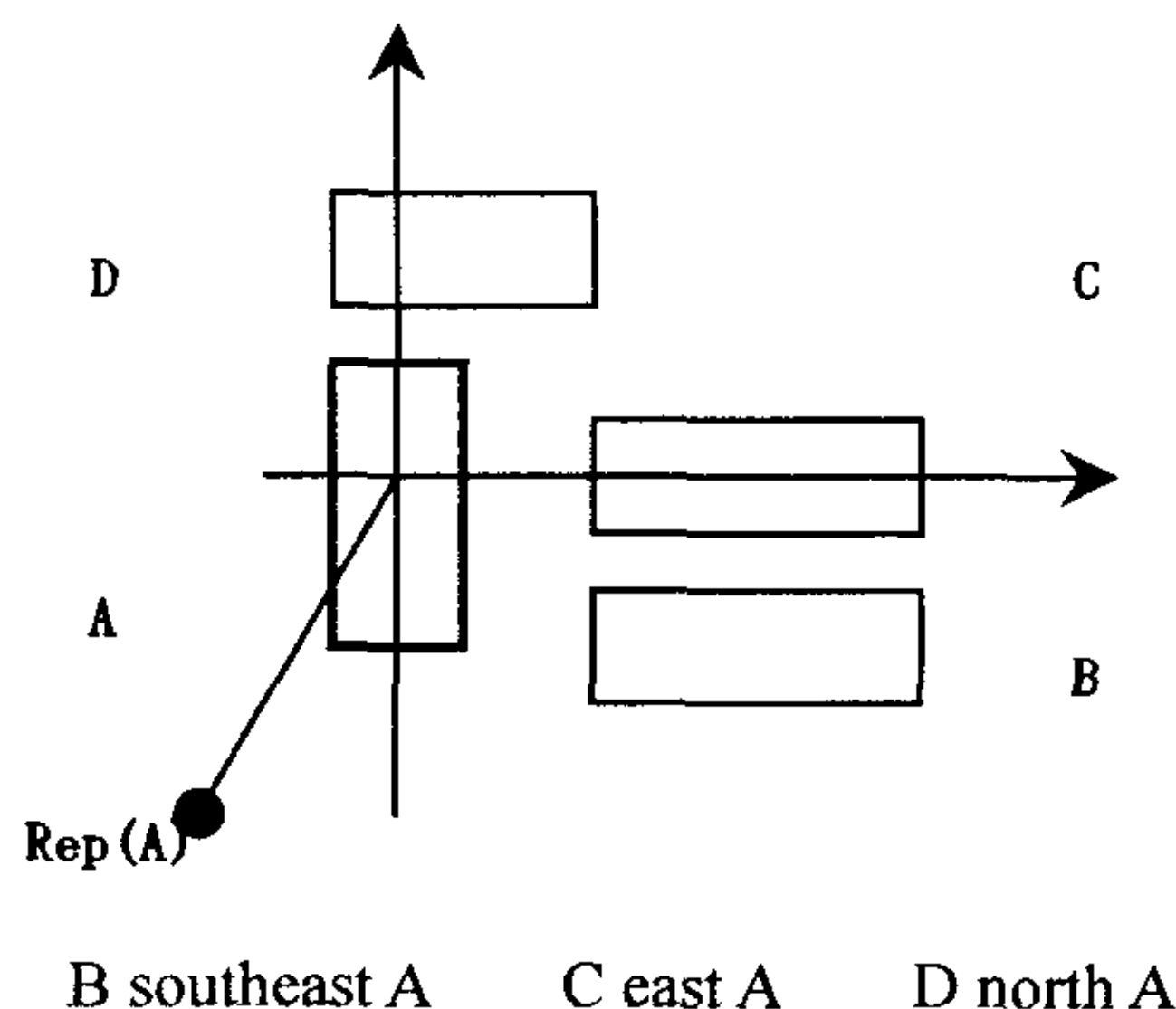


图2 空间物体间的方向关系

实际的空间物体之间的邻接关系往往比前面介绍的三种单纯关系更为复杂,但由简单的空间邻接关系通过与、或等逻辑运算可以表达复杂的空间邻接关系。

2.2 空间邻接关系的相关操作

空间实体可以是点或由点扩展得到的线、面等,因此,可以用点集来统一表示,设 2^{Points} 是二维点集, $DB \subset 2^{\text{Points}}$ 是一个空间数据库,neighbor是某种邻接关系。 $G = (N, E)$ 是DB中的空间物体依靠 neighbor 关系生

成的邻接图,其中结点集 $N = \{o_i \mid o_i \in DB\}$,边集 $E = \{(o_i, o_j) \mid o_i, o_j \in DB \text{ 且 } \text{neighbor}(o_i, o_j) = \text{true}, i \neq j\}$ 。

如果对所有 $o_i \in N (1 \leq i < k)$ 有 $\text{neighbor}(o_i, o_{i+1})$ 成立,则称结点序列 $\{o_1, o_2, \dots, o_i, \dots, o_k\}$ 是一条邻接路径(Neighborhood Path),结点数 k 就叫做邻接路径的长度。一条有效的邻接路径中不存在环路,即:

$$\forall i \leq k, j < k, \text{有 } i \neq j \Leftrightarrow o_i \neq o_j。$$

下面介绍几种基本的邻接关系相关操作:

(1) $\text{neighbors}(\text{graph}, \text{object}, \text{pred})$: 返回 graph 中与 object 存在路径连接且满足条件 pred 的所有物体的集合。pred 包括空间或非空间的属性关系。

(2) $\text{extensions}(\text{graph}, \text{paths}, \text{max}, \text{pred})$: 返回通过添加 graph 中至多 max 个点来延长邻接路径 paths 从而得到所有满足条件 pred 的路径的集合。pred 的作用是限制有效路径生成的数目,对提高 SDM 的效率有关键的影响。

(3) $\text{paths}(\text{Set of Objects})$: 生成由 Set of Objects 中某一元素组成的长度为 1 的路径的集合。

(4) $\text{objects}(\text{Set of Paths})$: 返回至少与 Set of Paths 中某一条路径中的一个结点相联系的物体的集合。

3 基于邻接关系的空间数据挖掘方法

根据空间数据的特点,介绍几种主要的基于邻接关系的空间数据挖掘方法:

(1) 空间聚类方法。空间聚类分析是要将空间数据库中的对象按照某些特征划分为不同的有意义的子类,同一子类中的对象具有高度相似的某种特征,并与不同子类的特征具有明显的差异。采用聚类分析的优点在于:想获取的结构或簇可以直接从数据中找到,不需要任何的背景知识。

目前已经提出了四种空间聚类方法:基于分割的方法、基于层次的方法、基于密度的方法和基于栅格的方法。

①基于分割的方法包括 K-平均法、K-中心点法和 EM 聚类法。它们都是采用一种迭代的重新定位技术,尝试通过对象在划分间移动来改进聚类效果。由于这类方法适用于发现大小相近的球状簇,故常用在设施选址等应用中。

②基于层次的方法固定数据对象的关系,只是对对象集合进行分解。根据层次的分解方式,这类方法可分为凝聚和分裂两种,Birch, Cure 和 Chameleon 是上述方法的改进。

③基于密度的方法的主要思想是:对给定类中的每个数据点,在一个给定范围的区域中必须包含超过

某个阈值的数据点,才继续聚类。它可以用来发现任意形状的簇,过滤“噪声”。代表性的方法有:DBscan, Optics, Denclue。

④基于栅格的方法把对象空间化为有限数据的单元,形成一个网格结构。该方法处理速度快,处理时间独立于数据对象的数目。该方法包括:Sting, Sting+, Wave Cluster 和 Clique^[5]。

(2) 空间分类方法。是指分析空间对象导出与一定空间特征有关的分类模式。空间分类的目的是在空间数据库对象的空间属性和非空间属性之间发现分类规则,是近些年来空间数据挖掘领域中比较活跃的一个方向。

在空间分类领域中常常使用决策树方法。Fayyad^[8]等人使用决策树方法对星形结构对象的图像进行分类,从而探测行星与银河系。他们的方法是使用 Focas 系统为选中的对象,例如天空图像,生成区域、方向等的基本属性。训练集中的对象由宇航员来分类。基于这些分类,构成用于决策树算法的十个训练集。决策树是通过学习算法得到的。最后,由决策树生成一个健壮、通用、正确的最小分类规则集合。该方法处理的是图像数据库,并应用于天文研究领域。但它却不善于处理常用于 GIS 中的向量数据格式。

Ester^[9]等人在邻接图理论的基础之上提出了一个基于 ID3 算法的空间分类算法。该算法不仅考虑了被分类对象的非空间属性,同时也考虑了其邻接对象的非空间属性。只要是满足任何一种邻接关系的对象都会被看成是邻接对象。但是,该算法不具备分析邻接对象非空间属性聚合值和进行相关分析的能力,并且没有考虑到非空间属性和空间属性之间可能存在的概念层次关系。因此, Koperski^[10]提出了一个高效的两步分类算法:第一步,通过较少代价的空间计算获得一个近似的空间谓词,在这个阶段同时进行相关分析;第二步,对模型进行进一步的精化计算,从而获得一个更小、更精确的决策树。

(3) 空间关联规则^[11]。是对传统数据挖掘中关联规则的扩展。空间关联规则是指空间邻接图中对象之间的关联。空间关联规则形如: $A \rightarrow B[s\%, c\%]$, A 和 B 是空间和非空间谓词的集合, $s\%$ 表示规则的支持度, $c\%$ 表示规则的可信度。由于空间关联规则的挖掘需要在大量的空间对象之间计算多种空间关系,因此,空间关联挖掘多采用逐步求精的优化思想,即首先用一种快速的算法粗略地对初始空间数据库进行一次挖掘,然后再在裁剪过的数据库上进行进一步的挖掘。

根据上述思想, Koperski^[2,3]提出了一个 5 步骤的算法:第一步,通过空间查询从初始空间数据库中获得

和任务相关的空间数据库;第二步,使用一些有效空间挖掘算法计算对象之间的空间连接,从而获得一个候选谓词集合;第三步,对第二步中所得到的谓词集合中的每一个谓词计算其支持度,并且将那些支持度小于最小支持度的谓词删除;第四步,对谓词集合进行进一步精化以确定准确的空间关系;第五步,以第四步所得的候选集作为输入,生成空间关联规则。

(4)空间趋势分析。空间趋势^[10]指离开一个给定的起始空间对象时,非空间属性的变化情况。例如,当离城市中心越来越远时经济形势的变化趋势。其分析结果可能是正向趋势、反向趋势、或者没有趋势。一般在空间数据结构和空间访问方法之上分析空间趋势,需要使用回归和相关的分析方法。由于空间对象自身的特殊性,传统的回归模型可能并不合适。例如,传统的线性回归模型($y = X\beta + \epsilon$)对空间对象就不适用,需要使用空间自回归 SAR 模型: $y = \rho W y + X\beta + \epsilon$ 。

在实际应用中,常常要综合运用上述方法。另外,空间数据挖掘方法要与常规的数据库技术充分结合,数据挖掘利用的技术越多,结果的精确性越高。

4 结束语

数据挖掘通过归纳推理的方法发现和建立模型并预测未来。传统的数据挖掘是针对关系和事务型数据,以数据独立作为挖掘前提,而空间数据挖掘不同,研究的对象可能要受到邻近对象的影响。空间数据库中的空间数据对象保留了各个对象间的空间关系,如空间对象间的拓扑结构、距离和方向等信息,空间数据

挖掘的算法中体现了这种关联特性,如聚类、分类、空间关联等数据挖掘方法。

参考文献:

- [1] Shekhar S, Chawla S. Spatial Databases[M]. 北京:机械工业出版社,2004.
- [2] 邱凯昌. 空间数据挖掘和知识发现的理论与方法[D]. 武汉:武汉大学,1999.
- [3] 李德仁,王树良,史文中,等. 论空间数据挖掘和知识发现[J]. 武汉大学学报:信息科学版,2001,26(6):491-499.
- [4] 毛克彪,田庆久. 空间数据挖掘技术方法及应用[J]. 遥感技术与应用,2004,17(4):199-204.
- [5] 王海起,王劲峰. 空间数据挖掘技术研究进展[J]. 地理与地理信息科学,2005,21(4):33-38.
- [6] 吴信才,刘少雄. 基于邻接关系的空间数据挖掘[J]. 计算机工程,2003,28(7):89-91.
- [7] 毛克彪,覃志豪. 空间数据与 GIS 集成及应用研究[J]. 测绘与空间地理信息,2004,27(7):14-17.
- [8] Fayyad. Advances in Knowledge Discovery and Data Mining[M]. Menlopark, CA: AAAI/MIT Press, 1996.
- [9] Ester M, Frommelt A, Kriegel H P. Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support[J]. Data Mining and Knowledge Discovery, 2000, 18(4): 193-216.
- [10] Ester M, Kriegel H P, Sander J. Spatial Data Mining: a Database Approach[C]//In: Scholl M V eds. Proceedings of The 5th International Symposium on Spatial Databases. Berlin: Springer-Verlag, 1997.
- [11] Han Jiawei. Data Mining: Concepts and Techniques[M]. 北京:机械工业出版社,2004.

(上接第 127 页)

- RYPT'2000 - Advances In Cryptology. Kyoto, Japan: International Association for Cryptologic Research, 2000: 443-457.
- [3] Wang C, Hill J, Knight J, et al. Software tamper resistance: Obstructing static analysis of programs[R]. Technical Report CS-2000-12. Virginia: Department of Computer Science, University of Virginia, 2000.
- [4] Low D. Java Control Flow Obfuscation[D]. Auckland: Department of Computer Science, University of Auckland, 1998.
- [5] Linn C, Debray S. Obfuscation of Executable Code to Improve Resistance to Static Disassembly[C]//In 10th ACM Conference on Computer and Communications Security(CCS). Washington DC: [s. n.], 2003: 290-299.
- [6] 段钢,王勇. 软件加密技术内幕[M]. 北京:电子工业出版社,2004.
- [7] Schwarz B, Debray S K, Andrew G R. Disassembly of Executable Code Revisited[C]//In Proc. IEEE 2002 Working

- Conference on Reverse Engineering(WCRE). [s. l.]: IEEE Computer Society, 2002: 45-54.
- [8] Theiling H. Extracting safe and precise control flow from binaries[C]//In Proc. 7th Conference on Real-Time Computing Systems and Applications (RTCSA). Cheju Island, South Korea: [s. n.], 2000: 23-30.
- [9] Szor P. The Art of Computer Virus Research and Defense[M]. Boston, USA: Addison-Wesley Professional, 2005.
- [10] Wroblewski G R. General Method of Program Code Obfuscation[C]//Proceedings of the International Conference on Software Engineering Research and Practice (SERP). Las Vegas, USA: [s. n.], 2002: 56-75.
- [11] Collberg C, Thomborson C, Low D. A Taxonomy of Obfuscation Transformations[R]. Technical Report # 148. New Zealand: Department of Computer Science, University of Auckland, 1997.