中/青/文/库

基于数据发布的隐私保护模型研究

刘英华◎著

中国社会科学出版社

中/青/文/库

本书得到中国青年政治学院出版基金资助

基于数据发布的隐私保护模型研究

刘英华◎著

中国社会科学出版社

图书在版编目(CIP)数据

基于数据发布的隐私保护模型研究/刘英华著. —北京:
中国社会科学出版社, 2015.9
ISBN 978-7-5161-6291-0

Ⅰ.①基… Ⅱ.①刘… Ⅲ.①计算机网络—隐私权—
安全技术 Ⅳ.①TP393.08

中国版本图书馆 CIP 数据核字(2015)第 131060 号

出 版 人 赵剑英
责任编辑 李炳青
责任校对 王 影
责任印制 李寡寡

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http://www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印刷装订 北京金瀑印刷有限责任公司
版 次 2015 年 9 月第 1 版
印 次 2015 年 9 月第 1 次印刷

开 本 710×1000 1/16
印 张 8.25
字 数 140 千字
定 价 29.00 元

凡购买中国社会科学出版社图书,如有质量问题请与本社营销中心联系调换
电话:010-84083683

《中青文库》编辑说明

中国青年政治学院是在中央团校基础上于1985年12月成立的，是共青团中央直属的唯一一所普通高等学校，由教育部和共青团中央共建。中国青年政治学院成立以来，坚持“质量立校、特色兴校”的办学思想，艰苦奋斗、开拓创新，教育质量和办学水平不断提高。学校是教育部批准的国家大学生文化素质教育基地，中华全国青年联合会和国际劳工组织命名的大学生KAB创业教育基地。学校与中央编译局共建青年政治人才培养研究基地，与北京市共建社会工作人才发展研究院和青少年生命教育基地。

目前，学校已建立起包括本科教育、研究生教育、留学生教育、继续教育和团干部培训等在内的多形式、多层次的教育格局。设有中国马克思主义学院、青少年工作系、社会工作学院、法律系、经济系、新闻与传播系、公共管理系、中国语言文学系、外国语言文学系等9个教学院系，文化基础部、外语教学研究中心、计算机教学与应用中心、体育教学中心等4个教学中心（部），轮训部、继续教育学院、国际教育交流学院等3个教学培训机构。

学校现有专业以人文社会科学为主，涵盖哲学、经济学、法学、文学、管理学5个学科门类。学校设有思想政治教育、法学、社会工作、劳动与社会保障、社会学、经济学、财务管理、国际经济与贸易、新闻学、广播电视学、政治学与行政学、汉语言文学和英语等13个学士学位专业，其中社会工作、思想政治教育、法学、政治学与行政学为教育部特色专业。目前，学校拥有哲学、马克思主义理论、法学、社会学、新闻传播学和应用经济学等6个一级学科硕士授权点和1个专业硕士学位点，同时设有青少年研究院、中国马克思主义研究中心、中国志愿服

务信息资料研究中心、大学生发展研究中心、大学生素质拓展研究中心等科研机构。

在学校的跨越式发展中，科研工作一直作为体现学校质量和特色的重要内容而被予以高度重视。2002 年，学校制定了教师学术著作出版基金资助条例，旨在鼓励教师的个性化研究与著述，更期之以兼具人文精神与思想智慧的精品的涌现。出版基金创设之初，有学术丛书和学术译丛两个系列，意在开掘本校资源与移译域外菁华。随着年轻教师的剧增和学校科研支持力度的加大，2007 年又增设了博士论文文库系列，用以鼓励新人，成就学术。三个系列共同构成了对教师学术研究成果的多层次支持体系。

十几年来，学校共资助教师出版学术著作百余部，内容涉及哲学、政治学、法学、社会学、经济学、文学艺术、历史学、管理学、新闻与传播等学科。学校资助出版的初具规模，激励了教师的科研热情，活跃了校内的学术气氛，也获得了很好的社会影响。在特色化办学日益成为当下各高校发展之路的共识中，2010 年，校学术委员会将遴选出的一批学术著作，辑为《中青文库》，予以资助出版。《中青文库》第一批（15 本）、第二批（6 本）、第三批（6 本）出版后，有效展示了学校的科研水平和实力，在学术界和社会上产生了很好的反响。本辑作为第四批共推出 12 本著作，并希冀通过这项工作的陆续展开而更加突出学校特色，形成自身的学术风格与学术品牌。

在《中青文库》的编辑、审校过程中，中国社会科学出版社的编辑人员认真负责，用力颇勤，在此一并予以感谢！

目　录

第一章　引言

随着计算机软件、硬件技术及互联网技术的迅速发展，各个领域已经积累了海量数据库，并且数据库的规模正以惊人的速率增长。随着知识发现与机器学习在诸多领域的深度应用和广度拓展，隐私保护数据挖掘（privacy－preserving data mining）已经成为知识发现领域的一个核心问题，特别是在国防、反恐、情报、社会网络分析、普适计算、语义 Web、病毒营销、医学、社会学等领域，隐私保护数据挖掘已经成为涉及每个国家、每个公司、每个部门、每位公民的首要问题。

数据挖掘是知识发现的核心步骤。数据挖掘就是从海量数据库中分析并发现未知的、潜在的、有价值的知识，它融合了人工智能、数据库、模式识别、统计学和数据可视化等多个领域的理论和技术。[1]随着数据挖掘技术在人类生活和社会生产中的广泛应用，带来了巨大的经济效益和社会效益。但是伴随着数据挖掘技术的逐步提高，功能越来越强大的数据挖掘工具的开发和使用，人们越来越关注数据挖掘可能导致的隐私泄露问题。传统数据挖掘技术是针对原始数据库进行数据挖掘，对数据挖掘者而言，原始数据库是完全裸露的，真实的数据完整地放在数据挖掘者面前。如银行客户数据挖掘者可以从原始数据库中获得客户的身份证号码，医疗病患数据挖掘者可以从原始数据库中得知患者的病情。使用传统数据挖掘技术可以通过一个或多个原始数据库挖掘出数据拥有者不希望他人得知的知识，如银行客户数据挖掘者可以挖掘出客户定期的消费习惯，数据挖掘者通过医疗病患数据库和上市公司数据库可以挖掘出某上市公司的重要人物患有不可治愈的疾病，若此信息被披露有可能对上市公司造成不必要的损失。随着传统数据挖掘技术的广泛使用，越来越多的传统数据挖掘技术在为某些客户提供未知的、潜在的、有价值的知识的同时，也造成了越来越多的隐私泄露。

数据发布是数据挖掘的基础和前提。数据发布是数据的拥有者将数据展示给用户（当然也包括数据挖掘者），随着信息化技术的快速发展，各种信息被不同的政府部门、研究团体、商业机构等广泛地获取、发布和分析，丰富的数据资源中难免会涉及隐私数据信息。如果数据的拥有者在发布数据时，不对数据采取任何保护措施，数据挖掘者可以获取完整的原始数据库后对其进行数据挖掘，则很有可能造成隐私泄露。如某上市公司将财务年报完整地、直接地展示给用户，数据挖掘者很可能挖掘出此上市公司的竞争对手感兴趣的知识。文献［2－4］研究表明，通过邮编、性别、出生日期三个属性对未采取任何保护措施直接发布的选民登记表和医疗信息表进行数据挖掘，可造成87%的美国公民身份被唯一标识。

隐私保护数据挖掘 PPDM（privacy－preserving data mining）就是在保证数据挖掘的同时尽可能地保护隐私数据，在不能完整或精确地访问原始数据的条件下，得到准确或基本准确的模型和分析结果。隐私保护数据挖掘作为数据挖掘领域中一个崭新、极其重要又富有挑战性的课题，其最终目标是实现隐私数据及敏感规则的保护又能得到准确的知识挖掘。

自 1995 年加拿大蒙特利尔市召开了第一次 KDD 国际学术会议并将“隐私保护数据挖掘”作为一个专门的研究主题开始，隐私保护数据挖掘就成为数据挖掘领域的研究重点之一。经过二十年的发展，隐私保护数据挖掘逐步得到各个国家各行各业的重视，并成为数据挖掘领域的研究热点。其主要的研究方向是：

1）数据发布中的隐私保护。针对发布的具体数据，采用不同的技术发布部分失真的数据、发布加密后的数据或阻止部分数据的发布，以达到隐私保护的效果。

a）数据扰动技术。数据扰动是采用数据交换、添加噪声等方法扰动原始数据，使敏感数据失真但能保证通过数据挖掘工具挖掘出的知识真实有效，如随机扰动、随机回答、阻塞和凝聚等算法。

b）数据加密技术。数据加密是对原始数据加密以保护隐私，如安全多方技术（SMC，Secure Multiparty Computation）。

c）查询限制技术。查询限制是通过限制数据的查询，避免数据挖掘者获取完整原始数据的方法，实现隐私保护，如通过抑制、泛化、数据抽样、数据划分等原则匿名化数据。

2）数据挖掘中的隐私保护。针对特定挖掘任务和数据的分布方式，

采用不同的技术实现在数据挖掘过程中的隐私保护。

a）针对特定挖掘任务。针对挖掘任务的不同研究相应的隐私保护数据挖掘算法。现在的数据挖掘技术可以解决关联规则挖掘、分类挖掘、聚类挖掘、数据流挖掘、序列挖掘、图挖掘、社会网络分析挖掘、多关系挖掘、对象挖掘、Web 数据挖掘、多媒体挖掘、空间挖掘等。不同的挖掘任务使用的隐私保护数据挖掘技术和算法差异很大，目前还没有通用的隐私保护数据挖掘算法。

b）针对数据分布方式。数据分布方式有两种：集中式和分布式。针对集中式数据分布，主要的隐私保护技术有启发式和重建式两种；针对分布式数据分布，主要的隐私保护技术是加密技术。隐私保护数据挖掘中的关联规则主要采用 Apriori 方法确定频繁项集；分类方法主要有决策树归纳分类、贝叶斯分类、支持向量机分类（SVM，Support Vector Machine）、k－最近邻分类、粗糙集分类和模糊集分类等；聚类方法主要有划分方法，例如 k 均值、k 中心点、期望最大化（EM，Expectation Maximization）等方法；层次方法，包含凝聚层次聚类和分裂层次聚类两种，例如 Birch 方法；基于密度的方法，如 DBSCAN 方法等。

c）隐私保护的度量。不存在某种隐私保护数据挖掘算法，在任何方面都胜过其他算法，优异算法是指在某些特定的方面优于其他的已有方法，如在性能或数据实用性等方面。度量一种隐私保护数据挖掘算法的优劣，主要的评测指标有：算法性能（主要是时间和空间代价）；算法效能（主要是隐私保护效果和信息遗失来衡量）；算法适用性（主要是适用不同类型的隐私、不同类型的数据、不同应用背景下隐私的保护能力）。

本课题的研究背景是北京科技大学知识工程研究所承担的国家自然科学基金面上资助项目“基于大规模复杂结构知识库的知识发现机理、模型与算法研究”（项目编号：60875029）和国家自然科学基金面上资助项目“基于多关系的模糊认知图挖掘模型、算法与评价机制研究”（项目编号：61175048）。

本书的文章组织结构如下：

第一章　引言。

第二章　引述本书研究的宏观背景。分析了 KDTICM 理论、隐私保护和数据挖掘，安全多方计算技术、数据匿名化技术、数据扰动技术的发展历程。

第三章　首先，总结了聚类隐私保护挖掘的研究历史沿革和现状，分析各种现存模型的特点及存在的问题，以此为基础提出了笔者解决问题的思路。其次，分析了针对数据加密技术，分布式聚类隐私保护挖掘模型。探讨了水平分布式聚类数据挖掘算法中隐私保护的可行性，在设计全新的完全同态加密算法的基础上，针对水平划分方法提出 FHE－DK－MEANS 模型和 FHE－DBIRCH 模型。最后，针对两个模型进行了实验分析、比较。

第四章　首先，总结了数据发布中隐私保护挖掘的研究历史沿革和现状，分析各种现存模型的优缺点及存在的问题，并在此基础上提出了个性化（$\alpha_{[s]},\ell$）－多样 k－匿名模型和个性化并行（$alpha_{[s]}$，k）－匿名模型。从理论证明和实验分析两个角度证明了这两种模型在数据发布中个性化隐私保护的优势。

第五章　首先，总结了前人的工作，发现了随着数据发布中隐私保护要求逐渐提高，现存模型无法满足要求的问题，分析了现存模型的特点，利用数据库系统中的有损分解思想，提出（$\alpha_{[s]}$，ℓ）－多样 k－匿名有损分解模型。从理论证明和实验分析两个角度证明了这种模型在数据发布中隐私保护的优势。

第六章　总结本书的研究成果并展望进一步的工作。

第二章　文献综述

随着计算机软件、硬件技术及互联网技术的迅速发展，各领域的海量数据库规模以惊人的速度增长。随着数据发布和数据挖掘在诸多领域的深度应用和广度拓展，隐私保护数据挖掘已经成为知识发现领域的一个核心问题，特别是在国防、反恐、情报、社会网络分析、普适计算、语义Web、病毒营销、医学、社会学等领域，隐私保护数据挖掘已经成为涉及国家和个人利益的首要问题。

本书研究的是基于数据发布的隐私保护模型。本书研究的内容与KDTICM理论、隐私保护、数据挖掘、数据发布技术密切相关。本章首先概述了KDTICM理论、隐私保护、数据挖掘、数据发布技术，以描述所研究内容的宏观领域；其次分析了数据发布领域及隐私保护数据挖掘领域的发展历程，以奠定本书工作的历史脉络；最后，分析了数据发布的隐私保护数据挖掘存在的问题，以阐明本书工作的意义。

第一节　KDTICM理论

1989年，在底特律第11届国际人工智能联合会议上提出了知识发现（KDD：Knowledge Discovery in Databases）的概念，Fayyad给出的定义是“从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程”。知识发现是一个受到来自各种不同领域的研究者关注的新兴、交叉、边缘学科领域。

1997年杨炳儒教授另辟蹊径，开创了研究KDD的新思路：从内在的认知机理考虑研究知识发现。[5-8]其核心思想是：把知识发现的过程视为一个认知过程，把知识发现的系统看作是一个认知的系统，用系统论和认知科学的思想和方法来研究复杂的知识发现过程。首次发现了知识发现系

统内在认知机理涵盖的三个机制（原理）：双库协同机制、双基融合机制和信息扩张机制，并分别给出其核心定理及其实现技术，此项研究揭开了国际上研究 KDD 的新方向。

随后杨炳儒又提出了基于知识库的知识发现（Knowledge Discovery in Knowledgebase，KDK）。基于知识库中的事实，首先基于归纳学习的方法生成归纳假设，然后使用卡尔纳普的归纳逻辑进行假设的验证与评价。针对知识库中的规则，首先采用广义概念格方法产生归纳假设，然后使用柯恩的归纳逻辑进行假设的验证与评价。

综合上述成果，2002 年杨炳儒教授在国际上首次提出了“基于内在认知机理的知识发现 KDTICM 理论（Knowledge Discovery Theory based on Inner Cognition Mechanism）”。KDTICM 的理论支柱是三个基本机制（原理），由基础理论层、内在认知机理层、过程模型层、技术方法层、智能系统层五个彼此相互联系的层面构成。[9-21]

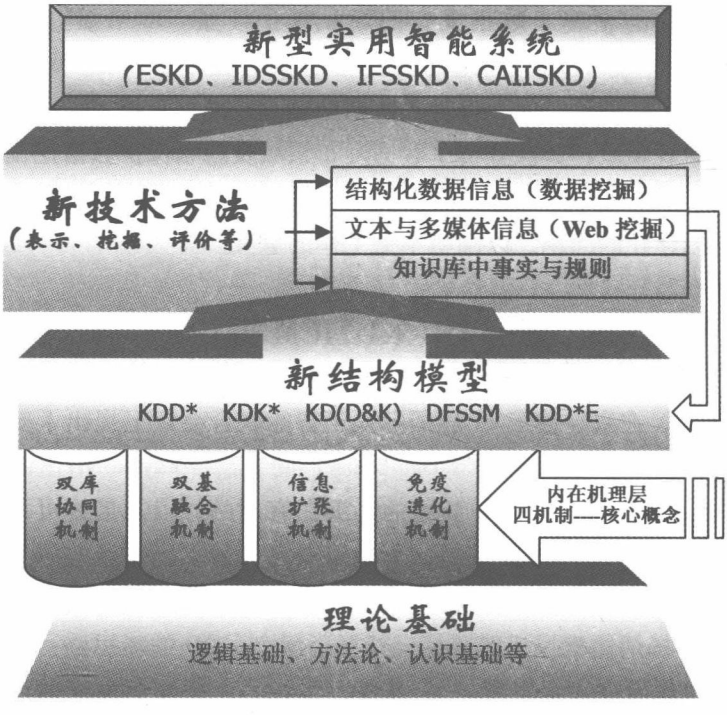

图 2-1 KDTICM 理论的总体结构图

第一层面（理论基础模块层）。包含了多个基础理论的研究成果。提出了多个层次的结构逻辑、广义归纳的逻辑因果；因果关系的判定方法、因果关系进行的定性推理的模型和方法；基于专家知识的归纳及获取；语言场与语言值结构的知识表示方法等。

第二层面（内在机理模块层）。由基于双库的协同机制、基于双基的融合机制、基于信息的扩张机制和基于免疫的进化机制四部分构成。包括：启发和协调维护算法、结构对应定理、可达关系概率估计定理；RST三类协调算法、过程模型逻辑等价定理；参数演化定理、不动点原理、突变性原理等。

第三层面（结构模型模块层）。包含内在机理的研究推导出来的多个新型结构模型。包括：KDD*过程模型（基于双库的协同机制的知识发现过程，用于处理结构化数据挖掘问题）；KDK*过程模型（基于双基的融合机制的知识发现过程）；KD（D&K）过程模型（基于双库的协同机制、双基的融合机制的具有全新特征的知识发现新系统，强调了知识发现过程的认知自主性）；DKD（D&K）过程模型［强调了分布式的KD（D&K）过程］；KDD*E过程模型（结合KDD*过程模型和信息扩张机制）；发现特征子空间模型DFSSM（用于复杂类型数据挖掘）；基于DFSSM的图像挖掘过程模型IMDFSSM。

第四层面（技术方法模块层）。由内在认知机理和新过程模型派生的多个新型技术、方法组成。包括：Maradbcm算法（应用于关联规则的挖掘算法）；源于KDD*的自动型评价、评测方法（应用于因果型关联规则）；基于Web的文本型挖掘算法（基于DFSSM模型）；基于相似模式的图像信息挖掘算法IARMA；混沌模式型挖掘算法；KDK*归纳挖掘算法（基于知识库中的事实与知识库中的规则）；多关系数据挖掘算法等。

第五层面（智能系统模块层）。包含新型结构模型及其相应的技术、方法的新型实用智能系统，它应用于现实的实用系统诱导出的多个基于内在型机理的研究。其中包括：知识发现理论中的专家型系统（Expert System based on Knowledge Discobery，ESKD）；知识发现理论中的智能决策型支持系统（Intelligent Decision Support System based on Knowledge Discovery，IDSSKD）；知识发现理论中的智能型支持预测系统（Intelligent Forecast Support System based on Knowledge Discovery，IFSSKD）；基于知识发现理论中的智能型计算机辅助设计、创新系统（Computer Aided Innovation In-

telligence System based on Knowledge Discovery，CAIISKD）等。

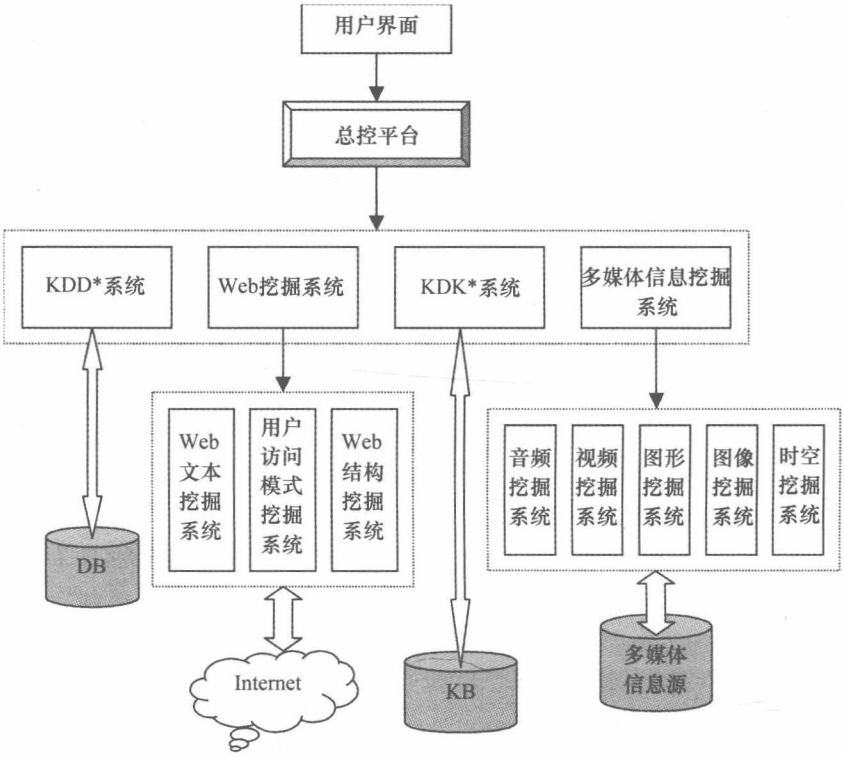

图 2－2　ICCKDSS 总体结构图

基于 KDTICM 理论研发出具有自主知识产权的集成化、组合构件式的大型知识发现软件系统 ICCKDSS。ICCKDSS 具有多个模块，可以根据需要选取、集成模块，且设计了完善的接口，可以单独系统集成到其他系统，方便功能的扩展和构件的重用。ICCKDSS 系统可以面向广泛的信息挖掘问题，如结构化数据挖掘、知识库挖掘、Web 挖掘、多媒体信息挖掘（图像、音频、视频、空间数据等）。

KDTICM 理论与软件系统 ICCKDSS 已在农业、铝电解、现代远程教育网、气象、国际商务等诸多领域中应用，通过软件系统验证理论，解决了大量应用领域中难以解决或尚未解决的典型问题。

笔者从事的专题研究，试图扩充已经构建的 KDTICM 理论的第四层面

（技术方法层）。在原有成果的基础上，借鉴其中一些有益的模型、方法的核心思想，研究基于数据发布的隐私保护模型。

第二节　隐私保护

隐私是个人、团体、国家等实体不想被其他个人、团体、国家等实体了解的信息。

一　隐私保护的定义

随着计算机处理能力、存储技术及互联网的快速发展，使得信息的数据化加快。针对数据化信息的隐私保护即保护数据的发布者不希望泄露的敏感信息。

敏感信息包括个人敏感信息和共同敏感信息。[22-23]个人敏感信息是可以确定特定个体或与确定特定个体相关的信息，如个人身份证号码、个人手机号码、个人住址等，也称为敏感数据。共同敏感信息是多个个体共同表现出来的，不想被其他个人、团体、国家等实体了解的信息，如某部门的平均工资、薪酬分布等，也称为敏感规则。

二　隐私的度量

隐私的度量是使用风险泄漏（Disclousure Risk）来描述的。风险泄漏表示数据挖掘者根据发布的数据并综合背景知识可能造成的隐私泄漏概率。背景知识（Background Knowledge）是数据挖掘者通过多种渠道、方式获取的与发布数据相关的信息和数据，在这些数据的配合下进行挖掘可能造成隐私泄露，由于网络的迅猛发展和数据发布的日益普及，背景知识的获取将更为容易和完整。[24-25]

定义2.1　设s表示敏感数据或敏感规则，S_k表示数据挖掘者在综合背景知识K后进行数据挖掘操作后泄漏s，泄漏风险r（s，K）表示如下[24]：

$$r(s, K) = P_r(S_k)$$

数据发布者发布数据集D，若所有敏感数据或敏感规则s的泄漏风险均小于阈值α（$\alpha \in [0, 1]$），则称数据集D的泄漏风险是α。若$\alpha = 1$，则数据集D是直接发布的数据，没有做任何的隐私保护处理，这样的数