中国高校
创意创新创业
教育系列丛书

李　涛 主编

网络安全中的数据挖掘技术

清华大学出版社

中国高校创意创新创业教育系列丛书

李　涛 主编

网络安全中的数据挖掘技术

清华大学出版社
北 京

内 容 简 介

本书以网络安全中主要子领域为主线，以数据挖掘算法为基础，搜集了大量基于数据挖掘的网络安全技术研究成果，汇编了数据挖掘技术在隐私保护、恶意软件检测、入侵检测、日志分析、网络流量分析、网络安全态势评估、数字取证等网络安全领域的应用，介绍了常用的网络安全数据集，并搜集了大量的网络安全资源，以供读者能将本书内容应用于实际的研究或学习中。

本书可作为研究人员、网络安全工程人员和对基于数据挖掘的网络安全技术感兴趣的研究生的参考书，也可作为高等院校高年级课程的教学用书，还可供相关领域工作的读者参考。

图书在版编目(CIP)数据

网络安全中的数据挖掘技术/李涛主编. —北京：清华大学出版社，2017
(中国高校创意创新创业教育系列丛书)
ISBN 978-7-302-45550-9

Ⅰ. ①网…　Ⅱ. ①李…　Ⅲ. ①数据采集—网络安全—研究生—教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 277489 号

责任编辑：谢　琛　薛　阳
封面设计：常雪影
责任校对：焦丽丽
责任印制：李红英

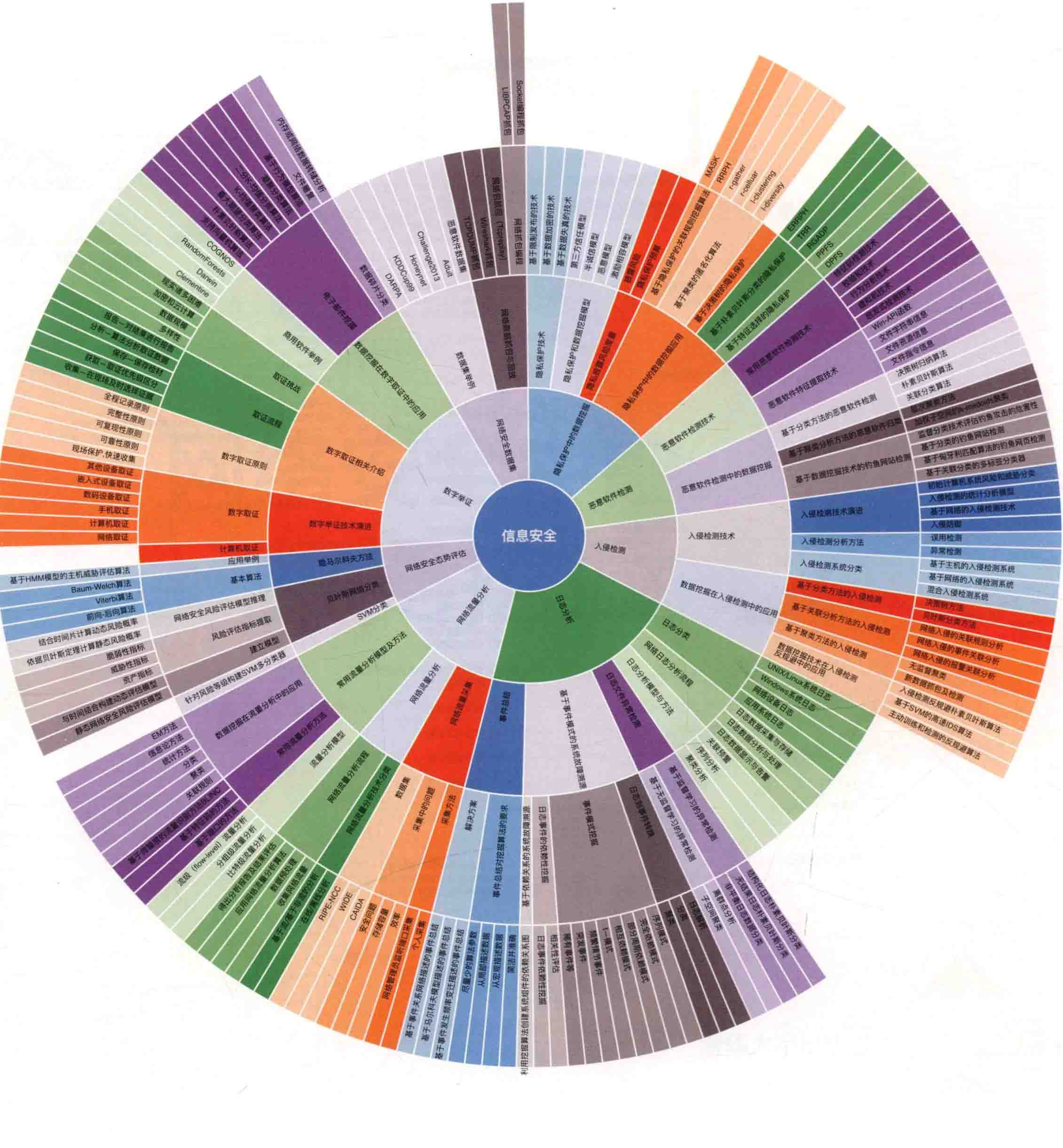
出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课　件　下　载：http://www.tup.com.cn，010-62795954

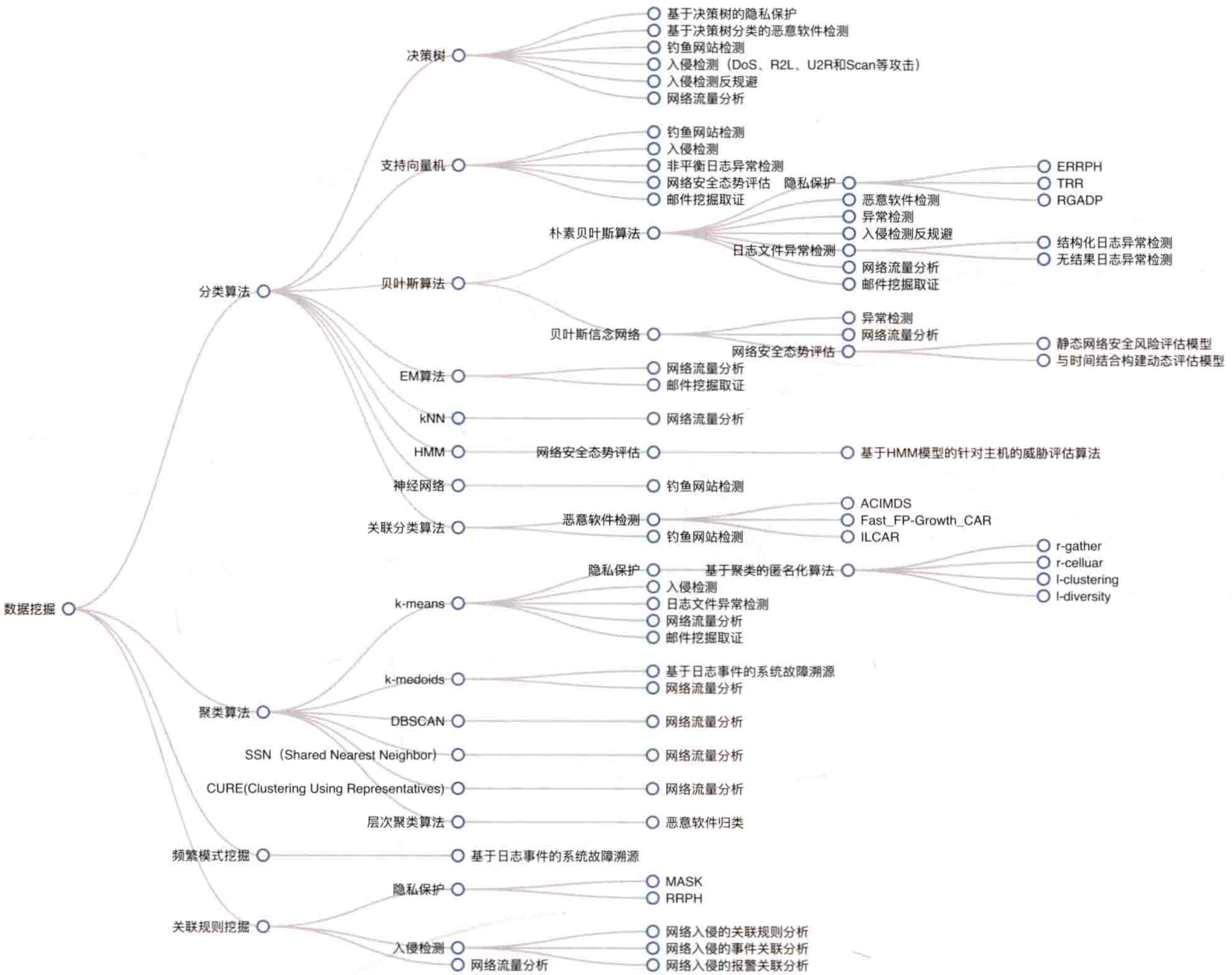
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：210mm×235mm　　**印　　张**：21.5　　**彩　　插**：1　　**字　　数**：447 千字
版　　次：2017 年 8 月第 1 版　　**印　　次**：2017 年 8 月第 1 次印刷
印　　数：1～2000
定　　价：69.00 元

产品编号：069416-01

作者介绍

李涛，美国佛罗里达国际大学计算机学院/南京邮电大学计算机学院教授。研究兴趣主要包括数据挖掘、机器学习和信息检索及生物信息学等领域，并在这些领域开展了一系列有相当影响力的理论与实证研究，取得了突出的成就。在基于矩阵方法的数据挖掘和学习、音乐信息检索、系统日志数据挖掘以及数据挖掘的各种应用等方面做出了具有开创性和前瞻性的研究。由于在数据挖掘及应用领域做出了成效显著的研究工作，李涛教授曾多次获得各种荣誉和奖励，其中包括2006年美国国家自然科学基金委颁发的杰出青年教授奖（NSF CAREER Award, 2006－2010）；2010年IBM大规模数据分析创新奖（Scalable Data Analytics Innovation Award, 2010）；多次获得IBM学院研究奖（2005、2007、2008）；2009年获得佛罗里达国际大学科研最高荣誉——最高学术研究奖；多次获得施乐公司学院研究奖（2011－2014）；并于2011年获得佛罗里达国际大学工程学院首位杰出导师奖（该奖2011年初次设立），2014年再获此殊荣。




数据挖掘
分类算法
决策树
基于决策树的隐私保护
基于决策树分类的恶意软件检测
钓鱼网站检测
入侵检测（DoS、R2L、U2R和Scan等攻击）
入侵检测反规避
网络流量分析
支持向量机
钓鱼网站检测
入侵检测
非平衡日志异常检测
网络安全态势评估
邮件挖掘取证
贝叶斯算法
朴素贝叶斯算法
隐私保护
ERRPH
TRR
RGADP
恶意软件检测
异常检测
入侵检测反规避
日志文件异常检测
结构化日志异常检测
无结果日志异常检测
网络流量分析
邮件挖掘取证
贝叶斯信念网络
异常检测
网络流量分析
网络安全态势评估
静态网络安全风险评估模型
与时间结合构建动态评估模型
EM算法
网络流量分析
邮件挖掘取证
kNN
网络流量分析
HMM
网络安全态势评估
基于HMM模型的针对主机的威胁评估算法
神经网络
钓鱼网站检测
关联分类算法
恶意软件检测
ACIMDS
Fast_FP-Growth_CAR
ILCAR
钓鱼网站检测
聚类算法
k-means
隐私保护
基于聚类的匿名化算法
r-gather
r-celluar
l-clustering
l-diversity
入侵检测
日志文件异常检测
网络流量分析
邮件挖掘取证
k-medoids
基于日志事件的系统故障溯源
网络流量分析
DBSCAN
网络流量分析
SSN（Shared Nearest Neighbor）
网络流量分析
CURE(Clustering Using Representatives)
网络流量分析
层次聚类算法
恶意软件归类
频繁模式挖掘
基于日志事件的系统故障溯源
关联规则挖掘
隐私保护
MASK
RRPH
入侵检测
网络入侵的关联规则分析
网络入侵的事件关联分析
网络入侵的报警关联分析
网络流量分析

前　言

1. 背景

网络安全事关国家安全，它已被多个国家纳入国家安全战略。在我国，网络安全已得到政府的高度重视，国家层面明确意识到网络安全对国家安全牵一发而动全身，并将保障网络安全提升至维护网络空间安全。2015 年 6 月经过国务院学位委员会的批准，网络空间安全也成为工学门类下的一级学科。

与此同时，危害网络安全的新手段正不断涌现，导致网络安全威胁与日俱增，全球的网络安全形势都不容乐观。在这种严峻的网络安全形势大背景下，大量研究人员正不断致力于寻求解决网络安全问题的新技术。而数据挖掘正是能够有效解决网络安全问题的技术之一，各种报道、文献显示，它已成为解决诸多网络安全难题的主力军。数据挖掘是理论技术和实际应用的完美结合，其研究源于真实世界中的实际应用需求，它在许多应用领域都取得了令人瞩目的成绩，利用其解决网络安全问题也顺理成章。在迅猛发展的网络空间中，大量的网络安全难题有待解决，正是这种实际的网络安全应用需求促使研究人员将经典的数据挖掘算法应用于网络安全领域。近年来，基于数据挖掘技术的网络安全研究成果不断出现在各种报道和文献中，这些研究成果在解决网络安全问题方面取得了良好的成效，尤其是在解决大数据背景下的网络安全问题方面，许多数据挖掘技术都凸显出了其解决网络安全恶疾的良好能力。

但是，目前基于数据挖掘的网络安全技术研究成果分散于各种文献中，还没有专门的中文书籍将这些研究成果进行整理汇编，导致从事该领域工作的人员难以综合地、全面地把握其研究进展。

2015—2016 年期间，在国家留学基金委的资助下，广州大学的彭凌西博士、乐山师范学院的杨进博士、刘才铭博士、张建东，以及福建省公安厅的黄君灿先后来到佛罗里达国际大学计算机学院的数据挖掘实验室访学。当不同研究方向的人员在一起讨论研究的时候，为了让数据挖掘的研究人员熟悉网络安全，也为了从事网络安全的研究人员深入了解数据挖掘，我们就萌发了编写此书的想法。

本书正是在以上背景和环境下编写的，其目的就是方便读者阅读或查阅，让读者能够较快地、广泛地掌握基于数据挖掘的网络安全技术。本书主编长期从事数据挖掘研究和教

学工作，在国际数据挖掘领域享有良好的声誉，他经历了数据挖掘技术在网络安全应用研究中的发展历程，对基于数据挖掘的网络安全技术具有深刻的体会。同时，本书编写成员还融合了来自网络安全领域和实际工程领域的研究人员和技术专家，他们编写的内容既涉及理论研究，又反映了大量的实际网络安全应用，全方位覆盖了经典的和最新的研究成果。

2. 主要内容

本书以网络安全中主要子领域为主线，搜集了大量基于数据挖掘的网络安全技术研究成果，这些研究成果既有经典的数据挖掘算法在网络安全中的应用，也有数据挖掘在网络安全热点问题中的最新前沿研究，包括发表于著名国际学术会议的论文。本书以数据挖掘算法为基础，汇编了数据挖掘技术在隐私保护、恶意软件检测、入侵检测、日志分析、网络流量分析、网络安全态势评估、数字取证等网络安全领域的应用，介绍了常用的网络安全数据集，并搜集了大量的网络安全资源，以供读者能将本书内容应用于实际的研究或学习中。

本书的目标群体是研究人员、网络安全工程人员和对基于数据挖掘的网络安全技术感兴趣的研究生，我们希望能够为这些在本领域工作的读者提供全面的参考。本书也可作为高级课程的教科书，能够为学习本领域的学生掌握数据挖掘和网络安全这个交叉领域提供便捷。同时，我们也希望本书能够为不熟悉基于数据挖掘的网络安全技术的读者提供一个好的起点，使得他们能更容易地、快速地、全面地把握数据挖掘技术在网络安全应用研究中的进展。

本书共 9 章，各章的内容介绍如下。

第 1 章介绍了网络安全的概念，包括网络安全的定义、面临的挑战，以及其重要意义。概述了网络空间安全学科相关情况。本章还简要阐述了数据挖掘的定义、作用和特点，介绍了数据挖掘十大算法。能够让读者对数据挖掘算法有宏观上的认识，也叙述了国内外数据挖掘这一领域的发展情况。这些内容为读者学习本书后面章节做了基础知识的铺垫。

第 2 章对数据隐私保护技术的概念及研究现状进行了介绍，并着重介绍了几种主流隐私保护技术及其特点，并列表进行了对比分析，然后介绍隐私保护中数据挖掘应用技术情况，进一步描述了隐私保护和数据挖掘模型，具体包括模型、算法及工作流程。

第 3 章概述了恶意软件及其危害性，介绍了几种数据挖掘技术在恶意软件检测中的应用，包括：分类技术在恶意软件检测中的应用原理；决策树、贝叶斯和关联分类方法在恶意软件检测中的应用实例；层次聚类方法和加权子空间的 K-medoids 聚类方法在恶意软件归类中的应用实例；多标签关联分类方法在钓鱼网站检测中的应用实例。

第 4 章概述了入侵检测的基本概念及其面临的严峻挑战，介绍了几种数据挖掘技术在入侵检测中的应用，包括：决策树分类方法在入侵检测中的应用实例；网络入侵的关联规则分析、事件关联分析和报警关联分析的应用实例；基于无监督聚类的入侵检测算法的应用

实例；入侵检测规避与反规避技术及数据挖掘技术在其中的应用情况。

第5章主要介绍了日志的概念与特点，日志分析的目的；日志文件的分类；日志分析的流程。重点给出了日志分析的模型和方法，根据实际网络管理的需求重点讨论了日志文件的异常检测以及基于事件模型的系统故障溯源。介绍了事件总结，并介绍了三种常见的事件总结方法。最后给出了日志分析重要的研究文献。

第6章介绍流量分析的概念，流量分析的目的、方法及现状，介绍了流量采集的方法，主流的流量分析模型及方法。简单介绍了基于端口的方法，基于特征码的方法，基于传输层的方法，统计特征的流量识别方法等。主要介绍了使用数据挖掘来进行流量分析的相关方法，包括关联规则、聚类和分类。关联规则方法主要介绍了 Apriori 算法；聚类算法主要介绍了 K-均值、K-中心点、DBSCAN、SNN、CURE 等算法；分类算法主要介绍了决策树、KNN、贝叶斯分类等。最后对数据挖掘方法进行了总结，同时给出了经典数据挖掘算法文献的介绍。

第7章阐述了网络安全态势评估相关的概念和重要意义；介绍了几种数据挖掘的常用算法应用到网络态势这一领域，包括 SVM 方法、贝叶斯网络方法、隐马尔思可夫方法等。这些方法在处理不确定信息的智能化系统中已得到了重要的应用，已成功地用于医疗诊断、统计决策、专家系统、学习预测等领域。

第8章首先对数字取证技术的概念及研究现状、面临的挑战进行了介绍，然后介绍数据挖掘技术在数字取证中的应用情况，进一步描述了几种典型应用模型、算法及工作流程，最后对数字取证技术未来的发展研究方向做了展望。

第9章首先从数据类型、属性构成和范例数据等方面对常见的网络安全数据集进行简介，接着介绍了网络数据包常见的抓包与回放工具，最后对网络抓包编程进行了范例描述。

附录部分包括常见的网络安全资源介绍以及相关网站链接等。

3. 其他

本书由李涛教授统筹，其研究团队成员执笔编写，欢迎读者积极反馈。各个章节的作者如下。

简介（杨进、李涛）

基于隐私保护的数据挖掘（彭凌西、李涛）

恶意软件检测（刘才铭、李涛）

入侵检测（刘才铭、李涛）

日志分析（张建东、李涛）

网络流量分析（张建东、李涛）

网络安全态势评估（杨进、李涛）

数字取证(黄君灿、李涛)

网络安全数据集简介及采集(彭凌西、李涛)

除第1章对网络安全和数据挖掘做概要介绍外,剩余的章节系统地介绍了数据挖掘在网络安全各个子领域的应用。这些章节大部分都是自包含的,使得读者能够按照任意的顺序读这些章节。由于同一种数据挖掘算法可以用于多个安全子领域,为了保证各章的独立性,本书在各个安全子领域里对数据挖掘算法的讲述可能有一定的重复。

由于受编者水平和能力所限,没能做到对本书中所涉及的每一个细节都十分精通。此外,推托为客观的因素,我们无法在断断续续的仓促时间内集中精力完成繁多内容的学习理解和组织整理。因此,越是在接近完成本书时,越感诚惶诚恐。所以,恳请读者一旦发现书中不妥或者疏漏之处请给予批评指正,并将意见反馈给我们,那更是求之不得。

网站和联系方式

与本书配套的网站地址为:http://users.cis.fiu.edu/~taoli/security-book。该网站不仅收录了多个相关资源的链接,还提供了一些相关程序和工具供读者下载使用。另外,也欢迎读者将更多的反馈意见和修改建议发邮件到 xiech@tup.tsinghua.edu.cn。

致　　谢

在本书的编写过程中得到了很多专家和朋友的大力支持和帮助。感谢佛罗里达大学计算机学院数据挖掘实验室的博士研究生(刘晓迁、夏彬、倪铭、周武柏、曾春秋、王文韬)和南京理工大学计算机学院李千目教授及硕士研究生(李建妹、王烁、吴丹丹、张文强)认真校对了本书的内容。刘晓迁和夏彬对于本书的目录编排,以及文字、图表格式的调整做了大量的编辑工作。他们认真、细致的工作让我感动,在此谨向他们表示最诚挚的感谢!

另外,本书中涉及的事件挖掘相关的研究项目得到了美国国家自然科学基金(National Science Foundation, NSF)项目(编号 IIS-0546280, CCF-0830659, HRD-0833093, DMS-0915110, CNS-1126619,IIS-1213026 和 CNS-1461926)、中国国家自然科学基金项目(编号 61300053、No. 61003310、No. 61103249、NO. 91646116)、中国博士后科学基金特别资助基金(No. 2012T50783)、四川省教育厅高校科研创新团队基金(No. 13TD0014)、中国博士后科学基金(No. 2011M501419)、广东省普通高校创新团队建设项目(No. 2015KCXTD014)、四川省应用基础研究计划项目(编号: 2014JY0036、2015JY0105),乐山师范学院科研培育计划项目(编号: Z1415,Z1412),江苏省科技支撑计划(社会发展)项目(BE2016776)、教育部-中国移动科研基金项目、江办省"六大人才高峰"创新人才团队项目以及美国国际商业机器公司研究中心(IBM Research)项目、华为技术有限公司研究项目的资助。同时,还得到了南京邮电大学、南京理工大学、厦门大学、厦门理工学院、广州大学、乐山师范学院和美国佛罗里达国际大学计算机学院(School of Computing and Information Sciences,Florida International University)的支持。

李　涛

2017 年 5 月

关于作者

李涛

2004年7月获美国罗彻斯特大学(University of Rochester)计算机科学博士学位。2004年至今先后任美国佛罗里达国际大学(Florida International University,FIU)计算机学院助理教授、副教授(终身教授)、正教授(Full Professor)、研究生主管(Graduate Program Director),博士生导师。2016年入选中组部创新类国家"千人计划"特聘专家,担任南京邮电大学计算机学院院长,南京邮电大学大数据研究院院长。同时他还是厦门大学、南京理工大学、厦门理工大学等国内多所高校的客座教授。长期从事数据挖掘、信息检索、大数据分析等方面的研究工作,在基于矩阵分解的数据挖掘和学习、智能推荐系统、音乐信息检索、系统日志挖掘等研究方向上开展了一系列有相当影响力的理论与实证研究,取得了突出的成就。由于在数据挖掘及应用领域做出了成效显著的研究工作,李涛教授曾多次获得各种荣誉和奖励,其中包括2006年美国国家自然科学基金委颁发的杰出青年教授奖,2010年IBM大规模数据分析创新奖,并于2009年获得佛罗里达国际大学最高学术研究奖。同时,他是数据挖掘和知识发现的国际权威期刊 *ACM Transactions on Knowledge Discovery from Data*(*ACM TKDD*)、*IEEE Transactions on Knowledge and Data Engineering*(*IEEE TKDE*)、*Knowledge and Information System*(*KAIS*)的副主编。

李涛教授在国际著名会议及期刊上已发表250多篇文章。根据Google Scholar的统计,李涛教授的引用指标H-index=55,总引用次数超过10000次。李涛教授已毕业14名博士学生和1名博士后,指导过20多名国家公派的访问学者。毕业的博士生在美国研究型大学(University of West Virginia和Florida Atlantic University)担任教授和博士生导师及在工业界(Microsoft, LinkedIn,Google,Facebook等)担任研发人员。他于2011年获得佛罗里达国际大学工程学院首位杰出导师奖(该奖2011年初次设立),2014年再获此殊荣。

刘才铭

2008年毕业于四川大学,获计算机应用技术博士学位,曾在西南交通大学从事博士后研究工作、美国佛罗里达国际大学(Florida International University)从事访学研究工作,现为乐山师范学院教授。长期从事网络安全方面的研究工作,研究兴趣主要集中于网络入侵检测、恶意软件检测、网络安全风险分析等研究方向,在网络安全领域发表论文五十余篇。

彭凌西

2008 年 6 月毕业于四川大学，获计算机应用专业博士学位，现为广州大学机械与电气工程学院教授，硕士生导师，主要研究方向为人工免疫和网络安全技术，共发表和录用论文超过 70 篇，其中以第一作者或通信作者发表 SCI 或 EI 收录论文三十多篇，主持国家自然科学基金、广东省自然科学基金等科研项目六项，以第一发明人申请并授权国家发明专利两项。

杨进

教授，1980 年 6 月出生，四川大学计算机博士学位，西南交通大学博士后，四川省教育厅科研创新团队带头人。四川省第十一批学术和技术带头人备人选；校级学术骨干；校学术委员会委员。主持国家级项目两项，省部级项目两项，市厅级项目三项。发表了 SCI、EI 检索论文三十多篇。获得“四川省优秀共产党员”“四川省省部级劳模”“乐山市第八批市级拔尖人才称号”“乐山市教师节优秀教师表彰”“市级共产党员示范岗”。获得国家专利九项、三项软件著作权。曾赴瑞典斯德哥尔摩大学学习交流，2015 年以访问学者身份赴美国学习交流。

张建东

2005 年毕业于成都理工大学信息工程学院，获工学硕士。现为乐山师范学院副教授，主要从事网络安全、数据挖掘等方面的研究，主持并参与多项国家级及省部级科研项目。发表论文十余篇。

黄君灿

1998 年毕业于福州大学，获计算数学专业硕士学位，曾在美国佛罗里达国际大学(Florida International University)从事访学研究工作，现为福建省公安厅刑事技术总队声像电子物证室主任、副调研员、高级工程师，全国公安刑事科学技术青年人才、全国刑事技术标准化技术委员会电子物证检验分技术委员，福建省安全技术防范专家、福建省公安厅信息化建设专家。主要从事公安信息化、声像电子物证检验工作，先后主持或参与“鞋类邦样整体级放系统”(获 2000 年福建省科技进步二等奖)、“福建省刑侦综合信息系统”(2002 年获“福建省科技进步二等奖”“公安部科技进步二等奖”)、“福建公安涉案人员信息采集室”(公安部科技进步三等奖)等项目的研发建设。共开展电子物证检验案件二百多起，发表论文近十篇。

目　录