21世纪 高等学校信息安全专业规划教材

网络信息内容安全

杨黎斌 戴航 蔡晓妍 编著

清华大学出版社

21世纪高等学校信息安全专业规划教材

网络信息内容安全

杨黎斌　戴　航　蔡晓妍　编著

清华大学出版社
北京

内 容 简 介

网络信息内容安全技术是信息安全领域的一个研究方向，有着广阔的应用前景。本书共8章，介绍与网络信息内容安全技术相关的基本概念、理论方法和最新研究进展。在编写中着重阐述信息内容安全的若干关键技术——信息过滤、话题检测与跟踪、社会网络分析、网络新闻评价、网络舆情分析等技术。

本书内容全面，既有对网络信息内容安全基础知识和理论模型的介绍，也有对相关问题研究背景、实现方法和技术现状的详细阐述，可作为高等院校计算机、信息技术等相关专业高年级本科生的教材或参考书，也可供从事信息技术、数据挖掘、人工智能、管理科学、战略研究等相关领域研究的教师、研究生和科研工作者参考，借以提供思路和技术支撑。

图书在版编目(CIP)数据

网络信息内容安全/杨黎斌，戴航，蔡晓妍编著. —北京：清华大学出版社，2017
(21世纪高等学校信息安全专业规划教材)
ISBN 978-7-302-45535-6

Ⅰ. ①网… Ⅱ. ①杨… ②戴… ③蔡… Ⅲ. ①计算机网络－信息安全－高等学校－教材 Ⅳ. ①TP393.08

中国版本图书馆CIP数据核字(2016)第277415号

责任编辑：郑寅堃　梅栾芳
封面设计：杨　兮
责任校对：梁　毅
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：185mm×260mm　　**印　张**：13.25　　**字　数**：320千字
版　　次：2017年2月第1版　　**印　次**：2017年2月第1次印刷
印　　数：1～1500
定　　价：29.50元

产品编号：070730-01

前　言

互联网已经成为人们获取信息、相互交流、协同工作的重要途径，但同时也带来一些负面影响，如色情、反动等不良信息在网络上大肆传播，垃圾邮件、广告等恶意营销行为泛滥，网络欺诈、钓鱼以及网络暴力、网络恐怖主义等恶意行为层出不穷。这些恶意信息和行为完全背离了互联网设计的初衷，也不符合广大网民的意愿，并且影响到现实的正常秩序和规范。因此研究网络信息内容安全，提供对互联网中各种不利信息的检测分析能力，是体现我国信息技术水平的重要环节，也是建设信息化社会的坚实保障。

互联网上各种不良信息和行为的产生，其原因主要在于互联网作为一个内容平台，人们可以更便利地获取、发布信息，而在互联网爆发性发展过程中，相关的规范制度、安全技术研究却未能同步发展。网络信息内容安全作为信息安全领域的一个研究分支，是上述问题的解决方案，它主要研究如何从包含海量信息的网络环境中，对特定安全主题的相关信息进行自动获取、识别和分析的技术。该研究分支涉及的相关技术包括信息安全、自然语言处理、网络理论、机器学习、模式识别等，直接或间接应用到这些研究领域的最新研究成果。结合网络信息内容安全的具体需求，本书全面介绍面向信息内容安全的网络信息处理技术的相关基本概念、理论方法和最新研究进展等，着重阐述网络信息内容安全的若干关键技术——信息过滤、话题检测与跟踪、社会网络分析、网络新闻评价、网络舆情分析等。本书既有对基础知识和理论模型的介绍，也有对相关问题研究背景、实现方法和技术现状的详细阐述。目前市面上缺乏对网络信息内容等技术进行系统介绍的书籍，本教材以研究型课程为特征，着重培养学生的思考能力和初步的研究能力，可以让授课教师和学生迅速了解网络信息内容安全的核心技术，同时让大家了解网络信息内容安全的实际应用。本书主要面向高等院校本科生，理论与应用相结合是本书的一大特色。内容编排时兼顾学科前沿研究和实际应用背景。该书有助于发掘学生的科研兴趣、提升学生的就业层次、满足人才市场的需求。

本书由杨黎斌担任主编，各章编写分工如下：杨黎斌编写第1、5章；蔡晓妍编写第2、3章；戴航编写第4章；慕德俊编写第6章；李梅编写第7章，张晓婷编写第8章。杨黎斌负责全书的策划、大纲的制定和统稿工作。

在本书编写过程中，参考了国内外许多公开发表的相关资料，在此对所涉及的各位专家、学者表示诚挚的感谢。研究生许红波、李飞、由文浩和陈志涛同学对本书的图表进行编辑，许晶晶、李敏、张嘉慧、韩亚敏、陈志涛和申昌同学对本书进行校对并提出宝贵的建议，特此表示感谢。由于编写时间紧迫，加之编者理论水平和实践经验有限，书中难免有不当和疏漏之处，恳请广大读者批评指正。

编　者

2016 年 7 月

目　录

第1章 绪 论

1.1 网络信息内容安全的背景

1.1.1 我国互联网发展现状

近几十年来，互联网的迅速发展，不仅促进了全世界范围内信息的有效传播与流通，而且对科学研究、工商行业的发展乃至人们的日常生活方式都带来了深远影响。自20世纪90年代开始，我国的互联网行业也经历了从无到有、从小到大的跨越式发展历程。根据《第37次中国互联网络发展状况统计报告》，截至2015年12月，我国网民规模达到6.88亿，互联网普及率达到50.3%，中国居民上网人数已过半。《报告》同时显示，网民的上网设备正在向手机端集中，手机成为拉动网民规模增长的主要因素。

在信息化已成为世界发展趋势的背景下，互联网有着应用极为广泛、发展规模最大、非常贴近人们生活等众多特点。一方面，互联网创造出巨大的经济效益和社会效益，如新兴的网络公司在互联网上建立业务并迅速发展，传统行业也纷纷将自身的业务和网络应用结合起来，它已经成为人们获取信息、互相交流、协同工作的重要途径；另一方面，互联网也带来一些负面影响，如色情、反动等不良信息在网络上大量传播，垃圾电子邮件等不正当行为泛滥，利用网络传播电影、音乐、软件等的侵犯版权行为，网络欺诈以及网络暴力和网络恐怖主义活动等问题层出不穷，这些行为完全背离了互联网设计的初衷，也不符合广大网络用户的意愿。因此，在建设信息化社会的过程中，提高信息安全保障水平及对互联网中各种不良信息的监测能力，是体现国家信息技术水平的重要一环，也是顺利建设信息化社会的坚实基础。

互联网上各种不良信息的流传和不规范行为的产生，其原因可归结为两类：一类是由于在互联网爆炸性发展过程中相关方面的规范和管理措施未能同步发展导致的。在互联网发展的初期阶段，用户数目很少，且多数用户是从事学术研究的工作人员，网络也没有涉及商业领域的应用，所以网络安全问题并不突出。如今，这种局势已经发生了巨大变化，一些原有的网络模式不再适应现在的发展需求。另一类是由于互联网作为一个新生事物，为人们提供了便利获取与发布信息的新途径，营造出前所未有的思想碰撞场所，相对于传统媒体，互联网上更容易出现一些另类、新奇、不易理解或不符合规范的行为和信息内容。互联网将整个世界变成了“地球村”，聚集了各种思想、观点的人和事物，以及各种形式的信息内容和安全问题，这也是一个长期存在的客观现实。面对这种挑战，人们不应“因噎废食”——因为互联网上存在的一些安全问题和不良信息而变得畏惧或排斥新技术、新事物；应当通过法律与技术等多方面的措施来抵制和消除不良现象，让互联网更好地为人们服务，使得人人都能更高效、更自由地利用互联网信息内容并为之所用，发挥更大的效益。

1.1.2 网络信息内容特点

与传统的信息资源相比，网络信息内容在数量、结构、分布和传播的范围、载体形态、内涵传递手段等方面都显示出新的特点。

1. 存储数字化，传输网络化

信息资源由纸张上的文字变为磁介质上的电磁信号或者光介质上的光信息，存储的信息密度高、容量大。以数字化形式存在的信息，可以通过信息网络进行远距离传送。传统的信息存储载体为纸张、磁带、磁盘。而在网络时代，信息的存在是以网络为载体，这大大提高了网络信息内容的利用与共享程度。

2. 表现形式多样化，内容丰富

网络信息内容包罗万象，覆盖了不同学科、不同领域、不同地域、不同语言的信息资源，还可以以文本、图像、音频、视频、数据库等多种形式存在。信息组织非线性化，超文本、超媒体信息资源成为主要方式。

3. 数量巨大，增长迅速

中国互联网络信息中心(CNNIC)于 2016 年 1 月发布的第 37 次《互联网络发展状况统计报告》全面反映了中国互联网络的发展状况。从该次报告中可以看出，截至 2015 年 12 月 30 日，中国网民规模达到 6.88 亿，网站数量达到 423 万，2015 年网页数量达到 2123 亿，增长迅速。网络信息量之大、增长速度之快、传播范围之广，是其他任何环境下的信息资源所无法比拟的。

4. 传播速度快、范围广，具有交互性

网络环境下，网络信息内容的传递和反馈快速、灵敏。信息内容在网络上的流动非常迅速，电子流取代纸张，加上无线电技术和卫星通信技术的充分运用，上传到网上的任何信息资源，都只需要短短数秒就能传递到世界各地的每一个角落。由于信息源增多，网络信息内容发布自由，网络信息内容呈爆炸性增长。随着网络的普及化，其传播范围将越来越广。与传统的媒介相比，网络信息传播具有交互性。它具有主动性、参与性和操作性，人们自己主动到网上数据库查找所需的信息，网络信息内容的流动是双向互动的。

5. 结构复杂，分布广泛

网络信息内容本身的组织管理没有统一的标准和规范，信息广泛分布在不同国家、不同区域、不同地点的服务器上，不同服务器采用不同的操作系统、数据结构、字符集和处理方式，缺乏集中统一的管理机制。

6. 信息源复杂、无序

网络的共享性与开放性使得人人都可以在互联网上索取信息和存放信息，由于没有质量控制和管理机制，这些信息没有经过严格编辑和整理，良莠不齐，各种不良和无用的信息大量充斥在网络上，形成一个纷繁复杂的信息世界。

网络信息被存放在网络计算机上，由于缺乏统一的控制，质量参差不齐，网络信息内容分布分散，开发显得无序化。

7. 动态不稳定性

Internet 信息地址、链接和内容处于经常性变化之中，信息源存在状态的无序性和不稳定性使得信息的更迭、消亡无法预测，这些都给用户选择、利用网络信息带来了障碍。

网络信息的这些特点决定了其容易成为网络欺诈、钓鱼以及网络暴力、网络恐怖主义等恶意行为的载体，因此研究网络信息内容安全，提供对互联网中各种不利信息的检测分析能力，是体现我国信息技术水平的重要环节，也是建设信息化社会的坚实保障。

1.2　网络信息内容安全的概念

1.2.1　网络信息内容安全的定义

网络信息内容安全是研究如何利用计算机从动态网络的海量信息中，对与特定安全主题相关的信息进行自动获取、识别和分析的技术。在研究文献和实际应用中，网络内容安全目前大致可分为两类：第一类是基于内容的访问控制，包括网络协议恢复、基于数据包的流量监测、特征码匹配的病毒防护、基于内容的反垃圾邮件等技术；第二类是基于信息传播的互联网安全管理问题，反映的是网络用户公开发布的信息所带来的社会公共安全问题，这里面所涉及的技术主要包括主题信息监控、舆情监控、社交网络社团挖掘等。本书认为，第一类内容安全应用所解决的问题，无论从技术还是表现形式上，更偏向于传统安全，可以依赖于传统信息安全技术解决。而第二类基于信息传播的内容安全问题在近几年显得尤为突出，并且涉及国计民生，对于社会和公民产生的影响更为直接和严峻，同时学术界在提到互联网信息内容安全时，也普遍默认为第二类内容安全是今后安全防范的趋势。因此，本书所研究的网络内容安全问题默认是指第二类内容安全。一般来讲，传统的信息安全体系中并不包括信息内容安全，但随着网络的大规模普及，信息内容安全所遭受的威胁日渐突出，从国家层面，公安机关和文化部门需要网络信息安全技术来保护社会稳定和文化安全，从单位层面，企事业单位需要维护单位形象、避免谣言和竞争对手的诽谤等带来的影响。近年来，网络信息内容安全越来越被认可，并已经纳入信息安全体系。传统信息安全层次包括物理安全、运行安全和数据安全，这 3 个层次所面临的安全问题十分严峻，但往往是普通用户肉眼所感受不到的潜在安全问题，而逐渐兴起的网络信息内容安全问题更为公开，可利用的人口资源更丰富，如图 1-1 所示。

图 1-1　信息安全层次结构

网络信息内容安全处在安全体系中的最上层，更倾向于信息自身的安全，因此更容易被利用。

1.2.2　网络信息内容安全的特点

网络信息内容安全作为一门新兴的课题，以互联网为载体，有着自身的特点。

(1) 网络内容安全既是一门新兴的课题，又需要多个学科进行交叉研究。在信息科学与技术领域，它不同于传统的信息安全问题，是一个综合交叉学科，所用到的技术涉及数据挖掘、话题识别与跟踪、信息过滤、社会网络计算、自然语言处理、数据存储技术等，涵盖计算机科学领域的很多方向。而在非计算机科学与信息安全领域，对于内容安全的研究又大量涉及法学、传播学、管理学、情报学、心理学、社会学等学科，这些学科使得网络信息内容安全不再仅仅像传统信息安全那样只局限于技术领域，对它的研究将更加复杂和丰富。

(2) 网络内容安全以互联网为研究载体。在互联网上发布和获取信息都十分便利，这也是网络内容安全问题的一个重要诱因，因此在网络信息内容研究中，从互联网技术角度入手仍然是对网络信息内容安全管理最有效的手段，尤其是对于新的应用应当格外关注。

(3) 网络信息内容安全问题面对的是海量信息。传统安全更关注封闭式网络安全，防止外界的攻击，相对来说数据流量规模较小。而互联网是一个开放的平台，信息来源广、传播途径多，因此在海量数据中挖掘出潜在的安全问题是对网络内容安全挖掘技术的考验。

(4) 网络信息内容安全虽不同于传统信息安全，但传统安全是信息内容安全的有力保障，例如维护网络和服务器的正常工作，保持数据传输的顺利进行。

网络信息内容安全的这些特点决定了其研究手段和方法与传统信息安全存在显著区别，需要加强网络信息内容安全技术的研究，以实现互联网的健康有序发展。

1.2.3 网络信息内容安全与相关学科的关系

作为新兴的边缘交叉学科，网络信息内容安全与相关学科，尤其是信息安全学科息息相关。本节从学科外延与内涵、学科科学研究方法以及《信息安全专业指导性专业规范》方面分析这两者的关系。

1. 学科外延及内涵的关系

信息安全学科是研究确保信息的完整性、可用性、保密性、可控性以及可靠性的一门综合性新型边缘学科。信息安全学科研究内容包括信息设备安全、数据安全、内容安全和行为安全四个方面问题。信息系统硬件结构的安全和操作系统的安全是信息系统安全的基础，密码、网络安全等技术是关键技术。只有从信息系统硬件和软件的底层采取安全措施，从整体上采取措施，才能比较有效地确保信息系统的安全。当前，信息安全学科的主要研究方向有密码学、网络安全、信息系统安全和信息内安全。可以预计，随着信息安全科学技术的发展和应用，一定还会产生新的信息安全研究方向，信息安全的研究内容将更加丰富。网络信息内容安全是以网络为主要研究载体，此外，报纸、杂志、广播、电视等传播媒体形式也涉及内容安全问题。对于所处理信息的判定方法和标准，与信息安全学科在原理上是一致的。但是在具体实现技术方面，网络内容存储在计算机上，更方便于利用计算机自动处理；而且由于网络信息量大、信息发布来源众多，对自动处理功能有更强烈的需求和更大的技术挑战。网络信息内容安全与计算机和网络系统安全相比较，着重强调的是网络上传输信息的内容安全问题，不等同于硬件设备、操作系统和应用软件的安全问题，但计算机与网络系统的正常工作，为信息内容安全系统的正常运行提供了基础。网络信息内容安全属于信息安全分析技术的一个分支。对特征选取、数据挖掘、机器学、信息论和统计学等多门学科的研究，不仅促进了信息分析技术的发展，也为信息内容安全的研究提供了技术支持。信息内容安全关注与安全相关的内容分析，在处理对象、研究方法的侧重点、对数据吞吐量及对处理

结果响应速度等方面的要求有其自身特点。由此可知，信息安全包含网络信息内容安全。

根据表 1-1，网络信息内容安全主要是研究禁止非法的内容进入和有价值的内容泄露的一门学科。网络信息内容安全关键技术主要包括：信息内容管理(Information Content Management，ICM)、信息内容过滤(Information Content Filtering，ICF)、信息内容监控(Information Content Monitoring，ICMO)和信息内容还原(Information Content Restore，ICR)。信息内容管理是根据设定的条件，用户受限浏览使用数字内容，但可以自由浏览使用非受限数字内容。信息内容过滤是指采用安全策略堵塞或过滤掉那些不良或恶意的数字内容。信息内容监控是由政府和军队执法机构(如公安、司法以及军队有关部门)采用安全策略监控和管理与国家安全、社会稳定、军队指挥紧密相关的数字内容，并有权直接处理与其安全策略不相符的内容。信息内容还原是指协议还原技术，为了保障网络安全高效地传输，在传输过程中包含了大量的协议，必须从有效信息中剔出协议数据，这就是网络协议还原。目前简要分析的协议主要包括 HTTP、FTP、SMTP、POP3、TELNET 和各类 IM 协议。综上，在外延上，信息安全学科包含了网络信息内安全学科；在内涵上，网络信息内容安全学科以网络信息为载体，研究问题更为具体，而信息安全学科研究问题更为体系化、结构化和全面化。

表 1-1　网络信息内容安全内涵

领域	内　涵	关键技术
政治方面	防止来自国内外反动势力的攻击、诬陷以及西方的和平演变阴谋，维护社会稳定	网络舆情分析、内容还原
安全方面	防止国家、军队和企业机密信息被窃取、泄露和流失	开源情报分析
宗教方面	防止法轮功等邪教组织利用宗教信仰传播不利于和谐社会的内容	话题检测与跟踪
破坏方面	防止病毒、垃圾邮件、网络蠕虫等恶意信息耗费或破坏网络资源	内容过滤、内容还原
健康方面	在传播过程中剔除色情、淫秽和暴力内容，使人们健康上网	网络内容过滤
生产方面	防止非生产力网络浏览，提高企业网络使用效率	内容管理
隐私方面	防止个人隐私被盗取、倒卖、滥用和扩散	开源情报分析

2. 学科科学研究方法区别

信息安全学科是综合计算机、电子、通信、数学、物理、生物、管理、法律和教育等学科发展演绎而成的交叉学科。信息安全学科是研究信息的获取、存储、传输和处理中的安全威胁和安全保障的新兴学科。信息安全学科已经形成了自己的理论、技术和应用，并服务于信息社会，信息安全学科归于工学，表 1-2 给出了信息安全支撑技术。由于信息安全理论与技术的内容十分广泛，信息安全学科仍在发展壮大中。

表 1-2　信息安全支撑技术

信息安全支撑技术	研究方向	关键技术
密码学	密码基础理论	密码函数、密码置换、序列及其综合、认证码理论、有限自动机理论等
	密码算法研究	序列密码、分组密码、公钥密码、哈希函数等

续表

信息安全支撑技术	研究方向	关键技术
安全协议	安全协议设计	单机安全协议设计、网络安全协议设计
	安全协议分析	经验分析法、形式化分析
信息隐藏	数字水印	数字版权保护、匿名通信等
	隐蔽通信	隐写术、隐通道、阈下通信等
安全基础设施	PKI/KMI/PMI	产生、发布和管理密钥与证书等安全凭证
	检测/响应基础设施	预警、检测、识别可能的网络攻击，响应攻击并对攻击行为进行调查分析等
系统安全	主机安全	访问控制、病毒检测与防范、可信计算平台、主机入侵检测、主机安全审计、主机脆弱性扫描等
	系统安全	数据库安全、数据恢复与备份、操作系统安全等
网络安全	网络硬件安全	防火墙、VPN、网络入侵检测、安全接入、安全隔离与交换、安全网关等
	信息内容安全	内容管理、内容过滤、话题跟踪与检测、社会网络分析、舆情分析、开源情报分析等
	网络行为安全	网络安全管理、网络安全审计、网络安全监控、应急响应等

网络信息内容安全以网络为主要研究载体，对信息处理速度要求高（近实时）、处理吞吐量大（达到 TB 级）、自动处理功能需求强烈。信息内容安全属于通用网络内容分析技术，对特征选取、数据挖掘、机器学习、信息论、统计学、中文信息处理等多门学科进行研究，不仅促进了信息分析技术的发展，也为网络信息内容安全研究提供了有力的技术支撑。

网络信息内容安全与信息安全研究方法的区别如下。

信息安全是使用密码学方法为信息制作安全的信封，解决信息的“形式”保护问题，而不需要理解信息的“内容”。换言之，采用密码学解决信息安全问题，使没有得到授权的人不能打开这个信封。

网络信息内容安全则需要“直接管理”信息内容，对海量、非结构化数据进行实时判断：哪些是“好消息”？哪些是“坏消息”？并尽可能地完成对坏消息的封堵和自动过滤处理。研究信息内容安全问题的首要条件，是必须由用户明确定义信息的“安全准则”，包括：安全领域（关注什么领域的信息内容安全问题）和安全标准（什么是安全的信息内容，什么是不安全的信息内容），这样才能据以判断具体的信息是否符合所定义的安全准则。可见，信息内容安全问题是“面向特定领域”的，取决于用当时的关注域，而不是“全方位”的。

研究信息内容安全问题的过程，是在“理解信息内容”基础上的“三分类”过程。

（1）句法分析：判断“信息是否为可读语句”，又称为语句分类。

（2）主题分类：判断“由可读语句表达的信息是否属于所关注的安全领域”，又称领域分类或主题分类。

（3）倾向分类：判断“落入某领域的信息是否符合所定义的安全准则”，又称安全分类。

这样，网络信息内容安全问题就可以归纳为“三分类”问题，“三分类”模型参见图 1-2。

3. 学科专业规范区别

信息安全基础（Information Security Base，ISB）是信息安全学科的一些基础内容。信

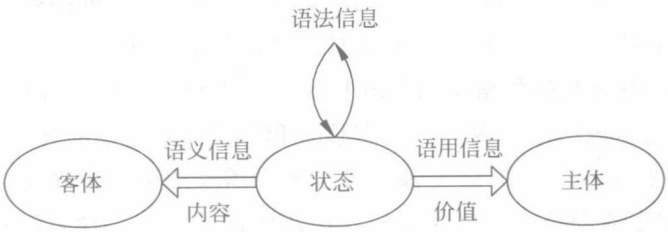

图 1-2　网络信息内容安全“三分类”模型

息安全基础知识领域由信息安全概念知识单元、信息安全数学基础子知识领域、信息安全法律基础知识单元和信息安全管理基础知识单元四个部分组成。而信息安全数学基础子知识领域又由数论、代数结构、计算复杂性、逻辑学、信息论、编码学和组合数学七个知识单元组成，如图 1-3 所示为它们之间的结构。

信息安全基础中的信息安全概念主要介绍对信息安全的威胁、信息安全的基本概念和确保信息安全的措施等基本知识。信息安全数学是信息安全学的理论基础之一，如数论、代数结构、组合数学、计算复杂性、信息论等是密码学的基础，逻辑学是网络协议安全的基础。信息安全法律基础介绍信息安全领域中的一些基本管理知识。信息安全法律和信息安全管理知识则是对整个信息安全系统的设计、实现与应用都有指导性作用的。

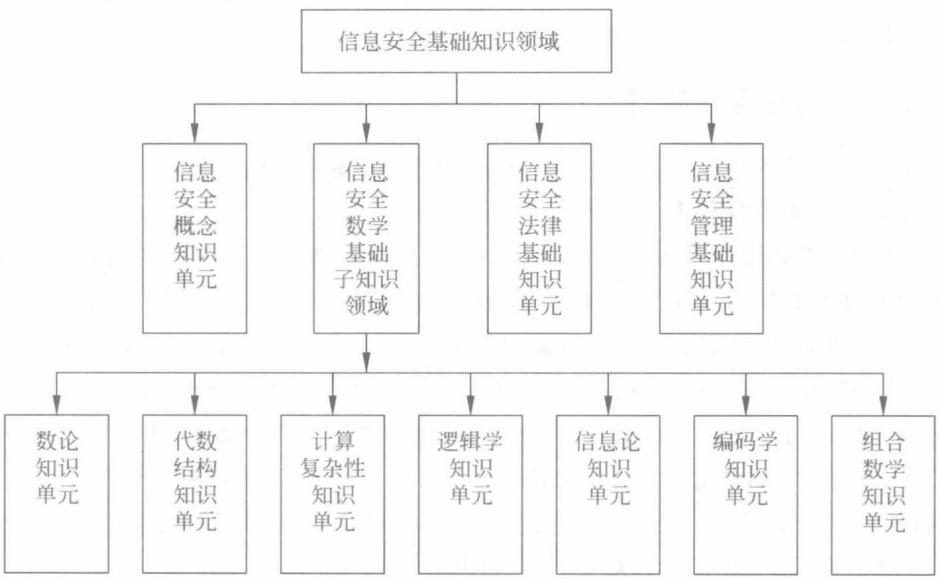

图 1-3　信息安全基础知识领域结构

网络信息内容安全旨在分析识别信息内容是否合法。确保合法内容的安全，防止非法内容的传播和利用。网络信息内容安全的知识单元包括：网络信息内容安全的概念、网络数据的获取、信息内容的分析与识别以及信息内容的管控等。因为不再单独设立信息内容安全法律法规课程，所以在安全概念中还包含了少量与信息内容安全相关的法律法规内容。

网络信息内容安全的重点是网络数据的获取、信息内容的预处理与过滤以及网络信息