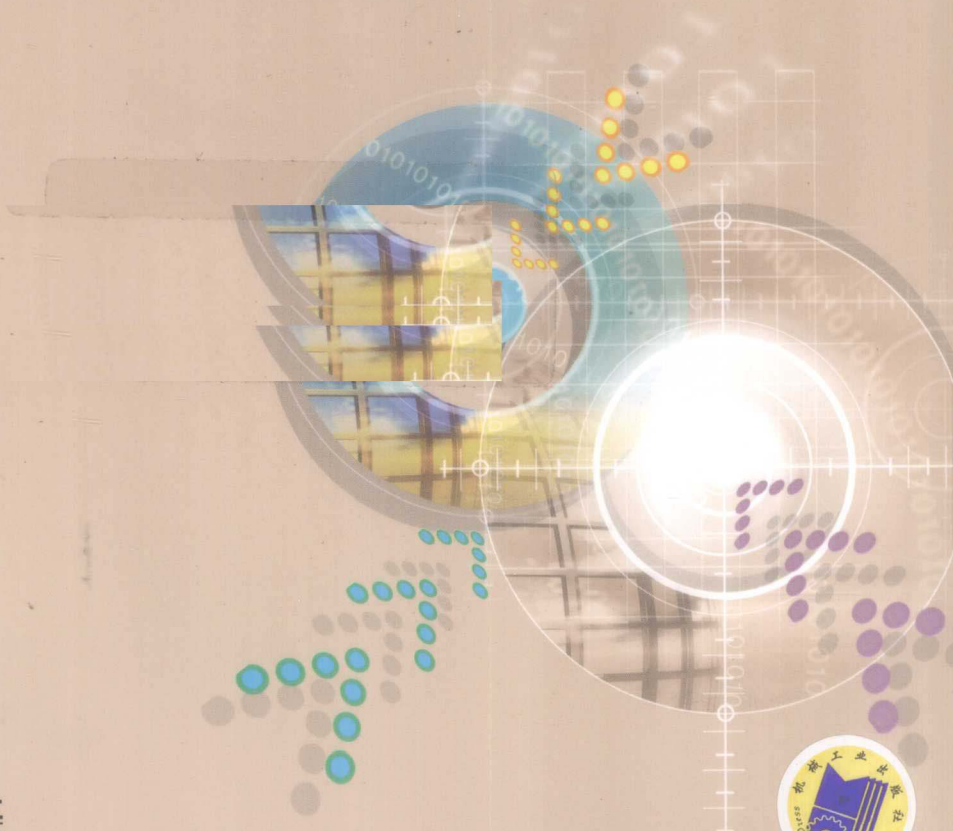
高等院校信息安全专业规划教材

信息内容安全管理及应用

- 信息采集的原理及方法
- 文本、图像、视音频的特征提取技术
- 面向内容安全的分类原理及方法
- 内容安全的典型应用案例

李建华　主编
李 翔　李生红　刘功申　马颖华　等编著



机械工业出版社
CHINA MACHINE PRESS

高等院校信息安全专业规划教材

信息内容安全管理及应用

李建华　主编

李　翔　李生红　刘功申　马颖华　等编著

机械工业出版社

本书从信息处理的基本理论开始讲解，通过几个具有代表性的信息内容安全应用实例，系统地介绍信息内容安全在目前的发展和现实水平。本书共9章，主要内容包括互联网信息内容获取、文本特征的抽取、音频和视频特征抽取、信息处理模型和方法、分类算法、信息过滤、数字水印和舆情系统等。

本书可作为高等院校信息安全相关专业信息内容安全课程的教材，也可作为从事信息内容安全工作的科技人员、工程技术人员以及其他相关部门人员的参考资料。

图书在版编目（CIP）数据

信息内容安全管理及应用 / 李建华主编. —北京：机械工业出版社，2010.5
（高等院校信息安全专业规划教材）
ISBN 978-7-111-29954-7

Ⅰ. ①信… Ⅱ. ①李… Ⅲ. ①信息系统－安全管理－高等学校－教材
Ⅳ. ①TP309

中国版本图书馆 CIP 数据核字（2010）第 036147 号

机械工业出版社（北京市百万庄大街 22 号　邮政编码 100037）
责任编辑：唐德凯
责任印制：杨　曦

北京四季青印刷厂印刷（三河市杨庄镇环伟装订厂装订）

2010 年 7 月第 1 版 · 第 1 次印刷
184mm × 260mm · 10.75 印张 · 264 千字
0001—3000 册
标准书号：ISBN 978-7-111-29954-7
定价：22.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务
社服务中心：（010）88361066
销 售 一 部：（010）68326294
销 售 二 部：（010）88379649
读者服务部：（010）68993821

网络服务
门户网：http://www.cmpbook.com
教材网：http://www.cmpedu.com

封面无防伪标均为盗版

高等院校信息安全专业规划教材

编委会成员名单

出版说明

信息技术的发展和推广,为人类开辟了一个新的生活空间,它正对世界范围内的经济、政治、科教及社会发展各方面产生重大的影响。如何建设安全的网络空间，已成为一个迫切需要人们研究、解决的问题。目前，与此相关的新技术、新方法不断涌现，社会也更加需要这类专门人才。为了适应对信息安全人才的需求,我国许多高等院校已相继开设了信息安全专业。为了配合相关的教材建设，机械工业出版社邀请了解放军信息工程大学、解放军理工大学通信工程学院、上海交通大学、西安电子科技大学、湖南大学、南京邮电学院等高校的专家和学者，成立了教材编委会，共同策划了这套面向高校信息安全专业的教材。

本套教材的特色：

1）作者队伍强。本套教材的作者都是全国各院校从事一线教学的知名教师和学术带头人，具有很高的知名度和权威性，保证了本套教材的水平和质量。

2）系列性强。整套教材根据信息安全专业的课程设置规划，内容尽量涉及该领域的方方面面。

3）系统性强。能够满足专业教学需要，内容涵盖该课程的知识体系。

4）注重理论性和实践性。按照教材的编写模式编写，在注重理论教学的同时注意理论与实践的结合，使学生能在更大范围内、更高层面上掌握技术，学以致用。

5）内容新。能反映出信息安全领域的最新技术和发展方向。

本套教材可作为信息安全、计算机等专业的教学用书，同时也可以供从事信息安全工作的科技人员以及相关专业的研究生参考。

机械工业出版社

前　言

近年来发生了很多安全事件，例如美国 9 · 11 事件、伦敦公交系统连环爆炸案、巴厘岛恐怖袭击、印度孟买恐怖袭击等。灾难的发生促使大众开始重新审视社会各个方面的安全性和可靠性。在这种环境下，计算机被认为是解决此类安全问题的一个有力工具，例如，它被广泛用来收集和分析情报。美国政府在 9 · 11 事件后，建立了全球联网的指纹系统及日趋严格的出入境管理体系，以期建筑严密的恐怖袭击防控网络，尽管由于对恐怖活动的规律性还缺乏清晰的认识，这些网络暂时还未发挥出预警和防范恐怖袭击事件的作用。

就计算机本身而言，无论从硬件到软件，还是从操作系统到数据管理系统，都存在严重的安全问题。网络所带来的计算机安全问题则更为严重。网络互连在方便信息传送的同时，也给连网计算机所保护的信息带来了威胁。除了基于网络和软硬件的安全问题以外，近几年来，互联网还暴露了其他的一些安全隐患，尤其是一些对于整个社会都起到负面影响的安全问题。

最为引人注目的是，自 2005 以来爆发的多起“人肉搜索”等网络暴力事件，把互联网中内容安全问题暴露在公众眼前。事实上，网络“暴力”由来已久，互联网上公开的信息及越来越强大的搜索功能，使原本隐在角落的信息被“曝光”到大众视野内，一些本不构成隐私的信息在互联网上任意传播，并在引发网络上的语言暴力后，造成了严重的后果。

还有数字信息的知识产权问题。由于数字信息复制及网络传播非常便利，造成信息自身具有的知识产权被有意或无意地侵犯。尽管在欧洲发生了几起因有意或无意的共享了具有知识产权歌曲而引发的多起诉讼和巨额的罚金，但法律毕竟是版权侵权的最后防范手段。目前，已经出现了在组织内部（局域网范围内）防范信息泄露的技术手段，尽管在整个互联网领域此类技术还很缺乏，但我们有理由相信计算机技术将能够起到更为重要的作用。

以上是一些计算机安全中的新型问题，大多是公共或私有信息的内容所带来的风险。这些风险中，有些是商业风险，有些是个人或者组织的危机，有些是社会的安全风险。相比于传统的信息安全问题，例如通信安全、计算机安全等与计算机网络和软硬件设备关系紧密的安全问题不同，对此类风险的评估及加强安全的防护是新的一类信息安全问题，我们把它称为“信息内容安全”，或称为“内容安全”。本书是对此类问题的分析及相关技术的总结和介绍。

本书目标

本书有三大目标:

第一个目标是强调信息内容安全与计算机安全和目前被广泛使用的信息安全等概念是不同的领域。现有大部分的计算机安全技术是通过密码学、存取控制等手段保护数据的保密性与完整性。但在互联网环境下的信息内容安全是出现在信息公开的前提和开放的环境下，由

信息的内容所引发的系列风险。所以对此类安全问题需要采用与其他的安全问题不同的解决思路。对信息内容安全领域安全问题的总结和思考是本书最重要的目标。

第二个目标是要探讨信息自身的特点和特征。随着计算机技术的发展，目前信息的格式多种多样，尤其以多媒体信息在信息中所占的比例日益增多。信息可以大体分为文本信息、视频信息、音频信息、图像信息、数字信息等多种类型，对于各类信息的处理方法也根据其内容的差异而有很大的不同。要了解信息内容安全技术，首先需要熟悉各类信息格式、特点、特征及其处理技术。

第三个目标是总结现阶段信息内容安全技术及其在各个领域中的应用。信息内容安全作为一个全新的安全课题，暂时还没有非常系统和完整的理论体系。另外，由于信息内容安全和其他计算机安全研究领域的存在目的及存在环境的不同，而无法直接借用以前的计算机安全体系或者策略等成果。因此，只能从现有的信息内容安全的实用系统出发，了解这些系统的原理和作用，希望以点概面地介绍目前阶段信息内容安全领域的进展。

本书结构

信息安全是一个得到当今社会越来越多重视并得到不断发展的学科。其中，信息内容安全在近几年起步，并日益受到越来越多领域的重视。

本书可分为三大部分，第 1 章即第一部分，介绍信息内容安全的基本概念；第二部分包括第 2～6 章，介绍信息处理的各项基础技术；第三部分包括第 7～9 章，针对当前较为典型，且应用较为广泛的几种信息内容安全应用领域进行详细的介绍。

本书体现了信息安全类专业课程改革和实践的方向。本课程建议授课学时为 36 小时。作为教材之用，每章后附有习题，有助于对知识的巩固。

感谢教育部高等学校信息安全类专业教学指导委员会信息安全类专业课程教学改革与实践课题组在本书编写过程中所给予的大力支持。

参加本书编写的有孙强、张文军、李翔、马颖华、苏贵洋、王士林、孙锬锋、刘功申、李生红。其中，第 1 章由孙强和马颖华编写；第 2 章由林祥和于朝阳编写；第 3 章由孙强编写；第 4 章由张文军编写；第 5 章由王士林编写；第 6 章由苏贵洋、王士林等共同编写；第 7 章由苏贵洋编写；第 8 章由孙锬锋编写；第 9 章由李翔编写。

由于时间仓促，书中难免存在疏漏或不妥之处，请读者予以指正。

编　者

目　录

第1章 绪 论

本章主要是从信息内容安全的产生、发展背景、应用环境、研究现状及其意义等角度对信息内容安全的某一个方面进行介绍。本章是学习本书后续内容的必要准备。

1.1 信息内容安全概述

互联网起源于20世纪60年代末70年代初。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。自上世纪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18次中国互连网络发展状况统计报告，到2006年6月，我国网民总数已超过1亿人，联网计算机总数超过5000万台。不久的将来，我国将成为世界上最大的互联网用户群体。

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、非常贴近于人们生活等众多特点。一方面，互联网创造出了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，互联网也带来了一些负面影响，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的泛滥，利用网络传播电影、音乐、软件等的侵犯版权行为，甚至通过网络方式欺诈网络用户，以及出现网络暴力和网络恐怖主义活动等问题，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，是国家信息技术水平中的重要一环，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。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，为人们提供了便利地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，互联网中更容易出现一些另类、新奇、不易理解或不符合规范的行为。但互联网将整个世界变成了“地球村”，使持有各种思想、观点的人聚集在一起，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不良现象，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与消除不良现象，让互联网更好地为人民服务，发挥更大的效用，从而使人人都能更高效、更自由地使用互联网进行信息沟通。

信息内容安全（Content-based Information Security）作为对上述问题的解决方案，它是研究如何利用计算机从包含海量信息且迅速变化的网络中，对与特定安全主题相关信息进行自

动获取、识别和分析的技术。根据所处的网络环境，它也被称为网络内容安全（Content-based Network Security）。信息内容安全是管理信息传播的重要手段，属于网络安全系统的核心理论与关键组成部分，对提高网络使用效率、净化网络空间、保障社会稳定具有重大意义。

信息化是当今世界发展的大趋势，是推动社会进步的重要力量。大力推进信息化，是覆盖我国现代化建设全局的战略举措，也是贯彻落实科学发展观、全面建设小康社会和建设创新型国家的迫切需要和必然选择。信息内容安全作为网络安全中智能信息处理的核心技术，为先进网络文化建设和社会主义先进文化的网络传播，提供了技术支撑，它属于国家信息安全保障体系的重要组成部分。因此，信息内容安全研究不仅具有重要的学术意义，也具有重要的社会意义。

1.2 信息内容安全威胁

从要解决的主要问题及其解决方案来看，和计算机安全一样，信息内容安全主要建立在保密性、完整性和可用性之上。由于安全问题所处的环境不同，对其解释也会有很大不同，本书主要从互联网角度来分析信息内容安全方面的几个大问题。

在分析信息内容安全的问题前，首先要搞清楚对安全的威胁来自何方。传统计算机安全面临的威胁有泄露（指对信息的非授权访问）、欺骗、破坏和篡改。但在互联网信息共享环境中，人们同样发现信息内容安全所面临的威胁也有泄露、欺骗、破坏和篡改。

在局域网连上互联网时，局域网内的敏感信息有可能泄露到互联网中。例如，由于局域网上的信息可能会保存在不同的系统中，造成无法进行或不可能实现可控的安全管理。这种安全管理上的缺失，造成了互联网信息内容的安全面临着各方面的威胁。下面对这些威胁进行详细描述。

1）互联网中有大量公开的信息，如某人的姓名、工作单位、住宅地址、电话号码等。由于这些公开信息的获取途径简单、成本非常低，在某些情况下，会被整合并可能被滥用，例如某些公司会将这些数据作为商业信息出售，还有些不法集团会利用这些信息进行诈骗。所以互联网上的信息泄露，还指将特定信息向特定相关人或组织进行传播，以妨碍特定相关人或组织的正常生活或运行。

2）互联网的开放性和自主性，可使信息由各个组织自发生成，并共享到互联网中。但这也带来了很多欺骗性的威胁。例如，互联网的地址和内容都存在被伪造的可能性。这些是由于互联网运行中无法保证信息完整性（尤其是信息来源）而造成的。

3）信息被非法传播。在网络中发现，很多具有知识产权的音乐和电影被广泛传播，从而造成了知识产权被侵犯的局面。

4）信息在传播过程中，也可能被篡改。篡改信息的目的，可能是为了消除信息的来源，使其无法跟踪；也可能是为了伪造信息的内容，影响正常的信息交流。此外，信息篡改后，还会被植入木马等病毒，这些程序代码不仅会对所在的信息载体带来破坏，还会直接危害到软硬件系统的安全。

1.3 信息内容安全特点及其与相关学科的联系

作为新兴边缘交叉学科，信息内容安全有其自身特点，同时也与许多学科有着密切的联系，具体分析如下：

1）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，在原理上是一致的。然而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，对自动处理功能有了更强烈的需求和更大的技校挑战。

2）信息内容安全和计算机与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。

3）信息内容安全属于通用网络内容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。信息内容安全关注于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。

1.4 信息内容安全研究现状

由于信息内容安全研究中有部分会涉及国家安全等敏感问题，因而相关资料较难获得。下面我们对收集到的典型项目进行讲解。

1.4.1 政府部门主导的项目

随着互联网应用的日益广泛，网上信息安全问题也逐渐突出。于是，各国政府均先后提高了对信息内容安全问题的重视程度。

在 9·11 恐怖袭击事件发生后，FBI 局长 Robert S. Mueller 在议会听证会上发言，认为政府花费了过多的精力用于案件侦查，以致没有足够的资源用于预防案件发生。Robert 认为，这是由于他们虽然获得了大量数据，但却缺少把数据进行整合与深度分析。此后，FBI 加大了对一些领域的研究力度，包括：整合不同来源、不同格式数据的技术；对犯罪及恐怖活动相关的网络链接进行分析与可视化显示的技术；能够对信息进行监控、检索、分析及做出主动响应的 agent 技术；对海量信息（TeraBytes）级别存储文档、网页和电子邮件的文本挖掘技术；利用神经网络对可能的犯罪活动或者新的恐怖袭击进行预测的技术；利用机器学习算法抽取罪犯描述特征与犯罪活动关系结构图技术等。

可见，信息内容安全影响的范围并不是仅仅局限于虚拟网络，而是与其他方面的安全问题密切联系、相互影响。政府主导的部分代表性项目见表 1-1。

表 1-1 政府主导项目

国 别	单 位	项目名称	简 介
美国	FBI	Carnivore	网络信息嗅探软件与相关软件配合，可实现信息还原和内容分析，主要用于监测互联网中的恐怖活动、儿童色情、间谍活动、信息战和网络欺诈行为等。运行于微软 Windows 平台，2005 年 1 月以后停止
美国	FBI	StrikeBack	与联邦教育部合作，用于查询可疑学生信息，每年有数百名学生信息被查询。5 年期计划，已结束
多国	UKUSA	ECHELON	以美英为主导，由多个英语国家参与。它是世界上最大的网络通信数据监听与分析系统。监听世界范围内的无线电波、卫星通信、电话、传真、电子邮件等信息后，应用计算机技术进行自动分析。每天截获的信息量约 30 亿条。最初，ECHELON 用于监控苏联和东欧的军事与外交活动。现在，其重点监听恐怖活动和毒品交易的相关信息

（续）

国　别	单　位	项目名称	简　介
英国		RIP	关于通信监听方面法律，是于2000年通过的。其该国政府被授权监控所有电子邮件通信，包括加密通信
美国	CIA	Oasis	以语音识别技术为核心，用于将电话、电视、广播、网络上面的音频信息转换为文本信息，以便于检索。目前，Oasis系统可以识别英语，下一步的目标是实现对阿拉伯语和汉语的处理
美国	DARPA	EELD	研究如何从海量的网络信息中，发现有可能威胁国家安全的关键信息提取技术
美国	DHS	ADVISE	建立在前述ECHELON项目的基础上，通过数据挖掘技术对互联网上的新闻网站、网志（Blog）、电子邮件（E-mail）进行分析，以发现其中各种网络标示之间的关系。该计划目的在于，尽早发现恐怖分子可能发动的恐怖活动。数据的三维可视化展示是该项目的一个特点，它提供了一种新型的数据展示方式

1.4.2　科研院所或公司的项目与产品

由科研机构主导的部分研究项目见表1-2。

表1-2　研究机构主导的研究项目

单　位	项目名称	简　介
UCLA	PRIVATE KEYWORD SEARCH ON STREAMING DATA	该项目需放置多台服务器到网络各处，收集网络上特定信息后传回信息处理中心，减轻了将所有信息直接传回信息处理中心的负担。项目特点在于，虽然这些放在信息源附近的机器，没有集中式服务器的物理性和系统安全性，甚至有可能为敌对方获取，但该系统会利用同态加密（Homomorphic Encryption）实现编码混淆（Code Obfuscation）。该技术保证了机器上面安装的软件不会被逆向工程侵犯，也即敌对方无法利用缴获的服务器来获取该服务器过滤的明确规则。另外，由于预先滤除了大量信息，系统在安全和隐私方面也取得了较好均衡 http://www.research.ucla.edu/tech/ucla05-487.htm
Autonomy	IDOL Server	Autonomy公司的产品IDOL Server是用途广泛的文本信息挖掘工具，具有能进行语义级别的检索、文本分类与推送等功能。支持多种自然语言，利用信息论的相关知识进行文本特征选择与提取，利用贝叶斯理论进行分类。在FBI与CIA中，有广泛应用 http://www.autonomy.com/content/Products/IDOL/index.en.html
Secure Computing	SmartFilter	用于阻止网络间谍软件与网络钓鱼软件对网络用户的侵害。在军事、民事领域，都有应用
NICTA	SAFE	澳大利亚国家信息与通信技术研究中心的紧急状态灵活应对系统计划，该项目通过人脸识别等机器视觉技术来分析可能的异常行为，从而实现预先判断，以阻止恐怖主义活动
Cornell	Sorting acts and opinions for homeland security	该项目由美国国土安全部资助，康奈尔大学联合匹兹堡大学和犹他大学负责实施。重点是通过信息抽取等多种自然语言理解与机器学习技术，从收集到的文本中判断各种信息所包含的观点，并且研究如何寻找信息的可能来源，利用这些信息进行辅助决策 http://www.eurekalert.org/pub_releases/2006-09/cuns-sfa092206.php

1.5　信息内容安全研究的意义

在信息化社会的建设过程中，信息内容安全研究有着广泛的应用。根据考察层次对象不同，可分为如下几个方面。

1）提高网络用户及网站的使用效率。网络用户经常遇到垃圾邮件、流氓软件等的恶意干

扰，网站中也存在某些用户发布一些广告或恶意言论的情况。信息内容安全研究有望提供技术上的解决方案，包括对电子邮件、论坛、Blog 回复和聊天室等进行信息过滤，通过预先过滤不良信息，减少手工处理各类无用信息所花费的时间与精力，从而有效提高网络的使用效率。

2）净化网络空间。互联网的迅猛发展，既满足了广大群众日益丰富的文化生活需求，成为人们获取信息、生活娱乐、互动交流的新兴媒体，同时也存在着传播各种不良信息的现象。例如，传播格调低下的文字与图片、侵犯知识产权的盗版影音或软件、不负责任的传播未证实的消息，甚至别有用心的散布虚假消息以制造恐慌气氛等。此外，随着网络的发展，上网的未成年人也越来越多，只有营造健康文明的网络文化环境，才有利于青少年的身心健康与顺利成长。清除不健康信息已成为社会的共同呼唤和强烈要求，也对信息内容安全相关课题的研究提出了迫切需要。

从建设国家信息安全保障体系的角度看，随着时代的发展，安全问题也拓展到网络这个看不见、摸不着的虚拟世界，提高国家信息安全保障水平是保障国家安全的重要环节。互联网作为信息传播和知识扩散的新式载体，加剧了各种思想文化的激荡与碰撞。各种观点与宣传在互联网上长期互存、互相影响，是一个客观现实。各种违法犯罪活动也利用网络作为传播的新场所，出现了各种网络诈骗活动与网络恐怖主义活动。上述种种情况，都需要更为完善的信息处理技术，尽早或尽量准确地发现安全隐患，以提高预防保护能力；降低各种不良活动发生的可能性或减少其带来的损失。

1.6 本章小结

信息内容安全是信息安全中一个较新的研究领域，它跨越多媒体信息处理、安全管理、计算机网络、网络应用等多个研究领域，直接和间接地应用各个研究领域的最新研究成果，结合信息内容安全管理的具体需求，发展出具有自己特点的研究方向和应用。随着网络在社会生活中占据越来越重要的地位，随着不断涌现出的各种类型的信息内容安全的具体应用，信息内容安全及其管理理论必将受到越来越多的重视，在日常生活和国家信息安全保障等方面也将起到越来越重要的作用。

1.7 习题

1．你认为信息内容安全的主要技术有哪些？

2．你认为信息内容安全技术的发展，能否解决所有的信息内容的安全问题？

3．你认为除计算机技术外，还有哪些领域需要协同工作，才能更好地保障信息内容的安全？

4．有序的疏导是解决水患的最好方法。同理，对于信息内容安全，你认为有哪些方法（包括技术、管理或法律等多个方面）可以对信息内容安全的隐患进行有效疏导？