第5章 内容安全基础

信息内容安全是日益受到越来越多的重视并得到不断发展的领域，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。本章将从信息内容安全概述、信息内容安全威胁、信息内容获取和分析处理基本技术，以及以网络舆情内容检测预警和内容中心网络为代表的信息内容安全应用等角度展开介绍。

5.1 信息内容安全概述

全球信息化的今天，互联网将朝着开放性、异构性、移动性、动态性、并发性的方向发展。通过不断演化，产生了下一代互联网、5G 移动通信网络、移动互联网、物联网等新型网络形式以及云计算等服务模式。同时，随着工业 4.0 影响全球和我国实施“互联网＋”行动，互联网与各个行业的融合也日益加深，创造了巨大的经济效益和社会效益，互联网已成为人们获取信息、互相交流、协同工作的重要途径，具有应用极为广泛、发展规模最大、贴近人们生活等众多优点。

伴随社会信息化和网络化的发展，当前全球数据正在呈现爆炸式增长，数据内容成为互联网的中心关注点。有统计表明，在每一分钟里 Facebook 用户会新共享 68.4 万比特的内容、Twitter 用户会新发出超过 10 万条推特、YouTube 用户会上传 48 小时的新视频、Instagram 用户会共享 3600 张新照片，2020 年的全球信息总量预计可达 35ZB。互联网中的数据和内容已经引起了学术界和产业界的广泛关注，2014 年，Gartner 新技术成熟周期分析报告显示，大数据技术正在逐步演化成生产力，已经成为诸多重要 IT 技术和应用领域的核心。近年来各国也已从国家战略视角高度重视通过互联网获取、掌握威胁国家政治、经济、文化乃至军事安全的情报信息。随着 5G 在全球范围内的部署以及 6G 技术的发展，到 2030 年，预计有万亿级智能设备接入网络，每秒太比特的数据量将被处理，进一步促进了数据驱动型的网络与社会的形成与发展。此外，随着网络技术和移动智能设备的快速发展，互联网逐步由传统媒体向社交网络等新型媒体演进，例如微信、QQ、新浪微博、Facebook 等社交网络工具。互联网和新兴媒体的发展带来了一些负面影响，不良信息在网络上大量传播，垃圾电子邮件、Sybil 攻击、网络水军攻击等不正当行为泛滥，

利用网络传播电影、音乐、软件侵犯知识产权，甚至通过网络钓鱼方式欺诈网络用户以及网络暴力和网络恐怖主义活动等。Facebook Live 自上线以来已经出现至少 45 起暴力事件，包括枪击、谋杀等恶性事件，给社会造成了极其恶劣的影响。因此，在以信息内容为中心的互联网环境下，网络信息内容的安全值得广泛关注和深入研究。

互联网上各种不良信息流传以及不规范行为的产生原因可归结为两类。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，多数是学术研究人员，网络也没有用于商业用途，网络安全的问题并不突出。如今这些情况都已经发生了巨大的变化，一些原有网络模式不再适应现在的情况。另外一类原因是互联网在为人们提供便利获取与发布信息的同时，也制造了前所未有的思想碰撞场所，因而在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。互联网将整个世界变成了“地球村”，将持有各种思想、观点的人聚集在一起，这也将是一个长期存在的客观现实。面对这种挑战，一方面，人们不应因噎废食，因为互联网上存在的一些不良现象而畏惧或排斥新技术、新事物；另一方面，应当通过法律与技术等多方面措施限制与消除这些不良现象，让互联网更好地为人民服务，发挥更大的效用，使得人人都能更高效、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）作为对上述问题的回答，是研究利用计算机从包含海量信息并且迅速变化的网络中对特定安全主题相关信息进行自动获取、识别和分析的技术。根据它所处的网络环境，也称为网络内容安全（Content-based Network Security）。信息内容安全是借助人工智能与大数据技术管理网络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。

大力推进信息化是我国现代化建设的战略举措，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，加强社会主义先进文化的网上传播提供了技术支撑，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。

5.2　信息内容安全威胁

从内容安全要解决的主要问题及其解决方案来看，内容安全和计算机安全一样，主要建立在保密性、完整性、可用性之上，典型的信息内容安全挑战如图 5-1 所示。

在分析内容安全的问题之前，首先要搞清楚对安全的威胁来自何方。在互联网、电信网、电视网等各类网络信息共享环境中，一方面，内容安全所面临的威胁有泄露（指对信息的非授权访问）、欺骗、破坏和篡夺等；另一方面，一些恶意用户产生并传播的恶意内容也是网络空间面临的潜在安全威胁。下面首先对泄露、欺骗、破坏和篡夺等威胁进行详细的描述。

首先，互联网中有大量公开的信息，例如某人的姓名、工作单位、住宅地址、电话号码

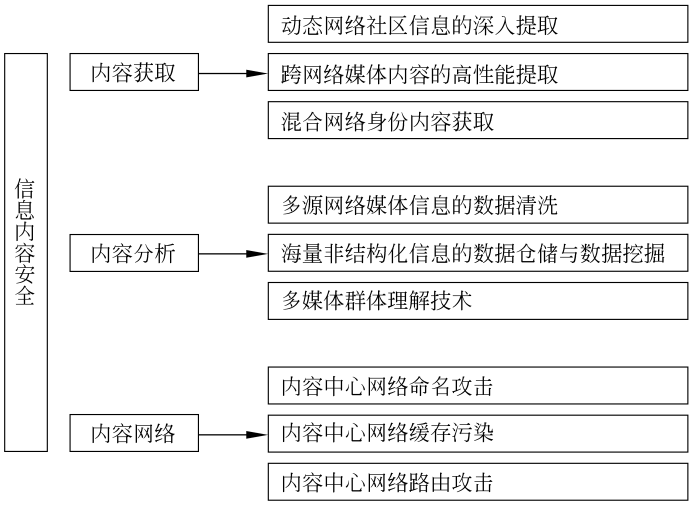

图 5-1　典型的信息内容安全挑战

等，由于这些公开信息的获取成本非常低，在某些情况下，这些信息会被整合，并可能会被滥用，例如，某些公司会将这些数据作为商业信息出售，还有些诈骗集团会利用这些信息进行诈骗。所以，互联网上的信息泄露可以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。其次，互联网的开放性和自主性导致信息由各个组织自发生成并共享到互联网中，这带来了很多欺骗的威胁，互联网的地址和 www 的内容都存在伪造的可能，这些是互联网中无法保证信息完整性（尤其是信息来源）造成的。再次，信息还会被非法传播，在很多网络中被发现具有知识产权的音乐和电影被广泛传播，造成了知识产权被践踏。最后，信息在传播过程中也可能被篡改，篡改信息的目的可能是消除信息的来源信息，使之无法跟踪；也可能是伪造信息的内容。此外，信息篡改后还会包括病毒或者木马，这些有害于计算机系统、数据的代码将不仅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。

随着 Facebook、微博等在线社交媒体平台的发展，人们交流、沟通、获取信息的方式产生了巨大变革。个人用户由传统的内容接收者转变为内容的创造者和传播者。在此过程中，除了上述基于保密性、完整性、可用性的安全问题，网络空间还面临着恶意用户制造传播恶意内容所带来的潜在安全威胁。下面分别介绍几种典型的互联网恶意用户行为威胁。①Spam 用户的恶意行为通常出现在邮件或者网页中，该行为表现为向一些合法的用户发布广告、色情、钓鱼等恶意信息。Spam 用户行为的主要攻击方法是通过创建大量的虚假账号，在邮件或者网页中推荐一些网页链接，来欺骗诱导用户进入推荐的网站或恶意的网站。最早的 Spam 行为始于邮件系统，可以追溯到互联网的产生。早在 1978 年，第一个邮件 Spam 就对阿帕网的几百个用户进行了攻击。近年来，无论是国外的 Twitter，还是国内的微博，都曾受到 Spam 行为的困扰。那些曾经在电子邮件领域横行的 Spammer 找到了新的乐土，在开放式在线社交网络上将恶意内容快速而大规模地传播出去。该方式比传统的定点群发邮件的传播方式更加有效。在线社交平台上的 Spam 行

为毫无疑问会破坏平台环境，威胁平台用户隐私和财产安全。②Sybil 攻击是目前兴起的另一种恶意行为攻击方式，即由少数节点控制多个虚假身份，并利用这些身份来控制或影响网络的大量正常个体的行为，以达到冗余备份的作用。Sybil 攻击最早出现在无线通信领域中，2002 年，美国学者 Douceur 第一次在点对点网络环境中提出了 Sybil 攻击的概念。这种攻击将破坏分布式存储系统中的冗余机制，达到削弱网络的冗余性，降低网络健壮性，监视或干扰网络正常活动等目的。后来学者们发现 Sybil 攻击对传感器网络中的路由机制同样存在着威胁。在线社交网络用户之间缺乏物理上的接触，这成为 Sybil 攻击在在线社交平台盛行的一个有利条件。Facebook 的一个调查报告表明，超过 8300 万个的 Facebook 用户都可能是 Sybil 用户。Sybil 用户在以信任为基础的社交网络上的恶意行为更加隐蔽，使得社交平台所面临的威胁愈加严峻。

水军用户的恶意行为是网络空间面临的另一大严重威胁。水军用户通过评论或者转发参与热点话题，以大量有情感倾向的评论影响舆情态势。以微博为代表的开放式网络平台聚集了大量用户创造的内容，同时用户之间建立起了错综复杂的巨大的关系网络，这些特点使得开放式社交网络媒体成为了网络水军的生存乐土，催生了集网络推手、网络打手、刷粉等功能于一身的网络新水军。水军营销是目前互联网平台常用的一种营销行为，对于企业来说，可以通过雇佣水军对自己的产品进行宣传。然而由于利益的驱使，水军营销渐渐地走向了歧途，也使得整个产业渐渐蒙上了阴影。水军对舆情态势的影响也不容小视。通过购买大量水军在热点微博下进行同一情感倾向的评论，可以达到混淆公众视听、影响公众情感态度的效果。这种对舆情态势的影响甚至会影响社会稳定和国家安全，需要引起足够的重视。

另一方面，以内容为中心的未来互联网旨在将内容名称而不是 IP 地址作为传输内容的标识符，从而实现信息的路由。内容中心网络更适合大数据的内容分发，可以在网络层实现高效的检索机制。事实上，内容中心网络为未来互联网带来了许多好处。首先，互联网中以信息为中心的内容将包含底层信息的内容、属性和关系，从而引入大量语义和情感特征。因此，可以实施更多优化表示来增强网络性能。其次，信息中心网络在大数据内容分发过程中能够提供更智能的分析，这种分析可以以提高未来互联网的智能水平的方式进行。内容中心网络具有许多独特的属性，如位置独立命名、网络内缓存、基于名称的路由和内置安全性。在内容中心网络体系结构中，除了可能对网络流量产生影响的旧式攻击之外，还出现了新的攻击。信息中心网络将安全模型从保护转发路径更改为保护内容使其可以为所有网络节点使用。内容中心网络攻击可以分为命名、路由、缓存和其他攻击。命名攻击可以分为监视列表和嗅探攻击。这些攻击允许攻击者审查和过滤内容。攻击者还可以获取有关内容流行性和用户兴趣的私人信息。考虑到信息中心网络的数据是根据名称进行路由和缓存的，发布者在向网络中发布内容时会依据相关的命名规则，将数据的有关属性、特征和内容包装为数据名称，从而暴露在网络中；订阅者在向网络中发布请求时，也会依次将所需要数据的相关信息包装为数据名称并将其以兴趣包的形式发布到网络中。因此数据名称本身携带了内容信息。通过对名称中暴露出的信息进行挖掘和延展，攻击者可以从中获得有关内容的信息，并通过语义方面的模糊化和替换，对需求进行混淆，从而可以将并非订阅者真实需要的内容发送给对方，以达到不同目的上的欺骗攻

击。内容中心网络的常见路由攻击是指恶意发布者和订阅者可以发布和订阅无效的内容或路由。内容中心网络缓存容易受到不同类型的攻击，这些攻击会污染或破坏缓存系统，此外还有缓存内容和未缓存内容之间的差异，这些攻击会侵犯信息中心网络隐私。其他路由攻击则表现为在传输过程中未经授权地访问和更改内容。

5.3 网络信息内容获取

正处于内容爆炸性增长的国际互联网、电信网、电视网等各类网络包含了琳琅满目、内容迥异的各式信息。在网络媒体信息与网络通信信息遍布世界各个角落的今天，面向海量网络信息实现全面或有针对性的内容获取，已经成为信息内容安全研究领域中的重要课题。

5.3.1 网络信息内容获取技术

与面向特定点的网络通信信息获取不同，网络媒体信息获取环节的工作范围理论上可以是整个国际互联网。传统的网络媒体信息获取环节从预先设定的、包含一定数量URL的初始网络地址集合出发，首先获取初始集合中每个网络地址对应的发布内容。网络媒体信息获取环节一方面将初始网络地址发布信息主体内容按照系列内容判重机制，有选择地存入互联网信息库，另一方面，还进一步提取已获取信息内嵌的超链接网络地址，并将所有超链接网络地址置入待获取地址队列，以“先入先出”方式逐一提取队列中的每个网络地址发布信息。网络媒体信息获取环节循环开展待获取队列中的网络地址发布信息获取、已获取信息主体内容提取、判重与信息存储，以及已获取信息内嵌网络地址提取并存入待获取地址队列操作，直至遍历所需的互联网络范围。

理想的网络媒体信息获取流程主要由初始URL集合——信息“种子”集合，等待获取的URL队列，信息获取模块，信息解析模块，信息判重模块与网络媒体信息库共同组成，如图5-2所示。

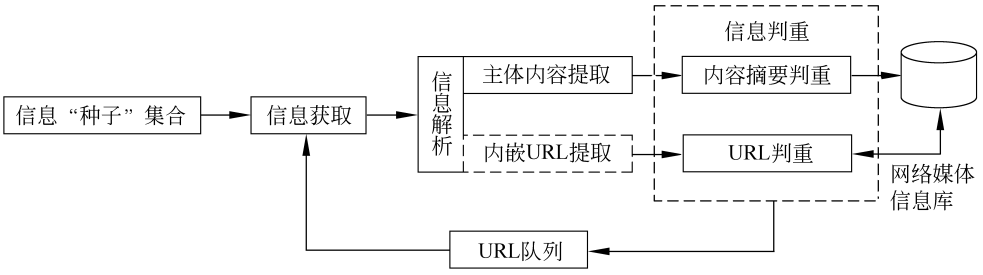

图5-2 媒体信息获取流程

早期传统网络媒体信息获取方法的技术实质，可以统一归属于采用网络交互过程编程重构机制实现网络媒体信息获取。在面向互联网实现公开发布信息获取过程中，网络交互过程编程重构完整实现网络信息请求/响应过程，应当说其属于网络媒体信息获取的一般性方法。理论上只要掌握网络通信协议的信息交互过程，就可以通过网络交互重构

实现对应协议发布信息获取。不过，随着网络应用的逐步深入，网络媒体发布形态不断推陈出新，不同网络媒体信息交互过程存在极大区别。需要对于不同网络媒体逐一进行网络信息交互重构，信息获取技术实现的工作量异常庞大。同时，新型网络通信协议正在不断得到应用，部分网络通信协议，尤其是视/音频信息的网络交互过程并未对外公开发布，无法直接通过网络交互重构实现对应协议发布信息获取。

正是由于通过网络交互过程编程重构机制，在实现媒体信息获取环节存在相当程度的技术局限性，在 Web 网站自动化功能/性能测试的启发下，浏览器模拟技术在网络媒体信息获取环节正得到越来越广泛的应用。基于浏览器模拟实现网络媒体发布信息获取的技术实现过程是，利用典型的 JSSh 客户端向内嵌 JSSh 服务器的网络浏览器发送 JavaScript 指令，指示网络浏览器开展网页表单自动填写，网页按钮/链接点击，网络身份认证交互，网页发布信息浏览，以及视/音频信息点播等系列操作。

在此基础上，JSSh 客户端进一步要求网络浏览器导出网页文本内容，存储网页图像信息，或在用于信息获取的计算机上对于正在播放的视/音频信息进行屏幕录像，最终面向各种类型的网络内容、各种形态的网络媒体实现发布信息获取，如图 5-3 所示。

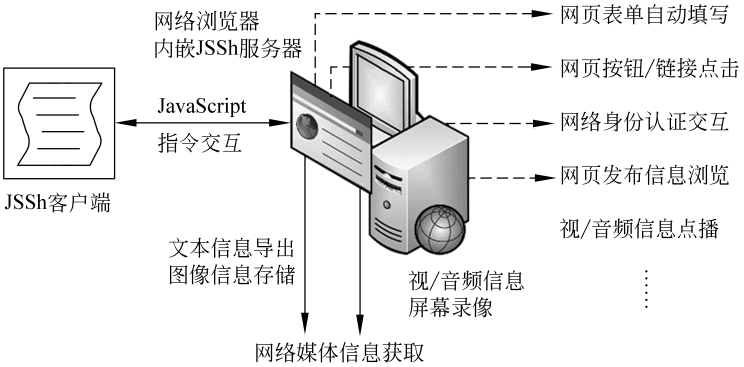

图 5-3　基于浏览器模拟实现网络媒体信息获取

5.3.2　信息内容获取的典型工具

网络爬虫是在互联网上实施信息内容获取的主要工具。网络爬虫是一种按照一定的规则，自动抓取互联网信息的程序或者脚本。互联网上的信息发布是分散的和独立的，但信息间又是相互连接的。爬虫就在超链接所建立的网上穿梭，这是爬虫又被称为蜘蛛的原因。

互联网信息资源非常庞大，在有限的网络资源的条件下，网络爬虫必须有选择性。针对不同的服务对象和行为，网络爬虫大体分为两类。一类是服务于搜索引擎等搜索类应用的网络爬虫，它的信息抓取规则是尽可能地覆盖更多的互联网网站，单一网站内的搜索深度要求不高。另一类是服务于针对性进行信息收集的应用的网络爬虫，例如，舆情分析系统要求它的网络爬虫具备高搜索深度和一定的主题选择能力。具有高搜索深度的爬虫被称为路径追溯爬虫，该类爬虫深入地尽可能抓取给定网站的全部资源；具有主题选择能力的爬虫被称为主题爬虫，该类爬虫会判断抓取的资源是否属于用户指定的主题，并持续

对有关给定主题的网页进行搜索和抓取。

通常，舆情分析系统采用的爬虫是以上介绍的两类爬虫的组合，并做一定的定制改动。随着网络技术的复杂化，网络爬虫也面临着越来越多的新问题，例如支持 Frame 的网页的处理，登录页面的处理等。其次，智能手持设备及相应应用（例如微信手机版）的发展，使得互联网资源的下载必须从单纯模拟浏览器浏览行为的爬虫，发展为能够模拟操作 APP 的爬虫。然后，对于个性化定制内容的网站（微博和微信都属于此类网站，每一个用户登录后所得到的信息内容均不相同），如何持续保持登录状态，如何自动修改定制（例如加关注）以得到更多信息，都是在此类网站抓取信息需要处理的问题。

网络爬虫通常采用分布式机制来保证信息获取的全面性和时效性。由于互联网资源规模巨大，而下载需要时间，所以网络爬虫都采用多进程或者多线程，甚至是分布式方式，同时下载多个网络资源（文本、图片、音频或者视频等），也就是说这是一项群体作业，爬虫们（下载器）集体一起完成抓取的任务（这也是网络爬虫也被称为蚂蚁的原因）。网络爬虫还需要避免过于频繁获取信息而被媒体网站判为“恶意”。一方面可通过适当选择周期遍历时间间隔，防止信息获取行为造成网络媒体负载过重；另一方面则可通过定期修改用于内容获取的网络客户端信息请求内容（内容协商行为），避免遭遇目标网络媒体的拒绝服务。

5.3.3 信息内容特征抽取与选择

信息内容的表示及其特征项的选取是数据挖掘、信息检索的一个基本问题，它把从信息中抽取出的特征词进行量化来表示文本信息。将它们从一个无结构的原始信息内容转化为结构化的计算机可以识别处理的信息，即对信息内容进行科学的抽象，建立它的数学模型，用以描述和代替信息内容，从而使计算机能够通过对这种模型的计算和操作来实现对信息内容的识别。

1. 文本信息内容的特征抽取与选择

对文本信息内容而言，由于文本是非结构化的数据，要想从大量的文本中挖掘有用的信息就必须首先将文本转化为可处理的结构化形式。目前人们通常采用向量空间模型来描述文本向量，但是如果直接用分词算法和词频统计方法得到的特征项来表示文本向量中的各个维，那么这个向量的维度将非常大。这种未经处理的文本矢量不仅给后续工作带来巨大的计算开销，使整个处理过程的效率非常低下，而且会损害分类、聚类算法的精确性，从而使得到的结果很难令人满意。因此，必须对文本向量做进一步净化处理，在保证原文含义的基础上，找出对文本特征类别最具代表性的文本特征。为了解决这个问题，最有效的办法就是通过特征选择来降维。

文本特征选择对文本内容的过滤和分类、聚类处理、自动摘要以及用户兴趣模式发现、知识发现等有关方面的研究都有非常重要的影响。通常根据某个特征评估函数计算各个特征的评分值，然后按评分值对这些特征进行排序，选取若干个评分值最高的作为特征词，这就是特征选择。特征选取的方式有 4 种：

① 用映射或变换的方法把原始特征变换为较少的新特征。

② 从原始特征中挑选出一些最具代表性的特征。

③ 根据专家的知识挑选最有影响的特征。

④ 用数学的方法进行选取，找出最具分类信息的特征，这种方法是一种比较精确的方法，人为因素的干扰较少，尤其适合于文本自动分类挖掘系统的应用。

特征选择已经有了很多成熟的方法，绝大多数都是基于统计的。信噪比(Signal-to-Noise Ratio)源于信号处理领域，表示信号强度与背景噪声的差值。如果将特征项作为一个信号来看待，那么特征项的信噪比可以作为该特征项对文本类别区分能力的体现。信息增益(Information Gain)是机器学习领域，尤其是构建决策树分类器时常采用的特征选择方法，信息增益也利用到信息熵的概念，依据特征项与类别标签之间的统计关系作为评价指标。卡方统计(Chi-Square Statistic)的判断依据是特征项与类别标签的相关程度。认为一个特征项与某个类别如果满足同时出现的情况，则说明该特征项能比较好地代表该类别。当单纯的特征选择无法满足信息表示的要求时，需要进行特征重构。特征重构以特征项集合为输入，利用对特征项的组合或转换生成新的特征集合作为输出。

2. 音频信息内容的特征抽取与选择

对于音频信息内容，充分地分析和提取其物理特征(例如频谱等)、听觉特征(例如响度、音色等)和语义特征(例如语音的关键词、音乐的旋律节奏等)，有效地实现音频信息的内容分类和检索至关重要。根据检索对象和检索方法的不同，国内外在音频检索方面的研究大致分为语音检索、音乐内容检索和音乐例子检索几类。音频检索第一步是建立数据库，对音频数据进行特征提取，并通过特征对数据聚类。然后检索引擎对特征向量与聚类参数集匹配，按相关性排序后通过查询接口返回给用户。音频信号的特征抽取指提取音频的时域和频域特征，将不同内容的音频数据予以区分。因此，所选取的特征应该能够充分地反映音频的物理和听觉特征，对环境的改变具有较好的鲁棒性。在进行音频特征抽取时，通常将音频划分为等长的片段，在每个片段内有划分帧。这样，特征抽取所采用的特征包括基于帧的特征和基于片段的特征两种。

基于帧的音频特征主要有以下几种：

① MFCC：语音识别中十分重要的特征，在音频应用中也有很好的效果，它是基于 Mel 频率的倒谱系数(Mel Frequency Cepstrum Coefficient)。由于 MFCC 参数将人耳的听觉感知特性和语音的产生机制相结合，因此得到广泛的使用。

② 频域能量：可以用来根据阈值判别静音帧，是区分音乐和语言的有效特征，通常语音中含有比音乐中更多的静音，因此语音的频域能量比音乐中的变化大得多。

③ 子带能量比：将频带划分为几个区间，其中每个区间称为子带，一般采用非均匀的划分方式，特别是 Bark 尺度或 ERB 尺度。不同类型的音频，其能量在各个子带区间的分布有所不同，音乐的频域能量在各个子带上的分布比较均匀，而语音的能量主要集中在第 1 个子带上，往往占 80%左右。

④ 过零率：描述音频信号通过过零值的次数，是信号频率的一个简单度量，可以在一定程度上反映其频谱的粗略估计。通常语音信号由发音音节和不发音音节交替构成，音乐没有这种结构；语音信号中，清音的过零率高，浊音的过零率低。所以过零率在语音

信号的变化要比在音乐的变化剧烈。

⑤ 基音频率：在周期或准周期音频信号中，声音的成分主要由基频（基音频率）及其谐波组成，而对于非周期信号则不存在基频。基音频率可以反映音调的高低，可以采用短时自相关方法进行粗略计算。

根据上面介绍的帧层次的基本特征，在音频处理中，常在片段层次上计算这些特征的统计值，作为该片段的分类特征。常见的基于片段的音频特征主要有以下几种：

① 静音帧率：如果一帧的能量和过零率小于给定的阈值，一般认为该帧是静音帧，否则该帧是非静音帧。静音帧率为静音帧数与片段中帧总数的比例。语音中经常有停顿的地方，所以其静音帧率一般比音乐的高。

② 高过零率帧率：根据对过零率特征的分析，语音由清音和浊音交替构成，而音乐不具有这种结构，因此，过零率在语音信号中要高于音乐信号中。对于一个片段来说，语音信号过零率高于阈值的比例高于音乐信号中的比例。

③ 低能量帧率：低能量帧率（Low Energy Frame Ratio，LER）是指一段音频信号中能量低于阈值的比例。一般来说，语音比音乐含有更多的静音帧，因此语音信号的低能量帧率高于音乐信号。

④ 谱通量：谱通量（Spectrum Flux，SF）也称为频谱流量，指片段中相邻帧之间谱变化的平均值。从整体上看，语音信号的谱通量数值较高，而音乐信号的谱通量往往较小，其他声音的谱通量数值介于两者之间。

⑤ 和谐度：如果一帧信号不存在基频，可以认为其基频为零。这样就可以用片段中基音频率不等于零的帧数所占的比例来衡量该音频片段的和谐程度。一般来说，语音在低频频带的和谐度较高，高频频带的和谐度较低；而音乐在整个频率范围内都具有较高的和谐度。

3. 图像信息的特征抽取与选择

相比文本信息而言，数字图像具有信息量大、像素点之间的关联性强等特点。因此，对于数字图像的处理方法与文本处理方法有较大的差别。图像的特征抽取和选择主要包含以下几个方面：

（1）图像颜色特征提取

所谓图像的颜色特征，通俗地说，即能够用来表示图像颜色分布特点的特征向量。常见的颜色特征有：颜色直方图、颜色聚合矢量、颜色矩等。

所谓颜色直方图（Color Histogram），即反映特定图像中的颜色级与出现该种颜色的概率之间关系的图形。颜色直方图仅仅从某种颜色出现的概率来描述图像的颜色特征。然而，完全不同的图像可能具有类似的直方图。为了能够方便区分该种情况，需要引入颜色以外的信息。颜色聚合矢量（Color Coherence Vector，CCV）的出发点在于引入一定的空间信息来进一步区分颜色分布类似而空间分布不同的图像。颜色矩（Color Moments）是一种统计特征，用来反映图像中颜色分布的特点，通过引入统计学中低阶矩（Moment）的概念，来描述整个图像的颜色变化情况。在图像分类、索引等应用中，可以通过计算颜色矩的距离来反映图像之间的相似程度。常见的颜色矩往往假定图像内的某

种颜色符合特定的概率分布，在此基础上选择有鉴别力的统计特征。

(2) 图像纹理特征提取

图像纹理特征提取能够用来表示图像纹理(亮度变化)特点的特征向量。纹理信息是亮度信息和空间信息的结合体，反映了图像的亮度变化情况。常见的纹理特征有：灰度共生矩阵、Gabor 小波特征、Tamura 纹理特征等。

- 灰度共生矩阵(Grey level co-occurrence matrix，GLCM)是最早期用于描述纹理特征的方法。灰度共生矩阵的元素 $P(i, j)$代表相距一定距离的两个像素点，分别具有灰度值 i 和 j 的出现概率。该矩阵依赖于这两个像素之间的距离(记作 dist)，以及这两个像素连线与水平轴的夹角(记作θ)，改变这两个参数能够得到不同的矩阵。共生矩阵反映了图像灰度分布关于方向、局部邻域和变化幅度的综合信息。一旦矩阵 P 确定了，就能够从中提取代表该矩阵的特征，一般可分为四类：视觉纹理特征、统计特征、信息特征和信息相关性特征。
- Gabor 小波特征(Gabor Wavelet Feature)是一种特殊的小波特征，其基本原理是通过小波变换对原有图像进行滤波(filtering)处理，然后对于滤波后的图像提取相关有鉴别力的特征。小波特征的鉴别力往往取决于小波基的选取。相比金字塔结构的小波变换(PWT)、树结构的小波变换(TWT)等，Gabor 小波更符合人眼对于图像的响应，故而常常用于描述图像的纹理特征。
- Tamura 等人根据人类视觉感知系统的特点，定义了 6 种与之相适应的纹理特征：Tamura 粗糙度(Coarseness)、对比度(Contrast)、方向性(Directionality)、线相似性(Line-likeness)、规则性(Regularity)和粗略度(Roughness)。

(3) 其他图像特征

除了以上两种常用的图像特征，现有的图像分类、检索系统中还使用边缘特征和轮廓特征。

边缘指的是灰度(颜色)存在较大差异的像素点，一般边缘点存在于目标/背景的分界处，或者目标内部纹理区域。这些信息都从一定侧面反映了图像的内容。因此，边缘特征也常常被用于图像分类、理解系统之中；轮廓特征是用来描述图像内某些目标物体的轮廓信息，从而为识别目标物体提供形状方面的信息，进而为理解图像内容提供线索。

5.4 信息内容分析与处理

海量信息内容分析的基本处理环节可以归结为分类和过滤，其他更加复杂的处理问题则是上述简单处理问题的组合。在信息检索和文本编辑等应用中，快速对用户定义的模式或者短语进行分类是最常见的需求。在文本信息过滤的处理中，分类算法也一直是人们所关注的。一个高效的分类算法会使信息处理变得迅速而准确，从而得到使用者的认可；反之，会使处理过程变得冗长而模糊，让人难以忍受。

5.4.1 信息内容分类

分类算法在图像分类、索引和内容理解方面都有直接的应用，其主要功能是：通过分