网络空间中的内容安全

杨锃涛[1]、秦子茗[2]、刘迅[3]

摘要：网络中的信息繁杂多样，很多时候我们难以直接断定信息真假，也难以在危害国家社会安全的信息传播之前就将其发现并封禁。那么常见的危险信息又有哪些，如何通过机器识别来识别这类内容，又如何将其彻底删除？本文将从这三部分粗浅的介绍网络中的内容安全。

引言：互联网上与日俱增的内容不仅显示着互联网在大众生活中的比重不断提升，而指数增长的数据量让信息内容安全的监测难度也随之增加。近年来，通过我国技术人员的不断努力，确实地消除了大部分不安全的信息内容，但是加强我国有害信息治理，是摆在当前的一项十分迫切和重要的任务。

1 存在于互联网中的危险信息

网络用户的基数不断增多，用户获取网络内容变得越来越便捷，新的内容编辑也变得越来越容易，新内容传播的途径越来越多、速度越来越快。这不仅帮助了许多重要信息的传递与传播，也助长了不实或是违法消息的快速传播加大了这些不法信息对国家对社会的危害。网络中的危险信息危害的程度和内容存在差异，但不可否认的是所有的网络内容形式中都会有危险信息的存在，下面将从六个危险信息内容种类展开说明。

1.1 淫秽色情类有害信息

这类有害信息对社会的危害比较大，主要传播方式有：通过微信、QQ 等即时通信工具发布大量淫秽色情音视频、图片；借助 QT、YY 等语音平台组织色情直播表演；利用网盘存储工具出售色情资源；开设色情游戏、动漫网站进行传播；开设色情小说、电子书网站，或者在部分网站设置小说频道，登载淫秽、伦理小说；部分网站提供招嫖、伴游等色情服务；部分情感、两性网站栏目内容存在色情内容；浏览网页时广告栏、侧边栏以及弹出色情窗口，点击后即进入色情网站；网站客户端及手机应用程序存在色情信息。

近年来国家对这类有害信息打击力度极大，但仍然没有做到全部消除。一方面是因为网站依托的服务器在境外，且存在多个备用网站可以随时跳转；另一方面，网盘涉及个人隐私，清理网盘中的色情资源存在难点。

1.2 政治类有害信息

这类有害信息主要类型有：散布危害国家安全、泄露国家秘密、颠覆国家

---

[1] 负责第一部分、摘要、引言
[2] 负责第二部分
[3] 负责第三部分、结语

政权、破坏国家统一的言论；散布煽动示威、游行等信息影响社会稳定；针对重大突发事件传播谣言；侮辱革命先烈，歪曲历史；捏造谣言，诬蔑、抹黑党和国家领导人；煽动民族仇恨、民族歧视、民族分裂，破坏民族团结；攻击国家宗教政策，宣扬邪教和封建迷信等。

1.3 诈骗类有害信息

网络的出现为诈骗提供了新方式，主要方式是冒充政府官方网站、各大银行、通信运营商、第三方支付平台等钓鱼网站，发布虚假信息，开设虚假功能页面从而得到用户的个人信息以及重要账户密码；开设游戏交易网站，网民充值后以信息输入有误为由冻结账户，诱骗网民再充相等金额方可解冻等。

相比于传统的诈骗，网络中的诈骗更偏向于守株待兔而不是像以前一样的通过话术来诱导受害人受骗。通过以假乱真的网站，往往比直接用话术诈骗更让人难以发现。

1.4 侵权类有害信息

最明显的例子就是侵犯著作权信息。未经著作权人许可，在互联网上发表其作品，歪曲、篡改、搬运并在互联网上发表他人单位作品。最常见的就是在各大视频网站互相转载视频，利用著作权人并未在全网注册账号的漏洞提前抢注并发布视频盈利；或是各大盗版网文网站将各大官方网文网站内收费的网文免费发布在网站上。

1.5 血腥暴力恐怖类有害信息

据统计，70%以上的暴恐类有害信息来源于境外，并在国内广泛传播。主要有：通过微博、贴吧、论坛等平台传播恐怖分子处决人质、近距离战争、车祸现场等视频和图片；开设网站大肆宣扬极端民族主义和宗教极端思想；通过文库网站发布教唆犯罪、教授杀人、处理尸首方法并附有相应图片的文章；通过即时通信工具、网站等出售枪支、管制刀具、毒品等违禁品，传授犯罪方法。

这一类信息尤其对青少年危害严重。在当下互联网获取途径越发便捷的情况下，对三观仍未确立的青少年来说，极有可能培养对方的暴力倾向。

2 如何判定网络上的图片，视频是否有害

互联网上与日俱增的内容不仅代表着更多的流量，也预示着巨大的内容风险藏身其中。色情、毒品、反动、暴恐、血腥、武器等等不良、有害信息不仅危害互联网平台的内容生态，更可能导致安全问题，使业务发展遭受损失。如何检测出其中的有害内容，避免携带病毒或者传播负能量甚至违法内容的图片或视频在互联网上传播，从而维护互联网用户的个人安全以及身心健康便成为了内容安全的一大任务。

2.1 检测图片或视频是否安全

图片与视频同样是带有信息的数据，在上学期的“计算机科学导论”课程中我们接触到了图像隐藏的方法，即将一段文本隐藏于图片之中，通过解析图片可以解析出该段文本。同理，黑客也可将一段恶意代码植入图片数据中，用户点击加载图片的同时也将木马下载并运行。例如，前些年爆出的“PNG 图片漏洞”，会影响 Firefox 以及部分使用 WebKit 内核的产品，使这些浏览器显示某些网页图片时自动崩溃中毒。据安全分析人士介绍，当 Firefox 等浏览器显示网页上的 PNG 图片时，图片中的恶意代码会利用 Libpng 漏洞复制到内存中，自动下载运行木马。

2.2 检测图片或视频是否违法

视频数据主要由一系列有序静态图像集合构成,同时也可能包含音频、文本等数据元素。视频数据通过多维特征描述内容信息,特征有颜色、纹理、时间及空间等，故图片内容安全的检测可归于视频安全检测的范畴之内。视频内容的评价指标体系可以由下图表示：

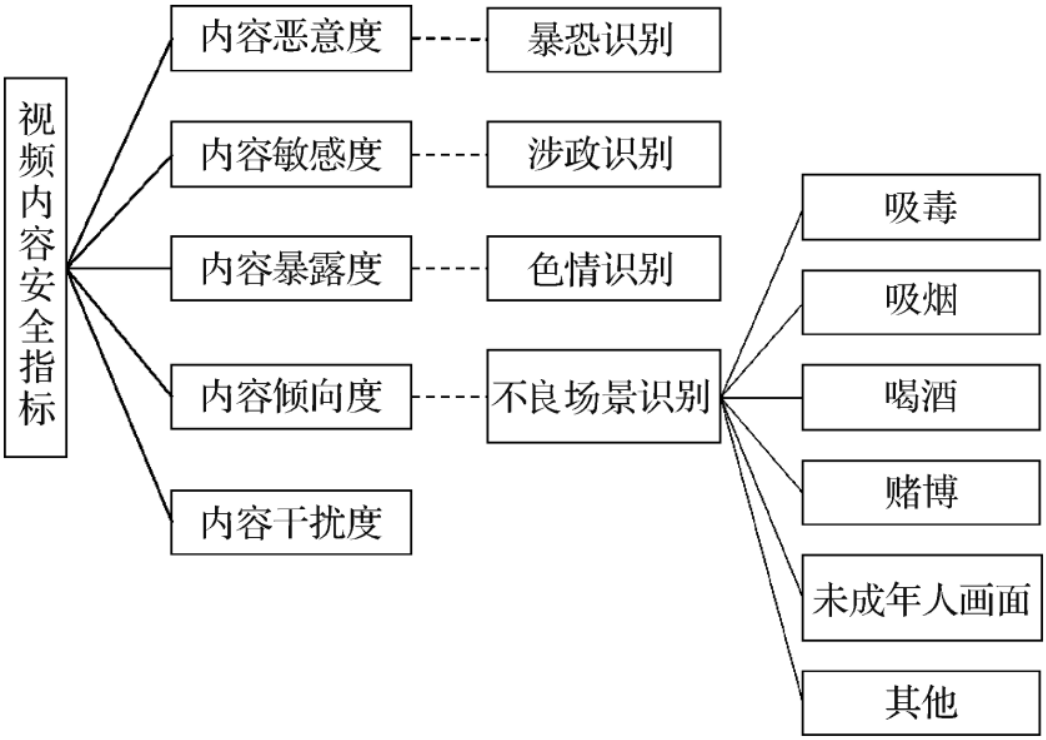

图 1：视频内容的评价指标体系

传统图像处理方法往往基于尺度不变特征变换（SIFT）：利用高斯差分函数来搜索兴趣点，再进一步精细拟合从中提取出图像的特征点，剔除其中非良性点后将剩余特征点用 128 维的特征向量进行描述。该种方法实用性强，已大范围应用于图像检测领域。

随着网络空间业务形式和内容越来越多元化，海量的多样化数据也在不断产生，对网络空间中数据及相互关系的分析和治理已成为必然，传统技术难以处理

现今海量的视图像数据。近年来，基于深度学习的视图像内容智能分析技术逐渐兴起，为该项工作的推进提供了极大的助力。在此我们介绍基于卷积神经网络（Convolutional Neural Networks, CNN）的视图像分类技术。该项技术将图像送入卷积神经网络中，然后开始对图像数据分类。比如输入一个大小为100×100的图像，只需创建一个大小为10×10的扫描输入层，扫描图像的前10×10个像素，向右移动一个像素，再扫描下一个10×10的像素。输入的数据被送入卷积层，每个节点只需要处理离自己最近的临近节点。卷积层也随着扫描的深入而趋于收缩。为了进一步提取高维特征并降低计算量，还需要对特征进行池化（即压缩图像）：假如得到一个局部特征，它是一个图像的一个局部放大图，分辨率很大，那么就可以将一些像素点周围的像素点（特征值）近似看待，然后统计平面内某一位置及其相邻位置的特征值，并将汇总后的结果作为这一位置在该平面的值。除此之外还有目标检测识别技术，跨媒体智能感知技术等前沿视图像内容智能分析技术。

但目前基于人工智能的视图像分析技术仍存在瓶颈，例如在视频直播中无法单独识别非法活动，难以处理全平台高流量的直播内容，对“数据投毒”识别能力弱等，还需要一线科研与工程人员的进一步努力。

3 彻底删除有害内容的策略

在上述内容的讨论中，我们界定了哪些内容是有害的、如何用自动化的方式识别出有害的图片视频内容。更进一步，我们需要考虑如何在网络中彻底删除这些有害内容。这一问题可以根据管理权限，分为内容自身和内容索引两方面考虑：在可管理的范围内，拥有读写权限的管理者主动地把有害内容删除；对于不可控的其他网络部分，管理者转而删除内容的索引，阻止有害内容扩散到其他的网络部分，从而实现对外的有害内容删除。

3.1 针对内容自身的删除

假如我们拥有完全的控制权限，在可以管理的范围内，希望彻底删除有害内容，最根本的方法便是删除目标内容自身。这一策略，要求避免内容在删除后被恢复。

3.1.1 物理删除

最原始的删除方法，莫过于物理上的彻底删除。由于一切网络世界都是依附于现实世界的实体存在，通过对内容存储介质的充分处理，所存储的内容也随之被彻彻底底地删除了。

常见的有盘片划损、硬盘回炉、外力破损等方法，但这些方法普遍存在费时、费力、效果不佳的问题，同时对资源的耗费大，并没有被广泛采用。[3]

3.1.2 重复覆写

重复覆写同样是针对物理介质的，但在物理删除的原始手段上更进了一步。由于存储介质的特性，内容被读写删除之后，还有可能被特殊手段恢复，因而需

要多次针对性的覆写来抹除内容的历史存储信息。

常见的规范有美国国防部的 DOD 标准。DOD 标准要求数据删除时需要重写七次，每次先用 0 或 1 写入，之后再用随机的 0 或 1 来覆写。

3.1.3 依赖时间的加密

从密码学的角度考虑，我们可以对数据加密，其中密钥依赖于当前时间。管理者拥有加密与内容分发的控制权，根据时间维护加密内容，对外只发布加密后的内容。这样访问内容便依赖于管理者提供的、随着时间更新的解密密钥。当内容被判定为有害后，我们随即停止对加密的维护，之后的访问者由于没有当前时间的解密密钥，便读取不了有害内容。

3.2 针对内容索引的删除

内容的生命在于传播。这提示着我们，不仅可以依靠主动地删除内容自身，还可以通过被动地删除内容与外界的联系，阻断外界访问这个内容，从而实现彻底删除有害内容。

3.2.1 主动：关键词屏蔽

关键词屏蔽是一种传统但仍十分有效的策略。假如我们拥有搜索引擎的控制权，那么在过滤器中设置对应的关键词，便可以阻止显示对应的有害内容。

与之相对应的对抗策略是加入无意义字符、替换同音或相近含义等隐晦指代词来混淆内容，从而躲避字符串匹配的过滤规则。但随着 NLP 技术的发展，我们可以用语义、词法分析等手段实现含关键词内容的屏蔽，既能够检测更多混淆内容，还能够避免单纯的字符串匹配关键词带来的误查杀问题。

3.2.2 被动：反爬虫技术

网络空间中绝大部分的内容访问，都是基于某种平台的，例如百度搜索、Google 搜索此类搜索引擎或是微博、知乎等内容平台的内容索引。而平台中内容索引的构建，几乎完全依赖于网络爬虫技术。[4]因而，我们可以采用反爬虫技术以阻断网络其他部分对有害内容的访问。

正规的搜索引擎，通常遵循一个约定俗成的规范：首先在网站下的 robots.txt 查询网站对爬取的限制。在这个文件中，网站指定了哪些内容可以被爬取，而哪些内容不可以，甚至这个网站自身能否被爬取。通过设定 robots.txt 文件，我们进而一定程度上阻断了有害内容被索引。

但还要考虑“不守规矩”的爬虫的情形。对于更一般的爬虫情形，常见的反爬虫措施有请求头的预处理、访问数量的并发限制、网页数据的异步加载、设置验证码等。通过反爬虫的技术手段，我们可以阻止内容被网络其他部分索引，也就是扼杀了有害内容的传播力、生命力。

结语：本文首先讨论有害内容的概念如何界定，进而讨论如何用传统图像处理技术、新兴人工智能技术自动化识别网络中的有害内容，最后讨论在界定、识别之

后如何彻底删除有害内容。跟随着有害内容的系列处理流程，小组成员认识到网络空间治理从现实的观念到网络的具体实践究竟是如何作用，以及从观念到实践中现今仍然面临的诸多挑战。然而，挑战，也即意味着机遇。一方面，这说明还存在很大发展进步的空间。另一方面，挑战之所为挑战，意味着这一问题是值得关注的。网络空间的治理，将会是未来很长一段时期内互联网在早期蓬勃发展阶段之后不可绕过的重要问题。现阶段的挑战，恰恰提示着我们，这一领域在未来大有可为。

参考文献：

[1]刘涵. 网络有害信息的类型及治理措施[J]. 管理观察,2016(17):53-56+59.

[2]沈宜,郭先会,石珺. 数据智能在内容安全治理中的应用 [J].通信技术,2022,55(8):1065-1072.

[3]尹燕彬,文伟平. 计算机数据安全删除和隐私保护[J]. 信息网络安全,2009(05):55-58.

[4]刘石磊. 对反爬虫网站的应对策略[J]. 电脑知识与技术,2017,13(15):19-21+23.