第9章 内容安全

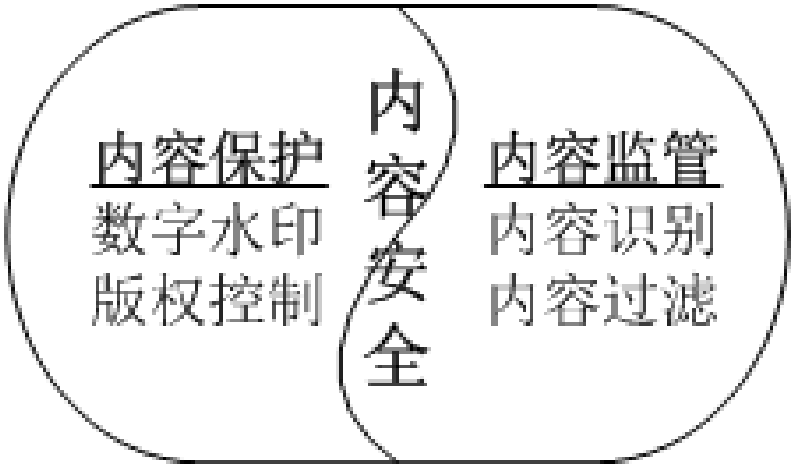
罗文坚

主要内容

- 9.1 概述
  - 9.1.1 内容保护
  - 9.1.2 内容监管
- 9.2 版权保护
  - 9.2.1 DRM概述
  - 9.2.2 数字水印
- 9.3 内容监管
  - 9.3.1 网络信息内容监管
  - 9.3.2 垃圾邮件处理

概述

- 信息内容安全有两方面内容:
  - 一方面是指针对合法的信息内容加以安全保护，如对合法的音像制品及软件的版权保护；
  - 另一方面是指针对非法的信息内容实施监管，如对网络暴力色情信息的过滤等。



内容保护

- 互联网的发展与普及使电子出版物的传播和交易变得便捷，侵权盗版活动也呈日益猖獗之势。

- 为了打击盗版犯罪：
  - 一方面，要通过立法来加强对知识产权的保护；
  - 另一方面，要有先进的技术手段来保障法律的实施。

- 内容保护技术大多数都是基于密码学和隐写术发展起来的：
  - 如数据锁定、隐写标记、数字水印和数字版权管理DRM等技术，其中最具有发展前景和实用价值的是数字水印和数字版权管理。

内容保护

- 信息隐藏和信息加密的区别：
  - 信息隐藏和信息加密都是为了保护秘密信息的存储和传输，使之免遭敌手的破坏和攻击，但两者之间有着显著的区别。
  - 信息加密是利用对称密钥密码或公开密钥密码把明文变换成密文；信息加密所保护的是信息的内容。
  - 信息隐藏是将秘密信息嵌入到表面上看起来无害的宿主信息中，攻击者无法直观地判断他所监视的信息中是否含有秘密信息。
    - 换句话说，含有隐匿信息的宿主信息不会引起别人的注意和怀疑，同时隐匿信息又能够为版权者提供一定的版权保护。

内容保护技术

- 数据锁定是指出版商把多个软件或电子出版物集成到一张光盘上出售，盘上所有的内容均被分别进行加密锁定，不同的用户买到的均是相同的光盘，每个用户只需付款买他所需内容的相应密钥，即可利用该密钥对所需内容解除锁定，而其余不需要的内容仍处于锁定状态，用户是无法使用的。
  - 在Internet上，数据锁定技术可以用于FTP服务器或Web站点上的数据保护，付费用户可以利用特定的密钥对所需要的内容解除锁定。

- 隐匿标记是指利用文字或图像的格式（如间距、颜色等）特征隐藏特定信息。
  - 例如，在文本文件中，字与字间、行与行间均有一定的空白间隔，把这些空白间隔精心改变后可以隐藏某种编码的标记信息以识别版权所有者，而文件中的文字内容不需作任何改动。

内容保护技术

- 数字水印是镶嵌在数据中，并且不影响合法使用的具有可鉴别性的数据。它一般应当具有不可察觉性、抗擦除性、稳健性和可解码性。
  - 为了保护版权，可以在数字视频内容中嵌入水印信号。
  - 如果制定某种标准，可以使数字视频播放机能够鉴别到水印，一旦发现在可写光盘上有“不许拷贝”的水印，表明这是一张经非法拷贝的光盘，因而拒绝播放。还可以使数字视频拷贝机检测水印信息，如果发现“不许拷贝”的水印，就不去拷贝相应内容。
- 数字版权管理DRM（Digital Rights Management）技术是专门用来保护数字化版权的产品。
  - DRM的核心是数据加密和权限管理，同时也包含了上述提到的几种技术。DRM特别适合基于互联网应用的数字版权保护，目前已经成为数字媒体的主要版权保护手段。

内容监管

- 在对合法信息进行有效的内容保护同时，针对虚假信息（如疫情期间的各种谣言）的监管，针对大量的充斥暴力色情等非法内容的媒体信息（特别是网络媒体信息）的内容监管，也是十分必要。

- 面向网络信息内容的监管主要涉及两类：
  - 一类是静态信息，主要是存在于各个网站中的数据信息，例如挂马网站的有关网页、色情网站上的有害内容以及钓鱼网站上的虚假信息等；
  - 另一类是动态信息，主要是在网络中流动的数据信息，例如网络中传输的垃圾邮件、色情及虚假网页信息等。
  - 无论是有害的网站静态信息，还是正在网络上传输的动态有害信息，都会对社会造成极大危害，因此，必须对它们进行有效监管。

内容监管技术

- 针对静态信息的内容监管技术主要包括网站数据获取技术、内容分析技术、控管技术等。
  - 控管技术是指对违法的网站实施有效的控制管理，将其危害性减少到最低程度，主要涉及阻断对有害网站的访问以及报警等技术。

- 对于动态信息进行内容监管所采取的技术主要包括网络数据获取技术、内容分析技术、控管技术等
  - 网络数据获取技术是指通过在网络关键路径上设置数据采集点，以监听捕获通过该路径的所有网络报文数据。
  - 有关内容分析技术和控管技术部分基本上与对静态信息采取的处理技术相同。

主要内容

- 9.1 概述
  - 9.1.1 内容保护
  - 9.1.2 内容监管
- 9.2 版权保护
  - 9.2.1 DRM概述
  - 9.2.2 数字水印
- 9.3 内容监管
  - 9.3.1 网络信息内容监管
  - 9.3.2 垃圾邮件处理

版权保护

- 版权（又称著作权）保护是内容保护的重要部分，其最终目的不是“如何防止使用”，而是“如何控制使用”，版权保护的实质是一种控制版权作品使用的机制。

- 数字版权保护技术DRM (Digital Rights Management)就是以一定安全算法实现对数字内容的保护。
  - DRM目的是从技术上防止数字内容的非法复制， 或者在一定程度上使非法复制变得很困难，用户必须在得到授权后才能使用数字内容。
  - DRM涉及的主要技术包括数字标识技术、安全和加密技术以及安全存储技术等。
  - DRM技术方法主要有两类，一类是采用数字水印技术，另一类是以数据加密和防拷贝为核心的DRM 技术。

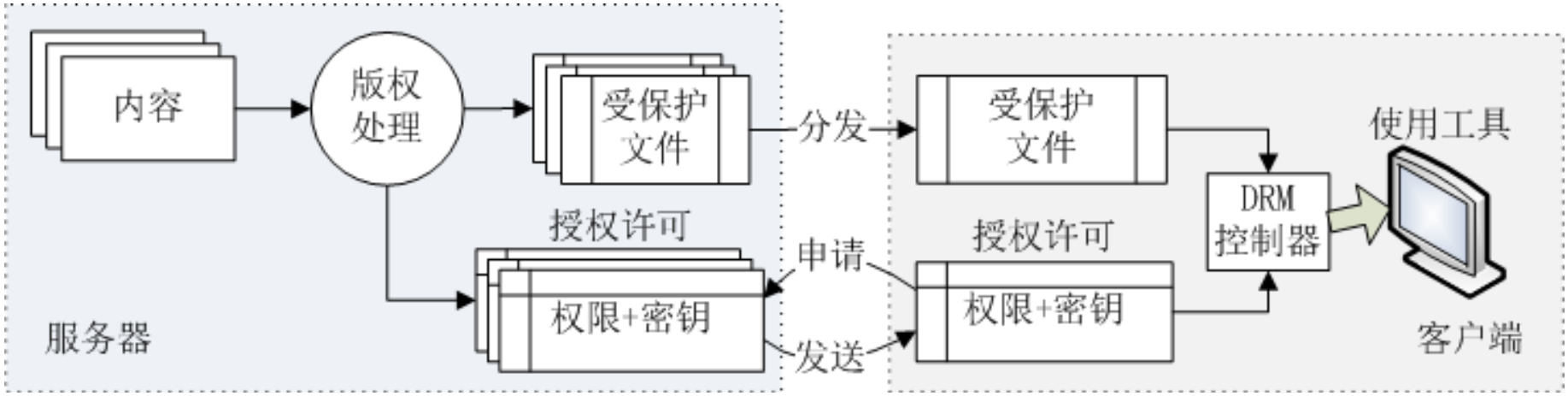
DRM概述

- DRM技术自产生以来，得到了工业界和学术界的普遍关注，被视为是数字内容交易和传播的关键技术。

- 国际上许多著名的计算机公司和研究机构纷纷推出了各自的产品和系统。
  - 例如，Microsoft WMRM、IBM EMMS、Real Networks Helix DRM以及Adobe Content Server等。
  - 国内的DRM技术发展同样很快，特别是在电子书以及电子图书馆方面，如北大方正Apabi 数字版权保护技术、书生的SEP技术、超星的PDG等。
  - Microsoft的Windows XP操作系统和Office XP等系列软件中也使用了DRM技术。

DRM工作原理

- DRM系统结构分为服务器和客户端两部分。



主要版权保护产品

- 目前DRM所保护的内容主要分为三类
  - 包括电子书、音视频文件和电子文档。

- Adobe 公司的ACS（Adobe Content Server）软件
- 方正的Apabi数字版权保护软件，主要由Maker、Rights Server、 Retail Server和Reader四部分组成。
- Microsoft公司于1999年8月发布了Windows Media DRM。
  - 最新版本的Windows Media DRM 10 系列包括了服务器和软件开发包SDKs。
- RMS（Rights Management Services）， Microsoft公司
  - 适用于电子文档保护的数字内容管理系统。

主要内容

- 9.1 概述
  - 9.1.1 内容保护
  - 9.1.2 内容监管
- 9.2 版权保护
  - 9.2.1 DRM概述
  - 9.2.2 数字水印
- 9.3 内容监管
  - 9.3.1 网络信息内容监管
  - 9.3.2 垃圾邮件处理

数字水印

- 原始的水印Watermark是指在制作纸张过程中通过改变纸浆纤维密度的方法而形成的，“夹”在纸中而不是在纸的表面，迎光透视时可以清晰看到的有明暗纹理的图像或文字。
  - 人民币、购物卷以及有价证券等，以防止造假。
- 数字水印（digital watermark）也是用来证明一个数字产品的拥有权、真实性。
  - 数字水印是通过一些算法嵌入在数字产品中的数字信息，如产品的序列号、公司图像标志以及有特殊意义的文本等。
- 数字水印分为可见数字水印和不可见数字水印。
  - 可见水印主要用于声明对产品的所有权、著作权和来源，起到广告宣传或使用约束的作用。例如，电视台播放节目时的台标既起到广告宣传，又可声明所有权。
  - 不可见数字水印应用的层次更高，制作难度更大，应用面也更广。

数字水印原理

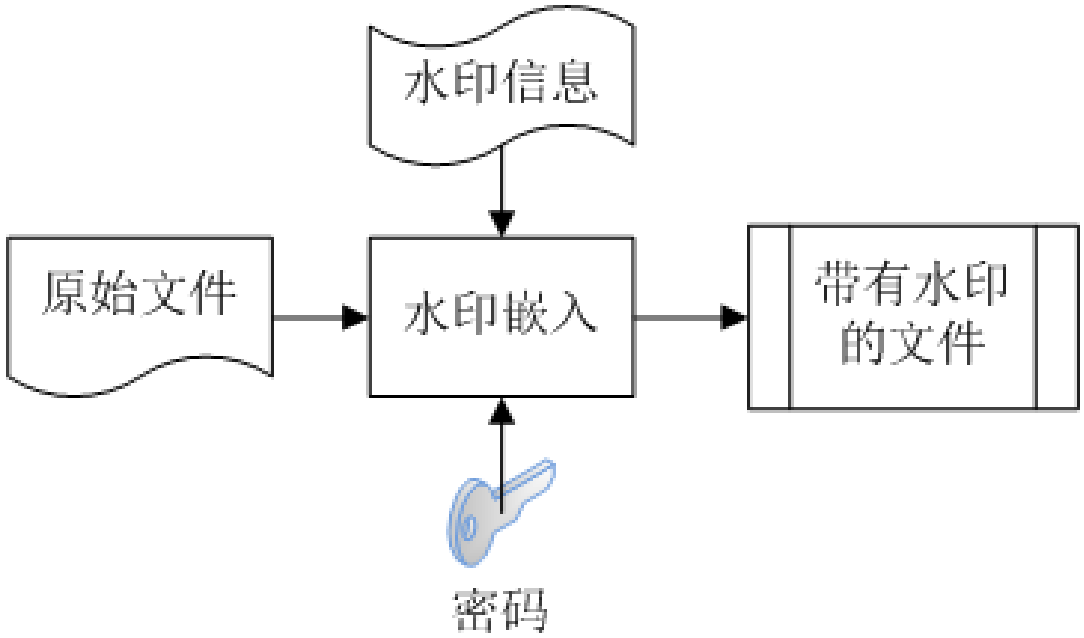
- 一个数字水印（后简称为水印）方案一般包括三个基本方面：水印的形成、水印的嵌入和水印的检测。

- 水印的形成主要是指选择有意义的数据，以特定形式生成水印信息，如有意义的文字、序列号、数字图像（商标、印鉴等）或者数字音频片段的编码。

- 一般来说，水印信息可以根据需要制作成可直接阅读的明文信息，也可以是经过加密处理后的密文。

水印的嵌入

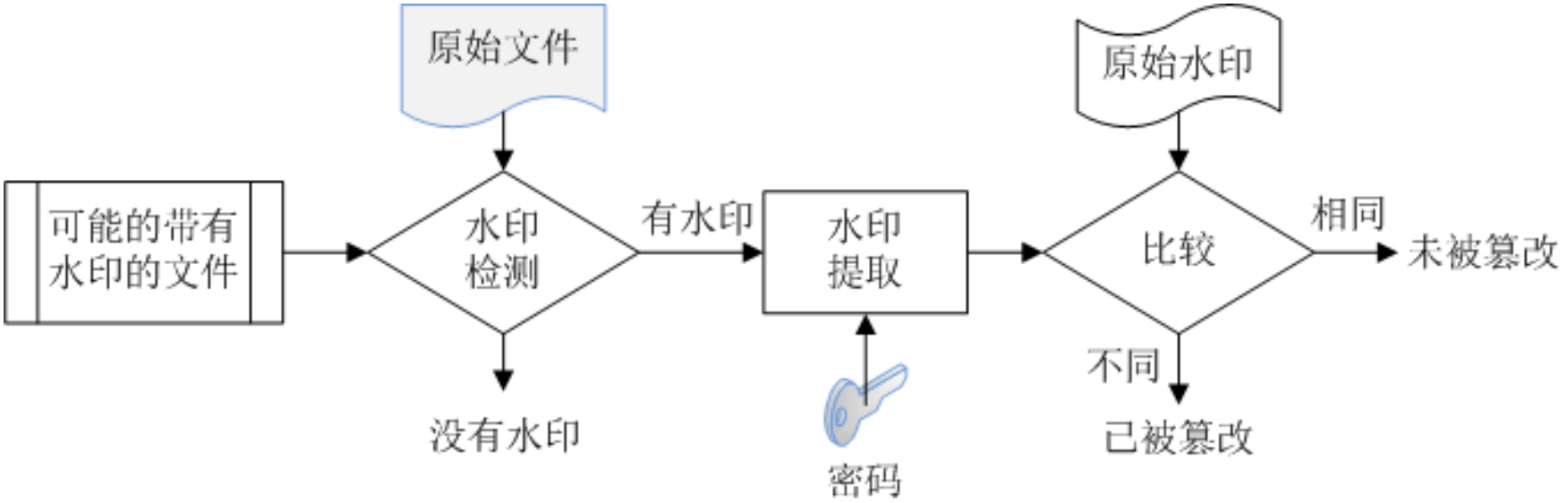
- 水印的嵌入与密码体系的加密环节类似，一般分为输入、嵌入处理和输出三部分。



- 嵌入处理完成的主要任务是对输入原始文件进行分析，选择嵌入点，在整个过程中可能需要密码参与。

水印的检测

- 水印的检测分为两部分：检测水印是否存在，提取水印信息。
- 水印的检测方式主要分为盲水印检测和非盲水印检测。
  - 盲水印检测主要指不需要原始数据（原始宿主文件和水印信息）参与，直接进行检测水印信号是否存在；
  - 非盲水印检测是在原始数据参与下进行水印检测。
  - 水印提取及比较主要针对不可见水印，一般可见水印可以直接由视觉识别。



数字水印的特征

- 数字水印的使用一般要以不破坏原始作品的欣赏价值和使用价值为原则，因此数字水印应具有以下基本特征：

1. 隐蔽性（不可见水印）：指水印与原始数据紧密结合并隐藏在其中，不影响原始数据正常使用的特性。

2. 鲁棒性：指嵌入的水印信息能够抵抗针对数字作品的各种恶意或非恶意的操作，即经过了各种攻击后是否还能提取水印信息。

3. 安全性：未授权者不能伪造水印或检测出水印。密码技术对水印的嵌入过程进行置乱加强安全性，从而避免没有密钥的使用者恢复和修改水印。

4. 易用性：指水印的嵌入和提取算法是否简单易用，主要指水印嵌入算法和水印提取算法的实用性和执行效率等。

数字水印算法

- 数字水印技术研究已经取得 很大的进步，出现了很多优秀的数字水印算法。

- 面向文本的水印算法
- 面向图像的水印算法
- 面向音视频的水印算法
- **NEC**算法
- 生理模型算法

面向文本的水印算法

- 纯文本文档，指ASCII码文档或计算机源代码文档。
  - 不存在可插入标记的可辨认空间，很难嵌入秘密信息，需要保护和认证的正式文档很少采用纯文本格式。

- 格式化的文档，一般指除了文本信息之外，有很多用来标记文字格式和版面布局的冗余信息，并可使用相关软件进行处理的文件。
  - 如Word文件、PDF文件等。
  - 对于这类文档，可以把水印信息嵌入到这类文档的格式化编排中。例如，行、字间距，字体，文字大小和颜色等不足以被人眼发现的微小变化都可以用来进行信息的隐藏。

面向文本的水印算法

- 基于文档结构微调的文本水印算法：主要指通过文本文档空间域的变换来嵌入数据。文档的空间域不仅包括文本的字符、行、段落的结构布局，也包括了字符的颜色和形状。
- 基于语法的文本水印算法：一类是按照语法规则对载体文本中的词汇进行替换来隐藏水印信息，另一类是按照语法规则对载体文本中的标点符号来进行修改隐藏水印信息。
- 基于语义的文本水印算法：将一段正常的语言文字修改为包含特定词语（如同义词）的语言文字，在这个修改过程中水印信息被嵌入到文本内。
- 基于汉字特点的文本水印算法：常见的基于汉字特点的文本水印有针对汉字的笔画特征（如倾斜角度）进行修改以嵌入水印信息，还有针对汉字的结构组合特征进行修改以嵌入水印信息，如将汉字看做二值图像，利用汉字结构中各部分的连通性嵌入水印信息。

面向图像的水印算法

- 空域数字图像水印算法主要是在图像的像素上直接进行的，通过修改图像的像素值嵌入数字水印的。
  - 经典的最低有效位LSB（Least Significant Bits）空域水印算法是以人类视觉系统不易感知为准则，在原始载体数据的最不重要的位置上嵌入数字水印信息。该算法的优势是可嵌入的水印容量大，不足是嵌入的水印信息很容易被移除。
- 变换域数字水印算法是在图像的变换域进行水印嵌入的。即，将原始图像经过正交变换，将水印嵌入到图像的变换系数中。
  - 常用的变换有离散傅里叶变换、离散余弦变换、离散小波变换等。
  - 在变换域中嵌入的数字水印能量可以扩展到空间域的所有像素上，有利于实现水印的不可感知性，还可增强水印的鲁棒性。

面向音视频的水印算法

- 根据音频水印载体类型，音频水印技术可分为基于原始音频和基于压缩音频两种。
  - 基于原始音频方法是在未经编码压缩的音频信号中直接嵌入水印。
  - 基于压缩音频方法指音频信号在压缩编码过程中嵌入水印信息，输出的是含水印的压缩编码的音频信号。

- 视频可以认为是由一系列连续的静止图像在时间域上构成的序列，因此视频水印技术与图像水印技术在应用模式和设计方案上具有相似之处。
  - 数字视频水印主要包括基于原始视频的水印、基于视频编码的水印和基于压缩视频的水印。

NEC算法

- NEC算法是由NEC实验室的Cox等人提出，在数字水印算法中占有重要地位。
- 水印信号应该嵌入到最容易让人察觉到变化的源数据部分。
  - 在频谱空间中，这种重要部分就是低频分量。这样，攻击者在破坏水印的过程中，不可避免地会引起图象质量的严重下降。
- 水印信号应该由具有高斯分布的独立同分布随机实数序列构成。这使得水印抵抗多拷贝联合攻击的能力大大增强。
  - 具体的实现方法：首先以密钥为种子来产生伪随机序列，该序列具有高斯N(0, 1)分布。密钥可以由作者的标识码和图像的哈希值组成，对整幅图像做离散余弦变换，用伪随机高斯序列来叠加该图像的1000个最大的DCT系数（除直流分量外）。
- NEC算法具有较强的鲁棒性、安全性、透明性等。

生理模型算法

- 人的生理模型包括人类视觉系统HVS（Human Visual System）和人类听觉系统HAS（Human auditory System）等。

- 生理模型算法的基本思想是利用人类视觉的掩蔽现象，从HVS模型导出的可觉察差异JND（Just Noticeable Difference）。
  - 利用JND描述来确定图像的各个部分所能容忍的数字水印信号的最大强度。

- 人类视觉对物体的亮度和纹理具有不同程度的感知性，可以调节嵌入水印信号的强度。
  - 亮度掩蔽特性：背景越亮，所嵌入水印的可见性越低；
  - 纹理掩蔽特性：纹理越复杂，所嵌入水印的可见性越低。

主要内容

- 9.1 概述
  - 9.1.1 内容保护
  - 9.1.2 内容监管
- 9.2 版权保护
  - 9.2.1 DRM概述
  - 9.2.2 数字水印
- 9.3 内容监管
  - 9.3.1 网络信息内容监管
  - 9.3.2 垃圾邮件处理

内容监管

- 内容监管是内容安全的另一重要方面，如果监管不善，会对社会造成极大的影响，其重要性不言而喻。

- 内容监管涉及到很多领域，其中基于网络的信息已经成为内容监管的首要目标。

- 一般来说病毒、木马、色情、反动、严重的虚假欺骗以及垃圾邮件等有害的网络信息都需要进行监管。

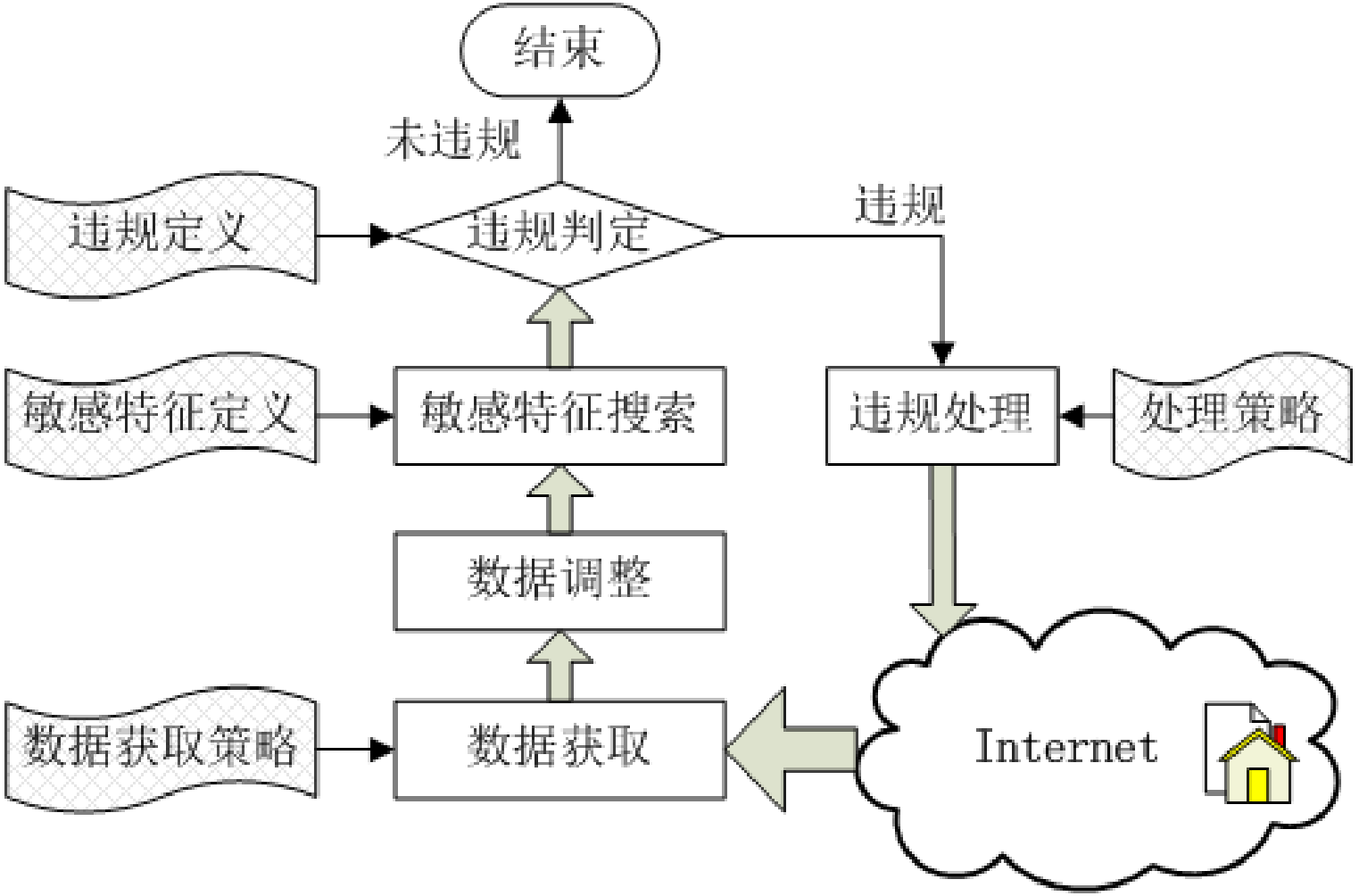
网络信息内容监管

- 内容监管首先需要解决的就是如何制定监管的总体策略。
  - 总体策略主要包括监管的对象、监管的内容、对违规内容如何处理等。

- 首先是如何界定违规内容（那些需要禁止的信息），既能够禁止违规内容，又不会殃及到合法应用。
- 其次是对于可能存在一些违规信息的网站如何处理。
  - 一种方法是通过防火墙禁止对该网站的全部访问，这样比较安全，但也会禁止掉其他有用内容；
  - 另一种方法是允许网站部分访问，只是对那些有害网页信息进行拦截，但此种方法存在拦截失败的可能性。

内容监管系统模型

- 内容监管系统模型可以分为监管策略和监管处理两部分。



内容监管策略

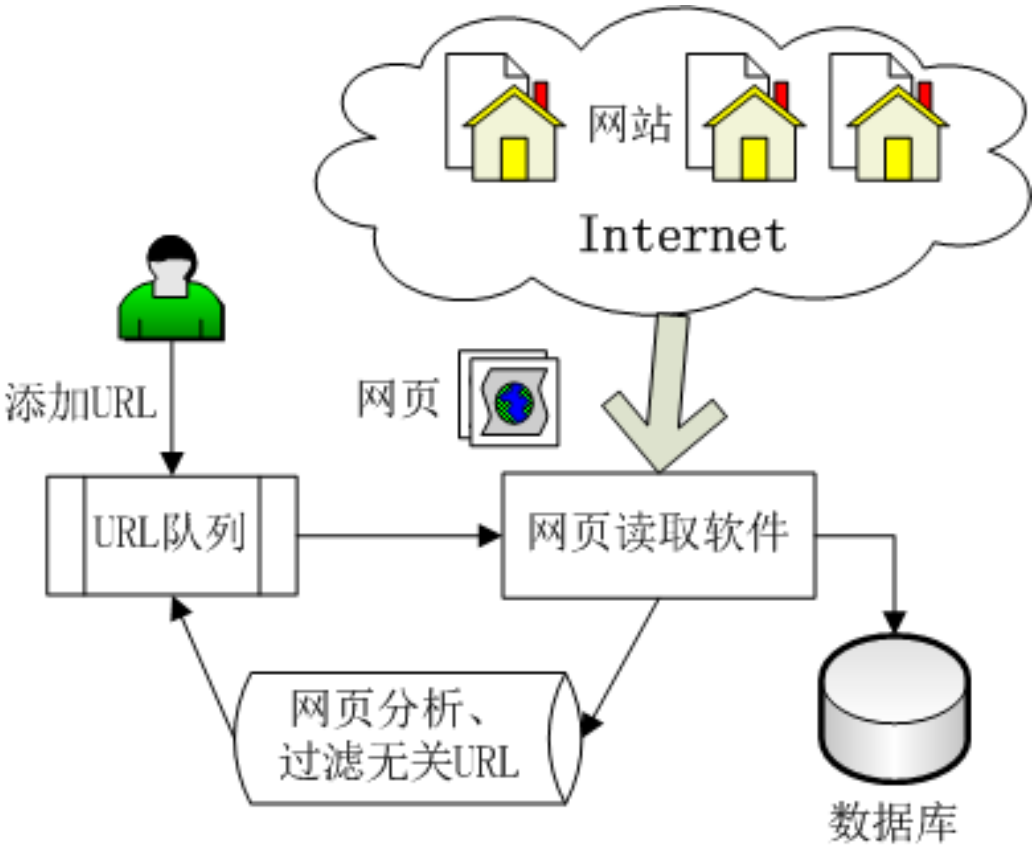
- 内容监管需求是制定内容监管策略的依据。
- 内容监管策略是内容监管需求的形式化表示。

- 内容监管策略包括：
  - 数据获取策略，主要确定监管对象的范围、采用何种方式获取需要检测的数据；
  - 敏感特征定义，是指用于判断网络信息内容是否违规的特征值，如敏感字符串、图片等；
  - 违规定义，是指依据网络信息内容中包含敏感特征值的情况判断是否违规的规则；
  - 违规处理策略，是指对于违规载体（网站或网络连接）的处理方法，如禁止对该网站的访问、拦截有关网络连接等。

数据获取

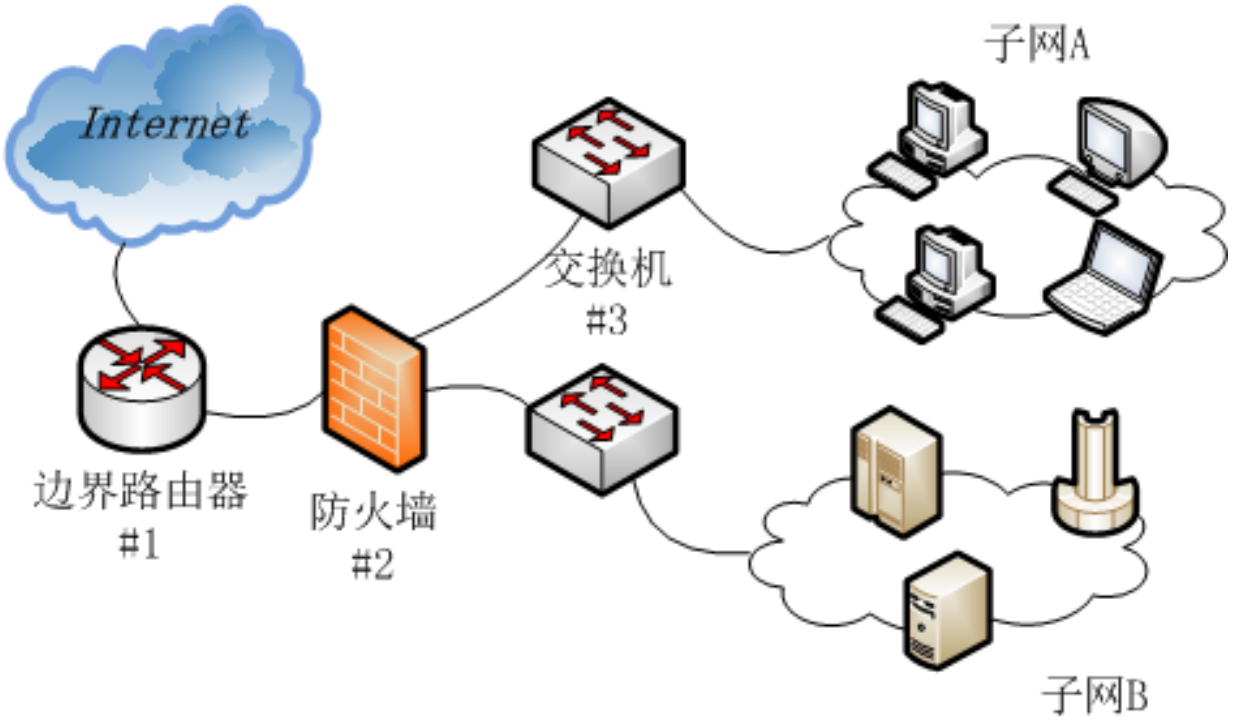
- 数据获取技术分为主动式和被动式两种形式。
- 主动式数据获取是指通过访问有关网络连接而获得其数据内容。网络爬虫是典型的主动式数据获取技术。

网络爬虫是如何工作的？



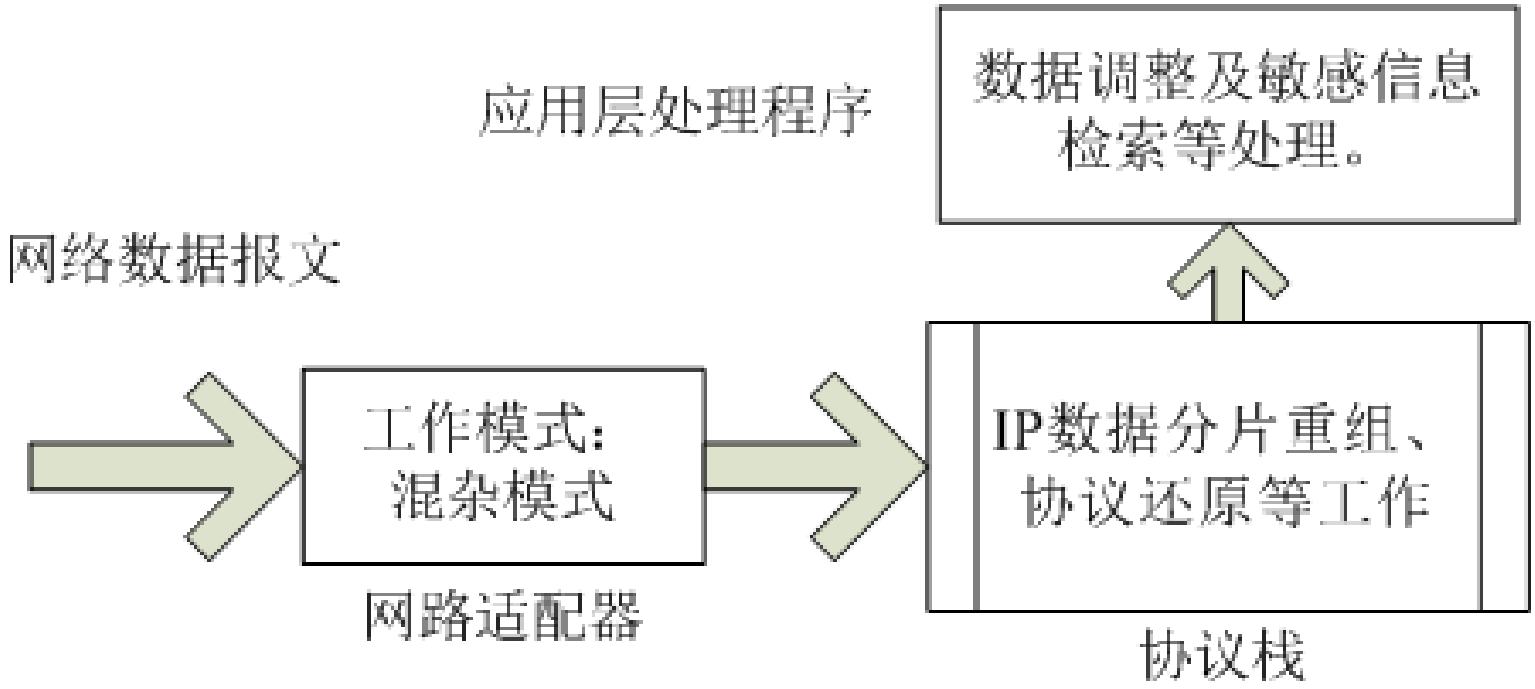
数据获取

- 被动式数据获取是指在网络的特定位置设置探针，获取流经该位置的所有数据。
- 被动式数据获取需要解决两个方面的问题：
  - 探针位置的选择：
  - 对出入数据报文的采集。



数据获取

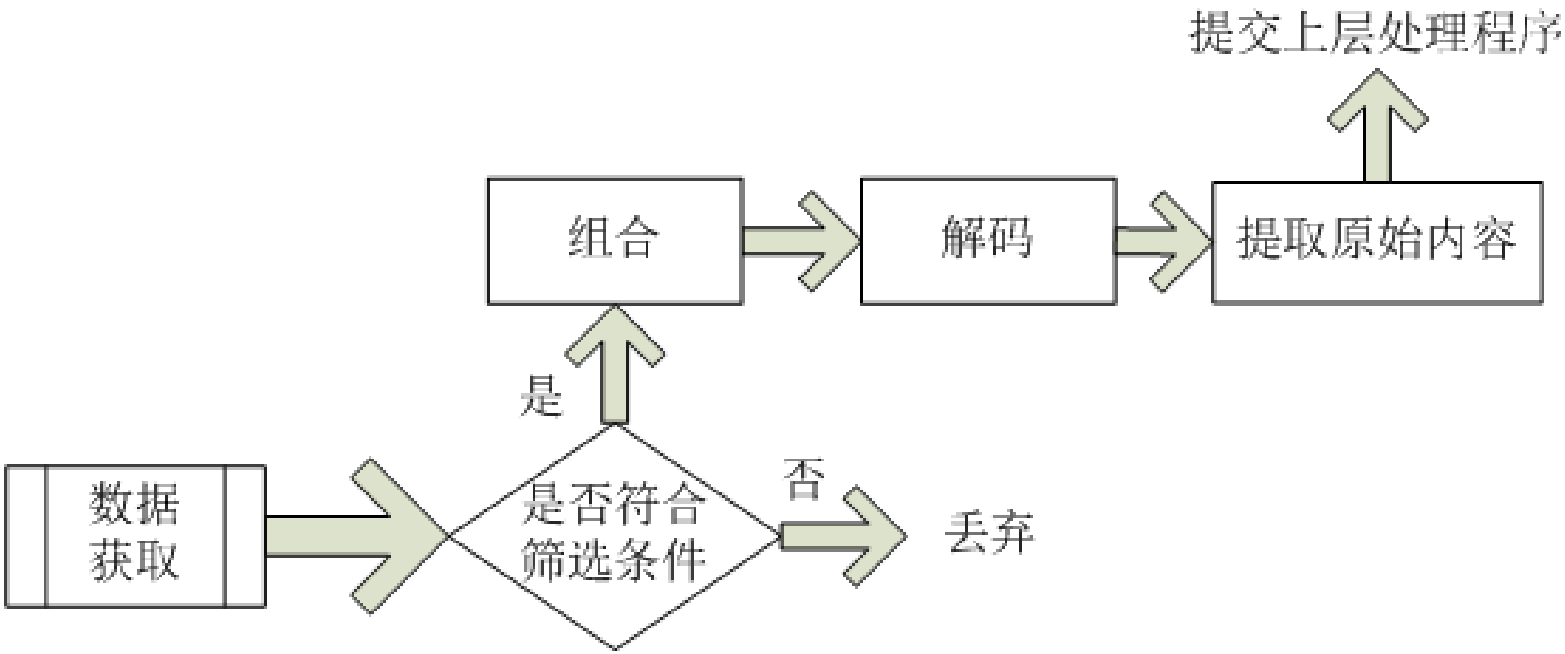
- 网络适配器必须工作在混杂模式下，这样才能保证所有接收到的数据报文被提交给协议栈。



网络报文处理流程

数据调整

- 数据调整主要指针对数据获取模块（主要是协议栈）提交的应用层数据进行筛选、组合、解码以及文本还原等工作；数据调整的输出结果用于敏感特征搜索等。



敏感特征搜索

- 敏感特征搜索实际上就是依据实现定义好的敏感特征策略，在待查内容中识别所包含的敏感特征值，搜索的结果可以作为违规判定的依据。

- 敏感特征值可以是文本字符串、图像特征、音频特征等，它们分别用于不同信息载体的内容的敏感特征识别。

- 基于文本内容的识别已经比较成熟并达到可实用化，而图像、音频特征的识别还存在着一些问题，难以实现全面有效的程序自动监管，更多时候需要人的介入。

- 基于文本内容的敏感特征又分为敏感字符串和敏感表达式两种形式。但无论哪种形式，均以串匹配为核心技术。

敏感特征搜索

- 串匹配又称为模式匹配，分为单模式匹配和多模式匹配。

- BF（Brute-Force）算法、KMP（Knuth Morris Pratt）算法、BM（Boyer-Moore）及BMH（Boyer Moor Horspool）算法均为经典的单模式匹配算法。

- 常见的多模式匹配算法有AC（Aho-Corasick）算法、ACBM（Aho-Corasick Boyer-Moore）算法、Manber-Wu算法等。
  - 这些多模式匹配算法的主要特点是通过一次扫描母串可以寻找到其包含的所有子串Patterns，其搜索速度与子串的数目无关，主要取决于对母串的扫描速度。

违规判定及处理

- 违规判定程序的设计思想：
  - 将敏感特征搜索结果与违规定义相比较，判断该网络信息内容是否违规。

- 违规定义是说明违规内容应具有的特征，即敏感特征。
  - 每个敏感特征由敏感特征值和特征值敏感度（某特征值对违规的影响程度，也可以看作权重）两个属性来描述。
  - 敏感特征的搜索结果具有敏感特征值的广度（包含相异敏感特征值的数量）和敏感特征值的深度（包含同一个特征值的数量）两个指标。
  - 违规判断算法针对上述内容进行计算，根据计算结果是否符合某个事先制定的标准来判断是否违规。

违规判定及处理

- 违规处理目前主要采用的方法与入侵检测相似。
  1. **报警**，就是通知有关人员违规事件的具体情况；
  2. **封锁IP**，一般是指利用防火墙等网络设备阻断对有关IP地址的访问；
  3. **拦截连接**，是指针对某个特定访问连接实施阻断。向通讯双方发送RST数据包阻断TCP连接就是常用的拦截方法。

主要内容

- 9.1 概述
  - 9.1.1 内容保护
  - 9.1.2 内容监管
- 9.2 版权保护
  - 9.2.1 DRM概述
  - 9.2.2 数字水印
- 9.3 内容监管
  - 9.3.1 网络信息内容监管
  - 9.3.2 垃圾邮件处理

垃圾邮件

- 垃圾邮件（Spam），目前还没有一个非常严格的定义。

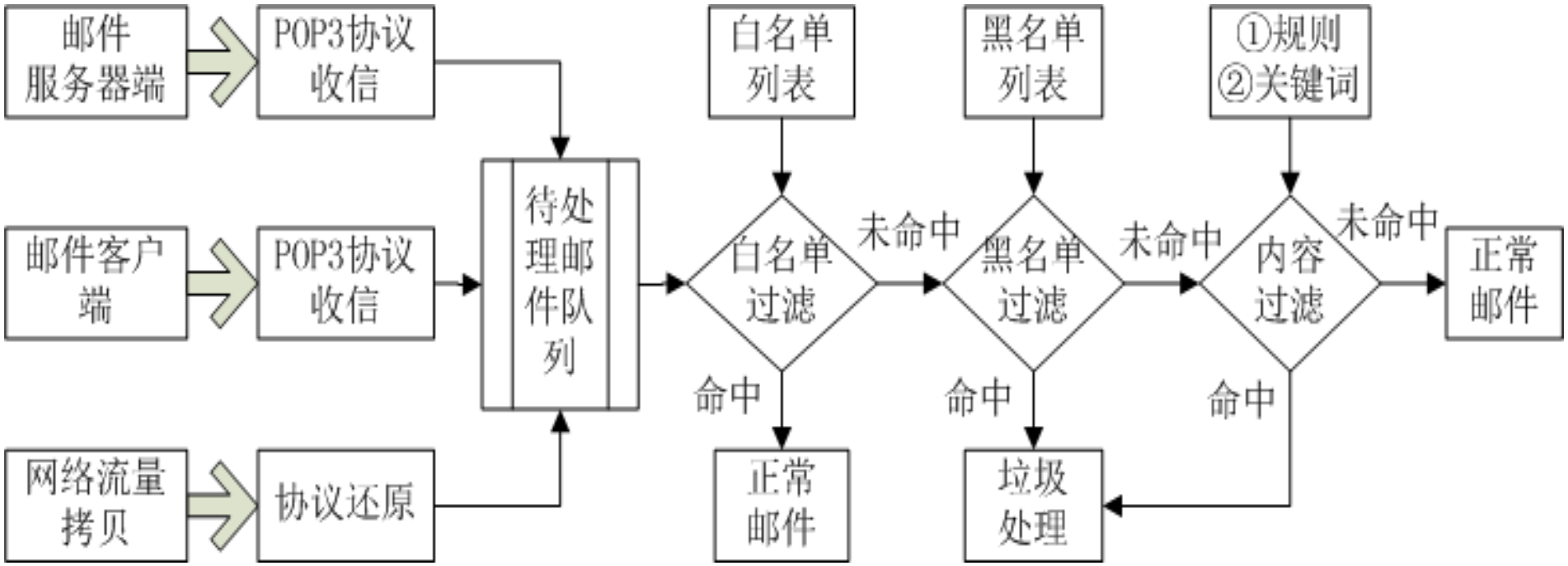
- 一般来说，凡是未经用户许可，就强行发送到用户邮箱中的任何电子邮件都属于垃圾邮件。

- 垃圾邮件，可以分为良性和恶性两种。
  - 良性垃圾邮件是指各种宣传广告等对收件人影响不大的信息邮件。
  - 恶性垃圾邮件是指具有破坏性的电子邮件。

垃圾邮件处理

- 目前主要采用的技术有过滤、验证查询和挑战。
  - 过滤（Filter）技术，是相对来说最简单、又最直接的垃圾邮件处理技术，主要用于邮件接收系统来辨别和处理垃圾邮件。
  - 验证查询技术，主要指通过密码验证及查询等方法来判断邮件是否为垃圾邮件。
    - 包括反向查询、雅虎的DKIM（Domain Keys Identified Mail）技术、Microsoft的SenderID技术、IBM的FairUCE（Fair use of Unsolicited Commercial Email）技术以及邮件指纹技术等。
  - 基于挑战的反垃圾技术，是指通过延缓邮件处理过程，来阻碍发送大量邮件。

基于过滤技术的反垃圾邮件系统



作业

1. 有人说“数字水印就是在视频或图像中可见的版权信息标识”，你认为正确与否，为什么？
2. 有人说“内容监管技术与入侵检测技术相同，可以通过简单改造IDS实现内容监管功能”，你认为正确与否，为什么？