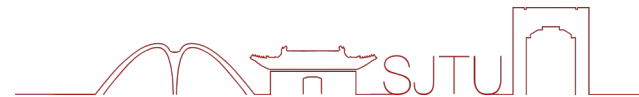




上海交通大学
SHANGHAI JIAO TONG UNIVERSITY



第五章 内容安全基础

主讲人：李建华 张全海
网络安全技术研究院

2024 年 12 月

—— 饮水思源 · 爱国荣校 ——



1

信息内容安全威胁

2

网络信息内容获取

3

网络信息内容分析与处理

4

舆情系统功能及内容分析

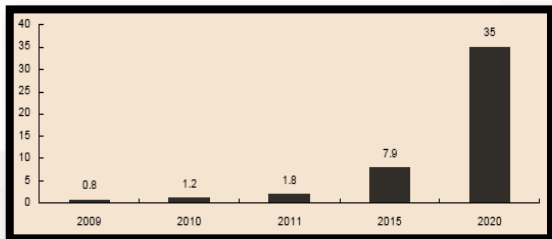
5

内容中心网络及安全

互联网中的数据和内容正成为互联网的中心关注点

- 视频新闻、播客、视频下载、图片照片、音频、论坛信息等相关内容产业快速发展，大量新数据源的出现则导致了**非结构化、半结构化数据**爆发式的增长；
- 互联网快速发展：5G 移动通信网络、移动互联网、物联网、云计算快速发展，成为人们获取信息、互相交流、协同工作的重要途径
- 当前数据呈现爆炸式增长，**数据驱动型的网络与社会**的形成和快速发展

互联网逐步由传统媒体向社交网络等新型媒体演进：微信、QQ、微博、Facebook、Twitter、YouYube。。。。。



数据
指数
增长



结构
日趋
复杂

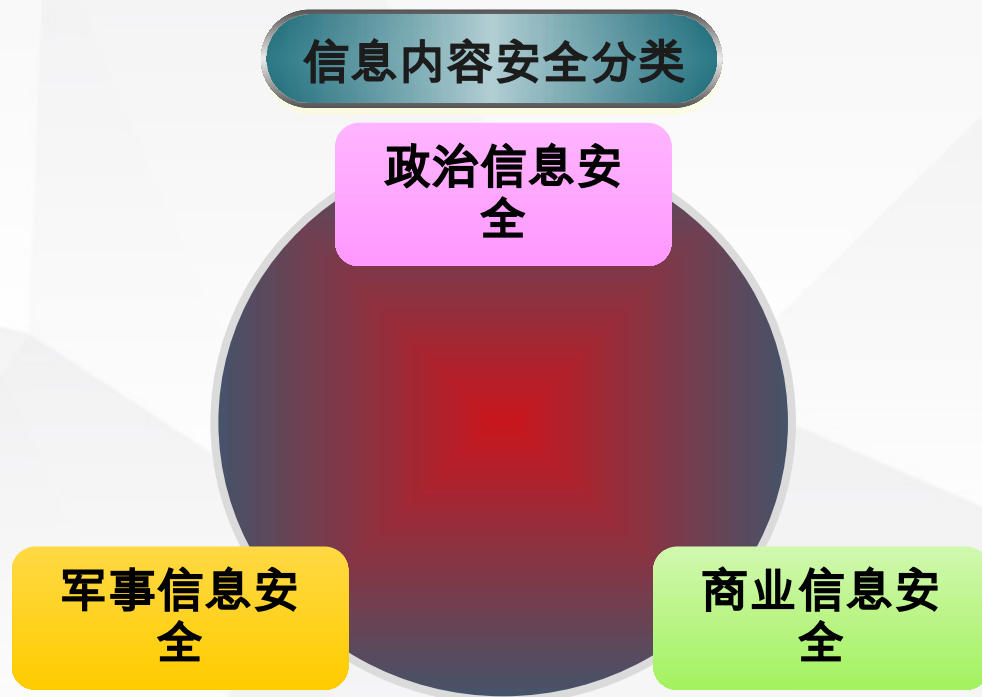




信息内容安全概述

信息内容安全 (Content-based Information Security) 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。

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。





信息内容安全概述

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	在传播过程中剔除侧请、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





网络信息内容安全的重要性



提高网络用户及
网站的使用效率



净化网络空间，
营造健康文明的
网络文化环境



提高国家信息
安全保障水平
是保障国家安
全的重要环节



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，和计算机安全一样主要建立在保密性、完整性、可用性之上。在互联网、电信网、电视网等各类网络信息共享环境中，信息内容面临两方面的威胁：

- 一方面，内容安全所面临威胁有泄露（指对信息的非授权访问）、欺骗、破坏和篡夺等；解决办法，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等技术。
- 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，解决办法，基于信息传播的互联网安全管理。

泄露

网络公开信息被恶意整合并滥用

欺骗

互联网的地址和网站内容被伪造

破坏

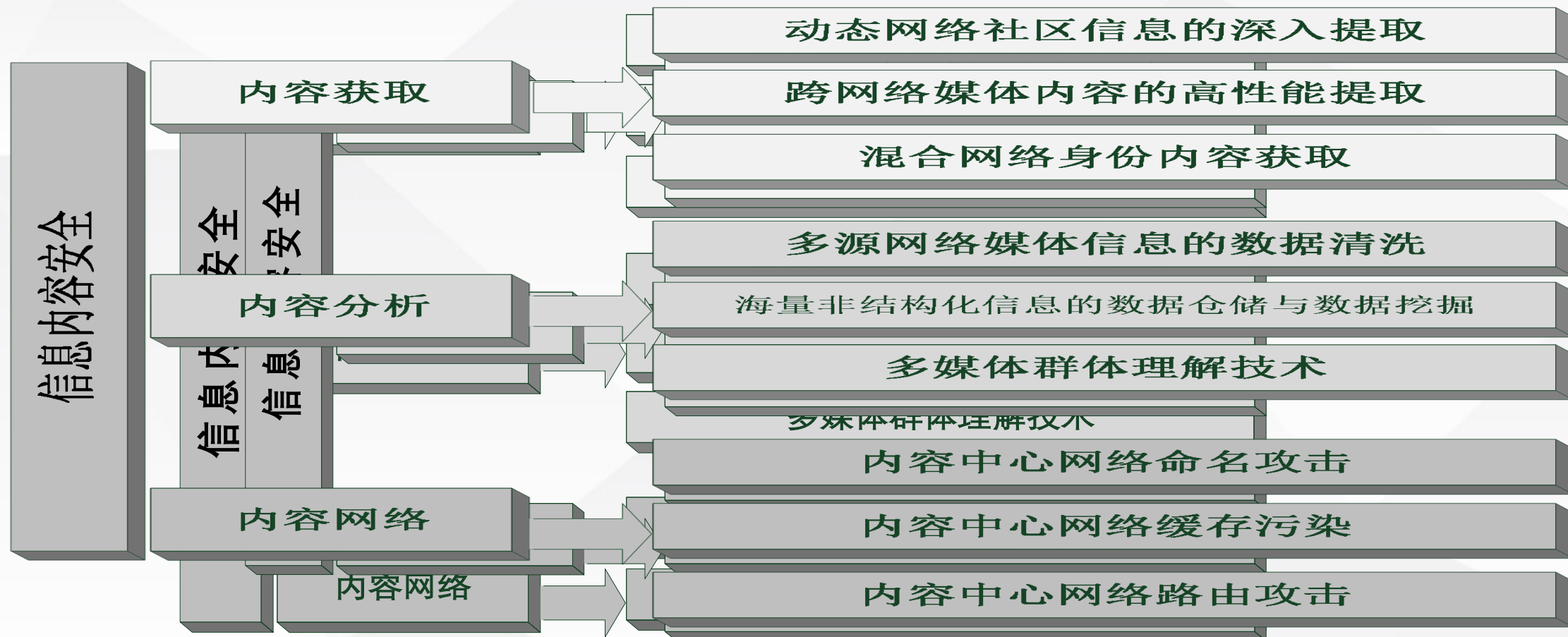
具有知识产权的音频视频非法传播

篡夺

信息在传播中被篡改伪造，被植入病毒






经典的信息内容安全挑战



从内容获取、内容分析和内容网络三个角度分析内容安全问题



典型的互联网恶意用户行为攻击

-  **Spam 用户的恶意行为**：向合法用户发布广告、色情、钓鱼等恶意信息，在开放式在下社交网络上将恶意内容快速而大规模传播；
-  **Sybil 攻击（女巫攻击）**：攻击者利用单个节点来伪造多个身份存在于 P2P 网络中，从而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。攻击者可以通过只部署一个实体，向网络中广播多个身份 ID，来充当多个不同的节点（Sybil 节点）。Sybil 节点为攻击者争取了更多的网络控制权，一旦用户查询资源的路径经过这些 Sybil 节点，攻击者可以干扰查询、返回错误结果，甚至拒绝回复。
-  **水军用户的恶意攻击**：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。网络推手、网络打手、刷粉。。。。。



以内容为中心的未来互联网

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 网络意义

- 实施更多优化表示来增强网络性能
- 提高未来互联网的智能水平

内容中心 攻击分类

- 命名：攻击者可以审查和过滤内容
- 路由：恶意攻击者可以发布 / 订阅无效内容或路由
- 缓存：污染或破坏缓存系统、侵犯中心网络隐私
- 其他：传输过程中未经授权地访问 / 更改内容



1

信息内容安全威胁

2

网络信息内容获取

3

网络信息内容分析与处理

4

舆情系统功能及内容分析

5

内容中心网络及安全



传统
网站
媒体

论坛 (BBS) , 博客
(Blog)



新型
网络
媒体

多媒体 (视 / 音频) 点播、
网上交友、直播等

- 主要包含新闻网站，论坛 (BBS)、博客 (Blog) 等形态；新兴的交互式媒体涵盖搜索引擎、多媒体 (视 / 音频) 点播、网上交友、网上招聘与电子商务 (网络购物) 等形态
- 可细分为文本信息、图像信息、音频信息与视频信息 4 种类型，其中，网络文本信息始终是网络媒体信息中占比最大的信息类型。

与面向特定点的网络通讯信息获取不同，**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。

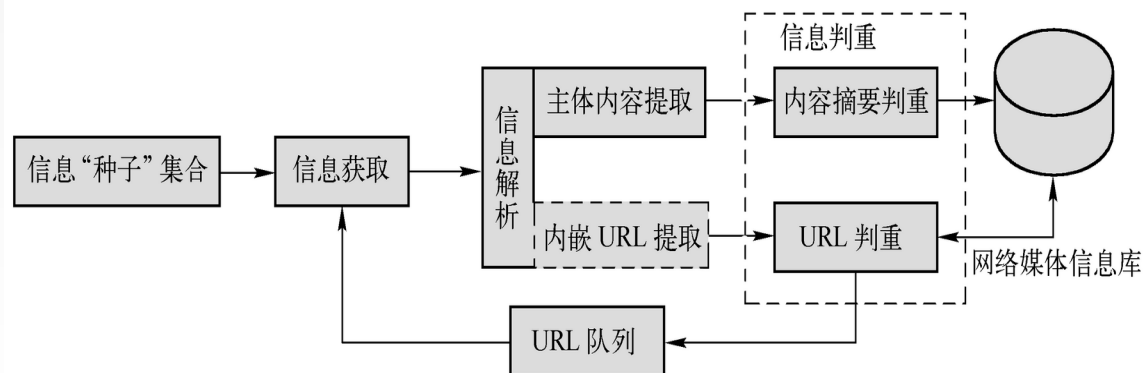
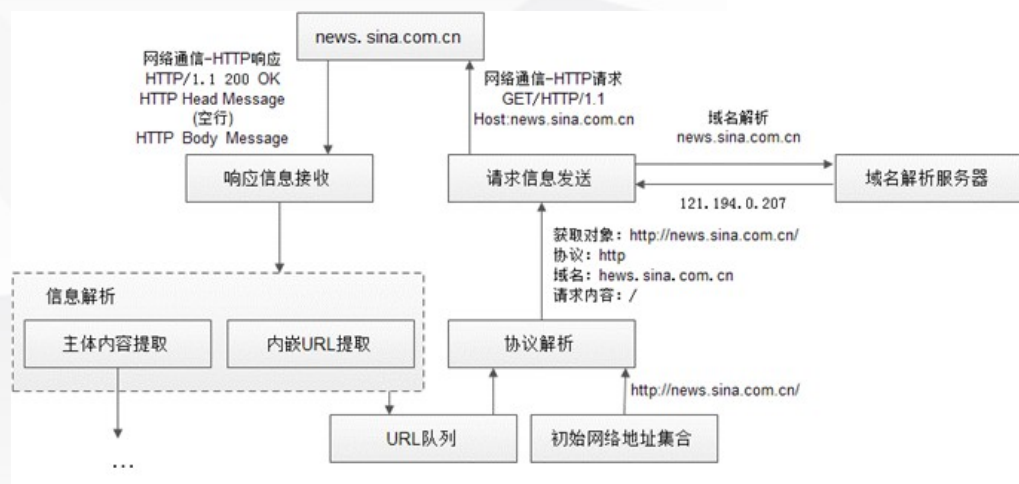
网络媒体信息获取流程

1) 初始 URL 集合

2) 信息获取

3) 信息解析

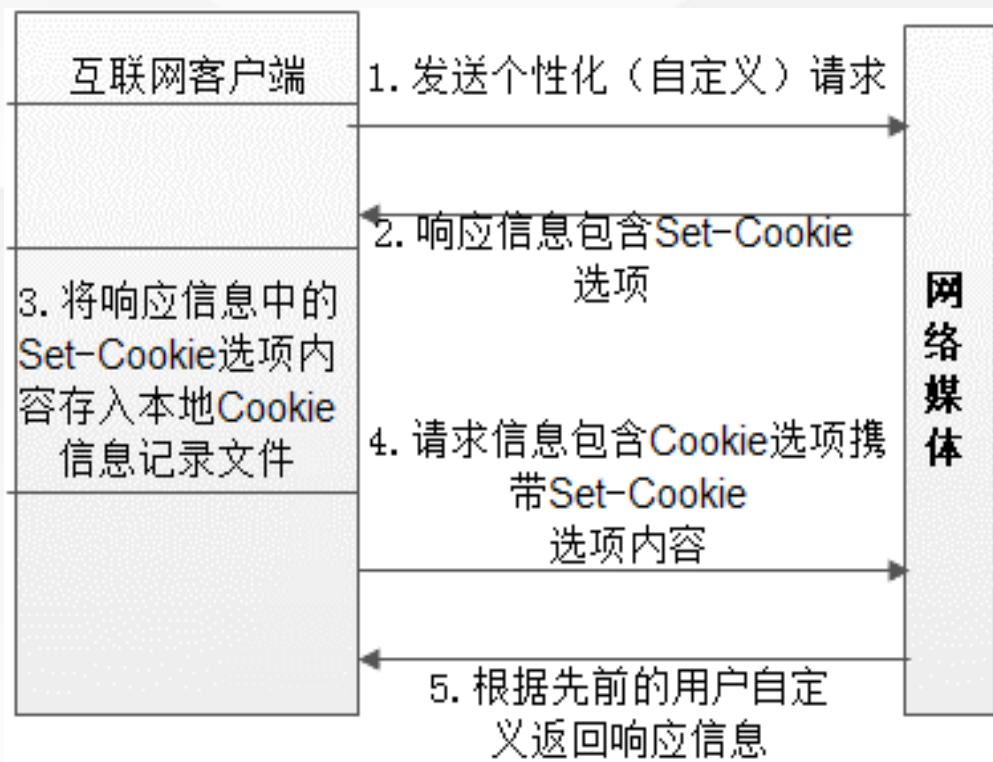
4) 信息判重



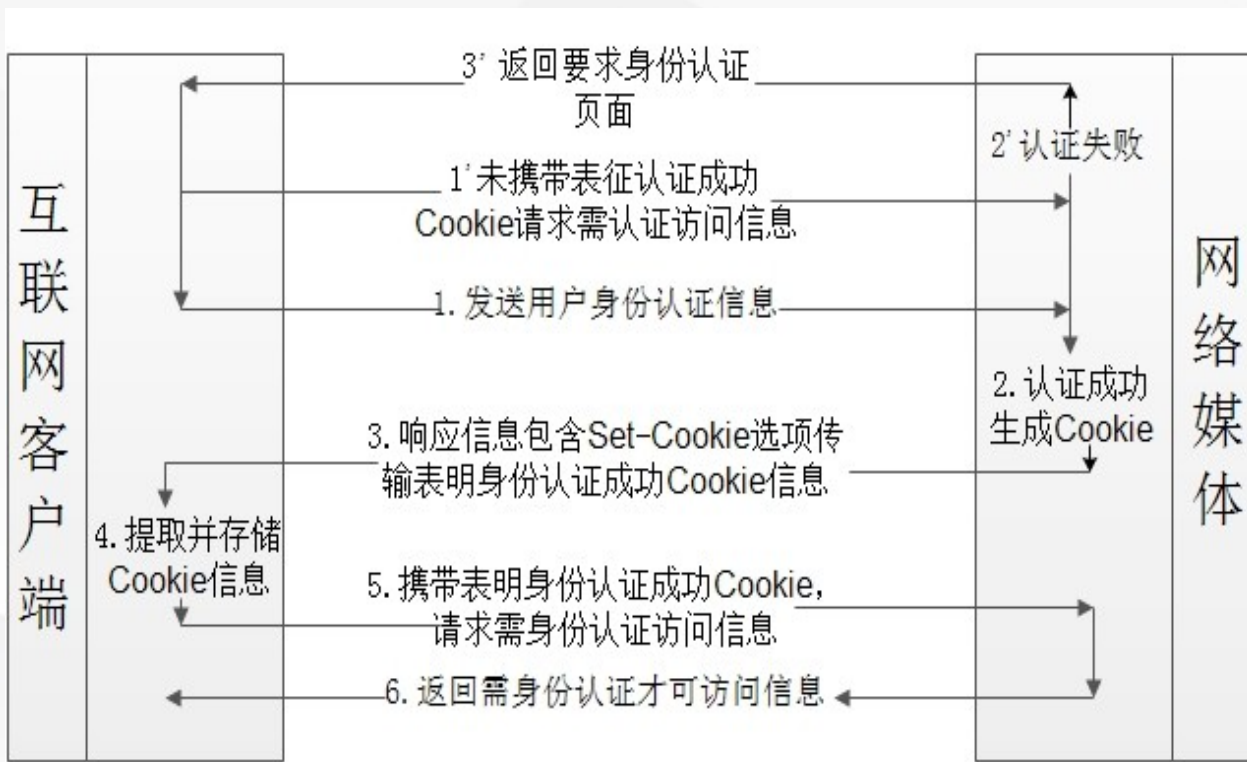
媒体信息获取流程



基于 Cookie 机制实现身份认证



基于 Cookie 机制的 HTTP 信息交互过程

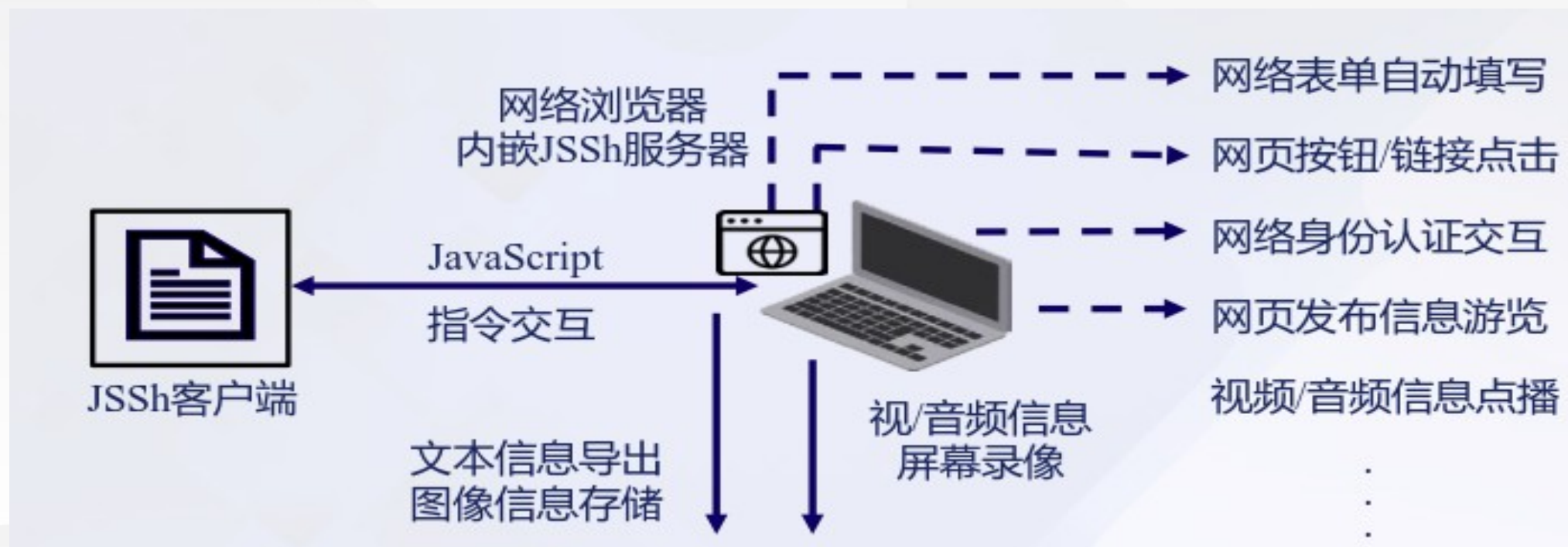


基于 Cookie 机制实现需身份认证才可访问信息请求



基于浏览器模拟实现信息获取的技术

- 身份认证表单自动填写，利用 **JSSh 客户端** 向内嵌 **JSSh 服务器** 的网络浏览器发送 JavaScript 指令，指示网络浏览器进行网络身份认证交互
- 网页内嵌 URL 逐一进行点击发布信息浏览和内容导出等操作。



网络媒体信息获取





信息内容获取的典型工具

网络爬虫

在互联网上的实施信息内容获取的主要工具，是按照一定的规则，自动的抓取互联网信息的**程序或脚本 (Heritrix , Nutch 和 Labin)**。

选择性

两类网络爬虫：
一是服务于搜索引擎等搜索类应用的网络爬虫；二是针对性进行信息收集的网络爬虫。

分布式机制

网络爬虫需要采用多进程或者多线程，甚至分布式机制来保证信息获取的全面性和时效性。



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，把从信息中抽取的**特征词**进行**量化**表示文本信息。

文本信息内容的特征抽取与选择

- 文本转化为可处理的结构化形式
- 文本特征选择以达到降维的目的，主要有基于统计的方法

音频信息内容的特征抽取与选择

- 提取音频的时域和频域特征。
- 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

图像信息的特征抽取与选择

- 图像颜色特征提取
- 图像纹理特征提取
- 其他图像特征提取

文本信息内容的特征抽取与选择

中文是以字为基本书写单位，单个字往往不足以表达一个意思，通常认为词是表达语义的最小元素。因此须对中文字符串进行合理的切分。

字符串	HERE IS A SIMPLE EXAMPLE
搜索词	EXAMPLE

字符串匹配

- 优点是：分词过程是跟词典作比较，不需要大量的语料库、规则库，其算法简单、复杂性小、对算法作一定的预处理后分词速度较快；
- 缺点是：不能消除歧义、识别未登录词，对词典的依赖性比较大，若词典足够大，其效果会更加明显。



基于统计方法

- 优点是：由于是基于统计规律的，对未登录词的识别表现出了一定的优越性，不需要预设词典；
- 缺点是：需要一个足够大的语料库来统计训练，其正确性很大程度上依赖训练语料库的质量好坏，算法较为复杂，计算量大，周期长，但是都较为常见，处理速度一般。

	f-o	f-g	o-g	f-b	o-b
$\psi(\text{fog})$	λ^2	λ^3	λ^2	0	0
$\psi(\text{fob})$	λ^2	0	0	λ^3	λ^2

$$k(\text{fog}, \text{fog}) = 2\lambda^4 + \lambda^6$$

$$k(\text{fob}, \text{fob}) = 2\lambda^4 + \lambda^6$$

$$k(\text{fog}, \text{fob}) = \frac{k(\text{fog}, \text{fob})}{\sqrt{k(\text{fog}, \text{fog})k(\text{fob}, \text{fob})}} = \frac{\lambda^2}{2\lambda^4 + \lambda^6} = \frac{1}{2\lambda^2 + \lambda^4}$$

www.voidcn.com

基于理解方法

- 优点是：由于能理解字符串含义，对未登录词具有很强的识别能力，能很好的解决歧义问题，不需要词典及大量语料库训练；
- 缺点是：需要一个准确、完备的规则库，依赖性较强，效果好坏往往取决于规则库的完整性。算法比较复杂、实现技术难度较大，处理速度比较慢。

词频是一个词在文档中出现的次数。通过词频进行特征选择就是将词频小于某一值或大于某一值的词删除，从而降低特征空间的维数。



文本信息内容的特征抽取与选择

- ◆ 特征抽取的主要功能是在不损伤文本核心信息的情况下**尽量减少要处理的单词数**，以此来降低向量空间维数，从而简化计算，提高文本处理的速度和效率。



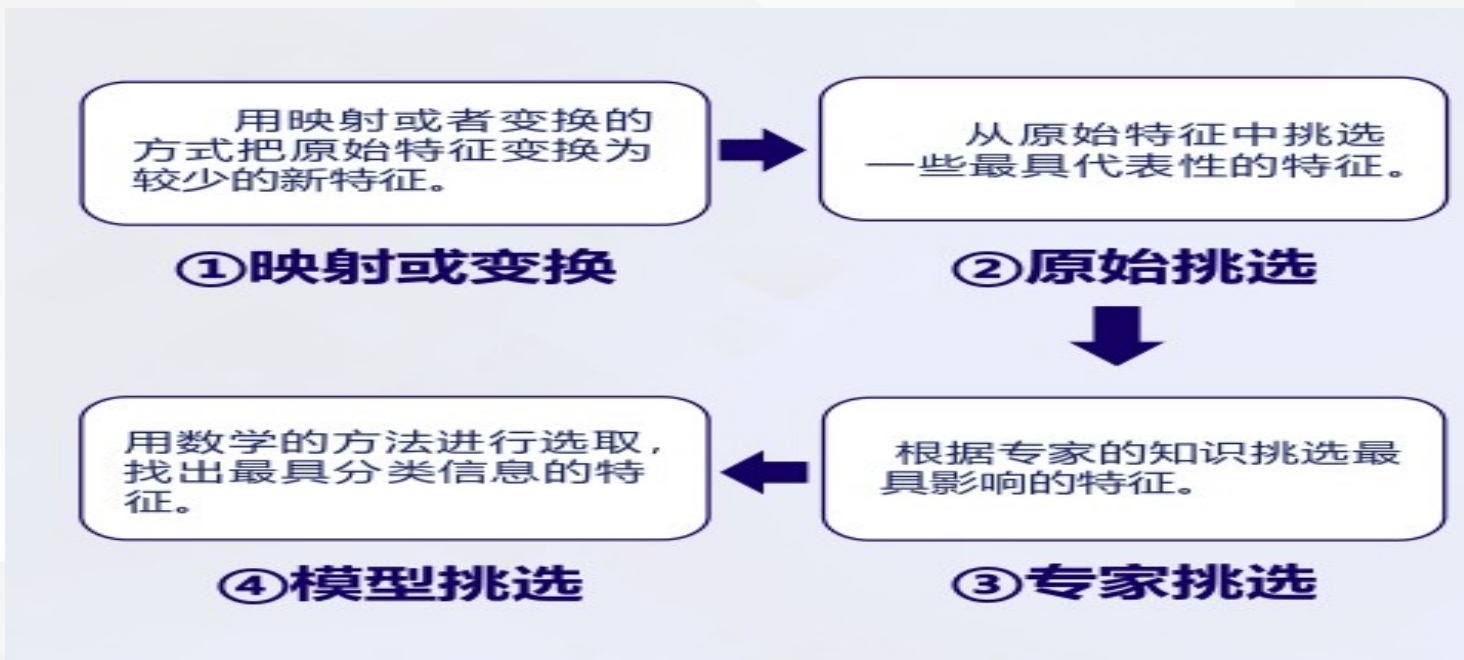
特征选取方式

- 采用向量空间模型来描述文本向量
- 通过特征选择来降维，找到代表性特征



特征选择过程

- 根据特征评估函数计算各个特征的评分值
- 按照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





1

信息内容安全威胁

2

网络信息内容获取

3

网络信息内容分析与处理

4

舆情系统功能及内容分析

5

内容中心网络及安全

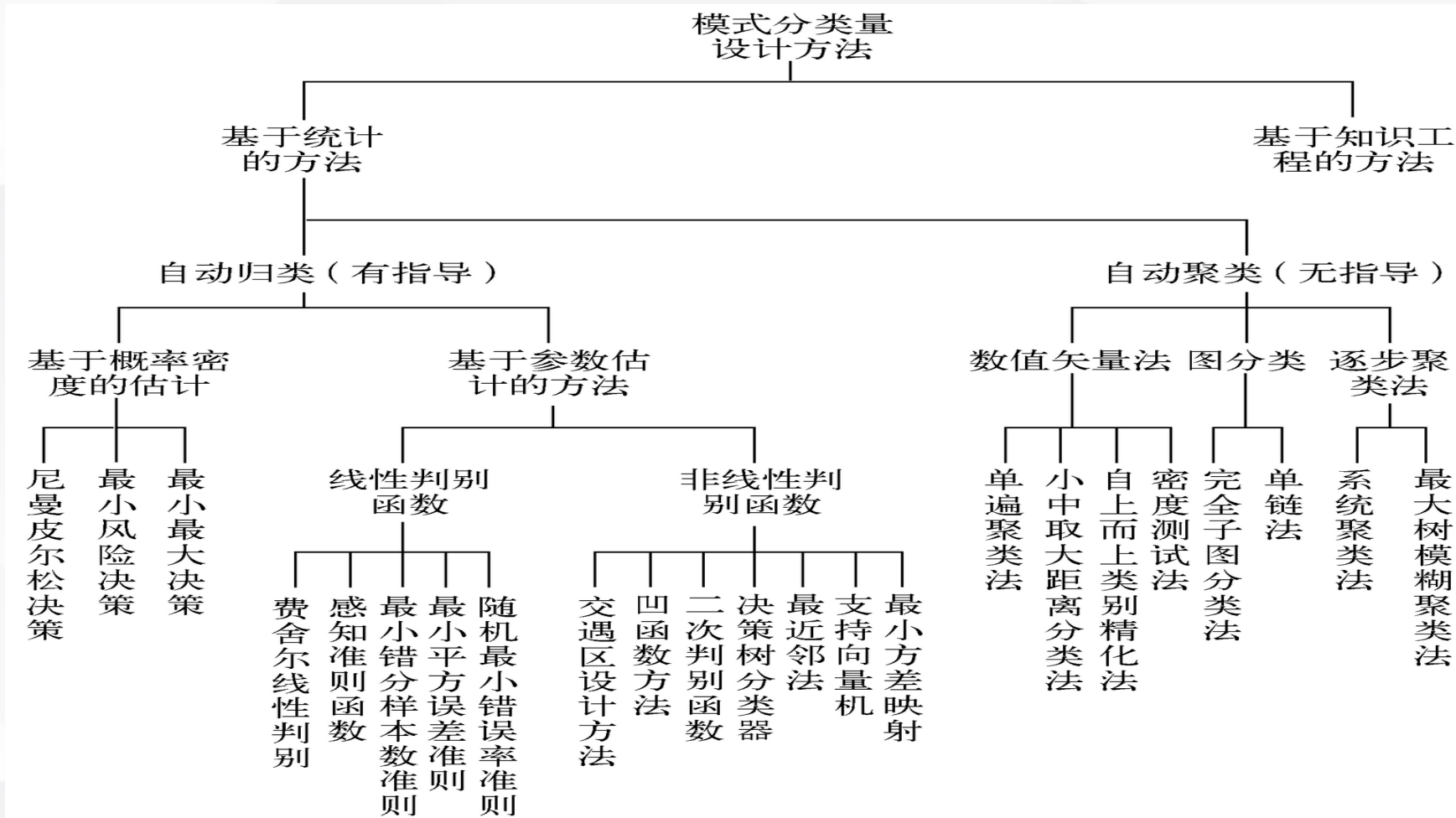


海量信息内容分析的基本处理环节可以归结为**分类和过滤**。其他更加复杂的处理问题则是上述简单处理问题的**组合**。

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。**高效的分类和过滤算法**能使信息处理变得迅速而准确，反之，则会使处理过程变得冗长而模糊。

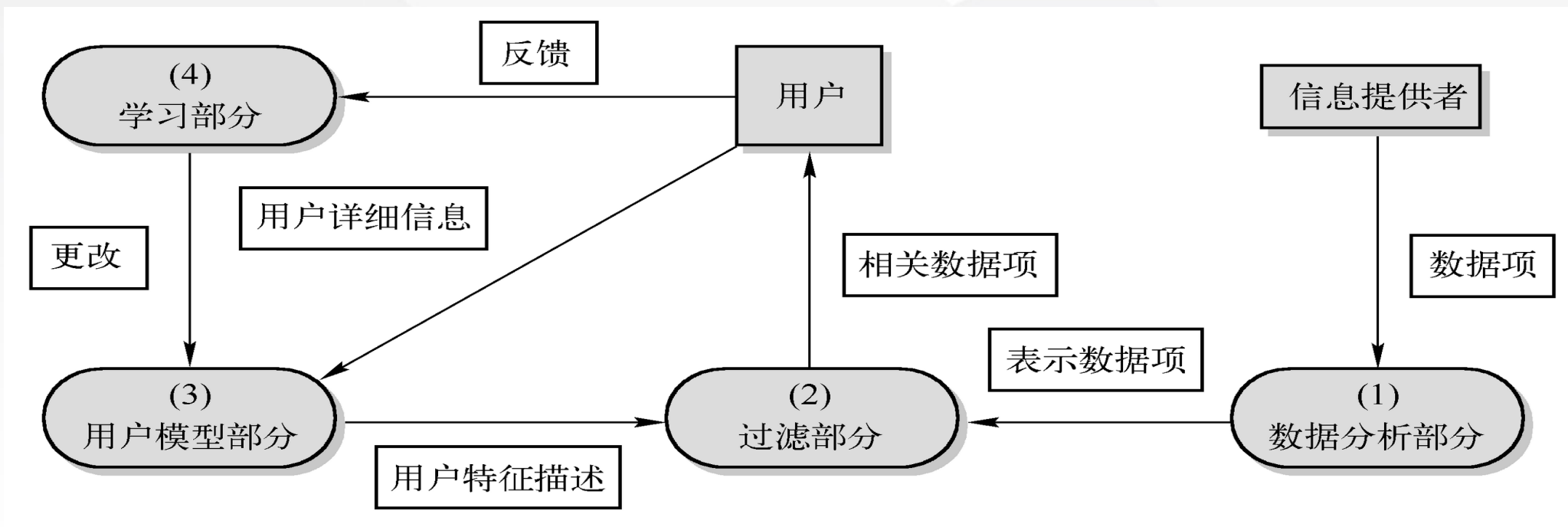


主要的分类方法和它们之间的基本关系。



信息内容过滤

信息过滤是大规模内容处理的一种典型操作，对陆续到达的信息进行过滤，是满足用户信息需求的信息选择过程。



通用信息过滤模型



信息内容过滤技术的分类

信息过滤是提供信息的有效流动，消除或者减少信息过量、信息混乱、信息滥用造成的危害。**为用户剔除不合适的信息**是当前信息过滤的主要任务之一。

基于内容的过滤
基于用户兴趣的过滤
协作过滤

根据过滤方法分类

信息的源头过滤
服务器和客户端过滤

根据过滤位置分类

根据操作的主动性分类

主动过滤
被动过滤

根据过滤的目的分类

用户过滤
安全过滤





信息内容过滤的主要方法



统计
方法

④ 向量中心法



简单实用，应用广泛

④ 相关反馈法



实现较易，受训练集合影响较大

④ K 近邻法



原理简单，需要确定 k 值

④ 贝叶斯法



机器学习中应用广泛

④ 多元回归模型



运用线性最小平方匹配算法

④ 支持向量机

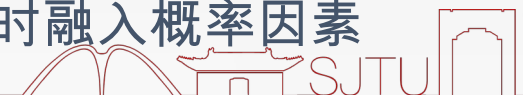


文本分类领域比较成功，训练过程效率不高

④ 概率模型



特征加权时融入概率因素





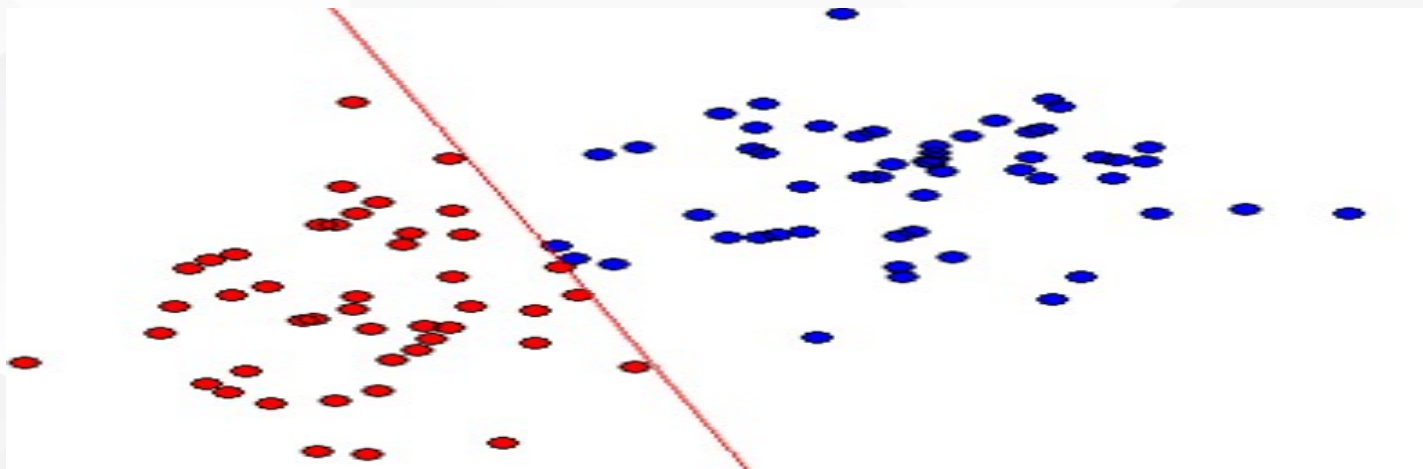
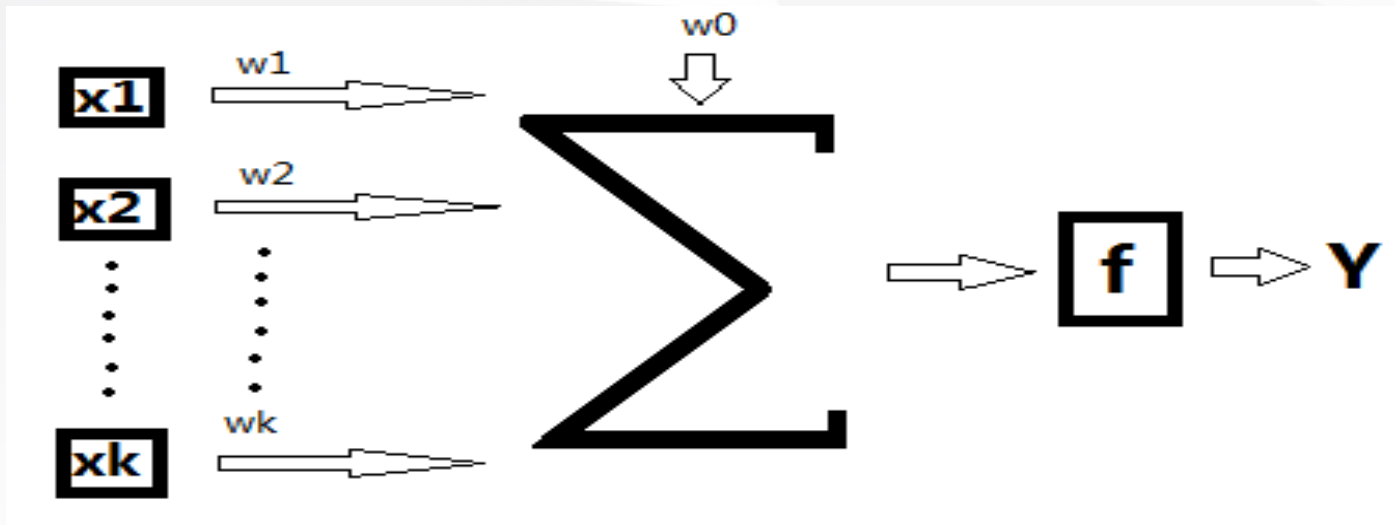
线性分类器

X 输入， X_i 表示的是第 i 个输入； Y 表示输出； W 表示权向量； w_0 是阈值， f 是一个判别函数。

线性分类器的基本思想：寻找一个合理的决策超平面（**确定投影方向和阈值 w_0** ）。故设这个超平面为 w ，

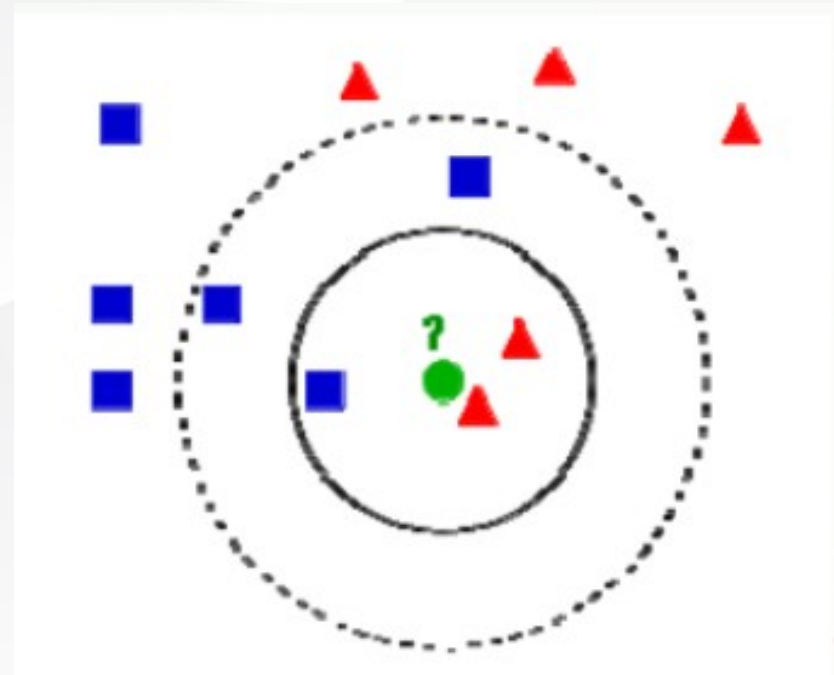
满足 $w^T * x > 0, \forall x \in \omega_1$
 $w^T * x < 0, \forall x \in \omega_2$

，即通过给定的训练数据**确定线性判别函数**。



K 最近邻分类算法的核心步骤如下：

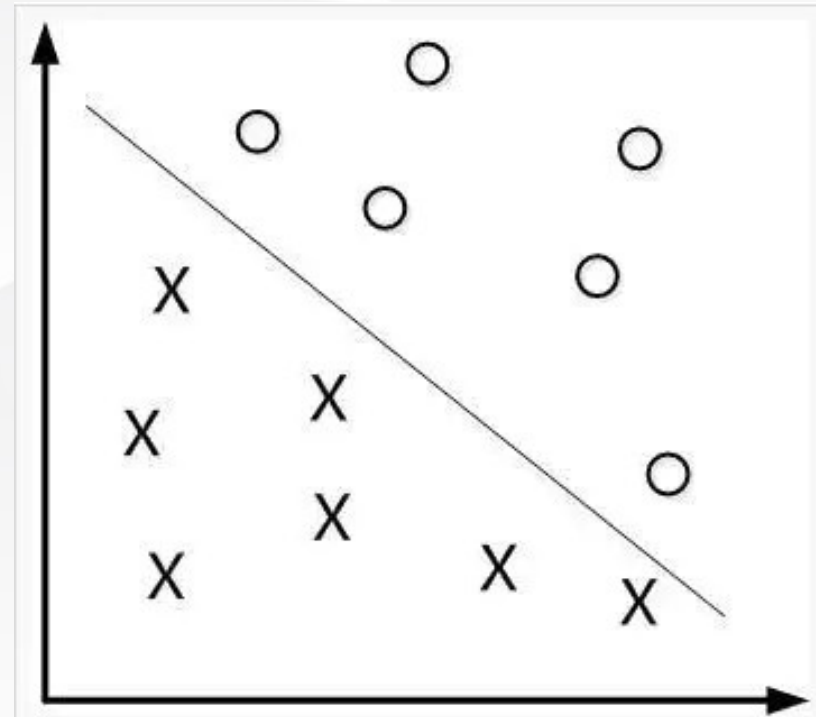
- 数据清洗：数据规范化；
- 确定临近度的度量，并计算临近度；
- 按照临近度递增次序排序；
- 选取与当前点距离最小的 k 个点；
- 确定前 k 个点所在类别的出现频率；
- 返回前 k 个点出现频率最高的类别作为当前点的预测分类。



K 最近邻 (k-Nearest Neighbor-KNN) 分类算法属于监督学习算法，多用于图像分类和识别领域，其核心想法非常简单明了，**确定一个临近度的度量，相似性越高，相异性越低的数据样本，可以认为是同一个数据类别**，即如果一个样本在特征空间中的 K 个最相近（特征空间中最邻近）的样本中的大多数属于某一个类别，则样本也属于这个类别



SVM 支持向量机 (英文全称 : **support vector machine**) 是一个分类算法 , 通过找到一个分类平面 , 将数据分隔在平面两侧 , 从而达到分类的目的。如图所示 , 直线表示的是训练出的一个分类平面 , 将数据有效的分隔开。



SVM 的分类基本思路是要找到最合适的分类平面 , 最直接的评估标准 : 被分隔的两边数据距离平面间隔最大 , 换句话说 , **SVM 就是获取最大间隔的超平面**。 $w * x + b = 0$ 确定的情况下 , $|wx+b|$ 表示点距离超平面的距离 , 而超平面作为二分类器 , 如果 $wx+b>0$, 判断类别 y 为 1, 否则判定为 -1 。



信息内容过滤常见应用





1

信息内容安全威胁

2

网络信息内容获取

3

网络信息内容分析与处理

4

舆情系统功能及内容分析

5

内容中心网络及安全



网络舆情系统概述

网络舆情：舆情指在一定的社会空间内，围绕中介性社会事项的发生、发展和变化，作为主体的民众对作为客体的国家管理者产生和持有的社会政治态度。如果把中间的一些定语省略掉，**舆情就是民众的社会政治态度。**

人大选举...
总统换届...
...
俄乌战争...

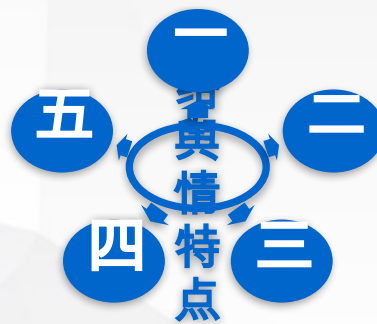
两会召开...
...
疫情...



网络舆情系统：对海量非结构化信息挖掘与分析；实现对网络舆情的热点、焦点、演变等信息的掌握；为网络舆情监测与引导部门的决策提供科学依据。



● 偏差性：网络舆情不等于全民立场



● 直接性：通过网络直接发表意见，传播迅速

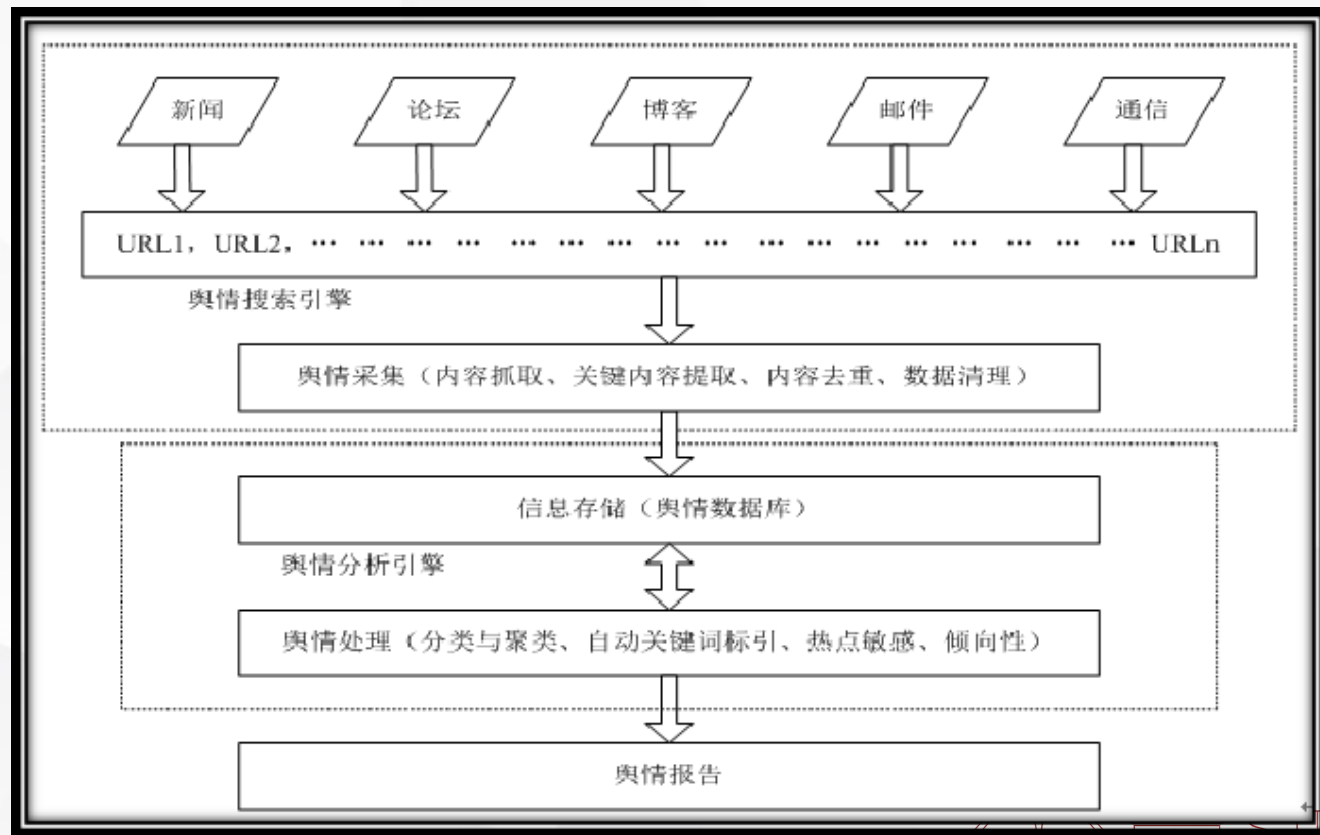
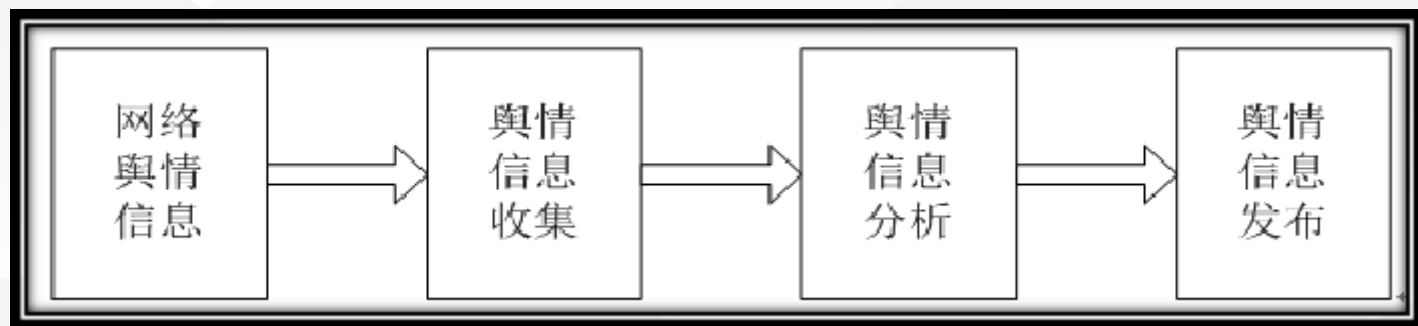
● 随意性和多元化：网民可匿名发表观点，健康观点和灰色言论并存

● 隐蔽性：虚拟网络空间中网民可以隐身发言

● 突发性：网络快速传播的特性使关注焦点迅速成长为舆论热点



网络舆情分析系统框架





网络舆情监测技术的发展趋势



针对信息源的 深入信息采集

问题描述：

传统搜索引擎一般采用**广度优先**的策略遍历Web并下载文档。

不能完全满足实际需求，其主要不足体现在——**互联网定点信息源信息的提取率过低**。

信息采集的深入性和全面性是重点解决问题。

异构信息融合 分析

问题描述：

互联网信息在**编码、数据格式**以及**结构组成**方面存在巨大差异。

信息分析与提取的**重要前提**是对信息在**同一表达或标准**的前提下进行有机的结合。

非结构信息的 结构化表达

问题描述：

非结构化信息对于阅读者而言比较容易理解，但对于计算机信息处理系统却**相当困难**。

目前已有优秀的技术与方法可以对结构化数据进行分析。



网络舆情分析关键技术



信息采集技术



热点发现



热点评估



主题跟踪

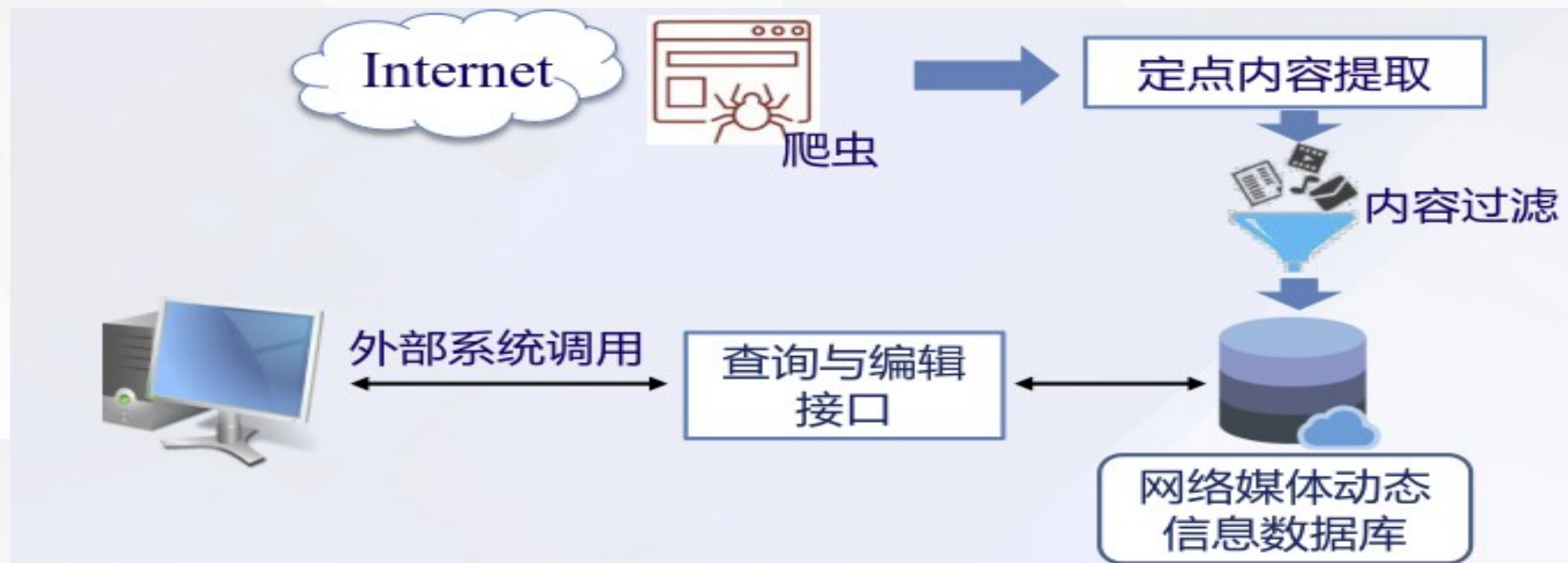


分析处理



高仿真网络信息（论坛、聊天室）深度提取技术是网络舆情监测预警系统建设的基础核心内容。重点研究**原创网络互动式动态信息提取**，形成高性能动态信息提取系统，组成舆情监控系统的信息获取模块

高仿真网络信息深度提取



基于语义的海量媒体内容特征快速提取与分类技术

- 实现信息特征提取和结构化转变功能，组成舆情监控系统的信息分析模块
- 为实现舆情的分析、监测与预警完成信息转化

基于语义的海量文本特征快速提取与分类

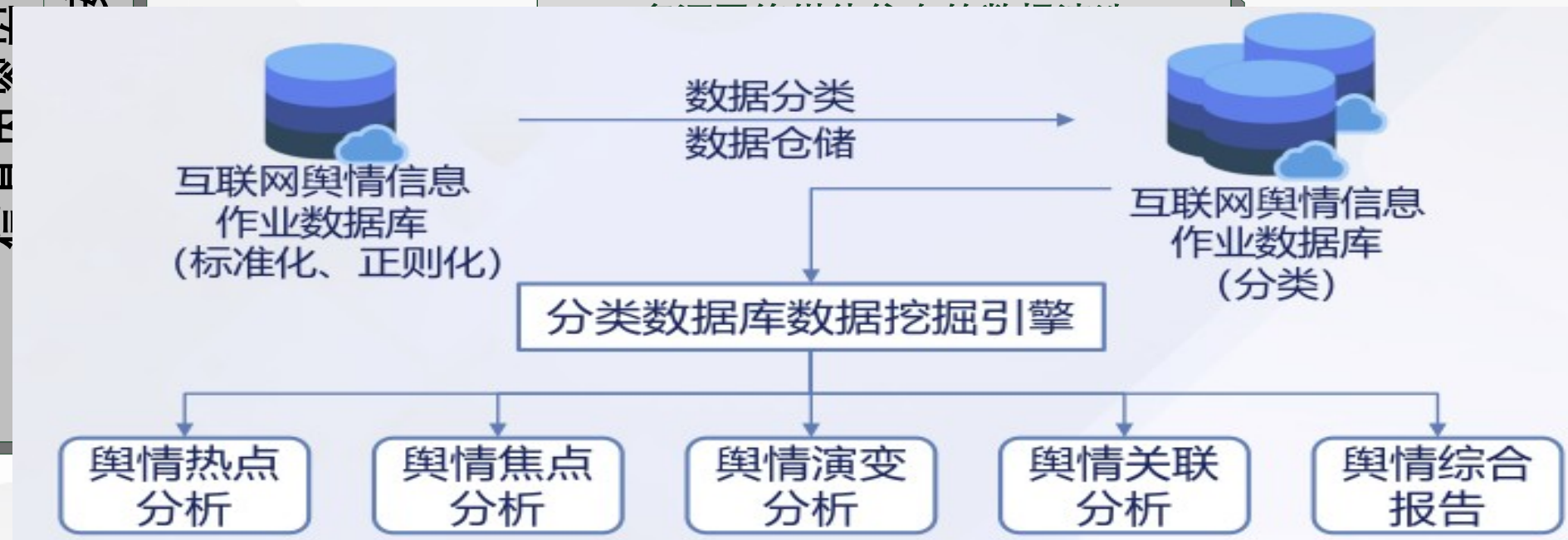
内容分析

海量非结构化信息的数据仓储与数据挖掘



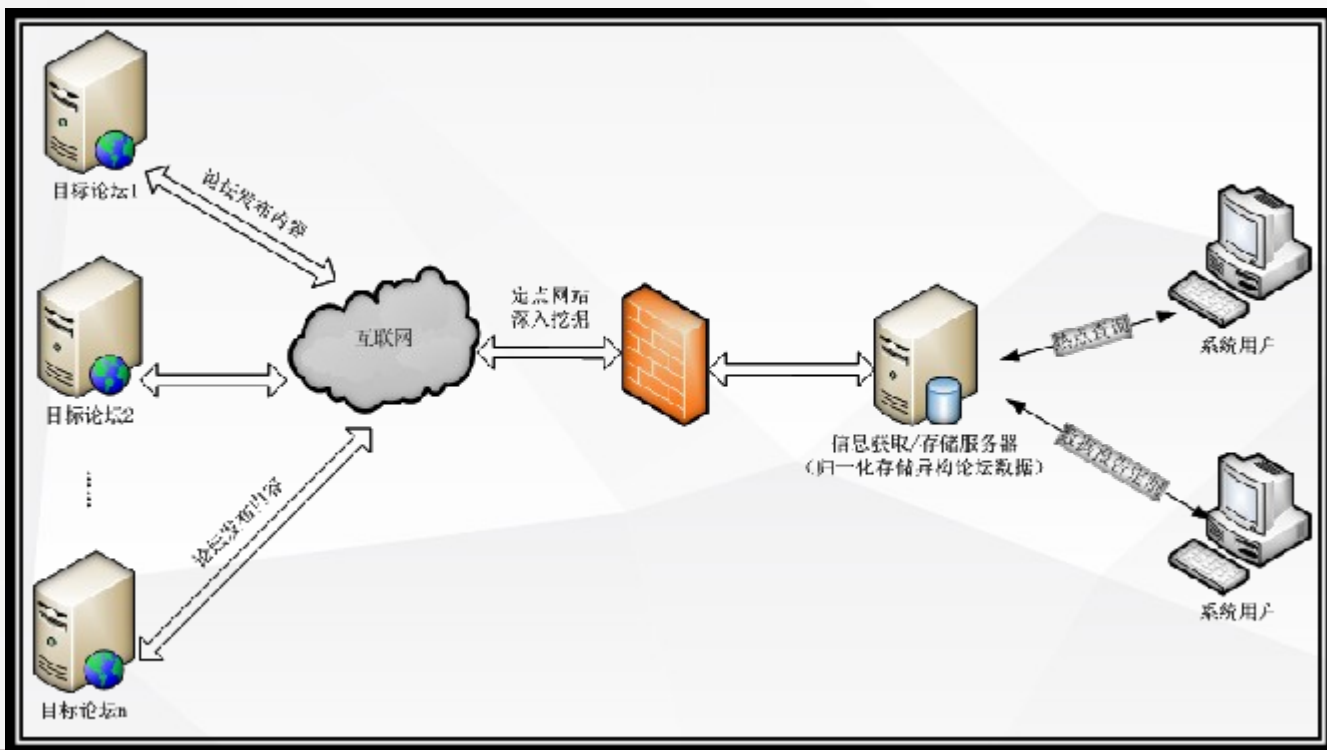
非结构信息自组织聚合表达技术能满足网络舆情监测预警基础设施与典型应用的实际需求，组成 数据分类、数据仓储和数据类型挖掘引擎模块。

非结构信息自组织聚合表达



典型应用 ---- 互联网舆情内容分析

深度挖掘技术：利用定向搜索手段完成针对指定信息源的深入、全面的内容提取操作，以面向结构迥异、风格多样的数据发布源实施互联网媒体信息监控工作。



关键技术



异构信息归一化



网络热点自动发现



网络协商与人际对话模拟



热点数据报告定制



1

信息内容安全威胁

2

网络信息内容获取

3

网络信息内容分析与处理

4

舆情系统功能及内容分析

5

内容中心网络及安全



- 内容中心网络 (Content Centric Network,CCN) : 是 2009 年美国帕罗奥多研究中心公司 (PARC) 的 Van Jacobson 教授等人提出来的新型 下一代网络体系结构，是一个基于内容的网络。
- CCN 中的核 心思想是它对**网络中的每个内容命名，而不是使用 主机和节点的 IP 地址**。当需要获取一个内容 / 服务时，网络节点将发送一个包含所需内容 / 服务名字的请求。该请求按照内容名字进行路由，而不是 IP 地址。
- CCN 目标是替代现有的以 IP 为核心的网络体系架构，“以数据为中心”将通信范式的重点从关注于“**where**”（地址、服务器、主机）转变到“**what**”（通信的内容）。以对数据命名的方式代替位置（IP 地址），将数据转变成网络的第一要素。



内容中心网络架构

内容中心网络设计的基本原理是 摒弃以 IP 地址为中心的传输架构，采用以**内容名称为中心的**传输架构。

内容中心网络通过以**内容为中心的订阅机制和语义主导的命名、路由和缓存策略**，在解决当前基于 IP 地址进行联网的模式上体现出了巨大的潜力

1 内容信息对象

2 命名

3 路由

4 缓存

5 应用程序编程接口



内容中心网络架构

1

内容信息对象

存储在计算机中并通过计算机访问的所有类型的对象都可以看作内容信息对象。



网页



文档



电影



照片



音乐

2

命名

内容的命名是信息对象的标识，具有全局性和唯一性。其地位与TCP/IP 架构的 IP 地址类似。



分层命名方案



扁平命名方案





3

路由



4

缓存

每个 CCN 节点维护**缓存表**，用于缓存 CCN 路由器接收的内容消息对象，以便**响应后继接收到的相同请求**。



5

应用程序编程接

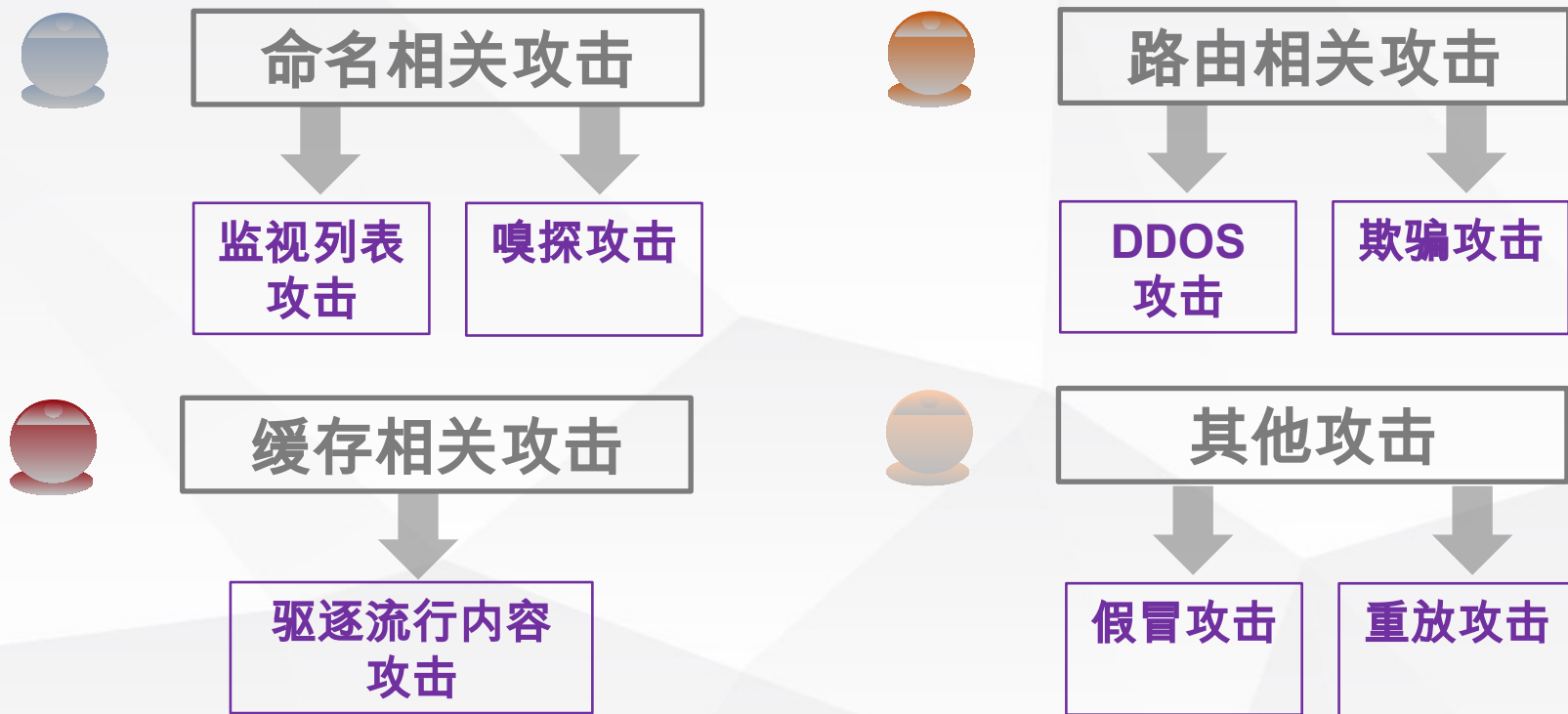
CCN 应用程序编程接口根据**请求**和**交付**内容信息对象定义，用于内容信息对象的**发布**和**获取**操作。





面向内容中心网络的攻击分类

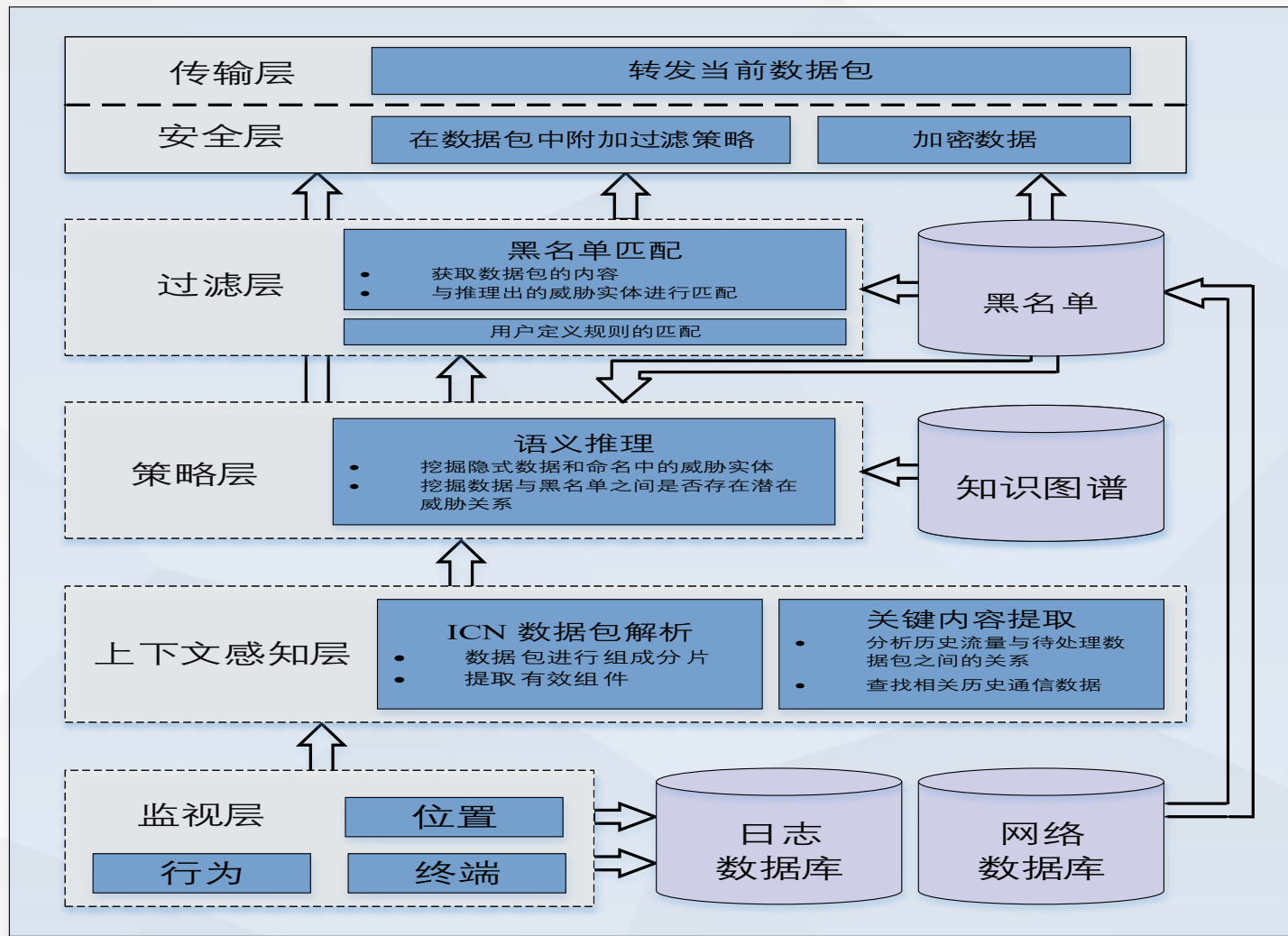
内容中心网络具有许多**独特的属性**。在内容中心网络体系结构中，除了可能对**影响网络流量**的**旧式攻击**之外，还出现了**新的攻击**。





* 基于雾计算的内容信息中心网络安全防护架构

基于雾计算的智能防火墙模型：利用**雾计算范式**在**网络边缘**实现了**隔离防御**系统；基于已有的安全策略，该防火墙能够实现对**兴趣包洪泛攻击**的智能感知和动态防御。





上海交通大学

SHANGHAI JIAO TONG UNIVERSITY

谢谢！

饮水思源 爱国荣校