信息内容安全

王巍

w_wei@hrbeu.edu.cn

泛在网络与信息安全团队

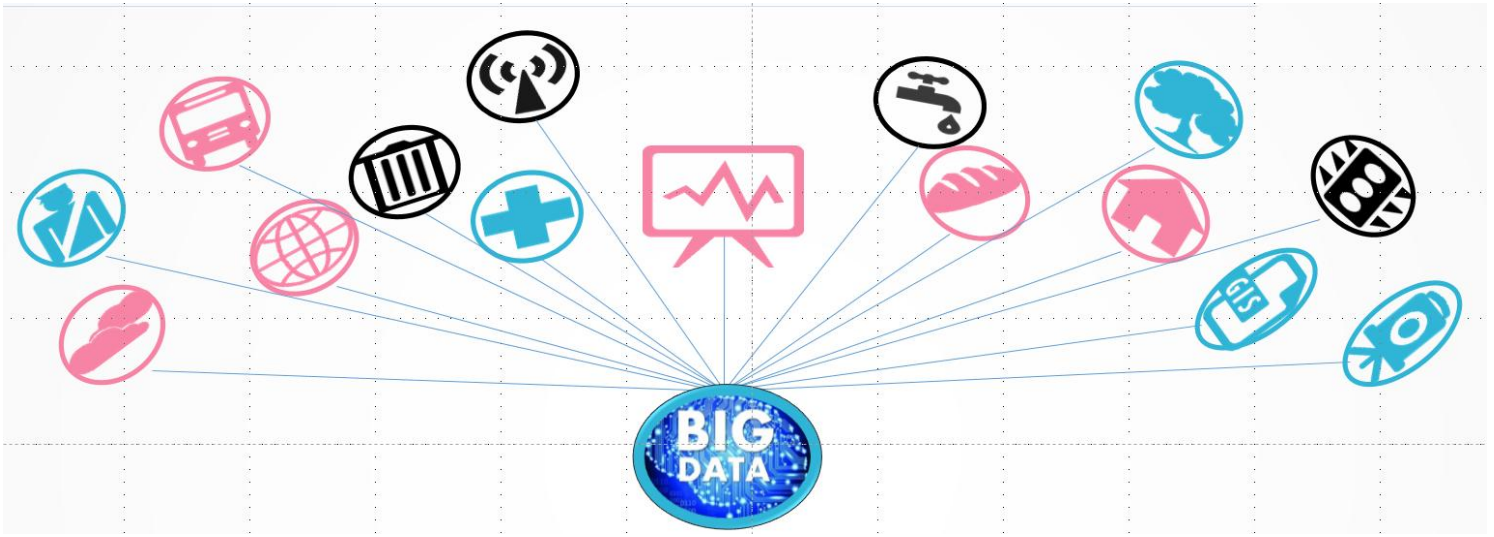
哈尔滨工程大学 HARBIN ENGINEERING UNIVERSITY

第2章 网络信息内容获取技术

本章首先给出互联网信息分类，然后深入研究信息内容获取技术；包括网络媒体信息的获取以及网络通信信息的获取。

2.1 互联网信息分类

- 网络媒体信息划分方法



- 网络媒体形态
- 发布信息类型
- 媒体发布方式
- 网页具体形态

网络媒体信息是指互联网公开发布的信息

2.1 互联网信息分类

- 根据网络媒体形态分类

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

- 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

2.1 互联网信息分类

- 按发布信息类型分类

主流信息

日趋增多

文本信息、图像信息

音频信息与视频信息

- 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。

2.1 互联网信息分类

- 按媒体发布方式分类

直接匿名浏览的公开发布信息

需要实现身份认证访问的网络媒体信息

- 按照网络媒体所选择信息发布方式的不同，网络媒体信息还可以分成可直接匿名浏览的公开发布信息，以及需要实现身份认证才可以进一步点击阅读的网络媒体发布信息

2.1 互联网信息分类

- 按网页具体形态分类

静态网页。

数据指数增长

动态网页

结构日趋复杂

- 按网页内容的具体构成形态，还可以对网络媒体信息中的静态网页与动态网页

2.1 网络通信信息分类

- 按网页具体形态分类

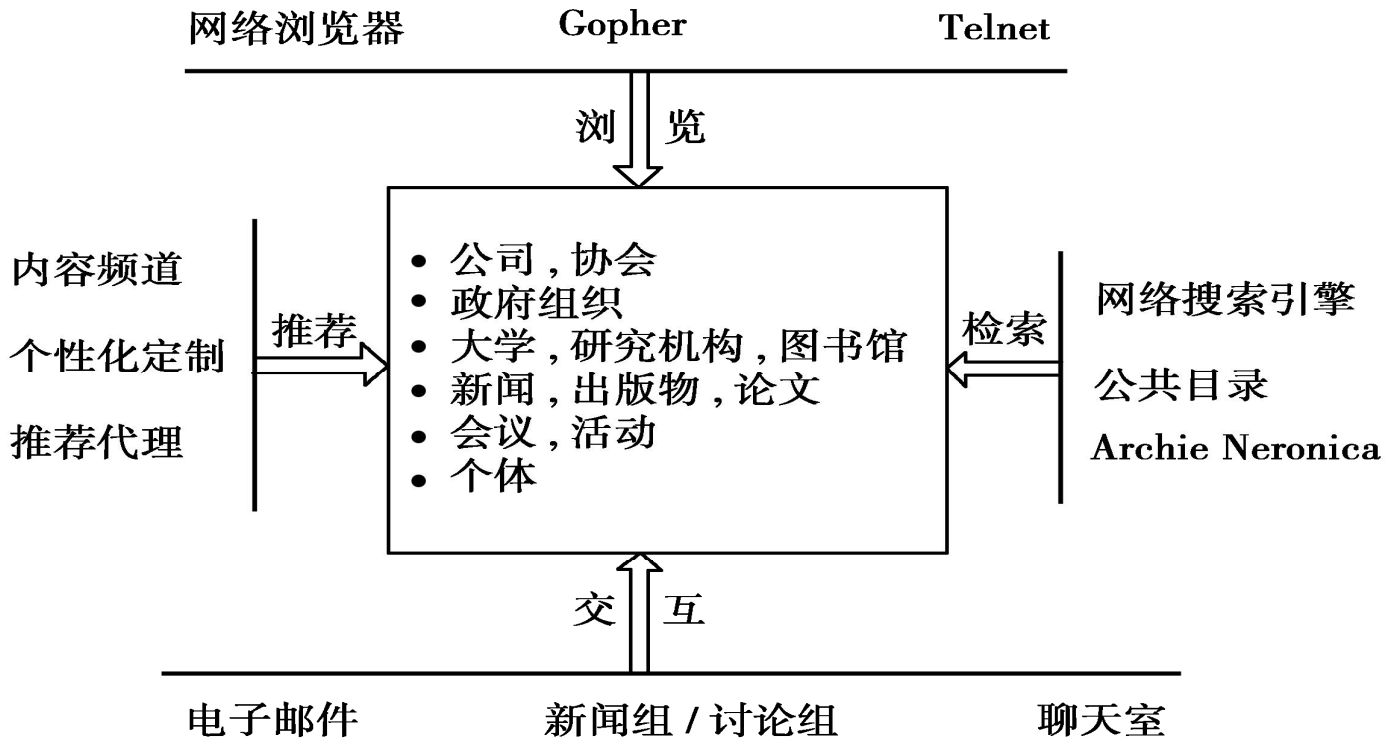
IM**通讯软件**

电子邮件客户端收发邮件

- 网络通信信息一般指互联网用户使用除网络浏览器以外的专用客户端软件，实现与特定点的通信或进行点对点通信时所交互的信息。常见的网络通信信息包括使用电子邮件客户端收发信件时通过网络传输的信息，以及使用即时聊天工具进行点对点交流时所传输的网络信息。

2.2 网络媒体内容获取

- 网络信息内容获取是指从网络收集数据的过程，分为信息检索、信息推荐、信息浏览和信息交互四种。



2.2 网络媒体内容获取

- 信息检索 （Information Search, IS）是信息的需求者主动地在网上搜寻所需要的信息。
- 1951年，Calvin Mooers首次提出了“信息检索（Information Retrieval, IR）”概念，并给出了信息检索的主要任务：即协助信息的潜在用户将信息需求转换成一张文献来源信息列表，而这些文献包含对用户有用的信息。目前通常使用搜索引擎技术完成信息检索功能。

2.2 网络媒体内容获取

- 信息浏览方式相当于传统情况下的阅读、观看、倾听等获取信息的行为。

2.3 搜索引擎技术

- 搜索引擎的祖先是1990年由蒙特利尔大学学生Alan Emtage发明的Archie。在因特网还没有出现以前，Archie可以帮助人们搜索散布在各个分散的FTP主机中的大量文件。1994年7月，Michael Mauldin首次将网络爬虫程序与文本索引程序相结合，创建了现在仍在提供服务的Lycos搜索引擎。
- 1995年，Stanford大学的两名博士生David Filo和杨致远共同创办了基于目录索引结构的雅虎（Yahoo!）搜索引擎，并成功地使网络搜索概念深入人心，从此，搜索引擎进入了高速发展时期。

2.3 搜索引擎技术

- 从2003年开始，中文网络信息服务的四大门户网站（新浪、搜狐、网易和腾讯）陆续推出了自己的搜索引擎服务，大大促进了中文信息检索技术的发展。

搜狐微博 维护中.....

对不起，服务器正在疯狂维护中，请您稍后再试。

Tencent 腾讯

智 慧 沟 通　灵 感 无 限

搜狐博客 POWERED BY SOHU

中国站长站

WWW.ZADMIN.ORG

新浪网

sina.com.cn

哈尔滨工程大学

HARBIN ENGINEERING UNIVERSITY

2.3 搜索引擎技术

- 从2003年开始，中文网络信息服务的四大门户网站（新浪、搜狐、网易和腾讯）陆续推出了自己的搜索引擎服务，大大促进了中文信息检索技术的发展。

搜狐微博 维护中.....

对不起，服务器正在疯狂维护中，请您稍后再试。

Tencent 腾讯

智 慧 沟 通　灵 感 无 限

搜狐博客 POWERED BY SOHU

中国站长站

WWW.ZADMIN.ORG

新浪网

sina.com.cn

哈尔滨工程大学

HARBIN ENGINEERING UNIVERSITY

2.3 搜索引擎技术

- 中文文本信息检索最早见于“ 748工程” 中的汉字情报检索。到了20世纪80年代中后期，中文信息检索研究在计算机处理能力的支持下进入实用化，经典代表是清华大学的《中国学术期刊（光盘版）》
- 2001年，百度搜索面世并开始逐渐成为中文搜索引擎市场的领头羊。

2.3 搜索引擎技术

- 中文文本信息检索最早见于“ 748工程” 中的汉字情报检索。到了20世纪80年代中后期，中文信息检索研究在计算机处理能力的支持下进入实用化，经典代表是清华大学的《中国学术期刊（光盘版）》
- 2001年，百度搜索面世并开始逐渐成为中文搜索引擎市场的领头羊。

2.3 搜索引擎技术

- 中文搜索引擎的关键技术包括网页内容分析、网页索引、查询解析和相关性计算。一个通用搜索引擎包括网上采集、索引、查询、排级和提交等算法，相关概念参下表。

术语	定义
URL	网页地址，例如：http：//www.google.com
采集（Crawling）	通过从一个种子开始递归地跟踪链接来穿越互联网
索引（Indexes）	允许快速确定爬过的、包含特定词或短语的网页数据结构
垃圾信息（Spamming）	发布为获取经济利益所设计的操纵搜索排名的、人为的网页材料
哈希函数（Hashing function）	一种算法，用于在所希望的范围内根据一个字符串计算出一个整数，使得所有的整数都是从很大的字符串集生成的，分布较均匀，例如URL

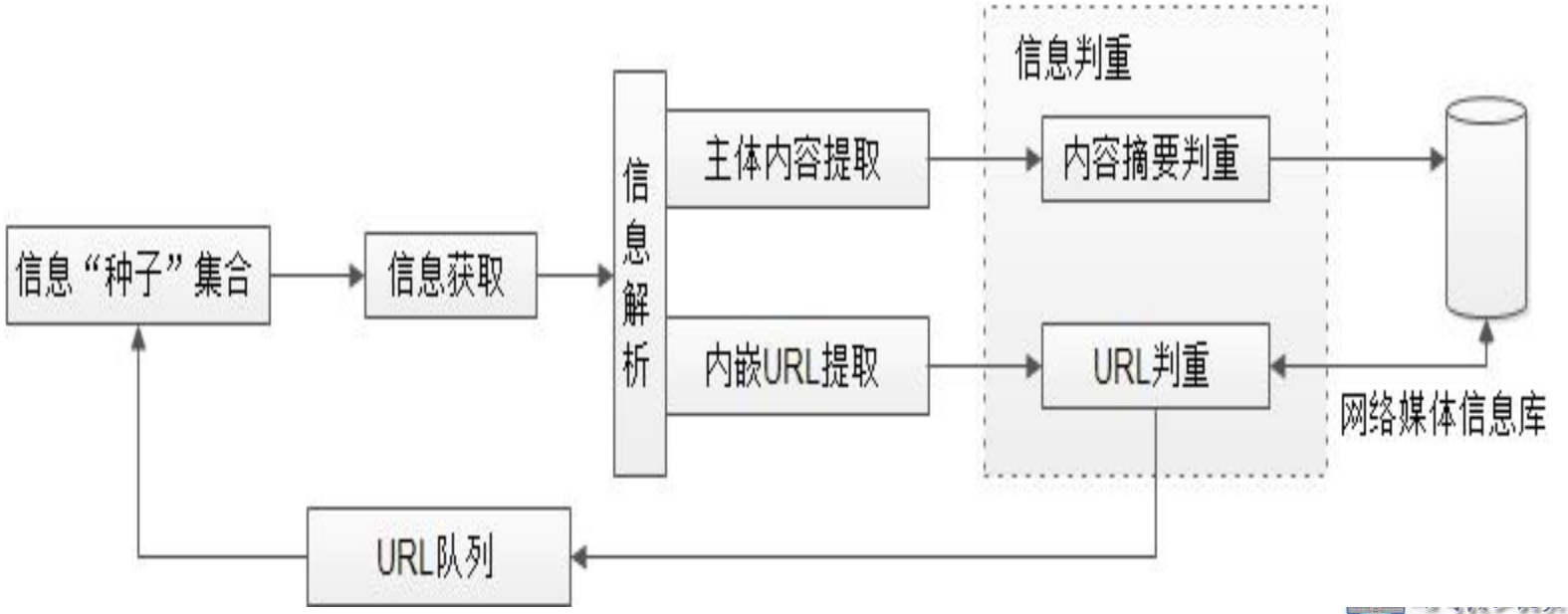
2.3 搜索引擎技术

- 中文搜索引擎的关键技术包括网页内容分析、网页索引、查询解析和相关性计算。一个通用搜索引擎包括网上采集、索引、查询、排级和提交等算法，相关概念参下表。

术语	定义
URL	网页地址，例如：http：//www.google.com
采集（Crawling）	通过从一个种子开始递归地跟踪链接来穿越互联网
索引（Indexes）	允许快速确定爬过的、包含特定词或短语的网页数据结构
垃圾信息（Spamming）	发布为获取经济利益所设计的操纵搜索排名的、人为的网页材料
哈希函数（Hashing function）	一种算法，用于在所希望的范围内根据一个字符串计算出一个整数，使得所有的整数都是从很大的字符串集生成的，分布较均匀，例如URL

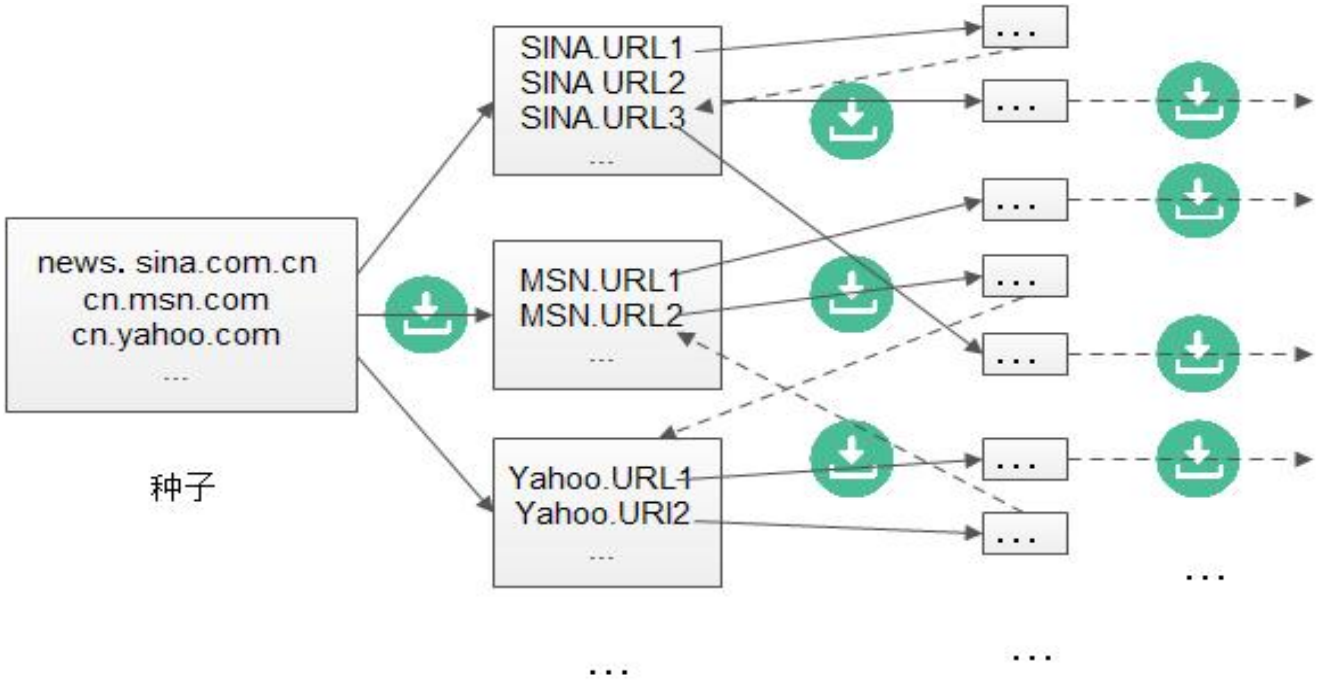
2.3 搜索引擎技术

- 网络媒体信息获取流程
  - 初始URL集合
  - 信息获取
  - 信息解析
  - 信息判重



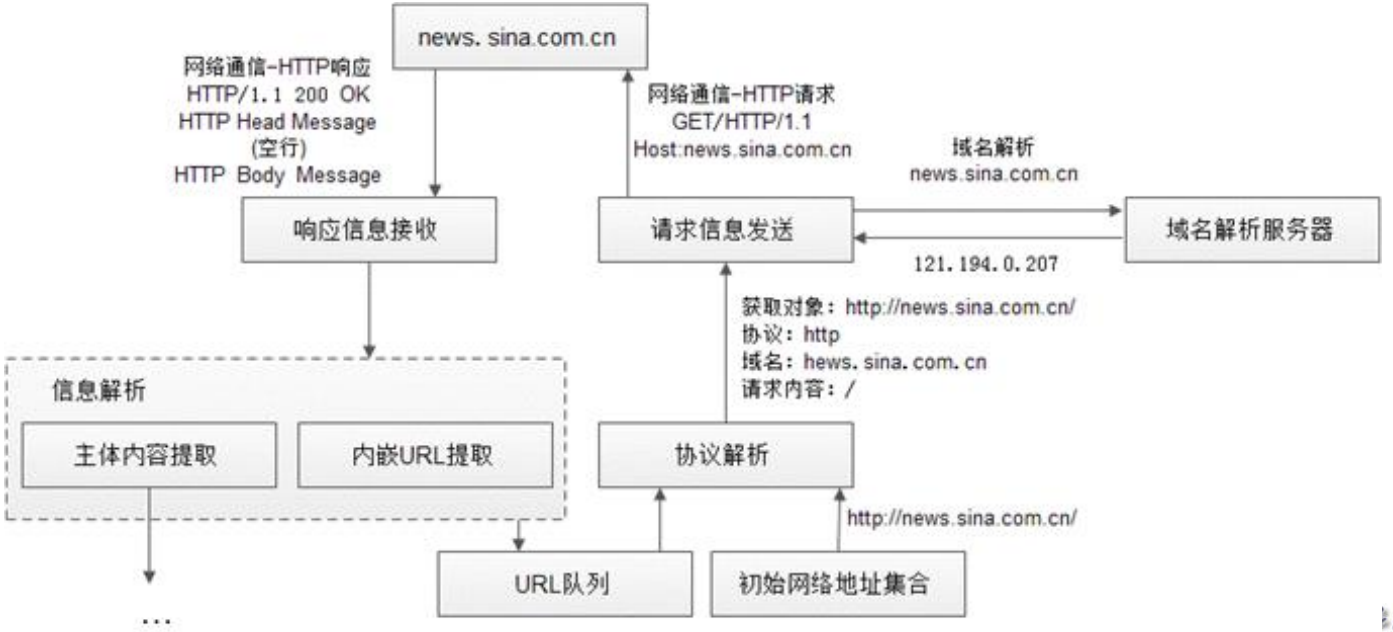
2.3 搜索引擎技术

- 初始URL集合
  - 维护相当数量初始URL集合
  - 网页内嵌地址的递归操作
  - 形象地称为信息“种子”集合



2.3 搜索引擎技术

- 信息获取
  - 维护相当数量初始URL集合
  - 向信息发布网站请求所需内容
  - 接收来自网站的响应信息
  - 传递给后续的信息解析模块

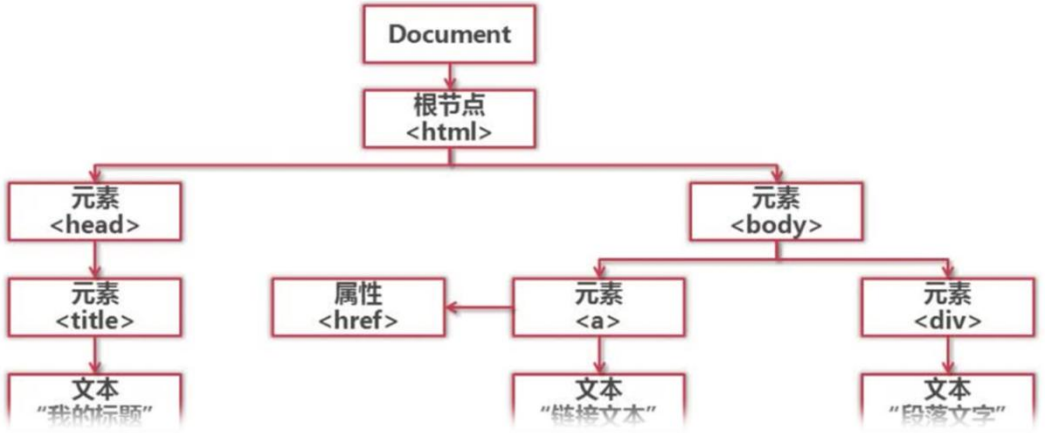


2.3 搜索引擎技术

- 信息内容解析
  - 提取发布信息的主体内容
  - 维护与网络内容的关键字段
  - 内容转交至信息判重模块
  - 关键字段存入信息库

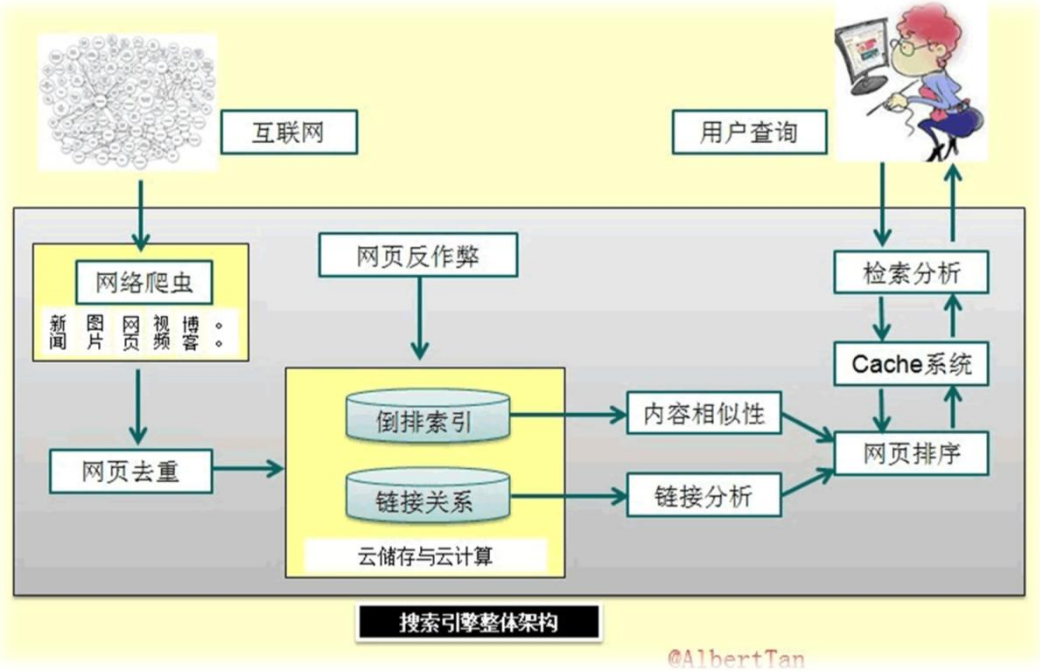
网页解析器

结构化解析-DOM（Document Object Model)树



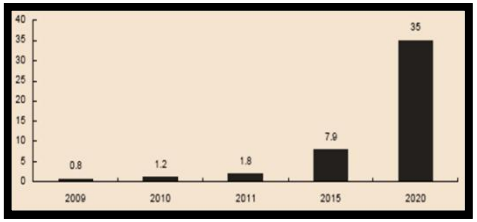
2.3 搜索引擎技术

- 信息判重
  - 判定是否已获取内嵌URL信息内容
  - 若是，注明信息失效时间及最近修改时间的URL
  - 否则重启完整的信息采集操作
  - 向对应的网络内容发布媒体发起信息查新获取操作

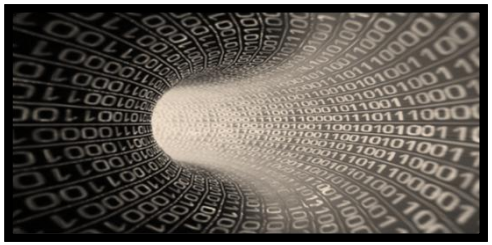


2.3 搜索引擎技术

- 网页媒体信息获取方法



带身份认证静态媒体发布信息获取

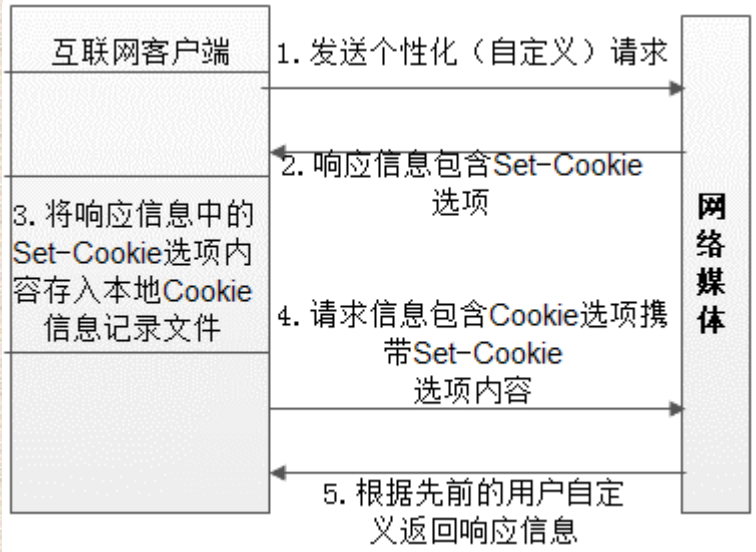


内嵌脚本语言片段的动态网页信息获取

- 按网页具体形态分类，网络媒体信息又可分成静态网页与动态网页两类，不同方法获取方式不一样

2.3 搜索引擎技术

- 基于Cookie机制实现身份认证

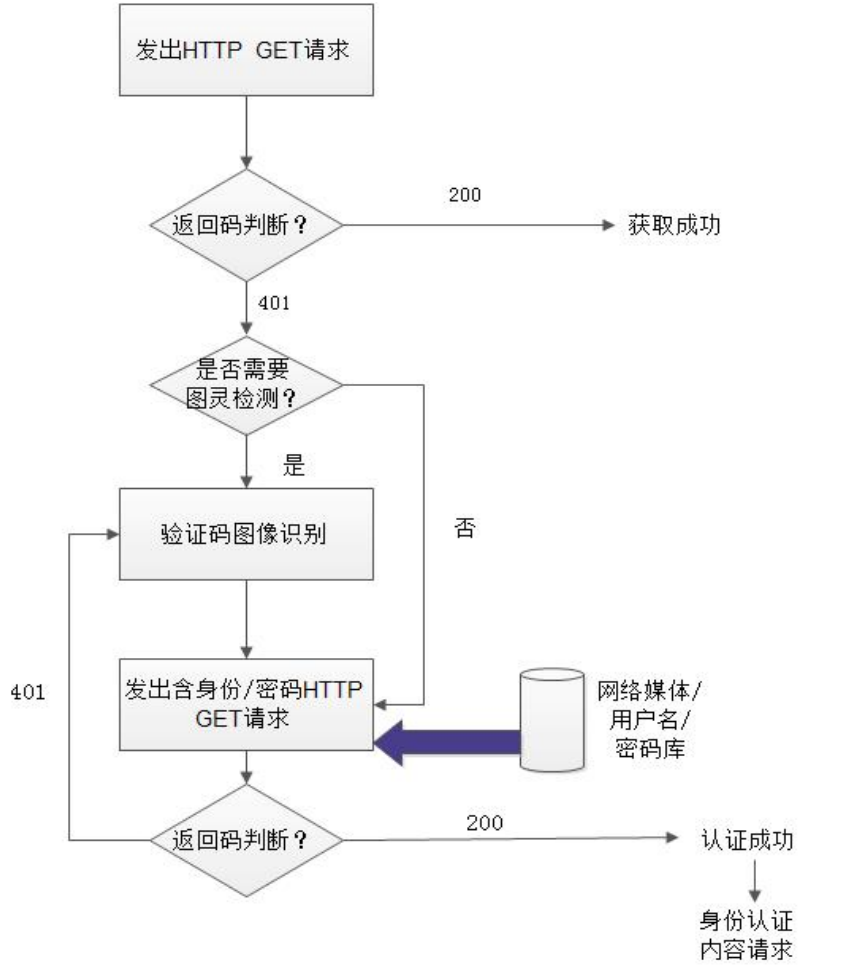


基于Cookie机制的HTTP信息交互过程

基于Cookie机制实现需身份认证才可访问信息请求

2.3 搜索引擎技术

- 基于网络交互重构实现信息获取
  - 通过网络编程顺序模拟网络媒体信息请求过程
  - 网络身份认证过程，都需要进行正确的网络交互过程模拟
  - 媒体信息获取环节是通过响应信息返回码判断信息获取请求是否成功的



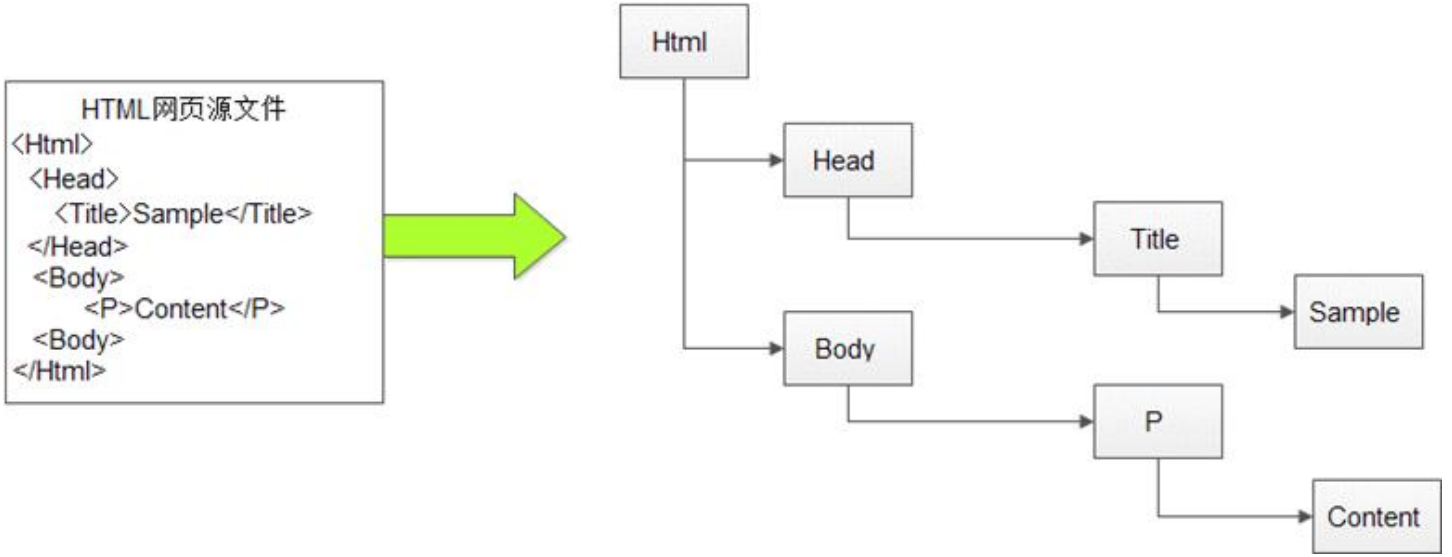
2.3 搜索引擎技术

- 动态网页信息获取

- 动态网页主体内容及其内嵌URL信息完全封装于网页源文件中的脚本语言片段内，无法直接使用基于HTML标记匹配方法提取网页主体内容及其内嵌URL信息。
- 可将脚本语言片段传递给Mozilla浏览器的脚本解释组件实现动态脚本解析并获得脚本片段所对应的静态网页内容

2.3 搜索引擎技术

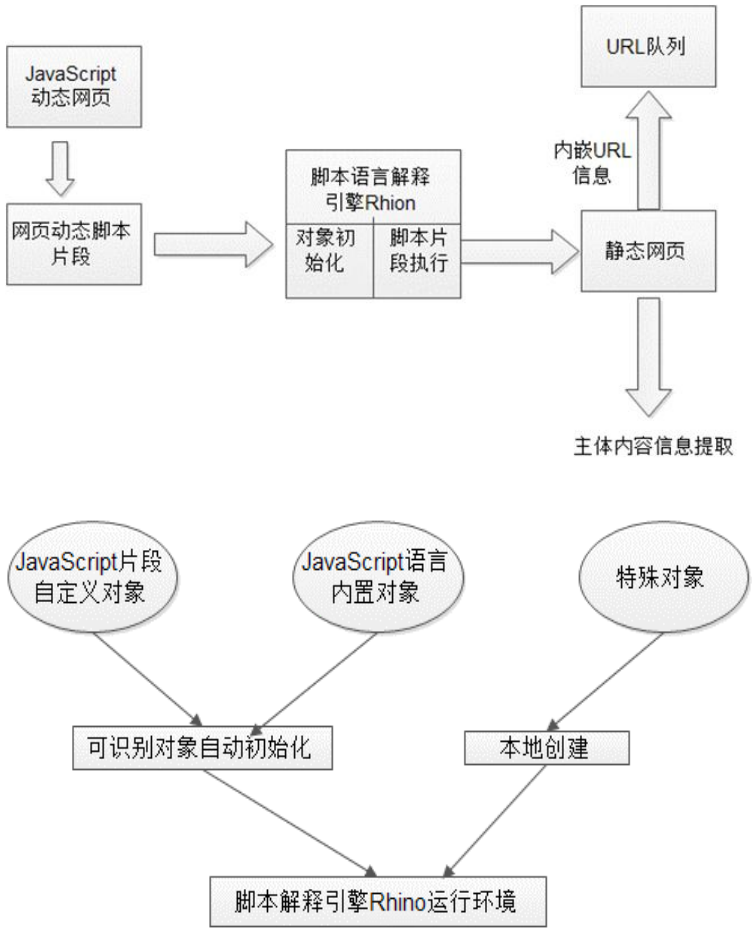
- 动态网页信息获取
  - 利用HTML DOM树提取动态网页内的脚本语言片段

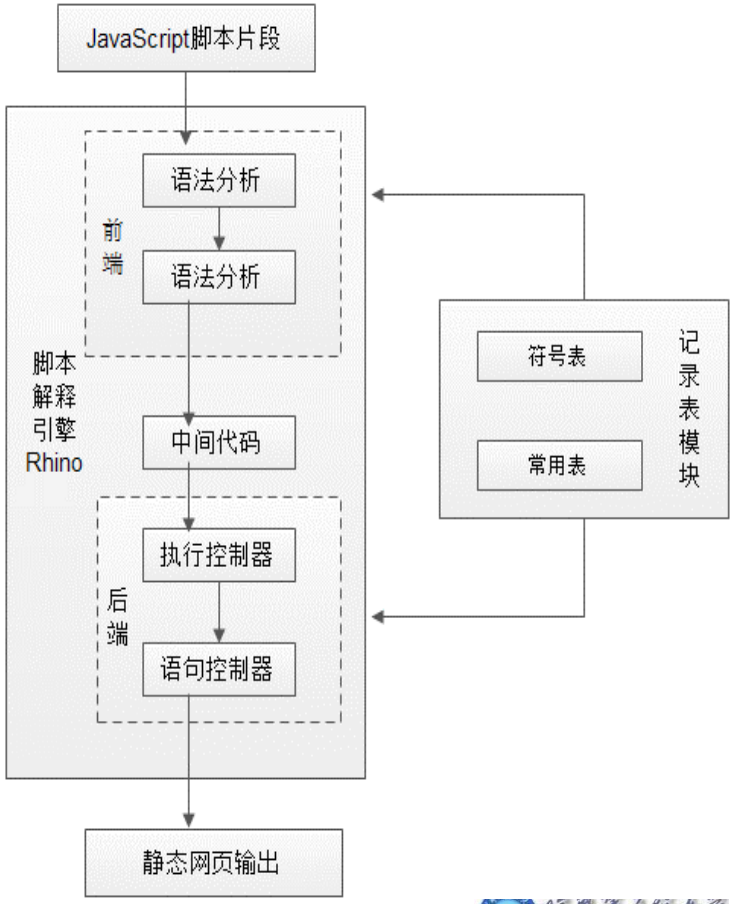


- HTML网页对应的HTML DOM树存储于浏览器内存对象中，该对象实现了包含若干方法的标准程序接口。网页开发人员可以通过相应接口，对HTML DOM树上的每个结点进行遍历、查询、修改或删除等操作，从而动态访问和实时更新HTML网页的内容、结构与样式

2.3 搜索引擎技术

- 基于Rhino实现JavaScript动态网页信息获取





2.3 搜索引擎技术

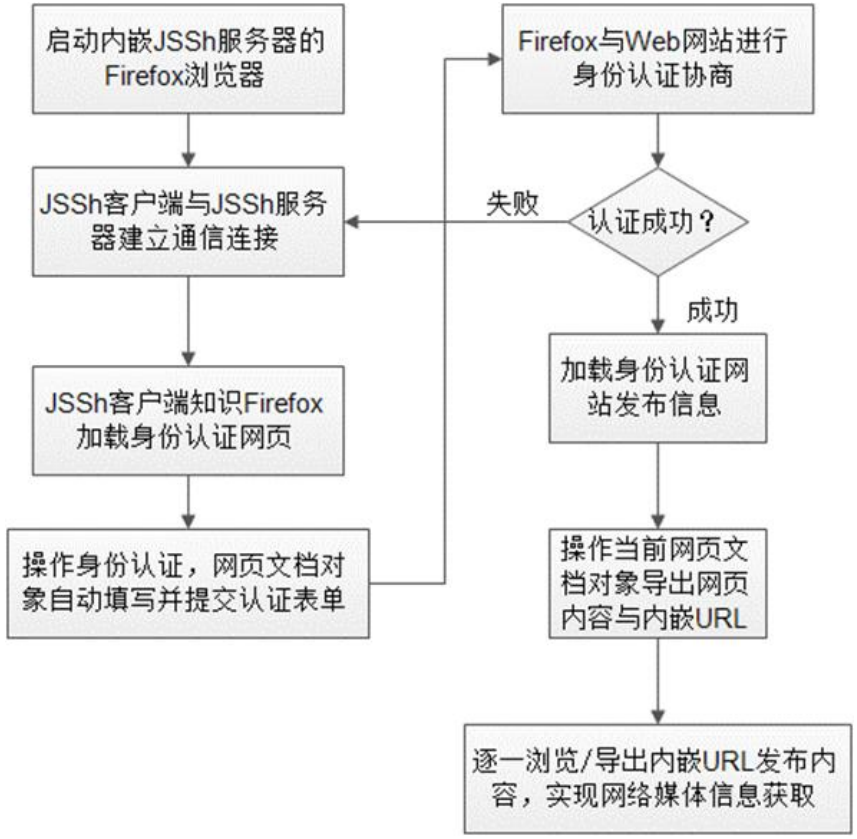
- 基于浏览器模拟实现网络媒体信息获取

- 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/链接被点击、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。

2.3 搜索引擎技术

- 基于浏览器模拟实现网络媒体信息获取

- 身份认证表单自动填写
- 向内嵌JSSh服务器发送指令，指示浏览器加载身份认证网站发布信息
- 对网页内嵌URL逐一进行点击浏览与内容导出



2.4 数据挖掘技术

- 1989年8月，第11届国际人工智能联合会议（IJCAI1989）在美国底特律举行，GTE实验室的G. Piatetsky-Shapiro牵头组织了一个名为“在数据库中发现知识（Knowledge Discovery in Database，KDD）”的研讨会[14]，标志着数据挖掘[15]成为一个新领域。

- 到1995年，在美国计算机年会（ACM）上，提出了数据挖掘（Data Mining，DM）概念，即通过从数据库中抽取隐含的、未知的、具有潜在使用价值信息的过程。

2.4 数据挖掘技术

- Web数据挖掘，即网络知识发现（knowledge discovery in Web, KDW）是一门交叉性学科，涉及数据库、机器学习、统计学、模式识别、人工智能、计算机语言、计算机网络等多个领域，其中，数据库、机器学习、统计学的影响无疑是最大的。Web挖掘是指从大量非结构化、异构的Web信息资源中发现兴趣性（interestingness）的知识，包括概念、模式、规则、规律、约束及可视化等形式的非平凡过程。这里，兴趣性是指有效性、新颖性、潜在可用性及最终可理解性。

2.4 数据挖掘技术

- Web数据挖掘，即网络知识发现（knowledge discovery in Web, KDW）是一门交叉性学科，涉及数据库、机器学习、统计学、模式识别、人工智能、计算机语言、计算机网络等多个领域，其中，数据库、机器学习、统计学的影响无疑是最大的。Web挖掘是指从大量非结构化、异构的Web信息资源中发现兴趣性（interestingness）的知识，包括概念、模式、规则、规律、约束及可视化等形式的非平凡过程。这里，兴趣性是指有效性、新颖性、潜在可用性及最终可理解性。

2.5信息推荐技术

- Resnick和Varian在1997年给出了信息推荐的非形式化定义[26]：利用电子商务网站向客户提供商品信息和建议，帮助用户决定应购买什么产品，模拟销售人员帮助客户完成购买过程。信息推荐有三个组成要素：推荐候选对象、用户、推荐方法。信息推荐过程如下：用户可以向推荐系统主动提供个人偏好信息或推荐请求；如果用户不提供，推荐系统也可主动采集；推荐系统可以使用不同的推荐策略进行推荐，推荐系统将推荐结果返回给用户使用。

2.5信息推荐技术

- Resnick和Varian在1997年给出了信息推荐的非形式化定义[26]：利用电子商务网站向客户提供商品信息和建议，帮助用户决定应购买什么产品，模拟销售人员帮助客户完成购买过程。信息推荐有三个组成要素：推荐候选对象、用户、推荐方法。信息推荐过程如下：用户可以向推荐系统主动提供个人偏好信息或推荐请求；如果用户不提供，推荐系统也可主动采集；推荐系统可以使用不同的推荐策略进行推荐，推荐系统将推荐结果返回给用户使用。

2.5信息推荐技术

- 协同过滤推荐（collaborative filtering recon… …技术之一，… …，计算对… …进行排序的

2.6网络通信信息获取

有线网络

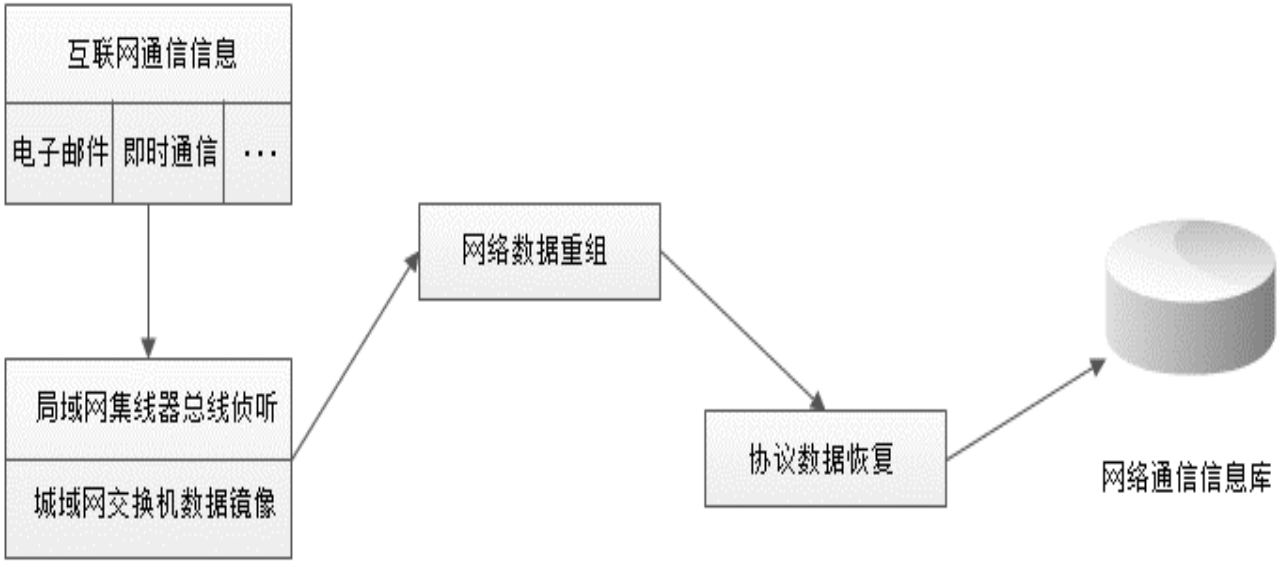
- 端口镜像复制模式
- 攻击交换机以得到所有的数据包
- 攻击模式包括：MAC Flooding攻击和ARP包欺骗

无线网络

- 设置成混杂模式，缺点不能获得802.11b的帧头
- 射频监听工作模式
- 能捕获到其所在的基本服务集中的所有数据包

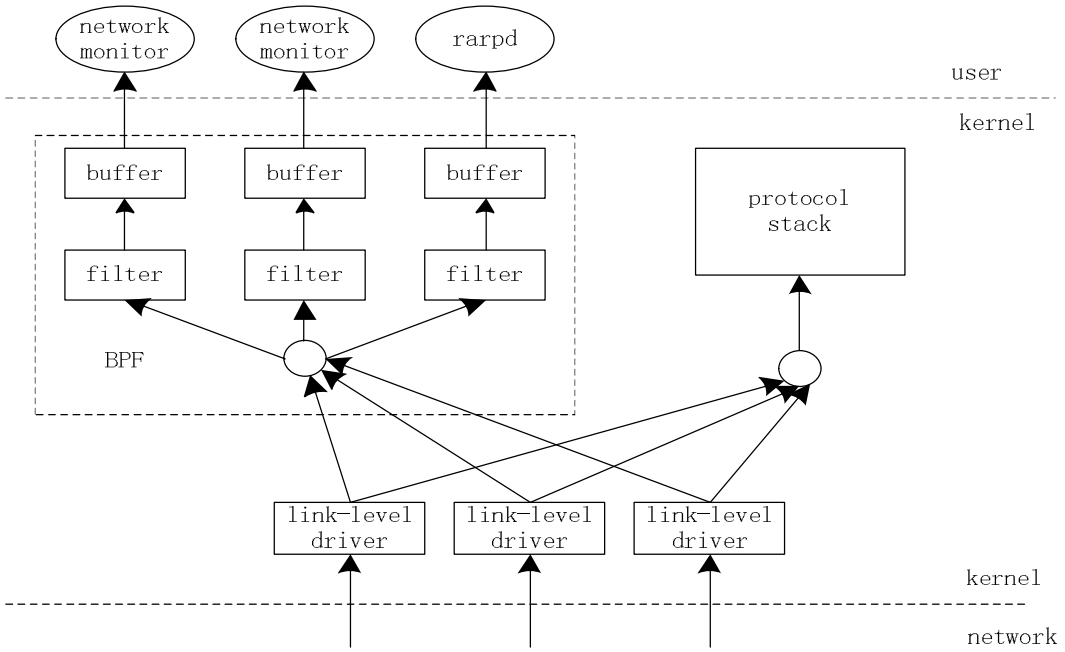
2.6网络通信信息获取

- 获取流程
  - 包括信息镜像、数据重组、数据恢复、信息存储等技术环节
  - 数据获取及交互主要通过三层交换机通信端口数据导出的方式
  - 避免获取个人隐私范畴的网络通信信息



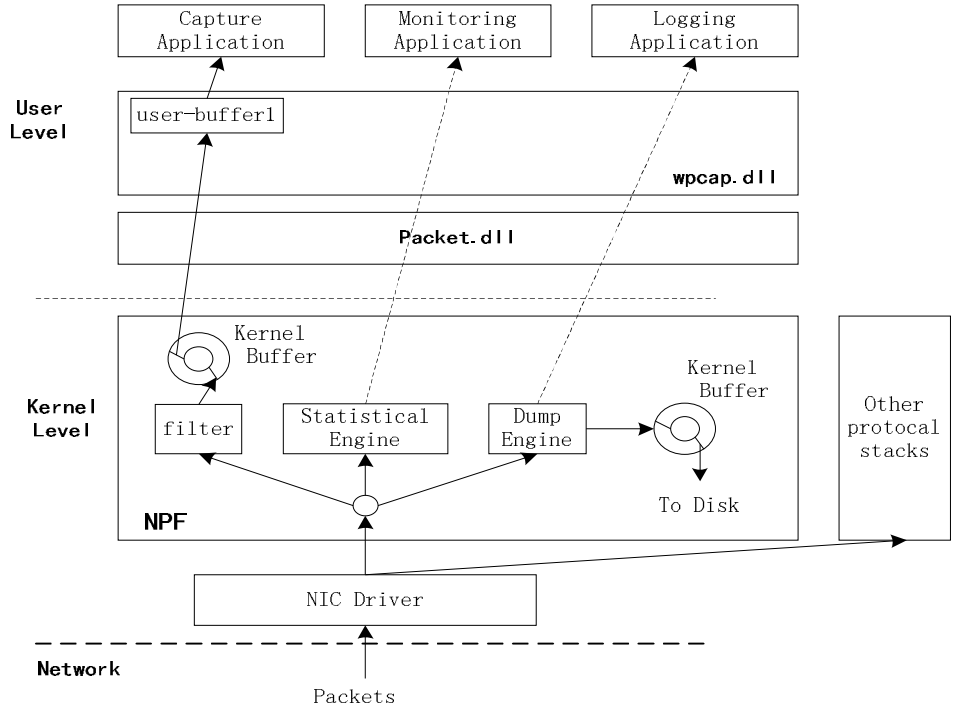
2.6网络通信信息获取

- Linux 主要采用BPF加载模块的方式实现数据包信息的俘获
  - 向用户程序提供了一套功能强大的抽象接口
  - 根据用户要求生成过滤指令
  - 管理用户缓冲区（User buffer）负责用户程序和内核的交互



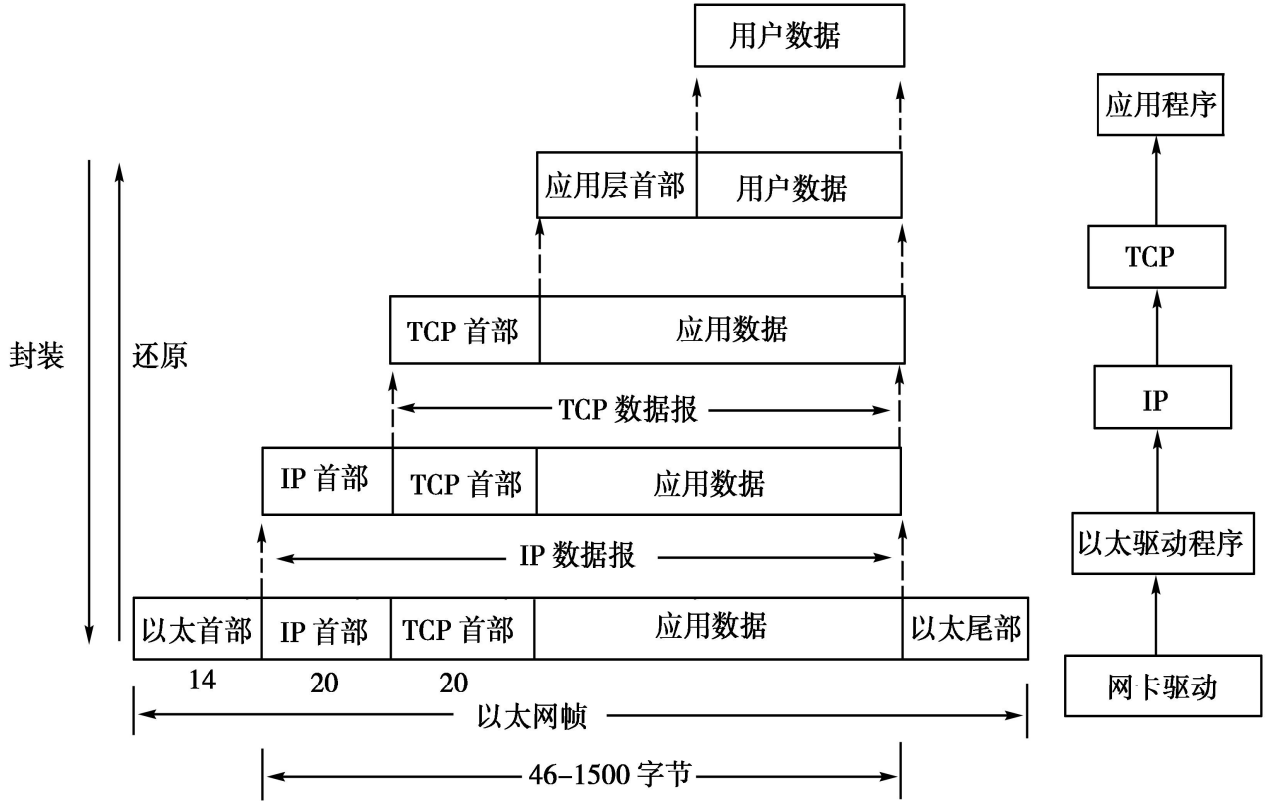
2.6网络通信信息获取

- 获取流程
  - NPF主要用于Windows系统平台需安装winpcap驱动安装包
  - 通过WinPcap将捕获过滤机制内置于操作系统
  - 高级系统无关库、低级动态链接库和内核级的数据监听驱动



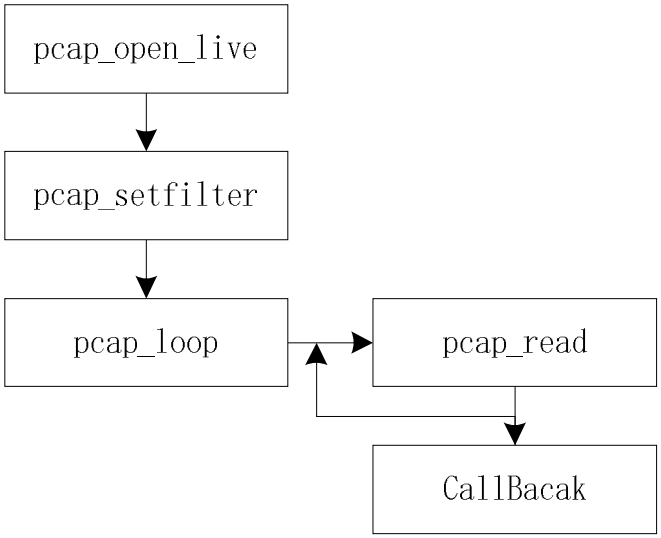
2.6网络通信信息获取

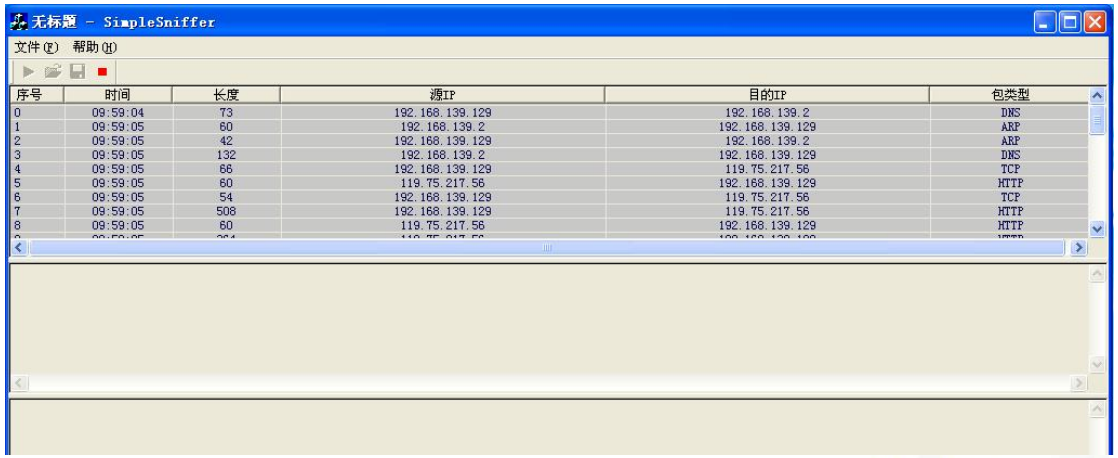
- 协议还原技术。当一个数据包从外部网络到达内部网络或者内部主机时，以链路层协议、TCP/IP协议、应用层标准的协议基本原理为依据，系统依次对链路层数据包、IP层数据包、TCP成数据包、应用层数据包进行的一系列数据包处理过程。



2.6网络通信信息获取

- 基于Winpcap的数据包捕获程序设计
  - 选择监听网络接口
  - 建立监听会话
  - 编辑过滤器
  - 设置过滤器
  - 捕获数据包





- 一些报文分析工具
  - Etheral、Wireshark