第一章 概述

1. 网络安全威胁的主体有哪些？公共基础设施安全运行、国家军事安全、经济金融安全、文化安全。**业余黑客 黑产组织 网络犯罪团伙或黑客组织 网络恐怖组织 一般/高级/超高能力国家/地区行为体**

2. 什么是网络空间？网络空间的四要素包括哪些？网络空间就是所有由可对外交换信息的电磁设备作为载体，通过与人互动而形成的虚拟空间。包括互联网、通信网、广电网、物联网、社交网络、计算系统、通信系统、控制系统等。四要素：**“网络角色”依托“信息通信技术系统”以进行“广义信号”的“交互”。**“信息通信技术系统”包括互联网、各种电信网与通信系统、各种传播系统与广电网、各种计算机系统、各类关键工业设施中的嵌入式处理器和控制器等声光电磁或数字信息处理设备，用“设施”来表征。“广义信号”是指基于声、光、电、磁等各类能够用于表达、存储、加工、传输的电磁信号，以及量子信号、生物信号等能够与电磁信号进行交互的信号形态，这些信号通过在信息通信技术系统中进行加工处理而成为“信息”，用“数据”来表征。“设施”与“数据”反映的是信息通信技术基础设施。“网络角色”是指产生、传输广义信号的主体，用“用户”来表征；“交互”是指用户借助广义信号，以信息通信技术设施为平台，以信息通信技术为手段，达到产生信号、保存数据、修改状态、传输信息、输出结果等表达人类意志的行为，用“操作”来表征。“用户”与“操作”反映的是“与信息通信技术相关的活动”。

3. 《网络安全法》中的主体、客体主要有哪些？试列举各主体的基本责任和义务。

主体：人：**管理者：政府部门；网络运营者：是指网络的所有者、管理者和网络服务提供者；产品和服务提供者(360、华为、阿里等公司)；信息传播者(微博、微信等)；信息提供者(新浪网、凤凰网等)**客体：**财物；网络数据：是指通过网络收集、存储、传输、处理和产生的各种电子数据。公民个人信息：是指以电子或者其他方式记录的能够单独或者与其他信息结合识别公民个人身份的各种信息。产品（软件、硬件）、服务。网络基础设施。责任义务**：（一）制定内部安全管理制度和操作规程，确定网络安全负责人，落实网络安全保护责任；（二）采取防范计算机病毒和网络攻击、网络侵入等危害网络安全行为的技术措施；（三）采取监测、记录网络运行状态、网络安全事件的技术措施，并按照规定留存相关的网络日志不少于六个月；（四）采取数据分类、重要数据备份和加密等措施；

4. 什么是网络空间？什么是***网络空间安全***？网络空间安全研究方向有哪些？

网络空间定义见第二题。**网络空间安全**主要是在信息通信技术的电磁设备、电子信息系统、运行数据、系统应用等系统与应用层面上，围绕信息获取、信息传输、信息处理、信息利用等核心功能，针对网络空间的设施、数据、用户、操作等核心要素来采取安全保护措施，以确保网络空间中信息通信技术系统及其所承载数据的机密性、可鉴别性（包含完整性、真实性、不可抵赖性）、可用性、可控性等元安全属性得到保障，从而保证信息通信技术系统能够提供安全、可信、可靠、可控的服务。**网络空间安全研究方向：网络空间安全基础、密码学及应用、系统安全、网络安全、应用安全、信息内容安全**

5. 国家***网络强国战略***的主要内容？我国政府中网络安全部门主要有哪些？

主要内容：重视互联网、发展互联网、治理互联网、人才培养**主要部门**：国家网信部门、国务院电信主管部门、公安部门和人民政府有关部门（网信办、工信部、公安部门）

6. 什么是***网络空间主权***？基本原则是什么？（网络独立权，平等，自卫，管辖）

一个国家的网络空间主权建立在本国所管辖的信息通信技术系统之上（领网）；其作用边界为由直接连向他国网络设备的本国网络设备端口集合所构成（疆界）；用于保护虚拟角色对数据的各种操作（政权、用户、数据）。网络空间的构成平台、承载数据及其活动受所属国家的司法与行政管辖（管辖权）；各国可以在国际网络互联中平等参与治理（平等权）；位于本国领土内的信息通信基础设施的运行不能被他国所干预（独立权）；国家拥有保护本国网络空间不被侵犯的权力及其军事能力（自卫权）。网络空间主权应该受到相互尊重（尊重主权）；国家间互不侵犯他国的网络空间（互不侵犯）；互不干涉他国的网络空间管理事务（不干涉他国内政）；各国网络空间主权在国际网络空间治理活动中具有平等地位（主权平等）。

基本原则：网络空间主权的基本原则也是来自于国家主权。尊重网络主权，就是要尊重网络独立权，不采取导致主权网络空间无法自主运行的行为；互不侵犯，就是不能对他国的网络空间实施网络攻击；互不干涉网络内政，就是对主权网络空间的管辖权不指手画脚；网络主权平等，就是主权国家之间具有平等共治网络空间的权力，而不是依靠“利益攸关方”的模式导致一些国家失去了参与网络共治的权力，而另一些国家则掌控了全球的网络空间。

7. 什么是信息内容安全？主要涉及对传播信息的有效审查监管，剔除非授权信息（非法信息、泄密信息、垃圾邮件等），保护授权信息。

8. 信息内容安全的***主要威胁***有哪些？有哪些典型事件？（净网 反恐 反腐 知识产权 邪教 反分裂 隐私保护）政治方面：防止来自国内外反动势力的分裂、邪教、暴恐行为，虚假内容传播；健康方面：剔除色情、淫秽和暴力内容等；保密方面：防止国家和企业机密被窃取、泄露和流失；隐私方面：防止个人隐私被盗取、倒卖、滥用和扩散；产权方面：防止知识产权被剽窃、盗用等；破坏方面：防止病毒、垃圾邮件、网络蠕虫等恶意信息耗费网络资源；经济方面：防止赌博、诈骗、骗贷等；净网行动 全民反恐 全民反腐 反邪教 反分裂 知识产权 隐私保护典型事件：全球反恐打黑、信息战、网络诈骗与舆论安全、**BT** 侵权、影视盗版

9. ***信息内容安全技术***主要包括哪些？信息理解、发现与追踪技术，信息识别、检索与筛选技术，信息监测与阻断技术，内容信息版权保护技术。大规模串匹配技术、面向海量文本信息的分类聚类技术、语音/图像识别技术、信息渗透及检测

10. 信息内容安全技术面临的***挑战***是什么？（数据量大，计算复杂度高，网络技术新，社会矛盾深）**规模巨大**：*非结构化数据*的超大规模，比结构化数据增长快 10 倍到 50 倍；产生高速：*实时分析*而非批量式分析，数据输入、处理与丢弃，立竿见影而非事后见效；形式多样：*异构性*（文本、图像、视频、机器数据），模式不明显，语法语义不连贯；信息价值：*大量的不相关信息*，对未来趋势与模式的深度复杂分析（机器学习、人工智能）

11. 网络与信息安全问题的本质根源是什么？物理安全问题、方案设计的缺陷、系统的安全漏洞、TCP/IP 协议的安全问题、人为的无意失误、人为的恶意攻击、管理上的因素

第二章 网络信息获取

	被动获取	主动获取
是否需要协作	需要授权	无需授权
获取范围	仅仅包括途径机器的流量(局部网络)	网上的开放信息(互联网)
数据处理实时性	实时性要求高(在线)	无需实时性要求(离线)
技术指标	吞吐量、丢包率	爬取效率
技术难点	理论分析	适应不同的发布方式

网络信息获取方式分为哪两类？试比较各自的主要特点。

被动获取需要获得管理员授权，协助接入网络设备；且要求高实时性，否则丢包。主动获取无需协助接入，只要客户端连入互联网即可获取，且数据处理可离线进行，无需实时性。 被动获取需要对应用层协议进行解析，一事一议；主动获取需要针对信息发布方式设计爬虫程序，一事一议

1. ***BPF 捕包原理***是什么？ Berkeley Packet Filter,是一个高效的数据包捕获机制，工作在操作系统的内核层。BPF 主要由网络转发部分和数据包过滤两部分组成。网络转发部分是从链路层捕获数据包并把它们转发给数据过滤部分，数据包过滤部分是从接收到的数据包中接收过滤规则决定的网络数据包，其他数据包被丢弃在操作系统的内核中完成，效率很高。使用了数据缓存机制，使捕获数据包缓存在内核中，达到一定数量再传递给应用程序。实际应用中，使用 libpcap。

2. ***被动捕包程序的主要流程***是什么？(1) 把网卡等同于文件进行 I/O ，查找网卡：Find_all_devices()，打开网卡：open ()；(2) 从网卡中读取数据，监听：loop()，数据回传给用户变量(3) 处理获取的数据，转用户程序执行：Handler()，(4) 释放 I/O 资源

- ***关于各种数据包：IP 首部：版本 4 头部长度 4 服务类型 8 总长 16 标识 16 分段偏移 16 生存期 8 协议 8 头部校验和 16 源 IP32 目的 IP32 选项; TCP 首部：源 16 目的 16 序列号 32 确认号 32 数据偏移 4 保留 6 Flags(6) 窗口 16 校验和 16 紧急指针 16 选项(可变) 填充; UDP 首部：源 16 目的 16 数据包长度 16 校验值 16; (0~31，单位都是 bit 比特)***

3. IP 首部主要包含哪些信息？在数据包分析时如何提取？版本，首部长度，区分服务，总长度，标识，标志，片偏移，生存时间，协议，首部检验和，填充字段，源 IP，目标 IP。用 struct ip_header 类型的结构体指针去做一个指针强制类型转换，然后就可以通过结构体的字段偏移提取出 ip 首部的各个字段。

4. TCP 首部主要包含哪些信息？在分析时如何提取？源端口号、目的端口号、序列号、确认号、数据偏移、URG/ACK/PSH/RST/SYN/FIN 标识、窗口大小、校验和、紧急指针、选项。用 struct tcp_header 去提取。

5. 大流量网络环境下***如何提高捕包系统***的性能？在操作系统的内核中完成，效率很高。使用了数据缓存机制，使捕获数据包缓存在内核中，达到一定数量再传递给应用程序。零拷贝技术。

6. 如何提高单机捕包能力？零拷贝捕包主要解决哪些核心问题?(1)减少系统调用和内存操作。数据报从网络设备到用户程序空间传递的过程中，减少数据拷贝次数，减少系统调用，降低 CPU 在这方面的负载。(2)数据先从网卡到内核，再到用户空间，这个拷贝过程仅仅起到“传输”作用，而数据包内容没有任何变化，零拷贝技术的核心：取消这两次拷贝。

7. 网站信息爬取的主要思路是什么？用伪代码描述***单机网页信息爬取***算法。

（1）选取一部分种子 URL 放入队列。（2）取出 URL，解析 DNS，得到主机 ip，下载存储入库。此外，将这些 URL 放进已抓取 URL 队列中。（3）分析已抓取 URL 队列中的 URL 得到新 URL，进入下一个循环。

```
DS:Url_QUEUE uqueue; History_LIST hlist; PROCEDURE: Crawler(seed_url) { in_queue (uqueue , seed_url); while (u=out_queue(uqueue)) { wpage=http_get(u); save wpage; for each url in wpage { if url not in hlist then in_queue (uqueue , url); } } }
```

8. 多机爬取多个网站，核心解决哪些问题？主要策略有哪几种？控制爬取范围、爬取深度、负载均衡、定期更新、爬取实时性；如何将爬取任务均匀的分配到各个机器上，如何让各个机器同步工作。按 URL 的散列值均匀分配抓取任务。深度优先；广度优先:在目前为覆盖尽可能多的网页，一般使用广度优先搜索方法。缺点：随着抓取网页的增多，大量的无关网页将被下载并过滤，算法的效率将变低；最佳优先:预测候选 URL 与目标网页的相似度，或与主题的相关性，并选取评价最好的一个或几个 URL 进行抓取。它只访问经过网页分析算法预测为“有用”的网页。缺点：在爬虫抓取路径上的很多相关网页可能被忽略，因为最佳优先策略是一种局部最优搜索算法；部分的 PageRank 的策略

9. 用伪代码描述***多机单网站网页信息爬取***算法。

```
DS: Url_QUEUE uqueue; History_LIST hlist; PROCEDURE: Crawler(seed_url) { in_queue (uqueue , seed_url); while (u=out_queue(uqueue)) { wpage=http_get(u); save wpage; for each url in wpage { if url not in hlist then { n=hash(url) mod N; if (n==my_node_ID) then in_queue (uqueue , url); else sendurl(url,n);}}}}
```

10. 简述基于***主动获取技术的网站信息监测系统***的构成。**搜索器、索引器、检索器和用户接口**

11. 简述 Pagerank 算法的主要思路。如果网页 B 存在一个指向网页 A 的连接，则表明 B 的所有者认为 A 比较重要，从而把 B 的一部分重要性得分赋予 A。这个重要性得分值为：PR(B)/L(B) ，PR(B)为 B 的 PageRank 值,L(B)为 B 的出链数; A 的 PageRank 值为一系列类似于 B 的页面重要性得分值的累加。**PR(A)=(PR(B)/L(B)+PR(C)/L(C)+…)q + (1-q)** q 为阻尼系数(一般为 0.85):在任意时刻,用户到达某页面后并继续向后浏览的概率.

12. 网页抓取的搜索策略有哪几种？各自适合哪类应用场景？**深度优先**在很多情况下会导致爬虫的陷入(trapped)问题，目前常见的是**广度优先**和**最佳优先**方法。广度优先搜索策略是指在抓取过程中，在完成当前层次的搜索后，才进行下一层次的搜索。该算法的设计和实现相对简单。在目前为覆盖尽可能多的网页，一般使用广度优先搜索方法。缺点：随着抓取网页的增多，大量的无关网页将被下载并过滤，算法的效率将变低。最佳优先：按照一定的网页分析算法，预测候选 URL 与目标网页的相似度，或与主题的相关性，并选取评价最好的一个或几个 URL 进行抓取。它只访问经过网页分析算法预测为“有用”的网页。缺点：在爬虫抓取路径上的很多相关网页可能被忽略，因为最佳优先策略是一种局部最优搜索算法。

13. 试比较主动获取与被动获取技术的异同。（如何网上违法活动的追踪溯源？）收集证据,数据处理,信息提取,数据分析,溯源追踪,联合打击

第三章 串匹配算法

1. 串匹配算法分类和评价指标有哪些？分类指标：匹配的模式数目(单模式/多模式)、匹配方式(精确/近似)、匹配的具体内容(单词/字符/正则表达式)、实时性(实时文本:动态更新~网络入侵检测/非实时文本:被查找文本是静态的~搜索引擎查找的数据) 评价指标：匹配次数、时间复杂度、是否需要回溯

2. 串匹配算法优化的主要思路？KMP 算法是充分利用已经比较过的字符信息来提高效率，而 BM 算法则是从利用匹配失败时获得的信息出发提高效率,基于(倒序)自动机/哈希/状态合并(自动机)/分治/预处理/并行的匹配算法,

3. 试比较不同串匹配算法的优缺点。KMP 算法时间复杂度 o(m+n)，并且不需要回溯；BM 算法时间复杂度 o(mn)，但是实际效果却常常是最快的.二者:每一次比较后，有明确的信息记录下一次比较的位置.AC 复杂度 o(n),时间复杂度与关键字的数目和长度无关,扫描文本时不需要回溯,wm 算法的复杂度 o(B*N/m)B:块字符数 N:文本长 m:模式最短长度,对最短模式长度敏感.

4. KMP 算法的基本流程？能够针对实际数据给出求解步骤。先求 next 数组：next[i]={-1,i=0;max{k|(1<k<i&& ’ x1…xk-1=’ xi-k+1…xi-1’ ’)},此集合不空;0,其他情况}(优化：若 p[i]=p[next[i]],那么 next[i]=next[next[i]])每次的跳跃长度: i-next[i] (yj 和 xi 匹配失败,就和 x_next[i]匹配) **注意**: next 数组最后比模式串多一个

```
void preKmp-opt(char *x, int m, int kmpNext[ ]){int i, j; i = 0，j=-1; kmpNext[0] = -1; while (i < m) { while (j > -1 && x[i] != x[j]) j = kmpNext [j]; i++; j++; if (x[i] == x[j]) kmpNext[i] = kmpNext[j]; else kmpNext[i] = j; } for （i from 0 to m-1） if (x[i]==x[0] && next[i]==0) next[i]=-1 }
void KMP-opt(char *x, int m, char *y, int n) { int i, j, kmpNext[XSIZE]; preKmp(x, m, kmpNext); i = j = 0; while (j < n) { while (i > -1 && x[i] != y[j])(i = kmpNext[i];
if i==-1 then i=0;}i++;j++;if (i >= m) {OUTPUT(j - i);i = kmpNext[i];}}}
```

5. BM 算法的基本流程？[**取 max**]（主要数据结构）能够针对实际数据给出求解步骤。算法从正文左端开始扫描,对齐后从模式最右端开始自右向左诸字符比较。在不匹配（或完全匹配）时,用两个预先计算的函数 bad character 和 good suffix 来确定指针在正文中移动的距离。计算出字符集中每个字符的偏移值 bmBC[i],对于未在模式中出现的字符，偏移为 m,否则取 m-i-1,(其中 i 为字符在模式中的位置),即取此字符在模式中出现的最右位置和文本中此字符位置对齐；若字符未在模式中出现，则可将模式向前推 m 个字符位置.但是,在某些情况下这种偏移可能是负的.实际的偏移值取得是两者中值较大者 ***坏字符转移的时候转移到的的是最右边的和坏字符相同的字符处，而且坏字符数组是针对每一个不同的字符的移动位数(为了相等)*** ***eg: GCAGAGAG bmBc[c]: A-1 C-6 G-2 T-8; suffix[i]:10020408;bmGs[i]:7727471*** ***shift(坏字符) = bmBc[T[i]]-(m-1-i)*** ****suffix[i]数组是为了找最长的后缀长度(在 i 为末尾分界的情况下)***

bmGs 构造方法: void preBmGs(char *x, int m, int bmGs[]) {int i, j, suff[XSIZE]; suffixes(x, m, suff); ***for*** (i = 0; i < m; ++i) bmGs[i] = m; j = 0; ***for*** (i = m - 1; i >= 0; --i) ***if*** (suff[i] == i + 1) for (; j < m - 1 - i; ++j) if (bmGs[j] == m) bmGs[j] = m - 1 - i; ***for*** (i = 0; i <= m - 2; ++i) bmGs[m - 1 - suff[i]] = m - 1 - i;} **bmGs 三种情况:模式串中没有子串匹配上好后缀,但找不到一个最大前缀； 模式串中没有子串匹配上好后缀,但找到一个最大前缀； 模式串中有子串匹配上好后缀.** **bmBc 构造:** bmBc[x[i]] = m - i - 1;(没出现就是 m 且 i≠m-1) **Suffix 构造方法:** void suffixes(char *x, int m, int *suff) {suff[m-1]=m; for (i=m-2；i>=0；--i){ q=i; while(q>=0&&x[q]==x[m-1-i+q]) --q; suff[i]=i-q;}}

■ AC 算法在千万量级的模式集上如何并行加速?:分组多线程并行,并行 tries 树,分布式计算(多计算节点),GPU 加速

6. AC 算法的基本流程？（失效函数）能够针对实际数据给出求解步骤。建立 goto、fail、output 三个函数。对每个使得 g(r, a) = s 存在的状态 s，执行以下操作记 state = f(r)。零次或者多次令 state = f(state)，直到出现一个状态使得 g(state, a) != fail（注意到，因为对任意字符 a，g(0,a) != fail，所以这种状态一定能够找到）记 f(s) = g(state, a)

WM算法: 模式集(thinner,shinning,church,touching,thinking)(取最小的6个-church)

字符块:	th	hi	in	nn	ne	sh	ni	ch	hu	ur	rc	to	ou	uc	nk	ki	Others
(mod29)	26	15	21	26	17	2	21	9	27	8	19	4	5	22	23	18	
shift表:	th	hi	in	nn	**ne**	sh	**ni**	**ch**	hu	ur	rc	to	ou	uc	nk	**ki**	*Others*
	4	3	2	1	0	4	0	4/0	3	2	1	4	3	2	1	0	5

hash[i]哈希表(元素为指向后面两个表的索引):hash[17]=0,hash[21]=1,...

pat_ptr[]: P0 P1 P2 P3 P4 *pattern_ptr记录每个模式串首地址*

prefix[]: 26 2 9 4 26 *prefix每个模式串前缀*

(对应块: th sh ch to th P.S. 不显示)

7. WM 算法的基本流程？能够针对实际数据给出求解步骤。建立 SHIFT、HASH、PREFIX 表。默认 SHIFT[i]=m-B+1；若块在某模式中出现，SHIFT[i]=m-q；(q 是块在模式中出现的最右位置的位置字符的下标，下标从 1 开始计数)SHIFT[i]=0 时，HASH 和 PREFIX 表才起作用

HASH[i]包含了一个模式列表的指针，其中的模式最后 B 个字符的哈希值为 i。HASH 表入口指针不仅指向有此后缀的所有模式的列表，还指向子串前缀在 PREFIX 表中的哈希值。首先计算当前文本子串（从向左移动 m-B'个字符位置的子串开始）的前缀哈希值，依次来过滤那些后缀哈希值相同而前缀哈希值不同的模式 1) 计算这 B 个字符的 hash 值,得到 h;2) 查 SHIFT 表找到 SHIFT[h]:如果大于 0,则根据这个值向后移动文本相应的长度,并转到 1);否则继续;3) 计算当前指针往左的 m 个字符串长度为 B' 的前缀 hash 值;4)查 HASH 表,找到 HASH[h]的指针 p,遍历模式链表,找到前缀 hash 值也相同的模式串;再将文本串和模式串逐一比较,判断是否匹配.如果匹配,则输出匹配模式串,并将文本向后移动一位,转 1),直到文本结束.

8. 如何进一步优化 AC 算法？（内存空间优化）倒序自动机,哈希,状态合并,适当缩短模式串；因为失配的概率更高，所以试图提高 goto 函数的查找效率，引入位图 bitmap。位图的内存大小 | Σ | * | {p} |

9. 如何进一步优化 WM 算法？（模式长度敏感）

WM 算法中哈希冲突较大时,算法效率较低. 通过应用随机指纹模型降低冲突(就是 shift 表冲突的时候),即用多项式指纹函数的 FPRINT 表替代哈希的 PREFIX 表示

10. 针对千万量级的模式集，如何优化串匹配算法?(1)预处理:在模式匹配之前对模式集合进行预处理,如去重/排序/合并相似模式等.这样可以减少匹配过程中的计算量.(2)为模式集创建索引,实现快速定位(3)分布式计算:分割成多个多个子集,在不同计算节点上匹配(4)多核处理器并行处理(5)对模式集数据压缩,降低存储空间内存需求.

（针对特定的 IP 地址或 URL 地址构成的模式集，如何优化串匹配算法？）多机并行。改进传统模式匹配算法。针对 URL 长度普遍较长的特点，对已有的传统模式匹配算法进行改进，使其尽可能适合于处理 URL 查找；对 URL 进行编码处理，将 URL 模式串的查找转化为数字或其它易处理的字符，从而简化匹配过程。借助于高性能硬件处理器，提高 URL 的查找效率。针对 URL 块状结构，利用多层哈希设计实现一个 URL 过滤算法.针对 URL 地址,使用后缀数组和最大公共后缀自动机等算法,这些算法可以直接定位到可能的匹配位置,效率较高.将 IP 地址和 URL 地址按块划分,每个块使用独立的匹配模型建立索引,这样可以最大限度利用本地性原理,提高 Cache 命中的机率.

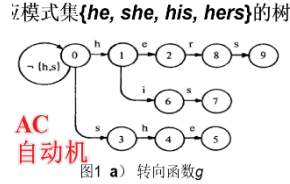


图1 a） 转向函数g

i	1	2	3	4	5	6	7	8	9
f(i)	0	0	0	1	2	0	3	0	3

图1 b） 失效函数f

i	output(i)
2	{he}
5	{she, he}
7	{his}
9	{hers}

图1 c） 输出函数output

2.4 TF-IDF

TF-IDF分数由两部分组成：第一部分是词语频率(Term Frequency)，第二部分是逆文档频率(Inverse Document Frequency)

$$TF(t) = \frac{\text{该词语在当前文档出现的次数}}{\text{当前文档中词语的总数}}$$

$$IDF(t) = log_e\left(\frac{\text{文档总数}}{\text{出现该词语的文档总数}}\right)$$

第四章 文本分类

1. 文本的量化表示？文本分类系统的主要流程？

词嵌入：向量空间模型(VSM)，VSM 中，每一文本 d 都被映像为由一组规范化正交词条向量所生成的向量空间中的一个点，将文本看作相互独立的词条组 d=(w1,w2,… wn)，对于每一词条 ti 都根据其在文本中的重要程度赋以一定的权值 ai。One-hot 向量/Bags of words/N-gram(相邻 n 个字成为一个元素的词袋)/TF-IDF

流程：预处理，比如分词，去掉停用词；文本表示及特征选择；分类器构造；分类器根据文本的特征进行分类；分类结果的评价。

2. 主要的分类方法有哪些？比较各自的优缺点。

(1)基于规则的分类器特点：优点-适合于规模小的数据集、具有良好的表达能力、易于构造、分类效率高、与决策树的性能相当;缺点-容易出现过拟合,分类规则过于复杂,对于大规模数据集,规则的生成和存储需要大量计算和存储资源(2)决策树分类算法(C4.5):优点-生成的决策树易于理解和解释,可以处理高维数据和非线性关系,可以处理缺失值和噪声数据,可以通过剪枝避免过拟合.缺点-容易出现过拟合,生成的决策树过于复杂;对于大规模数据集,决策树的生成和存储需要大量计算和存储资源,分类性能不如 SVM.(3)朴素贝叶斯分类器:优点-计算简单,速度快,适用于大规模数据集;可以处理高维数据和缺失值;对于噪声数据具有较好的鲁棒性;可以用于文本分类和推荐系统等应用.缺点-假设属性之间相互独立,可能会影响分类精度;对于连续属性,需要进行离散化处理.(4)SVM 分类器:优点-对于高维数据和非线性关系具有较好的分类效果;可以通过核函数扩展到非线性分类问题;可以处理少量的数据样本.缺点-对于大规模数据集需要较长的训练时间和较大的存储空间;对于噪声和缺失数据比较敏感;对于选择合适的核函数需要一定的专业知识和经验.

3. 基于规则归纳的分类方法？能够针对实际数据给出求解步骤.使用一系列“ if…then… ”对数据集分类规则. P.S.[1]如 R1 覆盖 hawk => Bird.[2]规则覆盖度:满足规则条件记录的百分比;规则的正确性:满足规则条件记录中,也满足规则结论的记录的百分比.[3]冲突解决:规模序-要求最严格(多)的规则优先,基于类的序-类的频繁性或错分代价,基于规则的序(决策表):优先级列表.有序规则集[4]有序规则集:满足多条,用排在前面的规则;规则都不满足使用默认类别

4. 基于决策树的分类方法（C4.5）？能够针对实际数据给出求解步骤(见下图)。**[1]**Info(D)= -9/14*1og2(9/14-5/14*log2(5/14)=0.940**[2]**Info(*tianqi*)=5/14*[-2/5*log2(2/5)-3/5*log2(3/5)]+4/14*[-4/4*log2(4/4)]+5/14*[-3/5*log2(3/5)-2/5*log2(2/5)]=0.694**[3]**Gain(*tianqi*)=0.940-0.694=0.246**[4]**H(*tianqi*)=-5/14*log2(5/14)-5/14*log2(5/14)-4/14*log2(4/14)**[5]**IGR(天气)=Gain(天气)/H(天气)=0.246/1.577=0.155 结果是天气的信息增益率 IGR 最大,选它

天气	温度	湿度	风速	活动
晴	寒冷	正常	弱	爬行
晴	适中	正常	强	触球
晴	炎热	高	弱	取消
晴	炎热	高	强	取消
晴	适中	高	弱	取消
阴	炎热	高	弱	爬行
阴	寒冷	正常	强	爬行
阴	适中	高	强	爬行
阴	炎热	正常	弱	爬行
雨	适中	高	弱	爬行
雨	寒冷	正常	弱	爬行
雨	适中	正常	强	爬行
雨	寒冷	正常	强	取消
雨	适中	高	强	取消

```
while（当前节点“不纯”）
  1. 计算当前节点的类别熵 Info(D)（以类别取值计算）
  2. 计算当前节点的属性熵 Info(Ai)（按照属性取值下的类别取值计算）
  3. 计算各个属性的信息增益 Gain(Ai) = Info(D) - Info(Ai)
  4. 计算各个属性的分类信息度量 H(Ai)（按照属性取值计算）
  5. 计算各个属性的信息增益率 IGR = Gain(Ai) / H(Ai)
end while
当前节点设置为叶子节点
```

5. 贝叶斯分类方法？能够针对实际数据给出求解步骤。

6. kNN 分类方法？能够针对实际数据给出求解步骤。考虑在全部训练样本中选出与待分类样本最相似(欧氏距离)的 K 个样本,将待分类样本标记为其 K 个邻居中所属最多的那个类别中

7. SVM 分类算法?一种监督学习方法,对数据进行二元分类的广义线性分类器,其决策边界是对学习样本求解的最大边距超平面,使用损失函数计算经验风险并加入了正则化项以优化结构风险,是一个具有稀疏性和稳健性的分类器,支持向量机将向量映射到一个更高维的空间里,在这个空间里建立有一个最大间隔超平面.在分开数据的超平面的两边建有两个互相平行的超平面,分隔超平面使两个平行超平面的距离最大化

8. 在样本集频繁更新的情况下，哪种分类方法更适合？用朴素贝叶斯或 kNN，因为他们不需要提前利用训练集训练模型。

9. 主要的聚类方法有哪些？各有什么特点？k-means，每个簇用该簇中对象的平均值来表示 k-medoids，每个簇用接近聚类中心的一个对象来表示(k 就是待分成的类的数量)

基于密度的方法：只要临近区域的密度超过某个阈值，就继续聚类。避免仅生成球状聚类。（DBSCAN，OPTICS，DENCLUE）

基于层次的方法：层次的方法对给定数据集合进行层次的分解。根据层次的分解如何形成，层次的方法可以被分为凝聚或分裂方法。 （Chameleon ，CURE，BIRCH）

基于网格的方法：基于网格的方法把对象空间量化为有限数目的单元，所有的聚类操作都在这个量化的空间上进行。这种方法的主要优点是它的处理速度很快。（STING，CLIQUE，WaveCluster）

基于模型的方法：为每个簇假设一个模型，发现数据对模型的最好匹配。（COBWEB，CLASSIT，AutoClass）

10. K-Means 聚类方法?

11. 层次聚类方法？基于分裂和聚合的聚类方法主要异同？

第五章 信息内容安全管理

1. C/S 与 P2P 发布方式各自的特点？[1]C/S:服务集中型,需要服务器稳定性高,压力大,服务器故障造成全局故障.[2]P2P:服务分散型;成本优势:带宽要求比较低;健壮性:天生具有耐攻击\高容错优点;高性价比:有效利用互联网络中散布的大量普通节点,将计算任务或者存储资料分布到所有节点上.

2. 舆情引导的基本原则？沟通时机,及时响应;沟通目标,告知真相,澄清和辟谣;沟通方式,因人而异,考虑接受能力;沟通内容,喜闻乐见,避免说教过于单调;沟通频度,加大正面宣传的强度;言行一致,树立权威.

3. C/S 模式下，如何响应？追踪中心化的 server.内容审核技术;黑名单技术;数字水印技术:在违规内容中嵌入特定的标识信息,用于识别和追踪该内容的来源和传播路径； 信源（发布端检测、行政管理）信宿（接收端防护、绿色上网）通道：旁路监测/网关监测、串接入网/防火墙、静态路由策略：IP 级、动态路由管理：连接级

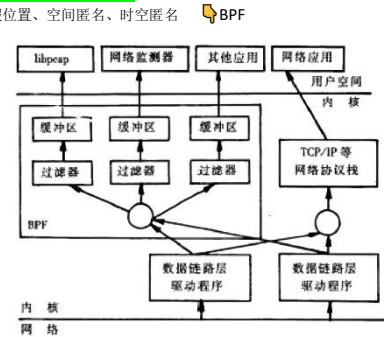
4. P2P 网络的典型体系结构？完全分布式的 P2P 网络、基于目录服务器的 P2P 网络、层次 P2P 网络

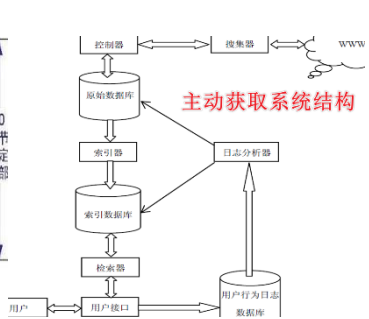
5. P2P 模式下，如何响应？基于用户来源交换的控管技术:每个节点都会维护一个信誉度和可信度值,这些值一方面可以用来评估该节点在网络中的贡献,另一方面也可以用来判断该节点是否可信;基于分布式哈希表的隔离控管技术:通过控制哈希函数和哈希值的范围，来限制恶意节点的访问和影响范围，从而实现隔离控制.(索引污染,资源占用,数据欺骗,数据块污染,其他方法)

6. 主要的隐私保护方法？数据加密:以一定的规则对数据进行变换，只有授权用户才能还原真实数据;同态加密:对经过同态加密的数据进行处理得到一个输出，将这一输出进行解密，其结果与用同一方法处理来加密的原始数据得到的输出结果是一样的;密文检索:它能够在信息资源加密存储的前提下，通过对其构建密文全文索引，提供高效安全的检索方法;数据分割:通过将准标识符属性集进行垂直分割，割裂其中的关联关系，使得以对数据做笛卡尔积的方式还原数据的精度不高 1/k;分权机制:在相关实体间引入分权机制防止系统中 “超级”节点的形成;匿名： 在数据发布时，隐匿用户 ID 的指示信息（如身份证号、姓名等）已经成为业界的惯例;假名、基于众包的轨迹隐私保护、数据泛化、数据加噪、位置模糊化、隐私信息检索 PIR:将用户的查询请求通过一个矩阵变换构造出 N-1 个与其不可区分的伪查询，使得攻击者无法精确的知道用户的真实意图

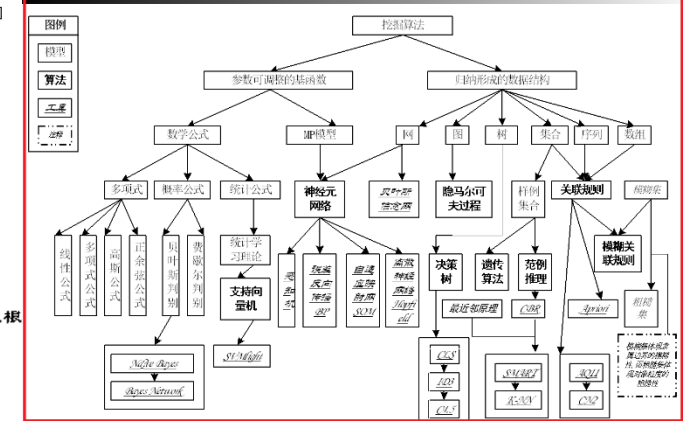
7. 位置隐私保护方法？

假位置、空间匿名、时空匿名 BPF





挖掘算法发展脉络图



$$P(buys\_computer = \text{“yes”}) = 9/14 = 0.643$$
$$P(buys\_computer = \text{“no”}) = 5/14 = 0.357$$

1， 2，我们计算下面的条件概率：

$$P(age = \text{“<30”} \mid buys\_computer = \text{“yes”}) = 2/9 = 0.222$$
$$P(age = \text{“<30”} \mid buys\_computer = \text{“no”}) = 3/5 = 0.600$$
$$P(income = \text{“medium”} \mid buys\_computer = \text{“yes”}) = 4/9 = 0.444$$
$$P(income = \text{“medium”} \mid buys\_computer = \text{“no”}) = 2/5 = 0.400$$
$$P(student = \text{“yes”} \mid buys\_computer = \text{“yes”}) = 6/9 = 0.667$$
$$P(student = \text{“yes”} \mid buys\_computer = \text{“no”}) = 1/5 = 0.200$$
$$P(credit\_rating = \text{“fair”} \mid buys\_computer = \text{“yes”}) = 6/9 = 0.667$$
$$P(credit\_rating = \text{“fair”} \mid buys\_computer = \text{“no”}) = 2/5 = 0.400$$

$$P(X \mid buys\_computer = \text{“yes”}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$$
$$P(X \mid buys\_computer = \text{“no”}) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019$$

$$P(X \mid buys\_computer = \text{“yes”})\, P(buys\_computer = \text{“yes”}) = 0.044 \times 0.643 = 0.028$$
$$P(X \mid buys\_computer = \text{“no”})\, P(buys\_computer = \text{“no”}) = 0.019 \times 0.357 = 0.007$$

叶斯分类预测 buys-computer = “ yes ”。

age=60, income=medium, student=no ， credit rating=fair

RID	age	income	student	credit_rating	Class:buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	Low	yes	fair	yes
6	senior	Low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no