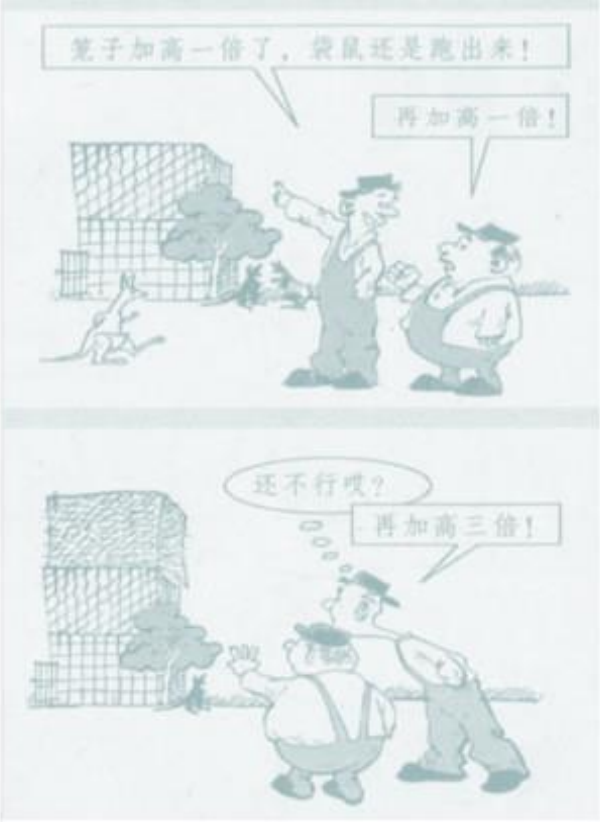
身份认证与访问控制技术

Identity Authentication and Access Control Techniques

目 录

入口安全问题



袋鼠从笼子里跑出来了，管理员决定加高笼子

第一天：加高10公尺。

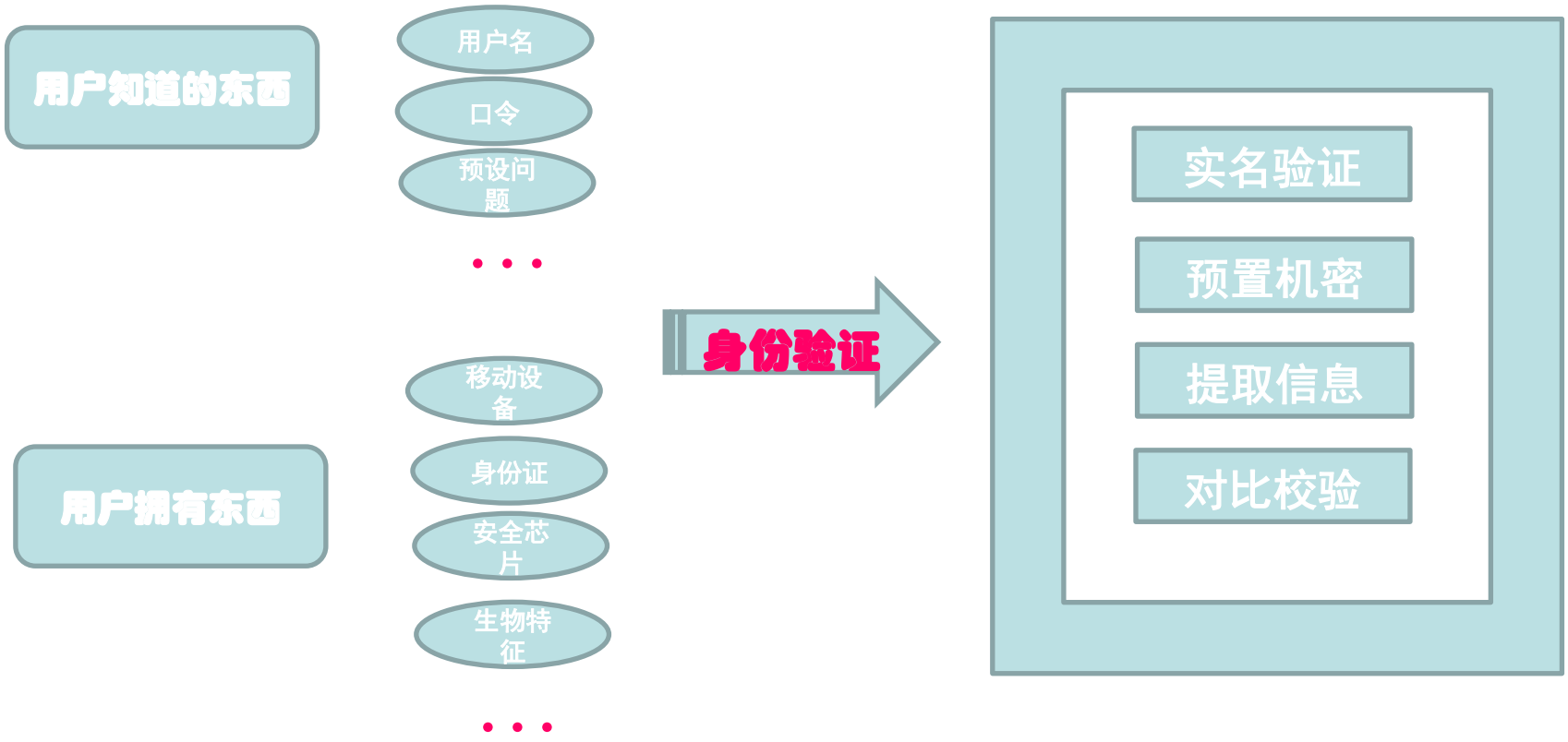
第二天：加高了20公尺。

第三天：加高到60公尺。

长颈鹿：“会不会再有人继续加高你们的笼子？”

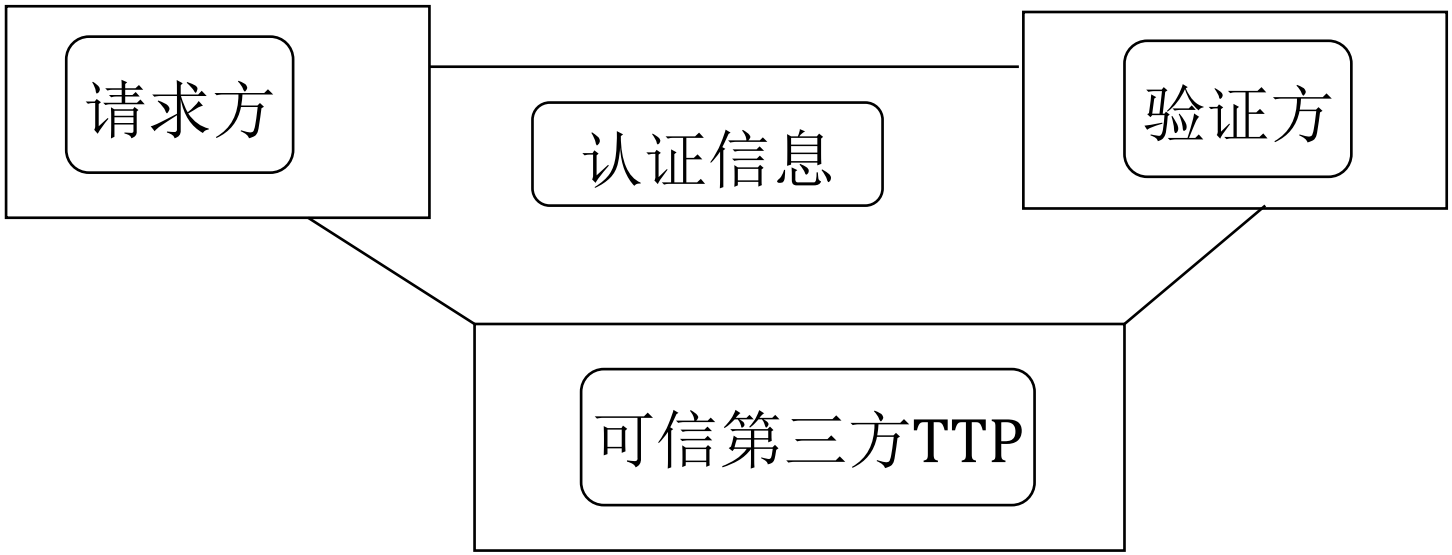
袋鼠：“很难说”，“如果他们再忘记关门的话。”

什么是身份认证?


用户知道的东西
用户名
口令
预设问题
• • •
用户拥有东西
移动设备
身份证
安全芯片
生物特征
• • •
身份验证
实名验证
预置机密
提取信息
对比校验

身份认证的基本模型

- **请求方（Claimant）**
- **验证方（Verifier）**
- **认证信息（Authentication Information）**
- **可信第三方TTP(Trusted Third Party)**



基于口令的身份认证技术（Password-based Authentication ）

- **口令的要求：**
  - 足够的长度，并包含各种不同的字符数;
  - 和ID无关;
  - 包含特殊的字符;
  - 大小写;
  - 不容易被猜测到;
  - 定期更改其口令;
  - 使用字典式攻击的工具找出比较脆弱的口令:
    - 网络管理员使用的工具：口令检验器
    - 攻击者破获口令使用的工具：口令破译器

口令泄露事件


中国版25个“弱密码”
*本项统计基于国内流行的密码字典软件破解列表
*标红密码同时也是国外网民常用的“弱密码”
简单数字组合
000000
111111
11111111
112233
123123
123321
123456
12345678
654321
顺序字符组合
abcdef
abcabc
abc123
a1b2c3
aaa111
临近字符组合
123qwe
qwerty
qweasd
特殊含义组合
admin
password
p@ssword
passwd
iloveyou
5201314
华军软件园
ONLINEDOWN
www.newhua.com
KaFan.CN
计算机安全
大气谦和


12306用户数据遭泄露
专家提醒尽快修改密码
O.CN 都市圈
Mapping your life

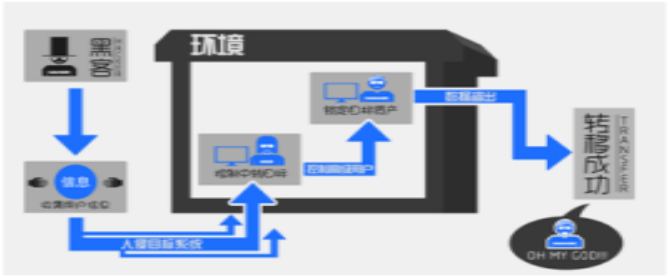

天涯社区
社会
深度
公益
文学
北京
天津
上海
南京
天涯聚焦
focus.tianya.cn
让


知名网站
IP
用户资料

口令攻击方式

网络监听

数据库入侵

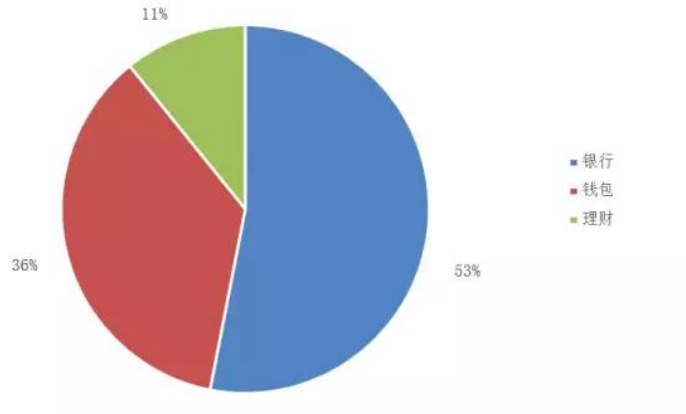


编程漏洞

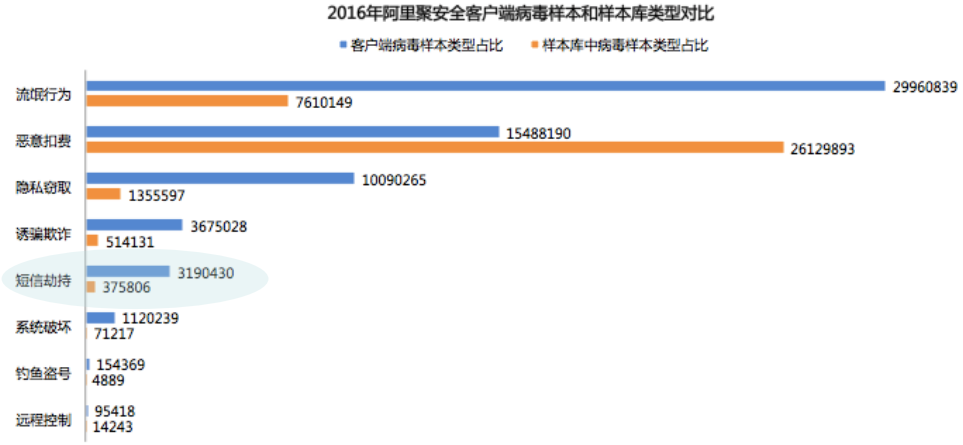
暴力攻击

短信认证问题

金融行业选取银行、钱包和理财3个子分类,分别选取10个热门应用进行分析,共发现仿冒应用407个。
银行类仿冒应用占53%,钱包类 仿冒应用占36%,理财类仿冒应用占11%。



2016年金融行业仿冒应用分布情况

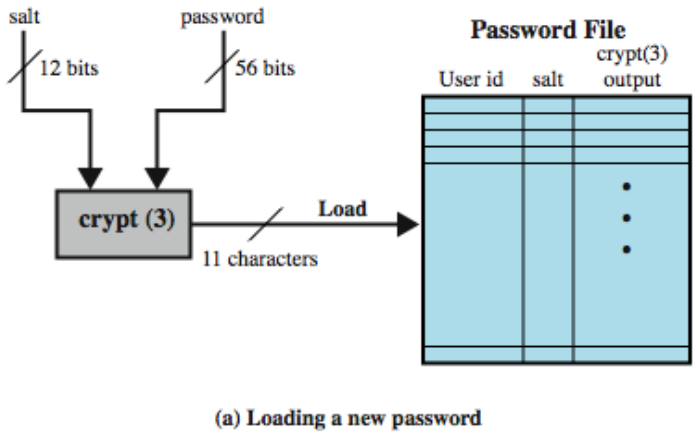


金融行业银行类仿冒居多,某银行仿冒应用全部具有短信劫持行为

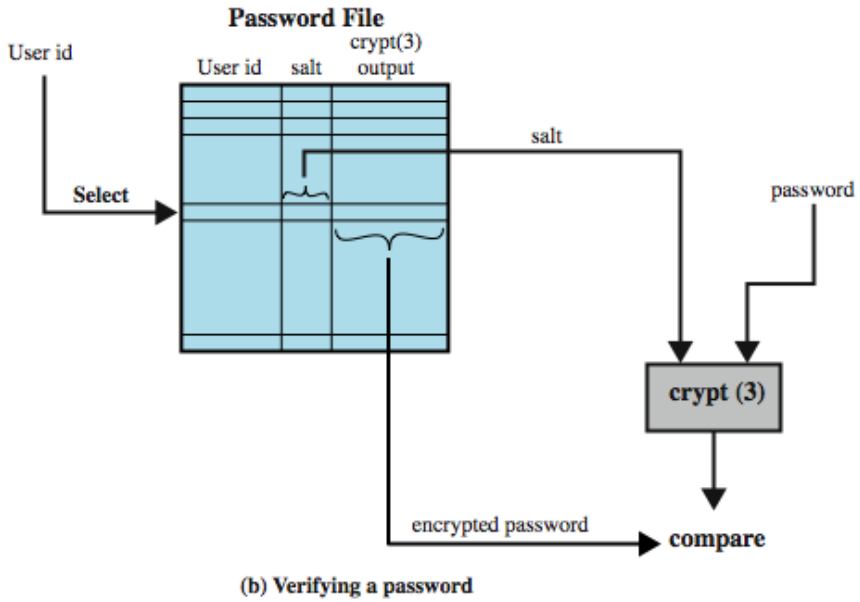
某银行共发现30个仿冒应用,全部具有短信劫持行为,感染设备量为33863台,感染用户主要分布在广东、北京和江苏等省份。

带杂凑函数(hash)和盐(salt)的口令认证——UNIX操作系统

- **操作系统登录认证：**
  - 将用户ID以及它们对应的口令，生成一个口令文件，例如Shadow影子口令文件
  - 当用户登录系统时，系统会将该用户输入的口令与口令文件中的口令信息进行比对，匹配成功才会允许用户进入系统
- **密码杂凑函数/散列函数(Hash)的作用：**
  - 防止口令文件直接泄露用户的口令
  - MD5/SHA等Hash算法，早期也用DES对称加密算法
- **加盐(salt)的作用：**
  - 为每个用户选择一个不同的随机数
  - 当不同用户采用相同口令时，因为“盐值”不同，所以口令文件中的杂凑值不同
  - 当同一个用户在不同系统中使用相同的口令时，在各个口令文件中的杂凑值不同
  - 增加字典攻击的难度，彩虹表攻击(rainbow table)



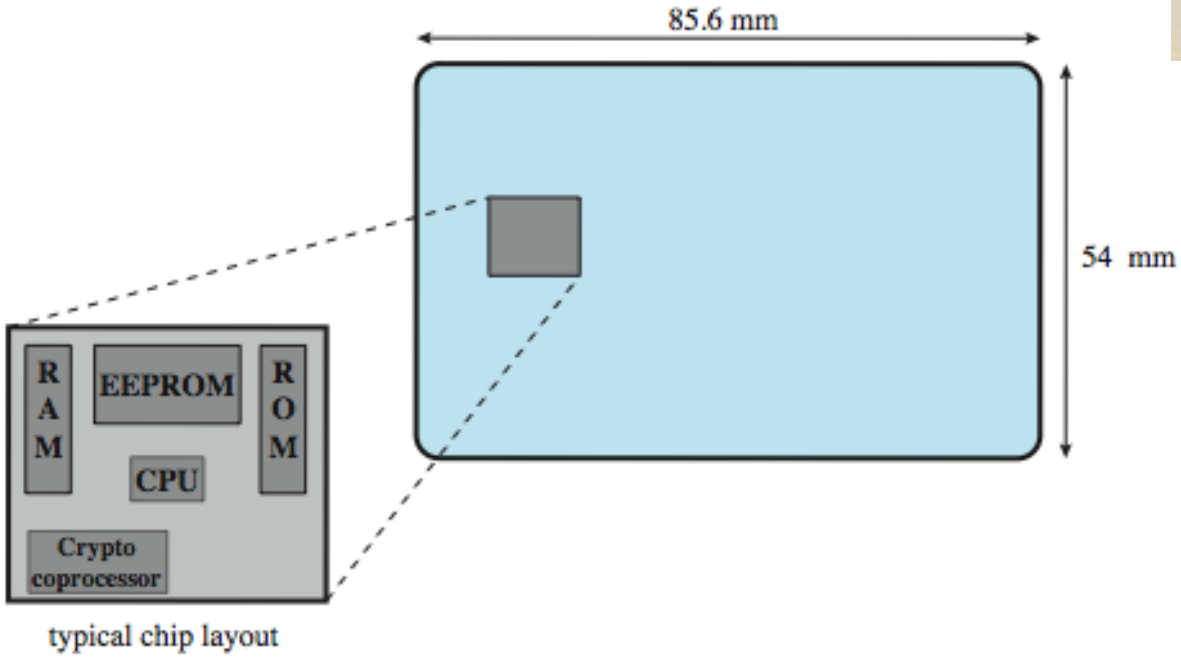
(a) Loading a new password



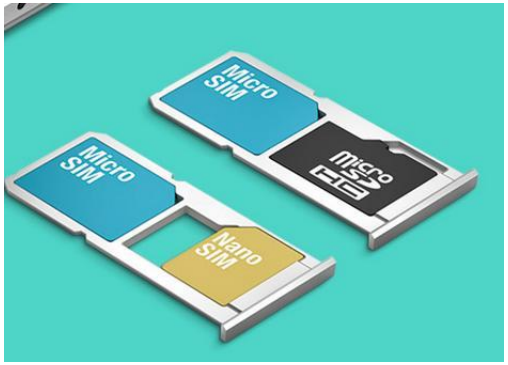
(b) Verifying a password

智能IC卡/智能密码钥匙

- 具有处理器、存储器和I/O接口
- 具备有线(串口/USB/7816等接口)或无线读卡器
- 可以带有密码协处理器crypto co-processor
- 存储器包括ROM, EEPROM, RAM等
- 需要认证持卡人的身份



typical chip layout



远程身份认证技术（Remote Identity Authentication）

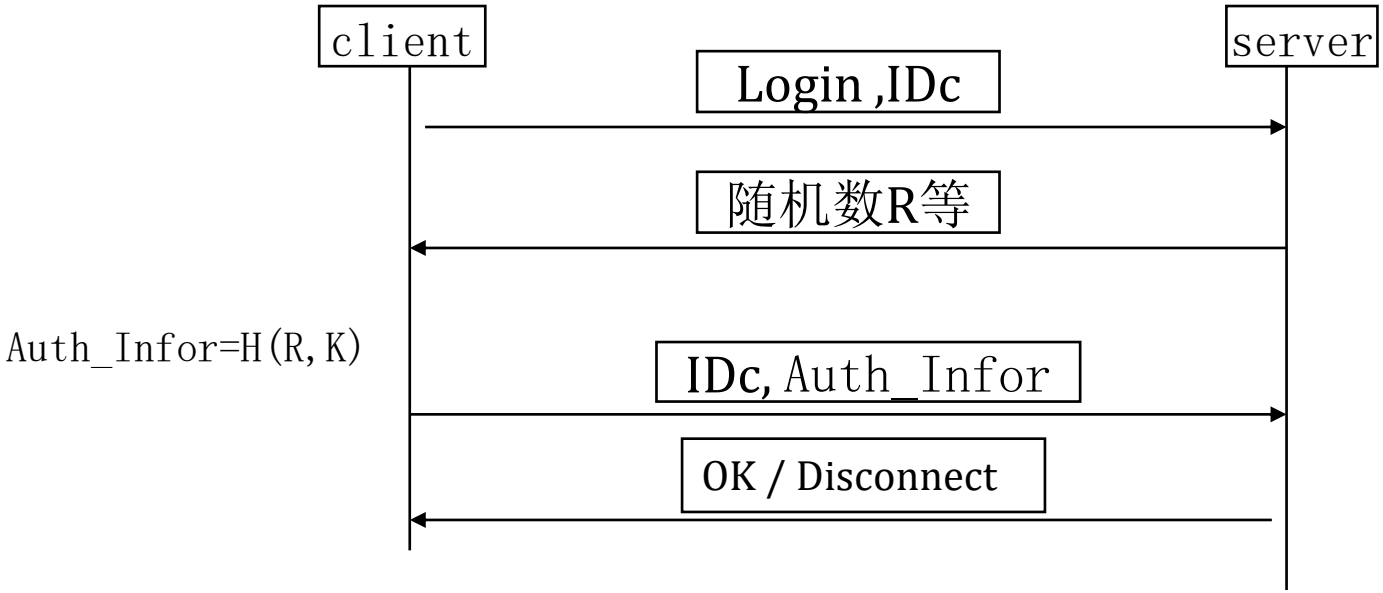
- 远程身份认证面临的风险和对策：

风险和攻击	解决方案
窃听	采用数据加密、密码杂凑算法等方法，对口令等敏感消息进行处理
重放攻击	采用随机数(Nonce)、时间戳等方法，挑战-应答协议(challenge-response protocol)
伪造消息	采用数字签名、消息鉴别码等方法
中间人攻击	采用数字签名等方法

- 远程身份认证面临的风险和对策：

- 远程口令认证协议
- 远程令牌认证协议
- 远程静态生物认证协议
- 远程动态生物认证协议
- 基于公钥密码算法的远程身份认证协议

挑战-应答协议(challenge-response protocol)



- Auth_Infor的计算方法可以采用Hash算法，对称密钥算法，公开密钥算法等
- Auth_Infor的计算输入包括随机数R、密钥K、口令、动态令牌、生物特征等

双因子认证的问题

企业采用各种身份验证方式来确保移动安全访问

- 动态口令
- USB Key
- 蓝牙Key
- 动态短信验证码
- 生物识别
- 金融IC卡
- OTP令牌
- 证书
- ……

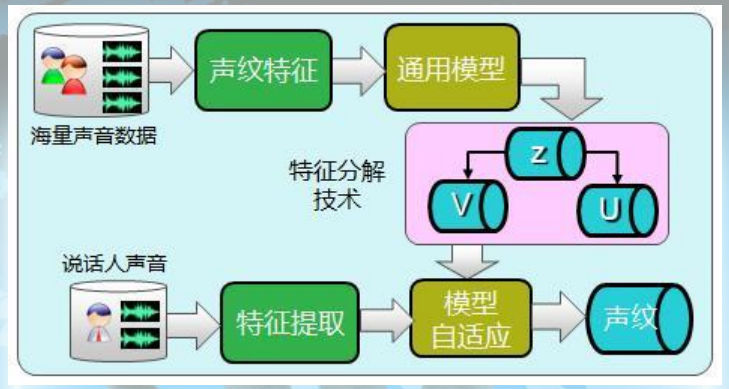
复杂的身份验证方式

- 增加了用户疲劳
- 降低了客户满意度
- 减低企业服务竞争力

生物识别技术

指纹识别：千元机标配，Android6.0开始开放指纹API。使用最为简单方便。

虹膜识别：最可靠的生物识别技术，部分高端手机已经适配虹膜摄像头。

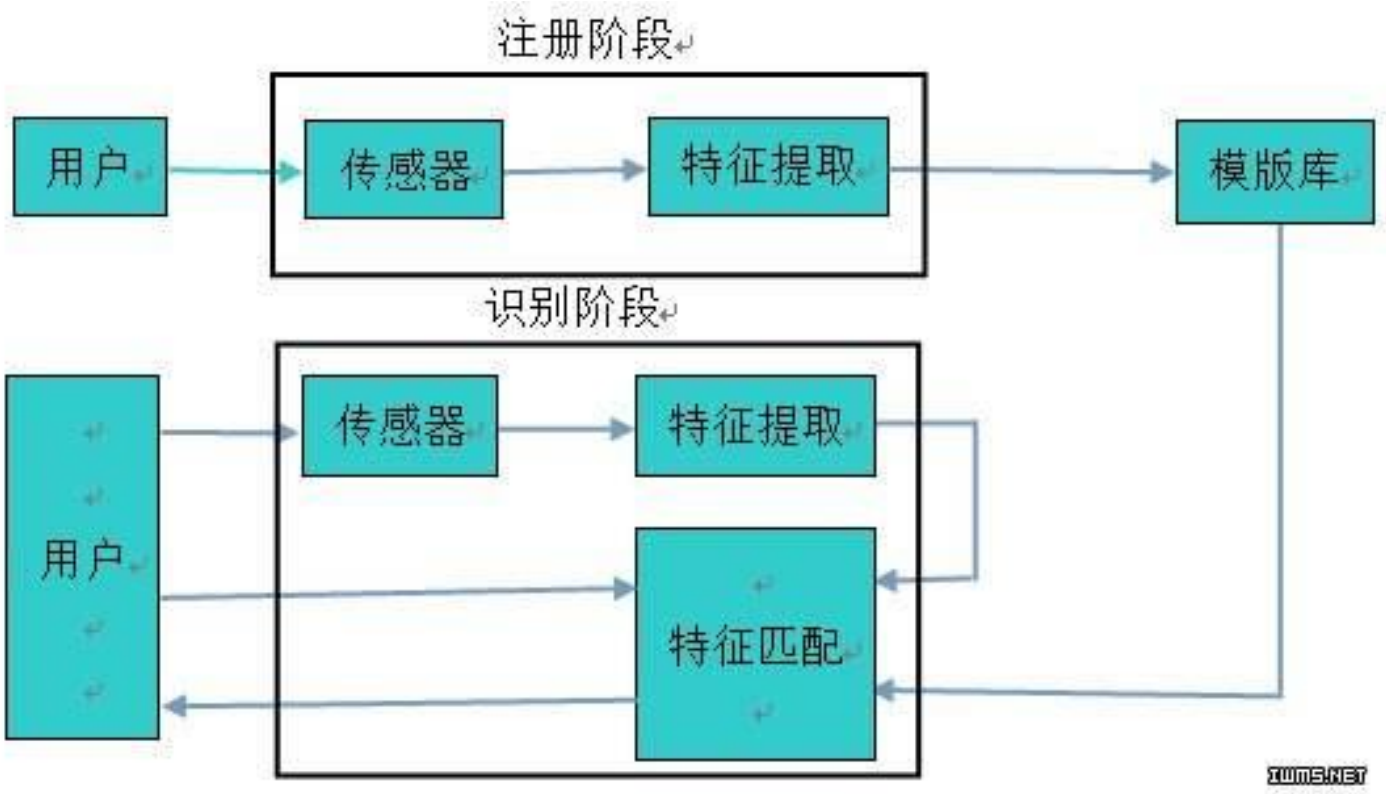


面部识别：精确度日益提高。

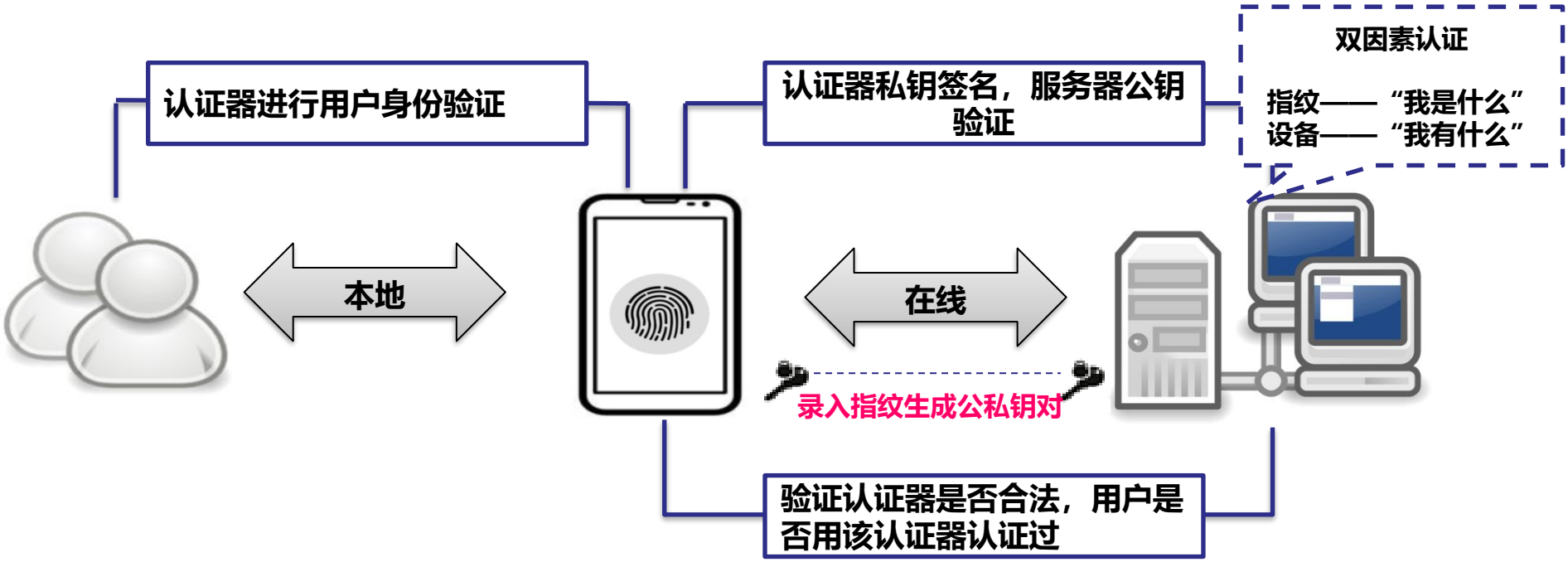
声纹识别：成熟易用，微信已经开始启用声纹识别登录。

静脉识别、掌型识别、眼纹……

生物识别产品通用流程



FIDO身份认证


双因素认证
指纹——“我是什么”
设备——“我有什么”
认证器进行用户身份验证
认证器私钥签名，服务器公钥验证
本地
在线
录入指纹生成公私钥对
验证认证器是否合法，用户是否用该认证器认证过

FIDO、IFAA、TUSI

认证平台	FIDO（国民认证）	IFAA	TUSI（领御守护计划）
成立时间	FIDO：2012年 国民认证：2015年	2015年6月	2016年4月
性质	FIDO：国际非营利性组织 国民认证：联想创投成立之后成功孵化的子公司	阿里巴巴牵头的联盟组织	腾讯移动事业部项目
官网	FIDO：https://fidoalliance.org/ 国民认证： http://www.noknoklabs.cn/	http://ifaa.alipay.com/index.htm	TUSI：http://tusi.qq.com/ Qkey：http://qkey.qq.com/
认证形式	指纹为主，人脸和虹膜即将面世	人脸为主，虹膜、指纹、可穿戴设备	可穿戴设备
成员数量	FIDO：截至2016年7月已经有252家企业、组织和政府机构参与。24家中国企业。（官方稿件透露）	联盟会员单位超过60家。（2016年7月官方公布数据）	参与企业目前已经达到了20多家。（MPSC 2016中国移动支付安全大会上腾讯演讲人透露）
相关数据	150多种设备已获FIDO的官方认证。	22余家手机厂商 100多手机型号 7000万用户。	无
落地情况	国内：翼支付、京东钱包、百度钱包、微众银行等产品支持FIDO	覆盖7000万用户，累计人脸识别5亿次。	Qkey未上市
高层情况	国民认证：国民认证总经理兼FIDO中国工作组主席柴海新	互联网金融身份认证联盟(IFAA)理事长、蚂蚁金服安全产品技术部总监冯春培	腾讯移动事业群副总裁吴宇（4月发布会项目最高层）
代表企业	联想集团、阿里巴巴集团、飞天诚信和台湾神盾、汇顶科技、华为终端有限公司、沃通电子认证、握奇、天地融、中国信通院、中国电子标准化院。	三星、华为、中兴、OPPO、酷派等手机厂商，以及高通、握奇、展讯等芯片厂商、安全厂商、算法厂商、检测机构等产业链。公安第一研究所曾在成立会中为其展台。	腾讯、握奇、飞天诚信、天地融、台湾神盾、复旦微电子、中钞信用卡产业发展有限公司等

认证技术的发展历程

一代

原始方式:

用户名 useradmin

密 码 •••••

1990

2000

二代

口令升级版：验证码、短信随机码、动态口令、
手势密码、刮刮卡。

2002

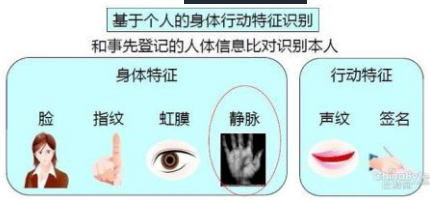
三代

双因子认证：U盾、指纹key、智能卡
液晶key、SD卡、

2015

四代

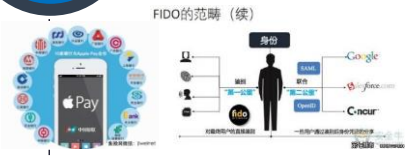
生物特征、行为特征：
刷脸、声纹、指纹、虹膜、步态等



2016

五代

多因素认证：
生物特征+密码技术
FIDO、IFAA、TUSI



未来

态势感知+零信任

目 录

第二章 安全存储与访问控制技术

2.1 早期访问控制技术

2.2 基于数据分析的访问控制技术

2.3 基于密码学的访问控制技术

2.1 早期访问控制技术

2.1.1 基本概念

2.1.2 访问控制模型

2.1.3 局限性总结

2.1.1 基本概念

访问控制(Access Control)：确保数据等资产只能经过授权的用户才能访问、使用和修改。

访问控制策略(Policies)：是对系统中用户访问资源行为的安全约束需求的具体描述。

访问控制模型(Model)：是对访问控制策略的抽象、简化和规范。

早期的访问控制技术都是建立在**可信引用监控机**基础上的，1972年由Anderson提出，它能够对系统中的主体和客体之间的授权访问关系进行监控。

一般来说，这类访问控制技术都涉及如下的概念：

① **主体：**能够发起对资源的访问请求的主动实体，通常为系统的用户或进程。
② **客体：**能够被操作的实体，通常是各类系统和数据资源。
③ **操作：**主体对客体的读、写等动作行为。
④ **访问权限：**客体及对其的操作形成的二元组<操作, 客体>。
⑤ **访问控制策略：**对系统中主体访问客体的约束需求描述。
⑥ **访问（引用）授权：**访问控制系统按照访问控制策略进行访问权限的赋予。

2.1.1 基本概念

⑦ **引用监控机（Reference Monitor，RM）**：指系统中监控主体和客体之间授权访问关系的部件。

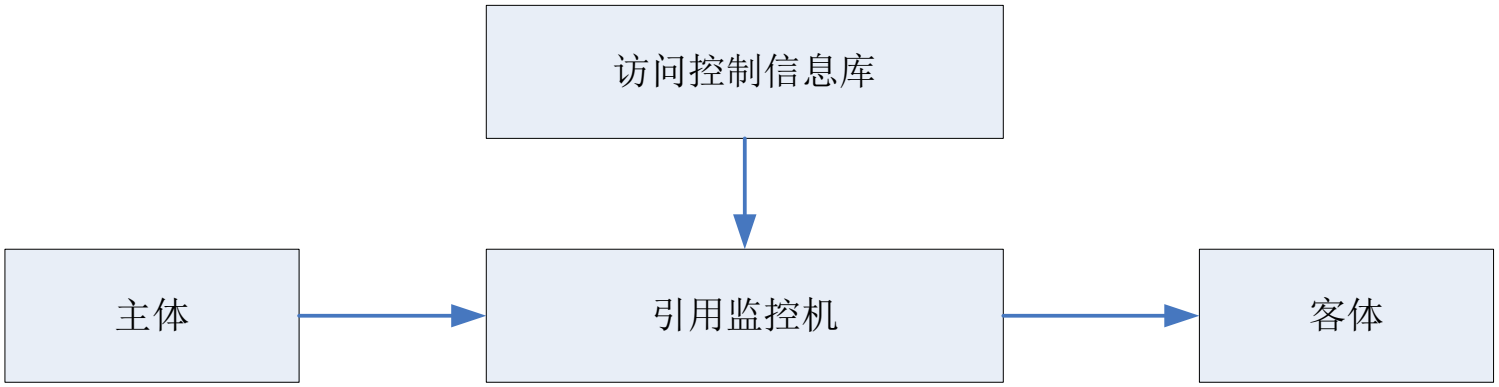


图2-1 引用监控机模型

⑧ **引用验证机制（Reference Validation Mechanism，RVM）**：是RM的软硬件实现。引用验证机制RVM是真实系统中访问控制能够被可信实施的基础。它必须满足如下三个属性：

(1) 具有自我保护能力；

(2) 总是处于活跃状态；

(3) 必须设计得足够小，以便于分析和测试。

- **访问控制模型的发展历史：**

  - **自主访问控制(Discretionary Access Control)和强制访问控制(Mandatory Access Control)：**在**20世纪70年代**，大型资源共享系统出现在政府和企业中。为了应对系统中的资源安全共享需求，访问控制矩阵等**自主访问控制模型**和BLP、Biba等**强制访问控制模型**被提出，并得到了广泛应用。
  - **基于角色的访问控制(Role-based Access Control-RBAC)：**在**20世纪80年代末到90年代初**，在商业系统按照工作或职位来进行访问权限的管理更加方便。因此，**基于角色的访问控制模型**被提出，并发展成为迄今为止在企业或组织中应用**最为广泛的访问控制模型之一**。
  - **基于属性的访问控制(Attribute-based Access Control-ABAC)：**在**21世纪初期**，互联网技术使得用户对资源的访问处于开放环境。开放环境往往**无法预先获得主客体身份的全集**，且存在身份隐藏的需求。因此，**基于属性的访问控制**被提出，它通过安全属性来管理授权，而**不需要预先知道访问者身份**。

2.1.2 访问控制模型

1、自主访问控制模型：客体的属主决定主体对客体的访问权限。

自主访问控制模型可以被表述为**(S,O,A)三元组**。

其中，**Subject表示主体集合**，**Oblect表示客体集合**。

Access matrix表示访问矩阵，$A(s_i,o_j)$则表示主体s_i能够对客体o_j执行的操作权限。

	o_1	…	o_n	s_1	…	s_m
s_1	$A(s_1,o_1)$	…	$A(s_1,o_n)$	$A(s_1,s_1)$	…	$A(s_1,s_m)$
…	…	…	…	…	…	…
s_i	$A(s_i,o_1)$	…	$A(s_i,o_n)$	$A(s_i,s_1)$	…	$A(s_i,s_m)$
…	…	…	…	…	…	…
s_m	$A(s_m,o_1)$	…	$A(s_m,o_n)$	$A(s_m,s_1)$	…	$A(s_m,s_m)$

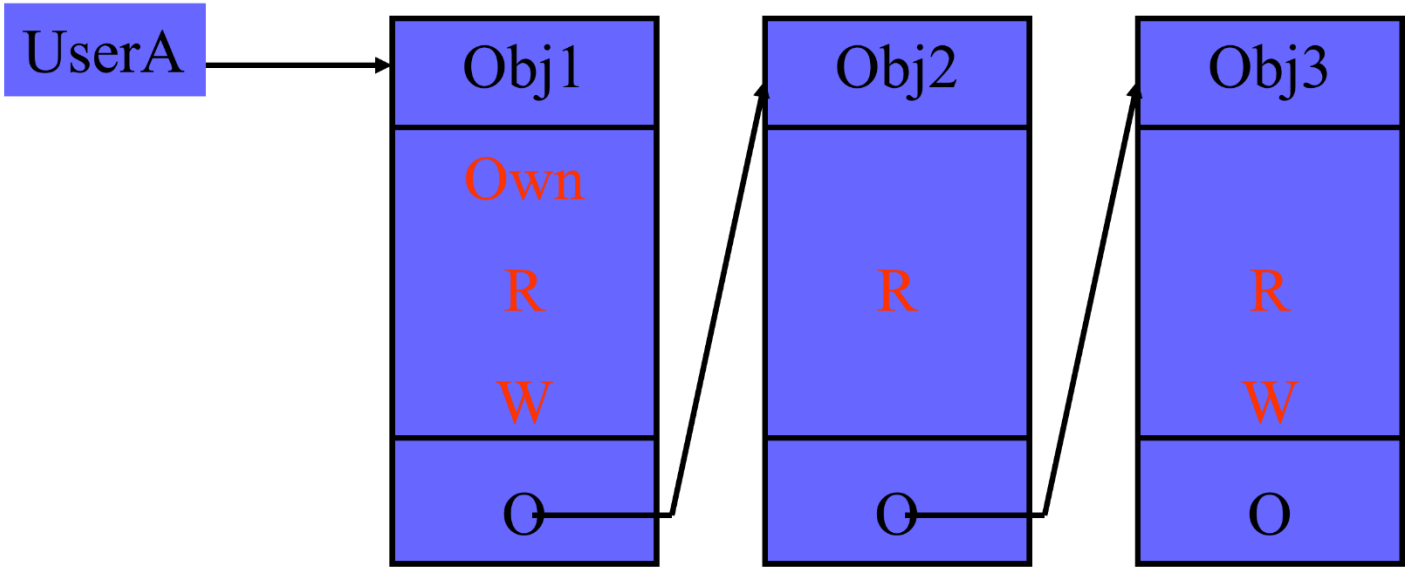
图2-2 访问矩阵

2.1.2 访问控制模型

自主访问控制模型：

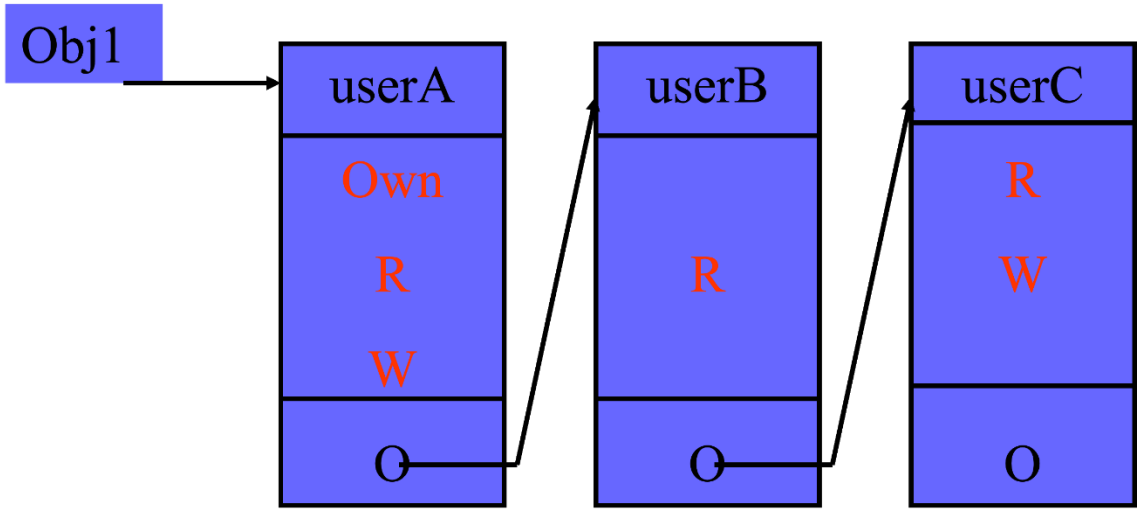
访问矩阵A在实际系统中主要有两种实现方式：

① **能力表（Capabilities List，CL）：基于主体的**自主访问控制实现，该表记录了每一个主体与一个权限集合的对应关系。权限集合中每个权限被表示为一个客体以及其上允许的操作集合的二元组。



2.1.2 访问控制模型

② **访问控制列表（Access Control List，ACL）：基于客体的**自主访问控制实现，该表记录了每一个客体与一个权限集合的对应关系。权限集合中的每个权限被表示为一个主体以及其能够进行的操作集合的二元组。



在大数据环境下，主体和客体数量巨大，无论哪种实现方式，**自主访问控制模型都将面临权限管理复杂度爆炸式增长的问题**。因此，直接采用自主访问控制模型是非常困难的。

2.1.2 访问控制模型

2、强制访问控制模型：其访问控制策略由安全管理员统一管理。

安全管理员为系统中每个主客体分配安全标记，然后依据主客体安全标记之间的支配关系来进行访问控制。由于安全标记之间的支配关系是满足偏序性质的，可以形成格结构，如下图所示。因此，强制访问控制模型又可**称为基于格的访问控制模型**。

最为经典的是**BLP模型(机密性)和Biba模型(完整性)**。

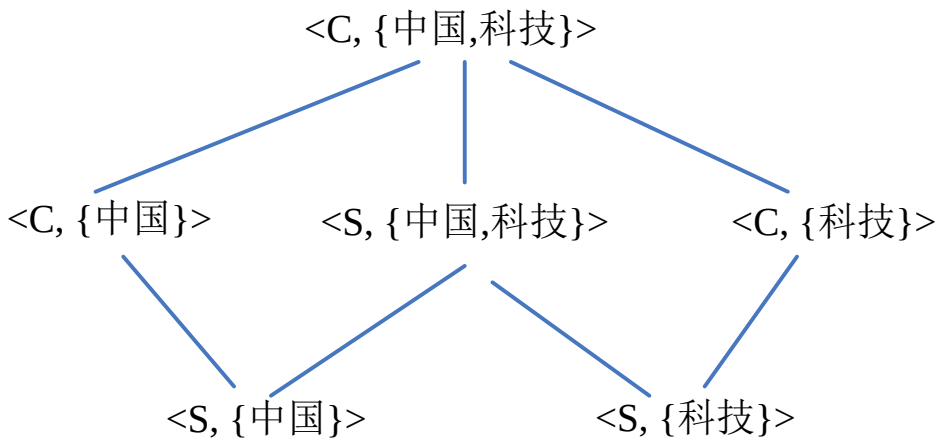
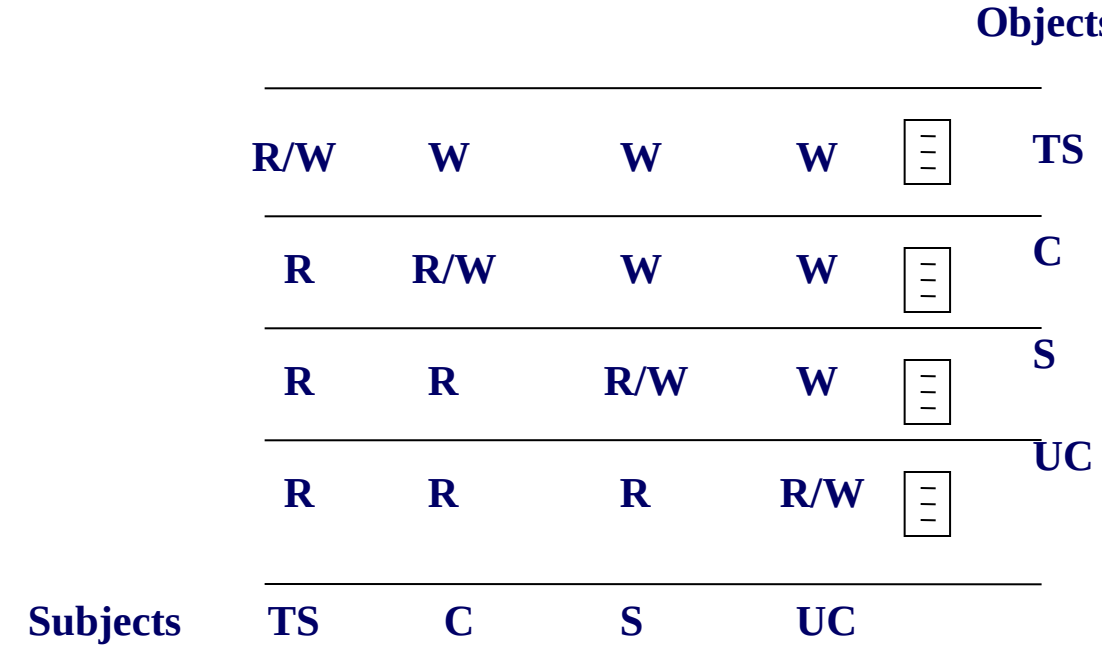


图2-3 安全标记之间的支配关系示意图

2.1.2 访问控制模型

强制访问控制模型：BLP模型被用于保护系统的**机密性**，防止信息的未授权泄漏。该模型的**核心规则是“不上读、不下写”**，即**低级别不能读取高级别的数据**，高级别不能修改低级别的数据，**保证数据只能从低级别往高级别流动**。



密级	英文
绝密TS	Top Secret
机密C	Confidential
秘密S	Secret
无密级UC	Unclassified

2.1.2 访问控制模型

① **安全级别Level：公开（UC）、秘密（S）、机密（C）、绝密（TS）**。它们之间的关系为UC≤S≤C≤TS。

② **范畴Category：**为一个类别信息构成的集合，例如｛中国，军事，科技｝。具有该范畴的主体能够访问那些以该范畴子集为范畴的客体。

③ **安全标记Label：**由安全级别和范畴构成的二元组<Level，Category>，例如<C,｛中国,科技｝>。

④ **支配关系dom：**安全标记A dom B，当且仅当$Level_A \geq Level_B$，$Category_A \supseteq Category_B$。

2.1.2 访问控制模型

强制访问控制模型：BLP模型中在为系统中每个保护范围内的主客体都分配了安全标记后，主体对客体的访问行为应满足如下两条安全属性：

① **简单安全属性：**主体S可以读客体O，当且仅当$Label_S$ dom $Label_O$，且S对O有自主型读访问权限。

② ***安全属性：**主体S可以写客体O，当且仅当$Label_O$ dom $Label_S$，且S对O具有自主型写权限。

从信息流角度看，上述两条读/写操作所应遵循的安全属性**阻止了信息从高安全级别流入低安全级别**，且使得**信息“仅被需要知悉的人所知悉”**，因此，能够有效地确保数据的机密性。

优点	缺点
1. 是一个最早地对多级安全策略进行描述的模型。 2. 是一个严格形式化的模型，并给出了形式化的证明。 3. 控制信息只能由低向高流动，能满足军事部门等一类对数据保密性要求特别高的机构的需求。	1. 上级对下级发文受到限制；部门之间信息的横向流动被禁止。 2. 只要信息由低向高流动即合法（高读低），不管工作是否有需求，这不符合最小特权原则。缺乏灵活、安全、细粒度的授权机制。 3. 低安全级的信息向高安全级流动，可能破坏高安全客体中数据完整性，被病毒和黑客利用。高级别的信息大多是由低级别的信息通过组装而成的，要解决推理控制的问题。

2.1.2 访问控制模型

强制访问控制模型：Biba模型是第一个关注**完整性**的访问控制模型，用于防止用户或应用程序等主体未授权地修改重要的数据或程序等客体。该模型可以看作是BLP模型的对偶。

① **完整性级别Level：**代表了主/客体的可信度。完整性级别高的主体比完整性级别低的主体在行为上具有更高的可靠性；完整性级别高的客体比完整性级别低的客体所承载的信息更加精确和可靠。

② **范畴Category：**若范畴$\text{Category}_A \supseteq \text{Category}_B$，则A能写入B；否则，A不能写入B。

③ **完整性标记Label：**由完整性级别和范畴构成的二元组<Level，Category>。

④ **支配关系dom：**完整性标记A dom B，当且仅当$\text{Level}_A \geq \text{Level}_B$，$\text{Category}_A \supseteq \text{Category}_B$。

2.1.2 访问控制模型

强制访问控制模型： Biba模型的严格完整性策略是BLP模型的对偶，也是不特别指明情况下所谓的Biba模型。它应满足如下安全属性。

① **完整性特性：** 主体S能够写入客体O，当且仅当$Label_S$ dom $Label_O$。
② **调用特性：** 主体S_1能够调用主体S_2，当且仅当$Label_{S1}$ dom $Label_{S2}$。
③ **简单完整性条件：** 主体S能够读取客体O，当且仅当$Label_O$ dom $Label_S$。

基于上述三条安全属性，**信息只能从高完整性级别的主客体流向低完整性级别的主客体**，从而有效避免了低完整性级别的主客体对高安全级别主客体的完整性的“污染”。

2.1.2 访问控制模型

强制访问控制模型：在大数据场景下，由安全管理员来进行强制访问控制的授权管理是具有挑战性的。

① 随着**主客体规模的急剧增长，安全标记的定义和管理将变得非常繁琐**；

② 来自**多个应用的用户主体和数据客体也将使得安全标记难以统一**。

2.1.2 访问控制模型

3、基于角色的访问控制：标准RBAC模型包括了RBAC0~3四个模型。

① **RBAC0模型（Core RBAC），**定义了**用户、角色、会话和访问权限**等要素。

② **RBAC1（Hierarchal RBAC）**在RBAC0的基础上**引入了角色继承**的概念。

③ **RBAC2（Constraint RBAC）**增加了**角色之间的约束条件**，例如互斥角色、最小权限等。

④ **RBAC3（Combines RBAC）是RBAC1和RBAC2的综合**，探讨了角色继承和约束之间的关系。

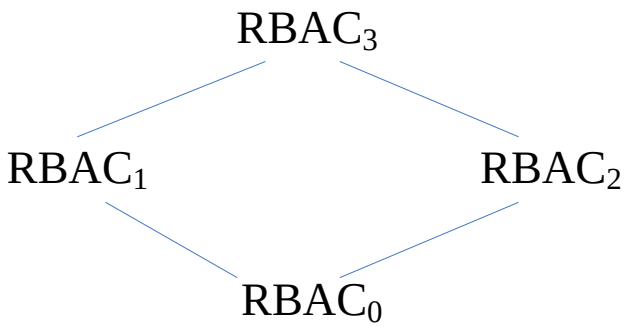


图2-4 标准RBAC模型框架

2.1.2 访问控制模型

基于角色的访问控制：CoreRBAC。

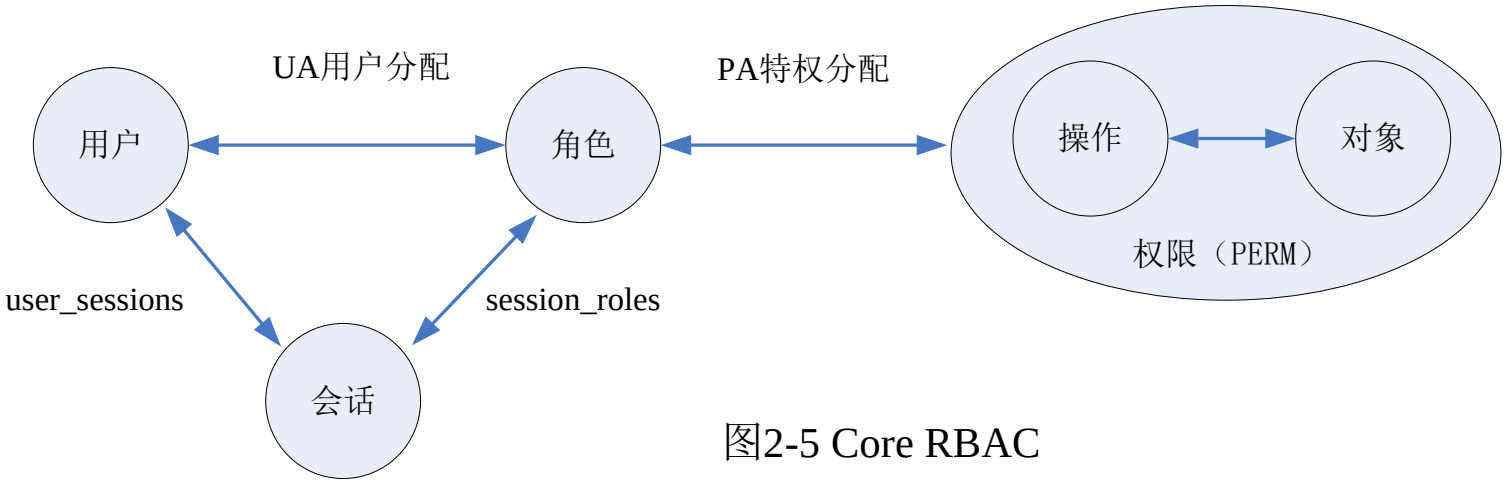


图2-5 Core RBAC

① **用户是访问控制的主体**，可以发起访问操作请求。
② **对象是访问控制的客体**，指系统中受访问控制机制保护的资源。
③ **操作是**指对象上能够**被执行的一组访问操作**。
④ **权限是**指对象及其上**指定的一组操作**。
⑤ **角色是权限分配的载体**，是一组有意义的权限集合。
⑥ **会话**用于**维护用户和角色之间的动态映射关系**。

2.1.2 访问控制模型

基于角色的访问控制：CoreRBAC。

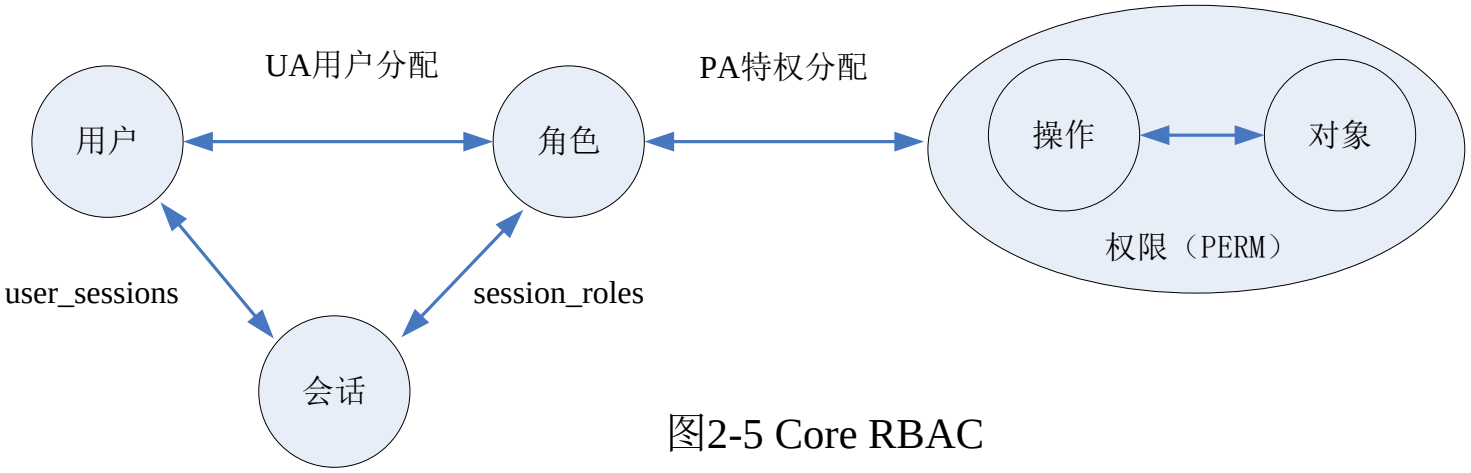


图2-5 Core RBAC

① **UA-用户分配：**用户和角色之间是多对多的映射关系，记录了管理员**为用户分配的所有角色**。

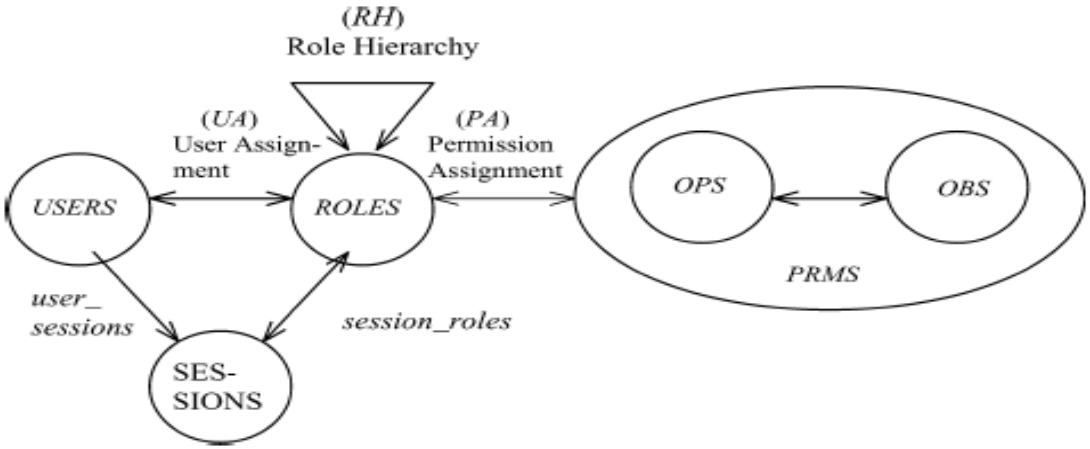
② **PA-特权分配：**角色与权限之间也是多对多的映射关系，记录了管理员**为角色分配的所有权限**。

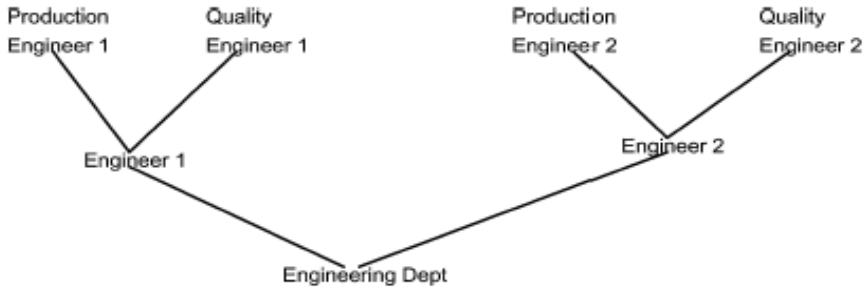
③ **user_sessions：**用户与会话之间的一对多映射关系。即一个用户可通过登录操作开启一个或多个会话，而每个会话只对应一个用户。

④ **session_roles：**会话与角色之间的多对多关系。即用户可以在一个会话中激活多个角色，而一个角色也可以在多个会话中被激活。

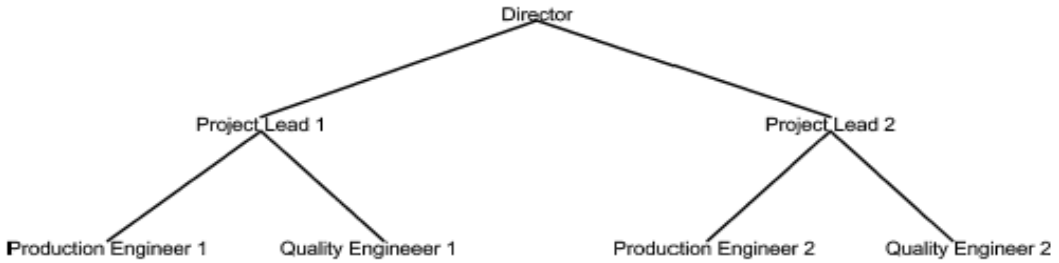
2.1.2 访问控制模型

角色层次





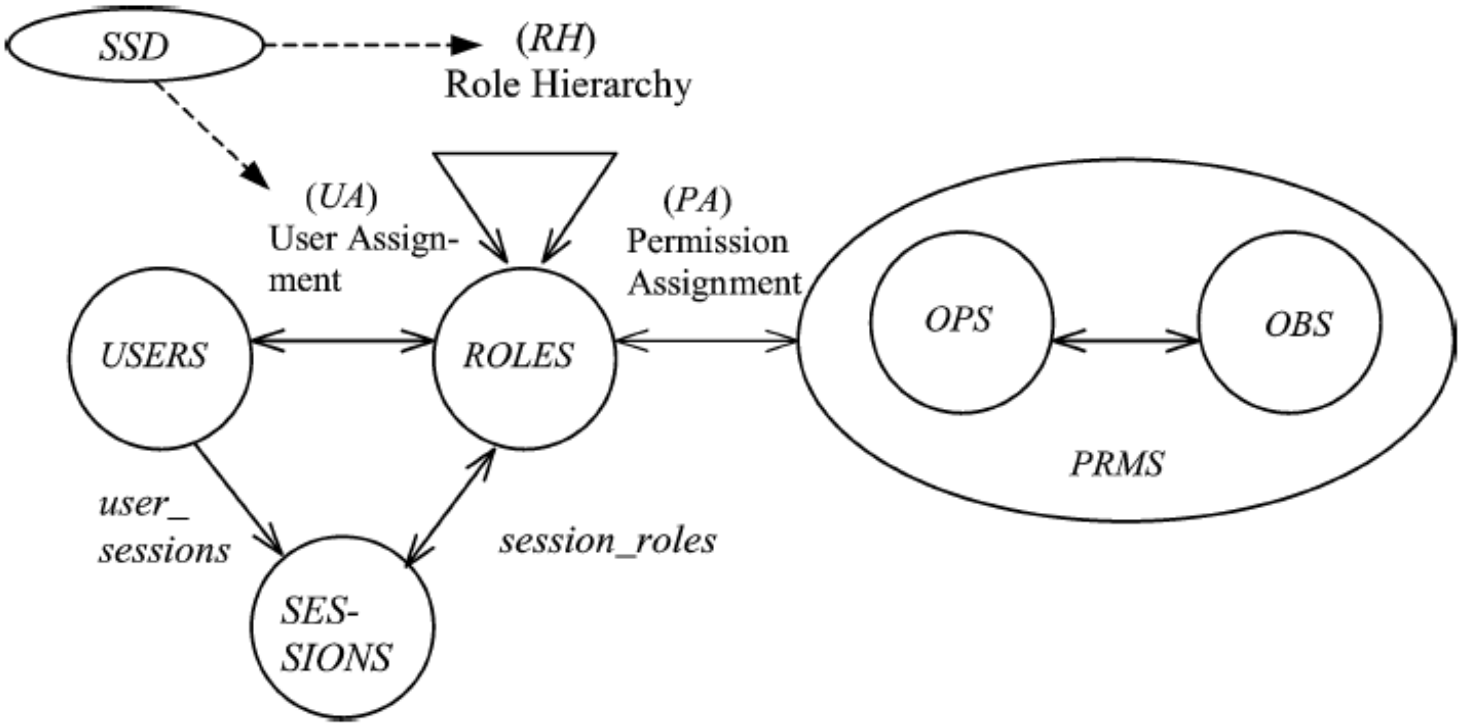
(a)



(b)

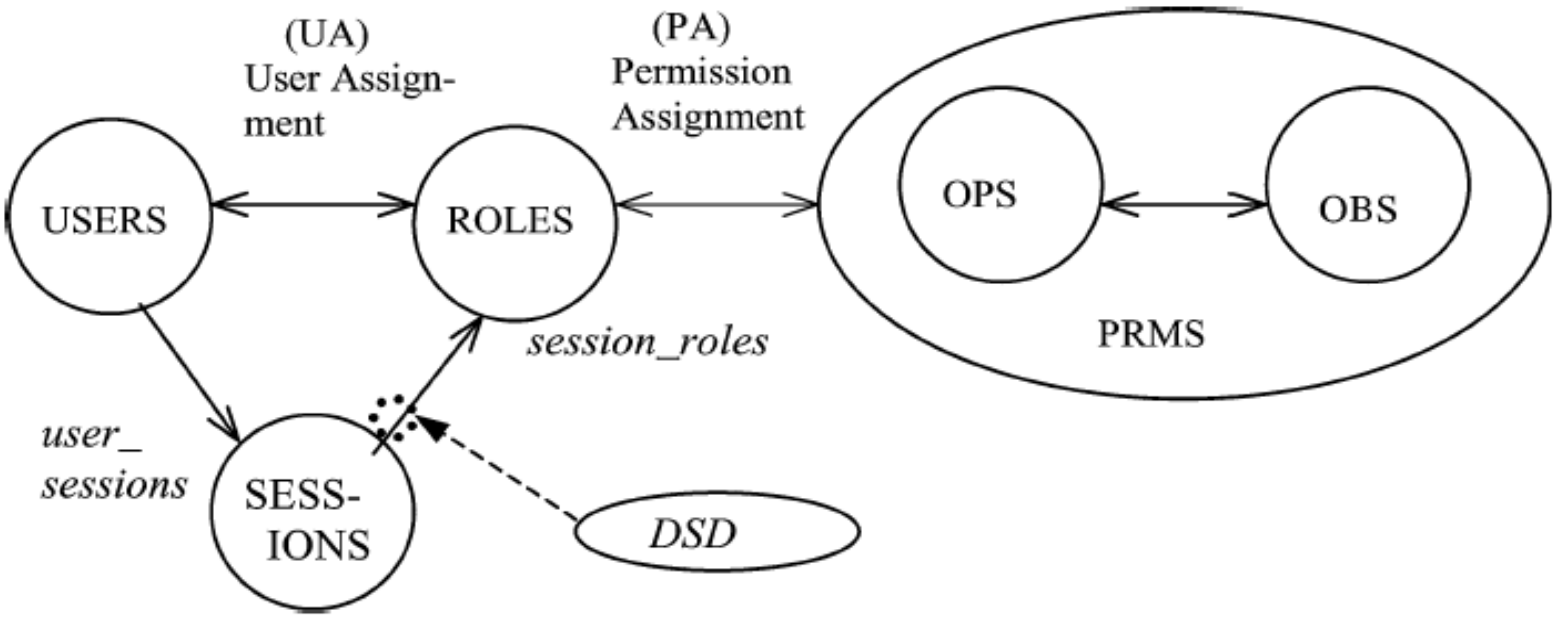
2.1.2 访问控制模型

带静态约束的角色分配



2.1.2 访问控制模型

带动态约束的角色分配



2.1.2 访问控制模型

基于属性的访问控制(ABAC)：通过安全属性来定义授权，并实施访问控制。由于安全属性可以由不同的属性权威分别定义和维护，所以具备**较高的动态性和分散性**，能够较好地适应开放式的环境。具体地，它包括如下几个重要概念：

① **实体：**系统中存在的**主体、客体**，以及**权限和环境**。
② **环境：**指访问控制发生时的系统环境。
③ **属性：**用于**描述上述实体的安全相关信息**。它通常由属性名和属性值构成，又可分为以下几类：
- ✓ 主体属性
- ✓ 权限属性
- ✓ 环境属性

2.1.2 访问控制模型

基于属性的访问控制：

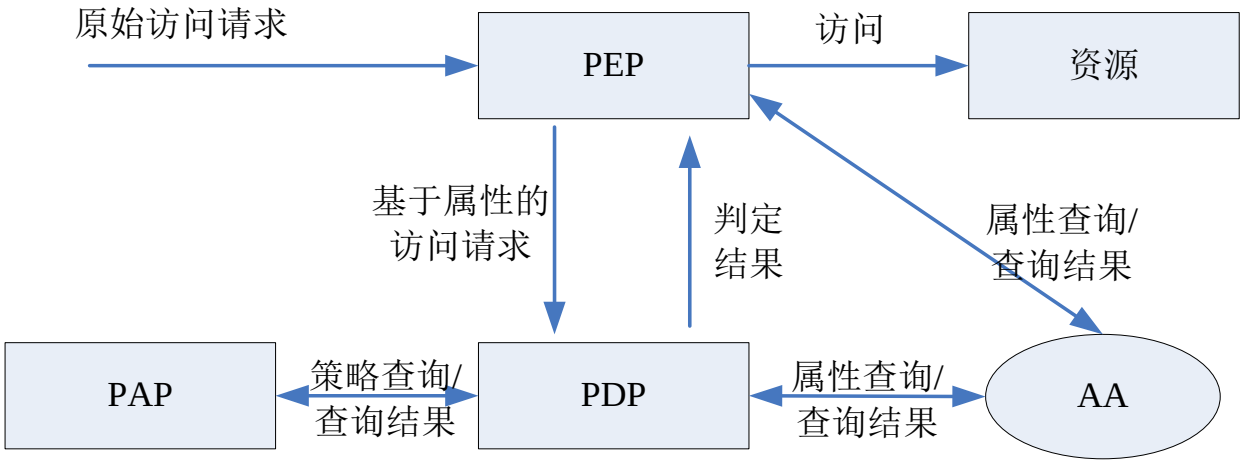


图2-8 ABAC框架示意图

① **AA为属性权威：**负责实体属性的创建和管理，并提供属性的查询。

② **PAP为策略管理点：**负责访问控制策略的创建和管理，并提供策略的查询。

③ **PEP为策略执行点：**负责处理原始访问请求，查询AA中的属性信息生成基于属性的访问请求，并将其发送给PDP进行判定，然后根据PDP的判定结果实施访问控制。

④ **PDP为策略判定点：**负责根据PAP中的策略集对基于属性的访问请求进行判定，并将判定结果返回PEP。

2.1.2 访问控制模型

基于属性的访问控制：

- **适合大数据的开放式数据共享环境**。
- 属性的管理和标记对于安全管理员来说仍然是劳动密集型工作，且需要一定的专业领域知识。
- 在大数据场景下，数据规模和应用复杂度使得这一问题更加严重。

2.1.3 局限性总结

早期访问控制模型和技术在大数据应用场景下，主要存在三方面问题：

安全管理员的授权管理难度更大。

- 工作量大
- 领域知识匮乏

严格的访问控制策略难以适用。

- 访问需求无法预知
- 访问需求动态变化

外包存储环境下无法使用。

- 数据所有者不具备海量存储能力
- 数据所有者不具备构建可信引用监控机的能力

2.2 基于数据分析的控制技术

2.2.1 角色挖掘技术

2.2.2 风险自适应的访问控制技术

2.2.1 角色挖掘技术

在基于角色的访问控制中，管理员需要解决两个问题：

① 创建哪些角色？
② 角色与用户、角色与权限如何关联？

经过前述分析，大数据场景下角色的定义将是大工作量，且需要领域知识的任务。安全管理员已经难以自上而下地分析和归纳安全需求，并基于需求来定义角色了。

2.2.1 角色挖掘技术

为了解决该问题，自底向上定义角色的方法被提出，即采用数据挖掘技术从系统的访问控制信息等数据中获得角色的定义，也被称为角色挖掘（Role Mining）。

目前经典的角色挖掘技术可以分为两类：

① 基于层次聚类的角色挖掘方法
② 生成式角色挖掘方法

2.2.1 角色挖掘技术

基于层次聚类的角色挖掘：

系统在初始情况下往往已经有了简单的访问权限分配——“哪些用户能够访问哪些数据”，例如授权信息表。基于层次聚类的角色挖掘方法将从这类数据中分析和挖掘角色的定义。

表2-1 系统的授权信息表

	权限1	权限2	权限3	权限4	权限5
用户A	✓	✓		✓	
用户B	✓	✓			
用户C	✓		✓		✓
用户D	✓	✓	✓	✓	✓

2.2.1 角色挖掘技术

基于层次聚类的角色挖掘：

以**凝聚式角色挖掘方法**为例。它将权限看作是聚类的对象，通过不断合并距离近的类簇完成对权限的层次聚类，聚类结果为候选的角色。基本定义如下：

① 类簇Cluster：由权限和持有这些权限的用户组成的二元组c=<rights，members>。
② 用户集合Persons：所有用户组成的集合。
③ 类簇集合Clusters：包含所有类簇的聚类结果集。
④ 偏序关系集合<：聚类之间的偏序关系构成的集合。
⑤ 无偏序关系类簇集合$T_<$：类簇集合中的类簇，两两间不存在偏序关系。即

$$T_< = \{c \in \text{Clusters}: \nexists d \in \text{Clusters}: c < d\}$$

且对于任意的类簇对<c,d>$\in T_<$有如下定义：

members(<c,d>)= members(c) ∩members(d)

rights(<c,d>)= rights(c) ∪rights(d)

2.2.1 角色挖掘技术

基于层次聚类的角色挖掘：

算法2-1 凝聚式角色挖掘算法

输入：所有权限及持有权限的用户。

输出：一个类簇Cluster构成的树结构，即Clusters和<。

1）初始化变量

$$\text{Clusters} := \emptyset$$

$$< := \emptyset$$

$$T_< := \emptyset$$

2）为所有单个权限r创建一个类簇c_r，并将其放入类簇集合Clusters和无偏序关系类簇集合$T_<$中

$$\text{rights}(c_r) = \{r\}$$

$$\text{members}(c_r) = \{p \in \text{Persons}: p \text{ has permission } r\}$$

$$\text{Clusters} := \text{Clusters} \cup \{c_r\}$$

$$T_< := T_< \cup \{c_r\}$$

2.2.1 角色挖掘技术

基于层次聚类的角色挖掘：

3）合并距离相近的类簇对产生新类簇。距离最近的类簇对的寻找方式为先寻找出拥有共同用户最多的类簇对集合S，再从S中选出包含权限最多的类簇对集合E。即

$$m = \max\{|members(< c,d >)|: c,d \in T_<\}$$

$$S=\{<c,d>:|members(<c,d>)|=m \wedge c,d \in T_<\}$$

$$r = \max\{|rights(< c,d >)|:< c,d >\in S\}$$

$$E = \{< c,d >: |rights(< c,d >)| = r \wedge <c,d> \in S\}$$

然后从E中选择任意一个<c,d>合并产生新的类簇e，其中

$$rights(e) = rights(c) \cup rights(d)$$

$$members(e) = members(c) \cap members(d)$$

4）更新Clusters、<、$T_<$变量

$$Clusters := Clusters \cup \{e\}$$

$$<:=< \cup \{< c,e >,< d,e >\}$$

$$T_< := T_< \backslash \{c,d\}$$

$$T_< := T_< \cup \{e\}$$

5）重复第3）步和第4）步，直到$T_<$为空。

2.2.1 角色挖掘技术

基于层次聚类的角色挖掘存在的问题：

它们是对已有的权限分配数据进行角色挖掘，所以挖掘出的角色定义的质量往往过多地依赖于已有权限分配的质量。而对于大数据应用这种复杂场景来说，已有权限分配的质量往往很难保证。

2.2.1 角色挖掘技术

练习：凝聚式角色挖掘的偏序关系图

	权限1	权限2	权限3	权限4	权限5
用户A	✓	✓		✓	
用户B		✓			✓
用户C	✓		✓		✓
用户D	✓		✓	✓	✓

2.2.1 角色挖掘技术

生成式角色挖掘：

将角色挖掘问题映射为文本分析问题，采用两类主题模型LDA（Latent Dirichlet Allocation）和ATM（Author-Topic Model）进行生成式角色挖掘，从权限使用情况的历史数据来获得用户的权限使用模式，进而产生角色，并为它赋予合适的权限，同时根据用户属性数据为用户分配恰当的角色。

2.2.1 角色挖掘技术

生成式角色挖掘：

基本定义：

① U是系统中用户的集合；

② P是系统中权限的集合；

③ UP是用户与权限的映射，$UP \subseteq U \times P$ 。

④ USAGE:$U \times P \rightarrow Z$，是一个函数，输入为（u,p）$\in U \times P$，输出为一个数值，该数值为用户u使用权限p的次数。

⑤ GUPA: $U \times P \rightarrow \{0,1\}$，是一个以USAGE为基础定义的函数，该函数的输入为（u,p）$\in U \times P$，输出为0或1。若USAGE(i,j)>0，则GUPA(i,j)=1，否则GUPA(i,j)=0。

生成式角色挖掘的结果为两个集合：

① PA是角色与权限的映射关系，$PA \subseteq R \times P$。

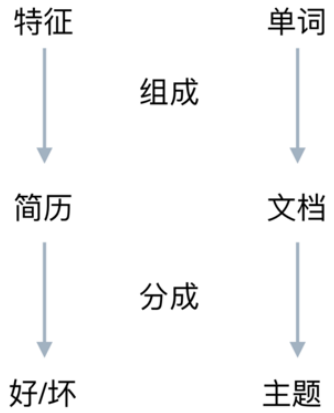
② UA是用户与角色的映射关系，$UA \subseteq U \times R$。

2.2.1 角色挖掘技术

2.2.1 角色挖掘技术

一个资深HR收到一份应聘算法工程师的简历，他想仅仅通过简历来看一下这个人是不是大牛，有可能是：

- 穿条纹衬衫
- 曾在BAT就职
- 做过大型项目



贝叶斯公式：$$P(B_i|A) = \frac{P(A|B_i) * P(B_i)}{\sum_j P(A|B_j) * P(B_j)}$$

$$P(\text{大牛}|\text{简历}) = \frac{P(\text{大牛})P(\text{简历}|\text{大牛})}{\sum P(\text{大牛})P(\text{简历}|\text{大牛})}$$

2.2.1 角色挖掘技术

例： 设有一批同类型的灯泡，是由灯泡一车间、二车间、三车间生产的，其中每个车间生产的灯泡数量分别占这批灯泡的50%，30%，20%，已知各厂的次品率分别为3%，4%，5%。厂长下来视察工作，任意抽查一灯泡，结果发现是次品，考虑该灯泡是一个车间生产的概率？

解： 设 A_i={任取一灯泡是i 车间产品}，B ={任取一灯泡是次品}。

$$P(A_1)=0.5 \quad P(A_2)=0.3 \quad P(A_3)=0.2$$

$$P(B\mid A_1)=0.03 \quad P(B\mid A_2)=0.04 \quad P(B\mid A_3)=0.05$$

$$P(A_i\mid B)=\frac{P(A_i)P(B\mid A_i)}{P(A_1)P(B\mid A_1)+P(A_2)P(B\mid A_2)+P(A_3)P(B\mid A_3)}$$

$$P(A_1\mid B)=\frac{0.5\times 0.3}{0.037}=\frac{15}{37}$$

2.2.1 角色挖掘技术

- **w代表单词；d代表文档；t代表主题；**
- **大写代表总集合，小写代表个体**

$$P(\text{词}\mid\text{文档})=P(\text{词}\mid\text{主题})P(\text{主题}\mid\text{文档})$$

用表达式如下：

$$P(w|d)=P(w|t)*P(t|d)$$

其实就是以主题为中间层，通过前面的两个向量（θ_d，ϕ_t），分别给出$P(w|t),P(t|d)$,它的学习过程可以表示为：

1. LDA算法开始时，先随机地给θ_d，ϕ_t赋值(对所有的d和t)
2. 针对特定的文档d_s中的第i单词w_i，如果令该单词对应的主题为t_j，可以把 上述公式改写为:

$$P_j(w_i|d_s)=P(w_i|t_j)*P(t_j|d_s)$$

3. 枚举T中的主题，得到所有的$pj(wi|ds)$.然后可以根据这些概率值的结果为d_s中的第i个单词w_i选择一个主题，最简单的就是取令$P_j(w_i|d_s)$概率最大的主题 t_j。
4. 如果d_s中的第i个单词w_i在这里选择了一个与原先不同的主题，就会对θ_d，ϕ_t有影响，他们的影响反过来影响对上面提到的$p(w|d)$的计算。

2.2.1 角色挖掘技术

生成式角色挖掘：

文档的生成：一篇文档包括了多个主题，文档中的每个词都是由其中一个主题产生的。也就是存在两个多项式概率分布θ和φ，θ是一个文档上的主题分布，φ是一个主题上的单词出现的概率分布。

角色挖掘问题被映射为文档生成问题，采用LDA模型就能够挖掘出角色定义。

文档生成问题	角色挖掘问题
包含多个文档的“语料库”	访问控制日志
一个文档u	用户u的权限使用记录
单词p	权限p
文档u中单词p的词频n	用户u对权限p的使用次数n
主题r	角色r

2.2.2 风险自适应的访问控制技术

访问控制是一种平衡风险和收益的机制。

① 传统访问控制：风险与收益的平衡被静态定义在访问控制策略中，即“满足策略约束条件的访问行为所带来的风险”被视为系统可接受的。

② 基于风险的访问控制：风险与收益的平衡是访问过程中动态实施的，而非预先定义在访问控制策略中。

2.2.2 风险自适应的访问控制技术

风险量化：风险量化是将访问行为对系统造成的风险进行数值评估，它是基于风险来实施访问控制的前提。

常见的风险要素：

① 被访问客体敏感程度是客体重要性的体现。

② 被访问客体的数量是指主体在一次访问请求中或一段时间内所访问的客体的规模。

③ 客体之间的互斥关系描述了多次访问行为的风险累加是非线性的。即两个客体存在如下关系：对其中一个客体访问后将不能访问另一客体，或者再访问另一客体时带来的风险会急剧增加。

④ 访问主体的安全级别是实施了强制访问控制的系统中对主体访问敏感客体时所能达到的安全性的评估。

⑤ 访问目的与被访问客体的相关性体现了在业务流程中主体对客体的需求程度。

2.2.2 风险自适应的访问控制技术

风险量化：

目前主流的风险计算方法分为基于概率论或模糊理论的静态方式，以及基于协同过滤的动态方式两类。

基于概率论的风险量化：

其核心思想是“风险量化值由危害发生的可能性和危害程度决定”，即

$$\text{Quantified Risk} = (\text{Probability of Damage}) \times (\text{Value of Damage})$$

其中，危害的值(Value of Damage)是一个对危害程度的量化度量，往往取决于信息资源的价值，只能由企业或组织根据业务背景自行评估。而危害发生的可能性（Probability of Damage）是指引发该危害的事件发生的可能性，通常采用概率论进行计算。

2.2.2 风险自适应的访问控制技术

风险量化：

基于协同过滤的动态风险量化：

基本思想是利用系统中用户的历史访问行为来构建正常用户的访问行为画像，并以此为风险量化的基准，然后计算每次用户访问行为与该基准的偏离程度作为风险量化值。即访问行为偏离基准越大，则该访问产生的风险越大。

其特点是通过行为异常的概率来衡量风险值，所以风险量化结果可以随着系统中整体用户的行为变化而动态变化，相比于静态计算方法更加灵活。然而这种计算往往需要大量的系统历史数据以确保风险量化的准确性。

2.2.2 风险自适应的访问控制技术

访问控制实施方案：

判定结果从“允许/拒绝”的二值向多值发展，引入了部分允许的概念。

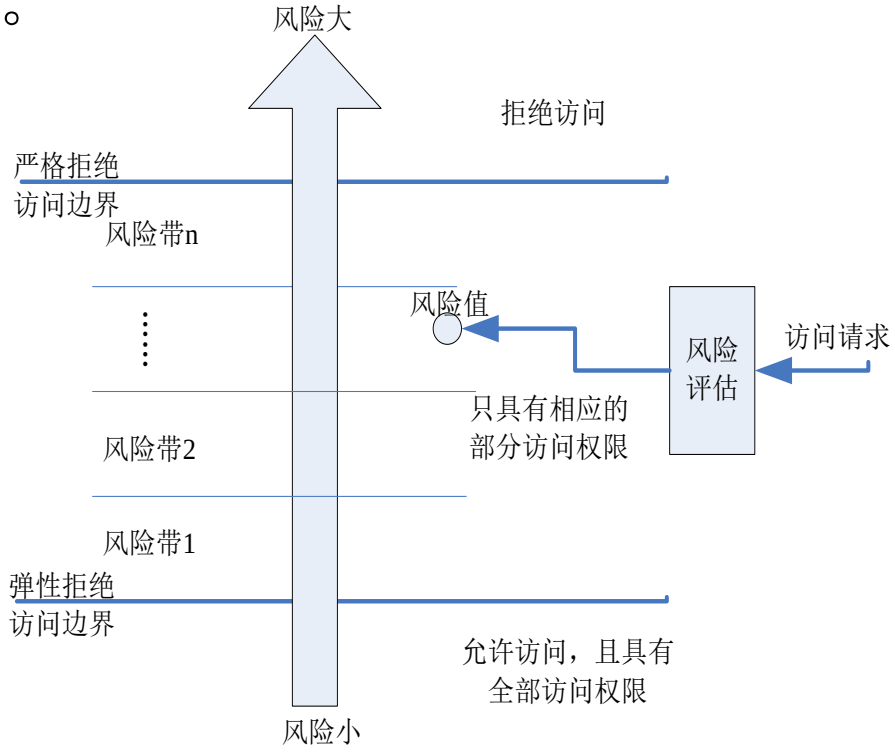


图2-10 采用风险带的访问控制

2.2.2 风险自适应的访问控制技术

动态借钱方案

9	跟我认识几年					
10	是否已婚		是	否		
11	有无借钱不还记录		有	无		
12	上次借钱记录		借钱时间			
13			还钱时间			
14	跟我借钱理由：(不低于1500字申请书)					
15	送我最贵重的礼物是什么：					
16	我们之间产生过几次矛盾：					
17	上次请我吃饭是什么时候：					

2.2.2 风险自适应的访问控制技术

访问控制实施方案：

实现整个系统风险与收益平衡的方法包括：

① 信用卡式：它为每个用户分配风险额度，并让用户在访问资源时根据访问带来的风险去消耗额度。当额度不足以支付新的访问时，系统将阻止用户的访问行为。

② 市场交易式：它将风险视为市场上的商品，而整个系统能够容忍的风险被视为可以交易的商品总量。作为商品的风险流通越充分，则越能够实现整体系统的风险与收益的最优化配置。

2.2.2 风险自适应的访问控制技术

访问控制实施方案：

风险访问控制通常采用与传统访问控制结合的实施框架。

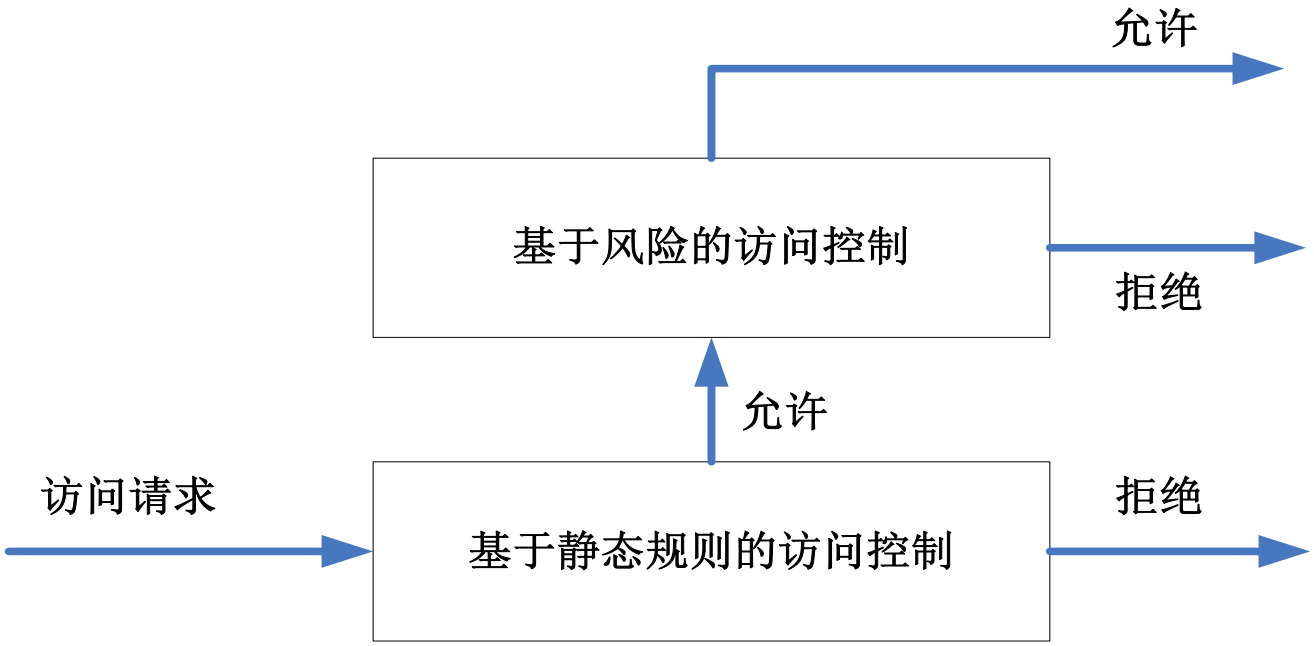


图2-11精化式访问控制实施框架

目 录

2.3 基于密码学的访问控制技术

- **依赖于密钥的安全性**，不需要可信引用监控机，有效解决大数据安全的问题：
  - 大数据分析架构，**很难建立可信引用监控机**
  - 大数据场景下，**数据经常处于所有者控制范围之外**
- **基于密钥管理**的访问控制技术：可信密钥管理服务器，或广播加密(broadcast encryption)
  - 基于**单发送者广播加密**的访问控制：广播发送者必须持有所有接收者的对称密钥，密钥和密文大小是参与方数量的对数级别
  - 基于**公钥广播加密**的访问控制技术：广播发送者必须持有所有接收者的公钥
  - 基于**属性加密**的访问控制技术：将属性集合作为公钥进行数据加密，包括基于密钥策略的KP-ABE和基于密文策略的CP-ABE。

2.3 基于密码学的访问控制技术

2.3.1 基于密钥管理的访问控制技术

- 基于单发送者广播加密的访问控制
- 基于公钥广播加密的访问控制技术

2.3.2 基于属性加密的访问控制技术

2.3.1 基于密钥管理的访问控制技术

- 基于密钥管理的访问控制技术，是**通过严格的密钥管理来确保授权用户才能有解密数据所需要的密钥**，来实现访问控制。
- 根据访问控制系统所支持的能够发送数据的用户数量，可以分为：

① 基于单发送者广播加密的访问控制

② 基于公钥广播加密的访问控制

2.3.1 基于密钥管理的访问控制技术

基于单发送者广播加密的访问控制

- **数据所有者：**拥有数据和完整的用户密钥树，负责根据数据分享的目标对象，有选择地从用户密钥树中选取加密密钥对数据进行加密，并将加密结果广播式发送给所有用户。
- **普通用户：**拥有用户密钥树中的与自己相关的部分密钥，负责接收数据密文并利用自己持有的密钥解密数据。

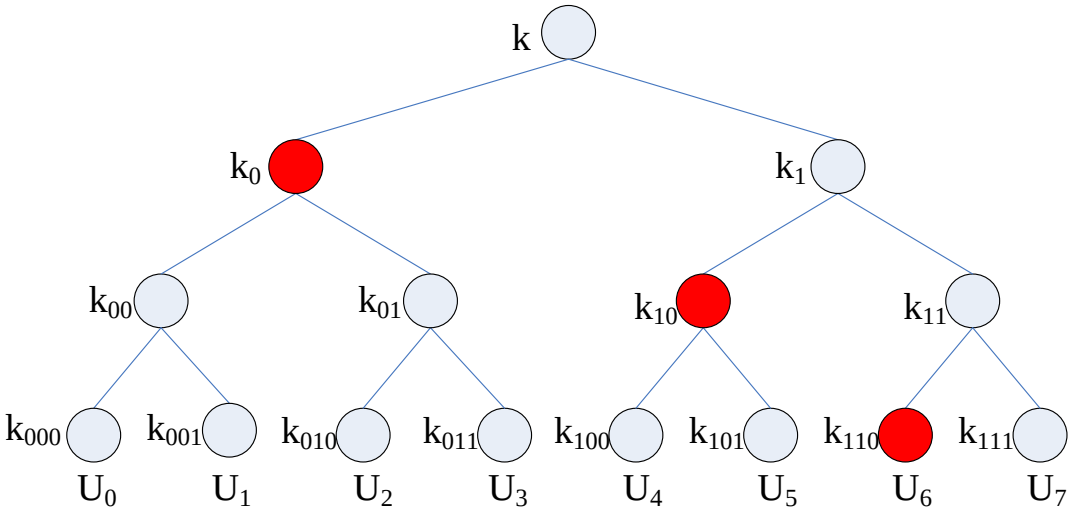


图2-12 一颗用户密钥树

选择红色节点处的密钥集{k_0，k_{10}，k_{110}}进行数据加密，则未授权的普通用户就是U_7，他将无法解密数据。

2.3.1 基于密钥管理的访问控制技术

基于公钥广播加密的访问控制

- **公钥服务器：**负责维护一个密钥集合。即将系统中的所有用户划分为子集，每个子集代表了可能的数据接收者集合。为每个子集产生公私钥对，并将私钥安全分发给其包含的用户。
- **数据所有者：**负责将数据加密，并采用基于公钥广播加密技术对加密密钥进行分发，以实现对授权接收者的限定。
- **数据服务者：**负责加密数据的存储，并向用户提供对数据的操作。
- **用户：**是数据的访问者。只有被数据所有者授权的用户才能获得数据的加密密钥，并进一步解密出数据。

由于采用公钥加密方式，所以**系统的所有用户都可以是数据所有者**，并向其他用户分享数据，消除了单发送者广播加密方案对发送者范围的限制。

2.3.1 基于密钥管理的访问控制技术

基于属性加密的访问控制

- 在**基于密钥管理的访问控制**中，系统**通过控制用户持有的密钥集合**来区分用户，进而实施授权和访问控制。
- **基于属性加密的访问控制**是通过更加灵活的属性管理来实现访问控制，**即将属性集合作为公钥进行数据加密**，要求只有满足该属性集合的用户才能解密数据。

2.3.1 基于密钥管理的访问控制技术

基于属性加密的访问控制

定义2-1（访问结构，Access Structure） 令$\{P_1,P_2,....,P_n\}$是一个参与者集合。令$A\subseteq 2^{\{P1,P2,....,Pn\}}$，若$\forall B,C$，有$B\in A$，且$B\subseteq C$，那么$C\in A$，则称A是单调的。若A是单调的，且是非空的，即$A\subseteq 2^{\{P1,P2,....,Pn\}}\setminus\{\emptyset\}$，则称A为一个访问结构。A中的元素被称为授权集，非A中的元素被成为未授权集。

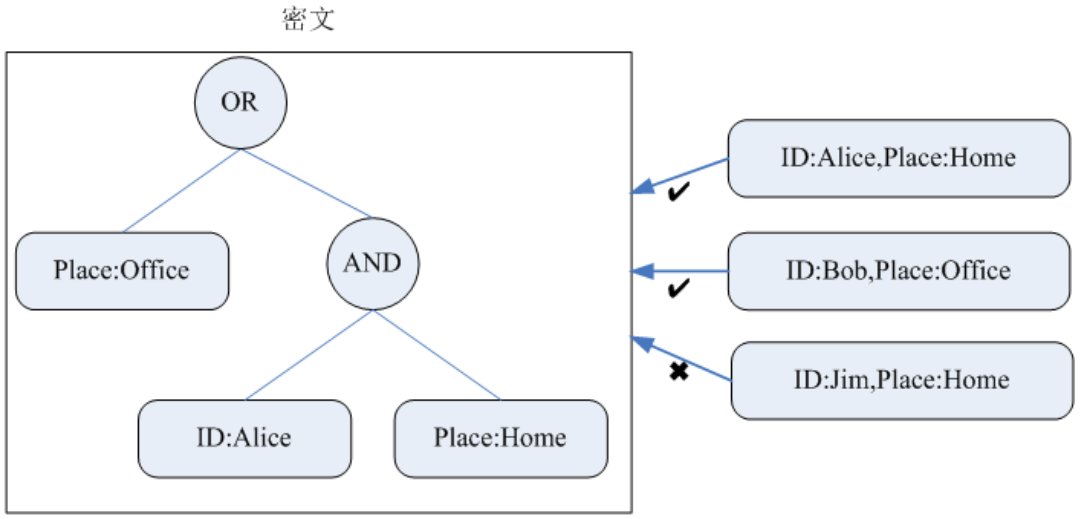


图2-18 CP-ABE访问控制结构示意图

访问树结构是常见的一种访问结构，它限定了授权集，即“Place属性为Office，或ID为Alice且Place为Home的用户能够解密数据”

2.3.1 基于密钥管理的访问控制技术

基于属性加密的访问控制

单属性权威的方案架构

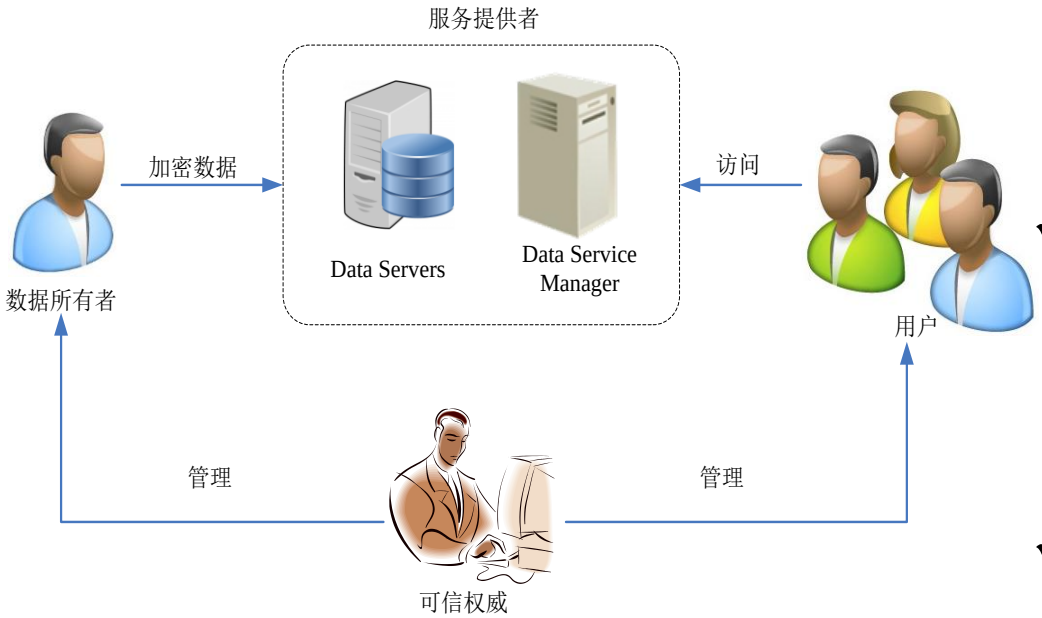


图2-19 基于CP-ABE的访问控制

- ✓ **可信权威**：维护了每个用户的属性与密钥的对应关系，为用户发布属性密钥。
- ✓ **数据所有者**：具有数据的所有权，负责访问策略（访问结构T）的定义，并产生与策略绑定的密文数据，然后发送给服务提供者。
- ✓ **用户**：是数据的访问者。若该用户具有满足密文数据所绑定策略中要求的属性，即持有恰当属性密钥，那么就可以解密出数据明文。
- ✓ **服务提供者**：负责提供数据的外包存储。

2.3.1 基于密钥管理的访问控制技术

基于属性加密的访问控制

多属性权威的方案架构

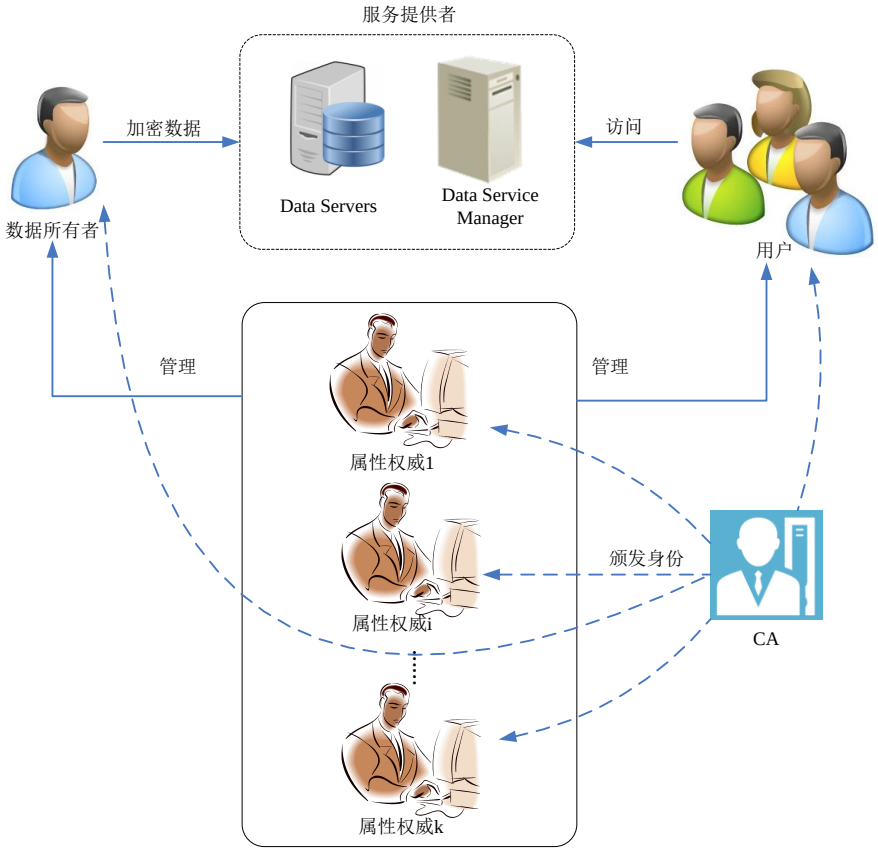


图2-20 支持多属性权威的基于CP-ABE的访问控制

- ✓ **CA**：负责为整个系统中所有用户和属性权威颁发和维护身份。
- ✓ **属性权威**：负责颁发、撤销和更新用户属性。
- ✓ **数据所有者**：具有数据的所有权，产生访问结构来描述授权用户的范围，并采用多权威的属性加密算法对数据加密。
- ✓ **用户**：是数据的访问者。如果用户的属性满足访问结构，则用户能够成功解密出数据。
- ✓ **服务提供者**：负责提供数据的外包存储。