

基于 Apriori 算法的网络取证设计

钟秀玉

(嘉应学院 计算机学院, 广东 梅州 514015)

摘 要:传统的计算机取证在事后收集证据,证据的法律效率低。网络取证把入侵发生后的被动调查转为事件发生之前的主动防御。基于 Apriori 算法的网络取证系统挖掘各种犯罪事件的关联,建立犯罪特征库。系统在获取、过滤网络数据包后,对原始数据进行协议分析,挖掘数据包间的关联信息,提取关联规则记录,再根据当前用户行为记录与犯罪特征规则的匹配结果来判定当前用户行为是否非法。为保证证据更具有原始性、完整性和法律效率,系统对获得原始数据进行加密传输,使用基于 SSL 的加密认证安全设计,防止证据泄露和被伪造。模拟实验表明,Apriori 算法的应用提高了非法入侵检测效率,可识别新的犯罪行为,系统完整地重构犯罪过程。

关键词:Apriori 算法;网络取证;数据获取;数据分析;模式匹配

中图分类号:TP309.08

文献标识码:A

文章编号:1673-629X(2011)05-0158-05

Design of Network Forensics Based on Apriori Algorithm

ZHONG Xiu-yu

(School of Computer, Jiaying University, Meizhou 514015, China)

Abstract: Because the traditional computer forensics collects evidence after events, the legal efficiency of evidence is low. Network forensics turns the passive investigation after events to the active defense before events. Network forensics based on Apriori algorithm mines the association of crime events to build crime characteristic database. After gaining and filtering the network data packet, the system carries on protocol analysis to the primary data, the association information between data packets are mined and the association rule records are extracted, and the current user behavior is illegal or not according to match result of the current user behavior records and the crime characteristic rules. In order to guarantee the primitiveness, integrity and legal efficiency of evidence, the system uses encryption transmission to the primary data and uses the SSL encryption authentication safe design to prevent evidence revealed and fabricated. Simulation results show that the application of Apriori algorithm increases illegal invasion detection efficiency and can identify new crime, and the system restructures criminal process completely.

Key words: Apriori algorithm; network forensics; data collection; data analysis; pattern match

0 引 言

网络安全威胁越来越严重,传统的取证方法对犯罪事件进行事后记录,没有对犯罪行为进行实时的监视记录,存在证据不准确、不完整等问题,特别是随着网络技术的发展,很多非法入侵通过计算机网络实施,传统取证方法无法及时、全面地获取网络入侵信息,不能满足取证要求。要从根本上解决计算机犯罪问题,就要将犯罪分子绳之以法,进行网络取证^[1,2]。

网络取证是使用科学的证明方法收集、融合、发现、检查、关联、分析和存档数字证据,证据涉及多层次的主动处理过程以及数据源的传递过程,以发现有预

谋的破坏行为或已经成功的非授权的犯罪行为,并为应急事件的响应和系统恢复提供有用的信息^[3,4]。文献[5]提出了基于计算机日志分析的取证原型系统,从数据挖掘的角度设计取证分析系统,系统包括日志处理、挖掘与分析。文献[6]提出了 CYDEST (The CYber DEfenSe Trainer) 应用于计算机取证,以提高取证的准确性。文献[7]提出了基于增量挖掘的大规模网络攻击检测的方法,通过网络流量的增量异常情况分析检测攻击行为。文献[8]提出了把日志当作主要信息源,对日志进行预解码、解码,根据日志与规则中的规则匹配情况来判定非法入侵行为,实现动态取证,但规则树的维护成本高。文献[9]利用 Bloom filter 数据结构的特点,提出了基于 Bloom filter 的网络取证系统结构,解决了海量数据传输和存储带来的资源瓶颈问题,但往往取证工作脱离需求。文献[10]提出了基于协议分析的入侵检测方法应用于取证,解决动态取

收稿日期:2010-11-02;修回日期:2011-02-10

基金项目:广东省自然科学基金项目(9151009001000043);广东省科技计划项目(2009B060700002)

作者简介:钟秀玉(1972-),女,硕士,副教授,CCF 会员,研究方向为网络安全、数据库应用。

证的实时性问题。文献[11]提出了用蜜罐技术发现入侵模式的方法,但入侵证据的法律性受质疑。基于上述参考文献的研究以及复杂的非法入侵往往是综合多变的情形,网络取证有必要将获取的大量网络信息和证据关联起来,形成证据链。文中提出基于数据挖掘的网络取证模型,应用数据挖掘技术提取与特定网络犯罪有关的特征,在取证分析阶段,将各种可能得到的信息关联起来,作为最后判断非法入侵的依据。Apriori 算法是一种最有影响的挖掘布尔关联规则频繁项集的算法,应用于网络取证,提高取证效率。

1 基于 Apriori 算法的网络取证模型

1.1 网络取证逻辑模型

基于 Apriori 算法的网络取证系统模型如图 1 所示。系统主要由五部分构成:数据获取、数据分析、数据挖掘、证据鉴定和证据提交,核心是数据获取和数据分析。数据获取模块的数据来源一方面来源于网络原始数据包,另一方面来源于入侵检测等网络安全产品,对每个数据包按协议进行解析,如对 IP、ARP、ICMP、TCP、UDP 以及部分应用层协议的解析,并定义过滤规则,对数据包进行底层过滤,对所有可能的计算机犯罪行为进行实时数据获取。数据分析模块主要应用 Apriori 算法建立犯罪知识库,同时根据法学和信息安全分析技术对过滤后的数据进行挖掘关联信息,对挖掘的关联信息记录与知识库中的规则进行模式匹配,提取证据。在确保系统及证据安全的情况下,进行证据鉴定与提交。

系统达到四个指标,一是实时记录入侵全过程,包括捕获账号、口令、IP 地址、MAC 地址、时间、操作序列、状态等;二是实施严格的证据保护机制,防止证据被篡改或删除;三是具有一定程度的通用性,可以针对不同种类的犯罪进行必要的定制,系统建立非法入侵规则库时,以常见犯罪行为特征作为初始依据;四是知识库不断更新,能识别新的犯罪行为。

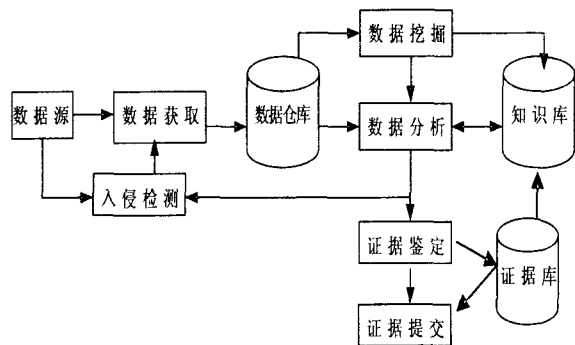


图1 基于 Apriori 算法的网络取证模型

1.2 物理设计

基于 Apriori 算法的网络取证系统在物理架构上

由被取证机、监控器和分析机构成,其中监控机配置双网卡,一网卡连接监控网络,另一网卡连接分析机所在内网。监控机记录原始网络数据,及时获取数据并对所获数据作初步处理:解析网络数据包,按自定义的数据模型存取各类包段信息。分析机对数据进一步处理与分析:一方面,根据知识库中定义的犯罪特征规则,分析、筛选出非法入侵行为,作为入侵证据。另一方面,对异常网络数据包进行事后分析、挖掘,发现新的犯罪方法和工具,生成新的犯罪特征规则并加入知识库,以便下次检测时能识别新的犯罪行为。网络取证的工作主要包括如下步骤:

- (1)在被取证机上启动收集系统信息的 Agent,并启动监控机的网络数据收集模块;
- (2)在被取证机与监控机之间建立认证;
- (3)原始数据从被取证机传输到监控机;
- (4)对监控机上的数据进行数字签名;
- (5)在监控机与分析机之间建立认证;
- (6)将需要的数据从监控机拷贝到分析机;
- (7)分析机对数据进行分析;
- (8)证据鉴定;
- (9)证据提交。

2 基于 Apriori 算法的网络取证设计

2.1 开发平台

在网络取证中,主要是数据获取与数据分析模块,数据获取的工具很多,如伯克利数据包过滤器 BPF 技术。本系统使用 Jpcap 作为开发工具,采用 Linux+Libpcap+Mysql 为开发平台。Jpcap 依赖本地库的使用,在 Windows 或 UNIX 上必须有必要的第三方库,即 WinPcap 或 libpcap。

2.2 数据获取模块

数据获取模块主要完成网络数据采集和存储功能。系统将取证机的网卡设置为混杂模式,以获取共享以太网中的所有通讯数据,监听监控网络的计算机。系统直接从数据链路层接收数据包,把获取的网络数据流以原始数据包的形式按时间段存储到数据库中,从而有力保证了证据的原始性。同时,存放在取证机中的网络数据包可实时或事后传送到分析机,进行在线或离线分析。系统应用 Jpcap 平台进行网络数据采集的过程如下:

- (1)数据获取。设置网卡为混合模式,获取监控网络的数据包。
- (2)数据预处理。函数 setFilter 实现网络数据的过滤,sendPacket 函数发送数据包;同时对数据包进行分类,如 TCP 包、UDP 包、ARP 包、ICMP 包等。
- (3)数据保存。对获取、筛选数据包的各种信息

字段封装到实体类,获取到的实体类存放到集合 Vector,并保存到 MySql 数据库,以备进一步处理与分析。数据包分析返回数据包的状态字符串,特别注意 TCP 数据包状态分析。

2.3 数据分析模块

从两方面进行非法入侵检测,一是在较高网段获取网络分组数据包,将其与异常数据库中的网络异常数据包的特征信息表条目进行模式匹配,相同则进入证据鉴定,并将相关信息存入结果数据库;二是分析主机日志,即将新的记录条目与异常数据库中日志特征信息表相比较,如匹配则进入证据鉴定,并将相关信息存入结果数据库。具体主要包括如下过程:

(1)在客户端和服务端导入密钥和证书,建立 SSL 安全套接层连接;

(2)将之前 Vector 集合中的数据包实体类通过 SSL 传输给分析机;

(3)对传输过来的数据包实体类进行备份;

(4)分析数据包实体类,筛选出一些异常数据,以作为网络非法入侵证据。

此外,系统还可从网络安全产品中直接获取报警数据。数据分析模块核心是第(4)步,主要对获取的数据进行挖掘和模式匹配。

2.3.1 数据挖掘生成强关联规则

Apriori 算法挖掘布尔关联规则频繁项集。关联规则具有如下两个重要的属性:支持度 $P(A \cup B)$,即 A 和 B 这两个项集在事务集 D 中同时出现的概率。置信度 $P(B|A)$,即在出现项集 A 的事务集 D 中,项集 B 也同时出现的概率。同时满足最小支持度阈值和最小置信度阈值的规则称为强规则。给定一个事务集 D ,先应用 Apriori 算法找出所有的频繁记录集,这些项集出现的频繁性至少和预定义的最小支持度相同^[12]。然后由频集产生强关联规则,用于网络取证分析阶段进行犯罪判定。

在本系统设计中,频繁记录集中的每个记录对应一个 Note 类实例,Note 类属性包括 count、column 和 data 等,各自的含义如下:(1)count 表示该记录出现的数量。(2)column 是一个 Vector 集合,表示维度的集合,存放着属于本记录的维度。(3)data 是一个 Vector 集合,存放维度对应的值的集合。应用 Apriori 算法生成频繁记录集的流程见图 2。函数 creatC1 获得第一个候选集;函数 getL 获得频繁集;函数 getC 获得候选集;函数 count 计算候选集的记录数。

挖掘生成关联规则的算法具体步骤如下:

(1)从 creatC1 函数获取候选集 C_1 ,其中项集的集合成员包括 { pid, timestamp, length, protocol, srcMAC, srcIP...state };

(2)根据设定的最小支持度阈值,从 C_1 中选出频繁 1 - 项集 L_1 ;

(3)生成候选集 $C_2 = L_1 \cup L_1$,根据最小支持度阈值从 C_2 筛选生成频繁 2 - 项集 L_2 ;

(4)生成候选集 $C_3 = L_2 \cup L_2$,根据最小支持度阈值从 C_3 筛选生成频繁 3 - 项集 L_3 ;

(5)重复上述过程,直到 $C_k = \emptyset$,生成所有频繁项集;

(6)对每个频繁项集 L ,生成 L 的非空子集;

(7)对每个非空子集,如果满足最小置信度阈值,则生成相应的规则。

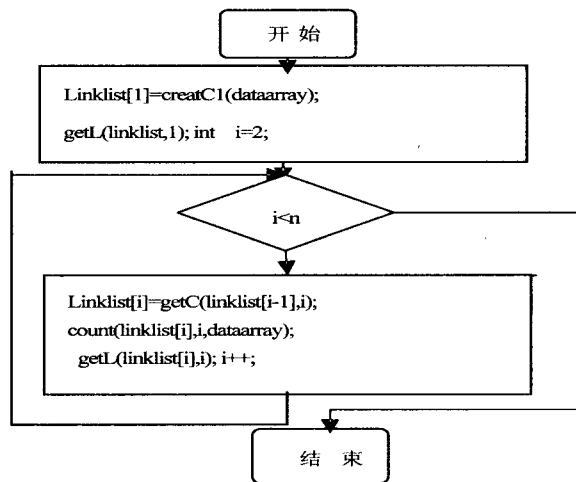


图 2 频繁记录集生成流程图

2.3.2 对当前网络行为进行模式匹配

Note 类对象对应频繁记录集中的每个记录,Note 实体类中设置了匹配方法 $\text{match}(\text{cha}:\text{String}):\text{Boolean}$,实现特征字符串匹配过程。参数 cha 表示要匹配的特征字符串,如匹配 SYS Flood 攻击的字符串 cha 为 “3=TCP,10=A,5”,3 表示存放数据库表中的第 3 列(protocol 所在的列),即 protocol = TCP,10 表示数据库表中第 10 列(即 state 属性),10 = A 表示 state 属性值为“A”,5 表示数据库表中的第 5 列,返回 Boolean 型的值表示是否匹配。系统在获取、预处理网络数据包后,通过 Apriori 算法生成频繁记录集,只对频繁记录集中的记录与非法入侵特征字符串进行模式匹配,这样,极大减少了匹配记录的个数,提高检测速度。

2.3.3 安全性设计

取证必须在确保系统安全的情况下最大限度地获取完整的入侵证据,保证证据的法律效率。本网络取证系统采用基于 SSL 协议的安全传输技术来保障数据传输的安全性。在监控器到分析器开始传输数据之前先建立分析器的 SSL 连接,在完成 SSL 设置、导入密钥和证书后才进行 SSL 传输。SSL 协议中的握手协议完成服务器端和客户端相互鉴别对方身份。此外,对系统得到的证据则加密后由 VPN 专线传输到证据库。

从而保证数据在传输过程中保持真实、完整和不可篡改性。

3 Apriori 算法应用

一方面,在网络取证的数据分析阶段,使用 Apriori 算法对海量的网络连接状态记录进行挖掘,将网络中连接状态记录的频繁项提取出关联规则记录,提取的关联规则记录与知识库的规则进行模式匹配,判断当前用户行为是否具有犯罪特征或与某一犯罪事件相关,提取可能的犯罪完整证据。另一方面,计算机犯罪行为内部、行为之间往往具有一定的关联性,相应地,非法入侵对应的网络数据包之间常常表现出一定的关联。故可应用数据挖掘技术来挖掘不同犯罪形式的特征、同一事件的不同证据间的关联特征,发现新的犯罪行为特征,形成相应的事件模式不断更新知识库,提高系统的犯罪识别率。

应用 Apriori 算法挖掘犯罪模式,建立、更新知识库,在网络取证的数据分析阶段应用知识库规则来判断当前用户行为的合法性,Apriori 算法的应用流程如图 3 所示。

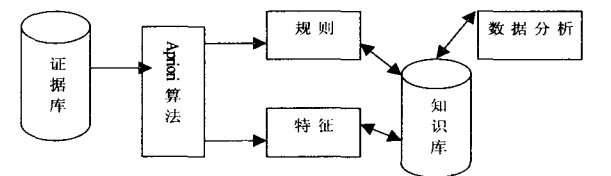


图 3 Apriori 算法在网络取证的应用流程

Apriori 算法应用于网络取证,提高匹配速度,识别新的非法入侵行为。如在 SYN Flood 入侵攻击中,入侵者利用 TCP/IP 协议的缺陷,在短时间内向目标主机发送大量的 SYN 包请求建立连接,但对目标主机返回的 SYN/ACK 应答包不进行 ACK 包确认,使目标主机不断对伪连接进行重试直到超时丢弃,造成目标主机的某个端口在一短时间内收到许多状态为“50”的半连接,相应的半连接列表很快被塞满,从而使得目标主机的资源耗尽,暂停服务。系统设计时把“服务”设为轴属性,“目标主机”设为引用属性,寻找相同目标主机的频繁序列服务模式,该模式表示某个主机的 HTTP 服务遭到了 SYN Flood 攻击,正常的服务请求不会在短时间内频繁出现半连接。再如对端口扫描攻击,攻击者在入侵前对目标主机的各端口进行扫描,试图查找漏洞,造成连接记录上出现在一短时间内目标主机接收到针对不同端口的多次连接,相应地连接记录上的状态有很多“REJ”标记。将“目标主机”设为轴属性,“连接状态”设为引用属性,找出相同目标主机的频繁序列模式。利用频繁序列挖掘可以得到多种表示存在入侵的模式,不断将新入侵模式加入知识库。

4 实验结果与分析

4.1 测试过程

在校园网中利用攻击工具进行模拟测试,测试过程如下:在抓包之前进行设置,建立数据库,监控器抓取数据包,把原始数据存入数据库;然后进行 SSL 设置,导入密钥和证书,进行监控器到分析器的 SSL 传输;对传输过来的数据可以进行保存或对已存在的数据备份进行读取;设置要分析的时间片;进行数据分析与模拟攻击。系统可以自行添加非法入侵特征字符串,新的犯罪特征可随时加入特征库中,提高系统的扩展性和犯罪识别率。

4.2 测试结果与分析

系统进行了功能测试,实现了犯罪检测、证据保存、犯罪过程重现和数据查询功能。犯罪检测以 Ping Flood、SYN Flood、ARP 欺骗和 TFN2K 工具模拟生成攻击数据包。证据保存功能对保存的相关记录进行进一步确认、过滤、加密保存、备份等操作。犯罪过程重现实现还原网络数据(如 WEB、E-mail、FTP 等),重现网络非法入侵过程,分析出新的犯罪方法和工具,以此作为诉讼的依据。数据查询功能根据 IP 地址、MAC 地址、端口号、协议类型进行查询,迅速分析出数据流量中的异常。

在性能测试方面,系统主要对 Ping Flood、SYN Flood、ARP 欺骗和 TFN2K 工具产生攻击进行多次模拟测试,特别是对 SYN Flood 攻击进行多次测试。系统的最小支持度设为 0.3,最小置信度设为 0.8,系统对不同持续时间攻击的检测率与误报率如表 1 所示。

表 1 不同持续时间不同典型攻击的检测结果(%)

攻击类型	持续时间(s)	检测率	误报率
Ping Flood	20	95.63	0.51
	40	97.81	0.64
	60	98.57	0.85
SYN Flood	20	96.33	0.66
	40	97.56	0.68
	60	99.28	0.72
ARP 欺骗	20	95.49	0.71
	40	96.92	0.87
	60	98.44	1.08
TFN2K 工具攻击	20	94.86	1.01
	40	96.52	1.02
	60	97.68	1.13

表 1 的实验结果表明,基于 Apriori 算法的网络取证系统能实时准确地检测,随着持续攻击时间的增加,系统检测率提高,具有较高的检测率和较低的误报率。此外,系统检测到一些新攻击,系统还对不同的最小支

持度下检测速度、入侵检测率进行了试验,实验表明,最小支持度越小,检测速度越慢,规则生成的时间越长,生成的规则越多,误报率可能性越大,若支持度太大,则相反,但检测率可能越低,造成漏报率较高,经多次实验,本系统将最小支持度设为 0.3,在检测速度、检测率和误报率上找到一个平衡点。

5 结束语

文中重点研究网络取证中的数据获取、分析与数据挖掘 Apriori 算法的应用,并进行了实验模拟。系统把数据挖掘、模式匹配和协议分析技术结合起来,先对获取的原始数据预处理、协议分析,再根据分析结果和挖掘生成的网络数据包关联规则记录,调用相应的犯罪特征库进行模式匹配,这样大大减少了匹配的次数,提高了检测效率。模拟实验表明,Apriori 算法的应用提高犯罪行为识别效率,发现新的犯罪行为,使取证工作处于主动状态;系统完整地重构犯罪行为,使证据更具完整性和法律效率。

参考文献:

- [1] 王 玲,钱华林. 计算机取证技术及其发展趋势[J]. 软件学报,2003,14(9):1635-1644.
- [2] Ayers D. A second generation computer forensic analysis system[J]. Digital investigation,2009(6):34-42.

(上接第 157 页)

性、可区别性的特点。

4 结束语

代理签名是一种特殊的签名,它广泛地应用于电子商务、电子政务等领域。在文中,以 Schnorr 提出的基于身份的签名方案为基础,合理引入代理委托协议,提出一种基于身份代理签名方案。在双线性群中 DLP、CDHP 和 BDHP 问题难解性的假设下,本方案被证明是安全的。本方案满足代理签名方案所具有的基本特性,具有重要的实际应用价值。

参考文献:

- [1] Mambo M, Usuda K, Okamoto E. Proxy signatures for delegating signing operation[C]//Proceedings of the 3rd ACM Conference on Computer and Communication Security. [s. l.]: [s. n.], 1996:48-57.
- [2] 杨义先,钮心忻. 应用密码学[M]. 北京:北京邮电大学出版社,2005.
- [3] Shamir A. Identity-based cryptosystems and signature schemes[C]//LNCS196: Advances in Cryptology: Crypto '84. Berlin:Springer,1984:47-53.

- [3] Wang Shiuh-Jeng, Kao Da-Yu. Internet forensics on the basis of evidence gathering with Peep attacks[J]. Computer Standards & Interfaces,2007,29:423-429.
- [4] 张有东,曾庆凯,王建东. 网络协同取证计算研究[J]. 计算机学报,2010,33(3):504-512.
- [5] 国光明,洪晓光. 基于日志挖掘的计算机取证系统的分析与设计[J]. 计算机科学,2007,34(12):299-302.
- [6] Brueckner S, Guasparia D, Adelsteina F, et al. Automated computer forensics training in a virtualized environment[J]. Digital investigation,2008(5):105-111.
- [7] Sua Ming-Yang, Yub Gwo-Jong, Lin Chun-Yuen. A real-time network intrusion detection system for large-scale attacks based on an incremental mining approach[J]. Computers & Security,2009,28:301-309.
- [8] 林 英,张 雁,欧阳佳. 日志检测技术在计算机取证中的应用[J]. 计算机技术与发展,2010,20(6):254-256.
- [9] 赵 蹇,崔益民,邹 涛. Bloom filter 在网络取证中的应用研究[J]. 计算机工程与应用,2010,46(14):90-94.
- [10] 杨卫平,黄烟波,段丹青,等. 基于协议分析的网络入侵动态取证系统设计[J]. 计算机技术与发展,2006,16(4):215-217.
- [11] Thonnard O, Dacier M. A framework for attack patterns' discovery in honeynet data[J]. Digital investigation,2008(5):128-139.
- [12] 范 明. 数据挖掘:概念与技术[M]. 孟小峰,译. 北京:机械工业出版社,2003.

- [4] 杨 波,肖国镇. 现代密码学[M]. 第 2 版. 北京:清华大学出版社,2007:124-128,186-200.
- [5] 徐茂智,游 林. 信息安全与密码学[M]. 北京:清华大学出版社,2007:178-197.
- [6] Boneh D, Franklin M. Identity-based encryption from the weil pairing[C]//LNCS 2139: Advances in Cryptology, Crypto 2001. Berlin:Springer,2001:213-229.
- [7] Boneh D, Lynn B, Shacham H. Short signature from the weil pairing[C]//LNCS 2248: Advances in Cryptology, Asiacrypt 2001. Berlin:Springer,2001:514-532.
- [8] Paterson K. ID-based signatures from pairing on elliptic curves. [EB/OL]. 2002. <http://eprint.iacr.org>.
- [9] 蔡光兴,陈 华. 高效的基于 ID 的无可信中心签名方案[J]. 计算机应用研究,2009,26(7):2752-2753.
- [10] 周 亮,李大鹏,杨义先. 基于身份的无需可信 PKG 的签名方案[J]. 通信学报,2008,29(6):9-11.
- [11] Stinson D R. 密码学原理与实践[M]. 第 2 版. 冯登国译. 北京:电子工业出版社,2003:233-261.
- [12] 王泽成. 基于身份的代理签名和盲签名[J]. 计算机工程与应用,2003,39(23):148-150.
- [13] 李 沛,王天芹,潘美姬. 基于身份的签名方案[J]. 计算机工程与应用,2008,44(14):103-106.