恶意软件鉴别技术及其应用

庄蔚蔚[1,2] 姜青山[2]
[1](厦门大学信息科学与技术学院 厦门 361005)
[2](中国科学院深圳先进技术研究院 深圳 518055)

摘 要 随着互联网技术的发展和安全形势的变化，恶意软件的数量呈指数级增长，恶意软件的变种更是层出不穷,传统的鉴别方法已经不能及时有效的处理这种海量数据，这使得以客户端为战场的传统查杀与防御模式不能适应新的安全需求，各大安全厂商开始构建各自的“云安全”计划。在这种大背景下，研究恶意软件检测关键技术是非常必要的。针对恶意软件数量大、变化快、维度高与干扰多的问题，我们研究云计算环境下的软件行为鉴别技术，探讨海量软件样本数据挖掘新方法、事件序列簇类模式挖掘新模型和算法及在恶意软件鉴别中的应用，并构建面向云安全的恶意软件智能鉴别系统原型以及中文钓鱼网站检测系统架构。

关键词 恶意软件鉴别；数据挖掘；特征表征；模型构建；分类集成；事件序列挖掘

Malware Identification Technique and its Applications

ZHUANG Wei-wei [1,2] JIANG Qing-shan [2]
[1](*Department of Cognitive Science, Xiamen University Xiamen* 361005)
[2](*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences Shenzhen* 518055)

Abstract With the development of the Internet technology and the changes of the situation of Internet security, we witness exponential increase of the number of malicious software and their endless variants. Traditional detection methods cannot effectively and timely deal with such mass of malicious software data, making traditional anti-virus platform running on PC client cannot satisfy current security requirements any more, thus some major Internet security venders have been launching their ‘cloud security’ program. Under such background, it is urgent to develop some new effective and efficient techniques for malware detection. In this paper, we investigate malware detection techniques based on cloud computing, including mining massive software samples, and applying new clustering models/algorithms for event sequences into malware detection, to deal with the critical issues of malware as being of large amount, fast change, high-dimension and noise-laden. Furthermore, we propose a prototype of intelligent malware detection system for cloud security.

Keywords malicious software identification; data mining; feature representation; model construction; classifier ensemble; event sequence mining

1 引 言

随着计算机技术的发展及互联网的普及，人们的生活越来越依赖于互联网：即时通信、电子邮件、电子商务、网络游戏、网上办公等与人们的日常生活息息相关。然而，在享受信息时代带来的各种好处的同时，也伴随着各种安全问题的困扰，而恶意软件（malware）作为信息安全中最主要的问题，已经对互联网造成了不可忽视的危害[1, 4, 5]：家庭及企业的计算机遭受黑客的非法入侵，为用户带来意想不到的损失；网上银行帐号被木马非法盗取，使互联网用户

基金项目：国家自然科学基金（面向软件行为鉴别的事件序列挖掘方法研究，NO.61175123）；深圳市生物、互联网、新能源产业发展专项资金（NO.CXB201005250021A）。**庄蔚蔚**，博士研究生，研究方向为数据挖掘，网络安全。**姜青山**，教授，博士生导师，研究方向为数据挖掘，信息安全。E-mail：qs.jiang@siat.ac.cn。

蒙受巨大的经济损失；通讯帐号及密码被恶意获取，威胁网民隐私；通过即时通讯工具接收或传输的文件，都有可能是恶意软件。通过四通八达、无所不及的互联网，每一个使用计算机和网络的人们，都可能成为受害者，也有可能成为继续传播恶意软件的施害者，种种危害不胜枚举。恶意软件的地下经济体系也已经非常成熟，恶意软件的生产速度几乎达到了安全厂商检测和分析能力的极限[1,5,16]。

恶意软件是指具有恶意功能的程序，如对计算机系统的安全构成威胁、破坏计算机系统或者在未经计算机用户授权许可的情况下获取计算机系统的敏感信息。根据制作目的和传播方法的不同，恶意软件分为以下几类[5,16]：计算机病毒、蠕虫、木马、后门、内核级、恶意移动代码、间谍软件、僵尸网络、广告软件和钓鱼网站。其中钓鱼网站作为恶意软件的一种新的表现形式，在近几年频繁出现，目前其导致的国内网民损失已达76亿元，严重影响了在线金融服务、电子商务的发展，危害公众利益，影响公众应用互联网的信心。事实上，对恶意软件类别的界定并不是严格的。在具体应用中，各类恶意软件的合理、综合应用，使得当前恶意软件形式日益多样化，功能日益强大。

传统的恶意软件检测方法主要包括特征码鉴别、启发式查杀、数字签名以及虚拟机技术等。特征码技术只能查找已知的、被彻底研究过的恶意软件，不能检测变形及未知的恶意代码，面对不断出现的恶意软件的变种或新型恶意程序，必须不断更新版本；启发式查杀比较依赖于专家经验，具有比较高的误报率，而且在不断更新和升级的操作系统环境下，不断升级的恶意软件始终与启发式查杀的对抗性在同一个层次上面，甚至超前；对于数字签名检测技术，如果一些具有颁发数字签名证书资格的机构不严加管理，有可能被木马作者恶意利用；虚拟机检测技术效率比较低，资源占用较大，很多恶意软件需要在一定条件下才能被触发执行。面对恶意软件爆发式的增长，对于收集的海量未知文件，经上述传统检测技术分析鉴定后，仍然存留数量相当大的无法识别的未知样本。传统的以客户端为战场的恶意软件鉴别技术已无法应对层出不穷和频繁变种的恶意软件。依据软件行为的“恶意性”，比如是否感染其它软件、窃取用户账号/密码信息等，对软件进行自动分类(正常软件/恶意软件两类)进而辨别其中的恶意软件成为保障互联网时代计算机安全的一种迫切需要[3,5,16]。在这个背景下，面向云安全的恶意软件鉴别应运而生，它基于云计算环境，利用大量客户端对网络中各种类型软件的行为进行监测，从而获取互联网中病毒、木马等恶意软件的最新信息，传送到云端进行自动分析和处理，再把病毒和木马的解决方案分发到每一个客户端。研究云计算环境下的安全技术、恶意软件鉴别理论方法是新形势下保障信息安全和互联网安全的迫切需要，具有广泛的应用前景。

研究恶意软件智能鉴别的关键技术，实现对海量未知软件行为进行自动、快速、准确的识别及分析处理是信息安全领域一项重要而又紧迫的研究内容。针对互联网安全的现状，作者及其团队近六年来一直致力于恶意软件检测新方法和新技术的研究[2,21,22,26]。为了解决云计算环境下海量软件行为自动识别和处理的难点问题，我们研究了海量软件行为的特征表征和选择方法、特征分类算法和分类集成学习技术、子空间聚类和病毒自动归类等恶意软件检测关键技术[3,8,9,12,13,15,21-26]。在已有的研究基础上，我们构建了面向云安全的软件行为智能鉴别系统原型，并研究钓鱼网站智能检测与防御技术以及系统架构。

2 传统恶意软件检测方法

本节介绍恶意软件的类型及其特点，并阐述传统恶意软件的检测流程及常见的恶意软件鉴别方法，最后概述传统检测方法在海量恶意软件鉴别中存在的问题。

2.1 恶意软件类型及其特点

恶意软件是指具有恶意功能的程序，如对计算机系统的安全构成威胁、破坏计算机系统或者在未经计算机用户授权许可的情况下获取计算机系统的敏感信息。根据制作目的和传播方法的不同，恶意软件包括以下几类[1,2,4,5]：

（1）病毒:依附于宿主文件，在宿主文件被执行条件下四处传播完成特定业务功能的一类恶意软件；

（2）蠕虫:具有自我繁殖能力，无需用户干预便可自动通过网络传播的一类恶意软件；

（3）木马:在正常程序掩盖下执行恶意程序以欺骗用户，并且受网络的另一端控制的一类恶意软件；

（4）后门:运行在目标系统中，通过某种特殊方式对目标系统未经授权的远程控制；

（5）内核级:指用于帮助入侵者在获取目标主机管理员权限后，尽可能长久地维持这种管理员权限的一类恶意软件。RootKit的作用是要尽可能长久地维

持对目标系统的远程控制；

（6）恶意移动代码:一类从远程系统下载并以最小限度调用或以不需用户介入的形式在本地执行的恶意软件；

（7）间谍软件:一种未经用户同意安装在个人电脑中用于暗中访问、拦截数据或获取计算机控制权的恶意软件；

（8）僵尸网络:僵尸网络是指采用一种或多种传播手段，将大量主机感染僵尸程序，从而在控制者和被感染主机之间形成的一个可一对多控制的网络；

（9）广告软件:指未经用户允许，下载并安装或与其他软件捆绑通过弹出广告或其他形式进行广告宣传的程序；

（10）钓鱼网站:引诱用户到一个通过精心设计与目标组织网站非常相似的钓鱼网站上，并获取用户在此网站上输入的个人敏感信息或骗取用户汇款。

2.2　恶意软件检测的一般流程

传统的恶意软件检测流程可归纳为两个阶段[2,4]：模型生成阶段与预测阶段，具体流程如图1所示：

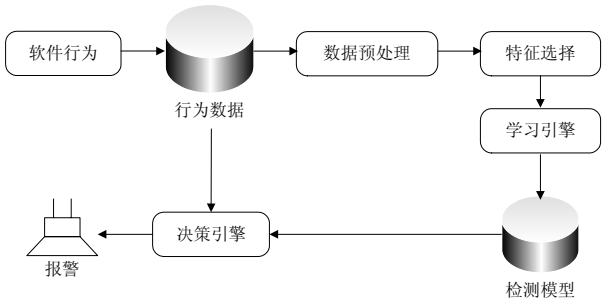

图1　传统的恶意软件检测流程

训练过程首先从已知的恶意软件和正常文件中抽取相应的特征来表示样本，然后将特征数据存储到数据库中，再通过相应的学习算法构建分类模型。模型在初次生成之后，还可以通过一些专家经验对模型中不够准确的部分进行修正，以提高检测精度。而对于未知文件，同样经过特征提取选择后，结合决策引擎及生成的检测模型对样本进行预测，判断出文件类别，并给出预测结果。

2.3　常见的恶意软件检测方法

当前恶意代码的分析方法有多种类型，一般可以分成基于代码特征的分析方法、基于代码语义的分析方法和基于代码行为的分析方法三种[4]。目前大部分反恶意代码软件所用的自动检测方法包括以下几种：

（1）基于特征码的检测法[19,20]:通过密罐系统提取恶意代码的样本分析它们独有的特征指令序列，当反病毒软件扫描文件时，将当前的文件与病毒特征码库进行对比，判断是否有文件片段与已知特征码匹配；

（2）启发式检测法[2,5]:为恶意代码的特征设定一个阈值，扫描器分析文件时，如果文件的总权值超出了设定值就将其看作是恶意代码；

（3）虚拟机技术[2,6]:用程序代码虚拟CPU寄存器，甚至硬件端口，用调试程序调入可疑恶意程序样本，将每个语句放到虚拟环境中执行，这样就可以通过内存、寄存器以及端口的变化来了解程序的执行，使变形病毒特别是加密病毒自动解密露出原形，而不用研究各个恶意程序的解密算法及密钥；

（4）数字签名检测方法[5,6]:正规软件开发者通常会对软件代码进行数字签名，证明软件没被非法篡改且来源可信，因此杀毒软件可以将具有真实数字签名的文件列入“白名单”；

（5）云查杀技术[5,22,23]:融合了云计算、数据挖掘、恶意软件鉴别、隐私保护、数据安全、入侵行为检测以及安全防御等新兴技术和概念。

2.4　拟解决的关键技术问题

对于互联网安全企业而言，“云安全”计划的核心和难点在于如何对收集的海量样本进行自动、快速、准确地识别、分析和处理。因此，要实现“云安全”中恶意软件的快速检测，需解决的主要问题包括：

（1）有效的软件行为描述[2, 8, 25]：文件特征表达的好坏对于恶意软件检测效果具有决定性影响，为了实现服务端对未知样本准确、有效的识别，特征表达应该有效、自动、高效；

（2）机器学习算法及分类模型的构建[9,12,13]：根据已知的样本文件，在文件特定特征表述的基础上，需要分析数据集的特点，选择合适的分类学习方法构建有效的分类模型以实现对未知文件快速、准确的识别；

（3）高效的归类方法[15,21,22]：对检测到的恶意软件，需要有高效、准确的聚类方法对其进行自动归类，并对属于同一个簇的样本文件提取通用特征，以缩小置放于客户端恶意特征库的体积，达到通过轻量级客户端实现对恶意软件快速识别的目标；

（4）客户端与云端服务数据交互时的动态隐私保护问题[27, 28]：面向云安全的恶意软件鉴别的一个基础是需要客户端收集、传送大量的信息到服务端，如同其它基于云计算的应用一样，对隐私的担忧将会损害人们对于这种恶意软件鉴别服务的信任。

3 恶意软件智能鉴别技术

针对传统恶意软件检测方法中存在的问题和不足，我们研究团队近6年来开展了基于数据挖掘和机器学习的海量软件样本分析和恶意软件鉴别研究，并提出了若干种新的恶意软件鉴别方法[2,21,22,26]。主要包括：在[2，22,24]中将出现在程序文件中属同一字符集的连续字符序列作为软件的特征，提出了一种结合粗糙集与基于聚类的遗传算法的特征选择方法，构建了一个基于SVM分类器融合的恶意软件检测方法，较好地解决了软件行为特征解释性差、恶意软件检测效率低的问题；在[8,12,21,25]中采用面向对象的关联规则挖掘技术进行基于静态API特征的恶意软件样本分析，同时也构建了各种特征分类模型，用来对未知软件行为进行预测，其中包括：基于Win API函数的增量关联规则分类模型、基于字符串特征的支撑向量机分类模型、基于资源特征的决策树分类模型以及基于指令特征的朴素贝叶斯分类模型；在[13,22]中研发了一种智能评分系统用于辨别疑似恶意软件，包括从程序文件预处理、特征抽取到原型系统构建等新技术。在已有的研究基础上，本文中我们将重点探讨海量事件序列挖掘模型、数据隐私保护技术，以及面向云安全的软件行为智能鉴别系统原型及相关技术。

3.1 海量事件序列挖掘模型与隐私保护模型

针对云计算环境下的软件行为鉴别关键技术的研究，我们将提出大规模事件序列簇类模式挖掘新模型、在线隐私保护模型，构建信息安全领域的背景知识库，并应用于面向云安全的软件行为鉴别核心技术和系统研发中。

3.1.1 在线隐私保护建模和动态匿名化算法

与传统的基于静态数据的隐私模型相比，在线交互是动态的，具有时间属性。基于[28]中提出的一个建模思路：在传统的隐私模型上添加一个参数w，该参数描述以某一个时间点t为中心的时间窗口大小。该隐私理论模型对用户在某一个时间点t上线的时候，提供在有效时间窗口w内的隐私保护，这种保护是动态的、暂时的。随着时间推移，如果用户需要继续进行数交互，则需要更新模型，否则模型自动失效。具体的在线匿名模型定义如下：一个在线查询$<q,t>$被认为是拥有在线隐私（online anonymity），如果该查询满足如下条件：在以t为中心的宽度为w的时间窗口内$p(q)<\theta$（用户定义的一个最小隐私阈值）。这实际上定义了一个基于时间窗口的动态数据集上的隐私模型，适合于动态的在线环境。

在这个新定义的动态隐私保护模型中，云服务是潜在的隐私破坏者，因而在实现层面一定要有一个第三方来负责动态的为在线用户的匿名化（anonymizer），对用户将提交的隐私信息进行隐藏，提供隐私保护；同时，这个新引入的第三方匿名化，也是一个潜在隐私破坏者，因此我们需要设计一个新的在线隐私保护框架。借鉴社会学的一个理论：隐私问题只有在攻击者同时拥有用户身份信息(user identity)和行为信息(user behavior)的才会存在，单单知道用户身份，却不知道行为，或者反之都不会构成隐私泄露。根据这一思路，我们设计一个新的框架，如图2所示。在这个新框架中，客户端将用户的身份信息(d)和行为信息(q)分开，然后将身份信息发给Anoymizer进行匿名化($d\rightarrow d'$)，在得到匿名化的身份信息之后，再将匿名化后的身份信息连同用户行为信息$<d',q>$一同发到云端以获得相应服务。

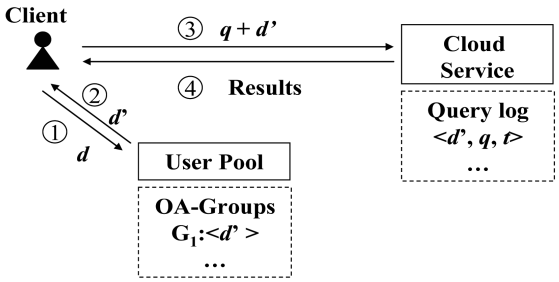

图2 在线隐私保护框架示意图

在此过程中，Aonymizer和服务提供商都没有同时拥有非匿名化的用户身份信息和用户行为信息，因此不存在隐私泄露问题。在此基础上，我们将解决在Anonymizer中如何对动态的用户信息进行匿名化的问题。这个问题和在线聚类有一定的相似性，因为匿名化的基本思路就是将用户某些特殊的特征信息及其组合进行泛化，我们通过在线聚类算法实现把一个用户藏到一个小的具有共同特性的用户组中的隐私保护服务。

3.1.2 软件行为的事件序列特征表示和提取方法

软件行为特征就是将软件运行时的行为或者部分重要的信息抽取出来，并使用格式化方法来表示这些特征数据。以计算机软件在执行过程中所调用的API函数序列来刻画软件的行为特征为例，这种序列数据可以抽象为一种(类属型)事件序列。预想的软件API行为序列提取过程分为三个步骤进行:首先，我们将基于云安全平台的集群计算机等，构建虚拟运行环境，使用DLL注入等手段捕获软件运行过程中调用的

有效的API函数序列；其次，根据基础研究阶段得出的领域背景知识，给一些特定的事件，比如修改操作系统注册表的操作、TCP网络连接等恶意软件为实施恶意行为需要依赖的API函数等，赋予较高的权重，结合无监督特征加权方法(如*idf*等)，从第一个步骤生成的原始序列中筛选保留具有最高权重的若干API函数组成新的序列。

第三个阶段也是最重要的一个阶段，我们将基于统计理论提取序列中重要的模式子序列集合，目的在于剔除序列中可能存在的噪声和随机子序列。在[29]中我们提出了一种无需序列对齐的重要模式子序列提取方法SCS，设X,Y表示两个序列，$E=\bigcup_{X,Y\in DB}E_{X,Y}$，其中 $E_{X,Y}$ 的定义如下：

$$E_{X,Y}=\left\{x,y\left|\begin{array}{c}|x|=|y|\\|x\cap y|>N_{X,Y}\\|x\setminus y|<N_{X,Y}\\\forall x',y'\in E_{X,Y}\Rightarrow(x\not\subset x')\vee(y\not\subset y')\end{array}\right.\right\}\quad(1)$$

其中x,y分别表示包含在X,Y中的子序列，$N_{X,Y}$为最小序列匹配长度(事件数目)。生成$E_{X,Y}$的过程并不严格要求在$N_{X,Y}$范围内的事件按照相同的顺序排列，且允许少量噪声存在，这一特性适用于提取软件行为特征模式子序列。许多恶意软件在传播过程中经常更改局部指令(对应于“事件”)顺序和引入“花指令”(导致子序列中的噪声)以逃避查杀，公式(1)的定义可适用于这种情况.通过将该方法进行改进，主要方面包括最小序列匹配长度参数的自动设置等，并保证提取的子序列具有统计意义上的高置信度。

3.1.3　海量事件序列挖掘模型及恶意软件分类和聚类算法

为了进行海量事件序列挖掘，一个简单、高效且能全面反映事件序列中重要模式子序列的信息以及其间蕴含的序关系信息的分析模型是必须的，这是因为，在一个有效的模型基础上，有关的分类、聚类算法等都可以更为方便地推导而来。因此，这里我们将基于海量事件序列挖掘的恶意软件鉴别方法划分为两个层次加以研究，分别着重于模型的建立和算法的设计上。

在模型的构造方面，我们提出一个类向量空间模型(VSM)作为事件序列簇类模式挖掘的数据模型，主要因为向量空间模型具有简单、易用、高效的优点，适用于海量数据的挖掘。为使向量空间模型能够处理特征间的顺序关系，我们首先将引入一阶HMM（hidden Markov models）假设，将特征(公式(1)中提取的重要模式子序列)的频度和条件概率(前后两个子序列的状态转移概率)相结合，定义新空间的属性的特征值。这种做法的另一个好处是，将事件序列/子序列的类属类型转换为数值型；其次，我们还将把子序列对之间的顺序关系通过一个约束矩阵附加到模型中，最终建立一个带序关系约束的类向量空间模型。恶意软件的分类和聚类算法将基于如上所述的数据分析模型来设计。由于我们所构建模型中向量的属性值已经转换为了数值型，因而可以使用常用的相似度度量(如欧式距离函数)衡量序列间的相似度，从这个角度说看，我们可以在现有的分类和聚类算法基础上开发恶意软件的鉴别算法。分类方法通过分析有标号的样本文件，将其中的特征或是知识结构抽取出来，构建分类模型，使用该模型对未知的文件进行预测，给出未知文件的分类类别，并且该模型还可以通过不断地学习，不断提高分类模型的有效性。与分类相比，聚类分析是无监督的学习过程，根据数据点之间的相似度度量，对数据集进行划分，其输出为数据点的簇类集合。

3.1.4　恶意软件行为事件序列特征及其约简方法

首先，我们将给每个簇类(也就是不同软件的类型：正常软件/恶意软件，或不同的恶意软件子类型)的行为特征一个形式化的描述，实际上，在(3.1.2)提出的软件行为事件序列分析模型中，软件的行为特征体现为带约束条件的事件子序列集合。从算法的角度看，事件子序列的集合对应于簇类所在特征子空间，因此，我们将通过对(3.1.3）分析的结果，也就是代表不同软件类型的簇类，进行子空间分析，以提取某种恶意软件最大程度区别于正常软件及其它类型恶意软件的特征子空间，即该型恶意软件的行为特征序列。其次，还需要对提取出来的行为特征序列进行“压缩”，以轻量化云安全客户端。我们将进行特征转换和特征选择这两种不同类型的方法进行特征约简，在特征转换方法中，拟采用SVD分解技术，这是因为SVD在文本挖掘等领域已经被证实了可以进行特征语义空间的变换。

3.1.5　面向云安全的软件行为鉴别技术应用

针对恶意软件频繁变种、数量暴涨的现状，以提高软件行为鉴别精度和降低恶意软件的误报率为目标，并有效解决客户端与云端动态数据交互过程中的隐私保护问题，研发基于云计算平台的软件行为鉴别技术实用系统以及在线匿名化隐私保护系统。本

部分研究的主要问题在于如何充分利用（3.1.1）—（3.1.4)提出的在线环境隐私模型、数据分析模型及挖掘方法的优势，解决在线动态隐私保护问题以及从海量软件样本中鉴别出恶意软件，是应对当前恶意软件层出不穷、频繁变种现状的一种现实选择，也必将是业界雄心勃勃的“云安全”计划的一条重要实现途径。基于云计算平台的软件行为鉴别系统结构如图3所示：

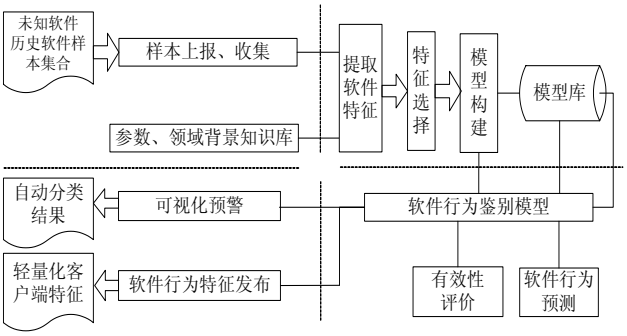

图3 恶意软件智能检测系统结构

对于进入检测流程的未知样本文件：（1）首先经过特征提取与选择；（2）然后经过四种不同特征分类器集成进行预测；（3）对这些基础分类器的预测结果采用异构分类器集成学习方法进行结论生成；（4）经上述预测环节，对于预测结果为“恶意软件”的样本文件经聚类模块进行恶意软件的归类，然后对于每一个恶意软件“簇”提取相应的通用特征发布到客户端；对于预测结果为“未知文件”的样本文件分流到人工鉴定环节。

3.2 恶意软件智能鉴别系统架构

在深入分析恶意软件行为的动态事件序列特征和隐私数据保护、海量数据挖掘特点的基础上，基于近年来我们在隐私保护研究、恶意软件鉴别技术研究、数据挖掘算法及软件系统方面的研究成果和经验，研发云安全平台中的恶意软件鉴别技术及实用系统。我们首先提出了面向云安全的恶意软件鉴别系统架构如图4所示。云安全检测将行为拦截放在云端完成，这样可以保证检测模型对恶意软件作者完全不可见，很难采取措施绕过或对抗。云安全服务端上部署多种以往无法在客户端上实现的、采用数据挖掘方法构建的智能鉴定、聚类系统，这些系统并行，既能相互纠错，而且随时可以不断增加新的恶意软件检测方法，以提升恶意软件检测的准确性。

该系统主要包括以下几个部分：

（1）终端用户：客户端负责终端用户文件扫描、用户进程扫描、内存扫描、漏洞扫描、用户访问

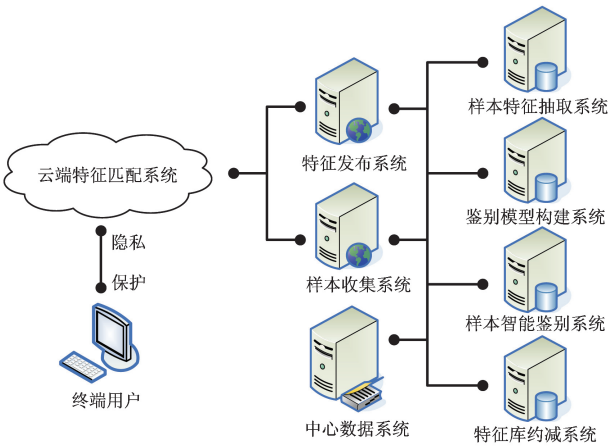

图4 可信云安全的系统架构

网页扫描等软件行为扫描，同时客户端应具备强杀、自保护、信息反馈、自动拦截以及应急处理等功能；

（2）云端特征匹配系统：客户端将行为特征上传至云端特征匹配系统进行特征匹配、鉴别，并反馈结果；对于无法鉴别的未知软件，将其上报至样本收集系统；

（3）特征发布系统：将研究的聚类和聚类融合的方法有效应用于恶意软件的自动归类中，对相应的恶意软件“簇”提取通用的特征，以缩小特征库的体积，并通过特征发布系统进行发布，云端特征匹配系统可从此取出特征进行匹配；

（4）样本收集系统：负责未知样本收集，以及将样本信息送至基于云计算基础平台构建的特征抽取集群系统、特征约减集群系统、模型构建集群系统、智能鉴别集群系统；

（5）中心数据系统:主要负责监控各个系统、集群系统的运维与数据状况，统计与发布系统数据信息；

（6）隐私保护技术：在客户端与云端的交互过程中，将研究的隐私保护模型用于解决动态数据交互的隐私保护问题；

（7）在云计算基础平台进行构建，部署多种以往无法在客户端上实现的、采用数据挖掘方法构建的分布式的特征抽取、特征库约减、智能模型构建以及智能鉴别系统，这些系统并行，既能相互纠错，而且随时可以不断增加新的恶意软件行为的鉴别方法，以提升恶意软件行为鉴别的准确性。包括：

①样本特征抽取系统：基于（3.1.2）软件行为的特征表示和提取方法，提取软件的动态行为事件序列特征；②鉴别模型构建系统：基于（3.1.3）数据挖掘模型及恶意软件分类和聚类算法，构建软件行为

鉴别模型；③样本智能鉴别系统：基于鉴别模型构建系统产生的鉴别模型对软件行为进行预测，以鉴别其是恶意软件还是正常软件；④特征库约减系统：基于（3.1.4）事件序列特征及其约简方法，对软件行为特征库进行压缩，然后将其送往特征发布系统。

云端对未知文件的智能鉴别系统的工作流程主要包括：未知样本软件通过数据输入、客户端与云端的通信模块将其传至云端；云端通过可疑文件、安全信息收集模块接收到文件后送入常规分析流程，利用动态行为序列检测抽取出动态行为事件序列特征；初步分析完毕后如果无法鉴别，就将特征送入样本智能鉴别模块利用数据挖掘鉴别模型、虚拟机等部署在云端的分布式的智能鉴别系统进行鉴定，并利用智能融合方法融合各个子鉴别系统的结果给出最终的判定结果；如果上述方法还是无法鉴别，则送入人工鉴别流程；最终，通过结果评估与反馈模块将特征保存到特征数据库，并将结果传回到客户端展示以及做相应的处理操作。

4　钓鱼网站智能检测与防御技术

钓鱼网站作为恶意软件的一种新的表现形式，在近几年频繁出现。例如一些仿银行、中奖类以及仿订票的站点，其严重影响在线金融服务、电子商务的发展，危害公众利益，影响公众应用互联网的信心。钓鱼网站具有恶意软件行为，可以使用恶意软件行为鉴定的方法对其进行识别与防御，但它又不同于传统的病毒，这给传统病毒查杀方法带来了巨大的挑战。

网页内容作为钓鱼欺骗信息的主要展示渠道，对钓鱼者意图具有较强的表达能力。为了弥补传统检测方法的不足，更有效的应对数量不断增长的未知钓鱼网站，不少研究学者开始结合各种页面元素作为特征进行钓鱼检测，通过定义和提取各种页面特征，采用各种机器学习方法进行建模来对钓鱼页面进行检测[17]。在页面特征表征方面，当前研究主要是利用部分页面元素，例如网页标题，Form表单，包含的链接等元素进行表征，部分研究还对页面Logo图标及包含的图片进行识别；在机器学习方法上，决策树、支持向量机、贝叶斯、神经网络等方法被应用于模型的训练和预测中。特征的表征方式与机器学习方法是其中的关键，因为钓鱼者可能试图通过增加歧义干扰词等方式构造特殊的页面来躲避识别系统的特征提取与最终的检测。此外，目前大部分研究均是在小数据集上进行的实验，算法的健壮性和实用性未能在真实的大数据集上得以验证，且当前大部分研究都是针对英文钓鱼站点的检测，基于英文特征识别的方法对中文钓鱼网站并不适用。

针对传统中文反钓鱼检测方法的不足，我们开展了以下工作[18]：

（1）实现了有效的钓鱼行为鉴别方法：提出一种新的基于分类融合的检测方法，根据基础特征分类器之间的关系，结合分类融合思想对钓鱼行为进行鉴别；

（2）提高检测准确率并降低误报率：研究不同特征对分类结果的不同影响，分别构建基于不同特征的异构分类模型，通过特征间的历史鉴定关系融合各个模型的鉴定结果，提高检测的准确率，并降低误报率；

（3）解决中文钓鱼网站的检测防御问题：针对中文钓鱼网站进行深入分析，研究有效的页面特征表达方式与特征选择算法，构建模型训练所需的各种节点信息，采用分类融合技术实现中文钓鱼网站的智能检测与防御。

4.1　智能反钓鱼系统架构

本文通过解析已知正常网站和钓鱼网站对应的网页内容，提取其相应的八种页面特征，并基于不同的特征表达构建相应分类器。对待检测网站，采用分类融合方法综合各个分类模型的预测结果，达到对钓鱼网站智能检测的目标。基于已有的研究工作，本文着重讨论钓鱼网站智能检测与防御系统IAPS(intelligent anti phishing system)架构如图5所示。

系统分为训练和预测两个部分：训练指基于大量已知正常网站和钓鱼网站的网页内容，提取其相应的网页标题、网页关键字、网页描述信息等特征构建相应的分类器。而预测则主要是对于待检测的网站，采用分类融合的方法，综合各个分类器的预测结果，实现对未知网站的钓鱼检测。

（1）特征提取模块

IAPS检测系统通过特征提取模块解析网站对应的页面内容并提取相应的页面特征，主要包括以下页面内容：Title（标题标签内容）,Keyword（Meta标签中提取的关键字信息），Description（Meta标签中提取的页面描述性信息），Copyright（Meta标签中提取的版权信息），Frame（页面中包含框架页面的URL地址），Img（图片链接URL地址），Alt（Alt标

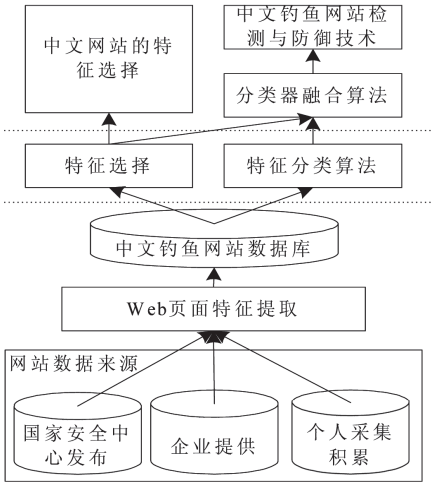

图5 智能反钓鱼检测防御系统研究架构图

签中的内容）和String（脚本中字符串和页面所有可见字符串集合）。

（2）特征分类模型构建

String特征经过分词后相比其它特征包含了更多的信息，维度高，是一种信息表达较强的特征，由于SVM对维度较高的数据集有较强的处理能力和较好的分类正确率，我们采用SVM分类算法进行分类器的构建。而标题、关键字、版权等其它特征，包含的信息相对较少，这里采用NBC对其构建分类模型，NBC算法简单，模型所需估计的参数很少，适用于这些特征的模型构建。朴素贝叶斯公式扩展如下[18]：

$$V'_{NBC}=\sum_{i=1}^{n}\log\frac{C(X_i,phishing)+1}{C(X_i,benign)+1}*\frac{C(benign)+1}{C(phishing)+1}+\log\frac{C(phishing)+1}{C(benign)+1} \quad (2)$$

其中n表示某个特征（8种页面特征）分词后词的个数，X_i表示第i个词，$C(X_i,phishing)$与$C(X_i,benign)$分别表示词X_i在训练的钓鱼网站和正常网站中出现的次数，$C(phishing)$与$C(benign)$表示训练集合中钓鱼网站和正常网站的个数。在式(2)中，采用加1平滑的方法避免概率为0的情况。

（3）特征分类器融合模块

考虑到钓鱼网站周期短、更新快、变化多等特点，本研究在每类特征上选择与当前数据分布最相似的N个分类模型进行样本预测，采用文献[13,18,22]中提出的结论生成方法CE（correlation based ensemble），通过历史鉴定结果中基础特征分类器两两之间的关系，对8个特征分类器得到结果进行进一步加权融合来提高检测的准确率和召回率。

4.2 实验结果与分析

在我们的实验验证中，通过收集的100,000个正常网站和100,000个钓鱼网站进行训练，同时以客户端安全产品连续4天上报的网站作为测试数据。

（1）不同基础分类器检测结果比较

表1显示了各特征分类器对4天上报数据的检测情况。

表1 各特征分类器鉴定情况

分类器	8/18		8/19		8/20		8/21	
	precision	recall	precision	recall	precision	recall	precision	recall
Title	91.41%	35.33%	86.35%	29.98%	83.85%	26.96%	78.89%	27.83%
Keyword	88.54%	21.45%	88.32%	15.24%	96.15%	18.44%	92.87%	13.85%
Description	92.06%	10.91%	93.94%	9.79%	96.70%	7.76%	93.94%	8.10%
Copyright	97.32%	5.04%	66.75%	12.04%	59.71%	7.05%	98.43%	3.05%
Alt	73.36%	5.74%	67.26%	3.76%	73.73%	5.28%	68.84%	3.51%
Frame	98.33%	2.50%	98.22%	2.39%	98.15%	3.38%	97.14%	1.86%
Img	85.64%	12.76%	92.12%	16.25%	87.38%	15.44%	88.44%	15.76%
String	81.46%	53.31%	85.45%	61.73%	89.74%	52.41%	91.35%	71.59%

从各个鉴定器对4天上报样本的检测结果来看，部分特征的检测准确率较高，如Title、Keyword和Description然而大部分特征分类器的检查召回率都普遍较低。可见，样本是由多个特征组成的，各个特征对分类的影响程度不同，单独采用某个特征都无法完整的表述整个样本，需要结合各个特征的优势对样本进行综合鉴定。

（2）不同分类融合算法的结果比较

在这部分实验中，进一步把各个分类器的鉴定结果进行融合，融合采用了常见的多数投票法（majority vote）、简单加权法（simple weight vote）、贝叶斯投票法（bayes vote）及CE融合方法，将融合后的检测结果与单个独立的特征分类器进行比较，并对这些融合方法进行比较。实验结果如表2所示. 对比表2可以看出，相比单个特征分类器，采用各种分类融合方法后模型对样本的整体识别能力提高了，检测召回率有很大幅度的提高，覆盖了绝大部分的钓鱼样本。此外，利用集成中基础分类器两两之间的关系，对各特征分类模型的权重进行进一步调整，CE融合方法在准确率和召回率上都要优于其它方法。

表2 分类融合鉴定情况

融合方法	8/18		8/19		8/20		8/21	
	precision	recall	precision	recall	precision	recall	precision	recall
多数投票	96.61%	91.35%	95.66%	93.45%	95.76%	94.38%	94.83 %	93.31%
简单加权	96.98%	95.14%	96.21%	95.43%	96.42%	96.13%	95.48%	95.42%
贝叶斯投票	96.86%	94.31%	95.92%	94.42%	96.34%	95.56%	95.18%	94.30%
CE融合	98.72%	97.84%	98.05%	98.76%	97.33%	97.31%	97.59%	98.53%

（3）与常见反钓鱼工具的比较分析

本节中将把IAPS系统的检测结果同Kaspersky，Mcafee SiteAdvisor和Netcraft三款流行的反钓鱼工具进行比较分析，验证IAPS系统对钓鱼网站的实际检测效果。同样以4天内真实上报的所有网址作为测试样本，实验结果如下：

表3 反钓鱼检查结果比较

Day	Kaspersky		Mcafee		Netcraft		IAPS	
	precision	recall	precision	recall	precision	recall	precision	recall
8/18	97.01%	1.77%	92.75%	4.36%	76.35%	30.95%	98.72%	97.94%
8/19	94.75%	1.82%	90.12%	6.43%	77.25%	24.21%	98.12%	94.88%
8/20	92.88%	1.80%	89.21%	4.45%	75.33%	29.30%	97.34%	95.86%
8/21	93.51%	1.73%	88.45%	4.53%	76.48%	31.84%	97.55%	97.95%

从表3的实验结果可以看出：IAPS系统具有最高的检测准确率，同时相比其它反钓鱼工具能够识别出更多的钓鱼网点。目前IAPS系统其每天要对50,000左右的未知网站进行鉴定（其中钓鱼网站的比例约为1%），IAPS系统能够较全面的检测出这些钓鱼网站。

5 结　语

本文从分析互联网安全的现状入手，剖析了当前恶意软件检测的研究现状，阐述了“云安全”计划面临的难点和需要解决的主要问题。针对恶意软件数量大、变化快、维度高与干扰多的问题，我们研究了云计算环境下的软件特征的表达与抽取方法、海量软件样本数据分析新方法、事件序列簇类模式挖掘模型、隐私保护建模和动态匿名化算法、恶意软件分类和聚类算法，并构建面了向云安全的恶意软件智能鉴别系统原型，最后，进一步利用恶意软件智能鉴别技术来解决钓鱼网站这种新形式恶意软件的智能检测与防御问题，具有重要的理论意义和实际的应用价值。

参 考 文 献

[1] 金山毒霸反恶意软件实验室. 2011年中国互联网安全情况整体分析 [R]. 2011.

[2] 赵新星. 面向云安全的恶意软件鉴别方法研究及其应用 [D]. 厦门:厦门大学硕士学位论文, 2011.

[3] Ye Y F, Li T, Zhu S H, et al. Combining file content and file relations for cloud based malware detection [C] //KDD. 2011: 222-230.

[4] Idika N, Mathur A P. A survey of malware detection techniques [R]. West Lafayette: Department of Computer Science, Purdue University, 2007: 3-10.

[5] 国家计算机病毒应急处理中心. 2011年中国计算机病毒疫情调查技术分析报告 [R]. 2011.

[6] Filiol E, Jacob, Liard G, et al. Evaluation methodology and theoretical model for antiviral behavioural [J]. Journal in Computer Virology, 2006, 3(1): 23-37.

[7] Oberheide J, Cooke E, Jahanian F. CloudAV: n-version antivirus in the network cloud [C] //17th USENIX Security Symposium. 2008: 91-106.

[8] Ye Y F, Li T, Huang K, et al. Hierarchical associative classifier (HAC) for malware detection from the large and imbalanced gray list [J], Journal of Intelligent Information Systems, 2010, 35 (1): 1-20.

[9] Ye Y F, Wang D D, Li T, et al. IMDS: intelligent malware detection system [C] //Proceedings of ACM International Conference on Knowledge Discovery and Data Mining, 2007.

[10] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy [C] //IEEE Trans. Pattern Analysis and Machine Intelligence, 2005: 27.

[11] 范明, 李川. 在FP-树中挖掘频繁模式而不生成条件FP-树 [J]. 计算机研究与发展, 2003,8(40):1216-1222.

[12] Ye Y F, Li T, Jiang Q S, et al. CIMDS: adapting postprocessing techniques of associative classification for malware detection [C] //IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 2010, 40(3): 298-307.

[13] Ye Y, Li T, Jiang Q S, et al. Intelligent file scoring system for malware detection from the gray list [C] //Proceedings of ACM International Conference on Knowledge Discovery and Data

Mining. 2009.

[14] Quinlan J R. Simplifying decision trees. Int'1 [J]. Man-machine Studies, 1987, 3(27): 221-234.

[15] Ye Y F, Li T, Chen Y, et al. Automatic malware categorization using cluster ensemble [C] //CD-Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining. 2010: 95-104.

[16] 金山互联网安全公司. 金山可信云安全:能体验到的全面云安全 [Z].

[17] Zhang , Liu, Chow. Textual and visual content-based anti-phishing: a bayesian approach [J]. IEEE TRANSACTIONS ON NEURAL NETWORKS. 2011, 22(10): 1532-1546.

[18] 庄蔚蔚, 叶艳芳, 李涛, 等. 基于分类集成的钓鱼网站智能检测系统 [J]. 系统工程理论实践. 2011, 31(10): 2008-2020.

[19] Filiol E. Malware pattern scanning schemes secure against blackb3ox analysis [J]. J. Comput. Virol, 2006, 2(1): 35-50.

[20] Filiol E, Jacob, et al. Evaluation methodology and theoretical model for antiviral behavioural detection strategies [J]. J. Comput.Virol. 2007, 3(1): 27-37.

[21] 黄锴. 基于数据挖掘技术的恶意软件检测方法研究及其应用 [D]. 厦门:厦门大学硕士学位论文, 2010.

[22] 叶艳芳. 恶意软件智能检测若干方法的研究及其应用 [D]. 厦门:厦门大学博士学位论文, 2010.

[23] Jiang Q S, Zhao X X, Huang K. A feature selection method for malware detection [C] //2011 IEEE International Conference on Information and Automation. 2011: 890-895.

[24] Ye Y F, Chen L F, Wang D D, et al. SBMDS: a behavioral string based malware detection system using svm ensemble with bagging [J]. Journal in Computer Virology, 2008.

[25] Ye Y F, Wang D D, Li T, et al. An intelligent pe-malware detection system based on association mining [J]. Journal in Computer Virology, 2008,.4(4): 323-334.

[26] 庄蔚蔚. 基于增量学习关联分类规则的病毒检测方法研究 [D]. 厦门:厦门大学硕士学位论文, 2009.

[27] Xu J, Sung A, Chavez P, et al. Polymorphic malicious executable scanner by API sequence analysis [C] //Proceedings of the International Conference on Hybrid Intelligent Systems. 2008: 378-383.

[28] Xu Y, Wang K, Yang G, et al. Online anonymity for personalized web services [C] //Proceedings of the Conf. on Information Knowledge Management (CIKM). 2009.

[29] Kelil A, Wang S, Jiang Q S, et al. A general measure of similarity for categorical sequences [J]. Knowledge and Information System (KIS), 2010, 24(2): 197-220.