第 31 卷第 10 期 系统工程理论与实践 Vol.31, No.10
2011年 10 月 Systems Engineering — Theory & Practice Oct., 2011

文章编号: 1000-6788(2011)10-2008-13 中图分类号: TP393.08 文献标志码: A

基于分类集成的钓鱼网站智能检测系统

庄蔚蔚[1], 叶艳芳[2], 李 涛[3], 姜青山[4]

(1. 厦门大学 智能科学系, 厦门 361005; 2. 金山互联网安全公司, 珠海 519015;
3. 美国佛罗里达国际大学, 迈阿密 33199; 4. 中国科学院 深圳先进技术研究院, 深圳 518055)

摘 要 近来, 通过仿冒真实网站的 URL 地址及其页面内容的 "钓鱼网站" 已严重威胁到互联网用户的隐私和财产安全. 为了应对这种威胁, 该文通过对大量已知正常网站和钓鱼网站的学习, 解析其对应的网页内容, 提取相应的网页标题、网页关键字、网页描述信息等 8 种特征来描述这些网站, 然后基于不同的特征表达方法构建了相应的分类器; 对于待检测的网站, 采用分类集成的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. 基于上述方法, 构建了钓鱼网站智能检测系统 IPWDS, 并将其集成于金山安全产品中. 在大量、真实数据集的基础上, 实验结果表明 IPWDS 系统对钓鱼网站的检测效果优于现有常见的钓鱼网站检测方法和常用的反钓鱼软件.

关键词 钓鱼网站; 分类器; 分类集成

Intelligent phishing website detection using classification ensemble

ZHUANG Wei-wei[1], YE Yan-fang[2], LI Tao[3], JIANG Qing-shan[4]

(1. Department of Cognitive Science, Xiamen University, Xiamen 361005, China;
2. Kingsoft Internet Security Corporation, Zhuhai 519015, China;
3. Florida International University Miami, FL 33199, USA;
4. Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China)

Abstract By counterfeiting the real URL address and the actual page content, phishing websites have been a serious threat to the Internet user's privacy and property. In this paper, the authors propose an automatic method for intelligent phishing website detection through learning from a large number of normal and phishing websites. In particular, given a website, the authors first parse and analyze its webpage content and extract 8 different types of features such as title, keywords and description information to represent the website. Classifiers are then built based on these different feature representations. Finally classification ensemble methods are used to combine the prediction results of individual classifiers together for phishing website detection. Using the proposed method, the authors developed an intelligent phishing website detection system IPWDS, which has already been integrated into the Kingsoft's security products. Experiments on real-world datasets demonstrate that IPWDS outperforms existing popular detection methods and commonly used anti-phishing software tools in phishing website detection.

Keywords phishing website; classifier; classification ensemble

1 引言

网络钓鱼 (phishing) 是通过大量发送声称来自于银行或其他知名机构的欺骗性垃圾邮件, 意图引诱收信人给出敏感信息的一种攻击方式 [1]. 最典型的网络钓鱼攻击是将收信人引诱到一个通过精心设计与目标组织的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, 通常这个攻击过程不会让受害者警觉 [2].

随着互联网应用的发展, 网络购物、网上交易越来越频繁, 钓鱼网站的数量也在急剧增长, 据中国反钓鱼

收稿日期: 2010-12-29
资助项目: 国家自然科学基金 (10771176); 广东省产学研重大科技专项 (2008A09030001)
作者简介: 庄蔚蔚 (1982–), 男, 博士研究生, 主要研究: 数据挖掘, 信息安全; 叶艳芳 (1981–), 女, 博士, 主要研究: 数据挖掘, 互联网安全; 李涛 (1975–), 男, 博士, 副教授, 主要研究: 数据挖掘, 机器学习; 姜青山 (1962–), 男, 博士, 教授, 主要研究: 数据挖掘.

网站联盟最新数据显示, 目前网络钓鱼导致网民的损失已达 76 亿元 [3]. 因此, 钓鱼网站的智能检测已成为网络安全领域最热门的话题之一 [4]. 为了应对网络钓鱼的威胁, 互联网厂商推出了一系列浏览器辅助工具 [5], 例如: eBay 提供了相应的浏览器插件, 可以有效防御对自身的仿冒; Google 推出了可以鉴别欺诈性网页的通用插件. 但是, 这些辅助插件对各种钓鱼网址的检测效果仍不尽人意 [6]. 因此, 迫切需要提出更有效的方法来抵御网络钓鱼的威胁.

目前有不少关于反钓鱼的研究 [7−10], 但多数集中在对英文钓鱼网站的识别, 对中文钓鱼网站检测的研究较少. 同时, 在网站的特征表征上, 多数研究以其 URL 地址、域名注册信息、网站排名信息等作为网站特征进行钓鱼网站的识别 [11], 以网页内容作为特征进行钓鱼网站检测的研究较少. 金山安全实验室对收集的大量钓鱼网站进行分析, 结果表明网页内容作为钓鱼欺骗的展示渠道, 对钓鱼者的意图具有较强的表达能力. 利用这些特征, 可以有效实现对钓鱼网站的智能检测. 因此, 本文通过对大量已知正常网站和钓鱼网站的学习, 解析其对应的网页内容, 提取相应的网页标题、网页关键字、网页描述信息等 8 种特征来描述这些网站, 然后基于不同的特征表达方法构建了相应的分类器; 对于待检测的网站, 采用分类集成的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. 基于上述方法, 本文构建了钓鱼网站智能检测系统 IPWDS, 并将其集成于金山安全产品中. 在大量、真实数据集的基础上, 实验结果表明采用本文所提出方法的钓鱼网站智能检测系统 IPWDS 对钓鱼网站的检测效果优于现有常见的钓鱼网站检测方法和常用的反钓鱼软件.

本文各章节安排如下. 第 2 节介绍目前反钓鱼研究的一些相关工作; 第 3 节描述了本文提出并实现的钓鱼网站智能检测系统 IPWDS 的系统架构; 第 4 节阐述了网站特征提取的方法、分类器的构造算法和分类集成的方法; 第 5 节通过与其它几种常用钓鱼检测方法进行比较, 验证本文提出方法及 IPWDS 系统对钓鱼网站的检测效果和性能; 第 6 节介绍 IPWDS 系统在实际中的应用, 并与常见的几种反钓鱼软件进行比较; 第 7 节对全文进行总结.

2 相关工作

为了防止网络钓鱼的攻击, 实现对钓鱼网站的有效检测, 近年来, 出现了许多以浏览器插件形式存在的反钓鱼软件, 如: Microsoft Phishing Filter[12]、Google safe Browsing[13]、Trustwatch[14]、Spof Guard[15]、Netcraft[16] 等, Kaspersky、Mcafee 等安全产品中也相继加入反钓鱼功能. 上述反钓鱼工具主要采用黑/白名单技术来识别钓鱼网站. 黑/白名单技术是指将所有已经发现的钓鱼站点和可信网站的 URL 记录到一个列表 (即黑/白名单) 中, 据此判断用户所访问的网站是否为钓鱼/安全网站. 这种技术实现简单, 但黑名单的及时更新十分困难, 因为钓鱼网站的平均在线时间仅有 4 天 [17].

为克服黑/白名单技术的不足, 研究领域开展了反钓鱼的相关工作, 机器学习的方法也被引入其中 [18−20]. 这些研究主要集中在如何对钓鱼的行为进行表征, 以及如何更有效的在低误报率前提下检测出更多的钓鱼网站.

Kang 等人 [21] 在白名单技术的基础上, 通过 URL 相似度计算以及对 DNS 域名欺骗的判断, 有效防止了混淆 URL 和修改 DNS 映射的钓鱼攻击. 但由于该类钓鱼所占比例极少, 大大限制了该方法在实际中的应用. Kim 等人 [11] 则通过对 GOOGLE PAGERANK、WHOIS 等第三方库的查询, 获取网站排名、注册时间、存活周期等信息作为页面特征, 并对各种特征赋予不同权重, 定义不同的风险级别, 通过计算页面综合风险指数来判定是否钓鱼. 由于大部分钓鱼网站存活周期短, 访问频率低, 能够被该方法识别. 但也因此, 该方法对大部分新增和访问频度较低的页面容易造成误报.

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, 不少研究开始结合各种页面元素作为特征进行钓鱼检测. Cao 等 [22] 针对账号欺骗的钓鱼网站进行分析, 采用贝叶斯分类算法来识别用户提交登录过程的合法性, 有效阻止钓鱼网站盗号欺骗. Layton 等人 [23] 通过计算钓鱼网站与正常网站间的差异程度进行钓鱼检测, 防御仿冒品牌网站的钓鱼攻击. 然而这些研究都是针对某种特定类型的钓鱼行为进行检测, 为了更好的识别出更多钓鱼页面, 文献 [5] 提出 CANTINA 算法, 在 URL、域名等信息基础上引入了页面的 Logo 和 Forms 等内容作为特征, 同时采用 TF-IDF 方法提取页面关键词进行钓鱼检测, 在与几款流行反钓鱼插件的比较实验中获得了较好的检测效果. Sanglerdsinlapachai 等人 [10] 在 CANTINA 算法基础上引入域名主页相似度特征度量定义, 并结合 SVM 等机器学习方法对钓鱼行为进行鉴定. 此外, 为

了加强特征的表征能力, 文献 [24] 引入时间和空间的概念, 结合 Internet Archive[25] 保存的历史域名页面以及对待检测页面相关页面内容的分析, 提取出更具表征性的关键词信息, 该方法虽然降低了基于内容检测方法的误报率, 但同时也消耗了大量遍历解析的时间开销.

由于上述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. 为了实现对大规模未知网站的检测, 本文通过对 100,000 个正常网站和 100,000 个钓鱼网站的学习, 解析其对应的网页内容, 提取其相应的网页标题、网页关键字、网页描述信息等 8 种特征来描述这些网站, 然后基于不同的特征表达方法构建了相应的分类器; 对于待检测的网站, 采用分类集成的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. 基于上述方法, 本文构建了钓鱼网站智能检测系统 IPWDS, 并将其集成于金山安全产品中. 在大量、真实数据集的基础上, 实验结果表明采用本文所提出方法的钓鱼网站智能检测系统 IPWDS 对钓鱼网站的检测效果优于现有常见的钓鱼网站检测方法和常用的反钓鱼软件.

3 系统架构

通过解析已知正常网站和钓鱼网站对应的网页内容, 提取其相应的网页标题、网页关键字、网页描述信息等 8 种特征来描述这些网站, 然后基于不同的特征表达方法构建了相应的分类器; 对于待检测的网站, 采用分类集成的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. 基于上述方法, 构建了钓鱼网站智能检测系统 IPWDS (intelligent phishing website detection system), 其系统架构如图 1 所示.

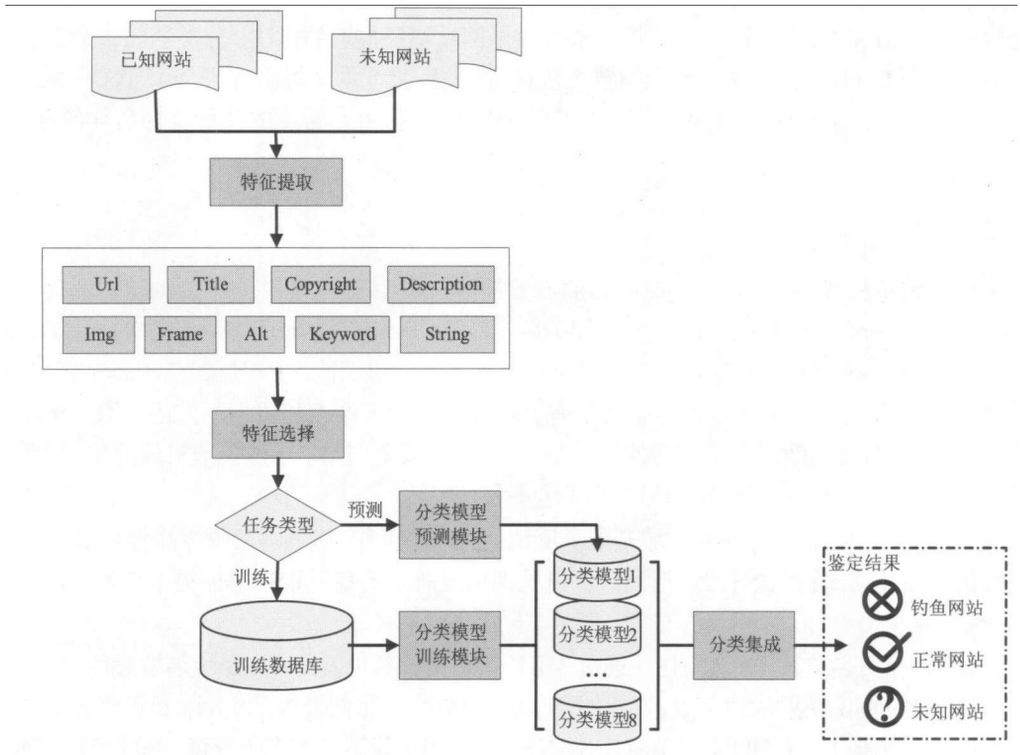

图 1 IPWDS 系统结构图

IPWDS 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, 采用分类集成的方法, 综合各个分类器的预测结果, 实现对未知网站的钓鱼检测. 系统的主要模块如下:

1) 特征提取模块: IPWDS 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 (如表 1 所示).

2) 基础分类器: 基于网站内容提取的不同特征构建相应的分类模型. 本文采用 NBC (naive Bayes classifier, 朴素贝叶斯分类器) 和 SVM (support vector machine, 支撑向量机) 作为基础分类器, 其中除 String 以外的其它特征分词后采用 NBC 作为基础分类器, 对 String 特征分词后采用 SVM 分类算法进行分类器构建.

3) 分类器集成模块: 对于待检测网站, 分类器集成模块根据基础分类器两两之间的关系, 计算各个特征分类模型预测结果的权重, 最后综合各个基础分类器的预测结果得出结论.

表 1　网站的特征表征方式

(http://www.taobao.com.mmand.co.cc/)

ID	特征名	说明	举例
1	Title	标题标签内容	淘宝网 - 魔兽世界 300 元 =3w 发货定单(承诺: 纯手工. 假一罚十)
2	Keyword	Meta 标签中提取的关键字信息	淘宝, 掏宝, 网上购物, 购物, 论坛, 联盟, 买, 卖,1 元, 一元, 一口价, 拍卖
3	Description	Meta 标签中提取的页面描述性信息	中国最大、最安全的网上交易社区, 尽享淘宝乐趣!
4	Copyright	Meta 标签中提取的版权信息	2001-2010 Taobao.com 版权所有
5	Frame	页面中包含框架页面的 URL 地址	http://assets.taobaocdn.com/tbra/1.0/tbra-widgets.js?t=20090409.js
6	Img	图片链接地址	http://img05.taobaocdn.com/bao/uploaded/i5/T1eVFwXm8CXXbnDi3Z_033008.jpg_160x160.jpg
7	Alt	Alt 标签中的替换文本	本店支持信用卡直接付款
8	String	Script 脚本中的字符串和页面所有可见字符串集合	SHOP NOTICE、信用、评价、购物、保障、收藏、快递、EMS、宝贝···

4 算法描述

本节将对 IPDWS 系统采用的相关方法和算法进行详细阐述, 包括: 特征提取的方法, 基础分类器的构造算法, 分类集成的方法等.

4.1 特征提取

在特征提取时, 利用 HTML 文档中的标签来提取相应的网页信息 (如表 1 中 1–8 所示), 其中, String 特征指所有标签的外部文本 (即浏览器中可见文本) 以及 Script 标签中的明文信息, 例如:

Alert("恭喜你有机会参加腾讯 2010 抽奖活动!");

"恭喜你有机会参加腾讯 2010 抽奖活动!" 的字符串信息将被作为 String 特征进行提取.

上述特征需要进一步进行分词操作, 具体做法是: 链接地址直接作为一个分词结果, 其它英文文本按单词分词, 中文字符串通过基于多层隐马尔科夫模型的汉语词法分析系统 ICTCLAS[26] 进行分词. 对上述例子, 分词后的结果如下:

恭喜 \你 \有 \机会 \参加 \腾讯 \2010\抽奖 \活动

通过分词操作后, 对于每个已知网站将获得 8 个新的特征词集合, 用于基础分类器的构建.

4.2 基础分类器

基于不同的特征构建了不同的基础分类器, 这是由于: 1) 每种特征的重要性不同, 例如: 网页标题的文本信息在多数情况下相比出现在页面中其它位置的大部分文本信息能够更好地概括出页面的主题信息; 2) 根据各种特征的不同特性, 应该采用不同的分类算法. String 特征经过分词后相比其它特征包含了更多的信息, 是一种信息能力表达较强的特征, 同时, 其维度也比较高, 由于 SVM (support vector machine, 支撑向量机) 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 [27], 本文采用 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, 维度较低, 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.

4.2.1 扩展贝叶斯分类器

朴素贝叶斯分类器 (NBC) 是数据挖掘中常用的分类方法之一 [28], 式 (1) 是朴素贝叶斯分类器通常采用

的公式:

$$V_{NBC} = \arg\max_{c_j \in V} P(c_j) \prod_i P(X_i|c_j) \tag{1}$$

其中 C_j 表示第 j 个分类, X_i 表示第 i 个特征. 在实际的应用中, 将式 (1) 扩展如下:

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

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.

4.2.2 改进支撑向量机

SVM (support vector machine, 支撑向量机) 方法是建立在统计学习理论的 VC 维理论和 Vapnik[29] 结构风险最小化原理基础上的, 对维度较高的数据集, 支撑向量机有较强的处理能力和较好的分类正确率 [27]. 本文采用线性 SVM 来完成模型的训练和预测.

线性 SVM 的基本思想就是寻找一个两类之间的最优分类平面, 其应该满足如下两个条件:

$$y_i[(W \cdot X_i) + b] \geq 1, \quad i = 1, 2, \cdots, |D| \tag{3}$$

$$\min \left(\frac{1}{2} \|W\|^2 + C \cdot \sum_{i=1}^{|D|} \xi_i \right) \tag{4}$$

其中 X 是输入向量, W 为权重向量, 式 (4) 中 $\frac{1}{2}\|W\|^2$ 使样本到超平面的距离尽量大, 此外, 考虑到可能存在一些样本不能被超平面正确分类, 因此引入松弛变量 ξ, $C \cdot \sum_{i=1}^{|D|} \xi_i$ 使得误差尽量小.

本文结合 TF-IDF 方法 [30] 计算 String 特征中分词后每个词的权重, 设 X_i 为某个词, $Count(j, X_i)$ 为文件 j 中该词出现的次数, $Count(j)$ 表示文件 j 包含的分词总数. 则 X_i 的词频 TF 表示如下:

$$TF(Xi) = \frac{Count(j, Xi)}{Count(j)} \tag{5}$$

设 $CountFile(X_i)$ 为出现 X_i 的文件数, $CountFileAll$ 为总文件数, 文件频率 DF 表示如下:

$$DF(Xi) = \frac{CountFile(Xi)}{CountFileAll} \tag{6}$$

$TF-IDF$ 值由下式计算:

$$TF-IDF(Xi) = \frac{TF(Xi)}{DF(Xi)} \tag{7}$$

此后, 对每个文件中所有词的 $TF-IDF$ 值进行归一化, 并将归一化后的结果作为向量输入.

SVM 训练后得到的模型为一组支撑向量, 对于线性 SVM 模型, 可以按以下公式将其转化为分词后词的加权向量:

$$Fi = \frac{\sum_{j=1}^{N} C_j \times F_{ji}}{N} \tag{8}$$

其中, N 为模型中的支撑向量总数, F_i 表示第 i 个维度值, C_j 表示第 j 个支撑向量的类型, F_{ji} 表示第 j 个支撑向量第 i 个维度的值. 通过上述转化, 预测的复杂度由 $O(N*k)$ 降低到 $O(k)$, 其中 k 为特征的维度, 即分词后词的个数.

4.3 分类集成

集成学习 (ensemble learning)[31] 是用有限个学习器对同一个问题进行学习, 集成在某输入示例下的输出由构成集成的各学习器在该示例下的输出共同决定. 当个体有较高的精度并且个体是互不相同的, 集成比构成集成的任何一个个体 (单个学习器) 有更好的预测效果 [31]. 集成学习主要包括两个部分:

1) 个体生成: 就是通过一定的策略生成具有较高正确率与差异性的集成个体 (分类器). 按照集成个体 (分类器) 之间的种类关系可以把集成学习方法划分为异态集成学习和同态集成学习两种 [32]:

a) 同态集成学习是指集成的基本分类器都是同一种分类器, 只是这些基本分类器之间的参数有所不同 [33].

b) 异态集成学习是指使用各种不同训练算法的分类器进行集成 [33].

如 4.2 节所述, 对于各个基础分类器, 由于其特征表达方式的不同, 采用的分类算法也不完全相同, 本文个体生成的方法主要采用异态集成学习的方法.

2) 结论生成: 就是将不同个体 (分类器) 的输出结果进行组合, 这种组合方法既可以是线性的也可以是非线性的 [34]. 以下是几种常见的结论生成方法:

a) 基于简单投票的结论生成方法: 简单投票的基本思想是多个基本分类器都进行分类预测, 然后根据分类结果用某种投票的原则进行投票表决. 按照投票原则的不同投票法可以有一票否决、一致表决、少数服从多数、阈值表决等 [33,35−36].

b) 基于贝叶斯投票的结论生成方法: 简单投票法假设每个基本分类器都是平等的, 没有分类能力之间的差别, 但是这种假设并不总是合适的. 贝叶斯投票法 [33,35,37] 是基于每一个基本分类器在过去的分类表现来设定一个权值, 然后按照这个权值进行投票, 其中每个基本分类器的权值基于贝叶斯定理来进行计算 [33,35,37]. 虽然理论上贝叶斯投票法在假设空间所有假设的先验概率都正确的情况下能够获得最优的集成效果, 但是实际应用中往往不可能穷举整个假设空间, 也不可能准确地给每个假设分配先验概率 [33,35,37]. 因此, 在实际使用中其他结论生成方法可能会优于贝叶斯投票法.

c) 基于线性组合的结论生成方法: 这种方法指的是使用各个基本分类器输出的线性组合作为分类结果.

d) 基于 D-S 证据理论的结论生成方法: Rogova[38] 提出了 D-S (Dempster-Shaffer) 证据理论结论生成方法, 其基本思想是通过识别率、拒绝率等一系列参数计算出每个目标分类的信任范围, 从而最终推断出分类结果 [33,35,39−41].

为了获取比单个基础分类更好的预测结果, 本文利用基础分类器两两之间的关系, 对 8 个基础分类器得到的预测结果进行进一步加权, 提出一种适合于二分类问题的结论生成方法 CE (correlation based ensemble). CE 方法的主要思想是根据基础分类器两两之间的相关关系, 对单个分类器的结果产生一个权值, 然后进行加权集成.

对一个待检测网站 s, 用 $C_i(s)$ 来表示第 i 个基础分类器的预测结果 (对于二分类问题, $C_i(s)$ 是 1 或 -1). 为了对 $C_i(s)$ 产生一个权值, 需要考虑其它基础分类器的预测结果对其产生的影响. $C_j(s)$ 表示第 j 个基础分类器的预测结果. 给定第 i 和 j 个基础分类器的预测结果为 $C_i(s)$ 和 $C_j(s)$, 分类集成的预测结果是 $C_i(s)$ 的可靠性有多大? 本文采用 $f(C_i(s), C_j(s), Class = C_i(s))$ 来表示这种可靠性, 它可以用下面的公式来估算:

$$f(C_i(s), C_j(s), Class = C_i(s)) = \frac{Count(C_i(s), C_j(s), Class = C_i(s))}{Count(C_i(s), C_j(s))} \tag{9}$$

$Count(C_i(s),\ C_j(s),\ Class = C_i(s))$ 表示在训练数据集中, 有多少训练样本满足以下条件: 第 i 个基础分类器的预测结果是 $C_i(s)$, 第 j 个基础分类器的预测结果是 $C_j(s)$, 分类集成的预测结果是 $C_i(s)$. $Count(C_i(s),\ C_j(s))$ 表示在训练数据集中, 有多少训练样本满足以下条件: 第 i 个基础分类器的预测结果是 $C_i(s)$, 第 j 个基础分类器的预测结果是 $C_j(s)$. 事实上, 通过式 (9), 我们可以计算出第 i 个基础分类器和第 j 个基础分类器之间的相关关系: 给定第 i 个基础分类器的输出 (记为 $C_i(s)$), 式 (9) 表示其与第 j 个基础分类器输出 (记为 $C_j(s)$) 一致的可能性.

为了得到分类集成的最终预测结果, 可以综合考虑每个基础分类器与其它所有基础分类器两两之间的相关关系, 计算如下权值:

$$Score(s) = \sum_{C_i(s)\neq 0} C_i(s) * \frac{\sum\limits_{C_j(s)\neq 0} f(C_i(s), C_j(s), Class = C_i(s))}{\sum\limits_{C_j(s)\neq 0} |C_j(s)|} \tag{10}$$

其中, $\sum\limits_{C_j(s)\neq 0} |C_j(s)|$ 表示与分类器 C_i(s) 相关的分类器的个数. 式 (10) 等式右边的第二项表示第 i 个基础分类器的权重是该基础分类器和所有与其具有相关关系的其它基础分类器输出一致的平均概率. 该权重刻画了一个基础分类器与其它基础分类器之间的一种一致性度量. 式 (10) 给出了对一个未知网站的最终预测结果是由所有基础分类器输出的加权和. 对一个未知网站进行预测时: 1) $Score(s) < 0$ 表明该网站经分类集成后的预测结果为钓鱼网站; 2) $Score(s) > 0$ 表明该网站经分类器集成后的预测结果为正常网站; 3) $Score(s) = 0$ 表明该网站经分类器集成后的预测结果为未知网站.

总之, CE 方法是一种加权集成的方法, 权重由基础分类器的相互关系来自动确定: 1) CE 方法不同于简单投票的结论生成方法, 它是一种基于基础分类器两两之间相关关系的加权投票. 只有当所有的基础分类器都两两不相关的时候, CE 方法才等价于简单投票策略. 2) CE 方法也不同于其它的组合策略, 因为每个基础分类器的权重是由其与其它基础分类器之间的相关关系自动生成的. 事实上, CE 方法可以认为是一种基于一致性 (consistency-based) 的结论生成方法.

5 实验结果与分析

本节实验数据来源于金山安全实验室, 实验内容主要包括以下几个部分: 1) 分析比较每个基础分类器对钓鱼网站的检测效果; 2) 检验本文所提出的分类集成方法的有效性; 3) 与其它常用钓鱼检测方法进行对比, 验证本文提出的钓鱼网站检测方法的有效性.

5.1 实验准备

本文的训练样本来源于金山安全实验室收集的 100,000 个正常网站和 100,000 个钓鱼网站, 同时以其客户端安全产品每日真实上报的未知网站作为测试数据 (这里取连续六天上报的未知网站作为测试集, 如表 2 所示).

根据各个分类器对钓鱼网站检测的准确率 (precision) 和召回率 (recall) 来评估其预测结果的好坏.

TP (true positive): 被分类器正确预测为钓鱼网站的个数;
TN (true negtive): 被分类器正确预测为正常网站的个数;
FP (false positive): 被分类器错误预测为钓鱼网站的个数;
FN (false negtive): 被分类器错误预测为正常网站的个数;
precision(准确率): $Precision = \frac{TP}{TP+FP}$;
recall(召回率): $Recall = \frac{TP}{TP+FN}$.

表 2 测试数据集说明

Day	Phishing Websites	Normal Websites	Total Number
2010-8-17	219	21338	21557
2010-8-18	618	46987	47605
2010-8-19	388	34878	35266
2010-8-20	297	28775	29072
2010-8-21	605	49407	50012
2010-8-22	638	53469	54107

5.2 不同基础分类器检测结果比较

本节主要对基于不同特征提取方法训练得到的基础分类器进行检测和比较, 表 3 和图 2 显示了各个基础分类器对钓鱼网站的检测结果.

从表 3 和图 2 可以看出: 1) 基于 Frame 特征的基础分类器对钓鱼网站检测的准确率最高, Frame 用于包含其它页面文件, 因此对于通过直接包含正常网站页面内容进行仿冒的钓鱼欺骗具有较高的检测准确率. 但由于该类型钓鱼欺骗数量不多, 其召回率只有 2%-3%. 2) 基于 String 特征的基础分类器对钓鱼网站检测的召回率最高, 其包含丰富的页面字符信息, 可以较充分地表达网页信息. 3) 此外, Title、Keyword 和 Description 与页面的主题相关, 例如: "QQ 中奖"、"幸运抽奖"、"廉价网购"、"游戏点卡充值" 等, 钓鱼网站主要通过它们来吸引和欺骗用户, 因此这些特征对钓鱼有较好识别能力, 检测准确率也都比较高; Copyright 代表页面的版权信息, 钓鱼网站版权信息往往与真实页面版权存在差异, 然而一些通用的版权信息, 如: "Copyright @ 2010", 容易对分类效果造成影响, 因此检测准确率变动较大; Img 代表了页面上的图片, 许多钓鱼欺骗采用大量图片来吸引用户, Alt 作为图片的说明文本, 如: "腾讯 QQ 幸运用户"、"非常 6+1 抽奖" 等, 对用户有较大的欺骗性, 因此, 通过这两个特征对钓鱼网站检测的准确率也比较高.

表 3　不同基础分类器检测结果比较

分类器	2010-8-17		2010-8-18		2010-8-19	
	precision	recall	precision	recall	precision	recall
Title	84.99%	26.41%	91.39%	35.34%	86.34%	29.99%
Keyword	82.99%	16.94%	88.55%	21.43%	88.31%	15.25%
Description	87.10%	9.00%	92.08%	10.93%	93.93%	9.78%
Copyright	55.30%	10.05%	97.33%	5.03%	66.73%	12.03%
Alt	69.94%	5.25%	73.38%	5.75%	67.25%	3.75%
Frame	96.59%	2.36%	98.30%	2.49%	98.21%	2.38%
Img	83.72%	15.32%	85.63%	12.78%	92.10%	16.24%
String	76.38%	62.47%	81.45%	53.30%	85.44%	61.72%
分类器	2010-8-20		2010-8-21		2010-8-22	
	precision	recall	precision	recall	precision	recall
Title	80.58%	20.98%	83.83%	26.97%	78.88%	27.82%
Keyword	93.96%	17.35%	96.16%	18.43%	92.85%	13.86%
Description	97.80%	10.40%	96.68%	7.78%	93.95%	8.07%
Copyright	59.61%	9.05%	59.70%	7.04%	98.42%	3.04%
Alt	74.81%	2.96%	73.75%	5.26%	68.82%	3.50%
Frame	98.54%	3.45%	98.05%	3.37%	97.12%	1.85%
Img	86.79%	12.51%	87.36%	15.43%	88.42%	15.77%
String	80.54%	75.63%	89.73%	52.40%	91.33%	71.58%

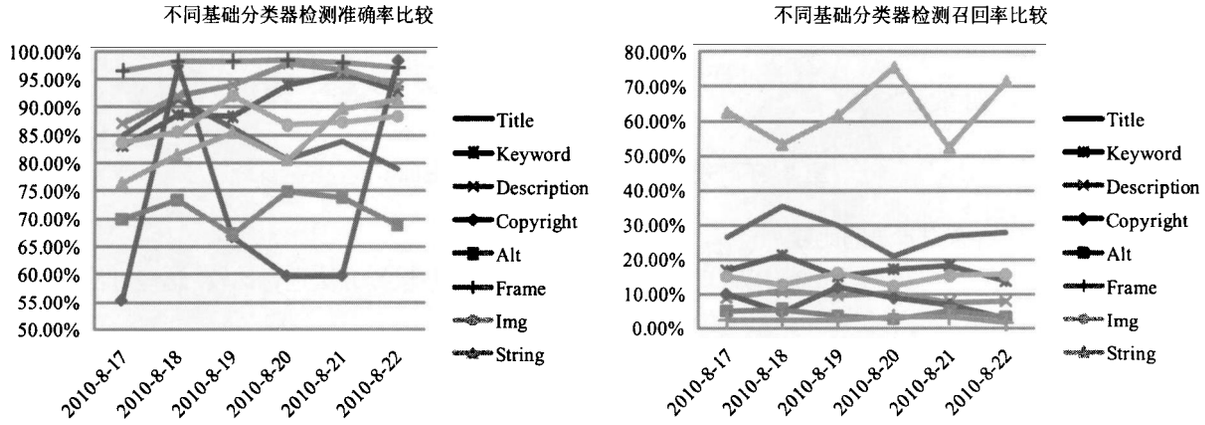

图 2　不同基础分类器准确率与召回率比较

从这组实验可以看出, 单独采用某个特征都不能完整地表述整个网页内容, 为了得到更好的检测效果, 需要对各个基础分类器进行分类集成.

5.3　不同分类集成方法的结果比较

本节实验将分类集成后的检测结果与单个基础分类器进行比较, 并对不同分类集成方法进行检测和比较. 比较的分类集成方法包括: 多数投票法 (majority vote)、简单加权法 (simple weight vote)、贝叶斯投票法 (Bayes vote) 以及本文的 CE (correlation based ensemble) 方法. 实验结果如表 4 和图 3 所示.

从表 4 和图 3 的实验结果可以看出: 1) 相比单个基础分类器 (如表 3 和图 2 所示), 对各个基础分类器进行分类集成后, 对钓鱼网站检测的准确率和召回率都有较大的提升; 2) 本文提出的 CE 方法比其它几种常用的分类集成方法对钓鱼网站的检测具有更高的准确率和召回率: CE 方法相比简单投票和其它结论生成方法, 考虑了各个基础分类器两两之间的相互关系.

表 4 不同分类集成方法检测结果比较

集成方法	2010-8-17		2010-8-18		2010-8-19	
	precision	recall	precision	recall	precision	recall
多数投票	95.73%	90.42%	96.62%	91.36%	95.65%	93.44%
简单加权	96.24%	93.31%	96.98%	95.15%	96.22%	95.40%
贝叶斯投票	96.02%	92.12%	96.85%	94.30%	95.91%	94.41%
CE 集成	97.62%	96.73%	98.71%	97.83%	98.02%	98.77%
集成方法	2010-8-20		2010-8-21		2010-8-22	
	precision	recall	precision	recall	precision	recall
多数投票	96.02%	92.19%	95.77%	94.37%	94.82%	93.30%
简单加权	96.64%	95.30%	96.41%	96.11%	95.50%	95.43%
贝叶斯投票	96.49%	93.98%	96.33%	95.55%	95.17%	94.28%
CE 集成	97.75%	98.53%	97.32%	97.29%	97.58%	98.52%

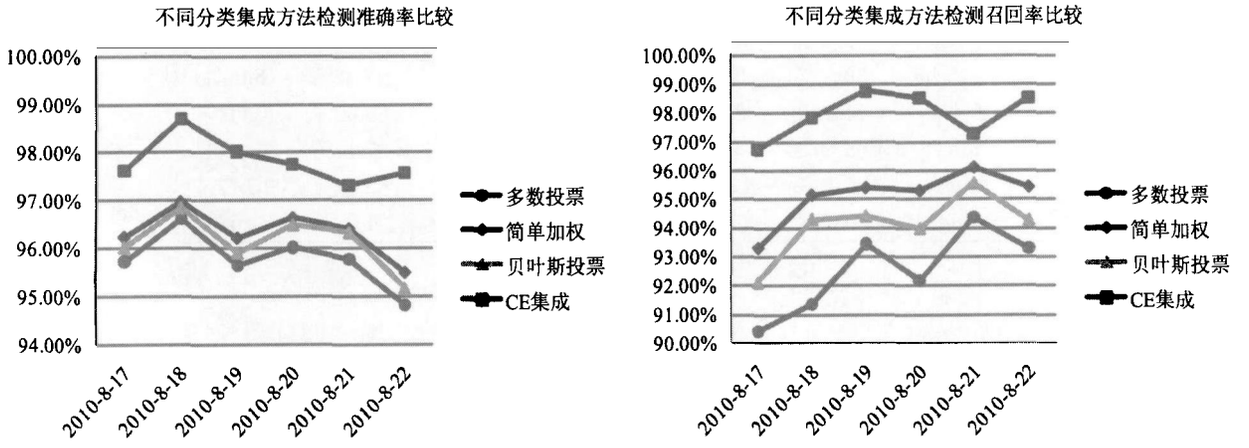

图 3 不同分类集成方法检测准确率与召回率

5.4 与常用钓鱼网站检测方法的比较

本节实验将采用本文提出方法的 IPWDS 系统与常用的钓鱼网站检测方法进行比较, 这些方法包括:

1) UDB: 基于 URL 混淆和 DNS 欺骗的检测方法 [22];

2) WSR: 基于第三方库特征查询的检测方法 [11];

3) TCML: 结合页面内容特征和机器学习的检测方法 [10].

对 5.1 节所描述的测试数据, 各种钓鱼检测方法的检测结果如表 5 和图 4 所示.

表 5 不同钓鱼检测方法检测结果对比

检测方法	2010-8-17		2010-8-18		2010-8-19	
	precision	recall	precision	recall	precision	recall
UDB	62.73%	5.28%	64.26%	4.73%	60.52%	5.11%
WSR	80.22%	61.30%	85.73%	54.33%	86.19%	59.30%
TCML	88.32%	67.71%	94.52%	65.30%	90.07%	67.56%
IPWDS	97.62%	96.73%	98.71%	97.83%	98.02%	98.77%
检测方法	2010-8-20		2010-8-21		2010-8-22	
	precision	recall	precision	recall	precision	recall
UDB	63.45%	6.01%	55.44%	5.31%	56.38%	5.77%
WSR	81.40%	66.01%	82.44%	63.81%	79.69%	65.47%
TCML	91.42%	72.31%	91.31%	70.43%	88.42%	74.83%
IPWDS	97.75%	98.53%	97.32%	97.29%	97.58%	98.52%

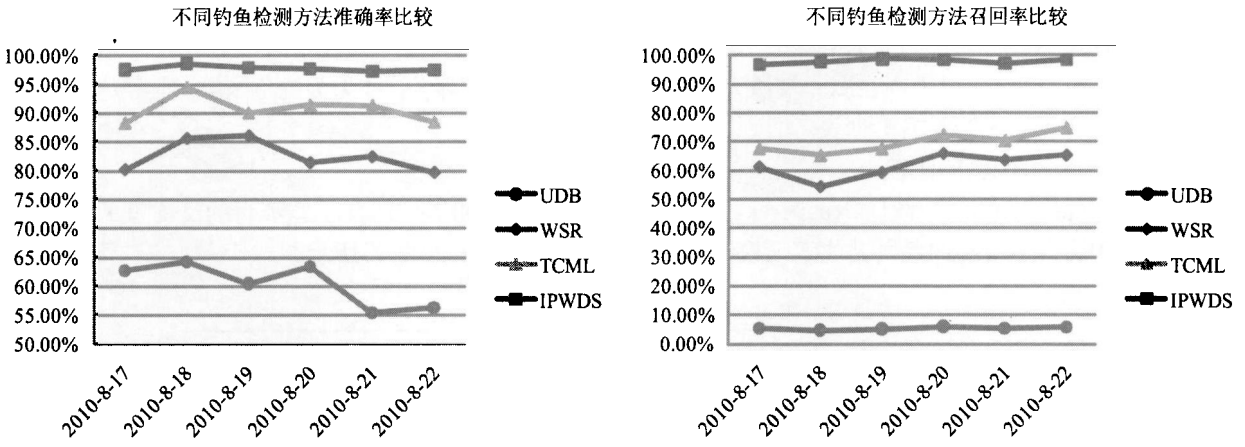

图 4　不同钓鱼检测方法准确率与召回率比较

从表 5 和图 4 中可以看出, 对每天上报的未知网站的检测, 采用本文提出方法的 IPWDS 系统比其它几种常用的钓鱼网站检测方法具有更高的检测准确率和召回率, 因为: IPWDS 使用了 8 种具有代表性的页面特征进行表征, 构造了不同的基础分类器, 并采用分类集成的方法进行最后的预测. 相比 UDB 和 WSR 方法, IPWDS 可以识别出更多新增和排名不高的页面, 此外, 对查无第三方库信息的页面, IPWDS 相比 TCML 方法具有更丰富的表征能力.

6　IPWDS 的实际应用

目前, IPWDS 已集成在金山安全产品后台中, 实现对每天上报的大量未知网站进行钓鱼检测.

6.1　IPWDS 系统应用

图 5 展示了 IPWDS 的实际应用部署及其对未知网站的检测流程.

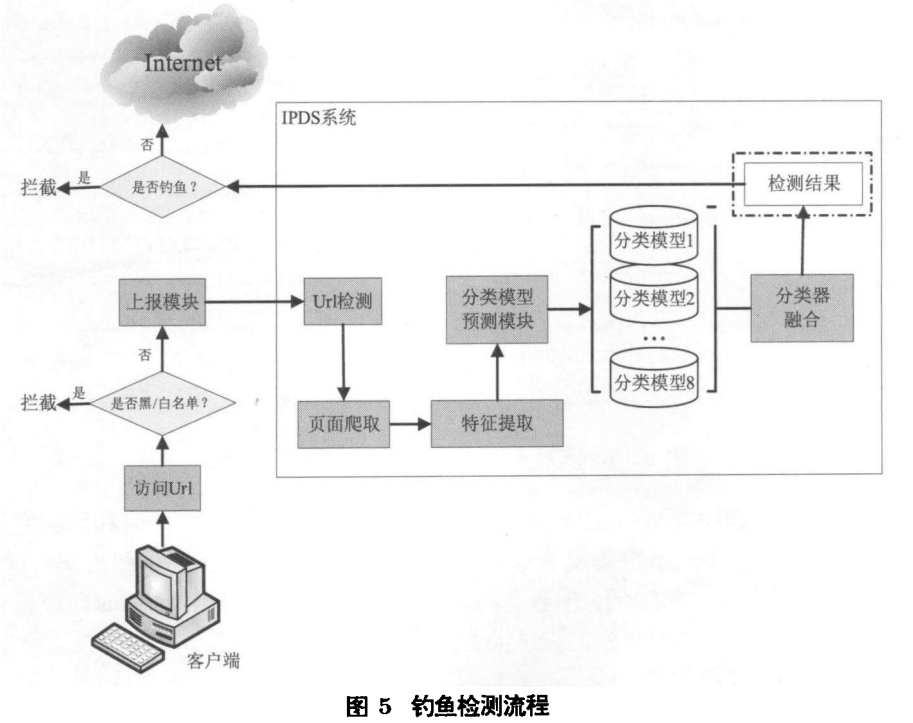

图 5　钓鱼检测流程

1) 客户端鉴定. 首先, 客户端简单的维护一张黑/白名单列表, 当访问网址时, 先在黑/白名单列表中查询, 匹配到白名单, 则验证通过; 若匹配到黑名单, 则拦截对该网站的访问. 否则, 客户端将该网址上报至服务器云后台的鉴定系统.

2) 页面特征鉴定. 如果 URL 检测模块无法鉴定出结果, IPWDS 将继续爬取页面内容, 然后进行特征提

取生成预测样本特征向量, 并交由 8 个分类器同时进行鉴定, 最后采用分类集成综合所有鉴定结果, 判断是否为钓鱼网站.

3) 服务端反馈. 一旦样本在服务端做完鉴定, 立即将结果反馈客户, 如果为钓鱼网址, 则客户端拦截用户访问, 并更新客户端的黑名单列表.

6.2 与常见反钓鱼工具的比较

本节实验将 IPWDS 系统同 Kaspersky、Mcafee SiteAdvisor 和 Netcraft 三款流行的反钓鱼工具进行比较, 验证 IPWDS 系统对钓鱼网站的实际检测效果. 实验结果如表 6 和图 6 所示.

表 6 不同反钓鱼工具检测结果比较

Day	Kaspersky		Mcafee	
	precision	recall	precision	recall
2010-8-17	95.18%	2.28%	88.72%	9.59%
2010-8-18	97.00%	1.78%	92.73%	4.37%
2010-8-19	94.77%	1.80%	90.10%	6.44%
2010-8-20	94.52%	2.36%	89.64%	7.41%
2010-8-21	92.89%	1.82%	89.19%	4.46%
2010-8-22	93.53%	1.72%	88.46%	4.55%
Day	Netcraft		IPWDS	
	precision	recall	precision	recall
2010-8-17	75.40%	28.31%	97.62%	96.80%
2010-8-18	76.32%	30.91%	98.71%	97.90%
2010-8-19	77.05%	24.23%	98.02%	94.85%
2010-8-20	74.11%	23.57%	97.75%	96.97%
2010-8-21	75.23%	29.26%	97.32%	95.87%
2010-8-22	76.58%	31.82%	97.58%	97.96%

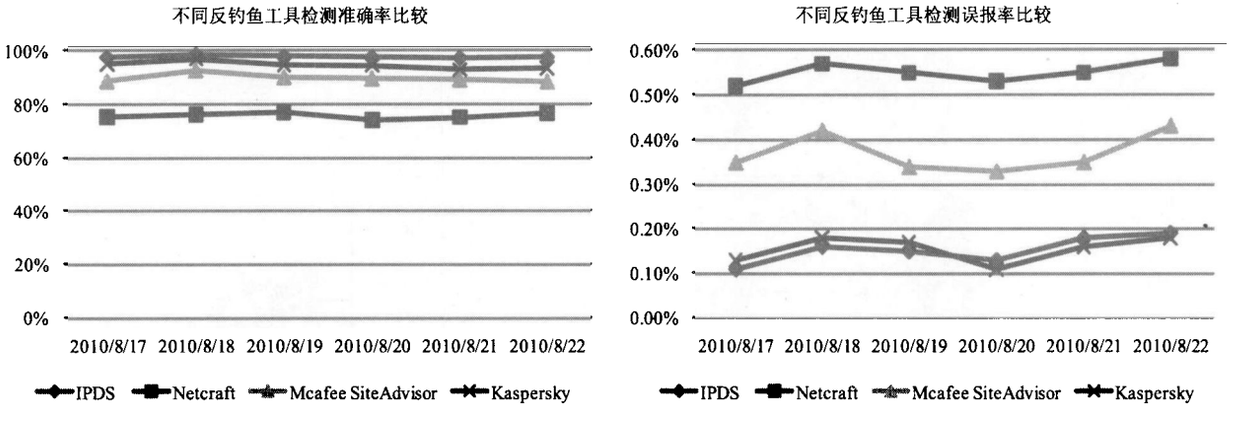

图 6 不同反钓鱼工具检测准确率与误报率比较

从表 6 和图 6 实验结果可以看出: IPWDS 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. IPWDS 系统集成了各种页面内容特征进行钓鱼识别, 相比 Kaspersky 和 Mcafee SiteAdvisor 可以识别出更多访问频度低、排名不高以及新增的页面, 此外, 对注册时间较长并有合法域名信息的钓鱼页面, 其相比 Netcraft 具有更强的表征能力. 因此, 与其它反钓鱼工具的比较实验中, IPWDS 系统具有更高的钓鱼检测准确率和召回率.

金山安全产品后台每天要对 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%), IPWDS 系统能够较全面的检测出这些钓鱼网站. 图 7 是某个仿淘宝网站的钓鱼网站, 采用上述各种反钓鱼工具进行检测后, 只有 IPWDS 系统能够对其进行有效拦截 (如图 7 所示).

图 7　IPWDS 系统识别钓鱼页面

7　总结

本文提出一种基于网站页面内容的智能钓鱼检测方法, 和传统方法相比, 有以下优点: 1) 采用多种特征表达方法: 通过对大量已知正常网站和钓鱼网站的学习, 解析其对应的网页内容, 提取相应的网页标题、网页关键字、网页描述信息等 8 种特征来描述这些网站信息; 2) 异构的基础分类器: 根据不同特征的特性, 采用了 NBC 和 SVM 两种不同的分类算法构造异构的基础分类器, 并成功应用于钓鱼网站检测中; 3) 分类集成: 根据基础分类器两两之间的关系, 采用分类集成方法综合各个基础分类器的预测结果, 提高最终的分类预测准确率和召回率; 4) 基于上述方法, 构建了钓鱼网站智能检测系统 IPWDS, 并将其集成于金山安全产品中. 在大量、真实数据集的基础上, 实验结果表明采用本文所提出方法的钓鱼网站智能检测系统 IPWDS 对钓鱼网站的检测效果优于现有常见的钓鱼网站检测方法和常用的反钓鱼软件. 在今后的工作中, 将进一步对检测出的钓鱼网站进行有效归类, 把更详尽的钓鱼信息反馈给用户.

致谢　在此, 我们向对本文工作给予数据和技术支持的金山安全实验室表示感谢.

参考文献

[1] Anti-Phishing Working Group[EB/OL]. http://www.antiphishing.org/.

[2] Liu G, Qiu B, Liu W Y. Automatic detection of phishing target from phishing webpage[C]//2010 20th International Conference on Pattern Recognition, Istanbul: IEEE Computer Society, 2010: 4153–4156.

[3] 中国反钓鱼网站联盟 [EB/OL]. http://www.cnnic.net.cn/html/Dir/2008/07/31/5246.htm.

[4] CNCERT/CC[EB/OL]. http://www.cert.org.cn/.

[5] Zhang Y, Hong J I, Cranor L F. Cantina a content-based approach to detecting phishing web sites[C]//Proceedings of the 16th International Conference on World Wide Web, USA: ACM New York, 2007: 639–648.

[6] Cranor L F, Egelman S, Hong J I, et al. Phinding phish: Evaluating anti-phishing tools[C]//Proceedings of the 14th Annual Network and Distributed System Security Symposium (NDSS'07), USA: ACM New York 2007: 88–99.

[7] Pan Y, Ding X H. Anomaly based web phishing page detection[C]//Computer Security Applications Conference, Miami: ACSAC 2006 22nd Annual, 2006: 381–392.

[8] Aburrous M R, Hossain M A, et al. Intelligent phishing website detection system using fuzzy techniques[C]//Information and Communication Technologies: From Theory to Applications, Damascus: ICTTA 2008, 2008: 1–6.

[9] Nakayama S, Yoshiura H, Echizen I. Preventing false positives in content-based phishing detection[C]//2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kyoto: IEEE Computer Society, 2009: 48–51.

[10] Sanglerdsinlapachai N, Rungsawang A. Using domain top-page similarity feature in machine learning-based web phishing detection[C]//2010 Third International Conference on Knowledge Discovery and Data Mining, Phuket: CPS, 2010: 187–190.

[11] Kim Y G, Cho S Y, Lee J S, et al. Method for evaluating the security risk of a website against phishing attacks[J].

Lecture Notes in Computer Science, 2010, 5075: 21–31.
[12] Microsoft Corp. Microsoft phishing filter: A new approach to build trust in e-commerce[EB/OL]. http://www.microsoft.com/downloads/, 2005.
[13] Paul R. Gone phishing: Evaluating anti-phishing tools for windows[EB/OL]. http:// www.3sharp.com/projects/antiphishing/gonePhishing.pdf, 2006.
[14] GeoTrust Corp. GeoTrust Introduces industry's first secure consumer search service[EB/OL]. http://www.geotrust.com/about/news_events/press/PR_TrustedSeacrh_092605s.pdf, 2006.
[15] Chou N, Ledesma R, Teraguchi Y, et al. Client-side defense against web-based identity theft[R]. Proceedings of the Network and Distributed System Security Symposium, 2004.
[16] Netcraft. Netcraft anti-phishing toolbar[EB/OL]. http://toolbar.netcraft.com/.
[17] Anti-Phishing Working Group. Phishing activity trends report[EB/OL]. http://antiphishing.org/APWG_Report_March_2007.pdf, 2007.
[18] Nimeh S A, Nappa D, Wang X L, et al. A comparison of machine learning techniques for phishing detection[C]//Proceedings of the Anti-Phishing Working Groups 2nd Annual Ecrime Researchers Summit, USA: ACM New York, 2007: 60–69.
[19] Sanglerdsinlapachai N, Rungsawang A. Using domain top-page similarity feature in machine learning-based web phishing detection[C]//2010 Third International Conference on Knowledge Discovery and Data Mining, Phuket: CPS, 2010: 187–190.
[20] Guo M Z, Yuan J S, Wang Y C. Phishing web page detection algorithm[J]. Computer Engineering, 2008, 34: 161–163.
[21] Kang J M, Lee D H. Advanced white list approach for preventing access to phishing sites[C]//2007 International Conference on Convergence Information Technology (ICCIT 2007), Korea: IEEE, 2007: 491–496.
[22] Cao Y, Han W L. Anti-phishing based on automated individual white-list[C]//Conference on Computer and Communications Security, USA: ACM New York, 2008: 51–60.
[23] Layton R, Watters P. Using differencing to increase distinctiveness for phishing website clustering[C]//Proceedings of the 2009 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, Brisbane: IEEE, 2009: 488–492.
[24] Nakayama S, Yoshiura H, Echizen I. Preventing false positives in content-based phishing detection[C]//2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kyoto: IEEE, 2009: 48–51.
[25] Internet Archive[EB/OL]. http://www.archive.org/.
[26] 中科院分词系统 ICTCLAS[EB/OL]. http://www.ictclas.org/.
[27] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273–297.
[28] Written I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation[M]. Seattle: Morgan Kaufmann Publishers, 2000: 265–314.
[29] Salton G, Buckley B. Term-weighting approaches in automatic text retrieval[J]. Information Processing and Management, 1988, 24(5): 513–523.
[30] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995.
[31] Dietterich T G. Ensemble Learning: The Handbook of Brain Theory and Neural Networks[M]. Cambridge: The MIT Press, 2002.
[32] Yu S X. Feature selection and classifier ensembles: A study on hyperspectral remote sensing data[D]. United States: Scientific Literature Digital Library and Search Engine, 2003.
[33] 梁英毅. 集成学习综述 [EB/OL]. soft.cs.tsinghua.edu.cn/ keltin/docs/ensemble.pdf.
[34] 张丽新. 高维数据的特征选择及基于特征选择的集成学习研究 [D]. 北京: 清华大学, 2004.
[35] Xu L. Methods of combining multiple classifiers and their applications to handwriting recognition[J]. IEEE Transactions on Systems, Man and Cybernetics, 1992, 22(3): 418–435.
[36] Bahler D, Navarro L. Methods for combining heterogeneous sets of classifiers[C]//17th Natl Conf on Artificial Intelligence (AAAI), Workshop on New Research Problems for Machine Learning, 2000.
[37] Dietterich T G. Ensemble methods in machine learning[J]. Proceedings of the First International Workshop on Multiple Classifier Systems, 2000, 1857: 1–15.
[38] Rogova G. Combining the results of several neural networks classifiers[J]. Neural Networks, 1994, 7(5): 777–781.
[39] 马少平, 朱小燕. 人工智能 [M]. 北京: 清华大学出版社, 2004.
[40] Al-Ani A, Deriche M. A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence[J]. Jair, 2002, 17: 333–361.
[41] Chen K, Wang L, Chi H. Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification[J]. International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), 1997, 11(3): 417–445.