doi:10.3969/j.issn.1000-436x.2014.09.004

轻量级的自学习网页分类方法

沙泓州[1,2,3]，周舟[2,3]，刘庆云[2,3]，秦鹏[2,3]
（1. 北京邮电大学 计算机学院，北京 100876；2. 中国科学院 信息工程研究所，北京 100093；
3. 信息内容安全技术国家工程实验室，北京 100093）

摘 要：提出了一种自学习的轻量级网页分类方法 SLW。SLW 首次引入了访问关系的概念，使其具有反馈和自学习的特点。SLW 从已有的恶意网页集合出发，自动发现可信度低的用户和对应的访问关系，从而进一步利用低可信度用户对其他网页的访问关系来发现未知的恶意网址集合。实验结果表明，在相同数据集上，相比于传统检测方法，SLW 方法可以显著提高恶意网页检测效果，大幅降低平均检测时间。

关键词：URL 分类；黑名单；访问关系；恶意网页；网页评价
中图分类号：TP393.8 **文献标识码**：A **文章编号**：1000-436X(2014)09-0032-08

Light-weight self-learning approach for URL classification

SHA Hong-zhou[1, 2, 3], ZHOU Zhou[2, 3], LIU Qing-yun[2, 3], QIN Peng[2, 3]
(1. Department of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China;
3. National Engineering Laboratory for Information Security Technology, Beijing 100093, China)

Abstract: A self-learning light-wight (SLW) is proposed. SLW is the first to introduce access relations and have the characteristics of feedback and self-learning. SLW approach starts from the seed set which includes known malicious pages. Then, it automatically figures out users with low credibility based on the seed set and the visit relation database. Finally, the access records of these users are used to identify other malicious pages. Experimental results indicate that SLW approach can significantly improve the efficiency of malicious pages detection and reduce the average detection time compared with other conventional methods.

Key words: URL classification; blacklist; access relation; malicious Web page; Web page evaluation

1 引言

作为一个开放式的共享平台，互联网在为人们提供便利的同时也为一些不法分子收集个人隐私信息、组织犯罪活动创造了新的机会。在一些已知的网络犯罪活动中，包含恶意代码[1]和网上诱骗[2]的网页（即恶意网页[3]，包括钓鱼网站、网页木马、色情网站等）常常扮演着十分重要的角色。卡巴斯基的统计数据显示[4]，恶意网页在 87.36%的网络攻击行为中出现并发挥作用。这类网页或者在用户不知情的情况下将恶意代码自动安装到用户的计算机中，或者协助不法分子冒充他人骗取用户个人信息及其他敏感信息。Google[5]的统计数据表明，平均每天拦截新的恶意网页数高达 9 500 个。这些恶意网页的存在，对 Web 的安全应用构成极大的威胁。

为了保护用户的计算机免受恶意网站攻击，一些主流浏览器（例如，IE 浏览器的 SmartScreen 筛选器[6]及谷歌浏览器的 safebrowsing[7]等）往往采用内置恶意网址列表的方法为用户提供安全服务。具体方法为：浏览器首先通过自动检测和人工举报的方式获得一份类似黑名单的恶意网址列表；然后，

收稿日期：2014-07-21；**修回日期**：2014-08-29
基金项目：国家高技术研究发展计划（“863”计划）基金资助项目（2011AA010703）；国家自然科学基金资助项目（61070026）
Foundation Items: The National High Technology Research and Development Program of China(863 Program) (2011AA010703); The National Natural Science Foundation of China (61070026)

在用户浏览某个网页前，浏览器通过扫描内置的恶意网址列表来判断该网页的 URL（uniform resource locator）是否为恶意网页，如果确定为恶意网页后，浏览器将向用户发出警告，以提示用户防止恶意代码和网上诱骗的攻击。这类安全服务的原理简单且易于实现，因此在工业界被广泛应用。然而，随着互联网的发展和网络攻击方式的层出不穷，这种方法逐渐面临一些新的挑战。

1) 大规模的网络数据环境。作为一个开放式的共享平台，互联网不断发展，网页规模不断扩大，新的网页不断涌现。由于第三方专业服务机构提供的恶意网址列表的更新速度远远跟不上恶意网页的更新速度，容易出现恶意网页漏判的情况。

2) 网页隐匿技术的使用。随着传统方法的广泛应用，很多攻击者开始寻找并逐步使用网页隐匿技术 [8]来躲避检查。例如，一些恶意站点通过伪装网页内容来逃避启发式爬虫的自动检测，以避免被加入恶意网址列表，进而常常导致网页错判的情况。因此，随着恶意网页隐藏技术的逐步应用，发现新恶意网页的难度也在逐步加大。

3) 不均衡的数据集特点。少量恶意网页往往淹没在海量的正常网页中。例如，Google 每天检查数亿的 URL 只能发现约 9 500 个不安全的站点 [7]，大部分网页的分析价值低，并且检测需要消耗时间长。同时，由于自动分析和人工报告需要消耗大量的计算资源，如果对每个网页都进行分析，资源利用率将十分低。

因此，如何设计一套自动化的工具快速准确地将新出现的恶意网站及其 URL 和其他大量正常网页区分开来成为一个迫切需要解决的问题。

针对此问题，本文提出了一种基于访问关系的 URL 分类方法 SLW (self-learning light-weight approach)。SLW 方法首次将存储在访问日志中的访问关系引入网页信誉评价问题中，以弥补恶意网页漏判和错判可能产生的不良影响。从文献[8]实验中观察到的现象可知，对频繁访问恶意网站的用户而言，他们对其他网站的访问交集也是可疑的。因此，通过引入访问关系的概念，可以有效地发现潜在的恶意网页。实验结果表明，在访问关系的基础上，SLW 结合黑名单的方法来区别不同的 URL，具有如下优点。

1) 自学习。和传统的基于黑名单的方法不同，系统可以通过用户对黑名单列表的访问记录确定哪些用户的可信度较低。然后，系统将利用低可信度用户的访问交集发现可疑 URL 集合。最后，对所有可疑 URL 进行逐一分析和检查，判断其是否属于恶意网址集合。如果是恶意网址，则将其添加到黑名单中，以保证黑名单的完整性和可用性，进而可找到新的低可信度用户。通过这一途径，SLW 方法可以适应于用户访问行为的变化以及恶意网页的更新。

2) 轻量级。在 Bando 等[9,10]工作中，常常需要抓取和检查大量的网页或 URL。但这些动作通常消耗大量的计算资源，其中绝大多数资源被浪费在良性 URL 的排查上。SLW 通过保存访问关系限制了 URL 检查范围，从而节省了大量的计算资源和时间开销。

本文通过抓取高校网关中的访问日志来验证这一方法。实验结果表明，和传统网页分类方法相比，SLW 方法能够有效提升检测中恶意网页所占比例(从 1.09%提升至 1.38%~1.94%)和识别效率（恶意网页检测所需的平均时间降低 6.36%~33.89%），以便快速有效地识别恶意网页。

2　相关工作

针对恶意网页分类问题，国内外学者进行了广泛的研究，如基于黑名单的方法[3,11]、基于深度包检测的方法[12]和基于机器学习的方法[13,14]等。

基于黑名单的方法相对比较简单，易于实现。它首先对恶意 URL 进行标注，然后利用字符串匹配等技术实现恶意 URL 的识别。而恶意 URL 标注可以通过人工标注和自动标注[11]的方法完成。人工标注比较准确，但需要标注人员有专业的领域知识，并且耗时较长，只适合低速、小规模的网络环境。而自动标注多利用启发式的网络爬虫技术 [11]进行标注，此类标注方法易于实现，可以进行并行化处理，但不够准确。这主要是因为很多恶意网页或者使用隐藏技术逃避检测；或者特征不够明显，出现标注错误的情况。

为了解决基于黑名单方法存在的网站漏判问题，Pak 等[12]提出基于内容检测的分类方法。相比基于黑名单的方法，这类方法更加准确，能够发现更多的恶意网页，且易于并行化处理。但在执行内容检测时，由于分析处理的网页内容较多，数据格式复杂，实践这一方法需要消耗很多计算资源和时间。此外，自动化分析依赖于一套由具体领域的专

业知识转换成的识别规则。这些领域知识的主观程度高，获取困难。

为了降低计算资源的消耗和减少对领域知识的依赖，Ma 等 [11]以 URL 词汇特征和主机特征为基础建立统一的分类模型（例如，SVM 模型 [15]等），进而根据已有标注集合识别恶意 URL。此类方法通过选取有代表性的语法特征进行判断，并不依赖特定领域的专业知识，分类速度快，资源占用少，是目前主流的 URL 分类方法。但它分类的准确性主要依赖于样本集的选取，并且部分主机特征受网络延迟影响较大。

上述方法从不同层面对 URL 分类问题进行了分析。在前人的研究基础[16~19]上，本文提出一种新的 URL 分类方法 SLW。该方法通过混合使用网页黑名单以及“用户—网站”间的访问关系，实现了网页的轻量级分类，并且有效提高黑名单的可扩展性，使其能够应用于动态网络环境中。与人工报告和启发式爬虫相比，SLW 提供了一种更好的动态黑名单的产生方法。一方面，它通过使用访问关系，限制了恶意 URL 的检查范围，避免了对访问流中所有未知 URL（规模过亿）的详细检查。另一方面，和人工报告相比，它需要更少的人为干预。

3 网页信誉评价

3.1 基本概念

根据本文的应用环境，对一些文中即将用到的概念定义如下。

定义 1 访问集合是指访问者所访问的所有网页资源所构成的集合，它包含了访问者访问过的所有网页资源。如图 1 所示，访问者 A 的访问集合 $U_A=\{p_1,p_2\}$。

定义 2 访问交集是指 2 个或多个访问集合的交集。一般地，对于给定的 2 个访问集合 A 和 B 的交集，是指含有所有既属于 A 又属于 B 的元素，而没有其他元素的访问集合。

定义 3 访问关系是一种建立在访问者和网页资源之间的关系，是访问者通过访问网页资源产生的一种对应关系。如图 1 所示，访问者 A 和网页资源 p_1 的访问关系 $r=<A,p_1>$。

定义 4 用户行为可信度是指恶意网页识别系统对访问用户根据用户此前访问记录而产生的信任程度。通常来说，从不访问恶意网页的用户行为可信度较高，经常访问恶意网页的用户行为可信度较低。因此，可以通过用户过去一段时间内的访问记录来预测用户未来行为的可信任程度。在区间 t (t 视具体应用而定，如 3 个月)内，假设用户 i 的访问次数为 n，则用户行为可信度可以通过式(1)进行计算。

$$D_i=\frac{\sum_{k=1}^{n}V(i,p_k)}{n} \tag{1}$$

其中，$V(i, p_k)$表示用户 i 第 k 次访问的网页 p 的评价结果，它通过式(2)进行计算。

$$V(i,p_k)=\begin{cases}1, & p_k\in malicious\\ 0, & p_k\in benign\end{cases} \tag{2}$$

本系统使用 Google 安全浏览 API [5]提供的恶意网页列表作为判断网页评价结果的依据。

式(1)和式(2)表明，用户可信度是通过对用户的访问次数和网页本身的评价结果加权计算进行度量的。

如图 1 所示，将 2 个用户(A, B)和 2 个网页资源(p_1，p_2)的访问关系进行了抽象，其中，集合 U_A 表示用户 A 的访问集合，U_B 表示用户 B 的访问集合，U_{AB} 表示用户 A 和用户 B 的访问交集。

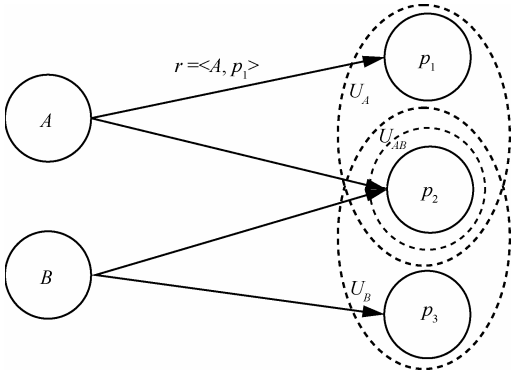

图 1 用户访问网页资源关系示例

3.2 网页评价

当用户访问一个网页时，网页评价结果是浏览器对用户行为执行不同操作（发出/不发出警示信息）的主要依据。浏览器可以依据恶意网址列表对用户进行直接评价。为了补充和完善评价信息，根据用户的浏览行为增加一种评价方式：访问行为评价[19]。

3.2.1 浏览器直接评价

浏览器根据先验知识（例如，恶意网址列表等）对网页进行直接评价，这种评价方式比较简单，使用方便，应用范围广泛，其不足之处在于网页评价分级方式比较粗略。此外，由于恶意网址列表的更

新速度较慢，部分新出现的恶意网页将无法得到正确评价结果。

3.2.2 访问行为评价

在用户浏览网页的过程中，可以利用用户的浏览行为对网页形成评价。一些行为评价方式主要使用用户访问网站的次数来评价网站的健康程度。这种方式没有考虑不同用户之间的信誉差异对网站评分结果的影响。通过在 3.2 节引入用户可信度这一概念，恶意网页识别系统记录访问某个资源的用户可信度和访问次数，以此作为参数获得对该网页资源的评价。网页 p 的评价结果可以通过式(3)来计算

$$V(p)=\frac{\sum_{k=1}^{m} D_k T(k,p)}{\sum_{k=1}^{m} T(k,p)} \tag{3}$$

其中，D_k 表示用户 k 的可信度，$T(k, p)$表示用户 k 对网页 p 的访问次数，m 表示用户总数。

3.2.3 访问行为评价因素

根据用户浏览行为的规律以及评价值的计算需要，为访问行为评价增加 3 个影响因素：最小访问间隔 I (interval)、访问日志保存周期 LC (log cycle)和行为评价更新周期 EC (evaluation cycle)，用以保证行为评价的准确性和有效性。

1) 最小访问间隔 I

行为评价随着访问行为的变化而变化。单个用户对网页的重复访问必须达到一定的时间间隔，统计得到的访问次数才有意义。如果忽视这一参数，则用户可以通过不断访问某个页面达到提升其评价值的目的。根据文献[19]的经验，本文将最小访问间隔 I 设置为一天，以防止出现上述情况。

2) 访问日志保存周期 LC

增加一定的访问日志对分析和识别恶意网页是必要的，但持续增加访问日志则会大幅增加存储负担，降低计算效率。因此确定合适的访问日志保存周期，只保存一段时间内的访问日志，能够避免造成存储和计算压力。本文中访问日志保存周期 LC 取 30 天，该周期可根据实际存储能力和计算资源进行相应调整。

3) 行为评价更新周期 EC

根据访问行为的变化，行为评价需要及时更新，以适应用户的访问需求。但是，对一些频繁访问的热门网址和站点而言，大量不同用户的连续访问可能造成行为评价不稳定和计算资源的浪费。为避免这种情况，本文将行为评价更新周期 EC 设置为 6 h。该周期 EC 可根据实际用户访问特点进行调整。

此外，对于同一个网页而言，如果浏览器给出了直接评价，则不再记录访问行为评价。否则，记录其访问行为评价结果。

3.3 基于访问关系的网页信誉评价

式(3)给出了一个网页综合评价值的计算方法。依据式(3)，以访问关系为基础，将用户可信度作为权重，计算网页综合评价值。表 1 为网页评价结果的示例。对比多个网页的评价结果，可以发现一个可信度高的用户多次访问某一网页，在该网页的综合评价中的贡献会多一些。

4 自学习轻量级分类方法 SLW

4.1 SLW 方法概述

URL 分类可以描述为一个二分类问题，其中阳性例子是可疑 URL，阴性例子是正常 URL。解决 URL 分类问题的关键是正确划分可疑 URL 和正常 URL。

SLW 方法主要依据黑名单和访问关系来划分 URL。具体过程如下：首先，使用黑名单和收集到的访问关系查明可疑用户集合并收集他们的访问

表 1　网页评价结果示例

用户 ID	可信度	访问次数				
		URL1	URL2	URL3	URL4	URL5
ID1	1.0	5	N/A	N/A	4	N/A
ID2	0.9	3	N/A	N/A	2	N/A
ID3	0.8	2	1	N/A	1	2
ID4	0.7	1	N/A	2	2	2
ID5	0.2	1	5	3	N/A	N/A
综合评价值	N/A	0.85	0.3	0.4	0.89	0.75

日志；其次，通过对这些访问日志中的网页进行评价，可以发现更多的可疑 URL；最后，对这些可疑 URL 进行详细的分析检查，以便准确识别恶意 URL。和其他方法相比，该方法需要增加额外的存储空间以保存部分访问日志，但保存部分访问日志可以有效地缩小内容检测范围并帮助发现潜在的恶意 URL。

4.2 SLW 架构和工作流程

图 2 展示了 SLW 方法的架构，从整体角度分析，SLW 方法包含 2 个主要步骤。

在第 1 步中，采用已有成熟的分析技术（如 DPI、关键词扫描等）对日志中出现的每个网页进行深入分析。通过这一步骤，SLW 可以积累一些恶意网页作为“种子”。在第 2 步中，SLW 方法从这些“种子”出发，对访问过这些已知恶意网页的用户的可信度进行打分。对这些用户进行分类筛选出一部分可信度较低的用户，利用它们的访问日志以发现潜在的恶意网页。下面详细叙述了 SLW 方法中使用的每个组件。

种子。本文的种子是指预先检测出来的恶意网页的集合。作为低可信度用户追踪器的输入，种子的质量对整个追踪过程至关重要。种子是由专业分析器产生的，当分析器分析出一个恶意网页时，就可以将这个恶意网页加入种子集合中，种子集合中常常包含 2 类网页。第 1 类网页是由攻击者或犯罪分子直接建立的，这些网页或者直接链接到一个恶意程序，或者包含一段可以在特定条件下执行的恶意代码。此外，这些网页之间常常存在链接关系，以便提高成功入侵的几率。第 2 类网页则属于良性网站的网页，和正常的良性网页不同，它们已经被攻击者挂马，通常会嵌套一段跳转程序将用户引导至恶意站点。SLW 方法通过把这 2 类网页加入到种子集合中，以便跟踪这些低可信度的访问者，从而从他们的访问日志中发现潜在的恶意网页。

低可信度用户追踪器。低可信度用户追踪器是 SLW 方法的核心。它的输入是种子（包含恶意网页的集合）以及“多对多”的访问关系。基于对已知恶意网页和访问关系的分析，低可信度用户追踪器产生低可信度用户的集合（如图 2 所示的过程①）。用户的可信度可以通过式(1)计算得到。低可信度用户往往访问过已知的恶意站点并且今后访问这些站点或者类似网页的可能性较大。因此，将低可信度用户集合提交至可疑 URL 收集器。通过在网络流中标识低可信度的用户，SLW 方法就有可能发现它们的访问历史并识别其他恶意站点。

可疑 URL 收集器。可疑 URL 收集器基于低可信度用户集合对他们的访问交集进行收集（如图 2 所示的过程②），即只有多个低可信度的用户访问的 URL 才会被收集。这个部件的功能是产生一个可疑 URL 的集合，并将它们发送给一组专业分析器。

专业分析器。专业分析器主要由 Google 提供的 safebrowsing 黑名单[5]组成。这个黑名单已经被 Google 用来实时处理数以亿计的网页,并有 API 提供给外部调用者使用。此外，它不断更新并且其误判率很低。

预处理。预处理模块的主要功能是对输入的 URL 中的已知良性 URL 进行过滤，是可选的。由于没有收集到有关良性 URL 的先验知识，因此没有具体实现预处理模块，但这不影响实验最终结果。

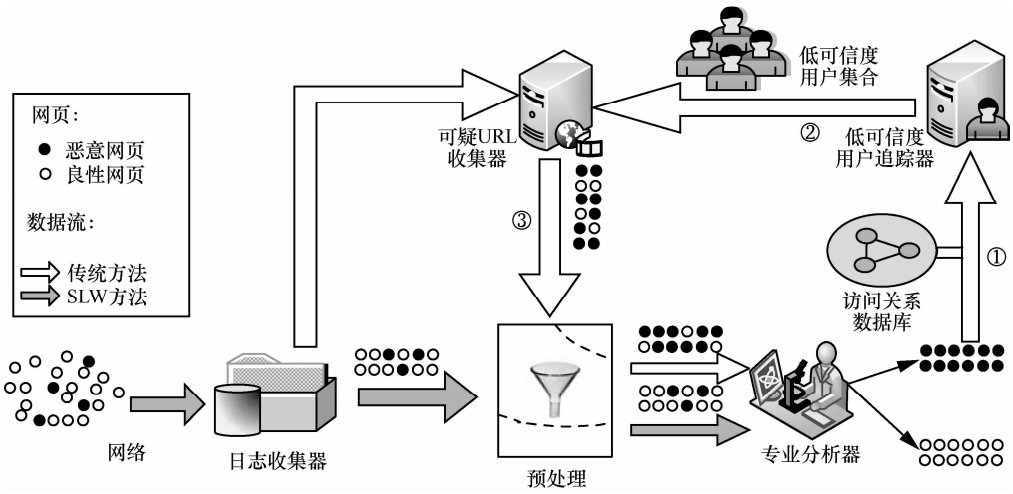

图 2 SLW 方法的架构

5　实验结果与分析

通过使用真实校园网关捕获的数据集来验证 SLW 方法的性能。首先介绍实验的评价指标，然后对数据集和实验环境进行分析，最后对实验结果进行讨论。

5.1　评价指标

由于真实校园网关捕获的数据集规模较大，并且呈现严重的不均衡性（数据不均衡性主要是指恶意网页资源和良性网页资源的数量相差很大，通常在一个数量级以上），不适合用传统的准确率和召回率来衡量本系统的好坏。因此，本文参考 Invernizzi 等的工作 [20]，选取了另外 2 个指标来验证系统的效果：浓度 (density)和扩展度 (expansion)。

浓度是指新发现的可疑 URL 中真正恶意的 URL 所占的比例。例如，如果一个可疑 URL 收集器提交 100 个可疑 URL 给专业分析器，并且其中 10 个 URL 最终被认定是恶意的，那么其浓度为 0.1。分类结果的浓度高表明分类系统的资源利用率高。

扩展度是指系统根据一个恶意网页 URL 分析可以得到的平均恶意 URL 数。扩展度高低体现了恶意网页资源是否得到有效利用。

浓度和扩展度是此消彼长的关系，需要根据实际情况加以权衡。考虑到恶意网页识别的实际应用环境，特别是所采用不均衡的数据集，获得更高的浓度对系统更加重要。

5.2　数据集

本实验在中国网站排名[21]上选取了 10 类共 6 353 个网址，并将它们作为良性 URL 的代表。此外，以现有的 URL 黑名单[6]作为恶意 URL 标注工具。通过在高校网关中捕获 130 GB 网络流量作为数据源，数据集的统计信息如表 2 所示。其时间范围是在 2013 年 3 月 22 日至 2013 年 4 月 8 日。其中，一个典型的访问行为如表 3 所示。本实验的实验环境为一台 8 核 2.13 GHz 主频 CPU，内存为 16 GB 内存的服务器。

表 2　数据集信息统计

指标	总数	恶意用户（网页）数	恶意用户（网页）占比/%
访问记录	12 347 243	98 134	0.79
用户	171 146	2 475	1.45
网页	3 155 234	40 581	1.29

表 3　用户访问行为的示例

用户 IP	时间	访问 URL
210.242.14.29	Mar22 14:52:47	ks.cn.yahoo.com/question/1.html

5.3　实验结果

本文设计了 2 个实验来分别验证 SLW 方法的浓度、扩展度和时间性能。第 1 个实验通过和传统检查方法进行对比来验证 SLW 方法发现恶意 URL 的能力。首先，对很小的一部分(实验中取 URL 总数的 0.2%，即 6 310 个 URL)网址进行深入分析和检查（如图 1 所示），得到一部分恶意 URL（又称“种子”，下同，本文中为 67 个 URL）；其次，充分使用“种子”来识别可信度低的用户并继续产生新的恶意 URL（实验组 1~3）；最后，将 SLW 方法同其他传统检测方法（半数检查是指检查一半的访问日志以发现恶意网址，全面检查是指检查全部访问日志以发现恶意网址，如实验组 4~5）进行比较，以分析 SLW 方法的优势和劣势。

如表 4 所示，初始阶段，SLW 从浓度为 1.06% 的 URL 库里，分析并识别出 67 个恶意 URL 作为“种子”。在对“种子”分析的基础上，SLW 提交了 18 440 条 URL 给专业分析器，其中 254 条 URL 被最终认定为恶意。由图 3 可知，其浓度由 1.29% 提高到 1.94%。此外，与全面检查访问记录相比，只对可信度低用户的访问日志进行分析，其恶意 URL 浓度分别上升 6.97%~50.38%（如图 3 所示）。即用户的可信度越低，其访问日志中包含的恶意 URL 浓度越高。

表 4　浓度和扩展度实验数据

实验组别	方法/步骤	种子网页数	低可信度用户数量(比例)	分析 URL 数	恶意 URL 数
1	SLW	67	223(10%)	11 215	218
2	SLW	67	1 115(50%)	16 974	247
3	SLW	67	2 230(100%)	18 440	254
4	半数检查	N/A	N/A	1 606 014	17 511
5	全面检查	N/A	N/A	3 155 234	40 581

表5 SLW方法时间开销数据

实验组别	方法/步骤	低可信度用户数量(比例)	恶意URL数	处理时延/s	平均时间/ms	降低比例/%
1	SLW	223(10%)	218	14.46	66.3	33.89
2	SLW	1 115(50%)	247	21.89	88.6	11.66
3	SLW	2 230(100%)	254	23.79	93.64	6.36
4	半数检查	N/A	17 511	2 071.63	118.3	−17.96
5	全面检查	N/A	40 581	4 070.12	100.29	N/A

扩展度实验。如表4和图3所示，只分析低可信度用户的日志，其实际产生恶意URL的数量初始恶意URL数量的3.25倍以上。

表5比较了不同方法的时间性能。如表5所示，检查低可信度用户访问记录是发现恶意URL的一种有效方法。采用这种方法可以减少约33.89%的平均检测时间。特别地，如果系统选取恶意用户的比例比较小，其用于发现一个恶意URL的平均时间将大大缩短。这种情况出现的可能原因是每次实验进行前，将用户按照可信度从低到高进行了排序。

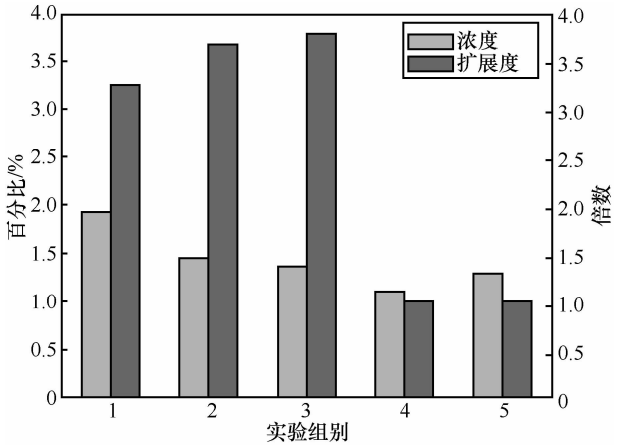

图3 浓度扩展度实验对比

5.4 分析和讨论

实验结果显示，相比于全部检查，SLW方法可以显著提高恶意URL的浓度（1.29% vs 1.94%），从而大幅度（33.89%）降低平均检测时间。此外，给定一定数量的恶意网页，SLW方法有能力发现大量额外的恶意网页。相比全面检查、检查同样数量的URL，SLW可以发现3倍以上的恶意网页。这一对比结果显示，SLW在提高URL浓度，降低平均检测时间方面具有较大优势。同时，该方法需要部分恶意网页作为“种子”以产生更多的恶意网页。因此，SLW方法的局限性在于，恶意网页识别能力的好坏依赖于初始种子的选择。为缓解这一局限性带来的问题，可以扩大“种子”恶意网页的选取范围，或者提升“种子”恶意网页质量。

6 结束语

恶意网页（例如，钓鱼网站[18]、网页木马[1]、色情站点等）是互联网健康发展的一个重要威胁。识别这类站点对抵制犯罪分子网络活动具有重要意义。然而，由于网页规模的不断扩大，数据集的不均衡性以及网页隐藏技术的使用等原因，网页分类问题变得更具挑战性。为了解决这一问题，本文提出了SLW方法以构建具有反馈和自学习机制的轻量级网址分类系统。实验结果表明，相比传统检测方法，SLW方法可以显著提高恶意网页浓度，大幅降低平均检测时间，并且具有反馈和自学习的特点。

在未来的工作中，将重点围绕以下几个方面开展研究：首先，引入URL白名单机制以避免一些不必要的资源消耗；其次，逐步引入URL的静态特征和语法规则作为分类依据，以进一步提升分类效果；最后，研究如何提高分类算法在动态环境下的顽健性。

参考文献：

[1] 诸葛建伟, 韩心慧等. HoneyBow: 一个基于高交互式蜜罐技术的恶意代码自动捕获器[J]. 通信学报, 2007, 12(28):8-13.
ZHU GE J W, HAN X H, *et al*. HoneyBow: an automated malware collection tool based on the high-interaction honeypot principle[J]. Journal of Communications, 2007, 12(28):8-13.

[2] PRAKASH P, KUMAR M, KOMPELLA R R, *et al*. Phishnet: predictive blacklisting to detect phishing attacks[A]. Proceedings of INFOCOM[C]. San Diego, CA, USA, 2010. 1-5.

[3] AKIYAMA M, YAGI T, ITOH M. Searching structural neighborhood of malicious urls to improve blacklisting[A]. Proceeding of the 11th Symposium on Applications and the Internet (SAINT)[C]. Munich, Germany, 2011.1-10.

[4] Kapaersky security bulletin. statistics 2012[EB/OL]. http://www.securelist.com/en/analysis/204792255/Kapersky, 2014.

[5] REPORT G T. Making the Web safer[EB/OL]. http://www.google.com/transparencyreport/safebrowsing/?hl=en, 2014.

[6] Smartscreen filter[EB/OL].http://windows.microsoft.com/zh-CN/ inter-

net-explorer/use- smartscreen-filter#ie=ie-9, 2014.

[7] Google chrome and google safe browsing[EB/OL]. http://www.google.com/chrome/intl/zh-cn/more/security.html, 2014.

[8] KOLBITSCH C, LIVSHITS B, ZORN B, *et al*. Rozzle: De-cloaking internet malware[A]. IEEE Symposium on Security and Privacy (S&P)[C]. San Francisco, USA, 2012. 443-457.

[9] BANDO M, ARTAN N S, CHAO H J. Scalable look ahead regular expression detection system for deep packet inspection[J]. Transactions on Networking, IEEE/ACM, 2012, 20(3):699-714.

[10] JIANG J, SONG X, YU N, *et al*. Focus: learning to crawl Web forums[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(6): 1293-1306.

[11] MA J, SAUL L K, SAVAGE S, *et al*. Beyond blacklists: learning to detect malicious Web sites from suspicious URL[A]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Paris, France, 2009.1245-1254.

[12] PAK W, CHOI Y J. High-performance packet classification for network-device platforms[J]. Communications Letters, IEEE, 2013, 17(6): 1252-1255.

[13] ZHANG F G. Preventing recommendation attack in trust-based recommender systems [J]. Journal of Computer Science and Technology, 2011, 26(5): 823-828.

[14] MA J, SAUL L K, SAVAGE S, *et al*. Learning to detect malicious URLs[J]. Transactions on Intelligent Systems and Technology, 2011, 2(3):30.

[15] CHANG C C, LIN C J. LIBSVM: a library for support vector machines [J]. Transactions on Intelligent Systems and Technology, ACM, 2011, 2(27):1-27.

[16] LEE S, KIM J. Warningbird: a near real-time detection system for suspicious urls in twitter stream [J]. Transactions on Dependable and Secure Computing, IEEE, 2013,10(3): 183-195.

[17] TSANG P P, KAPADIA A, CORNELIUS C, *et al*. Nymble: blocking misbehaving users in anonymizing networks[J]. Transactions on Dependable and Secure Computing, IEEE, 2011, 8(2):256-269.

[18] LE A, MARKOPOULOU A, FALOUTSOS M. Phishdef: URL names say it all[A]. Proceeding of the 30th IEEE International Conference on Computer Communications (IEEE INFOCOM 2011)[C]. Shanghai, China, 2011.191-195.

[19] 刘昕, 贾春福, 刘国友等. 基于社会信任的恶意网页协防机制[J]. 通信学报, 2013, 12(33):11-18.
LIU X, JIA C F, LIU G Y, *et al*. Collaborative defending scheme against malicious Web pages based on social trust[J]. Journal on Communications, 2013, 12(33):11-18.

[20] INVERNIZZ I, LUC A, *et al*. Evilseed: a guided approach to finding malicious Web pages[A]. IEEE Symposium on Security and Privacy (S&P)[C]. San Francisco, USA, 2012. 428- 442.

[21] China Webmaster[EB/OL]. http://top.chinaz.com/, 2014.

作者简介：

沙泓州（1988-），男，江苏淮安人，北京邮电大学博士生，主要研究方向为网络与信息安全、数据挖掘。

周舟[通信作者]（1983-），男，湖南长沙人，博士，中国科学院信息工程研究所助理研究员，主要研究方向为网络安全、高性能网络。E-mail:zhouzhou@iie.al.cn。

刘庆云（1980-），男，河北衡水人，硕士，中国科学院信息工程研究所高级工程师，主要研究方向为信息安全、网络安全。

秦鹏（1984-），男，内蒙古乌海人，硕士，中国科学院信息工程研究所工程师，主要研究方向为信息安全、网络安全。