



基于改进的 TrustRank 算法的钓鱼网站检测

韩浩, 刘博文, 林果园

(中国矿业大学计算机科学与技术学院, 江苏 徐州 221116)

摘要: 钓鱼检测方式一般只是通过比较网页间特征的相似度判断钓鱼网站, 容易被攻击者根据特征提取过程反检测。因此, 提出依据网页之间的关系来检测钓鱼网站, 通过设立网页集合、结合钓鱼网站与其他网站间的链接关系改进 TrustRank 算法来检测钓鱼网站。实验证明, 改进的 TrustRank 算法可以将钓鱼网站的信任值集中在一个范围内, 并且与其他方法相比具有较低的误判率、漏判率和较高的速度, 可以很好地检测钓鱼网站。

关键词: 钓鱼网站; TrustRank; 网页关系; 链接关系

中图分类号: TP182

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018046

Detection of phishing websites based on the improved TrustRank algorithm

HAN Hao, LIU Bowen, LIN Guoyuan

School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

Abstract: Anti-phishing methods just generally detect phishing sites based on comparing the similarity of pages features, which could be anti-detected when attackers are clear about the extraction process of features. Therefore, a method based on web pages relationship was proposed to detect phishing websites. According to this method, phishing websites were detected by an improved TrustRank algorithm which set up a collection of web pages and used the link relationship between phishing websites and other websites. Experiments show that the improved TrustRank algorithm can focus the trust values of phishing websites on some range with the lower false positive rate, the lower false negative rate and the higher speed compared to other methods.

Key words: phishing website, TrustRank, web page relationship, link relationship

1 引言

钓鱼网站是网络安全中的一个重大问题, 因钓鱼网站造成的经济损失和社会问题逐年增加。

2017年6月份, 由CNNIC牵头筹建的互联网域名管理技术国家工程实验室与国际反钓鱼工作组(APWG)、中国反钓鱼网站联盟(APAC)联合发布的《全球中文钓鱼网站现状统计分析报告(2016年)》^[1]

收稿日期: 2017-10-13; 修回日期: 2018-01-15

通信作者: 林果园, lingy@cumt.edu.cn

基金项目: 江苏省产学研前瞻性联合研究项目(No. BY2016026-04)

Foundation Item: Production and Research Prospective Joint Research Project of Jiangsu Province(No. BY2016026-04)

显示, 2016 年我国钓鱼网站数量为 147 211 例, 比 2015 年增长 150.96%。

常见的检测钓鱼网站的方法有基于页面特征的检测^[2,3]、基于 URL 的检测^[4,5]、基于图像的检测^[6,7]和结合机器学习进行检测^[8-10]。除了基于 URL 的检测方法, 其他方法都会处理大量的多媒体内容而影响检测效率, 并且随着网络钓鱼技术的发展, 传统的检测方法虽然能够有效地检测出钓鱼网站, 但容易被新的钓鱼技术反检测而影响检测的准确率。

不法分子往往通过电子邮件、即时通信工具、短信、论坛或者搜索页面等使受害者访问钓鱼网站^[11]。从受害者访问不法分子的推广链接到受害者受到损失过程中, 受害者所访问的网站中会出现以访问的钓鱼网站为中心链接在一起网站集合。而钓鱼者又不得不推广钓鱼网站以使得更多的人访问钓鱼网站, 因此以钓鱼网站为中心分析其和其他网页之间的链接关系从而检测钓鱼网站的方法是钓鱼者不容易躲避的。

TrustRank^[12]是通过检测网页间的链接关系并量化出网页的信任值, 从而判断网页性质的算法, 谷歌应用 TrustRank 屏蔽搜索结果中的垃圾页面。虽然钓鱼网站也是垃圾网站, 但是 TrustRank 所考虑的因素仅是网页之间的链接关系, 并不是针对钓鱼网站的特征, 因此不能够直接检测钓鱼网站。由此, 需要结合钓鱼网站的特征改进 TrustRank。

将可以判断一个网站是否是钓鱼网站的网站特征称为网站的钓鱼特征。对于钓鱼网站特征的选取也可以从 TrustRank 的角度选取, 即分析网站之间的链接关系, 这样贴近于 TrustRank 的算法思想, 可以使选取的钓鱼特征更有效地结合 TrustRank 算法。因此从以下 3 个方面考虑提取网站的钓鱼特征改进 TrustRank 算法:

- TrustRank 是针对链接的网页排序算法, 由此依据网页间的链接关系归纳出钓鱼网站和其他网站的 3 种链接关系, 并量化这

3 种链接关系作为网站的钓鱼特征;

- 由于钓鱼网站普遍存在时间短, 因此选取网站的注册时间作为网站的一个钓鱼特征;
- 为了避免误判新出现的正规网站, 考虑到 URL 检测的简单性和新出现的网站的 URL 并不会和黑白名单中的网站 URL 相似度过高, 因此, 将选取网页的 URL 与白名单中网页的 URL 相似度作为另一个网站的钓鱼特征。

2 TrustRank 基本原理

在 TrustRank 算法中, 将网络中的网页依据相互链接的关系组织成一个 Web 图^[13] $G_w = \langle P, R \rangle$, 其中, P 是网络中的网页集合, 也就是 Web 图中的顶点集合, R 是网页与网页之间的链接集合。

在 Web 图 G_w 中, 为了表示网页之间的链接关系定义 $R_{p,q} \in R$ 表示网页 p 到网页 q 的链接。用 No_p 表示网页 p 指出链接的数量, 即 Web 图中网页 p 的出度, 用 O_{pi} 表示网页 p 指出的链接, 其中, $i \in (0, No_p)$ 。用 N_{in_p} 表示指向网页 p 的链接的数量, 即 Web 图中网页 p 的入度, 用 IN_{pi} 表示指向网页 p 的链接, 其中, $i \in (0, N_{in_p})$ 。

TrustRank 算法是一种半自动分类的方法, 用于描述网页的信任程度从而对网页进行排名。TrustRank 假设好的网站很少会链接到坏的网站, 但是坏的网站会链接到好的网站以提高自己的信任度。TrustRank 算法首先设置一个好的网站种子集 Sp^g , 即认为集合中的网站是信任度高的网站, 一般将这些网页的初始信任度设置为 1, 之后将垃圾网页的信任度设置为 0, 其他网站设置为 $\frac{1}{2}$ 。

若种子集中的某网页到一个网页所有路径上都不包含垃圾网页, 则将此网页信任度设置为 1。然后采取信任分裂^[14]的传播方式传播信任值, 信任分裂是如果一个网页的信任值为 1, 且指向 n 个网



页, 那么它所指向的每一个网页得到的信任值为 $\frac{1}{n}$, 由此一个网页的信任值是它从所有链入到它的网页得到的信任值的总和, 信任分裂如图 1 所示。

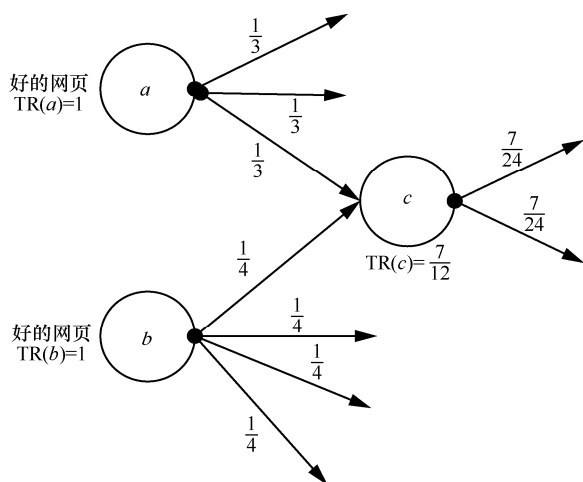


图 1 信任分裂

图 1 中的 TR 是页面的信任值, TrustRank 算法结合信任分裂由式 (1) 计算页面的 TrustRank 值。

$$TR(p) = (1-d) + d \sum_{i=1}^{N_{in_p}} \frac{TR(q_i)}{No_{q_i}} \quad (1)$$

其中, $TR(p)$ 是页面的 TrustRank 值, d 为阻尼系数; q_i 是含有指向网页 p 链接的第 i 个网页。

3 基于钓鱼特征改进的 TrustRank 算法

TrustRank 算法假设好网站不会链接到坏的

网站, 但是通常钓鱼网站的链接会出现在知名网站中, 也就是好网站, 诸如论坛、贴吧等。对钓鱼网站分析得出, 钓鱼网站与其他网站的关系有推广、假链接和恶链接 3 种: 推广关系是钓鱼网站的链接出现在了其他的网站并且该网站的可信度很高; 假链接是指钓鱼网站为了躲避检测, 在网站中加入自己所假冒网站的链接; 恶链接则是钓鱼网站含有的链接指向另一个钓鱼网站, 一般也是为了躲避钓鱼检测。

3.1 直接 TrustRank 值

对钓鱼网站分析可以得出钓鱼者在推广网站的时候, 通常选取用户数量大的社交网站对钓鱼网站进行推广。钓鱼网站的存活时间短, 但是钓鱼网站在其他正规网站中的推广链接不容易被删除, 因此随机选取 phishtank 上的 2 000 条钓鱼网站链接并使用网络爬虫爬取含有这些钓鱼网站链接的网站, 分别得出钓鱼网站推广链接的分布情况, 如图 2 所示。为了显示方便, 将网站进行分成几大类: 社交类网站、门户类网站、交易类网站、游戏类网站、小说类网站、企业政府类网站和其他类别。

图 2 中的百分比是此类网站中推广链接的数量占总的推广链接的比例。按照数量从多到少排序依次是: 社交类网站、交易类网站、游戏类网站、小说类网站、门户类网站、企业政府类网站和其他。可以看出, 社交类网站占有很大的比例,

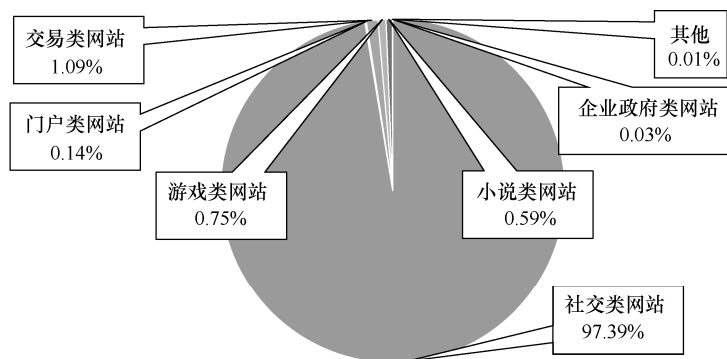


图 2 推广链接在各类网站的分布情况

高达 97.39%，也就是说钓鱼网站推广的目标网站大都属于社交类网站。进一步对社交类网站分析，每个网站所包含的推广链接数相近。然后再多次随机选取其他的 500 条钓鱼网站链接，最终发现有 156 个网站包含的钓鱼网站链接数量大并且接近。由此可以得到一个推广网站集合 p^- ，这个集合包含这 156 个网站。

TrustRank 算法假设好的网站不会链接到坏的网站，但是所分析出的 156 个网站都是 TrustRank 值较高的网站，根据此类情况，提出一个直接 TrustRank 值量化钓鱼网站和其他网站的推广关系。直接 TrustRank 值是只计算网页 p 和链入网页 p 的网页之间的信任关系，并且将包含在推广网站集合中的网页 q_i 的信任度 $S(q_i)$ 设置为 0，否则信任度 $S(q_i)$ 为 1，采取信任分裂的方式计算直接 TrustRank 值，如式 (2) 所示：

$$TR_D(p) = (1-d) + d \sum_{i=1}^{N_{in_p}} \frac{S(q_i)}{No_{q_i}} \quad (2)$$

其中， $TR_D(p)$ 是网页 p 的直接 TrustRank 值， q_i 是含有指向网页 p 链接的第 i 个网页。

直接 TrustRank 值有一个缺点，就是容易将与推广网站集合中的网页有密切关系的正规网站的直接信任值降低。因此提出链入网页 p 的所有网页中不在推广网站集合中的网页占有的比例来改善直接 TrustRank 的缺点。因为，正规网站相比较钓鱼网站而言，与正规网站产生相互间链入链出的网页应该是多样的。因此，当不在推广网站集合中的网页占有的比例越高时，说明此网页不是钓鱼网站的可能性越高。改进后的表达式如式 (3) 所示：

$$TR_{Dp}(p) = TR_D(p) \times \frac{S_{in_p}}{N_{in_p}} \quad (3)$$

其中， S_{in_p} 为指向网页 p 但不在推广网站集合中的网页的数量。

3.2 目标链接相似度

钓鱼网站为了躲避检测或者诱导受害者而在

自身网页设置链入正规网站的链接，而 TrustRank 算法并不能对一个网页的出链接做出检测。因此提出以下步骤。

步骤 1 设立正规网站集合包含易受攻击的一些网站，正规网站集合的网页来自于 Alexa 网站中排名靠前的 500 个网站以及谷歌搜索出的前 1 000 个被经常访问的网站，易受攻击的网站来自 millersmiles 的统计。

步骤 2 计算网页 p 的 URL 和正规网站集合中的 URL 的相似度。

步骤 3 计算得出的相似性越高，则网页 p 是钓鱼网站的可能性越高，因为钓鱼网站为了迷惑受害者会提高自己网站的 URL 与所要假冒的正规网站的 URL 相似度。

钓鱼网站会同时采取增加、修改和删除合法网站的 URL 中的字符来使自己的网站的 URL 看起来与合法网站一样。但钓鱼网站又要尽量与其所假冒的网站的 URL 相似，因此在对字符操作的过程中，字符之间的相对位置变化不大。然后提出用最长公共子序列匹配的方式匹配 URL 并计算匹配度以衡量一个网页 URL 与正规网站集合中的网页的 URL 的相似度，匹配度越高表明是钓鱼网站的可能性越高。

假设网页 p 为要检测的网站，网页 q_i 是正规网站集合的网页，则网页 p 的链接相似度 D_{umatch} 由函数 $U_{match}(U_p, U_{q_i})$ 决定。

$U_{match}(U_p, U_{q_i})$ 由以下步骤计算得出。

Function $U_{match}(U_p, U_{q_i})$

Input: U_p, U_{q_i}

//输入: U_p 为待检测的网页 p 的 URL, U_{q_i} 正规网站集合中的网页的 URL。

Output: D_{umatch}

//输出为 U_p 和 U_{q_i} 的匹配度

Begin:

Count=0;

$L_{q_i}=U_{q_i}.length$; // U_{q_i} 的长度



```

Num= Sum( $U_{q_i}$ ) //统计  $U_{q_i}$  的数量
String Maxstr;
For  $i:=1$  to Num step 1
{
    Maxstr=Maxtch( $U_p$ ,  $U_{q_i}$ , 2) //查找  $U_p$ 
    与  $U_{q_i}$  公共子序列 Master, 最短的子序列长度为 2,
    并存储在 Maxstr 中;
     $L_{\max}$  = Maxtch.length; //求 Maxtch 存储
    的数据数量
    For  $j:=0$  to Maxtch.length-1 step 1
        {
            String str = Maxtch[j];
            Count+= str.length;
             $D[i-1]=\text{count}/(L_{q_i} \times L_{\max})$ ;
        }
     $D_{\text{umatch}}=\text{Smax}(D[ ])$  //取数组  $D$  中最大数
}
Print  $D_{\text{umatch}}$ 
End

```

$U_{\text{match}}(U_p, U_{q_i})$ 就是统计出 U_p 和 U_{q_i} 的所有公共子序列, 然后计算 U_{q_i} 的每一个子序列长度占 U_{q_i} 长度的比例之和的平均值, 最终匹配完所有的 U_{q_i} 后取最大的值。因为钓鱼网站的假链接是和自己的 URL 相似度高的链接, 而正常网站除了自反链接外, 很难和其他网站的 URL 具有高的相似性。可以看出 D_{umatch} 值越大, 网页是钓鱼网站的可能性越高。

3.3 改进的 TrustRank

对于网站链出链接包含恶链接的, 则先判断该网站是不是钓鱼网站, 若不是则将网站添加到推广网站集合中。根据第 3.1 节和第 3.2 节中的分析, 结合钓鱼网站存活时间短的特征, 得出最终改进的 TrustRank 值如式 (4) 所示:

$$\text{TR}_{\text{ph}}(p) = \text{TR}(p) \times \frac{1}{D_{\text{umatch}} + \frac{1}{T_p}} \times \text{TR}_{\text{DP}}(p) \quad (4)$$

其中, T_p 是网站存在的时间, 单位为天。由分析得知 D_{umatch} 和 $\text{TR}_{\text{ph}}(p)$ 成反比, 但是若网页 p 和网页 q_i 没有大于 2 的公共子序列则 D_{umatch} 为 0, 因此不能直接将 D_{umatch} 取倒数。而 T_p 和 D_{umatch} 都属于网页的自身特征 URL 特征和时间特征, 因此可以由以上分析得出 D_{umatch} 和 $\frac{1}{T_p}$ 相加后取倒数可以

解决 D_{umatch} 为 0 的问题。而 $\frac{1}{D_{\text{umatch}} + \frac{1}{T_p}}$ 和 $\text{TR}_{\text{DP}}(p)$

是作为 $\text{TR}(p)$ 的权重因子存在的, 因此三者直接相乘。由此得出结合钓鱼网站改进后的网页的 TrustRank 值 $\text{TR}_{\text{ph}}(p)$ 。

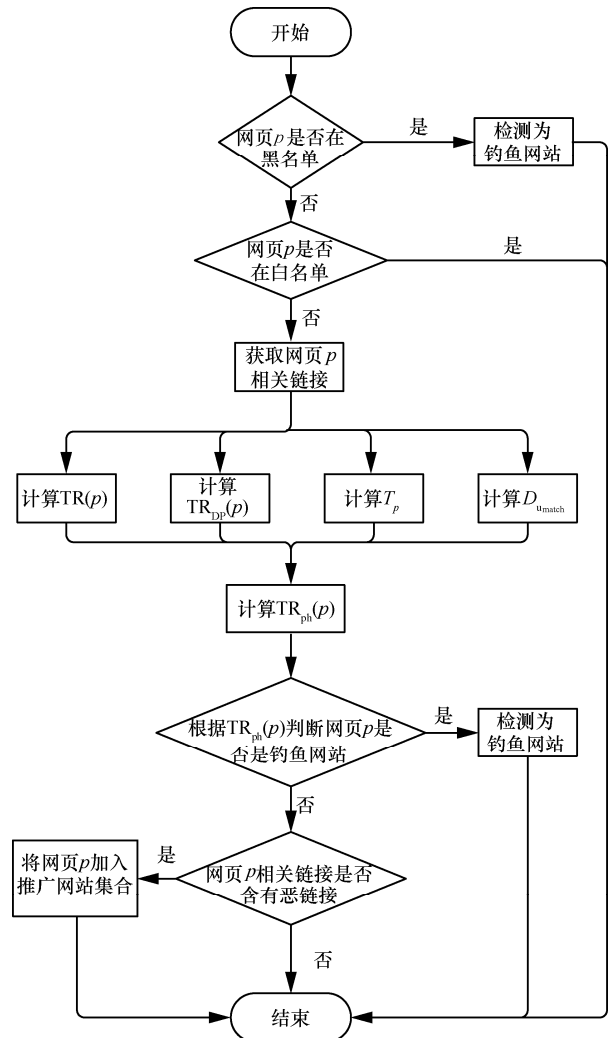


图3 改进的 TrustRank 钓鱼检测网站的流程

图3是基于钓鱼网站链接关系改进的TrustRank检测钓鱼网站的流程,其中黑名单中的网站是phishtank上的网站的URL集合,白名单为第3.2节中的正规网站集合中的网站的URL集合。

4 实验与分析

4.1 实验设置

为了检验基于钓鱼网站链接关系改进的TrustRank检测钓鱼网站是否有效,实验首先观察改进的TrustRank算法在含有钓鱼网站和正常网站的数据集上的准确率和速度,然后分析是否存在一个 TR_{ph} 范围可以将算法的准确率和速度确定在一个稳定和最优的范围内。在此基础上,与参考文献[14]提出的基于图形挖掘与信念传播的检测方法、参考文献[15]提出的通过线性分类器分析提取的页面异常特征检测方法、参考文献[16]提出的基于统计机器学习分类器来检测钓鱼网站的方法以及参考文献[17]提出的基于易受攻击网站的URL的相关性检测钓鱼网站的方法这4种方法,在另外一个随机选取的数据集上比较准确率和速度。

对于TrustRank所需要的好的网站的种子集 Sp^g ,采取参考文献[18]的方法确定。推广网站集合的初始集合由第3.1节中的156个网站组成并随着对钓鱼网站的检测逐渐增加。

4.2 参数确定

为了确定合适的 TR_{ph} 可以高效地检测钓鱼网站,计算包含钓鱼网站和正常网站的数据集中网页的 TR_{ph} 值并对其进行分析。在phishtank网站中随机选取500个钓鱼网站计算 TR_{ph} ,然后随机计算白名单中的网站和排名较高的网站总共2500个网站的 TR_{ph} 值。对这些 TR_{ph} 值进行分析,由于 TR_{ph} 值的范围分布较大,因此对 TR_{ph} 值进行归一化处理,因为检测钓鱼网站时只有当前网页的 TR_{ph} 可供参考,因此采取以自身值为基准归一化,如式(5)所示:

$$Norma(TR_{ph}) = TR_{ph} / (TR_{ph} + 1) \quad (5)$$

然后分析由改进得TrustRank得出的归一化后的 TR_{ph} 发现分布情况如图4所示。

从图4中数据分析可得,钓鱼网站的 $Norma(TR_{ph})$ 值集中分布在0~0.2中,占总的钓鱼网站数量的96.4%,0.2~0.3中有10个钓鱼网站,但是正常网站的数量占实验总网站数量的13.76%,数值过高因此舍掉0.2~0.3这个范围,选取0~0.2为钓鱼网站的 $Norma(TR_{ph})$ 值的范围,对应的 TR_{ph} 在0~0.25中,因此若网页的 TR_{ph} 在这个范围则判定网页为钓鱼网页。这个区间内,改进的TrustRank算法检测误判率ET按照式(6)计算为0.53%,漏判率LT按照式(7)计算为0.6%。

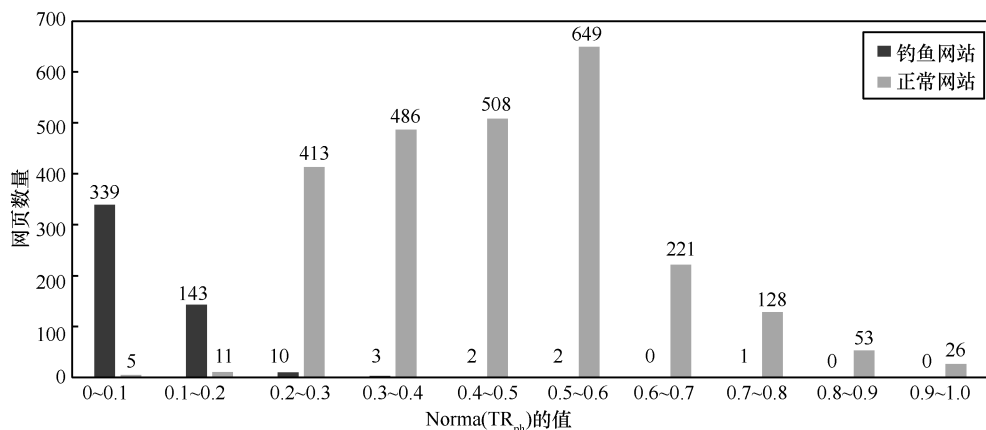


图4 由改进的TrustRank得出的 $Norma(TR_{ph})$ 分布情况



$$ET = \frac{N_E}{SN_p} \times 100\% \quad (6)$$

$$LT = \frac{N_L}{SN_p} \times 100\% \quad (7)$$

$$AC = \frac{N_T}{SN_{ph}} \times 100\% \quad (8)$$

$$D_{rate} = \frac{N_{ht}}{N_{httr}} \times 100\% \quad (9)$$

其中, N_E 是将正常网站判定为钓鱼网站的数目, SN_p 是所有网站的总数目, N_L 是将钓鱼网站判定为正常网站的总数目。

为了验证对 TrustRank 改进的有效性, 对以上网站集合的网页按照式(1)计算 TrustRank 值 TR, 然后计算 Norma(TR) 的值。由 TrustRank 得出的 Norma(TR) 的分布结果如图 5 所示。从图 5 中可以看出, TrustRank 并不能将钓鱼网站和正常网站很好地分开, 在每个区间内都分布有很多的正常网站, 在 0.1~0.2 这个区间, 正常网站的数目最少, 但是占总的网站数目的 3.63%, 远高于改进的 TrustRank 值选取 0~0.2 区间后误判的 0.53%。因此, 改进后的 TrustRank 比 TrustRank 算法能更有效地检测出钓鱼网站和正常网站。

4.3 实验对比

为了验证选取的 TR_{ph} 值的范围可以有效检测钓鱼网站, 将改进后的 TrustRank 算法与参考文献[14]、参考文献[15]、参考文献[16]以及参考文献[17]中的方法作比较, 并计算各种方法的准确率、误判率、漏判率以及速度并进行比较。其中准确率由式(8)计算, 检测速度由式(9)计算。

其中, AC 是检测钓鱼网站的准确率, N_T 是正确检测出的钓鱼网站的数目, SN_{ph} 是钓鱼网站的总数。 D_{rate} 是各方法检测钓鱼网站的有效速率, N_{ht} 是除了改进的 TrustRank 之外的其他方法每小时正确检测出的钓鱼网站的数目, N_{httr} 是改进的 TrustRank 每小时检测出的钓鱼网站的数目。

作对比所采取的钓鱼网站集合是从 anti-phishing 中的黑名单中随机选取的 150 000 个钓鱼网站, 正常网站则从谷歌搜索和 Alexa 网站中随机选取的 20 000 个正常网站。对于钓鱼网站集合的处理是随机分为 3 组, 每组包含 50 000 个钓鱼网站, 然后分别检测每组钓鱼网站集合和正常网站混合后的检测效果。最终对每组的准确率、误判率、漏判率、 D_{rate} 分别求和然后求取平均值。对于参考文献[16]实验中选取的特征为: URL 所在服务器是否是网络虚拟服务器; URL 中是否包含被攻击的特征; URL 中的长单词、关键词以及品牌词。其中, 关键词是包含登录、确认、注册等意思的单词, 品牌词是与现有的公司名有关的单词。训练的样本选取 phishtank 和 anti-phishing 的钓鱼网站。表 1 是 5 种方法在此方法中检测钓

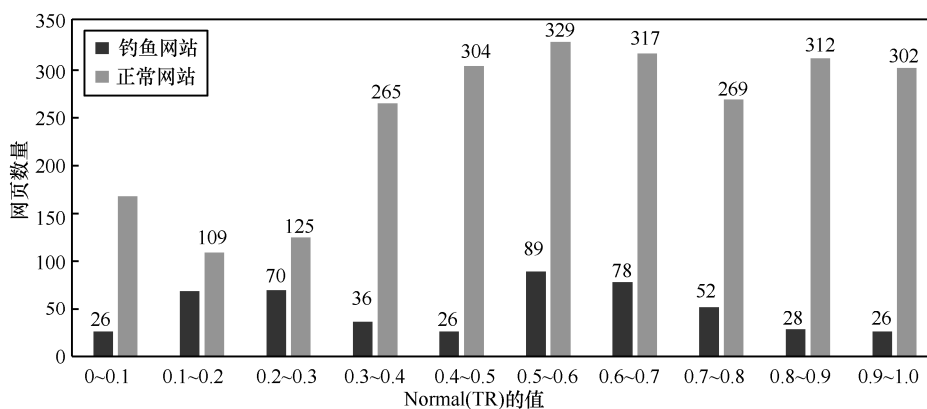


图 5 由 TrustRank 得出的 Norma(TR) 的分布情况

鱼网站的准确率、误判率、漏判率的平均值对比,表2是5组实验中各方法的 D_{rate} 值的平均值。

表1 各方法准确率(AC)、误判率(ET)以及漏判率(LT)的值

检测方法	准确率 (AC)	误判率 (ET)	漏判率 (LT)
改进的 TrustRank	98.42%	0.347%	1.124%
图形挖掘与信念传播	97.53%	0.992%	1.764%
页面异常特征检测	97.86%	2.023%	1.526%
统计机器学习分类器	98.03%	1.017%	1.407%
易受攻击网站的 URL 的相关性	97.68%	1.331%	1.653%

表2 各方法的有效速率 D_{rate} 值

检测方法	有效速率(D_{rate})
图形挖掘与信念传播	0.902
页面异常特征检测	0.887
统计机器学习分类器	0.851
易受攻击网站的 URL 的相关性	0.809

从表1的数据可以分析出,改进的 TrustRank 算法的准确率为 98.42%, 高于其他 4 种方法的准确率 0.89%、0.56%、0.39%和 0.74%。误判率低于其他 4 种方法 0.645%、1.676%、0.67%和 0.984%。同样漏判率也低于其他 4 种方法。因此可以得出改进的 TrustRank 算法能够很好地检测钓鱼网站。

由式(9)可知,当 D_{rate} 的值大于1的时候,可以认为所对应的检测方法的检测速度比改进的 TrustRank 快。但是表2中的数据分别为 0.902、0.887、0.851 和 0.809, 都低于 1, 因此这 4 种方法的检测速度都低于改进的 TrustRank 算法,由此可以得出改进的 TrustRank 具有较高的检测速度。

5 结束语

基于对钓鱼网站链接的分析以及钓鱼网站存在时间短的特点,分别计算了网页的直接 TrustRank 值、目标链接相似度以及钓鱼网站存在时间,然后结合 TrustRank 值改进得到网页的 TR_{ph} 来判断一个网页是否是钓鱼网站。因为结合钓鱼网站的特征以及 TrustRank 本身的基于链接分析

网页的特点,所以具有较高检测准确率和检测速度。但改进的 TrustRank 算法仍有不足,接下来的工作从以下两个方面展开:由于正规网站集合中网页的数量很大,可以将改进的 TrustRank 搭建在云平台中实现以提高检测的速度;结合机器学习的方法可以动态地调整选取的 TR_{ph} 以使检测准确率更高。

参考文献:

- [1] CNNIC. 全球中文钓鱼网站现状统计分析报告(2016年)[R/OL]. (2016-01-07)[2017-10-29]. <http://www.sohu.com/a/14765781-352856>.
CNNIC. Statistical report on current situation of global chinese fishing website(2016)[R/OL]. (2016-01-07)[2017-10-29]. <http://www.sohu.com/a/14765781-352856>.
- [2] SONOWAL G, KUPPUSAMY K S. PhiDMA – A phishing detection model with multi-filter approach[J]. Journal of King Saud University - Computer and Information Sciences, 2017.
- [3] MOGHIMI M, VARJANI A Y. New rule-based phishing detection method[J]. Expert Systems with Applications, 2016(53): 231-242.
- [4] TAN C L, KANG L C, WONG K S, et al. PhishWHO: phishing webpage detection via identity keywords extraction and target domain name finder[J]. Decision Support Systems, 2016(88): 18-27.
- [5] ZHANG D, YAN Z, JIANG H, et al. A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites[J]. Information & Management, 2014, 51(7): 845-853.
- [6] KANG L C, CHANG E H, SZE S N, et al. Utilisation of website logo for phishing detection[J]. Computers & Security, 2015(54): 16-26.
- [7] JAIN A K, GUPTA B B. Phishing detection: analysis of visual similarity based approaches[J]. 2017(4):1-20.
- [8] VARSHNEY G, MISRA M, ATREY P K. A phish detector using lightweight search features[J]. Computers & Security, 2016(62): 213-228.
- [9] RAJAB K D. New hybrid features selection method: a case study on websites phishing[J]. Security & Communication Networks, 2017(2):1-10.
- [10] HADI W, ABURUB F, ALHAWARI S. A new fast associative classification algorithm for detecting phishing websites[M]. Netherlands: Elsevier Science Publishers, 2016.
- [11] RATHORE S, SHARMA P K, LOIA V, et al. Social network security: issues, challenges, threats, and solutions[J]. Infor-



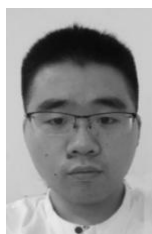
- mation Sciences, 2017.
- [12] SOUSA J, RIBEIRO L, MARQUES A S, et al. Locating leaks with trustrank algorithm support[J]. Water, 2015, 7(4): 1378-1401.
- [13] 刘阳, 张化祥. 基于结合内容特征的 TrustRank 算法改进[J]. 计算机工程与设计, 2013, 34(4): 1276-1279.
- LIU Y, ZHANG H X. Improvement of TrustRank algorithm based on combination of content features[J]. Computer Engineering and Design, 2013, 34(4): 1276-1279.
- [14] ZOU F, GANG Y, PEI B, et al. Web Phishing detection based on graph mining[C]// IEEE International Conference on Computer and Communications, Jul 1-2, 2017, Japur, India. Piscataway: IEEE Press, 2017.
- [15] 宋明秋, 曹晓芸. 基于敏感特征的网络钓鱼网站检测方法[J]. 大连理工大学学报, 2013, 53(6): 903-907.
- SONG M Q, CAO X Y. Detection method of phishing site based on sensitive feature[J]. Journal of Dalian University of Technology, 2013, 53(6): 903-907.
- [16] ZHANG J, PAN Y, WANG Z, et al. URL based gateway side phishing detection method[C]// Trustcom/bigdatasec/ispa Aug 1-4, 2017, Sydney, Australia. Piscataway: IEEE Press, 2017: 268-275.
- [17] XUE Y, LI Y, YAO Y, et al. Phishing sites detection based on Url Correlation[C]// International Conference on Cloud Computing and Intelligence Systems, 2016, Beijing, China. New Piscataway: IEEE Press, 2016: 244-248.
- [18] GYONGYI Z, GARCIA-M H. Seed selection in TrustRank.

Technical report [R]. Stanford University, 2004.

[作者简介]



韩浩 (1993-), 男, 中国矿业大学计算机科学与技术学院硕士生, 主要研究方向为信息安全。



刘博文 (1993-), 男, 中国矿业大学计算机科学与技术学院硕士生, 主要研究方向为云计算和信息安全。



林果园 (1975-), 男, 博士, 中国矿业大学计算机科学与技术学院副教授、硕士生导师、信息安全系主任、信息安全竞赛指导委员会副主任, 主要研究方向为网络空间安全、移动互联及其安全、云计算及其安全、信息系统及其安全、基于位置服务、矿井信息化等。