(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103843003 A
(43)申请公布日 2014.06.04

(21)申请号 201280039933.0
(22)申请日 2012.07.09
(30)优先权数据
61/505,630 2011.07.08 US
(85)PCT国际申请进入国家阶段日
2014.02.17
(86)PCT国际申请的申请数据
PCT/US2012/045979 2012.07.09
(87)PCT国际申请的公布数据
WO2013/009713 EN 2013.01.17
(71)申请人 UAB研究基金会
地址 美国亚拉巴马
(72)发明人 B·瓦德曼 W·哈顿克
(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 叶勇
(51)Int.Cl.
G06F 21/51(2013.01)
G06Q 10/10(2012.01)
H04L 29/06(2006.01)

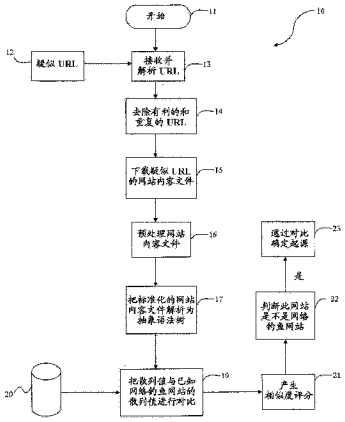
权利要求书3页 说明书20页 附图4页

(54)发明名称

句法指纹识别

(57)摘要

一种通过组成网站的构造组件，识别网络钓鱼网站和展示每个网站的起源的方法。本方法包括识别新观察到的网络钓鱼网站并使用本方法作为聚集网络钓鱼网站的距离度量。变更本方法内的阈值演示了网络钓鱼调查人员识别许多网络钓鱼网站源头以及各个网络钓鱼者的潜在能力。



CN 103843003 A

1. 一种识别网络钓鱼网站的方法，包括：

a. 提供计算机系统，具有操作系统、数据库系统以及用于控制通过因特网通信的通信系统；

b. 向计算机系统发送通信，包含多个疑似网络钓鱼 URL；

c. 检索多个网络钓鱼 URL 的每个疑似网络钓鱼 URL 的网站内容文件，该网站内容文件包括构造组件；

d. 预处理网站内容文件，从而为多个疑似网络钓鱼 URL 的每一个产生标准化的网站内容文件集；

e. 为标准化的网站内容文件集的每一个创建抽象语法树；

f. 为每个标准化的网站内容文件集的每个构造组件计算散列值，并且为每个标准化的网站内容文件集都从中构建散列值集；

g. 从第一个散列值集选择第一个散列值并且把第一个散列值与已知网络钓鱼网站构造组件的散列值进行对比，以定位匹配的散列值；

h. 如果匹配的散列值已定位，把第一个散列值集与匹配的散列值的散列值集进行对比并创建相似度评分；以及

i. 如果相似度评分达到或超过预定阈值，把导出第一个散列值的疑似 URL 指定为网络钓鱼网站。

2. 根据权利要求 1 的方法，其中，所述通信被发送自反垃圾邮件公司、反网络钓鱼公司、关机公司、在客户计算机系统上运行的自治程序，该自治程序被配置为自动地俘获疑似网络钓鱼网站的通信并把疑似网络钓鱼网站的通信发送到所述计算机系统。

3. 根据权利要求 1 的方法，其中，以电子邮件本体发送多个疑似网络钓鱼 URL 时，从采用第一解析程序的通信中提取所述多个疑似网络钓鱼 URL。

4. 根据权利要求 1 的方法，进一步包括在步骤 c 之前，从多个疑似网络钓鱼 URL 中去除以下情形的任何疑似网络钓鱼 URL：已知有利 URL、已知网络钓鱼 URL 或者所述多个疑似网络钓鱼 URL 中另一个网络钓鱼 URL 的副本的 URL。

5. 根据权利要求 1 的方法，进一步包括在所述计算机系统上存储网站内容文件。

6. 根据权利要求 1 的方法，其中，预处理包括一项或多项以下操作：从网站内容文件中去除空格、使网站内容文件不区分大小写或者从网站内容文件中去除动态内容。

7. 根据权利要求 1 的方法，其中，网站内容文件从所检索的网站内容文件的索引页中导出。

8. 根据权利要求 1 的方法，其中，创建抽象语法树包括解析标准化的网站内容文件集内的 HTML 标签并构建 HTML 实体的抽象语法树。

9. 根据权利要求 1 的方法，进一步包括在所述计算机系统上存储散列值。

10. 根据权利要求 1 的方法，进一步包括在所述计算机系统上把已知网络钓鱼网站的构造组件的散列值存储为散列值集表。

11. 根据权利要求 1 的方法，其中，使用 Kulczynski2 系数计算相似度评分。

12. 根据权利要求 1 的方法，进一步包括相似度评分达到或超过预定阈值时，把所述第一个散列值集添加到已知网络钓鱼网站构造组件的散列值。

13. 根据权利要求 1 的方法，其中，构造组件是 HTML 标签。

14. 根据权利要求 1 的方法，进一步包括确定网络钓鱼网站的起源。

15. 根据权利要求 14 的方法，其中，确定网络钓鱼网站的起源包括把网络钓鱼网站的散列值集与已知网络钓鱼网站的散列值集进行对比并对每个已知网络钓鱼网站计算相似度评分。

16. 根据权利要求 15 的方法，进一步包括识别最高的相似度评分并把所述网络钓鱼网站与从其算出最高相似度评分的已知网络钓鱼网站聚集。

17. 一种识别网络钓鱼网站的方法，包括：

a. 接收包含多个疑似网络钓鱼 URL 的通信；

b. 检索多个网络钓鱼 URL 的每个疑似网络钓鱼 URL 的网站内容文件，该网站内容文件包括构造组件；

c. 为每个网站内容文件创建抽象语法树；

d. 为每个网站内容文件的每个构造组件计算散列值，并且为每个网站内容文件集都从中构建散列值集；

e. 从第一个散列值集选择第一个散列值并且把第一个散列值与已知网络钓鱼网站构造组件的散列值进行对比，以定位匹配的散列值；

f. 如果匹配的散列值已定位，把第一个散列值集与匹配的散列值的散列值集进行对比并创建相似度评分；以及

g. 如果相似度评分达到或超过预定阈值，把导出第一个散列值的疑似 URL 指定为网络钓鱼网站。

18. 根据权利要求 17 的方法，进一步包括确定网络钓鱼网站的起源，方式为把网络钓鱼网站的散列值集与已知网络钓鱼网站的散列值集进行对比并对每个已知网络钓鱼网站计算相似度评分。

19. 一种识别网络钓鱼网站的方法，包括：

a. 提供计算机系统，具有操作系统、数据库系统以及用于控制通过因特网通信的通信系统；

b. 向计算机系统发送通信，包含多个疑似网络钓鱼 URL；

c. 在步骤 d 之前，从多个疑似网络钓鱼 URL 中去除以下情形的任何疑似网络钓鱼 URL：已知有利 URL、已知网络钓鱼 URL 或者所述多个疑似网络钓鱼 URL 中另一个网络钓鱼 URL 的副本的 URL；

d. 检索多个网络钓鱼 URL 的每个疑似网络钓鱼 URL 的网站内容文件，其中，网站内容文件包括构造组件并且从所检索的网站内容文件的索引页中导出；

e. 预处理网站内容文件，从而为多个疑似网络钓鱼 URL 的每一个产生标准化的网站内容文件集，其中，预处理包括一项或多项以下操作：从网站内容文件中去除空格、使网站内容文件不区分大小写或者从网站内容文件中去除动态内容；

f. 为标准化的网站内容文件集的每一个创建抽象语法树，其中，创建抽象语法树包括解析标准化的网站内容文件集内的 HTML 标签并构建 HTML 实体的抽象语法树；

g. 为每个标准化的网站内容文件集的每个构造组件计算散列值，并且为每个标准化的网站内容文件集都从中构建散列值集；

h. 从第一个散列值集选择第一个散列值并且把第一个散列值与已知网络钓鱼网站构

造组件的散列值进行对比，以定位匹配的散列值；

i. 如果匹配的散列值已定位，把第一个散列值集与匹配的散列值的散列值集进行对比并创建相似度评分；以及

j. 如果相似度评分达到或超过预定阈值，把导出第一个散列值的疑似 URL 指定为网络钓鱼网站。

20. 根据权利要求 19 的方法，进一步包括确定网络钓鱼网站的起源，方式为把网络钓鱼网站的散列值集与已知网络钓鱼网站的散列值集进行对比并对每个已知网络钓鱼网站计算相似度评分。

句法指纹识别

[0001] 相关申请的交叉引用

[0002] 根据 35U.S.C.§119，本申请要求 2011 年 7 月 8 日提交的临时专利申请序列号 61/505,630 的优先权，标题为“SyntacticalFingerprinting”，其全部内容在此引用作为参考。

技术领域

[0003] 本发明针对自动地识别在工具栏内新观察到的网络钓鱼网站、为了调查而正确地标注所述网络钓鱼网站以及确定所述网络钓鱼网站的流行和起源的方法。

背景技术

[0004] 研发人员已经提出许多不同的技术用于检测文件之间的相似性，比如确定源代码中变化的技术，例如普遍存在的 Unix 实用程序 diff，或者通过 ssdeep 取证地识别系统文件的变种或恶意软件。这些实用程序提供了确定文件变化的好处，比如对代码段的编辑或者诸如插入或删除若干字节的小变化。不过，需要几乎严格匹配并非总是切合实际，并且这些技术对文件的起源是否相同不提供指示。为了识别网站是不是恶意，文件不一定需要完全相同，网络钓鱼就是一个实例情况。常用组件如表单和 JavaScript 函数过程在计算机罪犯的这个子集之中被开发和重用，从而能够被用于识别新的网络钓鱼网站以及聚集类似的网站。

[0005] 网络钓鱼是社会工程攻击，通常通过模仿机构，典型情况下模仿金融机构的网站诱骗受害者提供敏感信息。收集的信息随后用于获得对账户信息的访问权限或用于身份窃取。2008 年，Gartner 的研究报告了 2008 年因网络钓鱼欺诈超过 500 万美国人损失平均 361 美元，损失大约将近 20 亿美元。有两种方法应对这些攻击：反应和主动行动。

[0006] 反应方式是用于许多金融机构的情况，其中恶意内容在被称为“拆卸”的过程中从因特网去除。典型情况下，机构把这个过程外包给“拆卸”公司。这些公司接收可能恶意的 URL 并判断这些 URL 是不是网络钓鱼。如果网站是网络钓鱼网站，那么联系该 URL 所在域的系统管理员随后要求他删除该内容。不过，某些机构已经开始了主动行动方式，通过起诉和定罪以阻止网络钓鱼者从事未来的攻击。

[0007] 机构的反应响应包括在恶意内容到达潜在受害者之前，经由电子邮件过滤器和浏览器工具栏阻止它。电子邮件服务提供商、邮箱软件比如微软的 Outlook 和 Mozilla Thunderbird 或者反垃圾邮件厂商使用了恶意内容所在的已知网站的 URL 列表（黑名单）、URL 内的特征以及统计技术（DSPAM，SpamAssassin 等）以阻挡网络钓鱼邮件到达潜在的受害者。为了适应垃圾邮件过滤器，网络钓鱼者经由 HTML 隐藏电子邮件消息内的内容，假冒发件人的电子邮件和 IP 地址，并且创建随机 URL 把受害者重定向到网络钓鱼网站。这些重定向可以协助使黑名单无效，因为每个 URL 都可以是随机地唯一的。此外，研究人员已经显示出：黑名单识别出足够百分比的 URL 要花两小时而这些 URL 的垃圾邮件活动——这是指为共同意图而发送简短、高容量分发的电子邮件消息——平均持续四至六小时。所以，到把

URL 列入黑名单之时，罪犯很可能已经转移到下一个网络钓鱼网站向新的 URL 发送垃圾邮件。

[0008] 浏览器工具栏是另一种反应措施，往往采用类似技术识别网络钓鱼网站。工具栏使用 URL 黑名单与网站内容的启发式的结合以警告用户网络钓鱼内容(Mozilla Firefox2011;Internet Explorer 2011;Netcraft2011)。这些基于内容的技术能够使用网站的文本分析、WHOIS 信息和图像分析用于识别。这是这些反应方式的一个主要弱点，并且是为何某些机构也已经开始采用主动行动方式的原因。

[0009] 机构的某些响应已经转向更为主动行动的方式，使用调查人员和法律实施以利用起诉和刑期的后果阻止网络钓鱼者。另一方面已经证明，对网络钓鱼的调查难以调查和定罪。调查人员往往缺乏必要的工具和分析数据对罪犯建立强有力的证据。研究人员已经试图收集关于网络钓鱼事件的集合信息，以提供关于这种犯罪行为流行程度的数据报告。2007 年，网络钓鱼者在相同的 IP 区块创建若干域并主管这些域。为了把由同一网络钓鱼者创建的网站分在一组，开发了聚集算法，根据 IP 地址或网络确定网络钓鱼网站的流行。因此根据这些网站所在之处指示网络钓鱼者的范围。不过，最近已经证明，网络钓鱼者正在共享常用的攻击工具，并且有可能使用相同的漏洞利用工具危害网络服务器；因此，如果若干网络钓鱼网站处在同一网络上，未必表明该网站由同一网络钓鱼者创建。

[0010] 在先前工作中，开发的聚集算法采用了称为 Deep MD5Matching 的文件匹配算法，通过文件集中类似文件的数量对网站集进行分组。这种技术展示了把由同一或类似网络钓鱼工具箱创建的若干网站组进行聚集的能力。这种技术的一个缺点是聚集以下网站的能力，它们由网站所在的域上的仅有一个文件组成。

[0011] 在阿拉巴马大学伯明翰分校的研究表明，大约 50% 的网络钓鱼网站包含所在域的服务器上的仅仅一个文件，而提供更网站外观和感觉的其它文件存在于另一台服务器上，比如目标机构的或商标的网络服务器。作为响应，需要开发新的方式用于这样的网站。

发明内容

[0012] 本发明针对识别在工具栏内新观察到的网络钓鱼网站、为了调查而正确地标注所述网络钓鱼网站以及确定所述网络钓鱼网站的流行和起源的方法。句法指纹识别计算网络钓鱼网站主索引文件的构造成分即组件的集之间的相似系数以确定相似度。所述方法能够用于识别、标注以及分组相似的网站，它们可以提供网络钓鱼的作者身份或起始地址的证据。

[0013] 确切地说，所述句法指纹识别方法被用于找到文件关系并确定文件相似度。如此做的方式为把文件和大的字符串集解析为片段并将这些片段与其他文件或文档进行对比而确定它们的相似度。句法指纹识别对网络钓鱼网站的识别能力部分地依赖于软件开发人员在其程序或网站的开发中重用结构和功能组件比如函数和类的实践。同样，人们重用论坛上的帖子和建议。

[0014] 除了确定网络钓鱼网站之间的关系，句法指纹识别还能够被应用到若干恶意软件样本以确定恶意软件家族和恶意软件版本。重叠的代码段或函数可能表明病毒作者重用了来自另一个源的代码，或者所述文件集全部来自同一文件家族(即从同一源创建)并随着时间推移或所述代码被分配给不同开发人员时被修改。论坛中成员往往重帖对用户的建议，

或传递来自其他论坛的新闻。在黑客或恐怖分子的情况下，论坛和论坛主题能够被指纹识别以确定帖子的起源或起点。另外，黑客创建新工具或漏洞利用工具箱闯入电脑。这些漏洞利用工具箱往往重用来自先前工具箱的漏洞利用工具。或许有可能显示出漏洞利用工具箱家族和这些工具箱随时间的演变。句法指纹识别也可以适用于分析因特网流量，无论通过网络日志还是即时数据包俘获。所述协议允许把流量解析为组件，并且这些组件可以对比以确定流量之间的相似度。加权的或白名单的方式能够被用于去除对流量相似度没有影响的常见组件。

[0015] 句法指纹识别的重要方面是其显示出有可能通向文件的起源或家族的文件之间关系的能力，尤其是当文件格式遵循特定的语法树或协议时。此外，句法指纹识别能够被用作聚集算法的距离度量，以展示文件或协议家族如何随时间演变。

[0016] 附图简要说明

[0017] 图 1 是流程图，展示了网络钓鱼网站的抽象语法树指纹识别的方法；

[0018] 图 2 展示了针对两种不同商标的两个网络钓鱼网站如何具有重叠的 HTML 构造，比如 JavaScript 函数；

[0019] 图 3 展示了两个源代码片段之间的代码变化；

[0020] 图 4 展示了关于两个训练集的句法指纹识别的 ROC 图；

[0021] 图 5 展示了使用句法指纹识别的集群。

具体实施方式

[0022] 本发明针对被称为抽象语法树指纹识别或句法指纹识别的新颖方法，用于对比相似的网络钓鱼网站文件结构组件或构造成分以确定相似度。预期这种技术可以应用于计算不同于网络钓鱼网站文件的若干文件类型之间的相似度。这种相似度能够被用于显示网络钓鱼网站文件为同一起源并可能来自同一文件家族。一般来说，本方法包括把网页比如网站索引页解析为抽象语法树。源代码构造成分可以包括网页的常用元素比如表单、表或 JavaScript 代码，但不限于仅仅这些组件。不是语法树的每个构造成分都被解析，因为某些网页可能包含数千个构造成分，可能导致比较和分析中的问题。下一步，对每个构造成分都计算散列值，并且构造成分散列值的集与其他钓鱼网页的构造成分集进行对比。最后步骤使用相似系数（如 Kulczynski2）产生相似度评分。取决于相似度评分的预定阈值，该网站被认为是与特定商标比如美国银行相关联的网络钓鱼网站。另外，根据相似度评分能够确定该网站的起源。

[0023] 参考图 1，所构建的系统 10 为了运行在计算机系统比如计算机服务器上，具有现代化的操作系统，象微软的视窗或 UNIX 的变种比如 Linux。数据库功能目前由 PostgreSQL 提供，它是强大的开源的对象 - 关系数据库系统，但是也可以用于其他数据库平台。目前，在系统中使用 PERL 控制经由因特网的通信并解析所收到的电子邮件。虽然本发明人目前使用解释型语言 PERL，但是预期编译语言比如 C 语言会最终实施所述系统的特征。

[0024] 启动 11 后，系统 10 接收 13 所供给的 URL12 的字符串并把它们解析 13 为文本文件，每行具有分开的 URL。URL12 由各种各样的来源提供，比如反垃圾邮件公司、反网络钓鱼公司、“关机”公司、受益人（如客户）、客户转发的电子邮件、积极预防网络钓鱼网站泛滥的其他实体的通知或者保存着由反垃圾邮件协会所维护的 URL 集合的自动化数据库发出的

通信。此外，消费者可能有在其 PC 上运行的自治程序，自动从疑似网络钓鱼网站俘获通信并把这些通信发送到系统 10 进行自动处理，或者消费者有可能手动地激活已安装的插件，它被设计为与消费者的电子邮件程序合作，转发疑似网络钓鱼通信的取证原始副本。另外，预解析程序（未显示）可以接收向本系统转发的电子邮件，并且提取电子邮件中出现的 URL 再把这些 URL 馈送到系统。典型情况下，编程语言 PERL 在其函数库中包括解析函数，能够用于成功地解析电子邮件，产生电子邮件本体中出现的 URL。

[0025] 决策步骤 14 提供了对重复 URL 和可能已被消费者报告为潜在网络钓鱼网站，但是被系统 10 的受益人先前确定为合法网站的 URL 的排除。例如，如果某特定域被预定义为保持受益人站点，那么利用该域名报告的全部 URL 将会被排除在系统的分析之外。去除有利的和重复的 URL 后，每个剩余疑似网络钓鱼网站 URL 的索引页的网站内容文件 14 都由系统 10 在因特网上检索并下载。然后系统 10 预处理 16 每个网站内容文件，包括去除网站内容文件中的全部空格，并使文件不区分大小写。预处理进一步包括去除在解压网站上钓鱼工具箱期间被添加到文件中的动态内容或定位。预处理产生标准化的网站内容文件。

[0026] 利用诸如 Beautiful Soup 的程序，识别了解析破损 HTML 的 Python 包、标准化网站内容文件内的 HTML 标签，比如 <form>、<script> 和 <table> 标签，并且为每个网站创建了抽象语法树 17。其他编程语言也可以用于解析网站文件。图 2 展示了示范内容文件。抽象语法树由标识的 HTML 实体构建，它们在树中排列的顺序与在导出它们的网站内容文件中呈现的顺序相同。

[0027] 把标准化网站内容文件解析为抽象语法树后，对每个标识的 HTML 实体计算 18 散列值。散列值集从每个网站内容文件的每个 HTML 实体的散列值构建并存储在数据库中。散列值通过计算 MD5 检查和而获得，利用了称为“md5deep”的已知库函数。Md5deep 是使用 MD5（消息摘要算法 5）的散列函数，产生唯一表示下载索引页的单一整数值。众所周知，散列函数是任何明确定义的过程序或数学函数，将大量的有可能可变规模量的数据转换为小数据项，通常是单一整数，可以用作阵列的索引。在这种情况下，MD5 散列函数被用来计算散列值，与其他存储的散列值进行对比。

[0028] 一旦已存储，从网站内容文件的散列值集中随机选择的散列值便与已知网络钓鱼网站的 HTML 实体的散列值进行对比 19。散列值呈现在按时间顺序排列的散列值表中并被存储在数据库 20 上。散列值从最新到最旧排列。在对比期间，随机选取的散列值与已知网络钓鱼网站的散列值按它们在表中呈现的顺序进行对比。这样，随机选择的散列值首先与近期添加的已知网络钓鱼散列值对比，然后再与较旧的散列值对比。如果对第一个随机选择的散列值在数据库 20 中没有找到匹配，便执行来自疑似网站内容文件的另一个散列值。如果在数据库 20 中没有找到匹配，反映所处理的 URL 没有匹配，此 URL 就能够被升级为由干预团队进行人工审核。

[0029] 如果在数据库 20 中找到了匹配，疑似网络钓鱼的 URL 的散列值集便与它已匹配的已知网络钓鱼的 URL 的散列值集进行对比，以产生相似度评分 21。Kulczynski2 系数产生相似度评分。Kulczynski2 系数在公式 1 中表达，其中 a 是集合 1 与集合 2 之间匹配的文件构造成分 MD5 或散列值的数量，b 是集合 1 中的构造成分没有与集合 2 中文件构造成分 MD5 匹配的数量，而 c 是集合 2 中的构造成分没有与集合 1 中文件构造成分 MD5 匹配的数量。

[0030] $$Kulczynski\ 2 = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) \quad (1)$$

[0031] 对公式 1 进行估算所提供的值度量了两个文件构造成分集或散列值集之间的相似度，方式为取两该集合之间匹配构造成分比例的平均。选择 Kulczynski2 相似系数是因为一个集合中匹配构造成分的百分比与另一个集合中匹配构造成分的百分比应当具有相等的权值，以便不歧视任一网页的集合。取决于相似度评分的预定阈值，如果疑似网络钓鱼网站达到或超过此阈值，其 URL 就被视为网络钓鱼网站。一旦网站被视为网络钓鱼网站，其散列值集便被存储到数据库 20 上的已知网络钓鱼网站的散列值表中。正如以下进一步详细介绍，当网站被视为网络钓鱼网站时，确定 23 其起源的方式为对比网络钓鱼网站的散列值集与在数据库 20 上存储的一定的已知网络钓鱼散列值集并计算相似度评分。

[0032] 句法指纹识别方法在以下实例中进一步详细介绍。

[0033] 实例

[0034] 构造成分的预处理。许多网络钓鱼的主索引文件包括在网络服务器上的网络钓鱼工具箱解压期间被添加到文件的动态内容(即对绝对文件路径的引用)。这种内容对过分简单化的和模糊散列值的函数都引起不匹配。从原始副本分开网络钓鱼网站的另一种尝试包括对字母大小写的编辑和插入或删除空格。参考图 3，描绘了两个美国银行网络钓鱼网站源程序的实例，其中差异在“onKeyPress”中存在。为了反击这些实例，对构造成分先进行预处理，方式为去除 URL 和空格并把构造成分改变为不区分大小写，再计算散列值。这些预处理步骤对其他形式的文件匹配算法也有效。

[0035] 数据集。这个实例利用了由 UAB 电脑取证研究实验室(CFRL)收集和标注的两个数据集。这些数据集合由来自许多不同疑似网络钓鱼的 URL 馈源的 URL 组成，使得每个数据集都为网络钓鱼网站的多变高质的集合。若干 URL 被发送到 UAB 网络钓鱼的数据宝库，在此重复的 URL 被去除以避免重复处理同一内容。与这些 URL 相关联的网站内容文件由采用 GNU 的 Wget Red Hat 修改版的定制软件下载。

[0036] 训练数据集由两个实验测试：第一个实验更新训练集，犹如 UAB 网络钓鱼的操作团队和自动方式深层 MD5 匹配每天一批地标注 URL，在每天结束时添加另外的已确认的 URL(即模仿从黑名单公司馈送已确认的 URL)。这个实验最接近地类似于热门网络钓鱼系统在其日常操作中遇到的问题。第二个训练数据集实验运行对由不断地对训练数据集更新的理想标签组成的系统进行仿真。这个实验最接近地类似于在学术研究中使用的假设数据集。

[0037] 数据集 1- 检测网络钓鱼网站。为了把网站标注为网络钓鱼或非网络钓鱼而收集数据集 1。为了确保结果的准确性，数据集的 49,840 个 URL 由人工检查以判断 URL 为有利还是网络钓鱼。结果发现 49,840 个 URL 中的 17,992 个是把 156 个不同的机构作为目标的网络钓鱼。由于这个数据集关于 URL 是网络钓鱼还是非网络钓鱼而不是关于商标的人工检查，所以 URL 商标标签可能不是一直准确的。

[0038] 数据集 2- 聚集网络钓鱼网站。在数据集 1 之后收集数据集 2。这个数据集具有把 230 个不同商标作为目标的网站，并且与数据集 1 相比关于商标更多样化，因为向 UAB 网络钓鱼数据宝库加入了 URL 馈源。这个数据集的一个限制是它不是人工检查准确性，并关于商标和网络钓鱼标签包含误标注的网站。

[0039] 方法。句法指纹识别使用网络钓鱼网站主索引文件的结构组件作为识别网络钓鱼的机制。除了网络钓鱼网站的检测，句法指纹识别也用于演示聚集网络钓鱼网站以及潜在地识别网站起源的能力。

[0040] 检测网络钓鱼网站。为了验证所提出的方法是检测网络钓鱼网站的可接受方法，设置了若干实验。数据集 1 被用于度量对训练数据进行句法指纹识别时检测与误报率，模仿使用人工确认的热门网络钓鱼系统和无瑕疵标注的数据集比如在学术研究中所使用的数据集都用这些训练数据。这些实验测试了改变文件组件的散列值集之间相似系数阈值的效果。

[0041] 聚集网络钓鱼网站。研究的第二阶段使用句法指纹识别作为聚集的距离度量。这个研究阶段对来自数据集 2 的 47,534 个网站即网站池进行测试。对网站池测试了三个实验，其中由 Kulczynski2 系数产生的阈值有变化，使用 10%、50% 和 85%。阈值的变化用于演示更低的阈值如何可以根据起源或源头聚集，而更高的阈值可以根据网络钓鱼者聚集。聚集算法的步骤如下：

[0042] 输入：URL 数据集(D)、阈值(tValue)

[0043] 输出：由相似系数聚集的集群集。

[0044]

```
for each URL Ui in D do
    if Ui is not in the set of clusters C then
  Cx = create_new_cluster_with_representative_URL (Ui) ;
  C.add (Cx)
  for each URL Uj not in the set of clusters C do
      score Sy = calculate_similarity (Ui, Uj ) ;
      if Sy >= tValue then
          add_to_cluster (Uj, Cx) ;
  end
end
```

[0045] 阈值的统计分析。为了设置阈值，必须确定误报率。例如，工具栏或拆卸公司或许仅仅能够接受小于 1% 误报率，而股份公司可能接受 5% 的误报率以保护其员工。考察了三个阈值表明，对于反网络钓鱼社区的某些派系，低值仍然提供了可管理的误报率。此外，考虑这样的技术的利用率时，误报种类存在差异。存在着关于网站是网络钓鱼网站还是有利网站的误报以及被标注为关于为同一目标机构的误报。前者会用于度量工具栏和电子邮件过滤器的准确度，而后者可以用于度量聚集算法的准确度。

[0046] 统计技术。在这项研究中使用的统计技术是对来自数据集 2 的有利和网络钓鱼的 URL 都进行系统性采样。在这个时段，有 47,534 个 URL 包含了与同一时段内另一个 URL 匹配的部分，建立了近 9650 万对。因为这么大的数量，所以系统性采样被用于减少需要手动验证以确定句法指纹识别置信度的 URL 对的数量。系统性采样方案把样本整体按 URL 被提交到 UAB 系统的时间排序。本采样技术计算该集合中前 i 个元素中的随机起始点，并在余

下已排序的样本整体中从该起始点每 i 个元素选择一个。i 的计算结果是总样本(N)除以采样规模(SS)的结果。

[0047] 当数据集关于商标和非网络钓鱼无序时选择这种方式。以 ±2% 的采样误差率实现 99% 置信度的统计公式表明,对于 1,000,000 的样本整体规模会需要对 4,143 个实例采样,而当样本整体规模为 100,000,000 时那么会需要对 4,160 个实例采样(只十三个以上)。公式 2 和公式 3 用于计算样本整体为 N 时所需的样本规模。Z 被定义为置信百分比的 Z 评分。P 是样本整体中网络钓鱼的比例。典型情况下,如果这个值未知,那么,使用 0.5,这将会使样本整体中需要采样的部分最大化。最后,C 指的是置信区间,意味着误差率在 ± 某百分比之内。在这个统计分析中置信百分比是 99%,Z 评分是 2.576,P 为 0.5,而 C 被设置为 1% 和 2%。

[0048] $$X = \frac{Z^2 * P * (1-P)}{C^2} \qquad (2)$$

[0049] $$SS = \frac{X}{1 + \frac{X-1}{N}} \qquad (3)$$

[0050] 选择 99% 置信度和 0.5 的 P 值是为了对显示句法指纹识别有效性所需要的网站数量进行过采样。

[0051] 统计分析。初步测试表明误报率和检测率随阈值变化而改变。所以,测试了三个阈值 10%、50% 和 85% 以确定在句法指纹识别方法中使用它们将引发的误报率。样本整体从 9650 万对中采集,有利和网络钓鱼网站都包括,其中算出的文件组件集的 Kulczynski2 系数大于等于这三个阈值。这些查询对于阈值 85% 产生了 10,548,665 对的样本整体,而对于 50% 产生了 19,282,737 对,对于 10% 产生了 88,999,846

[0052] 对。表 1 呈现了使用公式 3 对每个阈值的样本整体算出的样本规模。

[0053]

样本规模	85% 阈值	50% 阈值	10% 阈值
±1% 误差率	16,615	16,627	16,638
±2% 误差率	4,160	4,160	4,160

[0054] 表 1 :句法指纹识别的统计分析中每个阈值的样本规模

[0055] 为了测试所述采样方法的准确度,以及在研究中加入统计的优点,每组样本,1% 和 2% 的样本都随机地选择并且网站由人工检查以确定句法指纹识别方法的准确度。共计 62,360 对进行了准确度检查。这些样本显示出预期的在理想地标注的数据集上关于商标标注和网络钓鱼检测的误报率。表 2 和表 3 由对 ±1% 和 ±2% 采样误差率都使用 99% 的置信度的每个阈值的统计分析结果组成。

[0056]

误报率	85%阈值	50%阈值	10%阈值
±1%误差率的99%置信度	0.0%	0.0%	0.32%
±2%误差率的99%置信度	0.0%	0.0%	0.07%

[0057] 表 2 :使用句法指纹识别把有利网站标注为网络钓鱼网站的统计方法的结果

[0058]

误报率	85%阈值	50%阈值	10%阈值
±1%误差率的99%置信度	0.04%	0.69%	1.02%
±2%误差率的99%置信度	0.07%	0.07%	0.77%

[0059] 表 3 :使用句法指纹识别误标注网络钓鱼网站的统计方法的结果。

[0060] 结果。首先呈现的是当改变文件组件集之间 Kulczynski2 相似系数的阈值时,发生在数据集 1 上的检测和误报率的结果。其次,呈现了使用句法指纹识别作为距离度量的聚集方法的结果。呈现了 I2 分析师簿式视窗图表,以便视觉地展示在网络钓鱼网站整个进化中如何使用文件组件,无论是由相同的还是不同的网络钓鱼者。

[0061] 检测网络钓鱼网站。参考图 4,上述统计方法显示出,不同的阈值能够改变对网络钓鱼内容进行识别时的误报率水平。对数据集 1 测量了每个阈值的检测和误报率。正如在表 4 中所观察到的,用于句法指纹识别的不同阈值对检测和误报率都有影响。当在两个实验运行中都把阈值从 85% 降低至 10% 时,检测率实质提高 6-7%。85% 阈值的误报率分别为 1.9% 和 2.0%,然而,在两个实验中 85% 阈值的误报率增加了 12.5% 和 13.5%,这对于反网络钓鱼解决方案是高误报率。在这个数据集中有 1,981 个网站(11%)的主索引页不包含任何 AST 构造成分。

[0062] 标注似乎随训练数据(即网络钓鱼主页的数量)的规模增加而变得更好。表 4 表明,有众多有利网站包含着在网络钓鱼中出现的某些重叠段。许多这些有利网站几乎不重叠,如它们的评分所表明。

[0063]

技术	总数据集	
	DR	FP
AST（85%）训练集 1	88.1%	1.95%
AST（50%）训练集 1	93.0%	3.8%
AST（10%）训练集 1	95.1%	14.4%
AST（85%）训练集 2	89.5%	2.0%
AST（50%）训练集 2	93.4%	3.6%
AST（10%）训练集 2	95.4%	15.5%

[0064] 表 4 :使用数据集中主索引页的句法指纹识别的结果。

[0065] 聚集网络钓鱼网站。最后结果展示了如何使用组成这些网站的构造成分能够聚集网络钓鱼网站的主页。第一组群集根据数据集 2 中若干网站之间的 10% 重叠而分组。第二和第三组相似但是使用了 50% 和 85% 的阈值。阈值的变化有助于展示更低的阈值如何显示出识别同一起源或原始组件的网站的能力，而更高的阈值展示出找到由个别网络钓鱼者创建的一组网站的能力。

[0066] 第一组群集使用了 10% 的阈值，所以，如果在某网络钓鱼页面中包含部分的 10% 或更多出现在代表性 URL 中，那么向该群集加入候选 URL。一旦候选 URL 被添加到群集，它便从代表性的和候选 URL 池中去除。这个过程导致 4,033 个群集中的 2,182 个群集包含该群集中不止一个 URL。2,182 个群集中有 1,018 个包含至少一个网站具有群集中的商标，而这些群集中的 94 个包含同一群集中的多个商标。

[0067] 通过提高阈值增加聚集算法的选择性引起更小尺寸的更多群集。在 50% 阈值，有 6,791 个群集，包括 2,182 个带有不止一个 URL，而 85% 阈值引起 9,311 个群集，其中 2,948 个包含不止一个 URL。在 50% 水平，1,721 个具有至少一个网站带有已标注的商标且只有 87 个群集具有多个商标，而在 85% 水平，2,796 个群集包含带商标的网站且 106 个群集包含多个商标。

[0068] 如上所述，数据集 2 中的 URL 表示了 230 个不同的网络钓鱼商标。表 5 显示了关于 85% 和 10% 阈值结果中代表性 URL 的十个最知名商标的某些特征。表 5 展示了句法指纹识别中 Kulczynski2 系数的阈值变化如何能够改变群集的规模和数量。表 5 中重要的变化是 PayPal 网络钓鱼网站的减少，当阈值从 85% 移动到 10% 时减少 74.5%。以下聚集讨论部分中介绍了对发生现象的观察，以及对合成群集的更深入分析。

[0069]

目标	85%阈值		10%阈值		百分比差异	
	群集数	网站数	群集数	网站数	群集	网站
PayPal	595	4,322	186	1,104	68.7%	74.5%
美国银行	174	1,636	24	1,619	86.2%	1.0%
eBay	95	965	36	556	62.1%	42.4%
大通银行	72	1,216	18	1,093	75.0%	10.1%
HSBC	69	1,746	23	773	66.7%	55.8%
Visa	56	227	26	276	53.6%	-21.6%
Lloyds TSB	53	477	23	281	56.6%	41.1%
Craiglslist	48	102	38	92	20.8%	9.8%
Facebook	38	50	6	27	84.2%	46.0
巴西银行	35	263	11	341	68.6%	-30.0%

[0070] 表 5 :使用句法指纹识别根据代表性 URL 商标,关于目标机构的群集的最大数量。

[0071] 下文讨论了这些实验的结果以及如何调整这种技术用于不同目的。为了更清晰地表达句法指纹识别以及它如何工作,介绍了实例群集的 i2 分析师簿式视窗图表。

[0072] 图 5 是以 50% 阈值使用句法指纹识别所产生的群集之一的视觉表达。代表性 URL 的商标,该图中心的圆,是 NatWest。称为子集的正方形表示由网络钓鱼网站的图标所指示的每组网络钓鱼当中的公共构造成分集。弧线具有相关联的十进制数,是在每个子集中的网络钓鱼网站与代表性 URL 的相似度评分。有 13 个 JavaScript 和 2 个表单段被用于建造该群集。表 6 显示了所述实体存在于源代码之内的群集内所有网站的出现百分比。

[0073]

JavaScript 实体 1、2、3、4	91%
JavaScript 实体 5、6、7、8	82%
JavaScript 实体 9	64%
JavaScript 实体 10、11	46%
JavaScript 实体 12、13	18%
表单实体 1、2	27%

[0074] 表 6 :展示了包含每个实体的网站百分比

[0075] 正如表 6 中要素展示,与匹配的表单表单实体相比,匹配的 JavaScript 实体更为普遍。表单实体往往被用于提供网站的外观和感觉,而 JavaScript 实体能够影响网站的功

能。网络钓鱼网站的各个版本可能外观和感觉略有不同，但仍然需要相同的功能。因此，匹配的 JavaScript 实体的普遍性可能是由于网站的功能而不是外观和感觉。

[0076] 检测网络钓鱼网站。句法指纹识别实验的检测和误报率可以与之前研究人员的结果不相上下，与 10% 阈值相关联的高误报率例外。两个数据集的利用显示出令人吃惊的结果。使用每日批处理系统的训练数据与使用理想标签进行训练的方法之间没有显著差异。误报和漏报的分析减少了本技术当前实施方式的局限性。

[0077] 数据集内发生漏报由于两个主要原因。未识别网络钓鱼网站的第一个原因是因为在该数据集期间引入了在该数据集之内或该训练集之内先前并未出现的新的源代码。新的源代码不是对先前已观察到的网络钓鱼网站的变更或修改。利用更大的主索引页集，如同 UAB 网络钓鱼数据宝库中当前存在的，检测率能够得到提高。未划分网络钓鱼网站的第二个原因是由于网站的语法解析。在目前实验中，当元素字母大写时，比如搜索 <table> 标签但是不捕捉 <TABLE> 标签时，句法元素不被考虑。这种现象存在于许多被错过的网站。这种分析的另一个发现是，某些构造成分没有被解析并散列在语法树中。例如，目前的实验没有把 标签考虑为用于对比的构造成分。

[0078] 要解决误报率更加复杂。在某些情况下网络钓鱼者重用合法网站中存在的构造成分，给出逼真的外观和感觉。这些可重用的组件实际存在，比如用于登录用户的普通 JavaScript 函数和用来选择数据的表格。在未来的实施中可以给这些公共构造成分更小的权重，所以不会根据这样的构造成分标注网络钓鱼网站。

[0079] 聚集网络钓鱼网站。使用句法指纹识别作为距离度量已经显示出根据组成网站主索引页的公共结构组件聚集网站的能力。分析显示出在不同阈值的句法指纹识别可以根据网络钓鱼与非网络钓鱼、商标、或可能网络钓鱼者引起聚集。

[0080] 表 5 显示了使用三个阈值的句法指纹识别聚集得出的最高群集。显然，升高 Kulczynski2 系数的阈值引起群集数量的增大。在同高阈值群集中的成员可能由同一网络钓鱼者创建。

[0081] 表 5 中，更高阈值产生更多群集的实例在美国银行商标中观察到。使用 10% 阈值产生的 24 个群集与使用 85% 阈值产生的 174 个群集之间有 1% 的网站成员变化。群集数量的增大而商标成员差异的缺乏的解释可能是 174 个群集是由不同网络钓鱼者所编辑的网站组，而 24 个群集是来自同一文件起源的网站组。

[0082] 句法指纹识别可用于自动地标注网站商标。85% 阈值产生了 2,630 个单商标群集，包括 37,129 个网站。不过，85% 阈值也产生了 106 个跨商标群集，包括 18,457 个网站。106 个跨商标群集的分析显示出 88 个群集事实上不是跨商标群集。这 88 个群集的成员网站由人工和由 UAB 网络钓鱼数据宝库当前采用的自动标注而误标注。不仅如此，这些群集有可能被用于重新标注误识别的网络钓鱼内容。正如关于数据集 2 指出，数据集不是 100% 由人工检查标注。通过使用聚集方法，可以在数据宝库内识别并修复误标注的网络钓鱼网站。除了重新标注已知的网络钓鱼内容，句法指纹识别也显示出对过去错过的网络钓鱼网站进行更新的能力。一旦检测出新版本的网络钓鱼网站，现在就能够根据新模式或构造成分更新过去未检测出的网站。

[0083] 剩余 18 个 85% 阈值的跨商标群集中，9 个群集包含的网站使用 JavaScript 函数把用户重定向到附加内容。剩余 9 个跨商标群集每个群集都包含两个商标。所有这些群集

中，两个商标的网站使用同一构造成分组织和执行网络钓鱼内容。许多网站的源代码几乎相同，网页的标题和标志除外。

[0084] 以下索引页内容文件中可以观察到实例。

[0085] A. 桑坦德银行网站

[0086]

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>
<head>
<title>Santander - Mais segurança e praticidade no seu dia-a-dia</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<script language="JavaScript" type="text/JavaScript">
<!--
function MM_reloadPage(init) {  //reloads the window if Nav4 resized
  if (init==true) with (navigator) {if
((appName=="Netscape")&&(parseInt(appVersion)==4)) {
    document.MM_pgW=innerWidth; document.MM_pgH=innerHeight;
onresize=MM_reloadPage; }}
  else if (innerWidth!=document.MM_pgW || innerHeight!=document.MM_pgH)
location.reload();
}
MM_reloadPage(true);
//-->
</script>
</head>

<body leftmargin="0" topmargin="0" marginwidth="0" marginheight="0" scroll=no>
<div id="Layer14" style="position:absolute; left:502px; top:27px; width:306px;
height:16px; z-index:19"><font color="#FF0000" size="1" face="Verdana, Arial,
Helvetica, sans-serif">Ambiente
    Seguro e Critptografado</font></div>
<div id="Layer11" style="position:absolute; left:477px; top:22px; width:22px;
height:18px; z-index:18"><img
src="http://www.colorsfm.com/grupo/imagens/cadiado.png" width="20"
height="23"></div>
<div id="Layer13" style="position:absolute; left:295px; top:185px; width:22px;
height:18px; z-index:14"><img
src="http://www.colorsfm.com/grupo/imagens/padlock.jpg" width="22"
height="30"></div>
<div id="Layer10" style="position:absolute; left:295px; top:255px; width:22px;
height:18px; z-index:15"><img
src="https://www.colorsfm.com/grupo/imagens/padlock.jpg" width="22"
height="30"></div>
<div id="Layer1" style="position:absolute; left:9px; top:17px; width:181px;
height:64px; z-index:1"><img
```

[0087]

```
src="http://www.colorsfm.com/grupo/imagens/hello.png" width="178"
height="50"></div>
<div id="Layer2" style="position:absolute; left:5px; top:277px; width:109px;
height:122px; z-index:2"><img src="http://www.colorsfm.com/grupo/imagens/pc.png"
width="200" height="150"></div>
<div id="Layer3" style="position:absolute; left:237px; top:43px; width:463px;
height:44px; z-index:3; background-image: url(../imagens/barra.png); layer-
background-image: url(../imagens/barra.png); border: 1px none #000000;"></div>
<div id="Layer4" style="position:absolute; left:233px; top:11px; width:65px;
height:19px; z-index:4"><strong><font color="#FF0000" size="2" face="Geneva,
Arial, Helvetica, sans-serif">Passo
  1</font></strong></div>
<div id="Layer5" style="position:absolute; left:258px; top:46px; width:28px;
height:27px; z-index:5"><img
src="http://www.colorsfm.com/grupo/imagens/botaovermelho1.png" width="30"
height="40"></div>
<div id="Layer6" style="position:absolute; left:291px; top:56px; width:26px;
height:24px; z-index:10"><img
src="http://www.colorsfm.com/grupo/imagens/botaocinza2.png" width="25"
height="24"></div>
<div id="Layer7" style="position:absolute; left:322px; top:56px; width:26px;
height:23px; z-index:9"><img
src="http://www.colorsfm.com/grupo/imagens/botaocinza3.png" width="25"
height="24"></div>
<div id="Layer8" style="position:absolute; left:352px; top:56px; width:26px;
height:23px; z-index:8"><img
src="http://www.colorsfm.com/grupo/imagens/botaocinza4.png" width="25"
height="24"></div>
<div id="Layer9" style="position:absolute; left:384px; top:55px; width:26px;
height:26px; z-index:11"><img
src="http://www.colorsfm.com/grupo/imagens/botaocinza5.png" width="25"
height="24"></div>
<div id="Layer10" style="position:absolute; left:7px; top:104px; width:348px;
height:25px; z-index:12"><strong><font color="#CC0000" size="4" face="Verdana,
Arial, Helvetica, sans-serif">Selecione
  o tipo de sua Conta.</font></strong></div>

<div id="Layer13" style="position:absolute; left:276px; top:180px; width:330px;
height:39px; z-index:15"><a
href="http://www.colorsfm.com/grupo/bancos/fisico/2.php"><img
src="http://www.colorsfm.com/grupo/imagens/botao-pessoaprime.png" width="342"
height="35" border="0"></a></div>
```

[0088]

```
<div id="Layer14" style="position:absolute; left:276px; top:250px; width:326px;
height:33px; z-index:16"><a
href="http://www.colorsfm.com/grupo/bancos/prime/2.php"><img
src="http://www.colorsfm.com/grupo/imagens/botao-pessoaprivate.png" width="342"
height="35" border="0"></a></div>
<div id="Layer16" style="position:absolute; left:492px; top:51px; width:191px;
height:34px; z-index:17">
  <object classid="clsid:D27CDB6E-AE6D-11cf-96B8-444553540000"
codebase="http://download.macromedia.com/pub/shockwave/cabs/flash/swflash.cab#ve
rsion=6,0,29,0" width="195" height="25">
    <param name="movie" value="relogio_PF.swf">
    <param name="quality" value="high">
    <param name="wmode" value="transparent">
    <embed src="http://www.colorsfm.com/grupo/perfil/relogio_PF.swf" width="195"
height="25" quality="high"
pluginspage="http://www.macromedia.com/go/getflashplayer" type="application/x-
shockwave-flash" wmode="transparent"></embed>
  </object>
</div>
<script
src="http://www.colorsfm.com/grupo/perfil/dynActiveX_FineGround_vmzl5n2uiyseodqr
poaauzllb_FGN_V01.js" type="text/javascript"></script>
<table width="745" height="473" align="left">
  <tr align="center" valign="top">
    <td width="66%" colspan="2"
background="http://www.colorsfm.com/grupo/imagens/layout.png"><p
align="left"><font size="2" face="Geneva, Arial, Helvetica, sans-
serif"></font></p>
    <p align="left"> </p>
    <div align="left"><strong></strong> </div></td>
  </tr>
</table>
<div align="left"></div>
</body>
<script>

alert( "Segurança e tranquilidade - Você está operando em um ambiente seguro e
criptografado, a partir de agora." );

</script>
</html>
<?
}
```

[0089]

```
?>
```

[0090] B. 巴西银行网站

[0091]

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>
                <head>
<title>Bradesco - Mais segurança e praticidade no seu dia-a-dia</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<script language="JavaScript" type="text/JavaScript">
<!--
function MM_reloadPage(init) {  //reloads the window if Nav4 resized
  if (init==true) with (navigator) {if
((appName=="Netscape")&&(parseInt(appVersion)==4)) {
    document.MM_pgW=innerWidth; document.MM_pgH=innerHeight;
onresize=MM_reloadPage; }}
  else if (innerWidth!=document.MM_pgW || innerHeight!=document.MM_pgH)
location.reload();
}
MM_reloadPage(true);
//-->
</script>
</head>

<body leftmargin="0" topmargin="0" marginwidth="0" marginheight="0" scroll=no>
<div id="Layer14" style="position:absolute; left:502px; top:27px; width:306px;
height:16px; z-index:19"><font color="#FF0000" size="1" face="Verdana, Arial,
Helvetica, sans-serif">Ambiente
    Seguro e Critptografado</font></div>
<div id="Layer11" style="position:absolute; left:477px; top:22px; width:22px;
height:18px; z-index:18"><img
src="http://sprinters.ru/2011atual/imagens/cadiado.png" width="20"
height="23"></div>
<div id="Layer1" style="position:absolute; left:9px; top:17px; width:181px;
height:64px; z-index:1"><img
src="http://sprinters.ru/2011atual/imagens/hello.png" width="178"
height="50"></div>
<div id="Layer2" style="position:absolute; left:5px; top:277px; width:109px;
height:122px; z-index:2"><img src="http://sprinters.ru/2011atual/imagens/pc.png"
width="144" height="194"></div>
<div id="Layer3" style="position:absolute; left:237px; top:43px; width:463px;
height:44px; z-index:3; background-image: url(../imagens/barra.png); layer-
background-image: url(../imagens/barra.png); border: 1px none #000000;"></div>
```

[0092]

```
<div id="Layer4" style="position:absolute; left:233px; top:11px; width:65px;
height:19px; z-index:4"><strong><font color="#FF0000" size="2" face="Geneva,
Arial, Helvetica, sans-serif">Passo
 1</font></strong></div>
<div id="Layer5" style="position:absolute; left:258px; top:46px; width:28px;
height:27px; z-index:5"><img
src="http://sprinters.ru/2011atual/imagens/botaovermelho1.png" width="30"
height="40"></div>
<div id="Layer6" style="position:absolute; left:291px; top:56px; width:26px;
height:24px; z-index:10"><img
src="http://sprinters.ru/2011atual/imagens/botaocinza2.png" width="25"
height="24"></div>
<div id="Layer7" style="position:absolute; left:322px; top:56px; width:26px;
height:23px; z-index:9"><img
src="http://sprinters.ru/2011atual/imagens/botaocinza3.png" width="25"
height="24"></div>
<div id="Layer8" style="position:absolute; left:352px; top:56px; width:26px;
height:23px; z-index:8"><img
src="http://sprinters.ru/2011atual/imagens/botaocinza4.png" width="25"
height="24"></div>
<div id="Layer9" style="position:absolute; left:384px; top:55px; width:26px;
height:26px; z-index:11"><img
src="http://sprinters.ru/2011atual/imagens/botaocinza5.png" width="25"
height="24"></div>
<div id="Layer10" style="position:absolute; left:7px; top:104px; width:348px;
height:25px; z-index:12"><strong><font color="#CC0000" size="4" face="Verdana,
Arial, Helvetica, sans-serif">Selecione
 o tipo de sua Conta.</font></strong></div>
[redacted]
<div id="Layer13" style="position:absolute; left:276px; top:225px; width:330px;
height:39px; z-index:15"><a
href="http://sprinters.ru/2011atual/bancos/prime/2.php"><img
src="http://sprinters.ru/2011atual/imagens/botao-pessoaprime.png" width="342"
height="35" border="0"></a></div>
<div id="Layer14" style="position:absolute; left:276px; top:279px; width:326px;
height:33px; z-index:16"><a
href="http://sprinters.ru/2011atual/bancos/private/2.php"><img
src="http://sprinters.ru/2011atual/imagens/botao-pessoaprivate.png" width="342"
height="35" border="0"></a></div>
```

[0093]

```
<div id="Layer16" style="position:absolute; left:492px; top:51px; width:191px;
height:34px; z-index:17">
  <object classid="clsid:D27CDB6E-AE6D-11cf-96B8-444553540000"
codebase="http://download.macromedia.com/pub/shockwave/cabs/flash/swflash.cab#ve
rsion=6,0,29,0" width="195" height="25">
    <param name="movie" value="relogio_PF.swf">
    <param name="quality" value="high">
    <param name="wmode" value="transparent">
    <embed src="http://sprinters.ru/2011atual/perfil/relogio_PF.swf" width="195"
height="25" quality="high"
pluginspage="http://www.macromedia.com/go/getflashplayer" type="application/x-
shockwave-flash" wmode="transparent"></embed>
  </object>
</div>
<script
src="http://sprinters.ru/2011atual/perfil/dynActiveX_FineGround_vmz15n2uiyseodqr
poaauzllb_FGN_V01.js" type="text/javascript"></script>
<table width="745" height="473" align="left">
  <tr align="center" valign="top">
    <td width="66%" colspan="2"
background="http://sprinters.ru/2011atual/imagens/layout.png"><p
align="left"><font size="2" face="Geneva, Arial, Helvetica, sans-
serif"></font></p>
    <p align="left"> </p>
    <div align="left"><strong></strong> </div></td>
  </tr>
</table>
<div align="left"></div>
</body>
<script>

alert( "Segurança e tranquilidade - Você está operando em um ambiente seguro e
criptografado, a partir de agora." );

</script>
</html>
<?
}
?>
```

[0094] 在这个实例中，粗体字显示的标题包含不同的目标商标，但是标题的其余部分相同。每个网站涉及几乎完全相同内容的文件，不过当对这些文件进行预处理时文件处在不同的服务器上，正如以斜体字显示的类型所示以及上面指出。最后，桑坦德银行网络钓鱼涉

及

[0095] carlin.jpg，这不是巴西银行网络钓鱼所涉及的，因为巴西银行网络钓鱼涉及botoa-pessoafisica.png。carlin.jpg 的几行加了下划线而

[0096] botoa-pessoafisica.png 以加下划线的粗体字显示。这可以表明这种技术可以不仅用于识别某特定网络钓鱼者对一个机构的攻击，而且可以识别网络钓鱼者使用同一或类似内容攻击多个机构。

[0097] 最后，观察到了改变阈值导致的群集中网站数量的显著降低，如表 5 展示，它们的代表性 URL 被标注为 PayPal。85% 与 10% 阈值的对比显示出网站数量减少了 74.5%。更详细地说，10% 阈值时有 7,690 个 PayPal 网络钓鱼在其代表性 URL 被标注为有利的群集中发现，而 85% 阈值时 4,566 个 PayPal 网络钓鱼在类似群集中发现。这表明了两种情况之一。首先，重申了聚集方法可以被用于标注未被标注为网络钓鱼的 URL。另一方面，表明每个群集的代表性 URL 都应该由人工验证和标注以扩大商标标注。

[0098] 使用不同阈值的句法指纹识别能够由不同的网络钓鱼对策使用。URL 黑名单公司可以发现以 10% 或 50% 阈值的误报率可接受，然而，对于拆卸公司它们过高。另一方面，拆卸公司确实发现以 85% 阈值的误报率是可以接受的。典型情况下，这些公司采用人工判断某网站是否为网络钓鱼。考虑到这一点，可以把系统设置为把超出 85% 的全部网站标注为网络钓鱼，而把落入 85% 与 50% 之间的网站标注为更可能的候选者。从而减少检查所有潜在网络钓鱼内容所需要的人工量。

[0099] 局限性。本文介绍的方法具有局限性，如果专注解决，可以引起句法指纹识别更好的性能。第一个主要局限是用于收集网络钓鱼网站的提取过程。使用 Wget 是因为易于实施及其提供的特征。不过，在 HTML、PHP 或 JavaScript 内发生重定向时，关于提取内容 Wget 也有局限性。这样的重定向导致 Wget 无法检索恶意的网络钓鱼内容。数据集 1 中未检测出的许多网络钓鱼网站事实上为重定向并且网络钓鱼内容未被检索。要是网站被检索了，句法指纹识别的检测率很可能提高。一个解决方案将是开发定制网络爬虫，它具有跟随这样的重定向和捕捉网络钓鱼内容的能力。

[0100] 另一个问题是 Wget 提取与人工检查过程之间的时间差异。如果在 Wget 提取与人工检查过程之间的时间里网页已经改变或其可用性已经改变，那么人工检查网页将不会看到与 Wget 过程检索相同的网页。例如，如果 Wget 提取时钓鱼网页被挂起，但是进行人工检查过程时活化，那么人工检查将把 UAB 网络钓鱼数据宝库中的挂起页面标注为确认的网络钓鱼。产生的误标注能够通过系统级联导致许多未来的误标注。最后，对由句法指纹识别创建的群集只给出了粗略分析而不是深入分析。这些群集需要更多分析以理解它们如何组成，以及阈值的变化如何能够用于识别不同的网站层(即同一起源、文件家族、商标或网络钓鱼者的文件)

[0101] 未来的工作。本文介绍的方法提出了新颖的方法，用于网络钓鱼网站检测和分类。初始结果表明本方法有良好的检测性能并展示了链接类似网站的能力。在这两个领域都有改进空间，所以，需要未来的工作使这种技术更好。加入规模和 / 或构造成分类型作为相似系数中的加权调整，并使用抽象语法树的更多元素作为构造成分可以显示出在检测、商标标注以及作者身份即起源中的改进。

[0102] 检测阶段中，为了关于提高检测率同时降低误报率的更好性能，这种技术能够与

文件匹配的其他方法结合。与对照的文件匹配的其他技术相比，句法指纹识别可以证明是寻找候选文件的良好技术。这项研究还显示出句法指纹识别用作简单聚集的距离度量的能力；不过，未来的研究可以实施各种各样的聚集算法以提高性能。

[0103] 未来工作可以包括调查研究高阈值的群集，以及显示多种文件构造成分显现在由 UAB 网络钓鱼数据宝库所收集的数据中的时间。句法指纹识别或许能够显示文件构造成分的显现与目标机构网站的变化之间的相关性。最后，需要先进一步测试和分析，才能做出有关网络钓鱼主索引文件的起源的更多断言。

[0104] 正如将会被本领域的技术人员所理解，本发明可以以其他特定形式实施而不脱离其精神或本质特征。所以，本文的公开内容和说明旨在展示而不是限制本发明的范围，它在下面权利要求书中阐述。

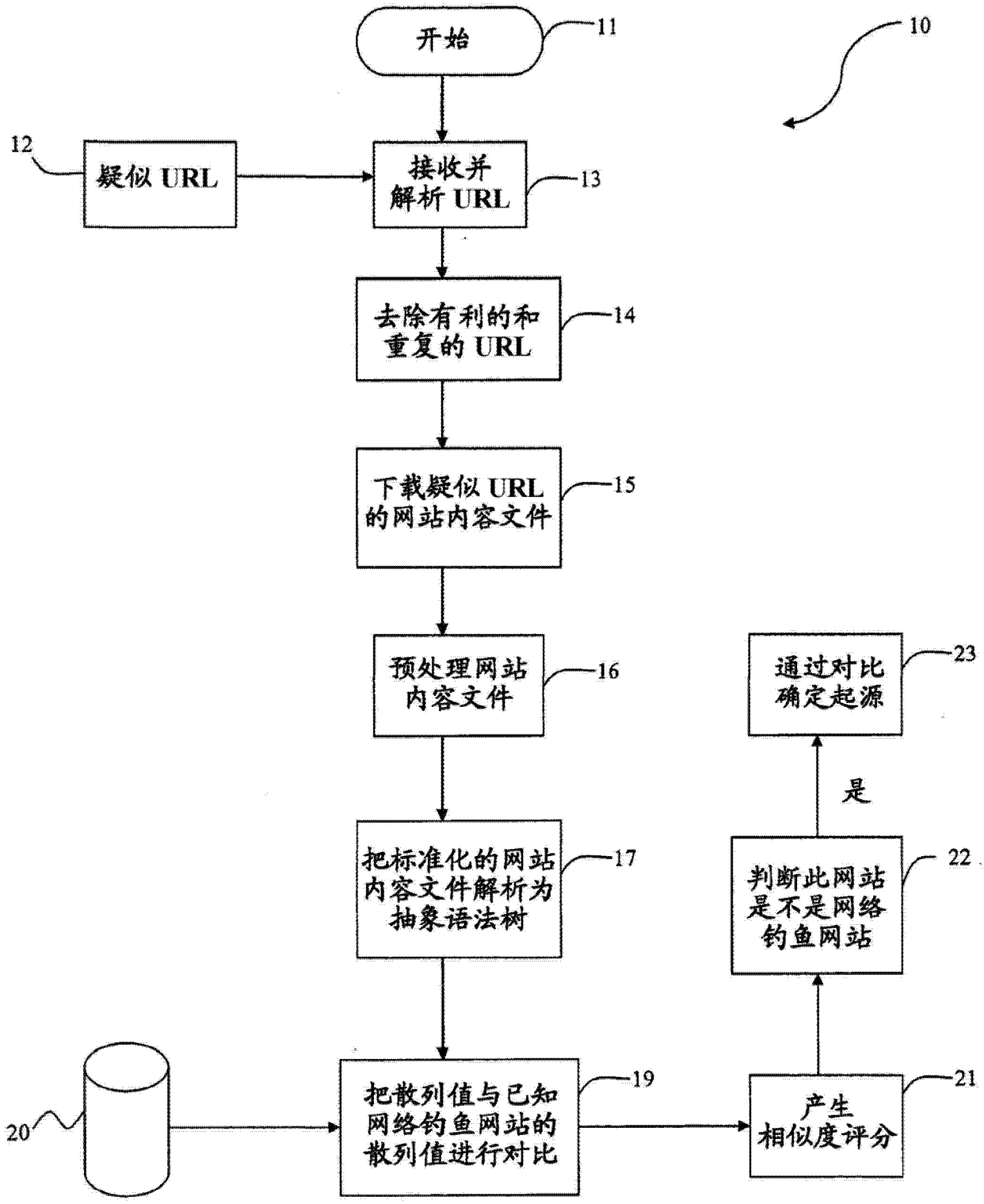

开始
11
10
12
疑似 URL
接收并解析 URL
13
去除有利的和重复的 URL
14
下载疑似 URL 的网站内容文件
15
预处理网站内容文件
16
通过对比确定起源
23
是
把标准化的网站内容文件解析为抽象语法树
17
判断此网站是不是网络钓鱼网站
22
20
把散列值与已知网络钓鱼网站的散列值进行对比
19
产生相似度评分
21

图 1

片段 1:

```
<title>Santander – Mais seguranca e particidade no seu dia-a-dia</title>
<meta http-equiv="Content-Type" content+"text/html; charset=iso-8859-1">
<script language="JavaScript" type="text/JavaScript">
<!—
function MM_reloadPage (init) {  //reloads the window if Nav4 resized
  if (init==true) with (navigator) {if
(( appName=="Netscape")&&(parseInt (appVersion)==4)) {
  document.MM_pgW=innerWidth; document.MM_pgH=interHeight;
onresize=MM_reloadPage; }}
  else if (innerWidth!=document.MM_pgW || innerHeight!=document.MM_pgH)
location.reload();
}
MM_reloadPage (true);
//-->
</script>
```

片段 2:

```
<title>Bradesco – Mais seguranca e particidade no seu dia-a-dia,/title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<script language="JavaScript" type="text/JavaScript">
<!--
function MM_reloadPage (init)  {  //reloads the window if Nav4 resized
  if (init==true) with (navigator) {if
((appName=="Netscape")&&(parseInt)appVersion)==4)) {
  document.MM-pgW=innerWidth; document.mM-pgH=innerHeight;
onresize=MM_reloadPage; }}
  else if (innerWidth!=document.mm_pgW || interHeight!=document.MM_pgH)
location.reload();
}
MM_reloadPage(true);
//-->
</script>
```

图 2

片段 1:

```
<td class="field-form"><spam class="hidden-left">Zipcode first 5 digits
</span><input autocomplete="off" class="align-font-width" name="zipcode5" value=""
maxlength=5" six=5" id="previous_zipcode5" onKeyPress="return
csv_isNumeric (event); " type="text"></td>
```

片段 2:

```
<td class="field-form"><spam class="hidden-left">Zipcode first 5 digits
</span><input autocomplete="off" class="align-font-width" name="zipcode5" value=""
maxlength=5" six=5" id="previous_zipcode5" onkeypress="return
csv_isNumeric (event); " type="text"></td>
```

图 3

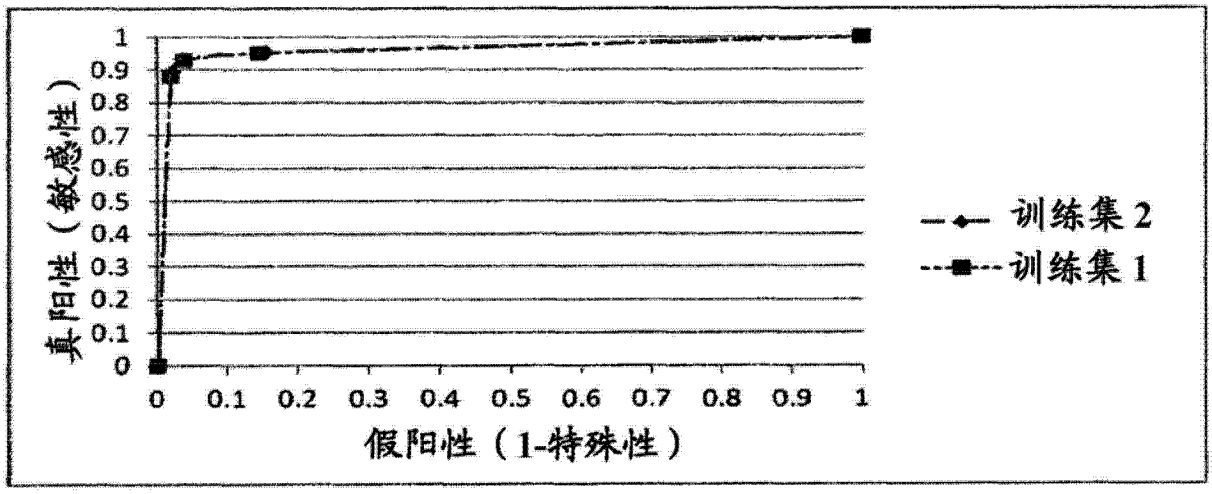


图 4

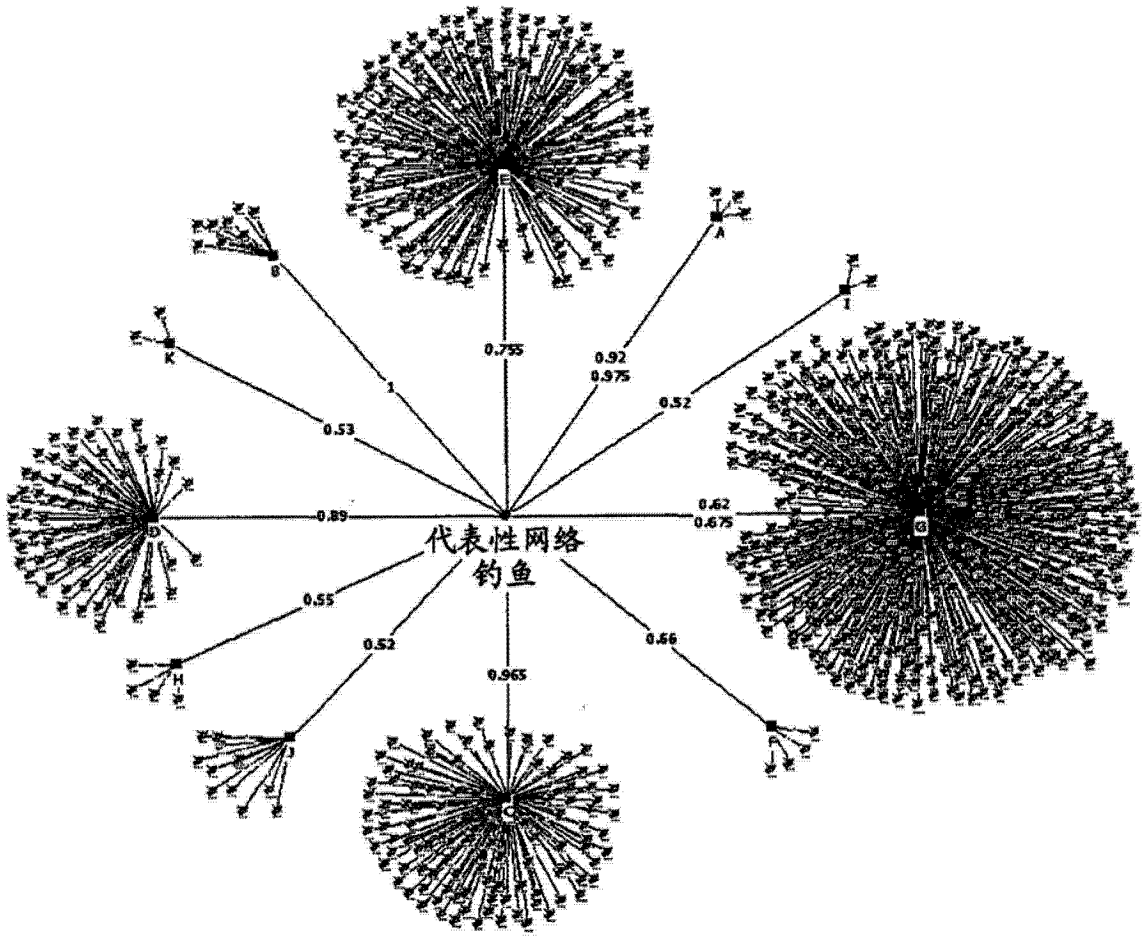


图 5