doi:10.3772/j.issn.1002-0470.2017.05.004

基于分布式 PageRank 算法的可疑目标挖掘①

李国玉② 周广禄③ 张兆心

(哈尔滨工业大学(威海)网络与信息内容安全技术研究中心 威海 264200)

摘 要 考虑到恶意木马、病毒等通过 URL 传播,已对网络安全构成了重大威胁,提出了一种高效的基于分布式 PageRank 链接分析的可疑 URL 目标筛选过滤算法。该算法通过构建 URL 危险性计算分析模型,迭代计算目标危险值,直至收敛状态。最后,根据得到的目标的危险值筛选可疑目标。通过实验验证了该算法的有效性。实验证明,分布式 PageRank 链接分析适应大矩阵计算,可以有效分析挖掘可疑 URL 目标。

关键词 网络安全,云计算,MapReduce,PageRank,可疑目标挖掘

0 引 言

随着互联网的快速发展,网络安全问题日受重视。恶意木马、病毒等通过统一资源定位符(uniform resource locator, URL)传播,致使正常的 URL 中混有恶意目标。这对日常的工作和生活构成巨大的威胁。如何从大量目标中分析出可能的恶意目标,是一个亟待解决的问题。这个问题同样也引起了国内外学者的关注。目前,国内外对于恶意目标的研究主要是通过目标的特征来分析检测恶意 URL。如林海伦[1]等人研究了基于 URL 字段的语义特征的分析方法,甘宏[2]等人研究了基于支持向量机(SVM)和词频-逆文件频率(term frequency-inverse document frequency, TF-IDF)的恶意 URL 识别分析采用的机器学习和内容分类的方法;刘健[3]等人研究了基于特征检测的过滤模型;Mašetic[4]等人研究了基于机器学习的特征分析方法。这些研究均是应用机器学习算法根据特征对恶意目标分析,这种分析方法分析精度高,但敏感性不够强,分析效率低,分析所需要的信息量较多。另外,根据不同的使用信息,恶意 URL 检测方法也不相同。大致可以分为两种方法:基于黑名单的方法和基于网页内容的方法。基于黑名单的方法[5,6]主要是通过查找 URL 黑名单判断给定的 URL 是否为恶意目标。如果命中,则该 URL 为恶意 URL;否则为正确的 URL。基于黑名单的方法只能根据黑名单比对来判断 URL 是否是恶意 URL,其没有可拓展性,其分析的准确性只和黑名单有关。分析恶意 URL 的网页内容具有某种特殊的目的或意义,因此另一种典型的检测恶意 URL 的方法是基于网页内容的方法,该方法借助网页包含的信息,判断给定的 URL 是否为恶意目标。Provos 等[7]提出一种利用网页标签特征检测恶意 URL 的方法;Moshchuk 等[8]利用反间谍软件来分析 URL 网页内容中是否包含木马可执行文件,以此来判定恶意 URL。基于网页内容的方法分析的精度较高,但是该方法需要获取的信息量较大,处理的异常类型单一。

无论是基于机器学习的 URL 目标检测,还是基于信息内容的 URL 检测,均是将每个 URL 作为一个独立的个体分析,并且 URL 特征分析方法和 URL 信息内容分析方法分析所需的信息量较多。本文基于网络互连的特性,研究发现 URL 彼此之间是存在着一定的关联关系的,而这种关联关系会在一定程

① 国家科技支撑计划(2012BAH45B01),国家自然科学基金(61100189,61370215,61370211, 61402137)和国家信息安全 242 计划(2016A104)资助项目。

② 男,1991 年生,硕士;研究方向:网络信息安全;E-mail:liguoyu_a@163.com

③ 通信作者,E-mail:30515509@qq.com

度上相互影响。本文在黑名单方法的基础上,结合URL 关联性提出了一种新的 URL 检测方法:基于PageRank(一种网页排名算法)链接分析的可疑目标 URL 挖掘方法,通过分析目标的整个 URL 链接网,进一步挖掘恶意 URL。该方法分析所需的信息量少,分析精度高。

1 相关技术

PageRank[9] 算法是 Google 排名运算法则的一部分,是用来标识网页等级或重要性的一种方法,是用来衡量一个网站好坏的标准。PageRank 是一个函数,它对 Web 中的每一个网页赋予一个实数值(PR 值),它表示网页的重要程度。也就是说,网页的 PageRank 值越高,那么它就越重要。PageRank 就好比一个投票系统,一个网页的 PR 值还受到其他有链接指向的网页的 PR 值得影响。

PageRank 的计算量大,空间占用过大是 PageRank 算法计算的一个重要问题。因此,分析中多采用分布式 PageRank 算法,而矩阵的计算是分布式 PageRank 算法的核心。目前已有的解决方法有矩阵分割[10] 和矩阵的关系标识方法[11] 等。本文采用矩阵的关系标识方法解决分布式 PageRank 算法矩阵计算问题。

矩阵的关系标识方法介绍:矩阵 $\boldsymbol{M}$ 中第 i 行第 j 列的元素记为 m_{ij} ,矩阵 $\boldsymbol{N}$ 中第 j 行第 k 列的元素记为 n_{jk} ,矩阵 $\boldsymbol{P} = \boldsymbol{MN}$,其第 i 行第 k 列的元素记为 p_{ik} ,其中 $p_{ik} = \sum_j m_{ij} n_{jk}$ 。

在上面的运算中矩阵 $\boldsymbol{M}$ 列数必须和矩阵 $\boldsymbol{N}$ 的行数相等,这时我们就可以把矩阵 $\boldsymbol{M}$ 看成关系,其元组为 $(i,\boldsymbol{M},k,m_{ik})$;也可把矩阵 $\boldsymbol{N}$ 看成关系,其元组为 $(j,\boldsymbol{N},k,n_{jk})$ 。

MapReduce[12] 是一种编程模型,用于大规模数据集的并行运算。概念“Map(映射)”和“Reduce(归约)”,和它们的主要思想,都是从函数式编程语言里借来的,还有从矢量编程语言里借来的特性。当前的软件实现是指定一个 Map 函数,用来把一组键值对映射成一组新的键值对,指定并发的 Reduce 函数,用来保证所有映射的键值对中的每一个共享相同的键组。

2 分布式 PageRank 可疑目标挖掘

可疑目标是指在多个恶意 URL 目标指向的危险性未确定的 URL 目标。PageRank 链接分析是通过链接分析方法,分析未知危险性的 URL 和恶意 URL 的链接关联关系,计算未知危险性的 URL 的危险值。本文对于可疑目标的分析采用 PageRank 的算法分析,通过 PageRank 算法计算各目标节点危险值,以分析目标(已知恶意目标和未知危险性的目标)的危险值作为可疑目标节点选择依据,将最终节点危险值大于预先设定的阈值的节点作为分析得到的可疑目标节点。

2.1 方法介绍

PageRank 基本思想指出:被大量高质量网页引用的网页也是高质量的网页。因此,并将 URL 的危险性作为连接关系,那么就有以下推论:被大量高危险性的 URL 网页连接的 URL 也同样具有较高的危险性。根据推论,算法采用所有 URL 投票表决来选出其中的高危险性的 URL。例如恶意目标 URL 有链接指向未知 URL 的链接 ,那么就对未知 URL 进行一次投票,设恶意目标 URL 危险性系数为 γ ,包含的链接有 μ 个,则从此处得到的危险系数值为 γ/μ ;将 URL 从其他所有投票的 URL 出获得的危险系性系数值相加即得到其危险系数值,通过 PageRank 算法不断迭代分析得到最后的近似危险性系数矩阵(其算法的收敛在 2.2 节中给予证明),最后选取危险系数值大于分析预设阈值的 URL 即是本研究的可疑目标。

可疑目标的 PageRank 算法分析,依据目标节点的危险程度值计算分析出未知目标集合中的可疑目标,但是还存在算法运行计算量大,占用内存空间大的问题。随着节点数目增加,节点链接信息的邻接矩阵表示下的计算量和内存空间占用也远高于线性增长,造成单节点机器无法处理或处理过慢的问题。为解决该问题,本文使用关系标识方法表示矩阵,同时采用 MapReduce 分布式计算模型进行矩阵的迭代计算。

2.2 分析模型

定义目标链接关系的邻接矩阵为 $\boldsymbol{M}=(\rho_{ij})$，其中如果目标 i 有指向 j 的链接，即对应的有向图存在从 i 到 j 的边，那么 $\rho_{ij}=1$，否则 $\rho_{ij}=0$。

设共 m 个目标 URL，将它们的危险系数分别记为 $\gamma_1,\gamma_2,\cdots,\gamma_m$，令 $\sum_i^m \gamma_i=1$，$\boldsymbol{G}=(\rho_{ij}/\eta_i)$（$\eta_i$ 表示目标 URL j 链接到其他目标的链接数，$\boldsymbol{G}$ 表示概率转移矩阵），目标 i 存在危险的概率为 $p_i(p_i<1)$，则有

$$\gamma_i=(1-p_i)/m+p_i\sum_{j=1}^{m}\rho_{ji}/\eta_j\cdot\gamma_j \tag{1}$$

设 $\boldsymbol{\gamma}=(\gamma_1,\gamma_2,\cdots,\gamma_m)$，$\boldsymbol{e}=(1,1,\cdots,1)$，根据式(1)于是有

$$\begin{aligned}\boldsymbol{\gamma}&=p\cdot\boldsymbol{\gamma}\cdot\boldsymbol{G}+(1-p)/m\cdot\boldsymbol{e}\\&=\boldsymbol{\gamma}(p\cdot\boldsymbol{G}+(1-p)/m\boldsymbol{e}^{\mathrm{T}}\cdot\boldsymbol{e})\end{aligned} \tag{2}$$

令

$$\boldsymbol{A}=p\cdot\boldsymbol{G}+(1-p)/m\cdot\boldsymbol{e}^{\mathrm{T}}\cdot\boldsymbol{e} \tag{3}$$

根据式(2)、(3)于是有 $\boldsymbol{\gamma}=\boldsymbol{\gamma A}$，得到结论：

结论 1 $\boldsymbol{A}$ 是一个状态转移矩阵，且有以下性质：所有元素都是整数；各行元素之和等于 1。

证明：令 g_{ij} 表示矩阵 $\boldsymbol{G}$ 中的第 i 行 j 列的元素值，由式(3)得

$$\boldsymbol{A}=\begin{bmatrix}\frac{pg_{11}}{\eta_1}+\frac{1-p}{m} & \frac{pg_{12}}{\eta_1}+\frac{1-p}{m} & \cdots & \frac{pg_{1m}}{\eta_1}+\frac{1-p}{m}\\ \frac{pg_{21}}{\eta_2}+\frac{1-p}{m} & \frac{pg_{22}}{\eta_2}+\frac{1-p}{m} & \cdots & \frac{pg_{2m}}{\eta_2}+\frac{1-p}{m}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{pg_{m1}}{\eta_m}+\frac{1-p}{m} & \frac{pg_{m2}}{\eta_m}+\frac{1-p}{m} & \cdots & \frac{pg_{mm}}{\eta_m}+\frac{1-p}{m}\end{bmatrix} \tag{4}$$

由 $p<1$ 得

$$\frac{pg_{ij}}{\eta_i}+\frac{1-p}{m}>0 \tag{5}$$

由式(5)得

$$\begin{aligned}\sum_{j=1}^{m}\left(\frac{pg_{ij}}{\eta_i}+\frac{1-p}{m}\right)&=1-p+\frac{p}{\eta_i}\sum_{j=1}^{m}g_{ij}\\&=1-p+\frac{p}{\eta_i}\eta_i\end{aligned} \tag{6}$$

结论 2 $\boldsymbol{\gamma}=\boldsymbol{\gamma A}$ 必有唯一解。

证明：由结论 1 易得 $\boldsymbol{A}^{\mathrm{T}}$（$\boldsymbol{A}^{\mathrm{T}}$ 表示矩阵 $\boldsymbol{A}$ 的转置）的各列之和为 1，所以 $\lambda=1$ 是 $\boldsymbol{A}^{\mathrm{T}}$ 的特征值，$\boldsymbol{e}=\boldsymbol{eA}^{\mathrm{T}}\Rightarrow\boldsymbol{e}(\boldsymbol{I}-\boldsymbol{A}^{\mathrm{T}})=0\Rightarrow|I-A^{\mathrm{T}}|=0$（$\boldsymbol{I}$ 表示主对角线全为 1，其他全为零的单位矩阵，I 表示矩阵 $\boldsymbol{I}$ 的行列式）。

由谱半径公式和 Perron-Forbennius 定理可得 $\theta(\boldsymbol{A}^{\mathrm{T}})=1$ 有且只有唯一的 $\boldsymbol{\gamma}^{\mathrm{T}}$ 使得 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\gamma}^{\mathrm{T}}=\theta(\boldsymbol{A}^{\mathrm{T}})\boldsymbol{\gamma}^{\mathrm{T}}=\boldsymbol{\gamma}^{\mathrm{T}}$。

2.3 PageRank 链接分析的可疑目标分析算法

上面结论 2 可知 $\boldsymbol{\gamma}$ 可以通过迭代求得。在实际应用中，由于求解的效率和实际需求，我们并不需要真正求得最终的 $\boldsymbol{\gamma}=\boldsymbol{\gamma A}$ 的精确解，只需求得等式的一个满足精度要求的近似解。根据分布式 PageRank 算法，对目标 URL 分析，算法流程图如图 1。

3 实验结果分析

实验在 Hadoop 分布式运算平台下测试基于 MapReduce 的 PageRank 算法对于 URL 目标挖掘可疑 URL 目标。通过对数据分析结果与权威恶意域名检测网站检测结果，资源占用情况，分析效率等对比得出结论。

3.1 数据来源

数据集中的恶意 URL 均来源于哈尔滨工业大学国家计算机信息内容安全重点实验室恶意域名数据库。获取 URL 数据有两种途径：一是通过爬虫爬取恶意 URL 获取的未知危险性的 URL 数据，二是通过实际爬取这些 URL 链接获得 URL 包含的链出关系数据集。在经过多次实验，收集到在单机矩阵运算和分布式计算不同数据量下分析程序单节点内存占用和计算消耗时间。

3.2 实验结果分析

如下图 2 和图 3 所示，对于可疑目标挖掘构建的矩阵，在 MapReduce 分布式计算模式下单节点运算占用的内存和 CPU 在随着矩阵维度的增长速率远小于普通矩阵运算随着矩阵维度增长速率。根据实验结果，可以看出在数据维度很大时，单节点的处理能力和资源不能满足计算分析的需求。

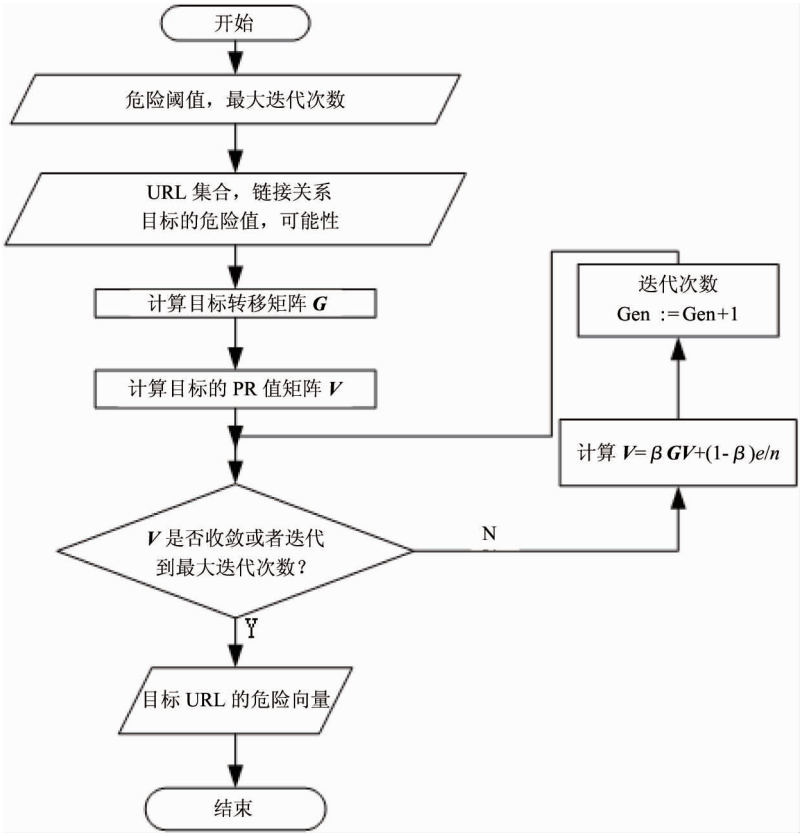

图 1　算法流程图

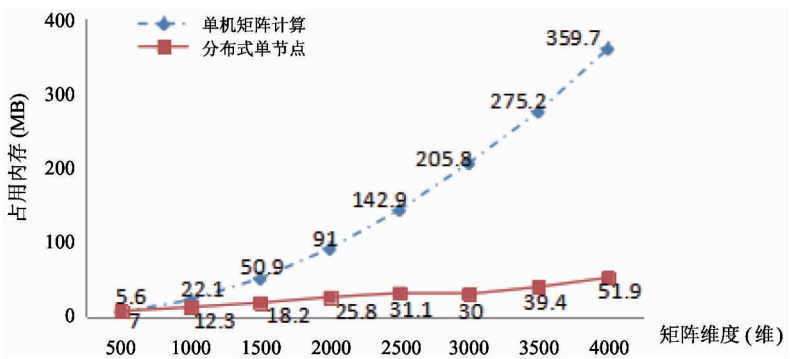

图 2　计算矩阵内存占用和矩阵维度关系折线图

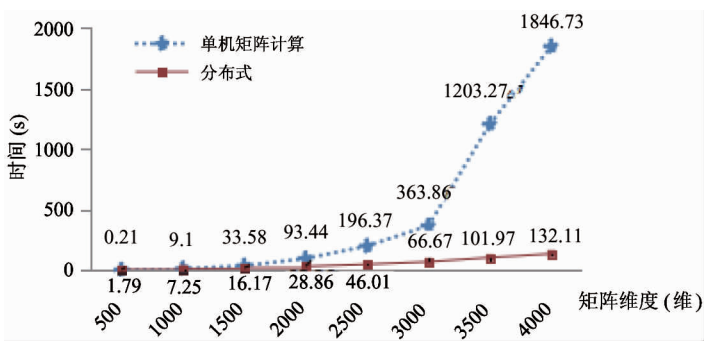

图 3　矩阵计算时间与矩阵维度的关系折线图

通过 PageRank 链接分析，从分析结果中提取 PR 值 top 5 的 URL 链接，然后将链接在 Phish Tank 和 360 网络安全等权威网站上进行安全检测，最终得到检测结果如表 1 所示。从表中可以看出，算法能够有效识别 Phish Tank 检测的钓鱼网站和 360 网络安全检测的其他类型的恶意 URL。通过结果分析可以得出，算法分析结果 PR 值较大的 URL 是恶意目标的可能性越大。

表 1　算法分析结果检验

URL（PR 值 top 5）	算法判定结果	Phish Tank 检测结果	360 安全检测结果
http://djchennai.com/stf/index.html	可疑	钓鱼	危险
http://santaclararecycling.com/tmp/adobe/adobes/ok/	可疑	钓鱼	高危
http://www.securityfocus.com/bid/94975	可疑	未知	危险
http://basniowakraina.leszno.pl	可疑	钓鱼	高危
http://www3.xcu.edu.cn/keyanchu/readnews.asp? newsid=377	可疑	未知	低危

如表 2 所示，该数据是根据实验分析结果和恶意域名检测网站检测结果对比统计出来的。根据统计结果可以看出 PageRank 算法的可疑目标分析具有较高的准确性，能够检测出 80% 以上的恶意 URL，这充分说明通过 PageRank 算法能有效发掘潜在的恶意 URL。本文通过与机器学习的特征分析方法的准确率对比分析。如图 4 对比所示，本文的 PageRank 链接分析准确率比单一站点特征分析的准确率高，但其分析精确率较多特征分析精确率要低。图 2 和图 3 显示 PageRank 算法的分布式实现方式能够有效解决分析中数据量过大造成的资源问题和运行效率问题。

表 2　实验分析结果统计表

	预测是恶意 URL	预测是非恶意 URL	总计
恶意 URL	1667	327	1994
非恶意 URL	1743	9195	10938
总计	3410	9522	12932

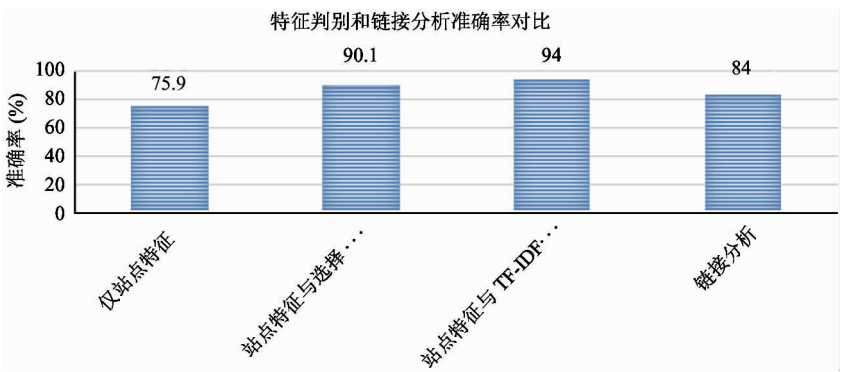

图 4　基于机器学习的特征分析方法和链接分析方法的分析精度对比

4　结　论

本文通过对可疑恶意目标挖掘方法的研究，采用基于分布式 PageRank 的链接分析算法挖掘危险目标 URL，并通过实验验证算法的有效性。实验结果证明，PageRank 的链接分析算法对恶意 URL 目标挖掘和发现具有显著的效果。本文还将分析结果与机器学习的特征分析方法对比，验证基于 PageRank 的链接分析算法具有较高的分析精度。同时其算法的分布式实现和矩阵的关系表示方法解决了算法单机运行的局限和集群实施的难题，使得算法分析过程更加高效。然而，链接分析的算法的可疑目标分析方法还只是一种尝试性的方法，目前实验是在假

设每一恶意 URL 的危险等级相同的情况下做出的分析,其精度和适应性方面可能仍存在一些需要改进的地方,这也是接下来本文要进行的工作。

参考文献

[1] 林海伦,李焱,王伟平等. 高效的基于段模式的恶意 URL 检测方法. 通信学报, 2015,36 (Z1): 141-148

[2] 甘宏, 潘丹. 基于 SVM 和 TF-IDF 的恶意 URL 识别分析与研究. 计算机与现代化, 2016,7: 95-97

[3] 刘健, 赵刚, 郑运鹏. 恶意 URL 多层过滤检测模型策略研究. 信息安全研究, 2016, 2(1): 80-85

[4] Mašetic Z, Subasi A, Azemovic J. Malicious web sites detection using C4. 5 decision tree. *Southeast Europe Journal of Soft Computing*, 2016, 5(1):68-72

[5] Prakash P, Kumar M, Kompella R R, et al. Phishnet: predictive blacklisting to detect phishing attacks. In: Proceedings of the IEEE INFOCOM, San Diego, USA, 2010. 1-5

[6] Likarish P, Jung E. Leveraging google safebrowsing to characterize web-based attacks. In: Proceedings of the Association for Computing Machinery, Chicago, USA, 2009. 3-5

[7] Provos N, Mavrommatis P, Monrose M A, et al. All your iframes point to us. In: Proceedings of the 17th USENIX Security Symposium, San Jose, USA, 2008. 1-16

[8] Moshchuk A, Bragin T, Gribble S D, et al. A crawler-based study of spyware in the web. In: Proceedings of the Network and Distributed System Security Symposium, San Diego, USA, 2006. 1-17

[9] Langville A N, Meyer C D. Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton: Princeton University Press, 2012. 68-69

[10] 安建瑞, 王海鹏, 张龙波等. 一种基于 MapReduce 的压缩矩阵关联规则挖掘算法. 重庆理工大学学报: 自然科学版, 2016, 30(2): 95-100

[11] Rajaraman A, Ullman J D. Mining of Massive Datasets. Cambridge: Cambridge University Press, 2011. 16-35

[12] Narasimhaiah M N, Sam R P. An introduction to map reduce approach to distribute work using new set of tools. *International Research Journal of Engineering and Technology*, 2015,2(3):1623-1625

Suspicious target mining based on distributed PageRank algorithm

Li Guoyu, Zhou Guanglu, Zhang Zhaoxin

(Network and Information Security Technology, Harbin Institute of Technology, Weihai 264200)

Abstract

Considering that Malicious Trojans, viruses and others spread through URL and this poses a major threat to network security, the study proposed an efficient algorithm for screening and filtering suspicious URL targets based on distributed PageRank link analysis. The algorithm through constructing a model for calculation and analysis of URL hazard iteratively calculates the valus of target risks, and this continues until the convergence state. Finally, it selects the suspicious target according to its dangerous value. The effectiveness of the algorithm was verified by experiment, and the experimental results show that the distributed PageRank link analysis can adapt to large matrix computation, and can effectively analyze suspicious URL targets.

Key words: network security, cloud computing, MapReduce, PageRank, suspicious target mining