doi:10.3772/j.issn.1002-0470.2017.08.004

基于 URL 文本特征及链接关系的钓鱼网站识别算法①

赵蹲宇② 张兆心③

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 为了提高对钓鱼网站的识别准确率,通过对钓鱼网站统一资源定位符(URL)文本数据的分析,结合钓鱼网站内部链接关系组成的网络拓扑结构特征,提出了基于 URL 文本特征及链接关系的钓鱼网站识别算法 FAUFL。该算法的原理是:以 URL 文本特征作为输入,采用随机森林算法生成基于 URL 文本特征的钓鱼网站判别算法;以链接关系作为输入构建相关网页群,采用基于最大流切割的相关网页群算法生成基于链接关系的钓鱼网站判别算法;将上述两种判别算法结果作为输入,采用 Bagging 算法进行进一步评估。测试结果表明钓鱼网站识别算法 FAUFL 算法的识别准确率为99.2%,比基于 URL 文本特征的算法的准确率提高3.9%,比基于链接关系的算法提高5.0%。

关键词 钓鱼网站,融合算法,统一资源定位符(URL),文本特征,链接关系

0 引 言

网络钓鱼已对网上交易构成重大威胁,如何高效识别钓鱼网站,是目前亟待解决的一个重要问题。钓鱼网站通过高度模仿真实网站来欺骗用户,其主要特点表现在网站统一资源定位符(uniform resource location, URL)极其相似,网页内容和链接关系与真实网页高度相似。目前判定网络钓鱼站点方法主要有三种,即基于黑名单的方法、基于网页内容的方法和基于 URL 的方法。基于黑名单的方法为建立黑名单集合,通过查找黑名单对给定 URL 进行判定。基于网页内容的方法是通过提取网页内包含的信息,如:HTML 标签,文本等进行特征提取,通过采用该特征集进行钓鱼网站的判定。基于 URL 的方法是通过对钓鱼网站 URL 的文本特征或链接关系进行钓鱼网站识别。

基于 URL 的钓鱼网站识别主要包括两方面:一是通过对 URL 文本特征对钓鱼网站进行识别,二是通过针对网站内部 URL 链接关系对钓鱼网站进行识别。2016 年赵加林[1]等人研究了基于 K-Means 和支持向量机(SVM)的钓鱼网站识别,2016 年 Mašetic[2]等人研究了“决策树恶意 Web 发现方法”,分析了采用机器学习和内容分类的方法,2013 年朱百禄[3]等人进行了基于 Web 社区的钓鱼网站检测研究,2014 年滕雯静[4]等人研究了基于链接分析的钓鱼网站检测方法。以上研究均从单一的维度对钓鱼网站进行识别,基于 URL 文本特征的钓鱼网站识别算法对新特征不敏感,在新特征出现时,准确率大大降低。基于 URL 链接关系的钓鱼网站识别算法识别准确率不够高。由于钓鱼网站危害极大,因此钓鱼网站识别准确率极为重要。无论通过 URL 文本特征对钓鱼网站识别,还是通过网页内 URL 链接关系对钓鱼网站识别,其准确率均未达到要求。URL 文本特征的判别算法对新特征敏感性差。网页链接关系判别算法识别准确率较低。因

① 国家重点研发计划(SQ2017YFGX110125-01),国家自然科学基金(61370215,61370211,61402137),国家科技支撑计划(2012BAH45B01)和国家信息安全计划(2017A065,2017A111)资助项目。

② 男,1994 年生,硕士;研究方向:网络信息安全;E-mail: dunyuzhao@163.com

③ 通信作者,E-mail: heart@hit.edu.cn

(收稿日期:2017-05-18)

此,本文提出将两个维度的判别方式进行融合,共同作为最终判别结果的依据。通过多维度的判别结果的融合来提高判别准确率。由于 Bagging 算法在集成学习中的效果比贝叶斯算法、加权算法以及加权贝叶斯更加显著,因此本文首次将 Bagging 融合算法用于钓鱼网站判定。本文提出了基于 URL 文本特征及链接关系的钓鱼网站识别算法(fishing website identification algorithm based on URL text features and link relation,FAUFL),该算法通过采用 Bagging 融合[5-7]学习算法将两个算法的判别结果进行融合而提高钓鱼网站识别准确率。

1 相关算法

基于 URL 文本特征的钓鱼网站识别方法主要包括支持向量机[8]、K-近邻等,因随机森林算法[9]对 URL 文本特征判别的准确率高于前几种算法,本文采用随机森林算法进行钓鱼网站判别算法构建。基于 URL 链接关系的钓鱼网站识别方法主要包括 HITS[10]、PageRank[11,12],Companion[13] 等。因基于最大流切割的相关网页群算法的相关网页群构建效果好于前几种,本文采用该算法构建链接关系钓鱼网站判别算法。同时使用 Bagging 集成算法,将 URL 文本特征和链接关系的判别结果进行综合,从而进一步提高判别的准确率。以下简要介绍各算法:

(1) 支持向量机。支持向量机(SVM)将向量映射到一个更高维的空间里,在这个空间里建立有一个最大间隔超平面。在分开数据的超平面的两边建有两个互相平行的超平面。建立方向合适的分隔超平面使两个与之平行的超平面间的距离最大化。其假定平行超平面间的距离或差距越大,分类器的总误差越小。

(2) K-近邻。K-近邻(nearest neighbor, NN)是一个理论上比较成熟的方法,也是最简单的机器学习算法之一。该方法的思路是:如果一个样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别,则该样本也属于这个类别。

(3) 随机森林。随机森林(Random Forest,RF)是采用随机方式去建立一个森林,该森林是由很多决策树构成,随机森林的每棵决策树之间是没有关联的。在得到一个随机森林之后,当有一个新的输入样本时,就让随机森林中的每一棵决策树分别进行判断,判断结果为该样本应该属于哪一类,然后通过计算判定结果中哪一类被选择最多,则该样本就被预测为该类。

(4) PageRank。PageRank 通过超链接关系来确定页面的等级。把从 A 页面到 B 页面的链接解释为 A 页面给 B 页面投票,根据投票来源(甚至来源的来源,即链接到 A 页面的页面)和投票目标的等级来决定新的等级。简单地说,一个高等级的页面可以使其他低等级页面的等级提升。

(5) HITS 算法。HITS 算法是超链接诱导的主题搜索(hyperlink induced topic search),HITS 算法是迭代算法,该算法从网络子图中抽取中心网页和权威网页。HITS 算法的重要思想是,一个好的中心网页一定会链接到很多好的权威网页,同时一个好的权威网页一定会被很多好的中心网页链接。被越多的中心网页链接的网页就越权威,链接到的权威网页越多的网页就越中心,中心网页与权威网页之间相互促进。

(6) Companion 算法。Dean 等人提出 Companion 算法,是以 HITS 算法为基础,输入种子网页,输出与种子网页相关的网页。Companion 是一种特殊的 HITS 算法,通过为链接加权以及网页上链接顺序来提高网络社区识别精确性。Companion 首先建立种子页面的网络子图,然后采用 HITS 算法抽取权威网页和中心网页,最后把权威网页作为相关网页返回。

(7) 基于最大流切割的相关网页群算法分析。基于最大流[14]切割的相关网页群(max-flow relevant web group, MFRG)算法是对 Companion 算法进行改进。通过迭代和筛选构造一个相互关联很强的相关网页群。在构造过程中应用最大流切割算法来提高相关网页群的关联强度。最后返回前五的权威网页和前五的中心网页。该权威网页和中心网页的集合,称为可疑网页潜在的目标网页群,即目标网页的

范围缩小到这个更小的范围内。

(8) Bagging 算法。Bagging 算法(bootstrap aggregating,引导聚集)基于数据随机重抽取的分类器构建方法,也称 Bagging 方法,是在从原始数据集选择 S 次后得到 S 个新数据集的一种技术。新数据集和原数据集的大小相等。每个数据集都是通过在原始数据集中随机选择一个样本来进行替换而得到的。这里的替换就意味着可以多次地选择同一样本。这一性质就允许新数据集中可以有重复的值,而原始数据集的某些值在新集合中则不再出现。

2 FAUFL

本文提出的基于 URL(统一资源定位符)文本特征及链接关系的钓鱼网站识别算法 FAUFL,通过 Bagging 融合算法将基于随机森林的 URL 文本特征钓鱼网站辨别算法和基于链接关系的钓鱼网站辨别算法进行融合,以得到更高的钓鱼网站判别准确率。

由于 URL 存在两种属性,即 URL 文本属性和 URL 链接属性。本文对于 URL 文本属性,采用随机森林算法对 URL 多个特征进行建模。本文对于 URL 链接属性可发现相应网页群,通过最大流切割的相关网页群算法建立建模。但上述单一算法准确率并未达到要求,因此通过 Bagging 算法将上述两种算法进行融合,得到更高准确率。

2.1 钓鱼网站 URL 文本特征提取

URL 文本特征是指出现在 URL 文本中的某些字符或单词。一般包括词汇特征和结构特征。通过不同维度提取相应的特征值。

钓鱼网站 URL 文本特征关系到钓鱼网站分类的准确率,本文对钓鱼网站 URL 文本进行了分析,如对 http://opersolutions. com. ar/img/spacer/Acrobat/index. php? ar = info@ gmail. com 等钓鱼网站的 URL 文本特征(包括特殊字符“@”,含有一定数量的“.”URL 路径等特征等)进行了分析。本文总结出 14 个文本特征,其中包括 4 个词汇特征、10 个结构特征,组成钓鱼网站 URL 敏感文本特征向量 $\boldsymbol{FV}$:

$$\boldsymbol{FV} = <F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12}, F_{13}, F_{14}>$$

特征 F_1 到 F_4 是 URL 的词汇特征,特征 F_5 到 F_{14} 是 URL 的结构特征。词汇特征为钓鱼网站 URL 文本中是否含有敏感词汇。w 为敏感词汇集合,$w \in \{\text{bank, sign, webscr, confirm, account, ebay, user, secure, login}\}$。在 14 个特征中,某些特征值类型为 bool 型时,其中赋“1”代表为钓鱼网站;赋“0”代表为合法网站。某些特征值类型为 int 型时,则记录相应特征情况。文本特征 F_1 到 F_{14} 见表 1。

表 1 钓鱼网站 URL 特征描述

特征变量	含义	特征值计算
F_1	域名中是否含有品牌关键字	bool 型,域名中含有 $F_1 = 1$
F_2	域名中是否含有模糊品牌关键字	bool 型,域名中含有 $F_2 = 1$
F_3	URL 中是否含有品牌关键字	bool 型,URL 中含有 $F_3 = 1$
F_4	URL 中是否含有模糊品牌关键字	bool 型,URL 中含有 IP $F_4 = 1$
F_5	URL 中特殊字符“@”个数	int 型
F_6	URL 长度	int 型
F_7	域名长度	int 型
F_8	域名的级数	int 型
F_9	URL 路径的级数	int 型
F_{10}	URL 中存在一定超长的数字串	int 型
F_{11}	URL 中含有一定数量的“.”	int 型
F_{12}	URL 中含有 IP	bool 型,URL 中含有 IP $F_{12} = 1$
F_{13}	URL 中含有 80 端口	bool 型,URL 中不含有 80 端口 $F_{13} = 1$
F_{14}	URL 中主机数目	int 型

2.2 算法设计

由于 URL 文本特征的判别算法对新出现的特征敏感性不够强,因此本文提出采用随机森林算法对 URL 文本特征进行建模,通过多个决策树进行表决来提高准确率。对于网页链接关系的判别算法准确率低的问题,本文应用基于最大流切割的相关网页群算法来提高准确率。再通过 Bagging 算法将两种算法的判别结果进行融合,从而进一步提高检测的准确率。

2.2.1 基于 URL 文本特征的钓鱼网站判别算法设计

由于钓鱼网站 URL 文本具有很多相似判别特征,因此本文采用基于随机森林的 URL 文本特征判别算法。首先对 URL 文本提取出相应文本特征,然后建立多棵随机决策树并对多个判别结果投票表决,最后投票数多的结果为最终判别结果。如图 1 所示。

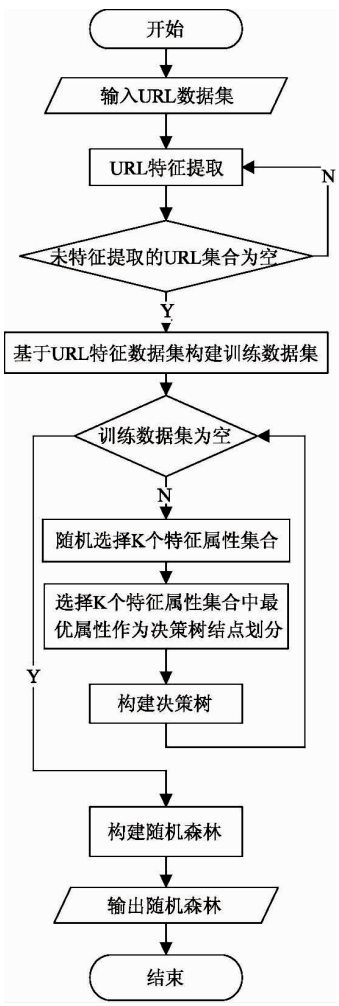

图 1 文本特征的随机森林算法流程图

具体算法如下:

定义:url:单个 URL;U:URL 数据和判别结果集合;$\boldsymbol{FV}$:URL 特征向量集合;$\boldsymbol{Fv}$:单个 URL 特征向量;y:判别结果(0/1);vector:(fv,y) 特征变量和判别结果的集合;D:特征向量和判别结果的集合;RandomTree:随机决策树;result:判定结果集合。

1)特征变量提取

Name: FeatureVariablesExtraction( )

Input: URL set $U = \{(url_1, y_1), (url_2, y_2), \cdots, (url_m, y_m)\}$;

Process:

for $url \leftarrow U = \{(url_1, y_1), (url_2, y_2), \cdots, (url_m, y_m)\}$;

$fv_i \leftarrow$FeatureExtraction (url_i)

vector$_i = (fv_i, y_i)$

$D \leftarrow D +$ vector$_i$

return D

Output: Data set $D = \{(fv_1, y_1), (fv_2, y_2), \cdots, (fv_m, y_m)\}$;

2)决策树构建

Name: RandomDecisionTree( )

Input: Data set $D = \{(fv_1, y_1), (fv_2, y_2), \cdots, (fv_m, y_m)\}$;

Feature subset size K.

Process:

RandomTree$\leftarrow$CreateTree(D)

If all _ instances in the same class

return RandomTree

$F \leftarrow$best _ feature( )

if $F ==$ NULL

return RandomTree

$\tilde{F} \leftarrow$random _ select _ features(K)

RandomTree. $f \leftarrow$feature _ best _ split _ point ($\tilde{F}$)//the feature which has the best split point in $\tilde{F}$;

RandomTree. $p \leftarrow$best _ split _ point (RandomTree. f)

$D_l \leftarrow$RandomTree. $f <$ RandomTree. p

$D_r \leftarrow$RandomTree. $f >=$ RandomTree. p

RandomTree$_l \leftarrow$RandomDecisionTree (D_l, K);

RandomTree$_r$←RandomDecisionTree (D_r, K);

return RandomTree

Output: A random decision tree

3）随机森林构建

Name: RandomForests( )

Input: RandomTrees = (RT_1, RT_2, RT_3,…, RT_n); url;

Process:

for i in RandomTrees (RT_1, RT_2, RT_3,…, RT_n)

$rf_$result = ← $rf_$result + RT_1 (url)

$y+$ ← $rf_$result. $y+$

$y-$ ← $rf_$result. $y-$

result←compare ($y+$, $y-$)

return result

Output: result

随机森林(Random Forests,RF)算法,首先通过特征提取(FeatureVariablesExtraction( ))对 URL 提取相应特征向量,然后对已提取的特征向量构建多个随机决策树(RandomDecisionTree( ))构成随机森林(RandomForests( )),最后在判别时通过随机森林(RandomForests( ))的判断结果,即多个随机树判定结果超过半数的结果为最终判断结果。随机森林算法输出结果“0”代表合法网站,“1”代表钓鱼网站。

2.2.2 基于网页链接关系的钓鱼网站判别算法设计

由于网页存在群居关系特征,因此本文采用基于网页链接关系的钓鱼网站判别算法。首先对可疑网页的页内链接关系计算相应链接关系,建立链接关系拓扑;然后对该拓扑采用最大流切割算法进行迭代地切割与计算,直到可疑网页在拓扑图中,即可疑网页被判定为合法网页、迭代次数达到上限或网络拓扑不再变化为止。如图 2 所示。

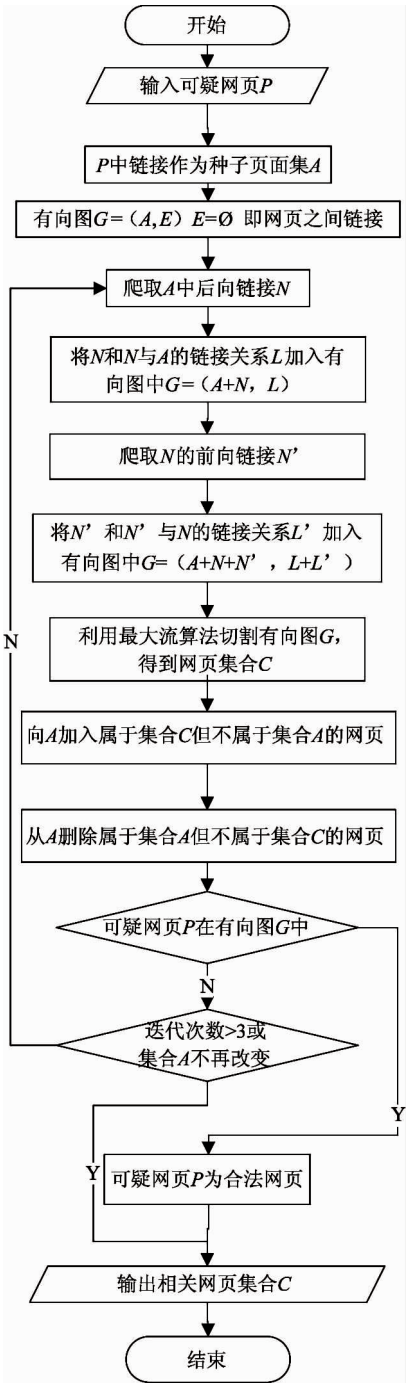

图 2 链接关系的相关网页群算法流程图

算法具体如下:

定义: $G(V,E)$:G 为有向图, V 为顶点即网页, E 为边即网页之间的链接关系; P: 可疑网页; A:P 中链接网页集; N:A 的后向链接集; L:N 与 A 的链接关系;N':N 的前向链接集;L':N' 与 N 的链接关系 L';C: 可疑网页的相关网页群集合。

Name:MaxFlowRelevantWebGroup( )

Input: P

Process:

A ←the link of P

G ← (V,E) $V=\varnothing$ $V \leftarrow V+A$

G ← (A,E) $E=\varnothing$

N ← A to link after N L ←link relation

$V \leftarrow V + N \quad E \leftarrow E + L$

$G \leftarrow (A + N, L)$

$N' \leftarrow N$ the forward link N' $L' \leftarrow$ link relation

$V \leftarrow V + N' \quad E \leftarrow E + L'$

$G \leftarrow (A + N + N', L + L')$

$C \leftarrow$ max-flow algorithm to cut G

$A \leftarrow A + (C - (C \cap A))$ // To join in A belongs to the set C but do not belong to A web page,

$A \leftarrow A - (A - (A \cap C))$ //removed from the collection of A belongs to set A but do not belong to the set C web pages;

if $P \in G$ or $n > 3$ or A not change//n (iterations)

then return C

else goto 4-10 $n++$//Step 4-10 iterations

Output: C

基于最大流切割的相关网页群(max-flow relevant web group,MFRG)算法。首先对可疑网页的页内链接统计相应的后向链接集;其次根据后向链接集,统计其前向链接,根据后向链接集合和前向链接集链接关系构成相应的链接关系拓扑图 G; 然后对该链接关系拓扑图 G 采用最大流切割算法进行切割,迭代统计与切割链接关系拓扑图 G, 最后直到可疑网页在拓扑图中,即可疑网页被判定为合法网页、迭代次数达到上限或网络拓扑不再变化为止。基于最大流切割的相关网页群算法输出结果"0"代表合法网站,"1"代表钓鱼网站。

2.2.3 FAUFL 的设计

由于钓鱼网站 URL 同时存在 URL 文本判别特征相似性和网页群居关系特征,因此本文提出基于 URL 文本特征和链接关系的 Bagging 融合钓鱼网站判别算法。首先建立多个大小相同训练集,然后不同数据集分别对基于随机森林的 URL 文本判别算法和基于最大流切割的网页链接关系判别算法进行训练,最后对多个上述两种算法的结果进行统计,投票结果最多的为最终判别结果。如图 3 所示。

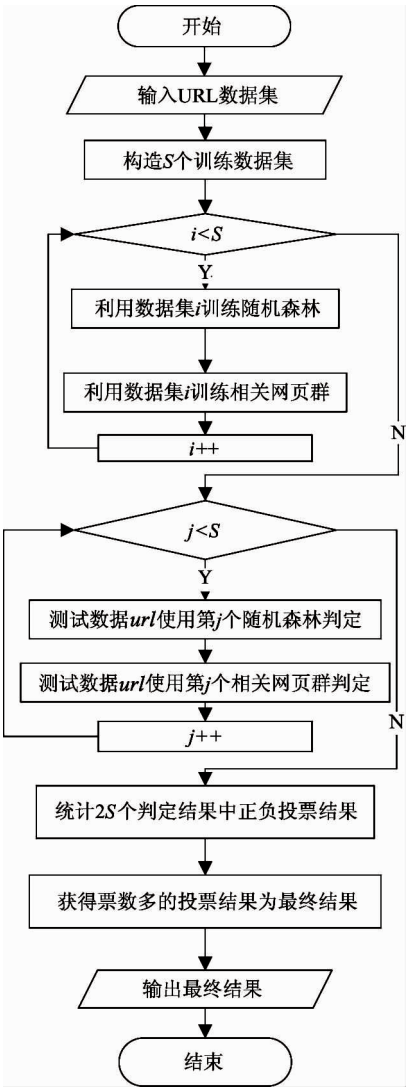

图 3 Bagging 融合算法流程图

算法具体如下:

定义: U: URL set $U = \{(url_1, y_1), (url_2, y_2), \cdots, (url_m, y_m)\}$; url: (url_i, y_i) URL 和类型集合; S: 训练数据集; result: 钓鱼网站的判定结果; rf: 随机森林钓鱼网站判别结果; $mfrg$: 基于最大流的网页链接关系钓鱼网站判别结果。

Name: BaggingBasedOnUTFAndLR()

Input: U: URL set $U = \{(url_1, y_1), (url_2, y_2), \cdots, (url_m, y_m)\}$;

Process:

Based on Bootstrapping sampling structure S data sets

for i in $S(0, \cdots S-1)$

$RF_i(S_i)$

MFRG $i(S_i)$

for j in $S(0,\cdots S-1)$

$rf\_result \leftarrow rf\_result + RFj(url)$

$mfrg\_result \leftarrow mfrg\_result + MFRGj(url)$

$y+ \leftarrow rf\_result.\ y+ + mfrg\_result.\ y+$

$y- \leftarrow rf\_result.\ y- + mfrg\_result.\ y-$

result←compare $(y+, y-)$

return result

Output:result

基于 URL 文本特征和链接关系的 Bagging 融合算法(BaggingBasedOnUTFAndLR())，首先通过有放回抽样构成 S 个训练集，然后对每个训练集分别采用 RF 和 MFRG 算法进行训练，训练结果总计 S 个随机森林(RF)算法模型和 S 个最大流切割相关网页群(MFRG)算法模型，最后判断测试集 URL 时，将 $2\times S$ 个结果集中超过半数的判断结果为最终判断结果。基于 RF 和 MFRG 的 Bagging 融合算法输出结果“0”代表合法网站，“1”代表钓鱼网站。

3 实验结果及分析

为验证本文提出算法的有效性，构建相应的测试数据集及测试环境，对于基于随机森林的 URL 文本特征钓鱼网站判别算法，基于网络链接关系的钓鱼网站判别算法和基于 URL 文本特征及链接关系的钓鱼网站判别算法进行测试。

实验代码的编写使用 python 语言，计算机硬件配置为英特尔酷睿 i5-4590，3.3Hz，8.00GB 内存，操作系统 Ubuntu14.04。

3.1 数据来源

本文数据集包括钓鱼网站 URL 集合和合法网站 URL 集合。钓鱼网站 URL 集合来源于 PhishTank 总计 30721 条，合法网站 URL 集合来源于爬虫抓取总计 12065 条。

3.2 实验及结果分析

本文对上述三种算法进行三组测试，每组测试数据集大小为 20000 条数据，第一组为合法网站 URL 为 5000 条，钓鱼网站 URL 为 15000 条；第二组为合法网站 URL 和钓鱼网站 URL 各 10000 条，第三组为合法网站 URL 为 15000 条，钓鱼网站为 5000 条。

将基于随机森林的文本特征算法与 K-近邻(K-NN)算法和支持向量机(SVM)算法进行比较，如图 4 所示。基于随机森林的判别算法仍有 5% 左右的错误率，实验发现譬如如下钓鱼网站 URL：http://www.zemo.org.ua/wp-includes/js/images/start.php 和 http://qus.sitey.me/等，从这两个 URL 文本可以看出如下特征：(1) URL 长度并没有超长；(2) 不包含或模糊匹配敏感特征词；(3) 没有“@”等特殊标识符；(4) URL 和 DNS 路径级数没有特别多；(5) 没有超长连续数字字符串等。即在本文的特征提取的特征中大多数特征均未出现，因此随机森林将该钓鱼网站 URL 定义为合法网站。

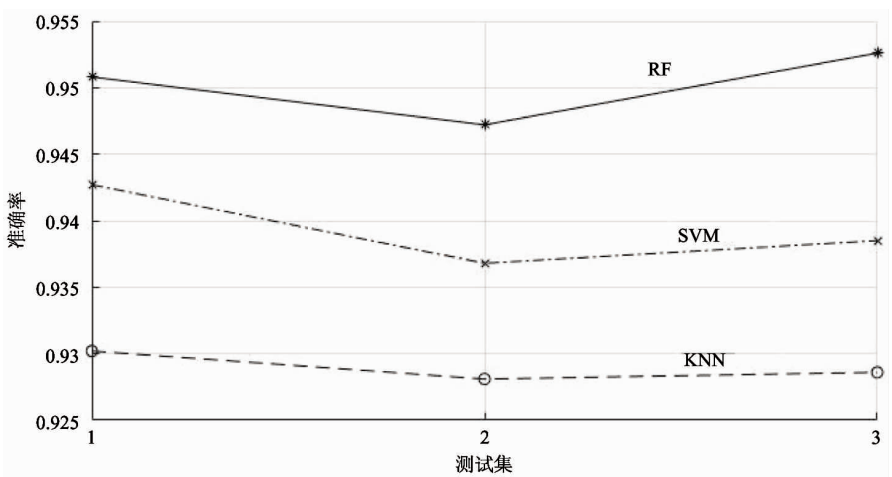

图 4 文本特征算法比较

将 MFRG 算法与基于语义链接分析法和基于页面内容检则方法(CANTINA)进行比较。语义链接分析法是基于可疑网页的架构和语义链接网络的推理,得到钓鱼网页与其目标网页之间的关系来检测钓鱼。CANTINA 是基于网页内容,以词频-逆文件频率(TF-IDF)信息检索为基础的钓鱼检测算法。

如图 5 所示,MFRG 算法虽然已经有 94% 的准确率,但仍有 6% 的错误率,该方法对于三种情况的结果并不理想:

(1) 钓鱼网页的相关网页中并不包含它的目标网页,导致误判成合法网页;

(2) 钓鱼网页与它潜在目标网页群中的网页的相关性强于与它的目标网页之间的相关性,导致误判成合法网页;

(3) 合法网页内链接较少的,导致误判成钓鱼网页。

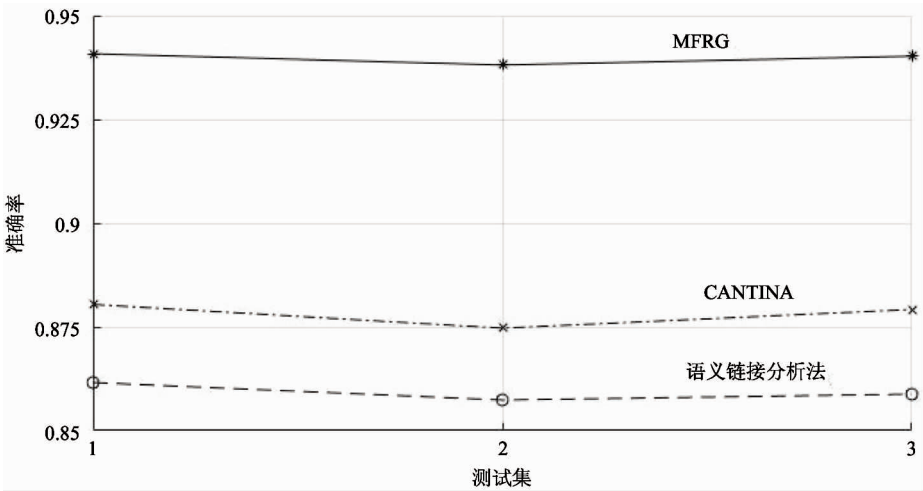

图 5　链接关系算法比较

本文上述的两个算法准确率有待加强,因此将两种算法通过 Bagging 算法进行融合,以达到更高的准确率。在 Bagging 算法中,S 取值不同与钓鱼网站判别准确率和计算消耗相关。随着 S 的增大,则判别准确率和计算消耗增高;随着 S 的减小,判别准确率和计算消耗减少。

经对 Bagging 算法的 S 进行测试,如图 6 所示。在 $S = 3$ 时准确率已经到达99.2%,但后续 S 的增加,准确率增加并不明显。因此本文选择 $S = 3$ 的情况下进行 Bagging 的融合算法。

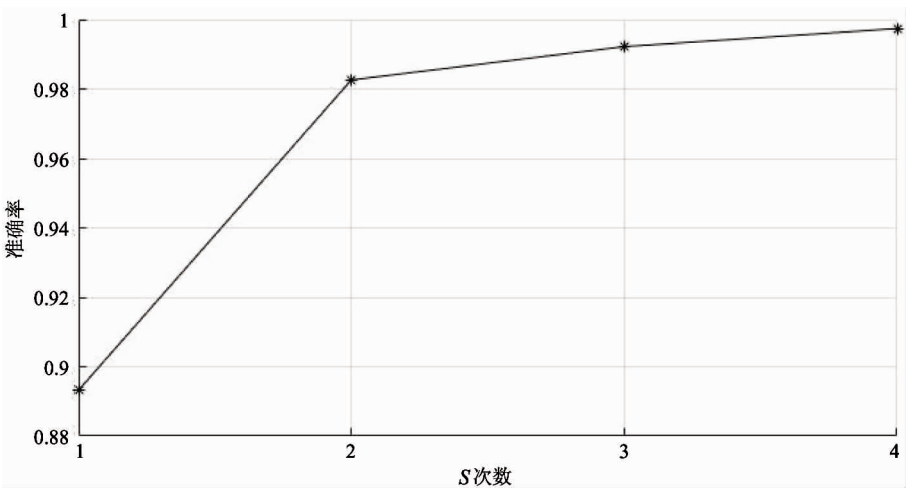

图 6　Bagging 算法 S 的选择

三种算法比较如图 7 所示,在三个测试数据集中基于 Bagging 融合算法的测试结果准确率均达到 99% 以上。

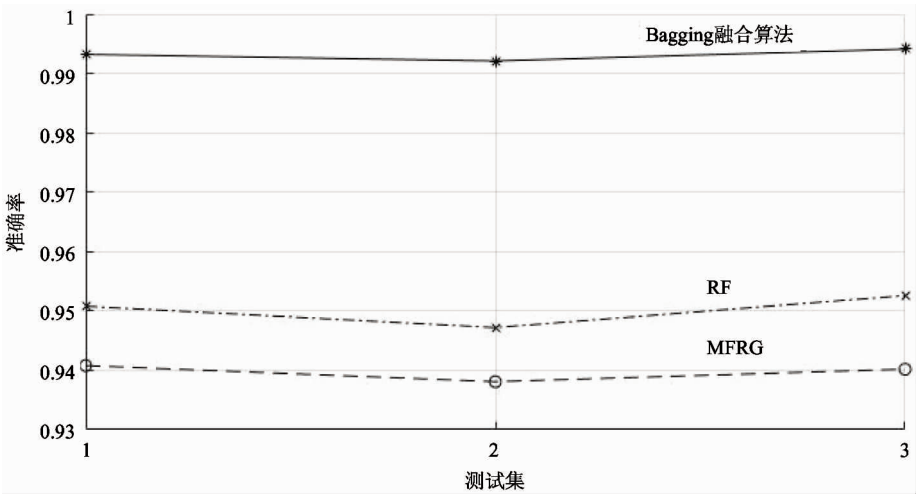

图 7 三种算法比较

检测钓鱼的评价指标形式多样,如表 2 所示,是钓鱼检测常用的四种指标。

表 2 钓鱼检测评价指标

	解释
TP	正确地把钓鱼网页检测为钓鱼网页
FP	错误地把合法网页检测为钓鱼网页
TN	正确地把合法网页检测为合法网页
FN	错误地把钓鱼网页检测为合法网页

本文主要用到的钓鱼检测的查准率 Precision 与查全率 Recall,它们的定义如下:

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

表 3 给出了不同检测方法的性能比较,从中可以看出,基于 RF 和 MFRG 的 Bagging 算法已经达到较高的准确率,但是仍有将近 1% 的错误率,因本文是通过 Bagging 算法进行结果融合,即 S 个随机森林算法和 S 个 MFRG 算法,只要 $2S$ 个判定结果中有大于 S 个判别结果为错误,则可导致最终的判断结果错误。

表 3 不同检测方法的性能比较

	FP (%)	Precision (%)	Recall (%)
基于 Bagging 的融合算法	1.2	98.7	99.3
MFRG	6.7	93.6	94.4
RF	5.3	94.9	95.5

如:http://ty42343ff. at. ua/confirmation/index.html,对上述 URL 进行分析:

在文本特征方面:该钓鱼网站 URL 在随机森林的判定结果为合法网站。因本文选取的诸多特征在该 URL 均没有出现,因此随机森林并不能判定正确,则有 S 个错误判定结果。

在链接关系方面:该钓鱼网页的相关网页中并不包含它的目标网页,导致误判为合法网页。因此 MFRG 的 S 个判定结果中有小于 S 个误判结果。

但最终误判结果总和大于 S 个,因此该钓鱼网站 URL 被误判为合法网站。

4 结 论

本文提出了一种基于 URL 文本特征及链接关系的钓鱼网站判别算法,通过 Bagging 融合算法将基于随机森林的 URL 文本特征钓鱼网站辨别算法

和基于链接关系的钓鱼网站辨别算法进行融合,以得到更高的钓鱼网站判别准确率。测试结果表明本文提出的算法准确率达到99.2%,对单一判别算法准确率提高4% ~5%左右。

参考文献

[1] 赵加林. 基于 K-Means 和 SVM 的钓鱼网站识别的研究:[硕士学位论文]. 成都:西南交通大学信息科学与技术学院, 2016. 21-32

[2] 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

[3] 朱百禄. 基于 Web 社区的钓鱼网站检测研究:[硕士学位论文]. 天津:天津理工大学计算机与通信工程学院, 2013. 13-33

[4] 滕雯静. 基于链接分析的钓鱼网站检测方法:[硕士学位论文]. 南京:南京邮电大学计算机学院, 2014. 17-28

[5] 唐伟, 周志华. 基于 Bagging 的选择性聚类集成. 软件学报, 2005, 16(4):496-502

[6] 谢元澄, 杨静宇. 删除最差基学习器来层次修剪 Bagging 集成. 计算机研究与发展, 2009, 46(2):261-267

[7] 刘韶涛, 李洪胜. 融合链接结构的主题爬虫算法. 华侨大学学报(自然科学版), 2017, 38(2):195-200

[8] Cao J X, Dong D, Mao B, et al. Phishing detection method based on URL features. *Journal of Southeast University(English Edition)*, 2013, 29(2):134-138

[9] 杨宏宇, 徐晋. 基于改进随机森林算法的 Android 恶意软件检测. 通信学报, 2017, 38(4):8-16

[10] Kleinberg J M. Authoritative sources in a hyperlinked environment. In: Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, San Francisco, USA, 1998. 668-677

[11] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. In: Proceedings of the 7th International Conference on World Wide Web, Brisbane, Australia, 1998. 107-117

[12] 谢月. 网页排序中 PageRank 算法和 HITS 算法的研究:[硕士学位论文]. 成都: 电子科技大学数学科学学院, 2012. 9-46

[13] Dean J, Henzinger M R. Finding related pages in the World Wide Web. *Computer networks*, 1999, 31(11): 1467-1479

[14] 赵礼峰, 纪亚宝. 最大流最小截问题的遗传算法研究. 计算机技术与发展, 2017, 27(4): 69-72

A fishing website identification algorithm based on URL text feature and link relation

Zhao Dunyu, Zhang Zhaoxin

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract

Based on the analysis of the uniform resource location (URL) text data of fishing sites and the characteristics of the network topology composed of fishing websites, a fishing site recognition algorithm based on URL text features and link relation (FAUFL) is proposed to improve the accuracy rate of fishing site recognition. The principle of the algorithm is as below: By using URL text features as input, the random forest algorithm is used to generate the fishing site discrimination algorithm based on URL text features. The related web page group is constructed by using the link relation as input, and the related web page algorithm based on the maximum flow cutting is used to generate the fishing website based on the link discriminant algorithm. By taking the above two kinds of discriminant algorithms' results as input, the further evaluation is conducted by using the Bagging algorithm. The test results show that the accuracy rate of the FAUFL is 99.2%, which is 3.9% higher than that of the URL text feature-based algorithm, and 5.0% higher than that of the link-based algorithm.

Key words: fishing website, fusion algorithm, uniform resource location (URL), text feature, link relation