doi:10.11959/j.issn.1000-436x.2016180

在线社交网络中 Spam 相册检测方案

吕少卿[1]，张玉清[1,2]，刘东航[1]，张光华[1,3]

（1. 西安电子科技大学综合业务网理论及关键技术国家重点实验室，陕西 西安 710071；
2. 中国科学院大学国家计算机网络入侵防范中心，北京 100190；
3. 中国科学院信息工程研究所物联网信息安全技术北京市重点实验室，北京 100097）

摘 要：提出一种针对 Spam 相册的检测方案。首先分析了 Photo Spam 的攻击特点以及与传统 Spam 的差异，在此基础上构造了 12 个提取及时且计算高效的特征。利用这些特征提出了有监督学习的检测模型，通过 2 356 个相册的训练形成 Spam 相册分类器，实验表明能够正确检测到测试集中 100%的 Spam 相册和 98.2%的正常相册。最后将训练后的模型应用到包含 315 115 个相册的真实数据集中，检测到 89 163 个 Spam 相册，正确率达到 97.2%。

关键词：社交网络安全；Photo Spam；Spam 检测；人人网

中图分类号：TP393 **文献标识码**：A

Detecting Spam albums in online social network

LYU Shao-qing[1], ZHANG Yu-qing[1,2], LIU Dong-hang[1], ZHANG Guang-hua[1,3]

(1. Information Security Research Center of State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China;
2. National Computer Network Intrusion Protection Center, University of Chinese Academy of Sciences, Beijing 100190, China;
3. Beijing Key Laboratory of IOT Information Security Technology, Institute of Information Engineering, CAS, Beijing 100097, China)

Abstract: A supervised learning solution to detect Spam albums instead of spammers in Photo Spam was proposed. Specifically, the characteristics of Photo Spam and the differences between Photo Spam and traditional Spam were analyzed. Then 12 features which were extracted easily and calculated efficiently were constructed based on the analysis. Next a classification model was built with a dataset of 2 356 labeled albums to identify Spam albums. The model provided excellent performance with true positive rates of Spam albums and normal albums, reaching 100% and 98.2% respectively. Finally, the detection model were applied to 315 115 unlabeled albums and detected 89 163 spam albums with a true positive rate of 97.2%.

Key words: social network security, Photo Spam, Spam detection, RenRen

1 引言

随着在线社交网络（以下简称为社交网络）的快速发展，人们的生活越来越依赖于社交网络，人们每天都在社交网络中查看好友最近的经历、分享自己的新鲜事以及获取最新资讯等。

社交网络给人们日常生活带来便利的同时，也带来了新的安全问题[1]。Spam 攻击就是其中最主要的安全威胁[2,3]。攻击者利用社交网络提供的各项功能来进行 Spam 攻击，如利用微博（状态）[4,5]、#[6]、@[7]和评论[8]等功能来传播广告、色情以及钓鱼等恶意信息，严重威胁到用户的个人信息安全以及社交网络的信誉体系。针对这些问题，学术界和工业界都提出了大量的方案来检测或防御 Spam 攻击[9~13]。但在利益的驱动之下，攻击者又提出了新的攻击方式来躲避或绕过这些检测和防御方案，

收稿日期：2016-03-14；**修回日期**：2016-06-14

基金项目：国家自然科学基金资助项目（No.61572460, No.61272481, No.61303239）；物联网信息安全技术北京市重点实验室开放课题基金资助项目；中国博士后科学基金资助项目（No.2015M582622）

Foundation Items: The National Natural Science Foundation of China (No.61572460, No.61272481, No.61303239), Open Fund of Beijing Key Laboratory of IOT Information Security Technology, China Postdoctoral Science Foundation (No.2015M582622)

Photo Spam 就是其中之一。Photo Spam 是攻击者利用社交网络中照片的分享功能来存储和传播 Spam 信息的新式 Spam 攻击。

Photo Spam 最早是 Wang 等[14]提出，他们发现人人网中的攻击者通过上传 Spam 照片，并在其中嵌入 Spam 信息来推广衣服、鞋子等物品。之后 Cao 等[15]发现在 Facebook 中攻击者通过短时间内上传大量照片来推销减肥药。同时这些工作针对 Photo Spam 提出了相应的检测方案，如 Wang 等[15]利用用户的鼠标点击模型来检测 Spam 账号，Cao 等[15]利用账号的行为模式和登录 IP 地址来检测 Spam 账号。这些工作虽然能够检测到一定的 Photo Spam 账号，但是这些方案都是针对 Spam 账号进行检测，而不是 Spam 相册本身。针对 Spam 账号进行检测的方案需要基于账号长期的行为数据，即需要在 Spam 账号执行大量的恶意行为操作之后才能够判断为 Spam 账号，而此时 Spam 账号的行为已经对正常用户造成了损害。而针对 Spam 相册的检测方案能够在用户上传照片之后就对相册进行检测，能够在攻击者传播 Spam 信息之前就检测到 Spam 相册，从而阻止 Spam 相册对正常用户的影响。

本文通过分析 Photo Spam 的特征提出了一种针对 Spam 相册的检测方案。首先，详细分析了 Photo Spam 的攻击策略，并对比了 Photo Spam 与传统 Spam 攻击的区别；基于此分析构造了 12 个提取及时且计算高效的特征，并对比了 Spam 相册与正常相册在这些特征方面的不同。之后利用这些特征设计了基于有监督学习的检测方案并通过包含 2 356 个相册的数据集进行训练，形成了针对 Spam 相册的分类器。实验表明本文的分类器能够正确区分测试集中 100%的 Spam 相册以及 98.2%的正常相册。随后本文分析了分类器在不同的训练样本比例、不同的分类算法以及不同特征子集情况下的分类结果。最后本文将训练后的分类器应用到包含有 315 115 个未标识相册的真实数据集中，检测到 89 163 个 Spam 相册，人工验证正确率为 97.2%。

虽然本文的实验只是针对人人网，但是本文的特征和模型只需要微小调整就能适用于所有的社交网络（如 Facebook 等）。

2　Photo Spam

Photo Spam 是一种新的 Spam 攻击方式。在 Photo Spam 中攻击者首先利用控制的账号（攻击者创建的虚假账号或盗用的账号）在社交网络中创建一个相册，然后在相册中上传 Spam 照片，并在照片的描述中加入恶意信息，如广告、钓鱼等 URL，最后攻击者通过分享该相册或其中的照片进行传播。攻击者为了吸引更多的用户，会用有吸引性的名称来命名相册，而且为了提高相册中 Spam 照片的访问量以及躲避检测，攻击者会在 Spam 相册中加入一些正常照片来伪装成正常相册。

具体的攻击实例如下。攻击者创建了一个标题为“世界 100 位用演技诠释电影的男人”的相册。在该相册中除了与相册名称符合的照片之外，还包含有 Spam 照片，这些 Spam 照片的描述中包含指向出售照片中商品的超链接。攻击者还可以通过对该相册进行分享来传播 Spam 信息，也可以用控制的多个账号分享该相册。任何正常用户如果被该分享的标题所吸引并访问该相册，就会在浏览的过程中不断接收到 Spam 照片以及 Spam 信息。这将会严重损害用户对社交网络的使用体验。

与传统 Spam 攻击相比，Photo Spam 具有以下特点。首先，Photo Spam 攻击中恶意信息的载体更加丰富。传统 Spam 攻击中攻击者将微博、评论、私信等作为恶意信息载体，每次传播都是将恶意信息文本进行复制。而在 Photo Spam 攻击中，攻击者可以将恶意信息存储在照片的描述、评论以及标题中，甚至可以作为水印直接嵌入到照片中。这种恶意信息的多种存在形式让 Photo Spam 更难被检测到。其次，Photo Spam 与传统 Spam 攻击在恶意信息的传播方式上也有区别。Photo Spam 进行传播的主要方式是分享，而传统 Spam 进行传播主要是将恶意信息进行多次重复发送。分享是一种有效的传播手段，因为分享只产生一个正常的链接，并不产生更多的恶意信息，而多次发送的方式每次都复制恶意信息，所以更容易被检测到，因此这种传播方式的不同使 Photo Spam 攻击更难被检测到。再次，在 Photo Spam 中恶意信息的存储与传播可以分离。由于 Photo Spam 的传播方式主要是分享，因此，攻击者可以利用新创建的账号来上传 Spam 照片，然后利用拥有更多好友的 Spam 账号来分享 Spam 相册，这样能够有效避免一些拥有更多好友的 Spam 账号被检测到，降低了攻击者的损失并使 Photo Spam 威胁到更多正常用户。最后，在 Photo Spam 中正常用户可能会无意参与到 Spam 信息的传播。

攻击者为了增加 Spam 照片的访问量会在 Spam 相册中混杂正常照片，正常用户可能会分享这些正常照片而没有察觉到在该相册中存在 Spam 内容，这样更加扩大了 Photo Spam 的危害范围。

通过以上分析，本文发现与传统 Spam 攻击相比，Photo Spam 对正常用户的危害更大而且更难被检测到。而 Spam 相册是 Photo Spam 整个攻击过程的核心，所有恶意信息的存储与传播都是基于 Spam 相册。因此，针对 Spam 相册的检测方案能够有效降低 Photo Spam 攻击对正常用户的影响，这就是本文工作的主要出发点。

3 数据获取和样本标识

本文的实验数据获取自人人网。人人网是中国最大的社交网络之一，2015 年，人人网的注册用户数已经达到 2.26 亿，月活跃用户数为 4 600 万。人人网与国外的 Facebook 具有类似的功能和特性，它们都属于实名制注册网站，用户在其中能够创建个人账号、填写个人信息、与其他用户建立双向好友关系，并能够发布自己的新鲜事、创建相册、上传照片，以及将各种有趣的事情分享到自己的新鲜事等。

本文获取数据的流程如图 1 所示。为了有效地获取数据并降低可能的取样误差[16,17]，本文首先在人人网的用户 *ID* 范围（1000~808834939）中随机抽取 10 000 个 *ID*，然后利用应用程序接口（API, application program interface）获取这些 *ID* 的个人信息。在这些 *ID* 中共有 4 875 个 *ID* 能够返回正确信息。然后通过 API 对这些账号 *ID* 获取相应的相册 *ID*。为了使样本具有代表性，本文没有获取网站自动创建的头像相册和应用相册，只获取用户自己创建的普通相册。同时为了保证样本的多样性，对于有多个普通相册的账号，本文随机选取其中 3 个相册 *ID*。通过用户 *ID* 和相册 *ID* 共同获取这些相册的具体信息以及相册中照片的绝对 URL 地址。最后通过定制的爬虫获取这些相册中所有的照片。本文数据获取的时间是从 2014 年 11 月至 2015 年 2 月。

接下来需要确定这些相册是否为正常相册或 Spam 相册。邀请 3 位志愿者手动确定这些相册的分类，这 3 位志愿者使用人人网或其他社交网络超过 3 年。3 位志愿者通过访问这些相册并浏览照片的内容和照片的信息来确定相册的分类。如果在相册中包含有恶意 URL，如指向广告、钓鱼、色情等网站，那么这个相册就被判断为 Spam 相册，否则为正常相册。每个志愿者都需要判断所有的相册，然后根据 3 位志愿者的综合评价来最终确定一个相册的分类。3 位志愿者如果对同一个相册的分类一致，那么就采用这个分类，如果对于一个相册的分类有分歧，那么就舍弃这一相册。最终共获得了 1 178 个 Spam 相册以及 5 693 个正常相册。

为了降低数据集中正常相册和 Spam 相册比例不同所带来的偏差，本文从 5 693 个正常相册中抽取了与 Spam 相册数相同的 1 178 个正常相册。这 1 178 个正常相册与 1 178 个 Spam 相册组成了本文的数据集 I。本文分别随机抽取了数据集 I 中 70%的正常相册和 70%的 Spam 相册作为模型的训练集，剩余 30%的数据作为测试集。数据集 I 的具体情况如表 1 所示。

表 1 数据集 I 汇总

相册	训练集	测试集	总数
Spam	825	353	1 178
正常	825	353	1 178

4 特征分析

根据对 Photo Spam 的分析，Spam 相册与正常相册存在多种差异。首先，相册使用目的不同，攻击者创建相册的目的是吸引更多用户的访问来更广泛地传播 Spam 信息，而正常用户创建相册的目的是与好友分享照片。其次，相册中照片性质不同，

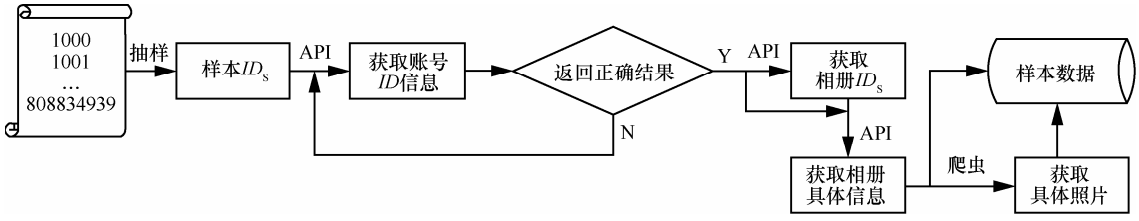


图 1 数据获取流程

正常相册中的照片一般是通过相机、手机等设备对自然景象的记录，而 Spam 照片大都是攻击者通过计算机软件人工合成。最后，相册后续操作不同，攻击者通常都是通过自动化工具对账号进行操作，因此在相册创建完成后很少有后续操作，而正常用户创建相册后会不定时上传照片。因此 Spam 相册和正常相册本身存在差异。

基于以上分析，本文设计了相关特征。这些特征可以分为 2 类，一类为相册相关的特征；一类为照片内容相关的特征。

相册相关特征主要是抽取与相册本身属性相关的信息，包含以下内容。

1) 相册标题长度：此特征是相册标题的字符长度。

2) 相册间隔时间：相册间隔时间是指从相册创建时间到相册最近更新时间的间隔天数。

3) 相册中照片数：即相册中包含的照片数。

4) 照片描述比率：即在相册中照片描述不为空的照片个数与相册中照片数的比值。

5) 照片描述 URL 比率：即照片描述中包含 URL 的照片数与相册中照片总数的比值。

6) 照片描述长度均值：相册中照片描述字符长度的均值。

7) 照片描述长度方差：相册中照片描述字符长度的方差。

照片内容相关特征是抽取相册中照片的内容相关属性，本文主要考虑以下特征。

1) 颜色均值：颜色均值包含 3 个值，即相册中所有照片红、绿、蓝三原色的均值。单张照片颜色均值中红色定义如式(1)所示，其中，L 和 W 分别为照片的长和宽，单位为像素，R_{jk} 为像素点(j, k)中红色的值。整个相册的红色颜色均值定义如式(2)所示，其中，N 为相册中照片数。绿色和蓝色的计算方式与红色类似。

$$C_R = \frac{\sum\limits_{\substack{0<j\leqslant L\\ 0<k\leqslant W}} R_{jk}}{LW} \tag{1}$$

$$F_R = \frac{\sum\limits_{0<i\leqslant N} C_R^i}{N} \tag{2}$$

2) 色饱和度均值：饱和度是指色彩的鲜艳程度。在 RGB 彩色图像中可以通过计算图片像素点中 $\max(R,G,B)$与 $\min(R,G,B)$的差值大于阈值 T 的像素点的数量与整体像素点数量的比值来表示饱和度[18]，如式(3)所示。单张照片饱和度 S 定义如式(4)所示，在实验中本文选取 T 为 50[19]。整个相册的饱和度均值特征 F_S 定义如式(5)所示，S^i 为相册中第 i 张照片的饱和度。

$$Q = \begin{cases} 1 & ,\max(R,G,B)-\min(R,G,B)\geqslant T \\ 0 & ,\max(R,G,B)-\min(R,G,B)<T \end{cases} \tag{3}$$

$$S = \frac{\sum\limits_{\substack{0<j\leqslant L\\ 0<k\leqslant W}} Q_{jk}}{LW} \tag{4}$$

$$F_S = \frac{\sum\limits_{0<i\leqslant N} S^i}{N} \tag{5}$$

3) 主色覆盖范围均值：主色就是照片中出现频率较高的颜色值。在实验中本文将出现频率最高的前 20 个颜色定义为主色[19]。主色覆盖范围主要是计算照片中出现频率最高的颜色所占的比例[20]，它可以捕捉到照片中大片的色块。由于 Spam 照片是通过软件创建，因此一般有大块的同色背景。如式(6)所示，其中，单张照片的主色覆盖范围 U 定义为式(7)，整个相册的主色覆盖范围均值 F_U 通过式(8)计算得到，其中，U^i 为相册中第 i 张照片的主色覆盖范围。

$$P = \begin{cases} 1 & ,(R,G,B)\in\{\text{主色}\} \\ 0 & ,(R,G,B)\notin\{\text{主色}\} \end{cases} \tag{6}$$

$$U = \frac{\sum\limits_{\substack{0<j\leqslant L\\ 0<k\leqslant W}} P_{jk}}{LW} \tag{7}$$

$$F_U = \frac{\sum\limits_{0<i\leqslant N} U^i}{N} \tag{8}$$

本文检测方案的目的是尽早地检测到攻击者上传的 Spam 相册，因此所设计的特征需要在照片上传之后就能够提取，所以舍弃了一些有良好的区分度，但需要经过一段时间才能够形成的特征的照片，如照片的访问次数以及评论数等，并且社交网络每一时刻都有大量照片上传，因此本文需要计算高效的特征，所以也舍弃了一些需要很大计算量的特征，如照片的角点、纹理等特征。

图 2 展示了 Spam 相册与正常相册在 4 个特征上的不同。为了更好地展示，本文从数据集 I 中随机抽取了 500 个 Spam 相册以及 500 个正常相册，并分别编号为 1~500，然后对这些相册的相关特征

进行了展示。

从图 2(a)中可以看出，Spam 相册的标题字符长度大于正常相册的标题长度。87.6%的 Spam 相册的标题长度大于 45 个字符，而正常相册中只有 38.8%的相册标题长度大于 45 个字符。这是因为攻击者为了吸引更多用户访问相册，会将相册命名为吸引用户的标题，因此标题长度比正常相册长。从图 2(b)中可以看出正常相册的间隔时间大于 Spam 相册。79.6%的正常相册的间隔时间超过 1 天，而 Spam 相册中只有 20.8%的相册间隔时间超过 1 天。这是由于攻击者通过自动化工具操作账号，在上传照片后基本没有后续操作，所以相册的间隔时间小于正常相册。图 2(c)展示了 Spam 相册与正常相册在照片描述长度方面的差别。82.6%的 Spam 相册的照片描述长度均值超过 90 个字符，而在正常相册中只有 42.6%的照片描述长度均值大于 90 个字符。这是因为攻击者需要在照片描述中添加 Spam 信息，因此照片描述的长度大于正常相册中照片描述。图 2(d)展示了 Spam 相册与正常相册在照片颜色均值中红色的差异。其中，Spam 相册中红色均值大于 150 的占 78.2%，而正常相册中只有 22.4%的照片颜色均值中红色大于 150。这是由于攻击者人工合成的 Spam 照片整体颜色更加亮丽，而且为了吸引更多正常用户，其所添加的正常照片也色度较高，因此 Spam 相册中照片的颜色均值更高。

从图 2 中可以看出，正常相册与 Spam 相册在这些特征上都具有差异，因此能够利用这些精心设计的特征来区分正常相册和 Spam 相册。

5　Spam 相册检测模型及实验结果

根据数据集 I 和所提出的特征，本文设计了基于有监督学习的 Spam 相册检测模型。有监督学习是机器学习算法中的一类，它根据标识过的训练集来训练模型[21]。训练集中包含有多条训练样本，每条训练样本都包括多个输入特征以及一个结果标识。通过对训练集的学习，有监督学习能够形成相应的分类器来对新的样本进行分类。

5.1　检测模型

本文所提出的 Spam 相册检测模型如图 3 所示。首先利用社交网络提供的应用程序接口 API 从社交网络中获取原始数据。然后通过样本标识对原始数据中的样本进行分类。接着利用特征抽取从标识后的样本中提取出相应的特征。这些特征样本集被随机分为训练集和测试集。通过训练集对分类器进行训练，最后通过测试集对训练后的分类器进行测

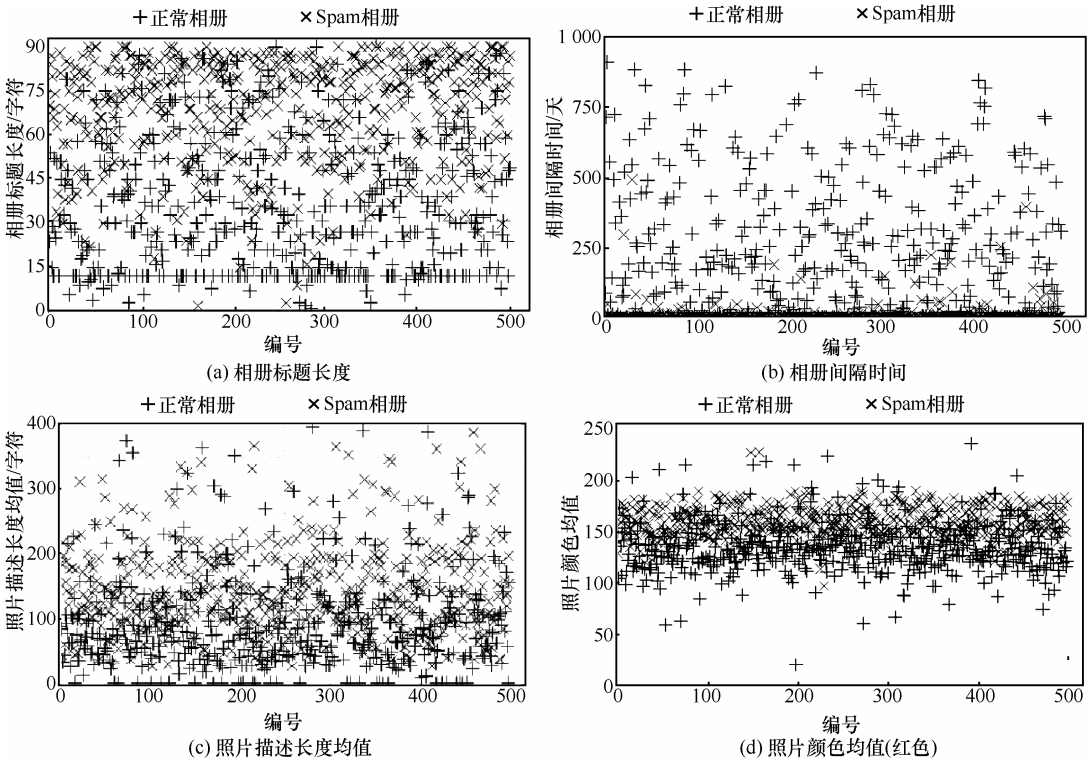


图 2　Spam 相册和正常相册在不同特征的区别

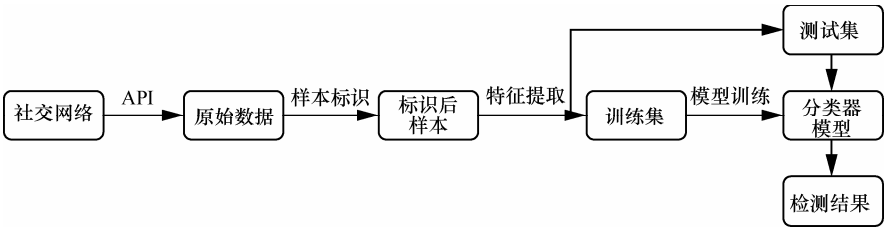


图 3　Spam 相册检测模型

试，获取相应的检测结果。

通过第 4 节的分析，Spam 相册和正常相册拥有不同的特征，本文采用第 4 节所设计的 12 个特征来区分正常相册和 Spam 相册。

5.2　SVM 分类器

本文主要采用支持向量机(SVM，support vector machine)[22]作为分类器，因为 SVM 在之前的工作中取得了很好的检测结果[9,10,14]。SVM 的最终目标是找到一个超平面能够以最大间隔将训练集中 N 维数据分为 2 类。在实验中本文采用非线性的 RBF (radial basis function)核来使 SVM 获得更好的效果。最终的算法采用 R 语言中 e1071 程序包的相关函数来实现。采用 RBF 核的 SVM 算法有 2 个训练参数：C 和 *gamma*。其中，C 控制模型的过载，*gamma* 控制模型的非线性度。本文采用 e1071 程序包所提供的参数选择函数来自动选取参数。该函数采用网格搜索策略，通过不同的 C 和 *gamma* 对来计算分类结果，然后选择获得最高准确率的参数对。最后，针对本文的训练集，获得的最优 C 和 *gamma* 值分别为 215.5 和 0.075。

为了提高模型训练结果的可信度，本文在训练时采用十折交叉验证法。十折交叉验证是将训练集平均分为 10 份，对其中 9 份进行训练，另一份做测试，循环进行 10 次实验。最后的结果是这 10 次实验结果的平均。

5.3　评价指标

为了评价本文的检测结果，本文采用如表 2 所示的混淆矩阵，其中，a 表示原本是 Spam 相册被预测为 Spam 相册的个数；b 表示原本是 Spam 相册但被预测为正常相册的个数；c 表示原本是正常相册但被预测为 Spam 相册的个数；d 表示原本是正常相册且被正确预测的个数。

根据表 2 的混淆矩阵，本文同时采用机器学习中的评价指标即准确率、召回率和 F_1 值。准确率(P)是指对于指定的一类，正确分类的个数与被分为这类的总体个数的比率。如在表 2 中 Spam 相册的准确率可以表示为 $P=\frac{a}{a+c}$。召回率(R)是指一类数据被正确分类的数量除以真实的属于这一类的数量。如表 2 中 Spam 相册的召回率即为 $R=\frac{a}{a+b}$。F_1 值是准确率与召回率的综合。表 2 中 Spam 相册 F_1 值的计算式为 $F=\frac{2PR}{P+R}$。

表 2　　**混淆矩阵示例**

真实值	预测值	
	Spam	正常
Spam	*a*	*b*
正常	*c*	*d*

5.4　Spam 相册与正常相册的比率

起初本文使用训练集中所有的样本对分类器进行训练，并且取得了 Spam 相册与正常相册的 F_1 值分别为 0.982 和 0.988，但这个结果不一定是最优结果。为了获得最优结果，本文对训练集中的样本采用下采样[23]的方式使 Spam 相册与正常相册的比例分别从 1:1 到 1:10，并且重复这个过程 10 次。在图 4 中本文展示了每次不同比例采样时相应的宏 F_1 值。宏 F_1 值是 2 类样本 F_1 值的算术平均值。从图 4 中可以看出，虽然每次同一比例采样的结果并不相同，但在整体趋势上，随着 Spam 相册与正常相册比例的降低，宏 F_1 值也在降低。其中，最优值是在第 7 次的 1:3 采样中获得，最大的宏 F_1 值为 0.991。因此，在之后的实验中，本文采用的训练集即为获得最大宏 F_1 值的训练集。

5.5　分类结果和比较

本文采用 SVM 算法获得的最优结果如表 3 所示。可以看出，本文的检测模型非常有效，能够正确区分测试集中 100%的 Spam 相册以及 98.2%正常相册，只有小部分的正常相册分类错误。表 4 中列出了准确率、召回率、F_1 的值。可以看出 Spam 相册和正常相册的 F_1 都高达 0.991。

进一步，本文比较了不同分类算法的检测结果。如表 5 所示。本文比较了 Naïve Bayes、Decision Tree、Random Forest 和 Neural Network 分类算法。这些算法分别采用 R 语言中程序包 e1071[24]、rpart[25]、randomForest[26]和 nnet[27]实现。对于每一种分类算法，本文都采用相同的评价指标（准确率、召回率、F_1 值）分别计算 Spam 相册和正常相册。可以看出，SVM 算法获得了最高 F_1 值，同时其他算法也都能获得较好的分类结果，这主要是因为本文根据 Photo Spam 的特点设计了合适的特征。

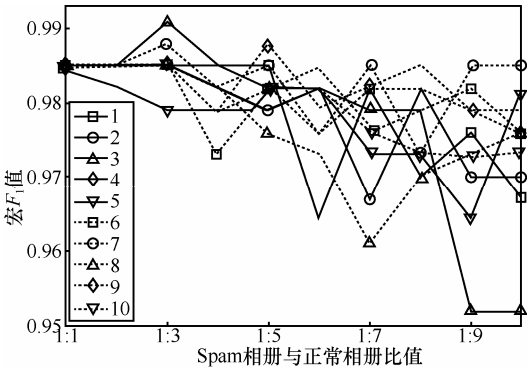


图 4 不同样本比例的分类结果

表 3 混淆矩阵

真实值	预测值	
	Spam 相册	正常相册
Spam 相册	100.0%	0.0%
正常相册	1.8%	98.2%

表 4 分类结果评价

相册	准确率	召回率	F_1 值
Spam	0.982	1.000	0.991
正常	1.000	0.982	0.991

5.6 特征的重要性

为了分析本文所设计的特征对检测结果的重要性，本文分别用卡方检验(Chi-square test)和信息增益(information gain)计算特征的重要性。表 6 列出了 2 种计算方式中对检测结果贡献最大的 10 个特征。其中，对检测结果贡献率最大的 2 个特征分别为照片描述 URL 比率和相册中照片数。照片描述 URL 比率说明 Spam 账号会在照片描述中嵌入恶意 URL 来存储 Spam 信息，与 Wang 等[14]的发现相符合。相册中照片数说明攻击者为了吸引更多用户会在相册中上传大量照片，与 Gao 等的发现相符合[15]。

表 6 特征贡献排名 TOP10

排名	卡方检验	信息增益
1	照片描述 URL 比率	照片描述 URL 比率
2	相册中照片数	相册中照片数
3	颜色均值—红色	颜色均值—红色
4	颜色均值—绿色	颜色均值—绿色
5	颜色均值—蓝色	颜色均值—蓝色
6	相册存在时间	相册存在时间
7	色饱和度均值	色饱和度均值
8	主色覆盖范围均值	主色覆盖范围均值
9	照片描述比率	相册标题长度
10	相册标题长度	照片描述比率

为了进一步分析这 10 个特征对检测结果的影响，本文将这 10 个特征按卡方验证结果的顺序每次从所有特征中移除一个特征，然后计算检测结果。图 5 所示为计算了不同特征子集的检测结果。可以看出，每次移除一个特征都会造成宏 F_1 值的微小降低。而影响最大是移除特征照片描述 URL 比率，不过即使在最差情况下，本文的分类器也能够获得很好的分类结果，Spam 相册和正常相册的 F_1 值分别达到 0.977 和 0.979。

5.7 在真实环境中评价

为了验证本文所设计的特征集和所提出算法的有效性，本文将训练好的分类器应用到一个真

表 5 SVM 与其他分类算法比较

分类器	准确率		召回率		F_1 值	
	Spam 相册	正常相册	Spam 相册	正常相册	Spam 相册	正常相册
SVM	1.000	0.982	1.000	0.982	0.991	0.991
Naïve Bayes	0.982	0.970	0.970	0.982	0.976	0.976
Decision Tree	0.988	0.988	0.988	0.988	0.988	0.988
Random Forest	0.982	0.994	0.994	0.982	0.988	0.988
Neural Network	0.960	0.988	0.988	0.960	0.973	0.973

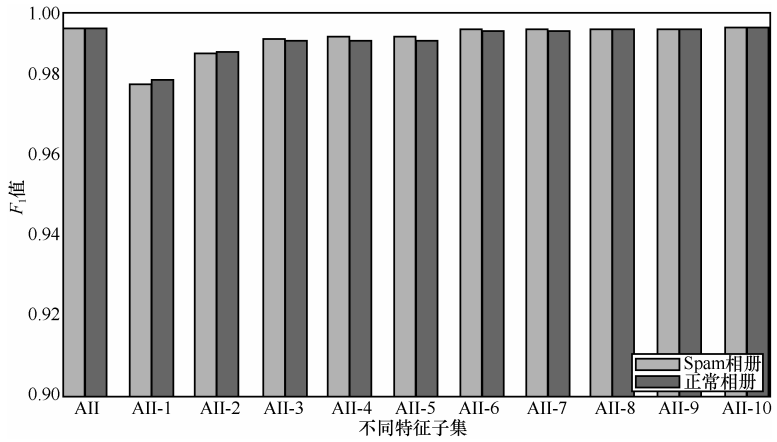


图 5　不同特征子集的分类结果 F_1 值

实的数据集中。为了区分，本文将其命名为数据集Ⅱ。数据集Ⅱ采用与数据集Ⅰ相同的获取方式。首先，本文从人人网的所有用户 *ID* 空间中随机抽取 100 000 个 *ID*，然后通过 API 来获取这些 *ID* 的个人信息和相册信息。本文过滤掉无效的 *ID*、已经被禁用的 *ID* 以及在数据集Ⅰ中已经包含的 *ID* 等，共获得了 49 156 个用户以及 315 115 个相册，将这 315 115 个相册作为数据集Ⅱ的样本，并抽取相应的特征。

本文利用已经训练好的检测模型对数据集Ⅱ进行检测。检测的结果如表 7 所示，其中，89 163 个相册被判断为 Spam 相册，剩余 225 952 个相册被判断为正常相册。这些 Spam 相册分别属于 5 242 个 Spam 账号。对所有相册进行人工验证需要耗费大量的时间，因此本文从判断为 Spam 相册的样本以及判断为正常相册的样本中分别抽取 1 000 个相册，然后让 3 位志愿者采用与第 3 节中相同的方法进行人工验证。结果如表 8 所示，其中，972 个 Spam 相册以及 969 个正常相册判断正确。这一结果更进一步说明本文的检测方案是有效的。

表 7　　**数据集Ⅱ汇总**

数据	值
总账号数	49 156
总相册数	315 115
Spam 用户	5 242
正常用户	43 914
Spam 相册	89 163
正常相册	225 952

表 8　　**数据集Ⅱ分类结果**

真实值	预测值	
	Spam 相册	正常相册
Spam 相册	97.2.%	2.8%
正常相册	3.1%	96.9%

6　相关工作

随着社交网络的快速发展，社交网络中 Spam 攻击也吸引了大量科研工作者的关注。因此有大量与 Spam 检测相关的工作。

在检测传统 Spam 攻击方面。Stringhini 等[11]利用机器学习和 6 个特征，如好友数、文本相似性等来检测 Facebook 和 Twitter 中发送 Spam 信息的账号。Benevenuto 等[9]利用 39 个基于内容的特征和 23 个用户行为特征来检测在 Twitter 热门事件中发布 Spam 信息的账号并针对 Spam 账号和正常账号分别获得了 0.79 和 0.912 的 F_1 值。Zheng 等[10]通过 SVM 分类器以及 18 个特征来检测新浪微博中发布 Spam 信息的账号。Lee 等[12]通过在 Twitter 中部署蜜罐来获取 Spam 账号信息并利用机器学习技术和设计的特征，如推文中@的比率、URL 比率等来检测 Spam 账号。与本文的工作类似，这些工作都是利用机器学习技术来检测 Spam 账号。但是这些工作都是针对传统 Spam 进行检测而且检测的对象是 Spam 账号，而本文的工作是检测 Photo Spam 中 Spam 相册。与传统 Spam 相比，Photo Spam 更难被检测到，本文针对 Photo Spam 设计了相应的特征，而且本文的工作能够获得更高的检测精度。

在检测 Photo Spam 方面，如前所述，Wang 等[14]

通过将鼠标点击模式相似的账号聚为一类来将账号分为不同的类别，即正常账号和 Spam 账号。通过这种方法检测到一类鼠标点击行为集中在照片分享的 Spam 账号，并发现这些照片中包含指向出售鞋子、衣服等购物网站的 URL。Cao 等[15]根据用户行为的相似性将用户聚为不同的类别来检测 Spam 账号，他们发现 Facebook 中一些 Spam 账号通过在短时间内上传大量照片来推广减肥药，并且这些账号共用少量 IP 地址。这 2 个工作都是针对 Photo Spam 中 Spam 账号进行检测，而本文的工作是针对 Photo Spam 中 Spam 相册进行检测。针对 Spam 账号进行检测需要在账号执行一定行为之后才能够将 Spam 账号检测到，而此时 Spam 账号的恶意行为已经对正常用户造成了威胁。而本文针对 Spam 相册进行检测是在账号上传照片之后进行，能够在 Spam 内容传播之前就检测到，从而避免 Spam 内容对正常用户的危害。Wang 等的工作只是利用了 Photo Spam 的传播特性，即分享操作，因此只能检测到传播 Spam 相册的账号，而无法检测到上传 Spam 相册的账号。通过本文检测到的大量 Spam 相册也表明，Wang 等的工作并不全面。针对 Cao 等的检测，攻击者同样可以通过代理来更换 IP 地址以及将照片上传行为分布在不同的时间段来绕过。本文的工作利用了 Photo Spam 中 Spam 信息存储和传播的核心，即 Spam 相册进行检测，构造了相册相关特征，并且这些特征都是易于计算且在照片上传之后就能提取的特征，能够在 Spam 信息传播之前就进行检测，从而有效避免 Photo Spam 对正常用户的威胁。

7 结束语

本文提出了一种利用有监督学习方法来检测 Photo Spam 中 Spam 相册的方案，与之前针对 Spam 账号的检测方案不同，本文针对 Spam 相册进行检测，能够及时检测到 Spam 相册并有效降低 Photo Spam 对正常用户的不良影响。本文首先分析了 Photo Spam 的攻击方式并对比了 Photo Spam 与传统 Spam 的区别，发现 Photo Spam 更难被检测到并且危害更大。基于此分析，本文设计了 12 个提取及时且计算高效的特征。然后提出了基于 SVM 分类算法的检测模型。通过包含有 2 356 个标记后相册的数据集的训练，检测模型能够对测试集中 Spam 相册和正常相册分别获得 0.991 的 F_1 值，随后本文分析了在训练集中不同样本比率、不同分类算法以及不同特征子集情况下的检测结果。最后，本文将检测模型应用到包含 315 115 个未标记相册的真实数据集中，共检测到 89 163 个 Spam 相册，人工验证正确率为 97.2%。

参考文献：

[1] GAO H, HU J, HUANG T. Security issues in online social networks[J]. IEEE Internet Computing, 2011, 15(4): 56-63.

[2] FIRE M, GOLDSCHMIDT R, ELOVICI Y. Online social networks: threats and solutions survey[J]. IEEE Communications Surveys and Tutorials, 2013, 16(4): 2019-2036.

[3] CAVIGLIONE L, COCCOLI M, MERLO A. A taxonomy–based model of security and privacy in online social networks[J]. International Journal of Computational Science and Engineering, 2014, 9(4): 325-338.

[4] GAO H, HU J, WILSON C. Detecting and characterizing social Spam campaigns[C]//The 10th ACM SIGCOMM conference on Internet Measurement. Melbourne, Australia, 2010: 35-47.

[5] LEE K, EOFF B D, CAVERLEE J. Seven months with the devils: a long-term study of content polluters on twitter[C]//The Fifth International AAAI Conference on Weblogs and Social Media. Barcelona, Spain, 2011: 185-192.

[6] MARTINEZ-ROMO J, ARAUJO L. Detecting malicious tweets in trending topics using a statistical analysis of language[J]. Expert Systems with Applications, 2013, 40(8): 2992-3000.

[7] MILLER Z, DICKINSON B, DEITRICK W. Twitter Spammer detection using data stream clustering[J]. Information Sciences, 2014, 260: 64-73.

[8] ZHANG J, GU G. NEIGHBORWATCHER: A content-agnostic comment spam inference system[C]//Network & Distributed System Security Symposium. San Diego, CA, United States, 2013.

[9] BENEVENUTO F, MAGNO G, RODRIGUES T. Detecting spammers on twitter[C]//The Collaboration, Electronic Messaging, Anti-abuse and Spam Conference. Redmond, USA, 2010:6-12.

[10] ZHENG X, ZENG Z, CHEN Z. Detecting spammers on social networks[J]. Neurocomputing, 2015, 159(2): 27-34.

[11] STRINGHINI G, KRUEGEL C, VIGNA G. Detecting spammers on social networks[C]//The 26th Annual Computer Security Applications Conference. TX, USA, 2010: 1-9.

[12] LEE K, CAVERLEE J, WEBB S. Uncovering social Spammers: social honeypots+ machine learning[C]//The 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Geneva, Switzerland, 2010: 435-442.

[13] ZHU Y, WANG X, ZHONG E. Discovering Spammers in social networks[C]//The 26th AAAI Conference on Artificial Intelligence. Toronto, Canada, 2012: 171-177.

[14] WANG G, KONOLIGE T, WILSON C. You are how you click: clickstream analysis for sybil detection[C]//The 22rd USENIX Security Symposium. Washington, USA, 2013: 241-256.

[15] CAO Q, YANG X, YU J. Uncovering large groups of active malicious accounts in online social networks[C]//The 2014 ACM SIGSAC Con-

ference on Computer and Communications Security. Scottsdale, USA, 2014: 477-488.

[16] LESKOVEC J, FALOUTSOS C. Sampling from large graphs[C]//The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia. USA, 2006: 631-636.

[17] GJOKA M, KURANT M, BUTTS C. Walking in facebook: a case study of unbiased sampling of OSNs[C]//IEEE Proceedings of INFOCOM. California, USA, 2010: 1-9.

[18] HU J, BAGGA A. Categorizing images in Web documents[J]. MultiMedia, IEEE, 2004, 11(1):22-30.

[19] WAN C, GENG J, CHENG H. Image spam identifying algorithm based on color and corner feature[J]. Computer Engineering, 2009, 35(15): 209-211.

[20] DREDZE M, GEVARYAHU R, ELIAS B A. Learning fast classifiers for image spam[C]//In Conference on Email and Anti-Spam(CEAS). California, USA, 2007: 1-9.

[21] GONEN M, ALPAYDIN E. Supervised learning of local projection kernels[J]. Neurocomputing, 2010, 73(10): 1694-1703.

[22] CORTES C, VAPNIK V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[23] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: one-sided selection[C]//International Conference on Machine Learning. Tennessee, USA, 1997:179-186.

[24] e1071: misc functions of the department of statistics, probability theory group[EB/OL]. http://CRAN.R-project.org/package=e1071. 2015.10.11.

[25] Rpart: recursive partitioning and regression trees[EB/OL]. http://CRAN.R-project.org/package=rpart. 2015.10.11

[26] RandomForest: breiman and cutler's random forests for classification and regression[EB/OL]. http://CRAN.R-project.org/package= randomForest. 2015.10.11.

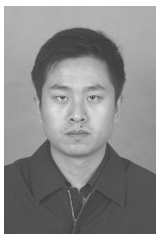
[27] Nnet: feed-forward neural networks and multinomial log-linear models[EB/OL].http://CRAN.R-project.org/package=nnet.2015.10.11.

作者简介：

吕少卿（1987-），男，山西五寨人，西安电子科技大学博士生，主要研究方向为在线社交网络安全。

张玉清（1966-），男，陕西宝鸡人，博士，中国科学院大学教授、博士生导师，主要研究方向为网络与信息系统安全。

刘东航（1990-），男，山西太原人，西安电子科技大学硕士生，主要研究方向为网络和信息安全。



张光华（1979-），男，河北石家庄人，博士，西安电子科技大学博士后在站，主要研究方向为信任管理、无线网络安全。