文章编号:1000-5471(2013)10-0001-06

基于多分形分析和主动学习反馈算法的图片垃圾邮件过滤①

周扬玲[1], 钟 剑[2], 邓 维[2]

1. 四川中医药高等专科学校,四川 绵阳 621000,2. 西南大学 信息中心,重庆 400715

摘要:图片垃圾邮件通常由随机变形技术制作,人眼认为内容相同但计算机认为不同,导致常规反垃圾邮件技术无法阻止这类垃圾邮件.根据图片垃圾邮件必然具有相似性、大量性和可变性的特点,提出了一种综合多向小波金字塔多分形分析算法和主动学习反馈驱动半监督支持向量机算法的创新图片垃圾邮件过滤方法.实验结果表明,该方法容易与常规反垃圾邮件系统相结合,而且该方法的效率高、准确性好、假阳性低,通过重复训练,可进一步提高准确性、降低假阳性,适合用于对抗性学习和垃圾邮件过滤.

关 键 词:图片垃圾邮件;多分形分析;主动学习聚类;反馈驱动半监督支持向量机

中图分类号:TP393.04 **文献标志码**:A

图片垃圾邮件是一种只有图片而没有文本或链接的邮件,即使包含文字信息也是以图片形式展现.常用的传统反垃圾邮件方法(如内容识别、贝叶斯、最大熵、支持向量机等)虽然在处理文本垃圾邮件方面具有准确率高、自学习能力强等特点,但是在过滤图片垃圾邮件时却无能为力.这主要是因为图片垃圾邮件发送者通常会使用各种新的图片生成与图片随机变形算法来制作图片垃圾邮件,这些算法通过改变图片形状或增加噪声背景来避免被反垃圾引擎所识别,但却不会影响收件人对邮件内容的理解.即使利用 OCR、图形文本识别技术等抽取图片中的文字信息后,再利用传统反垃圾邮件技术进行过滤,也很难达到较高精度.正是这些原因,导致图片垃圾邮件问题日益严重,已极大地影响到了人们的日常生活和工作.

近年来,研究人员积极探索通过获取图片标志性特征的方法来过滤图片垃圾邮件.Wu 等(2005)采用图片的可视化特征识别和 SVM 算法过滤图片垃圾邮件,能获得 81%的准确性和 1%的假阳性.Dredze 等(2007)用图形文本识别技术过滤图片垃圾邮件,能获得 90%的准确率.Wang 等(2007)用图片的颜色、纹理和小波转换系数进行垃圾邮件过滤,能获得 0.001%的假阳性.耿技等(2008)[1]用图片文本区域定位算法获取文本区域的文本信息,并用支持向量机(SVM)算法过滤图片垃圾邮件,能获得 85%的准确率.唐晓玲等(2008)[2]用信息接种技术改进垃圾邮件过滤器的准确率并保留邮件用户的偏好.虽然这些研究结果令人鼓舞,但是缺陷也非常明显,即都是使用图片的标志性特征作为过滤识别符号,而垃圾邮件制造者非常容易改变或不使用这些标志性特征,使得这些过滤方法失去应有的效果.

本文提出的创新方法是根据图片垃圾邮件大量性、可变性和相似性的本质特性进行过滤.大量性是指垃圾邮件发送者在发送同一批垃圾邮件时,总是用同一 IP 地址或 IP 地址段发出大量的邮件,这个特征容

① 收稿日期:2012-11-27
基金项目:四川省科技计划基金项目(10JJ8086);中央高校基本科研业务费专项基金项目(XDJK2013C005).
作者简介:周扬玲(1979-),女,四川绵阳人,硕士,讲师,主要从事网络与计算机应用方面的研究.

易被常规垃圾邮件过滤系统侦测到. 可变性是指邮件通过图片生成或随机变形算法处理后，垃圾邮件过滤器会认为图片的属性不相同而误判为正常邮件. 相似性是指同一批图片垃圾邮件虽然经过图片生成或随机变形算法处理后，图片属性不完全一致，但是一定相似. 因为只有图片基本相似，人才会对这些垃圾邮件的内容理解保持一致. 以上这些特性，决定了把常规垃圾邮件过滤规则和图片垃圾邮件过滤算法结合起来进行邮件过滤，不但能提高图片垃圾邮件过滤的准确性，而且能减少垃圾邮件过滤的经济投入.

1 图片垃圾邮件的特征选择与抽取算法

图片的特征包括颜色、布局、形状和纹理，但是，图片的颜色、布局和形状非常容易改变，而图片的纹理特征却不容易改变，因此，纹理分析在图片垃圾邮件过滤中扮演着至关重要的角色. 图片纹理是图片区域中的一系列灰度变化，图片纹理分析就是分析这些灰度的强度变化. 目前，用多重分形分析(Multifractal Analysis，MFS)数学框架分析图片纹理已经成为一种趋势[3-4]，然而，普通多分形算法分析图片纹理的稳定性较差. Xu 等(2010)提出了一种创新的多分形分析方法，即小波多分形谱(WMFS)方法. WMFS 的图片纹理描述符包括低频(MFS(D_j，θ))、高频(MFS(w_k，j，θ))和小波领导(MFS(L_j，θ))3 个部分，表示如下：

$$\mathrm{WMFS}=\{mean_\theta(\mathrm{MFS}(D_{j,\theta})),\ mean_\theta(\mathrm{MFS}(W_{k,j,\theta})),\ mean_\theta(\mathrm{MFS}(L_j,\theta))\}$$

小波多分形谱图片纹理描述符具有稳定性好、包含图片纹理信息丰富的优点. 但该描述符的维度高，在标度为 3($j=3$)和 3 个通道($k=1,2,3$)的情况下，图片纹理描述符的维度高达 338. 如果为了获得更多图片纹理信息而增加标度数和通道数的话，WMFS 的图片纹理描述符维度数将急剧增加.

高维度图片纹理描述符通常包含大量冗余信息和不相关信息，不但会增加计算复杂度，而且会导致图片垃圾邮件过滤器分类不准确. Unler 等(2011)[5]提出了一种利用互斥信息作为最大相关最小冗余测度的粒子群搜索最大相关最小冗余特征子集选择算法(mr2PSO)，互斥信息测度作为粒子群的搜索权重，通过粒子群智能搜索获取仅包含最大相关最小冗余特征的子集. Unler 的实证结果显示，通过 mr2PSO 算法抽取的最大相关最小冗余特征的子集不但能有效地提高计算效率，还能有效提高 SVM 的分类精度.

本文将在第三部分中，利用小波多分形谱图片纹理描述符算法抽取图片的纹理特征，用 mr2PSO 算法提取小波多分形谱图片纹理描述符的最大相关最小冗余特征子集，用提取的子集作为判断图片邮件是否是垃圾邮件的根据.

2 主动学习的反馈驱动半监督支持向量机分类算法

2.1 半监督支持向量机分类算法

目前流行的垃圾邮件过滤器通常采用机器学习和模式识别算法对垃圾邮件进行自动识别，这些算法都要求存在大量已经标记好的邮件作为训练集. 在实际使用中，每天的邮件数量巨大且千差万别，获取足够的训练集需要昂贵的代价. 相对于已标记好的邮件，未标记的邮件非常容易获取. 如果训练数据集中只有少量标记好的邮件，过滤分类器就会出现过拟合情形，使得垃圾邮件过滤器没有足够能力判别邮件是否为垃圾邮件. 为解决该问题，过滤器需要利用已标记的和未标记的邮件进行训练. 如果过滤分类器具有自学习能力，过滤分离器的精度将随着使用时间的增加而提升. Demiriz 和 Bennett(1998)[6]提出的半监督支持向量机算法(SSL-SVM)能实现这一目标，一些研究者利用 SSL-SVM 算法进行垃圾邮件分类处理，获得了较好的效果. 但是，SSL-SVM 仅能利用未标记邮件的很少一部分信息，即使长期训练也难以较大幅度地提高垃圾邮件过滤器的精度.

2.2 用户反馈重标记和主动学习聚类算法

为弥补 SSL-SVM 算法在分类垃圾邮件时的缺陷，本文提出一种创新的重复训练集合构造方法. 重复训练集合包含 3 个子集：用户反馈的已标记邮件(包括被垃圾邮件过滤器忽略或错误标记的邮件)、重新标

记的可疑邮件和过滤中发现的可疑未标记邮件. 许多垃圾邮件过滤器允许用户提交邮件过滤结果的反馈，针对被忽略的垃圾邮件或是被过滤器误判的邮件，用户能手动地进行确认或修改. 将用户反馈的结果用于重新训练邮件过滤分类器可大大提高邮件过滤器的过滤精度，但用户通常会因为过多反馈邮件而厌烦，导致反馈邮件的数量不足. 为获得足够数量的用户反馈标记邮件，可将聚类算法与主动学习算法相结合，使用户只需要标记最有代表性的某个邮件，图片垃圾邮件过滤器就能自动标记相似类型的所有邮件. 本文选择 DeBarr 和 Wechsler(2012)[7]提出的中心点分割算法(Partitioning Around Medoids，PAM)实现该目的，即利用主动学习算法对未标记的疑是垃圾邮件进行聚类后，选择每一类中处于最中心点的邮件作为典型代表提交给用户进行人工重新标记. 当用户标记后，以各类中心点的平均距离作为传染距离，传染距离内的所有邮件都自动标记为与中心点邮件一样的标记.

3　图片垃圾邮件过滤算法

为了充分利用图片垃圾邮件的大量性、相似性、可变性特征进行过滤，需要图片垃圾邮件过滤算法和传统垃圾邮件过滤算法相结合，其最为经济和方便的思路是将图片垃圾邮件过滤器算法模块以插件形式存在于常规垃圾邮件过滤器中. 本文以开源垃圾邮件过滤器 Spam Assassin 为例进行算法实现，其他常规垃圾邮件过滤器的实现方法类似. SpamAssassin 自带的探测规则如 honeypots、sender analysis 等已经能较好地识别普通垃圾邮件，图片垃圾邮件过滤器算法插件将与这些规则协同工作，共同完成图片垃圾邮件的过滤.

本综合算法的具体步骤是：

1) SpamAssassin 获取一封新邮件，通过常规规则对邮件的特性进行探测，如果该邮件是图片邮件且被 SpamAssassin 规则判定为垃圾邮件，则直接将其阻塞掉；否则，邮件会被提交给图片垃圾邮件过滤器插件进行进一步判断.

2) 插件首先使用多分形分析算法获取图片邮件的小波多分形谱图片纹理描述符，然后利用 mr2PSO 算法抽取仅包含最大相关最小冗余特征的子集.

3) 抽取的子集作为图片邮件的纹理描述符用于 SSL-SVM 分类. 如果邮件被归类为垃圾邮件，则作为疑是图片垃圾邮件，删除标签后放入重复训练池中.

4) 对重复训练池中的邮件进行主动学习聚类算法处理，提取该类中心的邮件作为代表请用户手动标记.

5) 插件获取用户的反馈后，利用 PAM 算法自动标记传染距离范围内的所有邮件，查看忽略的垃圾邮件数目(设定为 O)和被错误标记的邮件数目(设定为 E)是否达到预定的阈值(设定为 P). 如果达到阀值 P，插件就使用重复训练池(设定为 T)中的数据集对支持向量机进行重新培训，以提高过滤能力. 如果未达到阈值 P 的数量，则继续等待新图片垃圾邮件的到来. 算法 1 即是插件与 SpamAssassin 合作的图片垃圾邮件过滤算法.

算法 1：图片垃圾邮件过滤综合算法

```
1  Set     : re-training pool T ← □;
2  Set     : omitted spam O ← 0;
3  Set     : error labeled email E ← 0;
4  Set     : query Q ← □;
5  Input:  predefine threshold value P;
6  Input:  Image-based email I
7  If      Max(O, E)>=P And |T|>= Max(O, E) Then
8        re-training SSL-SVM use T;
9        T ← T − labeled emails;
10 else
```

续算法1：图片垃圾邮件过滤综合算法

```
11      get the relabel omitted spam n: O←O+n;
12      get the relabel error labeled email m: E←E+m;
13      If (O+E)/2 <= |{ T labeled emails}| then
14        PAM {T labeled} to (O+E)/2 clusters;
15        Q ← center emails in each clusters ;
16        Q to query the end users;
17      else
18        get the labeled query Q;
19        foreach Q_i in Q do
20          calculate average distance of center ad_i ;
21          propagation rate r=ad_i ;
22          Set label same as Q_i in extended emails <=r;
23  If I is spam identified by rules SpamAssassin then
24      block I;
25      goto 6;
26  else
27      if I is image-based email then
28        texture descriptor WMFS ←multifractal analysis;
29        feature subset FS ← mr2PSO← WMFS;
30        if FS is spam identified by SSL-SVM then
31          T ← I T
32        else
33          deliver I;
34          goto 6;
35    else
36          deliver I;
37          goto 6;
```

4 实验检验及效果评价

为检验算法的效果，检验使用 SpamAssassin 3.3.0 进行实验，并将图片过滤器以插件形式嵌入 SpamAssassin 中. 为了满足算法1的流程，对 SpamAssassin 的代码进行适当修改. 为了便于过滤中文垃圾邮件，在 SpamAssassin 添加了中文垃圾邮件过滤规则，中文垃圾邮件过滤规则来源于教育科研网.

4.1 图片邮件初始训练集

电子邮件通信有着很强的私密性，难于建立或获取初始图片邮件分类训练集合. 虽然目前一些研究机构和学者建立有普通邮件分类实验集合，但图片邮件的分类训练集合却很少. Dredze 等学者建立了一个图片邮件训练集，该训练集由图片非垃圾邮件(2 550 张图片，其中 2 359 张具有不相关的纹理特性)、图片垃圾邮件(3 239 张图片，其中 1 248 张具有不相关的纹理特性)和未标记的图片邮件(9 503 张图片中，2 173 张具有不相关的纹理特性)三部分组成，这些图片邮件数据被无私地公布在他们的网站上，可以免费下载公开使用. 本实验使用 Dredze 的数据集作为初始训练集，从本单位邮件服务器收集到的相关图片邮件作为测试集合，对图片垃圾邮件过滤方法的有效性和精确性进行评估.

4.2　实验设置及结果

本实验使用的 SpamAssassin，除为了满足对算法 1 的流程所做的少量代码修改和增加了中文过滤规则外，其余都使用 SpamAssassin 自带规则集合并采用默认设置. 由于对过滤器的再培训不但会花费一定的时间，而且会影响过滤器的实时精确度和有效性，因此再培训的预定义阈值(用 M 表示)非常重要. M 为一天中图片垃圾邮件的最可能数量，计算公式为 $M=etn\times sp\times isp$. 其中，etn 代表邮件服务器每天所处理的邮件总数量，sp 代表所有邮件中垃圾邮件的百分比，isp 代表垃圾邮件中图片垃圾邮件的百分比. 我单位的邮件服务器每天大概处理 100 000 封邮件，sp 值为 80%，isp 值为 25%，因此，M 值为 20 000.

实验过滤系统的效果评估时间期间为 2012 年 7 月 1 日到 9 月 1 日，每天处理约 100 000 封邮件. 前 10 天中，准确率为 85%，假阳性率为 4%；中间 10 天中，准确率为 90%，假阳性率为 3%；最后 40 天中，准确率达到 98%，假阳性率降为 1%. 与文献[1]、[4]中的方法相比，准确率更高，假阳性更低. 实验证明本方法是有效的，而且通过多次基于用户反馈的重复训练后，准确率会进一步显著提高，假阳性率会明显降低.

5　研究结论与应用展望

图片垃圾邮件占用的网络带宽大、耗费存储空间多、危害比文字垃圾邮件更大. 然而，图片垃圾邮件，特别是经过随机变形处理的图片垃圾邮件非常难于过滤. 本文从图片垃圾邮件的大量性、可变性、相似性等本质特征入手，首先利用多分形分析算法提取图片邮件的小波多分形谱图片纹理描述符，并利用 mr2PSO 算法去除图片纹理描述符中的冗余和不相关信息. 然后用 SSL-SVM 进行分类判断，判断后的结果通过主动学习聚类算法处理后提起典型代表邮件供用户手动重标记和 PAM 算法进行自动重标记而形成 SSL-SVM 的再培训集，用于提升分类过滤器的效率. 该综合算法流程容易编制成常规垃圾邮件过滤器的插件并与常规垃圾邮件过滤器协同工作，在保护原来投资的基础上提高了图片垃圾邮件的过滤能力.

随着计算机网络的普及和信息技术的发展，图片垃圾邮件发送与过滤、黑客入侵攻击和入侵检测、网络管理与网络阻塞等的对抗性信息安全问题会越来愈多，这些安全性问题通常包括大量性、可变性和相似性等基本特征. 对于这类对抗性信息安全问题，常规的算法难以充分利用其可变性和相似性特征，而多分形分析和主动学习反馈算法不但能有效利用这些可变性和相似性特征，而且能充分利用人的反馈因素，适用于处理对抗性学习与防御问题.

参考文献：

[1]　耿　技，万明成，程红蓉，等. 基于文本区域特征的图像型垃圾邮件过滤算法 [J]. 计算机应用，2008，28(8)：1904－1906.

[2]　唐晓玲，张自力. 一种基于 Agent 的信息接种反垃圾邮件模型研究 [J]. 西南大学学报：自然科学版，2008(5)：154－158.

[3]　XU Y，HUI J，FERMULLER C. Viewpoint Invariant Texture Description Using Fractal Analysis [J]. International Journal of Computer Vision，2008，83(1)：85－100.

[4]　WENDT H，ROUX S G，ABRY P，et al. Wavelet Leaders and Bootstrap for Multifractal Analysis of Images [J]. Signal Process，2009，89：1100－1114.

[5] ALPER UNLER，ALPER MURAT，RATNA BABU CHINNAM. mr2PSO：A Maximum Relevance Minimum Redundancy Feature Selection Method Based on Swarm Intelligence for Support Vector Machine Classification [J]. Information Sciences，2011，181(20)：4625－4641.

[6] KRISTIN P. BENNETT，AYHAN DEMIRIZ. Semi-Supervised Support Vector Machines [J]. Neural Information Processing Systems，1998(11)：368－374.

[7] DAVE DEBARR，HARRY WECHSLER. Spam Detection Using Random Boost [J]. Pattern Recognition Letters，2012，33(10)：1237－1244.

On Filtering Image Spam Based on Multifractal Analysis and Active Learning Feedback Algorithm

ZHOU Yang-ling[1], ZHONG Jian[2], DENG Wei[2]

1. Sichuan College of Traditional Chinese Medicine, Sichuan, Mianyang 621000, China;

2. Southwest University, Information Technology Center, Chongqing 400715, China

Abstract: Image spam has usually been produced by a variety of images randomized deformation algorithms. The spam can make the message fully legible by the human eye but undistinguishable by the common anti-spam engines. In this paper we have proposed a novel image spam recognition composite method which is a hybrid algorithm of multifractal analysis in multi-orientation wavelet pyramid algorithm and active learning feedback-driven semi-supervised support vector machine algorithm based on three natures of image spam: large quantity, similarity and variability. The experimental results demonstrate that our method is easy to plug into conventional anti-spam system with high efficiency, high accuracy and low false positive rate. The accuracy will be improved and the false positive rate reduced along with more and more retraining. So, the method is fit especially for an adversarial learning and processing like spam filtering.

Key words: image spam; multifractal analysis; active learning cluseting; feedback-driven semi-supervised support vector machine

责任编辑 汤振金