

一种基于最大熵模型的实时垃圾邮件判定及阻断技术研究

陈 炼

(贵州安顺职业技术学院, 贵州 安顺 561000)

摘 要: 垃圾邮件过滤是网络安全领域的重大研究课题和难题. 本文针对传统技术的不足, 使用机器学习领域经典的熵模型对垃圾邮件进行判定, 并结合其采用针对 SMTP 邮件通信协议的实时阻断技术. 实验结果表明, 该技术简单有效, 并适合实时在线处理.

关键词: 网络安全; 垃圾邮件; 机器学习; 最大熵模型; 实时阻断

中图分类号: TP309

文献标识码: A

文章编号: 1000-7180(2012)09-0091-04

Research on ME-Based Spam Identification and Real-Time Spam Blocking

CHEN Lian

(Anshun Vocational & Technical College, Anshun 561000, China)

Abstract: Spam filtering is a classic research topic in network security. This paper proposes a new mechanism, which based on the maximum entropy model for spam identification and can finish real-time blocking method in terms of the characteristics of SMTP. Experimental results demonstrate it is effective for spam identification and real-time blocking.

Key words: network security; spam filtering; machine learning; maximum entropy model; real-time blocking

1 引言

垃圾邮件问题已经消耗了大量的网络资源, 并给人们带来了极大的不便^[1]. 目前, 反垃圾邮件技术大致可以分成如下几类: 接入控制方法, 包括黑名单、白名单、灰名单以及延迟技术; 身份验证方法, 包括 DNS MX 查询、反向 DNS 解析; 内容过滤技术, 包括基于规则的 RIPPER 方法、基于统计的贝叶斯方法、Memory-based 方法、SVM 方法^[2]等; 以及, 2005 年最新出现的基于 URL 精确匹配过滤、基于 URL 聚类的过滤方法以及在开源社区比较著名的开源垃圾邮件过滤系统 Spamato 中的三种基于 URL 的过滤方法: Razor Filter、Earlgrey Filter 和 Domainator^[3]. 传统的反垃圾邮件技术虽然试图从各个层面上解决这个问题, 但是仍旧存在误报率高、处理的数量过大而导致实时性能不高、不能进行事

前处理等缺陷.

针对上述情况, 本文提出一种最大熵模型, 对垃圾邮件中 URL 所表现出的重要特征进行基于分类的机器学习, 形成垃圾邮件过滤器, 并尝试根据 SMTP 协议的特点, 进行实时在线地过滤垃圾邮件. 该方法旨在解决当前垃圾邮件过滤技术当中普遍存在的误报率高、实时性能不强等问题.

2 基于 URL 的特点进行垃圾邮件过滤

统一资源定位符 (Uniform Resource Locator, URL) 是当前垃圾邮件中最为稳定出现和起着重要作用的因素, 它比传统垃圾邮件过滤技术中采用的关键词更为普遍和具有区分性. 这是因为:

(1) 在众多的垃圾邮件当中, 相关的 URL 都会分布出现在邮件体当中 (邮件体的前面部分、中间部分和末尾部分), 这样可以起到一个重复地吸引用户

访问的目的;

(2) 从处理的计算量以及空间耗费方面来说, 处理 URL 远比传统的基于内容的垃圾邮件过滤方法来说耗费要小得多, 因而适用于实时的在线处理环境;

(3) 垃圾邮件中出现的 URL 和正常邮件中出现的 URL 在路径长度、特殊字符的出现频率等方面都有较大的区别, 因而在判定中以 URL 为依据具有较好的区分性。

3 基于最大熵模型(ME)的邮件分类器

3.1 最大熵模型原理

最大熵原理以信息熵为基础, 其基本思想是: 给定一个经验概率分布(从已知训练样本获得) p^* , 构造一个概率分布模型 p 尽可能与经验概率分布 p^* 相接近. 但构造这样的概率分布模型的方法很多, 文中选择一种使得熵最大的概率分布模型 p , 也就是说, 所求概率分布模型 p 是满足最大熵约束的最优解^[4].

鉴于篇幅关系, 有关最大熵模型的推导及相关原理请参见文献[4]. 在本文中, 使用最经典方法-拉格朗日乘数法, 可求得最优解为:

$$p^*(x|y) = \frac{1}{\pi(y)} \exp\left(\sum_i \lambda_i f_i(x, y)\right) \quad (1)$$

其中, 参数 λ_i 是对应特征函数 f_i 的权重, $\pi(y)$ 是归一化因子.

$$\pi(y) = \sum_a \exp\left(\sum_i \lambda_i f_i(a, y)\right) \quad (2)$$

如果通过在训练样本上进行学习, 就可求得式(1)中的 λ_i 值, 即可得到所求的概率分布, 这样就完成了满足最大熵模型的构造.

在本文所述的基于 URL 分类的垃圾邮件过滤任务中, 式(1)中的 $x = \{\text{spam}, \text{ham}\}$, 所讨论的 URL 分类问题本质上是一个在约束集合为: $P = \{p | p(f_i) = p^*(f_i), i=1, 2, 3, 4\}$ 的前提条件下, 求式(1)的最优解问题.

3.2 特征提取

根据文献[5]所述的 χ^2 测试方法, 确定选取如下四个重要特征作为邮件分类器的学习特征: URL 路径深度、URL 中是否包含数字字符串以及包含数字的数量、URL 中是否含有“http://www”字符串、URL 中是否包含特殊字符(“?”、“&”、“=”等)及其出现的数量.

3.3 基于 URL 的邮件分类器

对于选择特征 1, 对于类别为 x' 的 URL 给出如

下的特征函数:

$$f_{x'}(x, y) = \begin{cases} \text{nest}(y), & x=x' \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

式中, $\text{nest}(y)$ 表示 URL_y 的路径深度.

对于选择特征 2, 对于类别为 x' 的 URL 给出如下特征函数:

$$f_{\text{contno}, x'}(x, y) = \begin{cases} \text{no}(y), & \text{if } (x=x') \wedge \text{contno}(y) = \text{true} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

式中, $\text{no}(y)$ 表示 URL_y 中包含的表示 ID 的数字串的个数; $\text{contno}(y)$ 表示 URL_y 中是否包含数字串.

对于选择特征 3, 对于类别为 x' 的 URL 给出如下特征函数:

$$f_{\text{cont}, x'}(x, y) = \begin{cases} 1, & \text{if } (x=x') \wedge \text{cont}(w, y) = \text{true} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

式中, $\text{cont}(w, y)$ 表示 URL_y 中是否包括字符串 w .

对于选择特征 4, 对于类别为 x' 的 URL 给出如下特征函数:

$$f_{\text{cont}, w, x'}(x, y) = \begin{cases} \text{num}(w, y), & \text{if } (x=x') \wedge \text{cont}(w, y) = \text{true} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

式中, $\text{cont}(w, y)$ 表示 URL_y 中是否包括字符串 w , $\text{num}(w, y)$ 表示 URL_y 中包含的字符串 w 的个数.

4 在线垃圾邮件过滤模型

通过利对 SMTP 协议传输特性的分析发现^[6]: 该数据包(“DATA”包)分别传输, 并且一封邮件一般通过多个 DATA 包进行传送. 可以通过上述 ME 模型分类 URL 的分类器来在线(在接收邮件的 SMTP 服务器端)实时地判定垃圾邮件, 并实行即时的过滤工作(不等待所有的 DATA 包为本地邮件服务器接受). 这样, 不但可以极大地节省网络带宽以及系统的计算资源和邮件服务器的存储资源, 并且增加了网关处理邮件的吞吐率. 具体的在线垃圾邮件过滤模型如图 1 所示.

该模型包含如下几个组件:

(1) 邮件解码器主要用来对到来的邮件进行解码, 包括 Base64、quoted-printable、Unicode 等邮件编码格式的解码;

(2) 邮件处理器对解码后的邮件进行 URL 的提取;

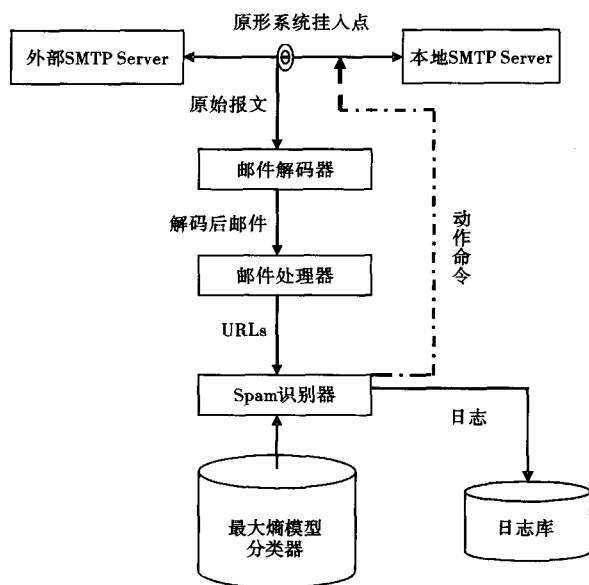


图1 基于URL的在线垃圾邮件过滤模型示意图

(3) 最大熵模型分类器提取邮件中的URL构造出分类器,以供Spam识别器在线垃圾邮件过滤时使用;

(4) Spam识别器采用离线构建好的基于最大熵模型分类器对提取的URL进行分类,从而识别垃圾邮件;

(5) 日志库用记录Spam识别器判定垃圾邮件的日志,以便于系统正确性和有效性的验证工作。

5 实验结果与分析

为了验证本文所述方法的有效性,我们首先将该方法的核心——基于ME模型分类器与经典的基于内容过滤的Naïve Bayes方法和SVM方法在收集的实际邮件集上的过滤效果进行了比较;另外,为了验证其在实际网络环境下的可用性以及高效性,文中开发了一个Linux下的在线实时垃圾邮件过滤原型系统。

5.1 对比实验

为了验证基于ME模型分类器的效果,收集了一段时间内的实际邮件共10 342封,其中垃圾邮件3 187封,正常邮件7 155封。将其分为训练集和测试集两个集合。采用十折交叉验证(Cross-Validation)的方法,重复实验10次,取实验结果的均值进行比较。在实验的结果中,将采用目前国际上较为流行的Precision以及Recall两个指标来衡量该方法的有效性。表1给出了详细的实验结果。

表1 对比实验结果

分类器参数	Precision	Recall
Naïve Bayes (keywords)	93.7%	92.6%
Naïve Bayes (URLs)	95.8%	93.3%
SVM (keywords)	96.8%	94.5%
SVM (URLs)	97.2%	96.7%
ME model (keywords)	98.4%	98.2%
ME model (URLs)	99.90%	98.3%

通过表1可以看出,ME模型在邮件过滤方面,比Naïve Bayes方法有较高的准确率,与SVM方法相差不多。另外,实验显示,采用基于URL的邮件过滤策略的准确率与基于关键词过滤方法相比,也有略微的提高。然而,采用基于URL的过滤方法,并不需要保存待判定邮件才能进行计算,而关键词为代表的传统内容过滤方法则需要对整个邮件进行保存和计算。

5.2 原型系统实验

文中在Red Hat Enterprise Linux 6.0下基于Sendmail邮件服务器开发了一个垃圾邮件过滤的原型系统,并部署该原型系统在测试邮件网关上对进出的邮件进行垃圾邮件过滤。通过用户反馈及数据分析,得到如表2所示实验数据。

表2 垃圾邮件过滤原型系统测试结果

处理邮件总数	28 327 封
正常邮件	18 718 封
垃圾邮件	9 609 封
Spam Precision	99.7%
Spam Recall	98.2%

6 结束语

本文提出了一种采用最大熵模型对邮件中提取的URL特征进行分类学习,并能够依据SMTP协议的传输特性,使用分类器对垃圾邮件进行在线过滤方法。实验结果证明该方法简单有效。如何将该方法与其他诸如DNS实时黑名单等技术有机融合,提高当前垃圾邮件过滤系统的可用性将是下一步研究的重点。

参考文献:

[1] 吴小红,李俊. 一种透明邮件过滤模型研究与实现

- [J]. 微电子学与计算机, 2004, 21(10): 63-67.
- [2] Zhang L, Zhu J, Yao T. An evaluation of statistical spam filtering techniques[J]. ACM Transactions on Asian Language Information Processing, 2004, 4 (13): 243-269.
- [3] Androutsopoulos I, Koutsias J, Chandrinos K, et al. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages[C]//Proceedings of the International ACM SIGIR Conference. Athens: ACM Press, 2000: 160-167.
- [4] Adwait R. A simple introduction to maximum entropy models for natural language processing[R]. University of Pennsylvania, 1997.
- [5] Yang Y, Pedersen J. A comparative study on feature selection in text categorization[C]//Proceedings of IC-ML-97, 14th International Conference on Machine Learning. San Francisco: IEEE Computer Society, 1997: 412-420.
- [6] Internet Engineering Task Force (IETF). RFC821: simple mail transfer protocol[EB/OL]. <http://www.rfc-editor.org/rfc/rfc821.txt>. 1982.

作者简介:



陈 炼 男, (1963—), 副教授. 研究方向为网络技术与网络信息安全.

(上接第 90 页)

- [5] Wu B, Wu J. k-Anycast routing schemes for mobile ad hoc networks [C]//Proceedings of IEEE IPDPS 2006 IEEE International Parallel and Distributed Processing Symposium (IPDPS 2006). 2006 Rhodes Island, Greece: Citeseer. 27-35.
- [6] Wang X N, Tang Z M, Qian H Y. Study and design of k-Anycast communication model in IPv 6 [J]. Journal of China Institute of Communications, 2009, 30 (6): 120-125.
- [7] 张园园, 余镇危. 基于覆盖网络的应用层选播服务的设计与研究 [J]. 微电子学与计算机, 2011, 28 (12): 72-74.
- [8] Santiago J, Casaca A, Pereira P R. Multicast in delay tolerant networks using probabilities and mobility information [J]. I. J. of Ad Hoc & Sensor Wireless Networks, 2009, 7 (1/2): 51-68.

作者简介:



张永晖 男, (1973—), 博士. 研究方向为移动互联网接入、容迟网络.

林漳希 男, (1953—), 博士, 博士生导师. 研究方向为 QoS 路由和网络安全.

蒋新华 男, (1956—), 博士生导师. 研究方向为智能交通通信、高速移动网络.

梁 泉 男, (1972—), 博士, 副教授. 研究方向为移动网络、机会网络.