

基于判别模型的垃圾邮件过滤方法

许 镇¹,王洪国²,冉玉梅¹,杨玉会¹

(1. 山东师范大学 信息科学与工程学院, 山东 济南 250014;

2. 山东省科学技术厅, 山东 济南 250014)

摘 要:垃圾邮件泛滥已成为网络时代的一个重要问题,随着垃圾邮件的伪装技术的不断更新,以前主要的几种垃圾邮件过滤技术面临着新的挑战。文中提出一种新的基于判别模型的垃圾邮件过滤方法,邮件分类器通过不断的学习来更新特征项的权重,当新的信息到达时,计算所有特征项的权重之和,并将其转化为一个概率值,如果此概率值超过某一阈值时,就认定此信息为垃圾邮件;同时将此方法应用到实时邮件处理环境中。实验结果表明,此方法能明显地提高准确度,有效地降低误判率。

关键词:互信息;判别模型;垃圾邮件过滤;梯度下降法

中图分类号:TP393.098

文献标识码:A

文章编号:1673-629X(2010)01-0181-04

Spam Filter Method Based on Discriminative Model

XU Zhen¹, WANG Hong-guo², RAN Yu-mei¹, YANG Yu-hui¹

(1. School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China;

2. Department of Science and Technology of Shandong Province, Jinan 250014, China)

Abstract: Spam e-mail is increasingly becoming a great problem in the Internet age. As the latest generation of spam incorporates sophisticated tactics, previous spam filtering technologies face a new challenge. Proposed a novel online spam filter based on discriminative model. Spam classifier updates the weights of features by continual learning. When a new message arrives, compute the sum of all weights and convert it to a probability. If that probability is over some threshold, predict that the message is spam, then applied the technique to online processing environment. Experimental results demonstrate that it can significantly raise the filtering accuracy, effectively reduce false positives.

Key words: mutual information; discriminative model; spam filter; gradient descent

0 引言

伴随着 Internet 的普及,电子邮件以其快捷、方便、低成本的特点日益得到了广泛地使用,成为互联网上最重要、最普及的应用。但是随之而来的垃圾邮件也越来越猖獗,不仅造成了邮件服务器拥塞,给用户带来不便,更严重的是一些邮件还含有色情和反动内容。除此之外,垃圾邮件还为企业带来了巨大的损失^[1]。据联合国贸发会议援引 Message Labs 的数据说,垃圾邮件给全球企业带来的损失高达 205 亿美元。因此,如何有效地遏制垃圾邮件迫在眉睫。

近年来,有关垃圾邮件过滤技术的研究逐渐兴起。最初是从电子邮件的半结构化特性出发,寻找垃圾邮件的特征,从邮件头、邮件体等各方面展开邮件过滤工作。常见的过滤方法有黑、白名单技术,过滤规则等,但由于邮件发送者在不断变化、规则难以维护、准确率不高等原因,这些方法都具有一定的局限性^[2]。随着机器学习技术的发展,越来越多的科研人员将此技术运用于垃圾邮件过滤并取得了不错的效果。

机器学习技术大体上可以分为两类:生成模型(如朴素贝叶斯模型)和判别学习模型(如支持向量机模型、最大熵模型和逻辑回归模型)。在大多数的文本分类中,如果存在丰富的训练数据,判别学习模型的运行效果要优于生成模型^[3]。文中提出的方法类似于机器学习中的线性模型(linear model),抽取邮件体中的单词以及邮件头部信息作为特征项,权值是借助于逻辑回归模型(logistic regression model)的梯度下降法不断训练得到的。

收稿日期:2009-04-13;修回日期:2009-07-08

基金项目:山东省自然科学基金(Q2006G03);山东省科技攻关项目(2009GG10001008);山东省软科学研究计划项目(2009RKA285)

作者简介:许 镇(1984-),男,山东济南人,硕士研究生,研究方向为文本挖掘、信息过滤;王洪国,博士,教授,研究方向为组合优化、数据挖掘、电子政务。

1 基于 Regression Model 的邮件分类

1.1 邮件预处理

目前,在信息处理领域,文本的表示主要采用向量空间模型。基本思想是以向量来表示文本:

$$d: (w_1, w_2, \dots, w_n)$$

其中, w_i 为第 i 个特征项在文本 d 中的权重。

根据实验结果,可以认为选取词作为特征项要优于字和词组。文中以词作为特征项,并进行了词根化和去除禁用词。

1.2 特征选择

过高的向量维数会对分类器的训练时间产生很大的影响,同时对存储空间提出更高的要求,所以在实际应用中仍然要进行特征选择,文中选择差分互信息作为特征选择策略。

互信息是单词与类别间共享信息的度量,定义:

$$I(t, c) = \log \frac{P(tc)}{P(t) * P(c)}$$

其中 $P(t)$ 表示单词 t 在所有文本中出现的概率, $P(c)$ 表示类别为 c 的文本在所有文本中的概率, $P(tc)$ 表示类别为 c 且包含单词 t 的文本在所有文本中的概率。互信息量越大,表明单词和类别之间的共现概率也越大。

互信息量的不足之处是偏向于选择稀有单词,这可以由互信息量的等同计算公式看出:

$$I(t, c) = \log P(t | c) - \log P(t)$$

因此为了防止这样的情况发生,互信息量还可以通过如下变形得出:

$$I(t, c) = P(t | c) \log \frac{P(tc)}{P(t) * P(c)}$$

在分别求出单词 t 与各个类别的互信息量后,总的信息量 t 可根据下面的公式计算:

$$I_{\max}(t) = \max_{i=1}^m \{I(t, c_i)\}$$

在互信息量方法中, $I(t, c)$ 即为特征 t 对类别 c 的“贡献”,而总的信息量则是特征 t 对所有类别的“贡献”的最大值。

定义 1 在二类文本分类问题中,单词 t 的差分互信息量(MID)为 t 与两个类别的互信息量之差的绝对值:

$$\text{MID}(t) = |I(t, c_1) - I(t, c_2)|$$

在训练数据集中,计算所有单词的 MID 值,将低于某个阈值的单词从特征集中移除,或选择 MID 值最高的若干个特征作为特征集^[4]。

1.3 特征向量的构造

通过特征选取确定特征空间后,就可以在此空间中对一封邮件的文本进行向量化,确定该文本向量在

每一维上的权重。根据有关实验表明,TF-IDF 和 Bernoulli 模型的性能在分类准确率上基本相同,但采用 TF-IDF 模型时所需要的训练时间更短,因此文中采用 TF-IDF 进行邮件向量化:

$$x_i = \frac{\text{TF}(w_i) \times \log(D/D_{w_i} + \alpha)}{\sqrt{\sum_{k=1}^n \text{TF}^2(w_k) \times \log^2(D/D_{w_k} + \alpha)}}$$

其中, $\text{TF}(w_i)$ 为单词 w_i 出现的次数, D 为训练集中所有文档的总数, D_{w_i} 为单词 w_i 曾在其中出现过的文档数量, α 为常数,文中选取其值为 0.01。

1.4 邮件分类

通过学习建立一张特征项的权值表,当新的信息到达时,提取信息中的所有特征项,并将它们相对应的权值相加,记作 $w \cdot x$,其中 w 代表权值向量, x 为 0 或 1, 1 表示在权值表中存在与信息中的内容相应的特征项。因此,可以理解为 $w \cdot x$ 表示新信息中所有特征项的权值之和,之后,通过下列函数(1)将权值之和转化为一个概率值:

$$p(Y = \text{spam} | x) = \frac{\exp(w \cdot x)}{1 + \exp(w \cdot x)} \quad (1)$$

当概率值超过了某个事先约定的阈值时,就可以认定此邮件为垃圾邮件,否则就认为是合法邮件。

下一步,更新权值。常用的方法即为梯度下降法,为更准确地计算出下降方向,通常通过计算函数的导数得出。对于逻辑回归模型而言,梯度就变得非常简单, $(1-p) \times x$ 或 $p \times x$,这取决于所渴望得到的结果,以及信息的种类,是垃圾邮件还是合法邮件^[5]。除此之外,存在一个正的常数叫做学习速率,它决定梯度下降搜索中的步长,一般选取学习速率 $\text{rate} = 0.02$ 。因此权值更新如下(其中 x_i, y_i 表示一组消息序列):

$$\begin{aligned} & \text{if } (y_i = 1) \\ & \quad w = w + (1 - p) \times x_i \times \text{rate} \\ & \text{else} \\ & \quad w = w - p \times x_i \times \text{rate} \end{aligned}$$

2 垃圾邮件判断流程

首先,在网络层对电子邮件进行检查,检查 IP 地址是否隐匿,检查 HELO 信息、路由信息的真实性。确定了 IP 地址的真实性后,进入 IP 地址白名单、黑名单处理,将白名单放在黑名单之前,首先保证合法邮件的通行。随即对邮件的发送行为进行检查,判断是否属于群发邮件,而且是否达到群发门限,若达到门限则判断为垃圾邮件。

其次是在应用层对邮件体的处理。先检查信体是否加密,若加密则判断为疑似垃圾邮件另作处理。对

未加密的邮件,先检查邮件的信箱是否匿名,若匿名,则判断为疑是垃圾邮件,加标签。随后进行病毒扫描,一经发现病毒即判断为垃圾邮件,之所以将病毒邮件单列出来,是因为病毒的判断较为复杂,而且邮件病毒库相对而言,较为庞大,邮件病毒数据库由系统服务商提供,并且提供实时更新。

若以上过程未完成对垃圾邮件、正常邮件的判断,则对邮件体的处理到此进入内容分析检测,相对而言内容处理速度慢,耗费时间,这里将邮件体处理分为邮件主题、邮件主体两大部分,在垃圾邮件特征数据库的辅助下,进行垃圾邮件判断。

SMTP的一个重要特点是可以可在交互的通信系统中转发邮件。SMTP提供了一种邮件传输的机制,当收件方和发送方在一个网络上,可以把邮件直接传给对方;当双方不在一个网络上,需要通过一个或几个中间服务器转发。SMTP首先由发送方提出请求,与接收方之间建立一个双向通道,通过3次握手实现邮件的传输。但是SMTP服务有一个缺点,就是没有任何的认证,即SMTP服务器无法确认SMTP客户机的合法性,SMTP客户机也无法确认SMTP服务器的合法性,从而导致了用户不经过认证就能发信、用户冒名发信、垃圾邮件的泛滥等^[6]。

依据前面所述的邮件分类方法以及SMTP的传输特性,文中提出了一种高效的垃圾邮件分类模型(如图1所示)。

此模型包括以下几个组成部分:

(1)译码器:用于将收到的邮件进行解码。

(2)E-mail处理器:处理刚刚解码的邮件包,然后抽取其中的特征项,之后这些特征项将会直接送入邮件分类器,以确定邮件的种类。

(3)垃圾邮件分类器:它采用的是基于判别模型的分类器。如果发现垃圾邮件,它会以XML的形式将垃圾邮件相关的信息记录到日志数据库中。作为补充性的功能,它同样可以向本地SMTP服务器发送命令以切断垃圾邮件的传输流。

这样就完成了对邮件的处理,初步分离出垃圾邮件和正常邮件,垃圾邮件加入到垃圾邮件副本和日志数据库,正常邮件加入到正常邮件副本和日志数据库,然后进入人工处理邮件阶段,将误判和漏判的邮件人工剔除。

3 实验结果

3.1 试验方法及评价指标

实验中采用由希腊 Androutsopouto 提供的 Ling - Spam 公共测试语料库^[7],它包含 2983 封邮件,481 封

垃圾邮件,2412 封合法邮件,将所有邮件分为 10 份。采用十次交叉验证方法对过滤算法测试,即将 10 份语料子集中的 9 份作为训练集,另一份作测试集,结果取 10 次测试结果的平均值。一般常用下列三个指标来衡量垃圾邮件过滤算法的性能,其定义如下:

$$\text{召回率:SR} = \frac{S \rightarrow S}{S \rightarrow S + S \rightarrow L}$$

$$\text{正确率:SP} = \frac{S \rightarrow S}{S \rightarrow S + L \rightarrow S}$$

$$\text{精确率:Acc} = \frac{L \rightarrow L + S \rightarrow S}{N_l + N_s}$$

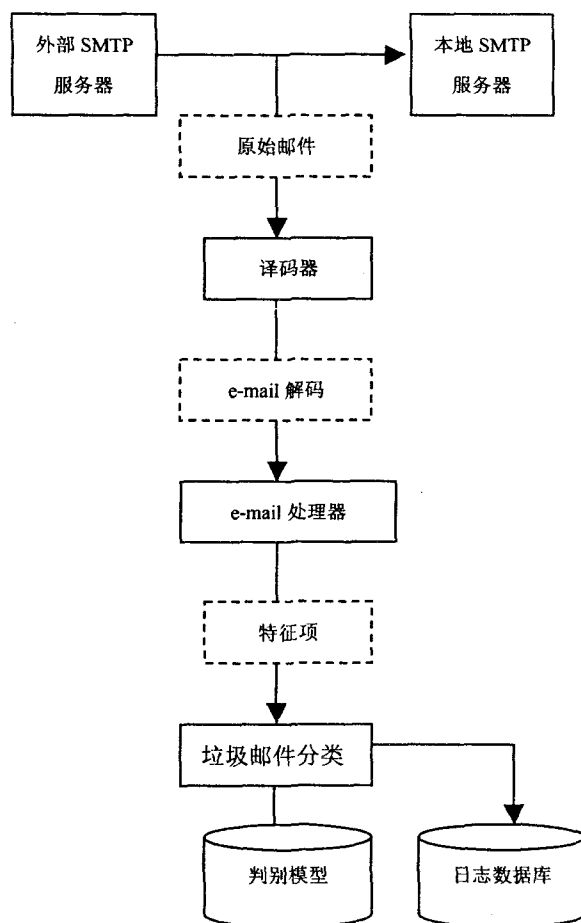


图1 垃圾邮件过滤模型

召回率反映过滤垃圾邮件的能力;正确率又称为垃圾邮件的检对率,值越大说明合法邮件误判为垃圾邮件越少;精确率是指所有邮件的检对率,它的值越大说明邮件误判率越小,也就是被错分的邮件数目越少。公式中, $S \rightarrow S$ 是被正确地分到垃圾邮件数目; $L \rightarrow S$, $S \rightarrow L$ 是被误分进垃圾邮件和合法邮件的邮件数目; N_l 是指所有的合法邮件类的个数; N_s 是指所有垃圾邮件的个数^[8]。

3.2 实验结果

第一个对比实验,采用文中提出的方法(简称

DM)与传统的朴素贝叶斯(Naive Bayes)分类法和支持向量机(SVM)方法在邮件正确率与召回率方面进行比较。

表 1 很直观地显示出基于判别模型的垃圾邮件分类效果要优于 Naive Bayes 与 SVM 分类方法(该实验采用的阈值是 0.9)。

表 1 正确率和召回率对比试验结果

分类方法	SP	SR
Naive Bayes	97.32%	95.08%
SVM	98.21%	96.54%
DM	99.06%	97.40%

第二个对比实验,采用文中提出的方法与 K 近邻算法(KNN)在邮件过滤的精确率方面进行比较,从图 2 可以看出基于判别模型的方法在过滤精度方面要优于 KNN。从算法的复杂度来看,KNN 没有训练过程,原理比较简单,它的过滤速度慢,因此不适用于过滤速度要求较高的场合。

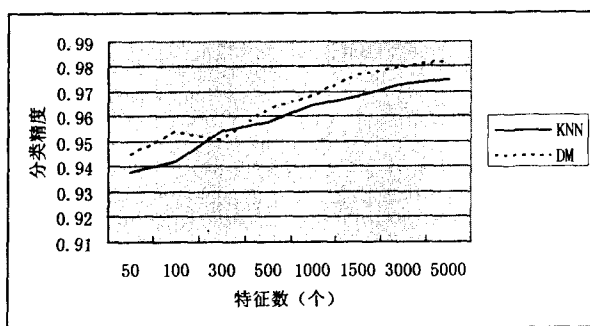


图 2 在不同特征个数下精确率的比较

4 结束语

基于判别方法的垃圾邮件过滤在现代研究中引起比较少的关注,文中提出了一个简单的方法,在某些情况下,它的分类效果要优于其他传统方法。实验很清楚地表明,基于判别模型的方法相对于传统的方法,在垃圾邮件的过滤方面,可以有效地提高正确率和准确率。

参考文献:

- [1] 中国互联网协会反垃圾邮件中心. 年度反垃圾邮件报告 [EB/OL]. 2007-04-05. <http://www.anti-spam.cn/>.
- [2] 潘文锋. 基于内容的垃圾邮件过滤研究[D]. 北京: 中国科学院研究生院, 2004.
- [3] Hulten G, Goodman J. Tutorial on junk email filtering [R/OL]. In ICML 2004; <http://www.research.microsoft.com/~joshuago/tutorialOnJunkMailFilteringjune4.pdf>.
- [4] 张文良, 黄亚楼, 倪维健. 基于差分贡献的垃圾邮件过滤特征选择方法[J]. 计算机工程, 2007, 33(8): 80-82.
- [5] Goodman J, Yih Wen-tau. Online discriminative spam filter training [C]//The Third Conference on Email and Anti-spam (CEAS). California: [s. n.], 2006.
- [6] 马莉, 柴乔林. 基于 Postfix 的垃圾邮件过滤技术的实现[J]. 计算机工程与设计, 2005, 26(4): 999-1001.
- [7] Androutsopoulos I, Paliouras G, Karkalexis V. Learning to filter spam E-mail: A comparison of a naive bayesian and a memory-based approach [C]//The Fourth Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). France: [s. n.], 2000: 1-13.
- [8] 程卫华, 尤晋元. 基于内容过滤的反垃圾邮件系统的设计与实现[J]. 安徽大学学报, 2007, 31(3): 30-33.

(上接第 180 页)

像自适应水印[J]. 计算机仿真, 2006, 23(7): 132-134.

- [5] Honsigner C W. Lossless recovery of an original image containing embedded data. US patent: 6278 791B1 [P]. 2001.
- [6] Jessica F, Goljan M, Du Rui. Invertible Authentication [C]//Proc. SPIE Photonics West, vol. 3971, Security and Watermarking of Multimedia Contents III. San Jose, California: [s. n.], 2001: 197-208.
- [7] Coltue D, Bolon P. Watermarking by Histogram Specification [C]//Proceeding of SPIE. Bellingham: Society of Photo-Optical Instrumentation Engineers, 1999: 252-263.
- [8] Coltue D, Chassery J M, Bolon P. Image Authentication by Exact Histogram Specification [C]//2001 IEEE Fourth Workshop on Multimedia Signal Proceeding. CA: Institute of Electrical and Electronics Engineers Inc, 2001: 701-710.
- [9] Pei S C, Zeng Y C. Hiding Multiple Data in Color Image by Histogram Modification [C]//Proceeding of the 17th International Conference on Pattern Recognition. Piscataway: Institute of Electrical and Electronics Engineers Inc, 2004: 799-802.
- [10] Cheng H, Isnardi M A. Spatial Temporal and Histogram Video Registration for Digital Watermark Detection [C]//Proceeding of 2003 International Conference on Image Processing. Piscataway: Institute of Electrical and Electronics Engineers Computer Society, 2003: 393-396.
- [11] 李妍, 张佑生, 张挺. 一种基于直方图的可逆数字水印算法[J]. 计算机技术与发展, 2006, 16(10): 122-124.