基于贝叶斯公式的垃圾邮件过滤方法*)

詹 川 卢显良 周 旭 侯孟书 袁连海
(电子科技大学计算机科学与工程学院 成都610054)

摘 要 伴随着电子邮件的广泛使用,垃圾邮件泛滥成灾,严重影响了人们正常的学习、工作和生活。本文提出了一种改进的基于贝叶斯公式垃圾邮件过滤技术。我们采用了基于词熵的特征项提取方法,并且使用特征项单词出现频率来表示向量,推导出相应的贝叶斯计算公式。实验表明,我们的方法使垃圾邮件过滤的整体性能都有明显提高。

关键词 贝叶斯公式,垃圾邮件过滤,特征项提取,向量

An Anti-Spam E-mail Filtering Method Based on Bayesian

ZHAN Chuan LU Xian-Liang ZHOU Xu HOU Meng-Shu YUAN Lian-Hai
(College of Computer Science and Engineering, UESTC of China, Chengdu 610054)

Abstract Along with wide application of e-mail nowadays, a large number of spam e-mails flood into people's life and bring catastrophe to their study and life. This paper presents a new Bayesian-based anti-spam e-mail filtering method. We adopt a way of attribute selection based on word entropy, use vectors which are represented by word frequency, and deduce its corresponding Bayesian formula. It is proved that our filter improves total performances apparently in our experiment.

Keywords Bayesian, Anti-spam e-mail filtering, Attribute selection, Vector

1 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的使用,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,严重影响和损害人们的工作、生活和学习。据美国 Bright Mail 公司的报告称,2002年美国人平均收到2200封垃圾邮件,若按垃圾邮件每月增长2%的速度递增,到2007年,这一数字将达到3600封。据英国贸易工业部官员称,垃圾邮件现在占到全球电子邮件流量的40%。更有甚者,据韩国信息保护振兴院统计,韩国国内电子邮件80%为垃圾邮件,其中60%含有淫秽内容。在中国,据中国互联网络信息中心2004年1月公布的第十三次《中国互联网络发展状况统计报告》[1]显示,中国网民平均每周收到13.7封电子邮件,其中垃圾邮件占据了7.9封,垃圾邮件数量超过了正常邮件数量。美国是受垃圾邮件危害最严重的国家,一年由于垃圾邮件给企业带来的损失高达90亿美元,而中国仅次于美国,排在第二,中国网民一年收到的垃圾邮件总数为460亿封,浪费处理在垃圾邮件的时间为15亿小时,2003年垃圾邮件浪费中国的 GDP 高达48亿元[2]。垃圾邮件的肆虐使得电子邮件系统本身的存在受到严重挑战,严重影响了电子邮件的健康发展。

关于垃圾邮件,《中国互联网协会反垃圾邮件规范》给出其定义,以下4种情况属于垃圾邮件:

1、收件人事先没有提出要求或者同意接受的广告、电子刊物、各种形式的宣传品等宣传性的电子邮件;

2、收件人无法拒收的电子邮件;

3、隐藏发件人身份、地址、标题等信息的电子邮件;

4、含有虚假的信息源、发件人、路由等信息的电子邮件。

在当前声势浩大的反垃圾邮件运动中,许多国家出台了反垃圾邮件法,如美国、日本、欧洲等,中国也在今年的全国政协十届二次会议提出了加快"反垃圾邮件立法"进程的提案。美国的 AOL 通过对垃圾邮件的发送者起诉,对控制垃圾邮件起到一定作用,但是据美国业界官员介绍通过立法来反垃圾邮件收效甚微,因为这些垃圾邮件发送者可通过国外来转发。据美国 Brightmail 公司统计,美国2004年2月收到的垃圾邮件已占总数的62%。因此反垃圾邮件不能光依靠立法,还要依靠技术手段。目前国内大部分邮件服务提供商都提供了一些简单的垃圾邮件过滤功能,如设置简单的规则,配置黑名单等等,但功能简单,其效果不太理想。

在智能过滤垃圾邮件方面,Sahami[3]等人提出采用机器学习方法来进行处理。他们采用二进制来表示邮件特征向量,通过特征属性的互信息量来提取特征项,用贝叶斯公式来计算邮件是垃圾邮件的概率来识别邮件。一些实验[4]证明用贝叶斯公式来进行垃圾邮件识别相当有效。本文提出一种改进过的基于贝叶斯原理的垃圾邮件过滤方法,采用了基于词熵的特征项提取方法,使用特征项单词出现频率来表示特征向量,则其对应的垃圾邮件过滤方法具有更高的垃圾邮件识别的准确性和查全性。本文第1节为引言,第2节介绍使用的邮件样本和相关数据的处理,第3节为本文重点,详细描述了贝叶斯分类的公式和原理,第4节为实验结果,最后是本文的结束语和工作展望。

2 邮件样本及数据预处理

2.1 邮件样本的收集

*)信息产业部电工业生产发展基金资助项目,编号:[2002]11006。詹 川 博士研究生,主要研究方向:计算机网络,信息安全,人工智能。卢显良 教授,博士生导师,主要研究方向:计算机网络,分布式系统,信息安全。周 旭 博士研究生。侯孟书 博士研究生。袁连海,博士研究生。

本次试验采用的训练和测试邮件样本来自于 http://www.spamassassin.org/publiccorpus。它是一个用于垃圾邮件过滤研究的公开可用的邮件样本。我们用的版本为20030228样本集，我们按中国目前的垃圾邮件大致比例从中选取了1000封邮件，其中，580封垃圾邮件，420封正常邮件。这些邮件全是英文邮件，其中的 html 标记和附件都已被除去。

2.2 邮件样本的预处理

对于邮件样本中的第 i 邮件我们用 e_i 来表示，则 e_i 对应的特征向量为：

$\vec{x}_{e_i}=(x_{1_i},x_{2_i},\cdots,x_{n_i})$，其中 $x_{1_i},x_{2_i},\cdots,x_{n_i}$ 分别为邮件 e_i 对应 $X_1,X_2,\cdots,X_n$ 特征项的特征值。

对于特征值的取值，在已有的垃圾邮件过滤算法的做法是采用二进制表示法，即当某一特征项 X_j 在邮件 e_i 中出现，则特征值 $x_j=1$，否则 $x_j=0$。这种表示使运算简单，但是没有充分表达出特征项在邮件中出现频率的信息。本算法中采用的是用特征项出现频率来表示邮件特征向量的方法。频率分为绝对频率和相对频率，相对频率为归一化的频率。此处我们采用的是绝对频率来表示，即 $N(X_j,e_i)$ 表示在 e_i 邮件中特征项 X_j 出现的次数。于是我们试验中的邮件对应向量表示为如下形式：

$$\vec{x}_{e_i}=(N(X_1,e_i),N(X_2,e_i),\cdots,N(X_n,e_i)) \tag{1}$$

特征向量的特征项可以选取邮件中的单词（如：price、adult、shop），短语（如：on sale、be over 21）以及非文本属性（如：是否带有附件，是否为 html 格式）。由于收集的邮件样本经过处理，都不带附件和 html 标志，且为了把注意力放在我们的算法上，让测试程序简单，因此本次试验全部选取以单词作为邮件的特征项。

2.3 特征项的提取

邮件由很多不同的单词组成，如果把所有的单词都作为特征项，则特征向量的维数相当大，其中有些单词对区分正常或垃圾邮件所起的贡献很小，完全可以忽略。因此进行相应的维数消减，不但可以大大提高程序的运行效率，而且几乎不会对识别效果有影响。在此我们提取特征项的方法是基于词熵的选取方法。其原理是不同的单词对邮件识别的贡献大小是不同的，单词具有的信息增益越大，则其对邮件分类的能力就越大。信息增益的计算公式如下：

设 S 为邮件样本数据集的总数，由正常邮件和垃圾邮件两类构成，其中，S_{legit} 为样本集中正常邮件数；S_{spam} 为样本集中垃圾邮件数；则对邮件样本集 S 进行分类所需要的信息熵为：

$$I(S_{\text{legit}},S_{\text{spam}})=-\frac{S_{legit}}{S}\log_2\frac{S_{legit}}{S}-\frac{S_{spam}}{S}\log_2\frac{S_{spam}}{S} \tag{2}$$

然后计算单词 X_j 对邮件分类所提供的信息熵 $E(X_j)$，在此处我们设定属性 X_j 取值为{0,1}。当单词 X_j 在邮件中出现时，为1；否则，为0。

$$E(X_j)=\frac{S_{x_j=0}}{S}I(S_{legit,x_j=0},S_{spam,x_j=0})+\frac{S_{x_j=1}}{S}I(S_{legit,x_j=1},S_{spam,x_j=1}) \tag{3}$$

其中，$S_{x_j=0,1}$ 分别表示样本邮件 S 中不含有或含有单词 X_j 的邮件数；$S_{legit,x_j=0,1}$ 分别表示正常邮件中不含或含有单词 X_j 的邮件数；$S_{spam,x_j=0,1}$ 分别表示垃圾邮件中不含或含有单词 X_j 的邮件数；则单词 X_j 对邮件分类所产生的信息增益为：

$$\text{Gain}(X_j)=I(S_{\text{legit}},S_{\text{spam}})-E(X_j) \tag{4}$$

提取特征项的具体步骤如下：

1. 从邮件样本中提取所有单词，除去停用词（如 a，an，the，of，数字等等），再对单词的不同形式如过去分词、动名词、动词过去式、名词复数、形容词比较级等进行标准化。本实验中采用的是停用词表和标准化方式都是参考的 ifile[5]源程序。

2. 计算每个单词对邮件分类所产生的信息增益 Gain (X_j)；

3. 按 Gain(X_j) 的大小，由大到小进行排序。我们注意到式(4)中对于固定的样本集其 $I(S_{\text{legit}},S_{\text{spam}})$ 是个固定的值，则 $E(X_j)$ 越小，Gain(X_j) 值越大。Gain(X_j) 由大到小的排序对应的是 $E(X_j)$ 由小到大的排序。实际我们只需求出 $E(X_j)$，按由小到大排序；

4. 按排序，从前抽取一定数量对邮件分类信息增益大的单词作为特征项，根据试验[6]表明维数取在100左右 Bayes 算法具有最佳的效果，因此本次试验选定的向量维数为100；

5. 把样本中的所有邮件，根据选定的特征项，表示成相应的特征向量。

3 贝叶斯邮件过滤器

3.1 贝叶斯公式

根据贝叶斯公式和全概率公式，某封邮件 e_i 其特征向量为：

$\vec{x}_{e_i}=(x_1,x_2,\cdots,x_n)$

则邮件 e_i 属于类 C_k（本试验中邮件分为正常邮件和垃圾邮件两类：正常邮件用 C_{legit} 表示，垃圾邮件用 C_{spam} 表示）的概率为：

$$P(C_k|\vec{x}_{e_i})=\frac{P(C_k)P(\vec{x}_{e_i}|C_k)}{\sum_{k\in\{legit,spam\}}P(C_k)P(\vec{x}_{e_i}|C_k)} \tag{5}$$

在实际计算中，$P(\vec{x}_{e_i}|C_k)$ 的概率是不容易求得的。为了大致估计出 $P(\vec{x}_{e_i}|C_k)$ 的值，我们假设特征向量中的各个特征项 $X_1,X_2,\cdots,X_n$ 之间是相互独立的，即某一特征单词在邮件中出现的事件独立于另一个特征单词在邮件中出现的事件。则邮件 e_i 属于类 C_k 的概率可表示为：

$$P(C_k|e_i)=\frac{P(C_k)\prod_{l=1}^{n}P(x_{l_i}|C_k)}{\sum_{k\in\{legit,spam\}}P(C_k)P(\vec{x}_{e_i}|C_k)} \tag{6}$$

尽管对特征单词间相互独立的假设过于简单，但是大量试验[7,8]证明用此假设构造的贝叶斯分类器确实相当有效。由于我们使用的是各单词在邮件 e_i 中出现的频率作为向量 $\vec{x}_{e_i}$ 中各特征项的值，则式(6)改写为：

$$P(C_k|\vec{x}_{e_i})=\frac{P(C_k)\prod_{l=1}^{n}P(x_l|C_k)^{N(X_l,e_i)}}{\sum_{k\in\{legit,spam\}}P(C_k)P(X_l|C_k)^{N(X_l,e_i)}} \tag{7}$$

其中，$P(C_k)$ 可以通过估算样本中 C_k 类的邮件占总样本邮件的比例来获得，

$$P(C_k)=\frac{S_{C_k}}{S} \tag{8}$$

$N(X_l,e_i)$ 为单词 X_l 在邮件 e_i 出现的频率，$P(X_l|C_k)$ 为单词 X_l 在 C_k 类的邮件中出现的概率，

$$P(X_l|C_k)=\frac{1+\sum_{i=1}^{|C_k|}N(X_l,e_i)}{n+\sum_{l=1}^{n}\sum_{i=1}^{|C_k|}N(X_l,e_i)} \tag{9}$$

式(9)中的 n 为向量的维数，即特征项单词的个数，$|C_k|$ 为 C_k 类的邮件的集合。

由贝叶斯公式构造的邮件过滤器可能存在两类错误：一类是把正常邮件误当作垃圾邮件，另一类是把垃圾邮件误识别为正常邮件，在这两种错误中，第一类错误给我们造成的损失更大，为了更加谨慎、准确地识别垃圾邮件，引入公式(10)

$$\frac{P(C_{spam}|\vec{x}_{ei})}{P(C_{legit}|\vec{x}_{ei})}>\lambda \tag{10}$$

表示邮件 e_i 是垃圾邮件的概率大于正常邮件的 λ 倍时，将其判断为垃圾邮件。当 λ 越大，其为垃圾邮件的可能性越大。在 Ion Androutsopoulos 等人的试验[6]中，当 $\lambda=999$ 时，分类器识别垃圾邮件的准确率高达100%。

3.2 评价指标

为了有效评价垃圾邮件过滤器的好坏，我们使用两个评定指标：SP(垃圾邮件识别的准确率)和 SR(垃圾邮件识别的查全率)

$$SP=\frac{n_{spam\to spam}}{n_{spam\to spam}+n_{legit\to spam}} \tag{11}$$

$n_{spam\to spam}$ 为正确识别出的垃圾邮件数，$n_{legit\to spam}$ 为正常邮件被误识别为垃圾邮件数。

$$SR=\frac{n_{spam\to spam}}{N_{spam}} \tag{12}$$

N_{spam} 为样本邮件中垃圾邮件的总数。

SP 和 SR 反映的是邮件过滤器质量的两个方面，为了综合考虑邮件过滤器的性能，我们引入一个新的评估指标 $F1$，它综合考虑了准确率和查全率两方面的指标，其数学公式如下：

$$F1=\frac{SP\times SR\times 2}{SP+SR} \tag{13}$$

4 实验及结果

我们采用10次交叉验证方法，把邮件样本 S 集随机地分为10个互不相交的子集 $S_1,S_2,\cdots,S_{10}$，每个子集大小相等。对邮件样本学习和测试分别进行10次，在第 i 次循环中，子集 S_i 作为测试集，其余9个子集合并构成一个大的训练数据集。通过学习获得相应的分类器，然后用对应的测试集对其进行验证。计算出每次分类器的准确率和查全率，最后求10次结果的平均值。这样避免了试验的随机性和偶然性。

在试验中，我们比较了邮件向量不同的表示方法对邮件过滤器效果的影响，一种是基于用特征单词是否在邮件中出现的二进制表示法，一种是用特征单词在邮件出现的次数表示的方法，对比见表1，同时我们还给出了在特征向量用单词出现次数的表示情况下，λ取不同值对邮件过滤器性能的影响，比较见表2。

表1

$SP(\%)$	$SR(\%)$	$F1(\%)$	
基于特征单词是否在邮件中出现	96.82	83.25	89.52
基于特征单词在邮件中出现的次数	98.65	86.48	92.16

表2

$SP(\%)$	$SR(\%)$	$F1(\%)$	
$\lambda=1$	98.65	86.48	92.16
$\lambda=9$	99.26	84.59	91.34

从表1可看出，基于单词在邮件中出现的次数的向量表示法在各方面的评价指标都明显好于基于单词是否在邮件出现的二进制向量表示法，因为它提供了很多的信息量给邮件过滤器，以至使其整体性能得到上升，在表2中，当 λ 越大，其识别垃圾邮件的准确率越高，但是付出了查全率下降的代价，F1稍微有所下降。

结束语及以后工作 本文应用贝叶斯公式进行垃圾邮件过滤，使用基于词熵的方法来提取特征项，并用特征项单词在邮件中出现的次数来表示特征向量，试验表明这种方法在整体上提高了邮件过滤器的性能。我们以后的工作：

1. 目前我们提取的特征项都是基于单词的，以后应该把基于短语和非文本属性的因素考虑进来，这样会进一步提高过滤器的准确率。

2. 目前我们的样本都是英文的，我们将自己收集中文邮件样本，以后进行关于中文垃圾邮件过滤的研究。

3. 进一步研究在垃圾邮件过滤上的其他先进算法。

参 考 文 献

1 中国互联网络信息中心. 第十三次《中国互联网络发展状况统计报告》. 2004,1

2 上海艾瑞市场咨询公司. 中国反垃圾邮件市场研究报告. 2003,11

3 Sahami M, Dumais S, et al. A Bayesian Approach to Filtering Junk E-Mail. Learing for Text Categorization -Papers from the AAAI Workshop, Madison Wisconsin, 1998

4 Chen Duhong, Tongjie, et al. Spam Email Filter Using Naive Bayesian, Decision Tree, Neural Network and AdaBoost. http://www.cs.iastate.edu/~tongjie/spamfilter/paper.pdf

5 http://www.ai.mit.edu/~jrennie/ifile/

6 Androutsopoulos I, Paliouras G, et al. Learning to filter spam e-mail : a comparison of a naive Bayesian and a memory-based approach. In: Proc. of the workshop "Machine Learning and Textual Information Access", 4th European Conf. on PKDD-2000, Lyon, France, Sep. 2000

7 Langley P, Wayne I, Thompson K. An Analysis of Bayesian Classifiers. In: Proc. of the 10th National Conf. on Artificial Intelligence, San Jose, California, 1992

8 Domingos P, Pazzani M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning, 1997, 29: 103~130