第 25 卷 第 2 期
2010 年 4 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.25 No. 2
Apr. 2010

基于内容的垃圾电子邮件过滤技术研究

马建斌[1]，薛博洋[2]
（1. 河北农业大学信息科学与技术学院，保定 071001；2. 河北农业大学现代科技学院，保定 071001）

摘 要：提出一种过滤垃圾电子邮件的方法. 通过 tf-idf 特征提取方法提取邮件的词汇特征,采用 χ^2 特征选择方法选取有效的特征,并抽取几个具有明显区分能力的结构方面的特征,利用支持向量机算法对垃圾电子邮件进行自动过滤. 对中科院中文垃圾邮件语料库(Cspam)的实验,识别正确率达到 82%以上,另外,tf-idf 词汇特征和结构特征搭配使用可以提高分类的正确率,表明此种方法能提高垃圾电子邮件过滤的准确性.

关键词：内容；垃圾电子邮件过滤；tf-idf；结构特征；支持向量机
中图分类号：TP393.098 **文献标志码**：A **文章编号**：1672-6510(2010)02-0072-04

Research on Technology of Spam Filtering Based on Contents

MA Jian-bin[1], XUE Bo-yang[2]
(1. College of Information Science and Technology, Agricultural University of Hebei, Baoding 071001, China;
2.College of Modern Science and Technology, Agricultural University of Hebei, Baoding 071001, China)

Abstract: One method to filter spam was proposed. The tf-idf method was used to extract e-mail's lexical features. χ^2 method was used to select effective features. The several structural features were extracted which could discriminate spam obviously. The support vector machine algorithm was adopted to filter spam automatically. By experimenting on dataset of Cspam, the evaluation value F is above 82%, the tf-idf lexical features and structural features combined can improve the classification accuracy, which proves that the method can approve the accuracy of filtering spam.

Keywords: contents; spam filtering; tf-idf; structural features; support vector machine

随着网上垃圾电子邮件现象日益严重,研究垃圾邮件过滤技术非常必要. 白名单与黑名单技术、规则过滤[1]及基于关键词匹配的内容扫描为常见的垃圾邮件过滤方法,黑白名单方法阻止可疑邮件地址用户发送邮件,但是,多人共用一个邮箱地址的情况比较常见,发送者在不断地变化,黑白名单方法有局限性. 规则方法就是人工制定一些规则,但是,这些规则需要人们不断发现和总结、更新,而且,还带有一定的主观性,准确率受到限制. 对电子邮件的内容进行分析,识别出垃圾电子邮件,无疑是一种较可靠的方法. k 近邻(k-Nearest Neighbor)[2]、贝叶斯分类器(Bayesian classifiers)[3–5]等机器学习算法都可用来识别垃圾电子邮件,支持向量机(Support Vector Machine,SVM)算法因其适合处理高维特征,训练速度快等优点在文本分类领域取得很大的成功,已经证明在垃圾邮件过滤领域效果较好[6–7]. 目前,大都采用 tf-idf 方法提取电子邮件词汇方面的特征,但是这种特征提取方法是不全面的,垃圾邮件和正常邮件在句子长度、段落长度等结构特征方面具有明显的区别. 所以,本文利用 tf-idf 特征提取方法广泛地抽取电子邮件的词汇方面特征,采用 χ^2 统计量特征选择方法对抽取到的特征进行必要地取舍,并提取几个具有明显区分能力的结构特征,采用支持向量机算法自动过滤掉垃圾电子邮件.

1 研究方法

过滤电子邮件实质上是二类分类的问题,垃圾邮

收稿日期：2009–08–26；**修回日期**：2010–01–05
作者简介：马建斌（1978—），男，河北临城人，讲师，博士研究生，majianbin@hebau.edu.cn.

件与正常邮件的不同之处就在于它表现出某种特征，比如某些词语出现的频率不同，如果能把这些特征抽取出来，并采用支持向量机分类算法识别这些特征，训练成一个自动分类器，就可以通过此分类器自动识别一电子邮件是否是垃圾邮件. 所以，可以把垃圾邮件过滤问题看成是机器学习问题，其基本原理如图 1 所示.

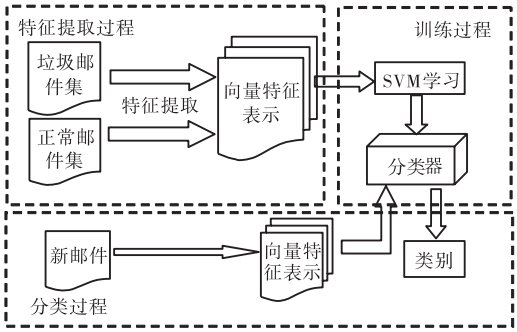

图 1　垃圾邮件过滤原理图

Fig.1　Principle of spam filtering

(1)特征提取

因为邮件文档具有半结构化的特点，计算机不能直接处理原始邮件，应把邮件转化为固定格式的结构化数据，也就是抽取正常邮件区别于垃圾邮件的特征，把邮件文档转化为空间向量模型的形式，可以被支持向量机分类器所接受.

中文与英文等其他语言不同，词与词之间没有边界，所以，从特定文本里提取关键词较困难，为了抽取代表文本特征的特征项，需要借助分词和词性标注工具对中文邮件的正文进行分词和词性标注处理. 现在的中文自动分词工具为中文信息处理提供可能，本研究采用厦门大学史晓东教授所开发的中文自动分词和标注工具 segtag 作为分词工具.

(2)训练过程

训练过程即通过机器学习算法对特征提取过程所提取的特征进行学习，形成分类器.

(3)分类过程

通过训练过程构造的分类器自动辨别某邮件是正常邮件还是垃圾邮件，如果是垃圾邮件，系统自动将其过滤掉.

2　支持向量机算法

支持向量机是由 Vapnik 等人根据统计学习理论导出的结构风险最小化原则基础上的机器学习算法[8]. 针对两类分类问题其主要思想是，在高维空间中寻找一个超平面作为两类的分割，以保证最小的分类错误率.

SVM 是从线性可分情况下的最优分类面发展而来的，基本思想见图 2. 分割线 1 和分割线 2 都能正确地将两类样本分开，有无数条这样的分割线，但分割线 1 为最优分类线(更高维即为最优分类面或最优超平面)，能使两类样本的间隙最大.

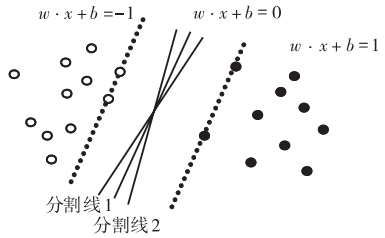

图 2　线性二类划分的最优超平面

Fig.2　Class linear separable optimization hyperplane

设线性可分训练集 $(x_1,y_1),\cdots,(x_l,y_l)$，其中，$x_i\in R^n, y_i=\in\{-1,+1\}$，是类别标号，$l$ 为样本数. n 维空间中线性判别函数的一般形式为 $g(x)=w\cdot x+b$，分类面的方程为 $w\cdot x+b=0$. 将判别函数归一化，等比例调节 w 和 b，使两类所有样本都满足 $|g(x)|\geqslant 1$，这样，分类间隔就等于 $2/\|w\|$，因此，求两类样本的间隔最大变为求 $\|w\|$ 最小.

其中，满足 $|g(x)|=1$ 的样本点，离分类线(平面)最近，它们决定了最优分类线(平面)，称之为支持向量.

可见，求最优分类面的问题转化为优化问题：

$$\min\varphi(w)=\frac{1}{2}\|w\|^2=\frac{1}{2}(W\cdot W) \tag{1}$$

约束条件为

$$y_i(w\cdot x_i+b)-1\geqslant 0, i=1,2,\cdots,l \tag{2}$$

本优化问题可以转化为

$$\max M(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j x_i\cdot x_j \tag{3}$$

约束条件为

$$\sum_{i=1}^{l}\alpha_i y_i=0,\alpha_i\geqslant 0 \tag{4}$$

通过求解，可得最优分类函数为

$$f(x)=\mathrm{sgn}(w\cdot x+b)=\mathrm{sgn}(\sum_{i=1}^{N_s}\alpha_i y_i x_i\cdot x+b) \tag{5}$$

其中，N_s 为支持向量个数.

对于线性不可分问题，Vapnik 引入了核空间理论：将低维的输入空间数据通过非线性映射函数映射到高维属性空间，这种非线性映射函数被称之为核函

数. 通过核函数,线性不可分问题将转化为线性可分问题.

3　特征提取方法

3.1　词汇特征提取方法

如何判别出一个邮件文档是垃圾邮件还是正常邮件,特征提取是其中的关键技术. 字、词或词组都可用来作为特征项,根据实验结果,普遍认为选取词要优于字或词组. 本文根据邮件正文文本的分词结果,通过 tf-idf 公式计算词的权重:

$$w(t,\vec{d})=\frac{t\times f(t,\vec{d})\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in\vec{d}}[t\times f(t,\vec{d})\times\log(N/n_t+0.01)]^2}} \quad (6)$$

式中: $w(t,\vec{d})$ 为词 t 在文本 $\vec{d}$ 中的权重; $f(t,\vec{d})$ 为词 t 在文本 $\vec{d}$ 中的词频; N 为训练文本的总数; n_t 为训练文本集中出现 t 的文本数.

3.2　词汇特征选择方法

训练集中包含了大量的词汇,并不是所有词汇都与类别有关,另外,过多的词汇会造成向量维数过大,计算机存储空间过大,处理速度慢. 因此,为了降低向量的维数要先对文本进行预处理,去掉的对分类用处不大的常用词,然后采用某种特征选择方法对所有的词排序,将排在前面的一定数量的词作为特征. 本文采用 χ^2 统计量特征选择方法.

χ^2 统计中 χ^2 用来度量词和类别之间独立性:

$$\chi^2(w,c_j)=\frac{N\times(AD-CB)^2}{(A+B)\times(B+D)\times(A+B)\times(C+D)} \quad (7)$$

式中: A 是 c_j 类中包含词 w 的文本数目; B 是不属于 c_j 类,但包含词 w 的文本数目; C 是 c_j 类中不包含词 w 的文本数目; D 是不属于 c_j 类,也不包含词 w 的文本数目; N 是文本总数. χ^2 统计为

$$\chi^2(w)=\sum_{j=1}^{K}P(c_j)\chi^2(w,c_j) \quad (8)$$

χ^2 越大,独立性越小,相关性越大.

3.3　结构特征提取方法

垃圾邮件通常包含广告信息,发送者为了充分展示其商品信息,往往包含大量文字,书面用语较多,句式较整齐,段落和句子较多,但是,正常电子邮件写作比较随意,篇幅较小,段落和句子较少,段落和句子中包含的字数也较少,这就是很明显的结构特征. 经过实验,抽取了 8 个具有明显区分能力的结构特征,见表 1.

表 1　结构特征提取方法

Tab.1　Structural feature extraction method

编号	特征及提取方法
1	邮件中包含的字数
2	邮件中包含的段落数
3	邮件中包含的句子数
4	平均句子长度 = 总字数/总句子数
5	平均段落长度 = 总字数/总段落数
6	空格出现的比率 = 总空格数/总字数
7	空行出现的比率 = 空行数/总行数
8	段落缩进数比率 = 段落缩进空格数/总字数

4　实　验

为了测试本文所提出方法,采用中科院计算所收集的中文垃圾邮件语料库(CSpam)[9]进行实验. 首先将中文邮件切分成词,然后进行特征提取,电子邮件中的特征通过向量空间模型映射到一个高维的空间中. 抽取垃圾邮件和正常邮件各 500 个进行训练,采用 3 交叉评估方法评价分类识别结果,即随机将样本集合分成 3 份,其中任意 2 份组合后作为训练集,剩余 1 份做测试集,共进行 3 次测试. 将 3 次测试的平均值作为最终结果,采用支持向量机 Libsvm-2.9 做分类算法,采用正确率指标来评价垃圾电子邮件过滤的性能,选取 1 000 个词汇特征和 8 个结构方面特征,支持向量机核函数选择线性核函数. 分别采用词汇特征和词汇特征与结构特征搭配使用的实验结果见表 2.

表 2　垃圾邮件过滤实验结果

Tab.2　Experimental results of spam filtering

特征	正确率/%
词汇特征(tf-idf)	82.68
词汇特征 + 结构特征	93.10

从表 2 可以看出,仅通过 tf-idf 方法提取出的词汇特征,垃圾邮件过滤的正确率为 82.68%,增加 8 个结构特征以后,正确率明显提高,达到 93.10%. 可以看出,本文方法可以明显地提高垃圾电子邮件过滤的正确率.

为了测试支持向量机算法不同核函数对分类识别的影响,考察不同核函数分类识别效果,选取词汇特征与结构特征搭配作为特征,多项式核函数的 degree 参数选择 1,高斯径向基核函数的 Gamma 参数选择 0.01,结果如见表 3.

表 3　不同核函数的分类识别结果

Tab.3　Classification results of different kernel function

核函数	正确率/%
线性核函数	93.10
多项式核函数	50.25
高斯径向基核函数	61.16

通过实验可以发现，支持向量机采用线性核函数具有较高的识别能力，该实验结果与文本分类的实验结果相同，表明支持向量机算法较适合垃圾电子邮件过滤识别.

5　结　语

本文提出一种垃圾邮件过滤的方法，利用 tf-idf 特征提取方法提取垃圾电子邮件的词汇方面的特征，采用 χ^2 统计量特征选择方法对抽取到的特征进行必要地取舍，并提取几个具有明显区分能力的结构特征，根据提取的特征，利用支持向量机算法对垃圾电子邮件进行自动过滤.

基于中科院中文垃圾邮件语料库（Cspam）的实验正确率达到 82%以上. 实验表明，词汇特征和结构特征搭配使用的垃圾电子邮件过滤的正确率较高，表明本文所提出的方法能明显地提高垃圾电子邮件过滤的性能.

参考文献：

[1] Cohen W W. Learning rules that classify e-mail[C]// Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access. Menlo Park：AAAI Press，1996：18–25.

[2] 田泽，颜松远，徐敬东. 基于改进 K 近邻的垃圾邮件过滤技术[J]. 计算机工程与应用，2007，43（25）：178–181.

[3] Sahami M，Dumais S，Hecherman D，et al. A bayesian approach to filtering junk email[C]// Proceedings of the AAAI Workshop on Learning for Text Categorization. Menlo Park：AAAI Press，1998：55–62.

[4] 宁绍军，邹恒明. 基于贝叶斯公式的自适应垃圾邮件过滤方法[J]. 计算机应用与软件，2007，24（11）：189–191.

[5] 李雯，刘培玉. 基于贝叶斯的垃圾邮件过滤算法的研究[J]. 计算机工程与应用，2007，43（23）：174–176.

[6] 王清翔，广凯，潘金贵. 基于支持向量机的邮件过滤[J]. 计算机科学，2007，34（9）：93–94.

[7] 王文剑，侯岩. 一种基于 SVM 的中文电子邮件过滤方法[J]. 山西大学学报：自然科学版，2007，30（3）：303–309.

[8] Cortes C，Vapnik V. Support-vector networks[J]. Machine Learning，1995，20（3）：273–297.

[9] 王斌. 中文垃圾邮件语料库 Cspam[EB/OL]. （2007–07–05）[2009–12–20]. http：//www. nlp. org. cn/categories/default. php?cat_id = 17.

（上接第 40 页）

[2] Jensen W L. Production of thionophosphorus：US，2662917[P]. 1953–10–15.

[3] Imashiro Y，Horie N，Yamane T. Process for producing 3-methyl-1-phenylphospholene oxide：US，5488170[P]. 1996–01–30.

[4] 徐元清，刘治国，王彦林，等. 苯基二氯化膦的合成工艺改进[J]. 化学世界，2001，42（12）：655–656.

[5] Wasserscheid P，Welton T. Ionic Liquids in Synthesis[M]. Weinheim：WILEY-VCH Verlag，2003.

[6] 李汝雄. 绿色溶剂——离子液体的合成与应用[M]. 北京：化学工业出版社，2004：56–70.

[7] 王忠卫，高亮，高军，等. [1，2-bisPyEt]Cl_2-$XAlCl_3$ 离子液体催化合成苯基二氯化膦方法研究[J]. 山东科技大学学报：自然科学版，2007，26（2）：57–59.

[8] 戈鹏，李巧玲. $AlCl_3$ 催化合成苯基二氯化膦及其反应机理的探讨[J]. 工业催化，2007，15（2）：44–46.

[9] 谢文杰，王鉴，曾群，等. 二氯化苯基膦的合成进展[J]. 精细化工中间体，2005，35（6）：19–21.