基于粗糙集的加权朴素贝叶斯邮件过滤方法

邓维斌[1,2] 王国胤[2] 洪智勇[2]
(重庆邮电大学电子商务与现代物流重点实验室 重庆 400065)[1]
(西南交通大学信息科学与技术学院 成都 610031)[2]

摘 要 邮件过滤中有两个关键问题,一是如何选择有效的邮件特征集,二是设计较好的邮件过滤算法。在对邮件特性进行分析的基础上,综合邮件头及邮件内容的主要形象特征给出了一种新的邮件特征集提取方法。用粗糙集的信息观点度量了各属性的重要性,并以此为权重进行加权朴素贝叶斯垃圾邮件过滤,有效地解决了朴素贝叶斯分类中的条件依赖性问题。通过在中英文邮件集上的测试实验,证明了所提出的邮件过滤方法的有效性。

关键词 垃圾邮件过滤,特征选择,粗糙集,加权朴素贝叶斯
中图法分类号 TP391 **文献标识码** A

Weighted Naïve Bayes Spam Filtering Method Based on Rough Set

DENG Wei-bin[1,2] WANG Guo-yin[2] HONG Zhi-yong[2]
(Key Lab of Electronic Commerce and Modern Logistics of CQUPT, Chongqing 400065, China)[1]
(School of Information Science & Technology, Southwest Jiaotong University, Chengdu 610031, China)[2]

Abstract Using a classifier based on a specific machine-learning technique to automatically filter out spam email has drawn many researchers' attention. In a spam filtering process, how to selecting the features of emails and how to design a good filtering algorithm are two key issues. A new method of features selecting was proposed, which include the head and the other main features of emails. Furthermore, the features' importance degree was measured according to information viewpoint of rough set. With it, a new weighted naïve Bayes spam filtering was put forward. It can solve the conditional dependence of naïve Bayes efficiently. Simulation results on two email data sets in English and Chinese respectively illustrate the efficiency of this method.

Keywords Spam filtering, Feature selecting, Rough set, Weighted naïve Bayes

1 引言

随着网络技术应用范围日益广泛,电子邮件正被越来越多的人所使用。然而,电子邮件在给人们带来极大便利的同时,垃圾邮件也在越来越频繁地侵袭着人们的工作和生活。因此,垃圾邮件的过滤已受到越来越多研究人员的注意,已成为信息安全领域研究的一个重要方向[1,2]。垃圾邮件的过滤主要有基于统计的方法和基于规则的方法[1]。基于统计方法主要有贝叶斯[3,4]和SVM[5]等,基于规则的方法主要有决策树和粗糙集[6]等。当然还有许多学者提出了基于n-gram等其它的过滤方法[7]。垃圾邮件过滤中一个关键步骤是进行特征提取,很多过滤方法是基于邮件内容的文本分类方法[1,2,4]。邮件内容特征反映了邮件的内容主体,是邮件分类的一个重要依据。根据邮件的文本内容短、所提取特征词的维数较高等特点,Lai Chih-Chin在文献[2]中通过实验证实了基于邮件内容的过滤方法在很大程度上影响过滤的时间效率和分类的准确性。此外,重庆邮电大学的李志君等提出了根据邮件头进行垃圾邮件过滤[6],邮件头反映了邮件在网络中传送的一些状态特征,而且根据邮件头提取的特征数量少,分类效率高。但是,如果垃圾邮件发送者经过巧妙伪装,则邮件头的特征信息会在很大程度上失真。如何将邮件头和邮件内容特征相结合,使其既能反映邮件传输过程中的特性,又能在保持较低的特征维数的情况下较好地体现邮件内容的主要形象特性,这是一个倍受研究人员关注但目前还没有很好解决的问题。

M. Sahami等在文献[3]中最早提出用朴素贝叶斯(Naïve Bayes, NB)方法进行邮件过滤,此后朴素贝叶斯成为了垃圾邮件过滤的主要方法。但它基于一个简单的假定,即在给定分类特征条件下属性值之间是相互独立的。在现实世界中,这种独立性假设经常是不满足的,故很多学者研究如何改进朴素贝叶斯的分类性能。由于粗糙集(Rough Set, RS)能有效处理不精确、不一致及不完备信息,邓维斌等提出了基于Rough Set的加权朴素贝叶斯分类算法[9],较好地克服了朴素贝叶斯分类中的条件独立性假设问题。为此,本文将邮件头及内容特性相结合提取邮件特征集,并根据粗糙集的信息观点度量各属性的重要性,再用加权朴素贝叶斯方法进行邮

到稿日期:2010-03-24 返修日期:2010-08-04 本文受国家自然科学基金(60773113),重庆市自然科学基金重点项目(2008BA2017),重庆邮电大学自然科学基金(A2008-38)资助。
邓维斌(1978—),男,博士生,讲师,主要研究方向为智能信息处理、电子商务,E-mail:dengwb@cqupt.edu.cn;**王国胤**(1970—),男,博士,教授,博士生导师,主要研究方向为Rough集理论、神经网络、机器学习、数据挖掘等;**洪智勇**(1978—),男,博士生,讲师。

件分类，以提高对垃圾邮件过滤的性能。

2 相关理论介绍

2.1 粗糙集理论

为了叙述方便，我们先给出粗糙集理论的一些基本概念[10,11]。

定义 1(决策表信息系统) 一个决策表信息系统(简称决策表)$S=\langle U,R,V,f\rangle$，其中 U 是对象的集合，也称为论域；$R=C\cup D$ 是属性集合，子集 C 和 D 分别称为条件属性集和决策属性集，$D\neq\emptyset$；$V=\bigcup_{r\in R}V_r$ 是属性值的集合，V_r 表示属性 $r\in R$ 的属性值范围，即属性 r 的值域；$f:U\times R\to V$ 是一个信息函数，它指定 U 中每一个对象 x 的属性值。

定义 2(条件类和决策类) 给定决策表 S，C 和 D 分别为决策表的条件属性集和决策属性集，$U|\mathrm{IND}(C)$ 和 $U|\mathrm{IND}(D)$ 分别为论域 U 在属性集 C 和 D 上形成的划分，条件类定义为 $E_i\in U|\mathrm{IND}(C)(i=1,\cdots,m$，其中 $m=|U/\mathrm{IND}(C)|$ 为条件类个数)；决策类定义为 $X_j\in U|\mathrm{IND}(D)(j=1,\cdots,n$，其中 $n=|U/\mathrm{IND}(D)|$ 为决策类的个数)。

定义 3(决策表的属性约简) 给定决策表 S，C 和 D 分别为决策表的条件属性集和决策属性集，对条件属性集的某个子集 C' 有 $\mathrm{POS}_{C'}(D)=\mathrm{POS}_C(D)$，则 C' 是条件属性集 C 的一个约简，其中 $\mathrm{POS}_C(D)$ 表示 D 的 C 正域。即在保持条件属性相对于决策属性的分类能力不变的条件下，删除其中不必要的或不重要的属性。一个决策表可能存在多个约简。

定义 4(熵和条件熵) 知识(属性集合)P 的熵 $H(P)$ 定义为：

$$H(P)=-\sum_{i=1}^{n}p(X_i)\log(p(X_i)) \tag{1}$$

式中，$p(X_i)$ 是属性集 P 的概率分布函数。知识 $Q(U|\mathrm{IND}(Q)=\{Y_1,Y_2,\cdots,Y_m\})$ 相对于知识 $P(U|\mathrm{IND}(P)=\{X_1,X_2,\cdots,X_n\})$ 的条件熵 $H(Q|P)$ 定义为：

$$H(Q|P)=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i)) \tag{2}$$

式中，$p(Y_j|X_i)=|Y_j\cap X_i|/|X_i|$，$i=1,2,\cdots,n$，$j=1,2,\cdots,m$。

定理 1(属性冗余定理) 设 U 是一个论域，P 是 U 上的一个条件属性集合，d 为决策属性，且论域 U 是在 P 上相对于 $\{d\}$ 一致的，则 P 中的一个属性 r 是 P 相对于决策属性 d 不必要的(多余的)，其充分必要条件为 $H(\{d\}|p)=H(\{d\}|P-\{r\})$。

定理 2(约简之充要条件定理) 设 U 是一个论域，P 是 U 的一个条件属性集合，d 为决策属性，且论域 U 是在 P 上相对于 $\{d\}$ 一致的，则 $Q\subseteq P$ 是 P 相对于决策属性 d 的一个约简的充要条件为 $H(\{d\}|Q)=H(\{d\}|P)$，且对任意的 $q\in Q$ 都有 $H(\{d\}|Q)\neq H(\{d\}|Q-\{q\})$。

定义 5(属性重要性的信息论观点定义) 设 $S=\langle U,R,V,f\rangle$ 是一个决策表系统，其中 $R=C\cup D$，C 是条件属性集合，$D=\{d\}$ 是决策属性集合，且 $A\subseteq C$，则对任意属性 $a\in C-A$ 的重要性 $SGF(a,A,D)$ 定义为：

$$SGF(a,A,D)=H(D|A)-H(D|A(\{a\})) \tag{3}$$

式中，$H(D|A)$ 表示属性集 D 相对于属性集 A 的条件熵。若 $A=\emptyset$，则 $SGF(a,A,D)=H(D)-H(D|\{a\})$，称为条件属性 a 和决策属性 D 的互信息记为 $I(a,D)$。$SGF(a,A,D)$ 的值越大，说明在已知 A 的条件下，属性 a 对于决策 D 就越重要。

2.2 朴素贝叶斯理论

2.2.1 朴素贝叶斯模型

贝叶斯分类基于贝叶斯公式，即：

$$P(C|X)=\frac{P(X|C)\times P(C)}{P(X)}$$

式中，$P(C|X)$ 为条件 X 下 C 的后验概率，$P(C)$ 为 C 的先验概率，$P(X|C)$ 为条件 C 下 X 的后验概率，$P(X)$ 表示 X 的先验概率。

为叙述方便，对符号作如下约定：用大写字母表示变量，C 表示类别变量，A 表示属性变量。假定共有 m 个属性变量，$A=\langle A_1,A_2,\cdots,A_m\rangle$；用小写字母表示变量取值，$Val(C)=\{c_1,c_2,\cdots,c_l\}$，$Val(A)=\{a_{i1},a_{i2},\cdots,a_{ik}\}$ 分别表示类别变量和属性变量的值域；用 X 表示待分样本集，$x=\langle a_1,a_2,\cdots,a_m\rangle$ 表示待分类样本；用 T 表示训练样本集，$t=\langle a_1,a_2,\cdots,a_m,c_l\rangle$ 表示训练实例。在朴素贝叶斯中假设各属性相对于类别条件独立，则有 $P(a_1,a_2,\cdots,a_m|c_l)=\prod_{i=1}^{m}P(a_i|c_l)$，从而后验概率公式为 $P(c_l|x)=\frac{P(c_l)}{P(x)}\prod_{i=1}^{m}P(a_i|c_l)$，测试样本(E)被分在后验概率最大的类中，由于 $P(x)$ 为一常数，则朴素贝叶斯分类模型为：

$$V_{nb}(E)=\underset{c}{\mathrm{argmax}}P(c)\prod_{i=1}^{m}P(a_i|c) \tag{4}$$

2.2.2 加权朴素贝叶斯模型及权值求解

朴素贝叶斯认为所有条件属性对决策属性的分类重要性是一致的(权重均为 1)，事实并非如此，有些因素对分类影响大一些，而另外的要小一些。因此，可对不同的属性根据其分类重要性赋不同的权值，使朴素贝叶斯得以扩展为加权朴素贝叶斯。加权朴素贝叶斯模型为[9]：

$$V_{wnb}(E)=\underset{c}{\mathrm{argmax}}P(c)\prod_{i=1}^{m}p(a_i|c)^{w_i}$$

式中，w_i 代表属性 A_i 的权值，属性的权值越大，该属性对分类的影响就越大。如何求解权值就成了加权朴素贝叶斯的关键问题。根据属性重要性的信息观点定义(定义 6)，可得出属性权值的求解方法。

定义 6(信息观下属性权值定义) 设 $I(a_i,D)$ 表示条件属性 A_i 与决策属性 D 的互信息，共有 m 个条件属性，则属性 A_i 的权值为：

$$w_i=\frac{I(a_i,D)}{\frac{1}{m}\sum_{i=1}^{m}I(a_i,D)} \tag{5}$$

式中，$\frac{1}{m}\sum_{i=1}^{m}I(a_i,D)$ 表示各条件属性与决策属性互信息的数学期望。

3 基于粗糙集的加权朴素贝叶斯邮件过滤方法

3.1 综合邮件头及邮件内容主要形象特征的特征集提取

一个用户在判断一封邮件是否为垃圾邮件时，往往首先从其主要形象特征就可作出判断，而不会仔细去读邮件本身的内容。这就给我们在邮件过滤的特征提取中一启发：在设计邮件过滤器时是否可只根据邮件的一些主要形象特性作出判断呢？电子邮件格式是由 RFC822 所定义的，是半结构化的文本文件，包括邮件头和正文。其中，邮件头中含有 From，Subject，Date 等关键信息，一封普通电子邮件的邮件头如图 1 所示。

Received:(eyou send program);Tue,06 Oct 2009 08:14:00 +0800

Received:from 202.202.32.41(HELO) cqupt.edu.cn)(202.202.32.41) by 202.202.32.45 with SMTP;Tue,06 Oct 2009 08:14:00 +0800

Received:(eyou anti_spam gateway 3.0);Tue,06 Oct 2009 08:13:57+0800

Message-ID:<454788037.14555@cqupt.edu.cn>

Received:from 218.15.46.204 by 202.202.32.41 with SMTP;Tue,06 Oct 2009 08:13:15 +0800

Received:from nbqlc8(unknown [110.181.63.16]) by smtp75 (Coremail) with SMTP id wpKYMKgtG1Oc7nC6.1 for <dengwb@cqupt.edu.cn>;Tue,06 Oct 2009 08:12:05 +0800(CST)

X-Originating-IP:[110.181.63.16]

Date:Tue,06 Oct 2009 08:12:05 +0800

From:=? gb2312? B? ZmY=? =<dgtu7@WE42345785754.net>

To:<dengwb@cqupt.edu.cn>

Subject:=? gb2312? B? uanTpsnMxsC5wNPrudyhq8DtMjAwOS0xMC02? =Mime-Version:1.0

Content-Type:multipart/mixed;

boundary="----=_NextPart_000_00C8_01CDF21C.47809875"

X-Mailer:Microsoft Outlook Express

Message-ID:<20091006060462.04205@WE42345785754.net>

图1 电子邮件头示例

文献[6]中只考虑了邮件头的特征,而没有考虑邮件内容的一些形象特性(如邮件的格式、邮件中的超链接情况等)。我们认为,虽然邮件头中包含了大量的邮件形象特征,但仅依靠邮件头的特征还是不够的。结合前人的研究[6,12],我们将邮件格式、附件数等特征加以考虑,给出了包含23个属性的综合邮件头及邮件内容主要形象特征的特征集,如表1所列。

表1 综合邮件头及邮件内容主要形象特征的特征集描述

特征	取值	特征说明
A1	整数	收件人个数
A2	整数	邮件中继次数,即邮件头中"Received"标签的个数
A3	整数	邮件路由信息的中断次数
A4	整数	"Received"中的各个域名与其IP不匹配的次数
A5	整数	"Received"中发送站点项缺少域名的次数
A6	0,1	若发件人所在域(MD)与邮件投递过程中所经过的域(ND)一致则取值为0,否则为1
A7	0,1	若"From"中的原始发送地址与"Received"中的原始发送地址一致则取值为1,否则为0
A8	0,1	若"To"中的目的地址与"Received"中的实际收信人地址一致则取值为1,否则为0
A9	0,1	若"Delivered-To"项和"To"项一致则取值为1,否则为0,缺省值为1
A10	0,1	若"Return Path"项和"From"项一致,则取值为1,否则为0,缺省值为1
A11	0,1,2	邮件类型,直接发送为0,回复为1,转发为2
A12	整数	附件(Attachment)数目
A13	0,1	标题有无:无标题为0,有标题为1
A14	整数	正文中http链接数
A15	0,1	邮件正文中是否包含表格:有表格为1,无表格为0
A16	0,1	邮件中是否包含图片:有图片为1,无图片为0
A17	0,1	邮件格式(Text/Html):Text为1,Html为0
A18	整数	邮件编码类型
A19	整数	邮件中字符"!"出现的频度
A20	整数	邮件中字符"$"出现的频度
A21	0,1	抄送(Carbon copy,CC)情况:有抄送为1,无抄送为0
A22	0,1	发送时间:8:00—23:00为0,23:00—8:00为1
A23	0,1,2	邮件长度,小于1M为0,大于5M为2,其余为1

3.2 基于粗糙集的加权朴素贝叶斯邮件过滤算法

本算法基于决策信息表,其前提条件是搜集电子邮件的相关信息,并进行相应的数据整理,形成电子邮件信息表。一个决策表 $T=\langle U,R,V,f\rangle$,其中 U 是对象的集合,也称为论域;$R=C\cup D$ 是属性集合,子集 C 和 D 分别称为条件属性集和决策属性集,C 是邮件特征属性集,即 $C=\{A1,A2,\cdots,A23\}$,D 是邮件的状态特征,即 $D=\{0,1\}$,其中0代表垃圾邮件,1代表正常邮件;$V=\bigcup_{r\in R}V_r$ 是属性值的集合,V_r 表示属性 $r\in R$ 的属性值范围,即属性 r 的值域;$f:U\times R\to V$ 是一个信息函数,它指定 U 中每一个对象 x 的属性值。邮件过滤过程具体如算法1。

算法1 基于粗糙集的加权朴素贝叶斯邮件过滤算法

输入:一个电子邮件信息表 $T=(U,C\cup D,V,f)$,其中 U 为论域,C,D 分别为条件属性集和决策属性集。

输出:邮件分类情况。

Step1 数据预处理:将训练样本和待分类样本进行补齐和离散化。

Step2 判断:如果是分类任务,则转Step7;如果是训练任务,则转Step3。

Step3 属性约简:

Step3.1 计算决策表 T 中决策属性集 D 相对于条件属性集 C 的条件熵 $H(D|C)$;

Step3.2 计算条件属性集 C 中相对决策属性集 D 的核属性集 C_0,并令 $C'=C-C_0$;

Step3.3 令 $B=C_0$,

(1) 如果 $|B|\neq 0$,则计算条件熵 $H(D|B)$,转(3);

(2) 对每个属性 $A_i\in C'$,选择使 $H(D|B\cup\{A_i\})$ 最小的属性 A_i,$C'=C'-A_i$,$B=B\cup\{A_i\}$;

(3) 若 $H(D|B)=H(D|C)$,则终止,否则转(2)。

Step4 属性重要性计算:根据式(5)计算约简后的每个属性 $A_j\in B$ 相对于决策属性集 D 的重要性,记为 w_j。

Step5 概率参数学习:扫描所有约简后的训练样本,计算所有的先验概率 $P(a_{jk}|c_l)$,即在类别 c_l 中属性 A_j 的第 k 种取值的概率;以及 $p(c_l)$,即取值为类别 c_l 的概率。

Step6 生成朴素贝叶斯概率表及属性权值列表。

Step7 分类:调用概率表及属性权值列表,得出邮件分类结果。

4 仿真实验

4.1 算法评价标准选择

用 N_L 表示实际的合法邮件数,N_S 表示实际的垃圾邮件数,$n_{L\to L}$ 表示正确查出的合法邮件数,$n_{L\to S}$ 表示被误判为垃圾邮件的合法邮件数,$n_{S\to S}$ 表示正确查出的垃圾邮件数,$n_{S\to L}$ 表示被认为是合法邮件的垃圾邮件数,那么召回率(Recall)和正确率(Precision)的计算公式为:

$$\text{Recall}=\frac{n_{S\to S}}{n_{S\to S}+n_{S\to L}}$$

$$\text{Precision}=\frac{n_{S\to S}}{n_{S\to S}+n_{L\to S}}$$

为了衡量正确分类的邮件情况,也常用准确率 Accuracy 来度量,$\text{Accuracy}=\frac{n_{L\to L}+n_{S\to S}}{N_L+N_S}$。在邮件过滤中,我们希望的是将垃圾邮件过滤掉,但绝不希望将正常邮件划分为垃圾邮件。因为有时哪怕将一封正常邮件划分为垃圾邮件,所带来的损失也是惨重的,为了衡量将正常邮件划分为垃圾邮件的比率,还采取了一种新的评估指标 F1,$\text{F1}=\frac{n_{L\to S}}{n_{L\to L}+n_{L\to S}}=\frac{n_{L\to S}}{N_L}$。

4.2 实验过程及结果

为了验证本文提出的邮件过滤方法的有效性，选择了两个通用的标准邮件测试集。英文语料选自 Spam Assassin 邮件库[14]，其中正常邮件 4150 封、垃圾邮件 1897 封，每封邮件均包含了邮件头及邮件内容的完整信息。中文语料选自中国教育和科研计算机网紧急响应组提供的电子邮件数据集 CCERT 2005-Jul[15]，该数据集包括一个垃圾邮件集和一个正常邮件集，我们从中选择了正常邮件 5000 封、垃圾邮件 2000 封，在 VC++ 6.0 环境下对本算法进行测试。

测试步骤如下。

Step1 邮件特征提取：根据表 1 中所描述的属性进行邮件的特征提取，形成一个邮件决策表。其中决策属性 $D=\{0,1\}$，0 代表垃圾邮件，1 代表正常邮件。

Step2 决策表补齐：由于并不是每封邮件均能提取得到完备的特征，对不完备的特征我们用重庆邮电大学计算科学技术研究所研发的“基于 Rough Set 的智能数据分析系统(RIDAS)[14]”的“条件组合补齐算法”进行补齐。

Step3 属性约简和权值计算：用算法 1 的 Step3 所述方法在 RIDAS 系统中进行属性约简，而后用定义 6 计算各属性的权值。

Step4 邮件分类方法的对比分析：测试采用 5 重交叉验证方法，先用本文提出的基于粗糙集的加权朴素贝叶斯邮件过滤方法进行测试，再分别用文献[6]中提出的基于 Rough Set 的邮件分类方法、文献[4]所描述的基于朴素贝叶斯的根据邮件内容的分类方法、基于 SVM 的根据邮件内容的分类方法及基于粗糙集的朴素贝叶斯方法在相同的邮件集上进行测试，其测试结果表 2 所列。

表 2 几种邮件分类方法的实验结果

邮件集	评价指标	RS(%)	NB(%)	SVM(%)	NBBRS(%)	WNBBRS(%)
Spam Assassin	Recall	87.41	83.13	85.76	88.72	90.81
	Precision	83.62	80.95	81.64	84.58	86.33
	Accuracy	85.80	82.33	83.26	87.13	88.45
	F1	6.63	6.76	6.14	5.76	5.63
CCERT 2005-Jul	Recall	87.55	83.88	84.68	90.21	91.56
	Precision	82.48	80.75	82.56	86.49	88.67
	Accuracy	86.34	82.63	85.48	88.83	90.78
	F1	6.40	6.53	6.84	5.94	5.73

注：RS 表示文献[6]中提出的邮件过滤方法；NB 和 SVM 分别表示文献[4]所描述的基于 naive Bayes 和 SVM 的根据邮件内容的分类方法，两者的特征数均取 500 维；NBBRS 为基于粗糙集的朴素贝叶斯邮件分类方法；WNBBRS 为本文提出的基于粗糙集和加权朴素贝叶斯的过滤方法。

通过对几种方法测试结果的比较分析，可以得出以下结论：

(1)从 WNBBRS 与 NB，SVM 方法的对比情况看，综合邮件头及邮件内容的主要形象特征提取的属性数量远比基于邮件内容的特征词方法少，过滤效果在 Recall，Precision 和 Accuracy 3 个指标上表现也较好。这表明本文邮件特征集提取方法既提高了邮件过滤的时间效率，也提高了过滤的准确性。

(2)WNBBRS 方法的过滤效果比 RS 方法要好，证明了我们在邮件头的基础上提取的其他邮件特征是很必要的。

(3)通过属性的权值求解，发现各属性的权重有较大差异，如邮件格式、附件数、发送时间及超链接数的权值较大，在本文提出的 WNBBRS 方法中我们对其赋予了较高的权重，过滤效果要比未加权的基于粗糙集的朴素贝叶斯方法(NBBRS)好，这表明了通过加权处理能较好地克服朴素贝叶斯邮件过滤中的条件独立性假设问题。

结束语 电子邮件过滤是当前网络安全研究的一个热点问题。本文对邮件过滤的特征进行了分析，提出了一种新的邮件特征提取方法。此外，为了克服朴素贝叶斯条件依赖的缺陷，给出了一种基于粗糙集的加权朴素贝叶斯邮件过滤算法，通过在标准邮件测试集上的仿真验证，说明了所提方法是有效的。在今后的工作中，将从以下方面展开进一步研究：在邮件特征的选取上，结合文本数据挖掘方法，选取更能反映邮件特征的条件属性集；由于将正常邮件当成垃圾邮件过滤掉的代价往往是很大的，如何既能提高对垃圾邮件的过滤效率，又能降低对正常邮件的“误杀”率也值得进一步研究。

参 考 文 献

[1] Guzella T S, Caminhas W M. A review of machine learning approaches to Spam filtering[J]. Expert Systems with Applications, 2009, 36(7): 10206-10222

[2] Lai Chih-chin. An Empirical Study of Three Machine Learning Methods for Spam Filtering[J]. Knowledge-Based System, 2007, 20(3): 249-254

[3] Sahami M, Dumais S, Heckerman D, et al. A Bayesian approach to filtering junk email[C]// AAAI Workshop on Learning for Text Categorization, AAAI Technical Report WS-98-05. Madison, Wisconsin, July 1998

[4] Kim J, Chung K, Choi K. Spam filtering with dynamically updated URL statistics[J]. IEEE Security and Privacy, 2007, 5(4): 33-39

[5] Chen Xiao-li, Liu Pei-yu, Zhu Zhen-fang, et al. A method of spam filtering based on weighted support vector machines[C]// IEEE International Symposium on IT in Medicine & Education. 2009: 947-950

[6] 李志君，王国胤，吴渝. 基于 Rough Set 的电子邮件分类系统[J]. 计算机科学，2004，31(3)：58-61

[7] Ciltik A, Gungor T. Time-efficient spam e-mail filtering using n-gram models[J]. Pattern Recognition Letters, 2008, 29(1): 19-33

[8] Pawlak Z. Rough sets and intelligent data analysis[J]. Information Sciences, 2002, 147: 1-12

[9] 邓维斌，王国胤，王燕. 基于 Rough Set 的加权朴素贝叶斯分类算法[J]. 计算机科学，2007，32(2)：204-206，219

[10] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001：1-147

[11] 王国胤，于洪，杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报，2002，25(7)：759-766

[12] Lai Gu-hsin, Chen Chia-mei, Laih Chi-sung, et al. A collaborative anti-spam system [J]. Expert Systems with Applications, 2009, 36: 6645-6653

[13] DEEPSOFT. SpamAssassin Project [DB/OL]. http://spamassassin.apache.org/publiccorpus/, 2007-05-10

[14] 中国教育和科研计算机网紧急响应组. CCERT 2005-Jul 数据集[R]. 2005

[15] Wang G Y, Zheng Zheng, Wu Yu. RIDAS-A Rough Set Based Intelligent Data Analysis System[C]// First IEEE International Conference on Machine Learning and Cybernitics (ICMLC 2002). Beijing, 2002: 646-649