# 垃圾邮件过滤中信息增益的改进研究

翟军昌[1] 秦玉平[1] 车伟伟[2]
(渤海大学 锦州 121000)[1] (沈阳大学 沈阳 110044)[2]

**摘 要** 针对垃圾邮件过滤中的特征项选择问题,提出了一种改进的信息增益方法。首先利用特征词的先验概率定义增益比,然后利用增益比对特征词为整个分类所提供的信息量进行放大或弱化,从而对特征词的类别条件熵计算作了改进,采用极大后验假设朴素贝叶斯决策方法在英文语料库上进行实验,通过召回率、正确率、精确率和错误率对算法进行评价分析。实验结果表明,改进后的算法提高了过滤器的分类精度,降低了过滤器对合法邮件的误判给用户带来的损失。

**关键词** 信息增益,特征选择,垃圾邮件,朴素贝叶斯
**中图法分类号** TP391 **文献标识码** A

## Improvement of Information Gain in Spam Filtering

ZHAI Jun-chang[1] QIN Yu-ping[1] CHE Wei-wei[2]
(Bohai University, Jinzhou 121000, China)[1] (Shenyang University, Shenyang 110044, China)[2]

**Abstract** The paper put forward a kind of improved information gain for the feature words selection in spam filtering. Firstly, defined gain ratio according to the probability of feature words, and then amplifed or weakened the amount of information of the feature words for classification, thereby improving the calculation method of category conditional entropy. Finally, combining with the naive Bayes decision method of maximum a posteriori hypothesis, carried out an experiment on the English Corpus to analyze the algorithm through recall, correct, accuracy and error. The experimental results show that the improved algorithm can enhance classification precision and reduce user loss.

**Keywords** Information gain, Feature selection, Spam, Naive Bayes

## 1 引言

电子邮件(E-mail)在人们日常工作和生活中发挥着越来越重要的作用。与此同时,大量包含欺诈、营销、暴力、色情和病毒等信息的垃圾邮件也随之产生。垃圾邮件日益泛滥,不仅占据了大量的网络带宽资源,而且产生一系列严重的网络安全问题。针对垃圾邮件问题的处理,目前主要以过滤技术为主,其中典型的是基于内容的过滤和基于身份标示的过滤两种类型。基于内容的过滤技术,以贝叶斯(Bayes)、支持向量机(SVM)和决策树(KNN)等机器学习方法为代表,该类方法的主要特点是以邮件的内容特征作为邮件分类的依据。基于身份标示的过滤技术,以基于黑、白名单过滤、反向DNS查询和基于用户信誉的过滤技术等为代表,该类方法的特点是根据邮件发件人的身份特征相关信息来判断邮件是否为垃圾邮件[1-7]。

邮件内容的特征反映了邮件的内容主题,是邮件分类的一个重要依据。目前基于内容的垃圾邮件过滤技术应用研究较多,该类方法首先收集大量合法邮件和垃圾邮件作为样本,然后指导过滤器对收集到的邮件样本进行学习,最后通过训练好的过滤器对新到达的邮件进行最终分类。过滤器通过对邮件样本的训练和学习可以自动获得垃圾邮件的特征,并根据垃圾邮件特征的变化准确地对垃圾邮件进行过滤。过滤器在学习阶段能否获得邮件样本的内容的有效信息从而建立有效的特征项词库,将直接影响过滤器的性能[4]。在实际使用中,用户宁愿接收更多的垃圾邮件,也不愿意将合法邮件误判为垃圾邮件,此外不同的用户对于同一封邮件的决策也不同,因此如何有效提取邮件样本的特征,降低对合法邮件的误判,显得尤为重要。

本文针对垃圾邮件过滤中特征项选择问题,提出了一种改进的信息增益方法。利用特征词的先验概率定义增益比,对特征词的类别条件熵计算做了改进,并采用了极大后验假设的贝叶斯决策方法。实验结果表明算法改进后过滤器的召回率变化与算法改进前召回率的变化比较接近,但是在改进后的算法中,正确率有明显的提高,而且正确率的变化比较稳定,表明改进后的算法使过滤器对合法邮件的误判数量在减少,对合法邮件的误判率在降低,降低了过滤器对合法邮件的误判给用户带来的损失。

## 2 相关知识介绍

### 2.1 特征选择与信息熵

特征选择是一种通过评价的方法,从高维向量空间中选

到稿日期:2013-09-02 返修日期:2013-12-27 本文受国家自然科学基金(61104106)资助。
翟军昌(1978—),男,博士生,讲师,主要研究方向为机器学习,E-mail:zhaijunchang@163.com;秦玉平(1965—),男,博士,教授,主要研究方向为机器学习;车伟伟(1980—),女,博士。

取对文本分类有效的特征词，从而达到对向量空间降维，提高文本分类效率的目的。常用的特征选择方法有文档频次(DF)、信息增益(IG)、互信息(MI)、相对熵和 $x^2$ 统计等。

利用以上方法可以选取出 $n$ 个特征词 $t_1,t_2,\cdots,t_n$ 构成向量空间，记为：$D=\{t_1,t_2,\cdots,t_n\}$，则在向量空间中，对于任意给定的邮件对应的特征向量，记为：$d=\{w_1,w_2,\cdots,w_n\}$，其中 $w_1,w_2,\cdots,w_n$ 代表特征词 $t_1,t_2,\cdots,t_n$ 的权重。

**定义 1**(熵) 假设随机变量 $X$ 可能的取值 $x_i$ 有 $n$ 种，如果每一种取值 $x_i$ 出现的概率为 $p(x_i)$，则随机变量 $X$ 的不确定性称为熵，记为 $H(X)$。

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

若随机变量 $X$ 的取值变化越多，则随机变量 $X$ 所携带的信息量越大，同时随机变量 $X$ 的熵 $H(X)$也就越大。

**定义 2**(条件熵) 假设有随机变量 $X$ 和 $Y$，随机变量 $X$ 和 $Y$ 的可能取值 $x_i$ 和 $y_i$ 分别有 $n$ 和 $m$ 种，每一种取值 $x_i$ 和 $y_j$ 出现对应的概率分别为 $p(x_i)$和 $p(y_j)$，$p(x_i|y_j)$表示观测到随机变量 $Y$ 后随机变量 $X$ 发生的概率。则在观测到随机变量 $Y$ 之后，随机变量 $X$ 的不确定性称为条件熵，记为 $H(X|Y)$。

$$H(X|Y)=-\sum_{j=1}^{m}p(y_j)\sum_{i=1}^{n}p(x_i|y_j)\log(p(x_i|y_j)) \tag{2}$$

### 2.2 贝叶斯理论

**定义 3**(贝叶斯法则) 在给定训练数据 $D$，以及 $D$ 的假设空间 $C$ 的情况下，$P(c)$表示在没有训练数据 $D$ 前假设 $c$ ($c\in C$)拥有的初始概率，称为 $c$ 的先验概率(prior probability)；$P(D)$表示将要观察训练数据 $D$ 的先验概率；条件概率 $P(D|c)$表示在假设 $c$ 成立的条件下观察到数据 $D$ 的先验概率；条件概率 $P(c|D)$表示在观察到数据 $D$ 后，假设 $c$ 成立的后验概率，$P(c|D)$反映了在看到训练数据 $D$ 后 $c$ 成立的置信度。贝叶斯法则给出了通过先验概率 $P(c)$和条件概率 $P(D|c)$来计算后验概率 $P(c|D)$的准则，即：

$$p(c|D)=\frac{P(D|c)p(c)}{p(D)},c\in C \tag{3}$$

**定义 4**(极大后验假设) 已知假设空间 $C$ 中所有假设 $c$ ($c\in C$)的先验概率分布，在给定训练数据 $D$ 的情况下，从假设空间 $C$ 中找出训练数据 $D$ 可能性最大的假设，称为极大后验假设(maximum a posteriori，MAP)，记为 $c_{MAP}$。

$$\begin{aligned}c_{MAP}&=\arg\max_{c\in C}P(c|D)=\arg\max_{c\in C}\frac{P(D|c)P(c)}{P(D)}\\&=\arg\max_{c\in C}P(D|c)P(c)\end{aligned} \tag{4}$$

在式(4)中，最后一步去掉了 $P(D)$，因为 $P(D)$是不依赖于 $c$ 的常量。

### 2.3 朴素贝叶斯分类模型

在贝叶斯分类模型中[8-11]，根据贝叶斯公式可知，计算任意邮件 $d=\{w_1,w_2,\cdots,w_n\}$属于 $c_j$($j=0$ 代表垃圾邮件，$j=1$ 代表合法邮件)类邮件的概率为：

$$\begin{aligned}p(c_j|D=d)&=\frac{p(D=d|c_j)p(c_j)}{p(D=d)}\\&=\frac{p(w_1,w_2,\cdots,w_n|c_j)p(c_j)}{\sum_{j=0}^{1}p(c_j)p(w_1,w_2,\cdots,w_n|c_j)},j=0,1\end{aligned} \tag{5}$$

朴素贝叶斯分类模型[7-11]是假设构成特征向量 $D$ 的 $n$ 个特征词 $t_1,t_2,\cdots,t_n$ 之间没有任何依赖关系，即假设特征词 $t_1,t_2,\cdots,t_n$ 之间是相互独立的，消除了特征词之间的相互依赖关系得到的一种简化贝叶斯模型。

根据朴素贝叶斯假设可知，在式(5)中条件概率 $p(w_1,w_2,\cdots,w_n|c_j)$可采用式(6)的方法计算：

$$p(w_1,w_2,\cdots,w_n|c_j)=\prod_{i=1}^{n}p(w_i|c_j),j=0,1 \tag{6}$$

由式(5)和式(6)可知，计算邮件 $d$ 属于 $c_j$ 类邮件概率的方法如式(7)所示：

$$p(c_j|D=d)=\frac{p(c_j)\prod_{i=1}^{n}p(w_i|c_j)}{\sum_{j=0}^{1}p(c_j)\prod_{i=1}^{n}p(w_i|c_j)},j=0,1 \tag{7}$$

由式(7)可以计算出邮件 $d$ 属于垃圾邮件的概率 $p(c_0|d)$和合法邮件的概率 $p(c_1|d)$，根据极大后验假设对邮件 $d$ 进行最终分类的公式如下：

$$\begin{aligned}c_{MAP}&=\arg\max_{c_{j=0,1}}P(c_j|D=d))\\&=\arg\max_{c_{j=0,1}}P(D=d|c_j)P(c_j)\\&=\arg\max_{c_{j=0,1}}p(w_1,w_2,\cdots,w_n|c_j)P(c_j)\\&=\arg\max_{c_{j=0,1}}P(c_j)\prod_{i=1}^{n}P(w_i|c_j)\end{aligned} \tag{8}$$

## 3 信息增益改进研究

### 3.1 信息增益改进分析

信息增益(information gain，简称 IG)，是指用一个属性 $t$ 去划分样本空间而导致期望熵降低的程度。在文本分类中，如果 $H(c)$代表类别 $C$ 的熵，$H(c|t)$代表观测到属性 $t$ 后属于类别 $c$ 的条件熵，则信息增益的定义如下：

$$\begin{aligned}IG(t)&=H(c)-H(c|t)=-\sum_{i=1}^{n}p(c_i)\log p(c_i)+\\&\quad\sum_{j=1}^{m}p(t_j)\sum_{i=1}^{n}p(x_i|y_j)\log(p(x_i|y_j))\\&=-\sum_{i=1}^{n}p(c_i)\log p(c_i)+p(t)\sum_{i=1}^{n}p(c_i|t)\log p(c_i|t)+\\&\quad p(\bar{t})\sum_{i=1}^{n}p(c_i|\bar{t})\log p(c_i|\bar{t})\end{aligned} \tag{9}$$

式中，$p(c_i)$表示 $c_i$ 类文本在训练样本中出现的概率；$p(t)$表示单词 $t$ 在训练样本中出现的概率；$p(\bar{t})$表示单词 $t$ 在训练样本中不出现的概率；$p(c_i|t)$表示在单词 $t$ 出现的情况下属于 $c_i$ 类的概率；$p(c_i|\bar{t})$表示在单词 $t$ 不出现的情况下属于 $c_i$ 类的概率。如果 $IG(t)$越大，则说明特征词 $t$ 对整个分类的作用越大。$IG(t)$反映了特征词 $t$ 为整个分类所提供的信息量。

在垃圾邮件过滤中，由于垃圾邮件分类属于二类问题，若令 $c_i$ 的取值为 $c_0$ 和 $c_1$，$c_0$ 代表垃圾邮件，$c_1$ 代表合法邮件，则式(9)变为式(10)：

$$\begin{aligned}IG(t)&=H(c)-H(c|t)=-\sum_{i=0}^{1}P(c_i)\log p(c_i)+\\&\quad p(t)\sum_{i=0}^{1}p(c_i|t)\log p(c_i|t)+p(\bar{t})\sum_{i=0}^{1}p(c_i|\bar{t})\log p\\&\quad(c_i|\bar{t})\end{aligned} \tag{10}$$

根据式(10)可以计算出每个词的 IG 值，然后取 IG 值最大的 $n$ 个词 $t_1,t_2,\cdots,t_n$ 构成向量空间，记为：$D=\{t_1,t_2,\cdots,t_n\}$。

在垃圾邮件过滤中，过滤器首先通过对合法邮件和垃圾邮件样本库进行学习，获取合法邮件和垃圾邮件样本的特征，从而建立特征词库，最后按照决策规则对新邮件进行分类。因此特征项词库的确定对最终过滤器的分类决策起到了至关重要的作用。

在利用信息增益计算特征词 $t$ 的信息增益值($IG(t)$)时，由于 $H(c)$可根据先验概率计算得到，因此对于一个确定的二

类分类集来说，类别 $c$ 的熵 $H(c)$是一个确定的值。因此每一个特征词 $t$ 的信息增益值($IG(t)$)的大小由条件熵 $H(c|t)$的估计值决定。$IG(t)$的大小反映了 $t$ 为整个分类所提供的信息量，所以条件熵 $H(c|t)$的计算将决定最终特征词的分布情况，从而直接影响过滤器最终的分类效果。

条件熵 $H(c|t)$在计算时，既考虑特征词 $t$ 出现时属于类别 $c$ 的熵，同时也考虑了特征词 $t$ 不出现时属于类别 $c$ 的熵。因此，当 $p(t)>p(\bar{t})$时，特征词 $t$ 出现的概率大于特征词 $t$ 不出现的概率，则特征词 $t$ 出现时对整个分类的贡献大于其不出现时对分类的贡献，从而特征词 $t$ 出现时对整个分类所提供的信息量大于其不出现时对整个分类所提供的信息量。当 $p(t)<p(\bar{t})$时，特征词 $t$ 出现的概率小于特征词 $t$ 不出现的概率，则特征词 $t$ 出现时对整个分类的贡献小于其不出现时对分类的贡献，从而特征词 $t$ 出现时对整个分类所提供的信息量小于其不出现时对整个分类所提供的信息量。当 $p(t)=p(\bar{t})$时，特征词 $t$ 出现的概率等于其不出现的概率，那么特征词 $t$ 出现和不出现时对整个分类的贡献是相同的，从而特征词 $t$ 出现和不出现时对整个分类所提供的信息量是相同的。

### 3.2 信息增益的改进

基于前面的分析，本文利用先验概率 $p(t)$和 $p(\bar{t})$定义增益 $K_1$ 和 $K_2$，对条件熵 $H(c|t)$的计算做了如下改进：

(1) 定义 $K_1=\frac{p(t)}{p(\bar{t})}$，$K_1$ 称为特征词 $t$ 出现时对特征词 $t$ 不出现时的增益比。

(2)定义 $K_2=\frac{p(\bar{t})}{p(t)}$，$K_2$ 称为特征词 $t$ 不出现时对特征词 $t$ 出现时的增益比。

(3)定义新的条件熵 $H'(c|t)$，利用增益 $K_1$ 和 $K_2$ 对特征词 $t$ 出现和不出现时对整个分类所提供的信息量进行放大或弱化。

$$H'(c|t)=K_1 p(t)\sum_{i=1}^{n}p(c_i|t)\log p(c_i|t)+K_2 p(\bar{t})\sum_{i=1}^{n}p(c_i|\bar{t})\log p(c_i|\bar{t}) \tag{11}$$

(4)根据贝叶斯法则，利用先验概率 $p(t|c_i)$、$p(\bar{t}|c_i)$、$p(t)$和 $p(\bar{t})$将式(11)中的条件概率 $p(c_i|t)$和 $p(c_i|\bar{t})$的计算方法进行变形，如式(12)所示：

$$p(c_i|t)=\frac{p(t|c_i)p(c_i)}{p(t)},\ p(c_i|\bar{t})=\frac{p(\bar{t}|c_i)p(c_i)}{p(\bar{t})},\ i=0,1 \tag{12}$$

则，改进后的信息增益记为 $IG'(t)$，$IG'(t)$定义如下：

$$\begin{aligned}IG'(t)&=H(c)-H'(c|t)=-\sum_{i=0}^{1}P(c_i)\log p(c_i)+K_1 p(t)\sum_{i=0}^{1}p(c_i|t)\log p(c_i|t)+K_2 p(\bar{t})\sum_{i=0}^{1}p(c_i|\bar{t})\log p(c_i|\bar{t})\\&=-\sum_{i=0}^{1}P(c_i)\log p(c_i)+K_1 p(t)\sum_{i=0}^{1}\frac{p(t|c_i)p(c_i)}{p(t)}\log\frac{p(t|c_i)p(c_i)}{p(t)}+K_2 p(\bar{t})\sum_{i=0}^{1}\frac{p(\bar{t}|c_i)p(c_i)}{p(\bar{t})}\log\frac{p(\bar{t}|c_i)p(c_i)}{p(\bar{t})}\end{aligned} \tag{13}$$

## 4 实验结果与分析

在 Windows XP 下，硬件配置 Pentium D 3.40GHz，内存 1.0G，硬盘 160G，以 VC++6.0 为实验环境，实验中使用 Androutsopoulos[8] 等人提供的 Ling-Spam 的语料库，选用了 lemm_stop 形式语料，其中包括 2412 封语言学家的合法邮件和 481 封垃圾邮件，将邮件样本分成 10 份进行交叉实验。采用召回率(Recall)、正确率(Precision)、精确率(Acc)和错误率(Err)作为过滤器的评价标准，其计算方法如下：

$$Recall=\frac{A}{S}\qquad Precision=\frac{A}{A+D}$$

$$Acc=\frac{A+C}{S+N}\qquad Err=\frac{B+D}{S+N}$$

其中，$A$ 代表被正确识别的垃圾邮件总数，$B$ 代表被误判的垃圾邮件总数，$C$ 代表被正确识别的合法邮件总数，$D$ 代表被误判的合法邮件总数，$N$ 代表合法邮件总数，$S$ 代表垃圾邮件总数。

实验中，对邮件样本学习时，利用式(13)改进后的信息增益方法建立特征词库 L。对新邮件进行分类决策时，首先在特征词库 L 中选取 $n$ 个特征词 $t_1,t_2,\cdots,t_n$ 建立 $n$ 维特征向量空间，其中向量空间维数 $n$ 从 100 变化到 1000，每次实验增加 100，然后采用极大后验假设的方法进行分类决策。分别采用改进前和改进后的方法对 lemm_stop 中 10 份样本进行交叉过滤实验，最后根据 10 份样本交叉实验的结果对召回率、正确率、精确率和错误率取平均值。实验结果如图 1—图 4 所示。

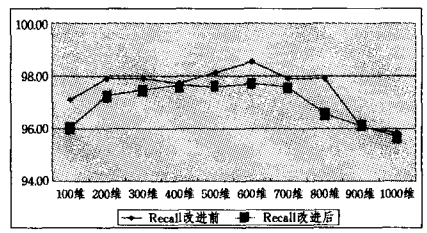


图 1 召回率变化对比

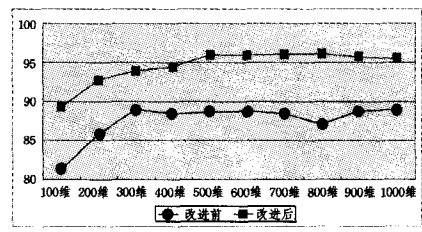


图 2 正确率变化对比

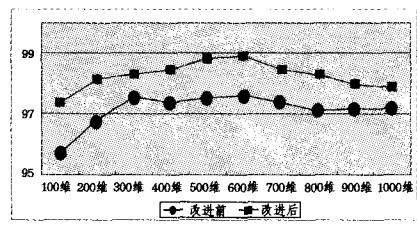


图 3 精确率变化对比

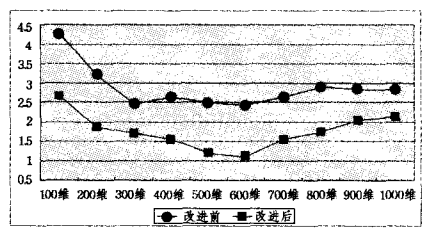


图 4 错误率变化对比

由图 1 可以看出，在改进后的算法中，召回率有所下降，

(下转第 224 页)

[5] 郑四发,曹剑东,连小珉. 复杂路网下多客户间最短路径的扇面Dijkstra算法[J]. 清华大学学报:自然科学版,2009(11):115-120

[6] 刘建美,马寿峰,马帅奇. 基于改进的Dijkstra算法的动态最短路计算方法[J]. 系统工程理论与实践,2011(6):200-206

[7] 李敬贤,厉小润. 求解震后最优路径的改进Dijkstra算法[J]. 计算机工程,2012(3):177-183

[8] 周竞文,程志全,金士尧. 基于Dijkstra距离剪枝的测地线求解算法[J]. 系统仿真学报,2009(10):92-98

[9] 吴若伟,楼佩煌. 基于Dijkstra算法的大型停车场最优泊车路径规划[J]. 工业控制计算机,2013(5):183-189

[10] Idwan S, Etaiwi W. Dijkstra algorithm heuristic approach for large graph[J]. J Appl Sci,2011,12:2255-2259

[11] Medeiros F L L,da Silva J D S. A Dijkstra Algorithm for Fixed-Wing UAV Motion Planning Based on Terrain Elevation[J]. Lecture notes in computer science,2010,6404:213-222

[12] Gunkel C, Stepper A, Muller A C ,et al. Micro crack detection with Dijkstra's shortest path algorithm[J]. Machine Vision and Applications,2012,23(3):589-601

[13] Li R. Utilizing Restricted Direction Strategy and Binary Heap Technology to Optimize Dijkstra Algorithm in WebGIS[J]. Key Engineering Materials,2010(419/420):557-560

[14] Tintor V, Radunovi A J. Distributed Dijkstra sparse placement routing algorithm for translucent optical networks[J]. Photonic Network Communication,2009,18(1):55-64

[15] Kimoto M, Tsuchiya T, Kikuno T. On the Time Complexity of Dijkstra's Three-State Mutual Exclusion Algorithm[J]. IEICE Transactions on Information and Systems, 2009, 92(8): 1570-1573

[16] Bauer R, Delling D, Sanders P, et al. Combining Hierarchical and Goal-Directed Speed-Up Techniques for Dijkstra's Algorithm [J]. Lecture Notes in Computer Science,2008,5038:303-318

[17] 耿素云. 离散数学[M]. 北京:清华大学出版社,2008

[18] 吕建华,王国仁,于戈. XML数据的路径表达式查询优化技术[J]. 软件学报,2003(9):1193-1199

[19] 孔令波,唐世渭,杨冬青,等. XML数据的查询技术[J]. 软件学报,2007(6):1400-1418

[20] 孟小峰,王宇,王小锋. XML查询优化研究[J]. 软件学报,2006(10):2069-2086

[21] 孔令波,唐世渭,杨冬青,等. XML数据索引技术[J]. 软件学报,2005(12):1000-1017

(上接第216页)

表明改进后的算法对垃圾邮件的识别能力有所降低,但是算法改进后的召回率变化与算法改进前召回率的变化比较接近。由图2可以看出,在改进后的算法中,正确率有明显的提高,而且正确率的变化比较稳定,表明改进后的算法中过滤器对合法邮件的误判数量在减少,对合法邮件的误判率在降低,从而也降低了过滤器误判给用户带来的损失。由图3和图4可以看出,在改进后的算法中,过滤器的精确率在上升,错误率在下降,表明算法改进后过滤器的分类精度在提高。

**结束语** 本文针对垃圾邮件过滤中的特征项选择问题,利用特征词的先验概率定义增益比,对特征词的类别条件熵的计算做了改进,利用增益比对整个分类所提供的信息量进行放大或弱化,从而提出了一种改进的信息增益方法来提取特征词。最后在英文语料库上进行实验,实验中采用了极大后验假设的朴素贝叶斯决策方法,实验结果表明改进后的算法虽然漏掉了一部分垃圾邮件,但是合法邮件误判率在降低,对合法邮件判断更加准确,从而降低了对合法邮件的误判给用户带来的损失。

本文下一步研究的工作是在提高过滤器的正确率,降低用户损失的基础上,提高过滤器的召回率。

## 参 考 文 献

[1] Guzella T S, Caminhas W M. A review of machine learning approaches to spam filtering[J]. Expert Systems with Application, 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] 黄国伟,许昱玮. 基于用户反馈的混合型垃圾邮件过滤方法[J]. 计算机应用,2013,33(7):1861-1865

[4] 邓维斌,王国胤,洪智勇. 基于粗糙集的加权朴素贝叶斯邮件过滤方法[J]. 计算机科学,2011,38(2):218-221

[5] Sanchez F, Duan Zhen-hai, Dong Ying-fei. Understanding Forgery Properties of Spam Delivery Paths[C]// CEAS 2010 Seventh annual Collaboration, Electronic messaging, AntiAbuse and Spam Conference (CEAS 2010). Redmond, Washington, US, July 2010

[6] 陈孝礼,刘培玉. 应用于垃圾邮件过滤的词序列核[J]. 计算机应用,2011,31(3):698-701

[7] Sahami M, Dumais S, Heckerman D, et al. A Bayesian approach to filtering Junk e-mail [C]// Learning for Text Categorization: Papers from AAAI Workshop. Madison, Wisconsin, 1998:55-62

[8] Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An Evaluation of Naive Bayesian Anti-Spam Filtering[C]// Proc of the Workshop on Machine learning in the New Information Age, 11th European Conference on Machine Learning (ECML'00). Barcelona, Spain, June 2000:9-17

[9] Schneider K. A Comparison of Event Models for Naive Bayes Anti-spam E-mail Filtering[C]// Procedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics(EACL'03). 2003:307-314

[10] Vangelis M, Androutsopoulos I, Georgios P. Spam filtering with Naive Bayes-which Naive Bayes? [C]// CEAS 2006 Third Conference on Email and AntiSpam(CEAS 2006). Mountain View, California USA, July 2006:27-28

[11] Chen Bin, Dong Shou-bin, Fang Wei-dong. Introduction of Fingerprint Vector based Bayesian Method for Spam Filtering [C]// CEAS 2007 Fourth Conference on Email and Anti-Spam(CEAS 2007). Mountain View, California USA, August 2007