中文垃圾邮件多层次过滤技术的应用研究

刘延华,陈国龙
LIU Yan-hua,CHEN Guo-long

福州大学 数学与计算机科学学院,福州 350108
College of Mathematics and Computer Science,Fuzhou University,Fuzhou 350108,China

LIU Yan-hua,CHEN Guo-long.Application research of multiplayer filter algorithm on Chinese spam filtering.Computer Engineering and Applications,2009,45(34):94-97.

Abstract: Aiming at the key issue of Chinese spam filtering,a multiplayer filter based on an improved minimum risk Bayes algorithm.The result of experiment shows this multiplayer filter but reach more better recall and precision,also can speed up the process of spam filtering.It is very good in practicability.

Key words: Chinese spam;multiplayer filter;minimum risk Bayes;feature selection

摘 要:针对当前中文垃圾邮件过滤中存在的问题,提出了一种基于改进最小风险贝叶斯算法的多层次垃圾邮件过滤方法,并研究了其中关键应用技术。实验结果表明,所设计的多层次过滤算法不但在召回率和准确率上具有一定优势,还具有较高的过滤速率,实际应用性较强。

关键词:中文垃圾邮件;多层次过滤;最小风险贝叶斯;特征选择
DOI:10.3778/j.issn.1002-8331.2009.34.029 **文章编号**:1002-8331(2009)34-0094-04 **文献标识码**:A **中图分类号**:TP393.08

1 引言

随着 Internet 及其应用的快速发展,电子邮件已经成为现代通信的一种重要手段,给社会生产和生活带来极大便利。与此同时,大量垃圾邮件在网络中迅速蔓延,给网络服务系统的运行和用户带来了严重危害。因此,研究垃圾邮件自动过滤技术具有十分重要的实际应用意义。

目前,垃圾邮件过滤技术主要集中在三个层面,即邮件地址、邮件标题和邮件内容。在邮件地址过滤方面,主要包括黑名单技术、白名单技术和实时黑名单技术。但由于黑白名单收集的不完备性,实验证明单纯采用黑白名单技术,会出现较高的误报率和漏报率。在邮件标题过滤方面,多采用快速模式匹配算法将邮件标题文本串(或关键词)与已建立的垃圾标题库进行逐一匹配。但高明的垃圾邮件制造者往往故意对标题进行"粉饰性"修改,(如插入固定间隔字符、随机间隔字符、故意拼写错误、同义词替换等)以图绕过标题过滤,导致基于标题过滤算法的漏报率大大增加。

从实际应用分析,邮件的主要信息载体应该是邮件内容,因此对邮件内容进行过滤被认为是目前最有效的垃圾邮件过滤方法。基于内容的邮件过滤方法主要分为基于规则匹配和基于概率统计两类[1]。已有研究表明,基于规则匹配的过滤方法准确度较高,但当邮件特征数量和规则数目较大时,会导致算法执行时间较长,显著降低了垃圾邮件过滤的实时性。相对来说,基于概率统计的垃圾邮件过滤方法在处理速率上具有较大优势,但其准确性却依赖于具体实现算法。

贝叶斯(Bayes)[2]作为一种经典的概率统计算法,由于它在文本分类方面具有良好性能,使得它在垃圾邮件过滤方面也得到了十分广泛的应用。有实验结果表明,Bayes 算法在英文垃圾邮件过滤中能够达到 90%以上的准确率,且计算速度较快,可见该算法在垃圾邮件过滤应用中具备一定优势。

中文邮件与英文邮件在垃圾邮件过滤方面存在较大的差异,一是中文分词和特征选择具体更大难度,二是中文语义理解还处于研究初期。这些差异使得 Bayes 算法在中文垃圾邮件处理中表现并不理想,需要更深入地研究 Bayes 算法在中文垃圾邮件处理中的应用模型及其优化问题,设计出更高效的中文垃圾邮件过滤方法。

从上述分析可知,Bayes 算法在中文垃圾邮件过滤研究需要解决的问题主要存在于两个方面:一是如何设计有效的中文邮件过滤 Bayes 应用模型;二是 Bayes 算法的进一步改进和优化。前者主要解决中文邮件的预处理问题;后者主要以提高邮件过滤的准确率和速率为研究目标。基于以上思想,在深入分析基本 Bayes 算法的基础上,提出了一个以改进最小风险贝叶斯算法为核心的多层次中文垃圾邮件过滤模型,并研究了其中的关键应用技术。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60673161);福建省科技厅重点项目(No.2007H0023);福建省科技创新平台计划项目(No.2009J1007)。

作者简介:刘延华(1972-),男,讲师,研究方向为网络信息安全、计算智能等;陈国龙(1965-),男,博士,教授,博士生导师,研究方向为计算智能、计算机网络等。

收稿日期:2009-08-10 **修回日期**:2009-10-13

2　贝叶斯分类器

贝叶斯分类器的基本原理是首先将文本划分为若干特征项的形式，然后根据这些特征项来计算文本属于每个类别的概率（也称相似度），最后根据得到的相似度值进行判定，将文本划分到相似度值最大的类别中。

定义 1　假设文本 x 可以划分为 m 个类，则定义文本类 $C=\{c_1,c_2,\cdots,c_m\}$，其中 $m\geq 2$。

定义 2　假设文本 x 的特征项有 n 个，则将文本 x 定义为 n 维特征向量形式：$x=(w_1,w_2,\cdots,w_n)$，其中 w_i 表示文本 x 的第 i 个特征项的值，$n\geq 1$。

定义 3　由贝叶斯公式，文本 x 属于 c_k 类的概率计算公式为：

$$P(c_k|x)=\frac{P(x|C_k)P(C_k)}{P(X)},k=1,2,\cdots,m \tag{1}$$

其中，$P(c_k)$是类 c_k 的先验概率，可通过对已有分类文本进行训练学习获得。$P(c_k|x)$称为类 c_k 的后验概率，即文本 x 属于第 c_k 类的相似度，是文本分类的判定依据。$P(x)$可由下列全概率公式求得：

$$P(x)=\sum_{k=1}^{m}P(c_k)P(x|c_k) \tag{2}$$

可见，贝叶斯分类算法的关键就是计算 $P(x|c_k)$，$P(x|c_k)$称为类 c_k 的条件概率，也叫似然函数。$P(x|c_k)$的计算方法有很多，但当文本 x 的特征项数目 n 较大且特征项之间关系复杂时，计算 $P(x|c_k)$的过程会相当复杂，将大大影响文本分类的速率。

假设不考虑文本 x 的 n 个特征项之间关系，即假设 w_i 之间相互独立，那么有：

$$P(x|c_k)=P(w_1|c_k)\times P(w_2|c_k)\times\cdots\times P(w_n|c_k)=\prod_{j=1}^{n}P(w_j|c_k) \tag{3}$$

其中，$P(w_j|c_k)$可以利用训练集计算估计计算。这就产生了朴素贝叶斯（Naïve Bayes，也称简单贝叶斯）分类器，它是贝叶斯分类器的基本形式。从上式可看出，朴素贝叶斯算法使得 $P(x|c_k)$的计算简单化，大大提高了计算效率，因此朴素贝叶斯算法得到了广泛应用。

3　垃圾邮件贝叶斯过滤模型

垃圾邮件过滤就是将邮件分为垃圾邮件和非垃圾邮件两种类型，即 $C_{email\_filter}=\{c_{spam},c_{ham}\}$，其中 c_{spam} 表示垃圾邮件类，c_{ham} 表示正常邮件类。

根据贝叶斯分类器原理，基于贝叶斯算法的垃圾邮件过滤基本过程如下：

步骤 1　捕获邮件数据包，根据邮件协议将邮件内容恢复生成邮件内容文本 x；

步骤 2　将邮件内容文本 x 进行特征词划分；

步骤 3　对划分形成的特征词进行筛选，去除影响因子较小的特征词，减少特征词数量，并将邮件内容文本 x 表示成特征向量形式；

步骤 4　由贝叶斯分类器分别计算 $P(c_{spam}|x)$和 $P(c_{ham}|x)$的值；

步骤 5　根据得到的两个值进行垃圾邮件判定。可采用以下三种方法来判定垃圾邮件：

（1）由 $P(c_{spam}|x)$的值独立判定：设定一个阈值 r_{spam}，满足 $0\leq r_{spam}\leq 1$，当 $P(c_{spam}|x)\geq r_{spam}$ 时，则判定 x 对应的邮件为垃圾邮件。

（2）由 $P(c_{spam}|x)$的值独立判定：设定一个阈值 r_{ham}，满足 $0\leq r_{ham}\leq 1$，当 $P(c_{ham}|x)\leq r_{ham}$ 时，则判定 x 对应的邮件为非垃圾邮件。

（3）由 $P(c_{spam}|x)$与 $P(c_{ham}|x)$的差值判定：定义 $\Delta p=P(c_{spam}|x)-P(c_{ham}|x)$，设定一个阈值 $r_{\Delta p}$，当 $\Delta p\geq r_{\Delta p}$ 则判定文本 x 对应的邮件属于垃圾邮件；反之，判定该邮件为非垃圾邮件。此时，$r_{\Delta p}$ 选择的优劣对于垃圾邮件判定将会产生重要影响。

通过分析实际邮件内容，发现邮件内容的特征项具有一定书写规律，特征项之间也具有较大的语义相关性。因此，朴素贝叶斯算法直接应用到垃圾邮件过滤中显得不尽合理。虽然如此，基于朴素贝叶斯的垃圾邮件过滤模型仍然被广泛应用，其邮件判定准确性达到了 90%以上，成为邮件内容过滤的主要方法之一。

4　基于改进风险贝叶斯的多层次邮件过滤模型

4.1　最小风险贝叶斯算法及其改进

从实际应用来看，朴素贝叶斯算法完全能够满足普通用户对邮件过滤的需求。但对于邮件安全性要求较高的机构组织或个人，一封重要邮件被误判为垃圾邮件而被过滤掉所带来的损失往往比 10 封（甚至更多）垃圾邮件被误判为合法邮件要严重的多，也就是说此类用户宁可接受多封垃圾邮件也不愿意漏掉一封正常邮件，以降低由于误判可能带来的风险和损失。

针对这种情况，有研究者引入了风险因子 λ 对朴素贝叶斯算法进行改进，这就产生了最小风险贝叶斯分类器模型[3]。风险因子 λ 表示把正常邮件错判成垃圾邮件的代价是把垃圾邮件判为正常邮件的 λ 倍，通常 $\lambda\geq 1$。只有当 $\frac{P(c_{spam}|x)}{P(c_{ham}|x)}\geq\lambda$ 时，才判定邮件 x 为垃圾邮件。

可见，风险因子 λ 的值越大则误判的可能性越小，那么误判的风险也就越小。当 $\lambda=1$ 时，则风险因子 λ 不起作用，即转化为朴素贝叶斯算法。因此，如果用户能够接受较多垃圾邮件，而误判风险承受能力较小，则考虑将 λ 设置为一个较大值；反之则可以设置为一个较小值，减少用户垃圾邮件的接收量。Androutsopoulos Ion 等人对最小风险贝叶斯算法在英文垃圾邮件过滤做了深入研究和实践分析，实验结果表明，当 $\lambda\geq 999$ 时，其准确率可高达 100%。

从上述分析可知，风险因子 λ 值的大小对最小风险贝叶斯分类器的过滤效果影响较大，因此 λ 的设置成为一个关键问题。目前，λ 的值多由专家指定或由已有邮件训练学习获得，缺乏动态更新性，这样可能导致 λ 值的设定出现滞后性，随着时间的推移，λ 值的滞后性会逐渐降低邮件分类的准确性。

针对风险因子 λ 值的更新问题，将上文中基于 Δp 的判定思想引入到最小风险贝叶斯算法中[4]，实现了 λ 值的动态更新。改进后的最小风险贝叶斯算法将垃圾邮件判定分为三个步骤：

步骤 1　当 $\Delta p\leq 0$ 时，判定该邮件为正常邮件；

步骤 2　当 $\frac{P(c_{spam}|x)}{P(c_{ham}|x)}\geq\lambda$ 时，判定该邮件为垃圾邮件；

步骤 3　当 $\frac{P(c_{spam}|x)}{P(c_{ham}|x)}<\lambda$ 且 $\Delta p>0$ 时，则系统不作垃圾邮件判定，而是将该邮件提交给用户，由用户进行人工判定。当用户判定结束后，系统根据人工判定结果将该邮件加入邮件训练集，执行贝叶斯训练算法。显然，训练后所得到的风险因子 λ 将更加能够适应用户的要求，这样就实现了 λ 的自适应更新。

4.2 多层次邮件过滤模型设计

与基于邮件内容的过滤方法相比,基于邮件地址和邮件标题的过滤方法执行速率更快,而黑白名单技术也具有更高的判定准确性。因此,为了同时获得较好的邮件判定准确性和执行速率,提出了一个多层次垃圾邮件过滤模型,具体结构如图 1 所示。

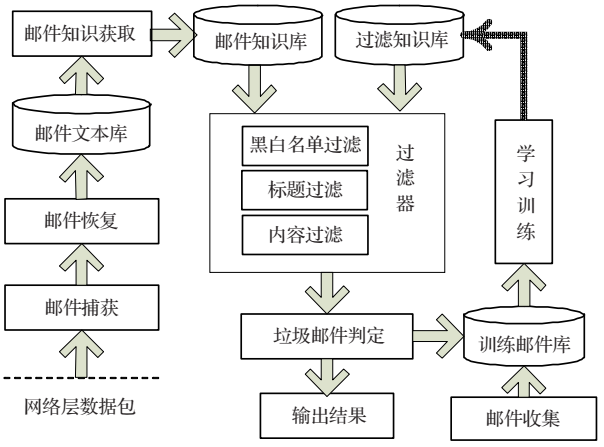

图 1 多层次垃圾邮件过滤模型

该邮件过滤模型主要包括三个部分。

4.2.1 生成过滤知识库

收集已有垃圾邮件和非垃圾邮件,对这些邮件进行规则学习,生成和优化邮件过滤知识库。过滤知识库包括黑白名单库、关键词匹配规则库、贝叶斯知识库。

4.2.2 邮件的获取和预处理

从网络中截获邮件协议数据包,恢复生成邮件内容文本,提取接收和发送邮件地址、邮件标题和邮件内容。进一步对邮件内容文本进行中文分词和特征词筛选,表示为邮件内容特征向量形式。

4.2.3 垃圾邮件判定

从图 1 可看出,垃圾邮件判定由三个过滤步骤构成,首先执行黑白名单过滤,其次执行基于标题的快速匹配,最后执行改进最小风险贝叶斯过滤,并给出判定结果。

多层次过滤模型的基本执行流程如图 2 所示。

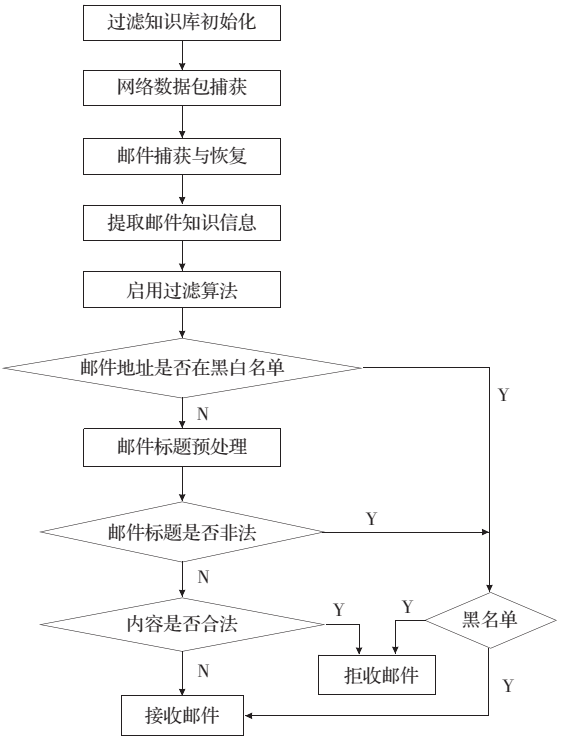

图 2 多层次过滤的基本流程

5 关键模块的设计

5.1 中文分词

为了实现中文邮件内容的分析过滤,首先要对还原后的邮件内容文本进行特征提取,即对邮件文本进行中文分词,并转换得到邮件文本的特征向量形式。由于中英文分词具有较大差异性,通过实验比较,借鉴了中文分词工具 SharpICTCLAS,并将其中的一些关键模块进行了改编,集成到多层次垃圾邮件过滤系统中。实验中,该算法达到了较高的中文分词准确率,这有力保证了中文邮件内容过滤的准确性。

所实现的中文分词算法基本流程[5]如图 3 所示。

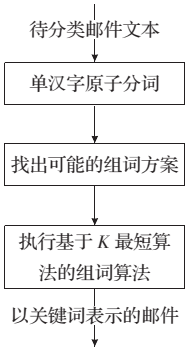

图 3 中文分词基本流程

5.2 邮件特征项的选择算法

经过中文分词处理后,每一封邮件的邮件内容被划分为若干个中文词语,每个中文词语作为一个邮件内容特征项。当一封邮件较大时,中文词语的个数也会增大。当中文词语数量巨大时,若直接进行垃圾邮件分析则花费的时间代价较大。因此,在垃圾邮件判断之前,要将邮件划分得到的中文词语进行有效筛选,去除对垃圾邮件判定影响较小的词语,以降低邮件内容特征向量的维数,加快垃圾邮件分析速率。

该文采用了 TFIDF 算法[6],该算法是以特征词在文档 D 中出现的次数与包含该特征词的文档数之比作为该词的权重,即

$$w_i=\frac{TF_i(t,D)}{DF_i(t)} \tag{4}$$

其中,w_i 表示第 i 个特征词 t 的权重,$TF_i(t,D)$表示特征词 t 在文档 D 中的出现次数,$DF_i(t)$表示其他文档包含 t 的次数。w_i 越大,表示特征词 t 在文档 D 中出现的概率越高,而在其他文档出现的概率就越低,说明特征词 t 对于文档 D 的影响大。根据垃圾邮件判定需求,系统应保留 w_i 值大的特征词,去除 w_i 小于设定阈值的特征词。该算法的基本处理流程如图 4 所示。

5.3 贝叶斯算法的功能设计

把贝叶斯过滤模型从功能上划分为邮件预处理、贝叶斯算法实现、分类过滤器等几个主要模块,如图 5 所示。

5.3.1 邮件预处理模块

该模块根据 RFC822 和 MIME 协议对邮件进行解析,得出邮件的主题和内容,并将邮件内容表示成邮件特征词向量。对于较大邮件进一步执行特征词选择,去除影响系数较小的特征

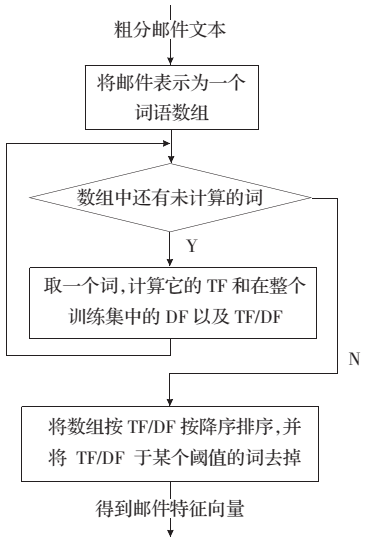

图 4　特征选取过程

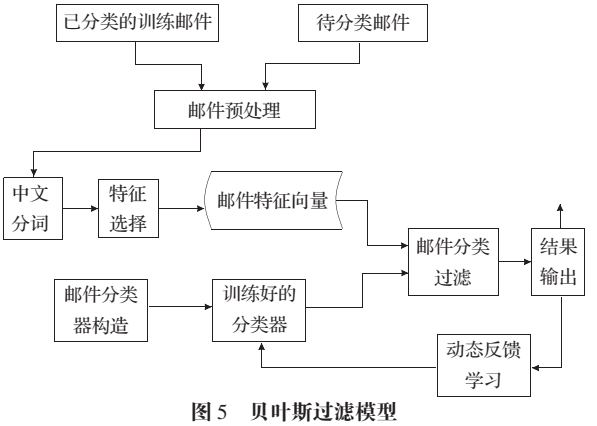

图 5　贝叶斯过滤模型

词，得到垃圾邮件判定要处理的邮件特征向量。

5.3.2　贝叶斯分类器及训练

该模块根据改进的最小风险贝叶斯模型构造出贝叶斯邮件分类器，并进行参数的初始化。利用该分类器将已分类的邮件进行分类学习，训练该贝叶斯分类器，优化各种参数，提高分类器的效率。

5.3.3　新邮件过滤

将新接收的邮件向量输入训练好的贝叶斯分类器，得到判定结果。对于疑似垃圾邮件，进行人工判定后由系统学习模块进行反馈型学习，实现系统运行中的动态更新，使得贝叶斯分类器具有更好的分类性能。

6　实验与分析

6.1　分类器评价体系

为了能客观、定量地评价一个邮件分类器的分类性能，通常借用文本分类和信息检索领域的相关技术指标来对垃圾邮件分类器的性能进行评价。

设测试集合中共有 N 封邮件，定义变量 A、B、C 和 D，满足 $N=A+B+C$ 各变量的含义如表 1 所示。

表 1　评价体系变量表　　封

	实际为垃圾邮件	实际为合法邮件
系统判定为垃圾邮件	A	B
系统判定为合法邮件	C	D

下面定义两个技术评价指标：

定义 4　召回率(Recall)：$Recall=\frac{A}{A+C}\times100\%$，即垃圾邮件检出率。这个指标反映了过滤系统发现垃圾邮件的能力，召回率越高，“漏网”的垃圾邮件越少。

定义 5　准确率($Precision$)：$Precision=\frac{A}{A+B}\times100\%$，即垃圾邮件检对率。这个指标反映了非垃圾邮件被误判为垃圾邮件的可能性。准确率越高，说明系统将正常邮件误判为垃圾邮件的可能性就越小。

6.2　实验结果分析

实验采用 CCERT 提供的中文邮件数据集 CDSCE(CCERT Data Sets of Chinese Emails)作为语料库。从 2005 年 7 月份数据集中随机抽取了 1 800 封垃圾邮件和 1 200 封正常邮件，即共 3 000 封实验样本邮件。将这些邮件分为 10 份，每份 300 封，每次取一定份数作为训练集，另在剩余的邮件中取一份作为测试集，进行实验。

测试过程中，分别单独使用朴素贝叶斯算法、最小风险贝叶斯算法和多层次过滤算法进行实验比较，其中最小风险贝叶斯和多层次过滤算法中的风险因子 λ 值设置为 λ=999。实验中不断调整训练样本集的大小，对比三种算法的实验结果。其召回率和准确率对比如图 6、图 7 所示。

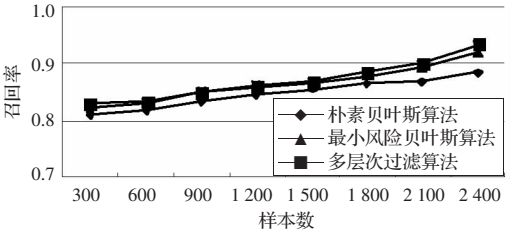

图 6　三种算法召回率结果对比

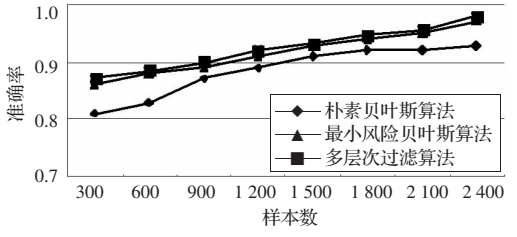

图 7　三种算法准确率结果对比

从实验结果可以看出，当风险因子 λ=999 时，最小风险贝叶斯算法和多层次过滤算法在召回率和准确率上都比朴素贝叶斯算法更具有优势。另外，多层次过滤算法由于 λ 值较大而使得有些垃圾邮件被判定为疑似垃圾邮件，比单纯的最小风险贝叶斯降低了查全率，但由于多层次过滤方法集成了基于邮件地址和邮件标题的过滤方法，因此总体上多层次过滤算法还是稍微优于最小风险贝叶斯算法。在训练样本选取方面，由实验结果可看出，随着训练样本数的增多，三种算法的过滤效果都得到相应提高。根据最小风险贝叶斯算法模型，当风险因子 λ 的值增大时，所设计算法在召回率上会出现一定程度的下降，而其准确率会相应提高；反之，当风险因子 λ 的值减小时，会提高算法的召回率，而相应降低了判定的准确率。实验中发现，当风险因子 λ 的值为 999 时判定准确率能够达到 95%以上；当 λ 值大于 999 时，算法对于中文垃圾邮件的判定正确率的提高不

(下转 129 页)