基于改进贝叶斯原理的垃圾邮件过滤算法研究*

袁连海　李湘文　徐　晶
（成都理工大学工程技术学院　乐山　614000）

摘　要　为了提高垃圾邮件过滤系统的对邮件过滤的准确性和返回率，论文改进了传统的贝叶斯定理。提出一种改进的垃圾邮件过滤方法，该方法使用基于单词提取特征值和使用特征向量来描述频率。模型降低了垃圾邮件的错误率，总体上提高了系统的过滤性能。与传统贝叶斯公式的假设不同，系统为垃圾邮件样本的每个特征值分配不同的权值，降低了的垃圾邮件判断误差。实验结果表明，论文提出的垃圾邮件过滤方法能够显着提高准确性和返回率，系统性能得到了较大改进。

关键词　贝叶斯原理；邮件过滤；特征向量
中图分类号　TP301.6　　**DOI**：10. 3969/j. issn. 1672-9722. 2020. 03. 002

An Improved Anti-Spam Filtering Method Based on Bayesian

YUAN Lianhai　LI Xiangwen　XU Jing
（Engineering & Technical College，Chengdu University of Technology，Leshan　614000）

Abstract　In order to improve the accuracy and return rate of the spam filtering system for mail filtering，the paper improves the traditional Bayes´ theorem. An improved spam filtering method is proposed，which uses word-based feature extraction and feature vectors to describe frequency. The model reduces the error rate of spam and improves the overall filtering performance of the system. Different from the assumption of the traditional Bayesian formula，the system assigns different weights to each feature value of the spam sample，which reduces the spam judgment error. Experimental results show that the spam filtering method proposed in this paper can significantly improve the accuracy and return rate，and the system performance has been greatly improved.

Key Words　Bayesian principle，spam filtering，feature vector
Class Number　TP301.6

1　引言

电子邮件因为其价格低廉、使用方便、支持非实时通信等优点，逐渐成为当今主要的网络应用。电子邮件流行的结果是出现了大量的垃圾邮件，从而大大影响了正常的通信，垃圾邮件严重制约了人们使用电子邮件。中国互联网协会将具有以下特点的电子邮件划分为垃圾邮件[1]：

1）垃圾邮件能够隐藏发件人身份、地址和标题等信息；

2）垃圾邮件对于收件人不能拒绝；

3）垃圾邮件包含欺诈信息；

4）具有传播性质的电子邮件，如广告、电子杂志和宣传材料，以任何形式发送给收件人，但未事先征得他们的同意。

许多国家制定了反对垃圾邮件的法律法规，我国也在过去推出了一些限制垃圾邮件的法律，但是，因为垃圾邮件具有一定的利益驱动，目前垃圾邮件依然很严重。除了国家制定法律防止垃圾邮件发送外，很多邮件服务器使用技术方法来进行垃圾邮件过滤，例如增加黑名单、采用敏感词语过滤规则、使用白名单等方法。目前比较流行的垃圾邮件过滤方法有决策树、Boosting、K 近邻、支持向量机、贝叶斯原理等[3]。

* **收稿日期**：2019 年 9 月 3 日，**修回日期**：2019 年 10 月 10 日
基金项目：国家自然科学基金面上项目（编号：11375055）资助。
作者简介：袁连海，男，硕士研究生，讲师，研究方向：计算机技术，计算机信息安全。李湘文，男，硕士研究生，副教授，研究方向：人工智能，机器人技术。徐晶，男，硕士研究生，助教，研究方向：计算机技术，计算机信息安全。

对垃圾邮件进行过滤本质上是将邮件按照一定的规则进行分类,然后将邮件分类成正常邮件和垃圾邮件,当今最常用的过滤方法基本上都是基于贝叶斯原理模型进行设计的。该算法是自适应具有统计功能的机器学习算法。目前包括几种基于贝叶斯原理的过滤算法:基于多变量、基于多项式和基于布尔类型属性,这些方法总体上是采取对邮件进行分类,从而判断收到的邮件是不是属于垃圾邮件。

朴素贝叶斯需要假设所有的特征值都有同样的重要性,而且相互独立。这种假设前提和现实实际情况是有一定差别,从而导致判断错误的情形出现。研究人员为了解决垃圾邮件误判问题,在各个方面对贝叶斯原理进行提高和改进。网络分类器法通过直接对邮件结构进行一定的扩展,采用增加不同特征值间的依赖关系;特征值权重法是一种基于权向量的过滤算法。还有一些研究人员研究了使用二进制方法表示邮件特征向量,使用特征值提取特征项,贝叶斯原理通过计算一个邮件是否属于垃圾邮件的概率来进行垃圾邮件判断,具有简单、有效等特点。

文章提出了改进的贝叶斯公式的垃圾邮件过滤系统,采用特征项的单词出现的频率来表示特征向量,使用单词的特征项,系统在垃圾邮件识别准确性方面有很大提高。

2 过滤器设计

2.1 贝叶斯分类器

传统的贝叶斯分类器的功能是将邮件进行分类。图1表示传统的贝叶斯分类器示意图,图中节点M表示类别集合的某个元素,节点$x=(x1,x2,x3,\cdots,xn)$是分类的特征向量。假设有某一待分类的样本S,分类特征值向量是$x=(x1,x2,x3,\cdots,xn)$,那么样本S属于类别M_k的概率为$P(M=M_k|X=x)$

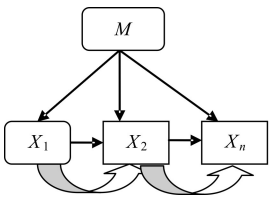

图1 贝叶斯分类器

所以,样本S属于类别M_k需要满足以下式(1):

$$P(F=M_k|X=x)=\mathrm{Max}(P(M=M_1|X=x),P(M=M_2|X=x),\cdots,P(M=M_n|X=x_1)) \quad (1)$$

根据贝叶斯公式和全概率公式,推到出式(2):

$$P(M=M_n|X=x)=\frac{P(X=x|M=Mn)P(M=Mn)}{P(X=x)} \quad (2)$$

在式(2)中,因为$X=x$的概率和类别M_n无关,所以只需要计算出$P(X=x|M=Mn)$以及$P(M=Mn)$两个概率。通常第一个值通过似然函数获得,而第二个数值可以通过经验取得[7]。在实际应用中,常常假设特征向量的分量是相互独立的,也就是说各个分量之间是没有关联的。所以,图1贝叶斯分类器可以简化成图2结构。

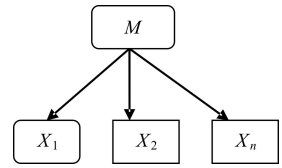

图2 简化的贝叶斯分类器

根据$P(X=x|M=M_m)=\prod_{i=1}^{m}P(X_i=x_i|M=M_m)$,我们推导得到式(3):

$$P(M=M_m|X=x)=\frac{\prod_{i=1}^{m}P(X_i=x_i|X=x)P(M=M_m)}{P(X=x)} \quad (3)$$

2.2 贝叶斯过滤器改进

贝叶斯过滤器将邮件划分成两类:正常邮件和垃圾邮件。假定某个电子邮件M_i的特征向量表示为$X_{mi}=(x_1,x_2,\cdots,x_n)$,则邮件属于类别$M_n$(在本论文邮件系统中邮件只分为两类:垃圾邮件用$M_{垃圾}$表示,正常邮件用$F_{正常}$表示)的概率计算公式(4):

$$P(M_k|x_{mi}=\frac{P(M_k)P(x_{mi}|M_k)}{\sum_{km(正常,垃圾)}P(M_k)P(x_{mi}|M_k)} \quad (4)$$

由于朴素贝叶斯公式中假定了特征向量的各个特征值是独立的,所以式(4)进一步简化为式(5)。

$$P(M_k|x_{mi})=\frac{P(M_k)\prod_{i=1}^{n}P(x_{mi}|M_k)}{\sum_{km(正常,垃圾)}P(M_k)P(x_{mi}|M_k)} \quad (5)$$

上述贝叶斯公式邮件过滤器存在两种情况:第一种误判是把垃圾邮件当成正常邮件,另外一种是将正常邮件看成是垃圾邮件。其中后一种判断会带来更大的损失,为了提高垃圾邮件的判断正确率,避免把正常的电子邮件看成是垃圾邮件,我们引入一个参数,只有当垃圾邮件的概率大于是正常邮件的一定倍数时,才会将正常邮件判断为垃圾邮件。

2.3 系统评价指标

所有贝叶斯公式的垃圾邮件分类基本上都是将邮件文本划分成两种类别。通常使用的指标包括正确率以及查全率。正确率表示系统正确判断的垃圾邮件数量和垃圾邮件识别为正常以及正常邮件识别为垃圾的比值。查全率表示为正确识别的垃圾邮件的数量和样本中垃圾邮件总数量的比值。

正确率=正确识别的垃圾邮件数量/(垃圾识辊为正常+正常识别为垃圾)

查全率=正确识别的垃圾邮件数量/垃圾邮件总数

3 系统邮件预处理

3.1 邮件样本选择

本系统采用的样本数据是成都理工大学工程学院工程训练中心邮件服务器中包含的历史邮件作为样本。其中总共邮件数目为2000 封电子邮件,经过人工识别分类1100封电子邮件为垃圾邮件,剩下的900封为正常邮件。研究中为了重点测试系统的过滤性能,我们将忽略附件和标签。

3.2 邮件预处理

实验过程中使用MIME以及RFC822电子邮件协议进行了邮件预处理。处理内容包括邮件的格式和正文内容进行分析提取,分离出电子邮件的正文以及邮件主题。大部分邮件过滤方法在挑选邮件特征值的数值的时候常常使用二进制数据来表示,其优点显然简单方便,缺点是对某个特征值在邮件中出现的频率不能清楚表示。

邮件的特征值通常使用电子邮件的英文单词、短语以及其他非文本参数。系统因为已经对样本数据做了处理,没有附件以及其他的HTML标记,本系统采用邮件单词作为特征值可以重点研究系统性能是否提升。

3.3 提取邮件特征值

在对样本进行处理过程中,为了减少特征向量的维度,系统对邮件提取特征值,这样可以大幅提高邮件过滤系统的过滤速度和准确率。经常使用的特征值提取方法包含词条和类别相互信息、词条统计、词条期望的熵、文本权重等方法[8]。

系统采用的提取特征值的步骤为

1)首先,在样本邮件里面提取出所有单词,然后删掉冠词、介词和数词(例如,邮件里面的如in,of,the,an,a,he以及数量词汇等),接着把单词的不同表现形式如复数形式、现在时、动名词、过去式、形容词的比较级以及最高级等形式进行标准化处理;

2)接着统计每个单词在邮件中产生的权重,并对权重降序排列;

3)依据排序结果,将排列在靠前的单词作为邮件的特征值;

4)最后把全部邮件样本的特征值,得到特征向量。

4 实验结果及分析

系统中我们将所有邮件样本数据随机分为10组相互没有关联的集合,每个集合数据中包含的邮件数目相同。系统对样本数据进行了10机器学习以及测试,实验中挑选一个集合当成测试集,其他9个集合合并作为训练集。

系统实验学习来得到邮件分类器是为了防止测试的随机和偶然性,并且使用相应的测试集和对该子集进行验证测试。每次测试中分别计算其邮件分类器的准确率以及查全率,最后采用10次测试的平均值作为测试结果。

表1以及表2是采用朴素贝叶斯算法得到的测试数据。在系统邮件测试过程中,使用两种表示方法:一是通过判断特征单词是否出现在邮件中表示,另一种采用特征单词在样本数据邮件中出现的频数来表示,其目的是为了比较不同的邮件特征向量表示方法对邮件过滤系统性能的影响,实验结果见表3所示。

表1 基于朴素贝叶斯

	正常邮件数目(封)	垃圾邮件数目(封)
判断为正常邮件	486	38
判断为垃圾邮件	42	495

表2 改进的贝叶斯

	正常邮件数目(封)	垃圾邮件数目(封)
判断为正常邮件	499	21
判断为垃圾邮件	30	502

表3 不同特征词的性能参数

	正确率	查全率
基于特征词二进制	97.47	80.12
基于特征词频数	96.32	70.14

根据上述实验结果我们可以得出结论,改进的贝叶斯原理和朴素贝叶斯方法相比,改进的算法的性能相对要好一些,改进后的垃圾邮件过滤系统在正确率和查全率两个方面都有一定的提升。虽然朴素的贝叶斯公式能够降低计算工作量,能够在一定程度上提升邮件系统的运行效率,但是,因为没

有考虑很多其他对邮件过滤比较有用的信息,而且,传统的贝叶斯垃圾邮件过滤方法仅仅把邮件的内容作为关键词的没有顺序的向量空间,也没有将词和词之间的相互关系考虑进去。改进的贝叶斯过滤系统将词和次之间部分依赖关系进行考虑。还有基于单词频数的向量方法比二进制向量的表示法要好一些,这是由于这种方法将较多的信息传递给了邮件过滤器,从而系统的整体过滤性能有了一定提升。

本垃圾邮件过滤系统的不足之处是:改进后的贝叶斯垃圾邮件过滤器的系统时间开心要比传统的方法要大,根据实验结果可以知道,改进的系统时间开销可以达到10倍的传统的过滤方法,也就是说,如果测试的邮件内容只有一条语句,改进的系统将退化成朴素的贝叶斯方法。

5 结语

系统使用单词的特征值在邮件中出现的频数作为特征向量,使用改进后的贝叶斯原理方法设计的垃圾邮件过滤器,根据单词频数顺序来提取邮件的特征值,实验研究结果显示,系统的正确率和查全率都有一定提升。

本次实验在选取样本数据时,没有选择邮件附件以及其他信息,只是对文本的正文内容的进行过滤。另外,垃圾邮件过滤系统还没有对携带病毒的电子邮件以及在附件中包含病毒的电子邮件进行考虑,上述问题是垃圾邮件过滤系统应该要解决的,下一步的研究思路包括如何将目前基于规则的过滤、白名单和黑名单、机器学习算法在垃圾邮件中的应用等方面进行研究,另外,系统的安全性和如何处理邮件病毒也是需要考虑的问题。

参考文献

[1] http://www.isc.org.cn/[EB/OL].

[2] 郑炜,沈文,张英鹏. 基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究[J]. 西北工业大学学报,2010,4(28):622-627.
ZHENG Wei, SHEN Wen, ZHANG Yingpeng. Research on spam filter based on improved naive Bayes algorithm [J]. Journal of Northwestern Polytechnic University, 2010,4(28): 622-627.

[3] 詹川,卢显良,周旭,等. 基于贝叶斯公式的垃圾邮件过滤方法[J]. 计算机科学,2005,2(32):73-75.
ZHAN Chuan, LU Xianliang, ZHOU Xu, et al. Spam filtering method based on Bayesian formula [J]. Computer Science, 2005, 2(32): 73-75.

[4] 计宏. 改进贝叶斯垃圾邮件过滤技术的研究[J]. 计算机测量与控制,2013,8(21):2181-2184.
JI Hong. Research on improving Bayesian spam filtering technology[J]. Computer Measurement & Control, 2013, 8(21): 2181-2184.

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

[6] 林伟,柳荣其,徐熙. 一种基于N-GRAM的垃圾邮件过滤方法研究[J]. 计算机应用与软件,2010,27(2):121-123.
LIN Wei, LIU Rongqi, XU Xi. A N-GRAM based spam filtering method[J]. Computer Applications and Software, 2010, 27(2): 121-123.

[7] GRAHAM P. A plan for spam[EB/OL]. [2012-02-01]. Http: //Paul graham. com/spam.htm1.

[8] 王涛,裘国永,何聚厚. 基于改进Naive Bayes的垃圾邮件过滤模型研究[J]. 计算机工程与应用,2007,4(13):186-190.
WANG Tao, YAN Guoyong, HE Juhou. Based on improved Naive Bayes spam filtering model [J]. Computer Engineering and Applications, 2007, 4(13): 186-190.

[9] 陈晋川,陈治璋,贾洪明,等. 基于模式的贝叶斯垃圾邮件过滤的研究与实现[J]. 计算机工程与应用, 2006(6):172-175.
CHEN Jinchuan, CHEN Zhizhen, JIA Hongming, et al. Research and implementation of pattern-based Bayesian spam filtering [J]. Computer Engineering and Applications, 2006(6): 172-175.

[10] 王申. 基于内容的垃圾邮件过滤技术若干研究[D]. 北京:中国科学院,2005.
WANG Shen. Research on content-based spam filtering technology[D]. Beijing: Chinese Academy of Sciences, 2005.

[11] Richard O Duda, Peter E Hart, David G Stork. Pattern Classification[M]. Second Edition, 2003.

[12] 谢冲锋,李星. 基于序列的文本自动分类算法[J]. 软件学报,2002.
XIE Chongfeng, LI Xing. Sequence-based automatic text classification algorithm[J]. Journal of Software, 2002.

[13] 丁文斌,李斌,罗浩. 基于改进贝叶斯的垃圾邮件过滤系统设计与实现[J]. 计算机工程与应用,2005,18 127-130.
DING Wenbin, LI Bin, LUO Hao. Design and implementation of spam filtering system based on improved Bayesian [J]. Computer Engineering and Applications,

(下转第712页)

[J]. 计算机集成制造系统,2014,20(5):1133-1140.
SHI Hui, ZENG Jiangcao. Preventive maintenance strategy based on life prediction[J]. Computer Integrated Manufacturing Systems,2014,20(5):1133-1140.

[4] De Smidt-Destombes K S, Van Harten A. Availability of k-out-of-N systems under block replacement sharing limited and spares and repair capacity[J]. General Information,2007,107(2):404-421.

[5] DOUER N, YECHIALI U. Optimal repair and replacement in Markovian systems[J]. Stochastic Models, 1994, 10(1):253-270.

[6] 胡飞. 保修产品最优预防维修策略的建模[J]. 统计与决策,2008(24):57-60.
HU Fei. Modeling of Optimal Preventive Maintenance Strategies for Warranty Products[J]. Statistics and Decision,2008(24):57-60.

[7] WIJNGAAR J. The effect of interstage buffer storage on the output of two unreliable production units in series with different production rates[J]. American Institute of Industrial Engineers Transactions,1979,11(1):42-47.

[8] BOUSLAH B, GHARBI A, PELLERIN R. Joint optimal lot sizing and production control policy in an unreliable and imperfect manufacturing system [J]. International Journal of Production Economics,2013,144(1):143-156.

[9] 吕文元,郑睿. 基于时间延迟维修理论检查模型的比较研究[J]. 管理科学,2007(01):18-21.
LU Wenyuan, ZHENG Rui. Comparative Study on the Theory Check Model Based on Time Delay Maintenance [J]. Management Science,2007(01):18-21.

[10] Jack N, Murthy D N P, Iskandar B P. Comments on "Maintenance policies with two-dimensional warranty" [J]. Reliability Engineering and System Safety,2003,82(1):105-109.

[11] JACK N, VAN DER DUYN SCHOUTEN F A. Optimal repair-replace strategies for a warranted product[J]. International of Production Economics,2000,67:95-100.

[12] DIMITRAKOS T D, KYRIAKIDIS E G. A semi-Markov decision algorithm for the maintenance of a production system with buffer capacity and continuous repair times [J]. International Journal of Production Economics, 2008,111(2):752-762.

[13] SALAMEH M K, GHATTAS R E. Optimal just-in-time buffer inventory for regular preventive maintenance[J]. International Journal of Production Economics, 2001, 74(1/2/3):157-161.

[14] 桂云苗,龚本刚,程幼明. 保修期限促销策略下供应链协调[J]. 软科学,2012,26(05):67-70.
GUI Yunmiao, GONG Bengang, CHENG Youming. Supply Chain Coordination under the Warranty Period Promotion Strategy[J]. Soft Science,2012,26(05):67-70.

[15] DAHANE M, CLEMENTZ C, REZG N. Effects of extension of subcontracting on a production system in a joint maintenance and production context[J]. Computers and Industrial Engineering,2010,58(1):88-96.

(上接第516页)

2005, 18:127-130.

[14] Sahami, Dumais S, et al. A Bayesian Approach to Filtering Junk E-Mail. Learning for Text Categroization Zhen Chenggang, Li Baocai. Data Storage Scheduling in Cloud Computing Environment Algorithm research [J]. Information Security and Technology, 2015, 6(12):55-62.

[15] http://www.baidu.com[EB/OL].

[16] Androutsopoulos I, Sakkis G, Paliouras G, et al. Learning to Filter Spam E-Mail[C]//European Conference on Principles and Practice of Knowledge Discovery in Databases.Lyon,France,2000:1-13.

[17] Aawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large DataBases[C]//Proceedings of 20th International Conference on Very Large Data Bases (VLDB 1994). Santiagode Chile, Morgan Kaufmann, 1994:487-499.

[18] Park J S, Chen MS, Yu P S. An Effective Hash-Based Algorithm for Mining Association Rules[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'95). San Jose, 1995:175-186.

[19] Savasere A, Qmiecinski E, Navathe S. An Efficient Algorithm for Mining Association Rules in Large Databases [C]// 21st VLDB Conf. Zurich, Switzerland, 1995: 432-444.