国内图书分类号：TP391.1

工学硕士学位论文

基于在线学习的垃圾邮件过滤技术研究

硕士研究生：沈跃伍
导　　师：孙广路
申请学位级别：工学硕士
学科、专业：计算机应用技术
所在单位：计算机科学与技术学院
答辩日期：2012年3月
授予学位单位：哈尔滨理工大学

Classified Index: TP391.1

Dissertation for the Master Degree in Engineering

Research on Online Learning Based Spam Filtering

Candidate:	Shen Yuewu
Supervisor:	Sun Guanglu
Academic Degree Applied for:	Master of Engineering
Specialty:	Computer Applied Technology
Date of Oral Examination:	March, 2012
University:	Harbin University of Science and Technology

哈尔滨理工大学硕士学位论文原创性声明

本人郑重声明：此处所提交的硕士学位论文《基于在线学习的垃圾邮件过滤技术研究》，是本人在导师指导下，在哈尔滨理工大学攻读硕士学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文研究工作做出贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签名：沈跃伍　　日期：2012 年 3 月 14 日

哈尔滨理工大学硕士学位论文使用授权书

《基于在线学习的垃圾邮件过滤技术研究》系本人在哈尔滨理工大学攻读硕士学位期间在导师指导下完成的硕士学位论文。本论文的研究成果归哈尔滨理工大学所有，本论文的研究内容不得以其它单位的名义发表。本人完全了解哈尔滨理工大学关于保存、使用学位论文的规定，同意学校保留并向有关部门提交论文和电子版本，允许论文被查阅和借阅。本人授权哈尔滨理工大学可以采用影印、缩印或其他复制手段保存论文，可以公布论文的全部或部分内容。

本学位论文属于

保密 □ ，在　　年解密后适用授权书。

不保密 ☑ 。

（请在以上相应方框内打 √）

作者签名：沈跃伍　　日期 2012 年 3 月 14 日

导师签名：孙广路　　日期：2012 年 3 月 15 日

基于在线学习的垃圾邮件过滤技术研究

摘 要

电子邮件给人们的生活和工作带来极大的便利，但大规模的垃圾邮件严重影响了邮件正常使用。垃圾邮件消耗大量网络资源，损害用户利益，还会被一些别有用心的人用来散播虚假消息，危害社会安定。因此，垃圾邮件过滤技术已经成为当前研究普遍关注的热点问题。

本文研究了基于机器学习理论的垃圾邮件过滤技术。由于该过滤技术具有正确率高，成本低等特点，已成为解决垃圾邮件过滤问题的主流方法。本文的研究内容主要分为以下几个部分：

首先，研究了基于在线学习的垃圾邮件过滤技术的框架和过滤模式，并实现了基于朴素贝叶斯、基于逻辑回归和基于在线支持向量机等三种模型的垃圾邮件过滤器，并从过滤器消耗的时间和过滤性能等方面评价三种过滤器的优缺点。

其次，研究了面向邮件过滤的特征工程，其中包括两部分内容：邮件的特征提取和特征选择。在特征提取部分，研究了基于词的特征提取方法和基于字节级 n-grams 的特征提取方法。在特征选择部分，研究特征选择方法，提出基于信息增益的特征选择方法和基于朴素贝叶斯统计的特征选择方法来解决在线支持向量机模型消耗时间过大的问题。同时，本文从过滤器核心评价指标 1-ROCA 的角度优化过滤器模型，提出了一种基于在线排序逻辑回归学习算法的垃圾邮件过滤器。

最后，研究了含有噪声数据集对过滤器性能的影响。在实际系统中，用户给过滤器的反馈邮件不一定是完全正确的，必然存在噪声邮件。本文了创建了噪声邮件数据，分析了含有不同噪声数量的数据对过滤器性能的影响。

关键词 垃圾邮件过滤；在线学习；特征选择；排序学习；噪声用户反馈

Research on Online Learning Based Spam Filtering

Abstract

Email provides lots of convenience to people's life and work. But a mass of spam greatly affects the use of email. Spam occupies too much network resource and harms the interests of users. Some people who have ulterior motives use it to spread false news. So spam filtering is hot problem in current research.

This paper studies spam filtering based on machine learning methods. These types of methods which have the features of high accuracy and low cost, have become the mainstream methods to tackle spam filtering. This paper is mainly divided into four parts.

Firstly, we study the framework and filtering model of spam filtering based on online learning. We realize three spam filters which respectively utilize Naive Bayes, Support Vector Machines and Logistic Regression. Their advantages and disadvantages are compared in aspect of CPU time and accuracy.

Secondly, we study the feature engineering in spam filtering, including feature extraction and feature selection. In feature extraction, we introduce words-based method and n-grams method with bytes level. In feature selection, Information Gain and Bayesian statistics methods are proposed to reduce computational cost and improve the filter performance a little. At the same time, we suggest that spam filtering can be treated as an online ranking task. Online ranking logistic regression method is presented to settle spam filtering.

Finally, we show that noisy data sets harm or even break state-of-the-art spam filters. The spam filter based on machine learning methods attains near-perfect performance when filters are given accurate labeling feedback for training. However, users perhaps give incorrect feedbacks in real-world settings. The noisy data sets are created and used to analyze the changes of the filtering performance with the number of noisy emails.

Keywords spam filtering, online learning, feature selection, ranking learning, noisy user feedback

目 录

第1章 绪论

1.1 课题研究的目的和意义

大规模的垃圾邮件严重影响了电子邮件正常使用。垃圾邮件消耗大量网络资源，损害用户利益，还会被一些别有用心的人用来散播虚假消息，危害社会安定。2011 年三季度中国网民平均每周收到垃圾邮件数量为 14.9 封，环比（与上季度调查相比）增长了 2.1 封，同比上涨 1.0 封。中国网民平均每周收到垃圾邮件比例为 33.7%，同比下降 3.9 个百分点，环比增长 0.5 个百分点。全球的平均水平更是高于上述数据。根据英国著名安全公司 Sophos 最新发布的数据显示，2011 年第三季度全球垃圾邮件数量中源于美国的垃圾邮件数据仍高居首位，占全球垃圾邮件产量份额的 11.3%，其次是韩国(9.6%)、印度(8.8%)、俄罗斯(7.9%)、巴西(5.7%)和中国台湾(3.8%)[1]。垃圾邮件的特征越来越隐蔽，技术也更复杂。传统的过滤技术已经越来越没有效果。基于机器学习方法的垃圾邮件过滤技术是目前最常用的过滤技术，能够在一定程度上解决基于内容的垃圾邮件过滤问题。

1.2 垃圾邮件过滤技术研究现状

近年来，垃圾邮件给电子邮件行业带来了很多问题，给人们生活造成了影响，个人和公司由于接收垃圾邮件和区分垃圾邮件而占用大量网络资源和时间。同时垃圾邮件也是一个有利可图的商业模式，因为垃圾邮件发送者只需要付出很小的代价就能得到丰厚的回报[2]。由于垃圾邮件导致了经济上的损失，有些国家制定相关法律来制裁垃圾邮件发送者。然而，由于技术上的限制，无法跟踪某些垃圾邮件的来源（如国家或某个地点），从而无法对垃圾邮件发送者进行法律制裁[3]。

很多研究人员提出了多种解决垃圾邮件的方法[4]。从 2006 年起，通过垃圾邮件过滤器的帮助，用户收到的垃圾邮件数量逐渐减少[5]。一种有效阻止垃圾邮件发送的方法是使用垃圾邮件过滤器。早期的过滤技术是根据邮件发件人的地址，IP 信息、邮件的主题等内容过滤邮件，我们将此方法称为基于黑白名单的过滤技术[6]。该方法实现简单，运行速度快，但准确率较

低。随着垃圾邮件发送技术的发展，垃圾邮件的特性大部分体现在邮件的内容上，所以垃圾邮件发送者发送的邮件很容易躲避基于黑白名单过滤技术的检测。为了解决此类问题，研究人员通过分析邮件的内容和附件来判别垃圾邮件。例如，建立一个垃圾邮件词库，将邮件内容与库中词进行匹配，若匹配成功，判断为垃圾邮件。我们将这种技术称为基于规则或匹配的垃圾邮件过滤技术。该技术针对性很强，词库也便于更改。但随着电子邮件的发展，垃圾邮件的数量非常庞大，规则库或词库也将变得非常庞大，从而导致系统匹配速度变慢[7]。

为了更好地解决基于内容的邮件过滤问题，基于机器学习理论的垃圾邮件过滤技术成为当前最常用的方法。该方法针对邮件内容进行过滤，从邮件的内容获取特征，并通过这些特征以及对应的标注来训练和优化过滤器。该方法准确率高，成本较低，已成为当前解决垃圾邮件过滤问题普遍采用的方法。

为了使研究者们更好地交流，并验证目前研究的方法在实际系统中的性能，国内外举办了多个评测，具有影响力的包括：TREC[8](Text Retrieval Conference)、CEAS[9](Conference on Email and Anti-Spam)以及国内的SEWM(Search Engine and Web Mining)。TREC 会议分别于 2005、2006、2007 年举办了垃圾邮件评测活动，在 2006 年由清华大学提供了中文评测数据[10]。CEAS 会议于 2007 和 2008 年举行了垃圾邮件评测。国内的 SEWM 会议的垃圾邮件评测是由华南理工大学的董守斌教授组织，于 2007、2008、2010 以及 2011 年分别举行了评测。这些评测促进了垃圾邮件过滤技术的发展，积累了完整的评价指标和大量的测试数据集。

垃圾邮件过滤的目的是将垃圾邮件和正常邮件区分开来，所以邮件过滤问题可以当作二值分类问题来处理。基于机器学习的垃圾邮件过滤技术是通过机器学习方法来区分垃圾邮件和正常邮件。过滤器中主要应用的机器学习技术可以分成两类：生成模型，如朴素贝叶斯(Naïve Bayes，NB)和判别模型，如支持向量机(Support Vector Machines，SVM)和逻辑回归(Logistic Regression，LR)。在垃圾邮件过滤任务中，判别模型在性能上优于生成模型[11]。

国外的垃圾邮件过滤技术研究取得了很多的成果。Sahami 于 1998 年首次提出了将朴素贝叶斯方法用于垃圾邮件过滤问题[12]。Graham 于 2002 年对贝叶斯模型做了一定改进，尤其是对在基于词的特征选择做了很深的研究，并完成多个基于贝叶斯方法的系统[13]。该系统实现简单，计算效率高，并

在不同数据集上都能取得较好的分类效果。在模型训练部分，Sahami 提出的模型适合离线特征选取，而 Graham 提出的模型适合在线特征选取。Metsis 在 2006 年提出了多种贝叶斯模型[14]。Segal 提出了基于不确定样本近似方法的贝叶斯过滤器[15]。Kim 提出通过分析邮件中的 URL 链接来判别垃圾邮件，该方法是基于朴素贝叶斯模型[16]。Ciltik 使用基于 n-grams 词模型的朴素贝叶斯分类器，该分类器中只取邮件的前一部分特征，大大提升了过滤器运行时间[17]。

Drucker 在 1999 年首次使用支持向量机模型来处理二值的邮件分类问题，并采用 tf-idf 特征模型在两个私有数据集上进行测试[18]。Haider 提出了基于增量式 SVM 模型的垃圾邮件过滤器，该模型在垃圾邮件过滤性能有了很大的提升，与传统的 SVM 相比，该模型性能提升到 40%[19]。Kanaris 在 SVM 模型中使用 n-grams 特征提取方法，在本文第三章我们将详细介绍该方法[20]。D.Sculley 首次提出宽松在线支持向量机(ROSVM)模型，该模型很大程度上提升了过滤器运行速度，并采用 n-grams 特征提取方法[21]。ROSVM 模型在 2007 年的 TREC 评测中取得非常好的成绩。

对于其它模型，Goodman 提出了在线逻辑回归模型来处理垃圾邮件过滤问题[11]。该模型实现简单，准确率高，运行速度快。Gordon 详细解释了在线学习模式和离线学习模式，并对两种模式进行性能上的比较[22]。Wu 采用了基于人工神经网络(Artificial Neural Networks，ANN)方法的垃圾邮件过滤技术[23]。Oda 提出使用人工免疫系统(Artificial Immune Systems，AIS)来处理垃圾邮件过滤问题等[24]。

国内对基于机器学习方法的垃圾邮件过滤方面的研究也取得很多成果。在 2006 年的 TREC 评测中，中文数据集 TREC06c 是由清华大学提供。而国内的 SEWM 的评测会议是由华南理工大学董守斌老师主持，通过近几年的发展，该会议完善了评测流程、过滤器评价指标等内容，并提供了大量的测试数据集，在 2011 年评测大会上，部分学者提出了 F_1 评价指标。浙江大学、大连理工、山东大学、黑龙江工程学院以及哈尔滨理工大学等高校积极参加了近几年的评测。其中大连理工大学大学采用了基于支持向量相机模型和朴素贝叶斯模型过滤器，山东大学研究的是基于规则匹配和机器学习方法混合的过滤器。浙江大学对朴素贝叶斯模型做了改进。黑龙江工程学院和哈尔滨理工大学主要研究在线逻辑回归模型和在线支持向量机模型，这两个模型在评测中取得了非常好的成绩。在学术研究方面，中科院计算所的王斌老师在 2005 年对国内外垃圾邮件现状、评价指标以及数据集等方面做了综述

[25]。黑龙江工程学院的齐浩亮老师研究了基于字节级的 n-grams 特征提取方法的在线逻辑回归模型。浙江大学徐从富老师基于朴素贝叶斯模型提出了 NSNB 模型过滤器[26]，并申请了基于逻辑回归的中文垃圾邮件过滤方法的专利等[27]。

1.3 本课题研究的主要内容

本文主要研究基于机器学习理论的垃圾邮件过滤技术，具体研究内容如下：

1．研究机器学习方法并实现基于机器学习方法的垃圾邮件过滤器

研究了国内外基于机器学习理论的垃圾邮件过滤技术，提出了基于在线学习的垃圾邮件过滤技术的框架和过滤模式，实现了基于朴素贝叶斯、基于支持向量机和基于逻辑回归的垃圾邮件过滤器，并通过实验来比较和评价三种垃圾邮件过滤器的优缺点。

2．研究邮件过滤的特征工程

研究了邮件过滤的特征工程，包括邮件的特征提取和特征选择方法。在基于机器学习方法的垃圾邮件过滤系统中，邮件的特征空间向量对机器学习模型的分类性能起决定性作用。在特征提取部分，我们研究了基于词的特征提取方法和基于 n-grams 的特征提取方法，重点比较这两种方法之间的优缺点；在特征选择部分，我们分别提出了基于贝叶斯统计和基于信息增益的特征选择方法，用以减低支持向量机模型消耗的时间，同时提升过滤器的性能。

3．提出了基于在线排序逻辑回归学习算法的垃圾邮件过滤模型

传统过滤器模型优化目标是使分类错误个数达到最低。由于过滤器的核心评价指标 1-ROCA 与过滤器的错误分类个数并没有直接的联系，本文从评价指标(1-ROCA)%的角度优化垃圾邮件过滤模型，提出了基于在线排序逻辑回归学习算法的垃圾邮件过滤器。

4．研究噪声数据集对过滤模型的影响

在理想状态下，邮件过滤模型中训练数据集的标注是完全正确的。然而，在实际系统中用户的反馈必然存在噪声。因此，过滤器的训练数据中必然存在错误标注的邮件。基于上述问题，本文将研究噪声数据集对过滤器模型的影响。

1.4 本文的组织结构

第 1 章阐述了本课题研究的意义、国内外在垃圾邮件过滤方面的研究现状。

第 2 章研究了基于机器学习理论的垃圾邮件过滤技术，提出了在线垃圾邮件过滤器的框架及在线过滤模型，并实现了基于三种机器学习模型的邮件过滤方法，并通过实验比较它们之间的优缺点。

第 3 章研究了邮件过滤的特征工程，其中包括邮件的特征提取和特征选择研究，深入研究了中文邮件分词和英文邮件分词，研究了常用的邮件特征提取方法，并提出了基于信息增益和贝叶斯统计的方法来提升在线支持向量机模型过滤器。

第 4 章分析了过滤器的邮件过滤目标和评价指标之间不一致的原因，研究了从 1-ROCA 角度优化过滤器模型，并提出了基于在线排序逻辑回归算法的垃圾邮件过滤器。

第 5 章研究了垃圾邮件过滤系统中噪声产生的原因及背景，创建了噪声数据集，并详细分析了含有不同噪声数量的噪声数据对过滤器性能的影响。

第2章 基于在线学习的垃圾邮件过滤概述

基于机器学习理论的垃圾邮件过滤技术具有实用性强、实现简单、运行速度快等特点，已成为当前普遍采用的方法。目前很多机器学习方法被用来解决垃圾邮件过滤问题，如朴素贝叶斯、支持向量机、人工神经网络(Artificial Neural Networks, ANN)、逻辑回归、人工免疫系统(Artificial Immune Systems, AIS)和 KNN 邻近算法等。本章介绍了基于机器学习方法的垃圾邮件过滤系统的框架以及在线过滤模型，讨论和实现了基于朴素贝叶斯模型的垃圾邮件过滤器、基于逻辑回归模型的垃圾邮件过滤器以及基于支持向量机模型的垃圾邮件过滤器。本章还详细描述了实验应用的数据集和过滤器评价指标，并分析了上述三种过滤器的性能。

2.1 垃圾邮件过滤器系统框架

基于机器学习方法的垃圾邮件过滤系统框架如图 2-1 所示，系统的运行流程为：过滤器收到邮件流发送过来的邮件，然后对邮件进行分类，即判断邮件属于垃圾邮件还是正常邮件，并放到相应的文件夹下。过滤器在分类过程中需要调用内存中的特征库数据。用户根据文件夹中的邮件判断过滤器分类是否正确。若过滤器分类错误，即垃圾邮件被放到正常邮件文件夹中或正常邮件被放到垃圾邮件文件中，过滤器将根据这些分类错误的邮件更新邮件特征库。

在垃圾邮件过滤系统之中，过滤器使用机器学习方法对邮件进行分类和训练，其运行流程图如图 2-2 所示，过滤器的运行可以分为以下四个模块。

1．邮件的特征提取模块。

该模块将邮件流中发送过来的邮件进行提取特征。提取特征的方法很多，如基于词、基于 n-grams 等特征提取方法。

2．邮件的特征选择模块。

邮件中含有一些对分类器分类没有影响的特征，如 From、Subject、body、the、at、to 等特征，这些特征增加特征向量空间的维度，消耗了机器学习模型训练和分类的时间，所以有必要进行特征选择。

3．建立特征向量模块。

该模块是将选择好的特征转化为机器学习模型能够识别的空间向量。

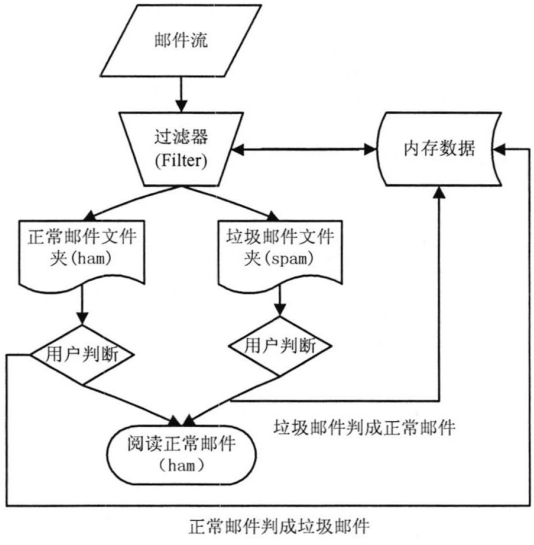

图 2-1 垃圾邮件过滤系统框架

Fig.2-1 Framework of spam filtering

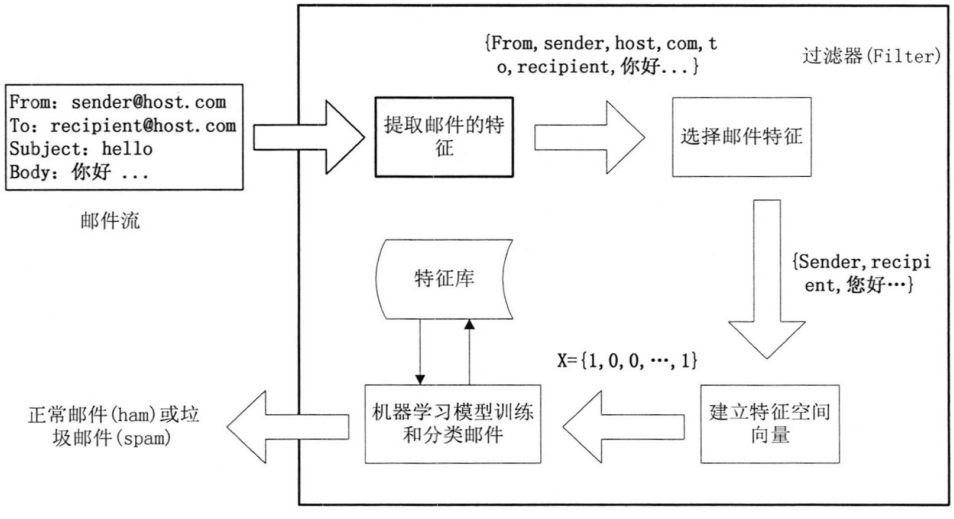

图 2-2 整个过滤系统中过滤器流程图

Fig.2-2 Flow chart of filter system

4．机器学习模块。

该模块对邮件进行分类和训练，即对建立好的特征空间向量进行计算和按照一定的算法更新特征库。

2.2 基于在线学习的垃圾邮件过滤模式

传统的机器学习模型是离线学习(Batch Learning)模式，即模型先训练样本，然后再分类样本，在分类样本的过程中不再进行训练。也就是说，样本的训练和分类是不同步进行的。对垃圾邮件过滤系统来说，离线模式就是过滤器在分类之前先训练一部分邮件，然后就不再训练邮件，即模型的所有参数将不再变化，直接分类。然而随着邮件的变化，垃圾邮件的特征是逐渐变化的，如果不及时更新过滤器，新出现的垃圾邮件将很难被过滤掉。

为了解决此问题，人们提出了在线学习模式。在在线学习模式中，过滤器的训练和分类是同步进行的，即每封邮件通过分类器分类之后立即进行训练，即更新邮件的特征权重值或模型参数等。在线学习模式能识别最新出现的垃圾邮件，能够满足系统实时性需求。在线学习流程图如 2-3 所示。其步骤：第一、邮件流通过过滤之后，过滤器将邮件分类为垃圾邮件和正常邮件；第二、用户根据过滤器分类的结果（包括分类错误和分类正确的结果）反馈给过滤器中训练模块；第三、训练器根据用户反馈的结果进行训练，训练完成之后，将结果返回给过滤器。本文所使用的过滤器均为在线模式。

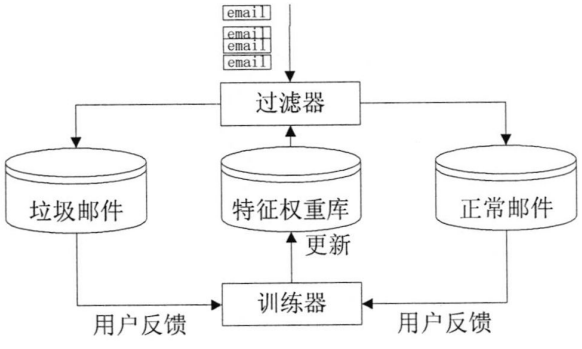

图 2-3 垃圾邮件过滤在线学习模式

Fig.2-3 Online learning model for spam filtering

2.3 机器学习方法

本节将详细介绍三种机器学习方法：朴素贝叶斯、支持向量机和逻辑回归，以及该三种方法在垃圾邮件过滤中应用。

2.3.1 朴素贝叶斯方法

朴素贝叶斯方法是贝叶斯学习方法中实用性很高的一种方法，在某些研究领域内，它的性能可以与神经网络和决策树相当[13]。本小节介绍朴素贝叶斯方法。

朴素贝叶斯的训练任务中，每个实例 x 可由每个属性描述，而目标函数 $f(x)$ 从某有限机会 C 中取值。训练器提供一系列目标函数的训练样本，以及新的样本，然后要求预测新样本的目标值（或分类）。

贝叶斯方法的新样本目标是在给定描述样本的属性值 $<x_1,x_2,...,x_n>$ 下，得到最有可能的目标值 P

$$P=\arg\max_{c_j\in C} P(c_j \mid x_1,x_2...x_n) \tag{2-1}$$

可使用贝叶斯公式将公式(2-1)改写为

$$\begin{aligned} P&=\arg\max_{c_j\in C}\frac{P(x_1,x_2...x_n \mid c_j)P(c_j)}{P(x_1,x_2...x_n)} \\ &=\arg\max_{c_j\in C} P(x_1,x_2...x_n \mid c_j)P(c_j) \end{aligned} \tag{2-2}$$

现计算公式(2-2)中两个数据项的值，估算每个 $P(c_j)$ 很容易，只要计算每个 c_j 目标值出现在训练数据中的次数就可以。但估算不同的 $P(x_1,x_2...x_n \mid c_j)$ 项的值是行不通的，除非训练数据非常庞大。更严重的问题是 $P(x_1,x_2...x_n \mid c_j)$ 项数量等于可能样本数量乘以可能的目标值的数量。因此为了得到合理的估算，样本空间中的每个属性必须多次出现。

朴素贝叶斯分类器基于一个假设，在给定目标值时每个属性之间的条件是相互独立的。换句话说，该假设说明在给定样本的目标值情况下，观察到的 $<x_1,x_2,...,x_n>$ 的概率值刚好是对每个属性的概率值乘积：

$$P(x_1,x_2...x_n \mid c_j)=\prod_{1\le i\le n} P(x_i \mid c_j) \tag{2-3}$$

将公式(2-3)带入(2-2)中。可得到朴素贝叶斯分类器方法：

$$P_{NB}=\arg\max_{c_j\in C} P(c_j)\prod_{1\le i\le n} P(x_i \mid c_j) \tag{2-4}$$

其中，P_{NB} 表示朴素贝叶斯分类器得到的目标值。估算不同的 $P(x_i \mid c_j)$ 项数量只是不同属性值得数量乘以不同目标值数量，这要比估算 $P(x_1,x_2...x_n \mid c_j)$ 项的数量小很多。

总的来说，朴素贝叶斯方法估算不同的 $P(c_j)$ 和 $P(x_i|c_j)$ 项，计算它们在训练数据上的值。这些估算对应待学习的假设，然后用公式(2-4)中的规则来分类新样本，但要满足所需条件相互独立性。

在垃圾邮件过滤中，使用朴素贝叶斯方法来分类垃圾邮件(spam)和正常邮件(ham)。假设一封邮件的特征组成的特征向量 $<x_1,x_2,...,x_n>$，邮件的类型 $c\in\{spam,ham\}$，根据公式(2-4)可得：

$$P_{NB}=\underset{c_j\in\{spam,ham\}}{\arg\max}\ P(c_j)\prod_{1\le i\le n}P(x_i|c_j) \tag{2-5}$$

邮件只有两类垃圾邮件和正常邮件，所以目标值为 $P(spam)\prod_{1\le i\le n}P(x_i|spam)$ 和 $P(ham)\prod_{1\le i\le n}P(x_i|ham)$ 的大小，即根据这两项值的大小判断邮件所属类型[14]。

2.3.2 逻辑回归方法

逻辑回归模型在 TREC 和 SEWM 评测中表现了很好性能。该模型具有学习速率快，准确率高等特点，很方便被运用在实际系统中[9]。

对于一封邮件，我们将提取邮件的特征，并将这些特征对应的权重相加，用数学公式表示为 $\overline{w}\cdot\overline{x}$。其中，$\overline{w}$ 表示为特征向量。$\overline{x}$ 表示只含有 0 和 1 的向量，其中 0 表示该特征在这封邮件中没有出现，1 表示该特征在这封邮件中出现。数学公式 $\overline{w}\cdot\overline{x}$ 可以简单的表示为一封邮件中所有特征对应的权重相加。为了将一封邮件的权重和转化为概率值，将使用以下逻辑方程：

$$P(Y=spam|\overline{x})=\frac{\exp(\overline{w}\cdot\overline{x})}{1+\exp(\overline{w}\cdot\overline{x})} \tag{2-6}$$

公式(2-6)将一封邮件的权重和 $(-\infty,+\infty)$ 转化为 0 到 1 之间的概率值。假如这个概率值超过一个阈值，如 0.5 或 0.9，我们将该邮件判断为垃圾邮件，否则为正常邮件。

下面我们将更新邮件的特征，使用常用的梯度下降算法来更新邮件的特征库。该算法如算法 2-1 所示，其中 TRAIN_RATE 代表算法学习速率，gap 代表 TONE 的设定阈值。

逻辑回归模型取得了较好的过滤性能，但仍存在一些问题。当过多的邮件内容相近时，这些邮件中的特征在特征库中的权重很高，如果再继续多训练此类邮件会产生准确率降低。同时也加重了系统的负担。本文我们采用了 TONE 主动学习方法来解决此类问题。该方法不仅训练被过滤器判断错误的邮件，同时，邮件的分值靠近分类阈值的邮件需要也进行训练。如，邮件的

阈值为 0.5 时，TONE 值为 0.2，则邮件分值在 0.3 到 0.7 之间的邮件需要进行训练。通过实验证明该方法具有很好的效果。分类及权重跟新算法如下：

W=0; //初始化权重向量为 0

for each $\vec{x_i}$， y_i

$$p=\frac{\exp(\vec{x_i}\cdot\vec{w})}{1+\exp(\vec{x_i}\cdot\vec{w})}$$

if (p>0.5)

predict spam;

else

predict ham;

if(abs(p-0.5)<TONE or prediction error) //TONE 主动学习方法

if($y_i=1$)

$w=w+(1-p)\cdot x_i\cdot TRAIN\_RATE$

else

$w=w-p\cdot x_i\cdot TRAIN\_RATE$

2.3.3 支持向量机方法

1．支持向量机

支持向量机是在高维空间中使用一个线性函数的超平面将两类样本分开[28]。在线性情况下，间隔是指两类样本中最靠近分类面的两个不同类型样本之间的距离。给定一个线性、相互独立的样本 $S=\{(x_1,y_1),...,(x_n,y_n)\}$， x_i 表示样本的特征空间向量， y_i 的值 1 和-1，1 表示为垃圾邮件，-1 表示为正常邮件。分类函数如下：

$$f(x)=w\cdot x+b \tag{2-7}$$

其中 w 表示超平面向量， b 是偏移项， x 是邮件的特征向量。当 $f(x)=0$ 时， w 为超平面，距超平面最近两个不同样本符合 $f(x)=\pm1$。因此距超平面最近的两个不同类型的样本的距离为 $1/\|w\|^2$。所以最大间隔的优化问题如下(2-8)形式：

$$minimize_{w,b}:\frac{1}{2}\|w\|^2$$
$$subject\_to:y_i(w\cdot x+b)\geq 1-\xi_i,\forall i,\xi_i\geq 0 \tag{2-8}$$

其中，x_i表示第 i 个训练样本，y_i表示此样本的所属类型。

然而并不是所有的样本都是线性可分的，即不能找到线性超平面，当训练样本不是线性可分的情况，我们引入松弛变量ξ_i。当最大分类间隔变大时，最少错分样本个数会增加，当最小错分个数减少是，最大分类间隔变小。最大分类间隔和最少错分个数之间是矛盾，所以平衡参数 C，调节两者之间的个数。优化形式如下：

$$minimize_{w,b,\xi}:\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i$$
$$subject\_to:y_i(w\cdot x+b)\geq 1-\xi_i,\forall i,\xi_i\geq 0 \tag{2-9}$$

其中，ξ_i是松弛变量，C 是平衡因子。参数 C 的值选择很重要，它决定了过滤器的分类性能和消耗的时间。

2．在线支持向量

通常，SVM 采用批量学习(batch learning)的方式进行训练，批量学习就是事先使用一部分训练数据对 SVM 进行训练，之后再采用另一部分数据作为测试集对训练好的模型进行测试。在测试的过程中模型不再进行学习[9]。

由于垃圾邮件过滤通常采用在线的方式进行，训练数据随着时间源源不断的到来，当模型遇到新的训练邮件后，必须将该训练邮件加入训练数据集，重新对模型进行学习。

对于一封新的邮件，过滤器首先对邮件进行分类，分类后等待用户给予的反馈，即用户会告诉过滤器这封邮件是垃圾邮件还是正常邮件，过滤器获得用户的反馈后，根据反馈结果调整模型的参数。由于不停地有新邮件添加到训练集中，一个简单地加快训练速度的方法是，每次训练时候，都以上次训练得到的参数作为本次训练的初始参数。

本文的在线支持向量机分类器使用 Platt 的 SMO 算法作为求解器[29]，因为 SMO 方法对线性支持向量机来说是最快的方法。

3．提升在线支持向量机

在线的学习算法的训练样本是源源不断的到来的，随着时间的推移，训练样本集合会达到很大的规模，当训练规模很大时，在线支持向量机模型的训练速度就会急剧下降，从而导致模型不可用。因此，应该采取相应的算法

提升模型的训练速度。D.Sculley 提出了三个简化措施[21]。

（1）减少训练集合大小

在线支持向量机训练时，将训练之前出现的所有邮件。邮件数量很大时，训练时间将是非常昂贵的。本文提出只训练最新的 n 封邮件，减少过滤器训练时间，同时，新的邮件训练时并不需要训练之前的数据。每次训练完之后保存当前的权重向量，下次训练时只需要在此向量进行训练。

（2）减少训练的次数

根据 KKT(Karush-Kuhn-Tucker)条件，当 $y_i f(x_i) > 1$ 时 x_i 被认为是一个很容易正确分类的样本。所以当样本 x_i 满足 $y_i f(x_i) \le 1$ 时，该样本需要重新训练。现在我们放宽条件来降低重复训练的更新数量，当样本满足 $y_i f(x_i) \le M$，$(0 \le M \le 1)$ 时，该样本进行重新训练。这样就降低了训练样本的次数。

（3）减少迭代次数

减少学习过程的迭代次数。SVM 模型中最优分类面是通过多次迭代获得，迭代次数的多少直接影响到模型的运行速度，如果次数过大，模型训练将很慢。然而在垃圾邮件过滤中，该模型并不需要过多的训练。所以，通过降低 SVM 模型的迭代次数来提升模型的运行速率，从而提升过滤器整体性能。

经过这三个方面的简化，在线支持向量机模型在一定程度提升了运行时间。该模型的进一步提升我们将在第三章详细介绍。

2.4 实验数据集及评价指标

本文中所用的数据集是有 TREC、CEAS 和国内的 SEWM 评测会议提供的 9 个评测数据集。数据集情况如表 2-1 所示。TREC 分别于 2005、2006 和 2007 年举行垃圾邮件评测，其中 2006 年举行了中文垃圾邮件评测，其数据集由清华大学提供。CEAS 会议分别于 2007 和 2008 年举行了垃圾邮件评测活动。国内的 SEWM 会议的评测是有华南理工大学董守斌老师组织，并于 2007、2008、2010 以及 2011 年分别举行了评测活动。本文实验部分使用的数据全部来这些会议的公开数据集。

本文使用 $(1-ROCA)\%$ 和 $lam\%$ 作为过滤器的评估指标。$hm\%$ 为正常邮件的误判率，$sm\%$ 为垃圾邮件的误判率。由于 $hm\%$ 值很小时，并不能保证 $sm\%$ 的值也很小，所以本文使用 $(1-ROCA)\%$ 和 $lam\%$ 做过滤器评价指标[30]。

$lam\%$ 逻辑平均误判率，其它的定义为如下

$lam\% = logit^{-1}(\frac{logit(hm\%)+logit(sm\%)}{2})$ 其中，$logit(x) = log(\frac{x}{1-x})$

以 $hm\%$ 为横坐标，以 $sm\%$ 为纵坐标，组成 ROC 曲线，$(1-ROCA)\%$ 的值取不同的阈值时的 ROC 曲线上方面积。该值的范围为 0 到 1 之间。$(1-ROCA)\%$ 和 $lam\%$ 的值越小表示过滤器性能越好。

表2-1 测试数据集

Table 2-1 Test Corpus

数据集名称	语言	正常邮件	垃圾邮件	邮件总数
TREC05p-1	英文	39399	52790	92189
TREC06p	英文	12910	24912	37822
TREC07p	英文	25220	50199	75419
CEAS08	英文	167989	41285	209274
TREC06c	中文	21766	42854	64620
SEWM07	中文	15000	45000	60000
SEWM08	中文	20000	50000	70000
SEWM10	中文	15000	60000	75000
SEWM11	中文	15000	45000	60000

2.5 实验结果及讨论

在本小节，我们将讨论三种模型的性能情况。在逻辑回归模型中，学习速率 TRAIN_RATE=0.03，TONE 的值为 0.45。在支持向量机模型中，回看样本的参数 n=10000，即只训练最新出现的 10000 封邮件，KTT 条件中，M 的参数为 0.8，优化迭代次数设为 1。这三种机器学习方法在上节中介绍的九个数据集上进行测试，测试结果如表 2-2 所示。

从表 2-2 我们可以看出，基于机器学习理论的垃圾邮件过滤模型具有极高的性能，其中 SVM 表现的性能最好，其次是逻辑回归模型，而朴素贝叶斯模型性能偏低。其中 SVM 模型和逻辑回归模型的性能基本相近，这两种模型都属于判别模型。而朴素贝叶斯模型属于生成模型。从实验结果我们可以看出，判别模型要优于生成模型。本结论在 TREC 评测中得到了论证。

表 2-2 三种机器学习方法在数据集上的性能

Table 2-2 The performance of three machine learning methods in corpus

数据集名称	Naïve Bayes		Logistic Regression		SVM	
	lam%	1-ROCA%	lam%	1-ROCA%	lam%	1-ROCA%
TREC05p-1	0.54	0.0451	0.42	0.0124	0.33	0.0093
TREC06p	0.51	0.0837	0.55	0.0295	0.49	0.0263
TREC07p	0.42	0.0251	0.14	0.0058	0.10	0.0096
CEAS08	0.11	0.0183	0.11	0.0021	0.08	0.0014
TREC06c	0.15	0.0068	0.07	0.0010	0.05	0.0003
SEWM07	0.30	0.0544	0.00	0.0000	0.00	0.0000
SEWM08	0.77	0.1183	0.01	0.0000	0.01	0.0000
SEWM10	0.32	0.0168	0.00	0.0000	0.02	0.0000
SEWM11	0.15	0.0260	0.01	0.0000	0.00	0.0000

2.6 本章小结

本章我们介绍了垃圾邮件过滤器系统框架，过滤器在线学习模式，以及三种机器学习方法，并介绍了垃圾邮件常用的测试数据集和性能评价指标。通过实验表明，基于机器学习技术的垃圾邮件过滤取得了非常好效果。在机器学习方法中，判别模型的过滤性能要优于生成模型的性能。

第3章 面向邮件过滤的特征工程研究

在基于机器学习方法的垃圾邮件过滤系统中，邮件的特征空间向量对基于机器学习模型的垃圾邮件过滤器分类性能起决定性作用。如果没有好的特征空间向量模型，再好的机器学习方法，也不能取得很好的分类效果。所以在基于机器学习方法的垃圾邮件过滤中，邮件的特征是研究重点之一，尤其在工业运用中。本章将深入研究垃圾邮件的特征提取和特征选择，并通过降低特征空间向量的维度来提升在线支持向量模型过滤器的运行时间和过滤性能。

3.1 邮件过滤的特征工程研究背景

在垃圾邮件过滤器早期，过滤器只要通过邮件的主题、发送人地址以及IP 等特征即可判断，然而随着发送垃圾邮件技术的不断变化，这些特征已远不能判断垃圾邮件。近几年，在基于内容的垃圾邮件过滤技术已成为研究的重点。基于内容的垃圾邮件过滤器技术中，基于词的特征提取方法是目前用得最多的方法之一，即一个英文单词或中文词组作为邮件的特征。然而，垃圾邮件发送者往往对邮件内容中的词进行变形，如“money”变为“money”，或“胡锦涛”变为“胡 j 涛”等形式，基于词的特征提取方法很难过滤这些特征。基于这些问题，研究人员提出了基于字节级 n-grams 的特征提取方法，该方法能够克服上述问题。本章将重点介绍基于词的特征提取方法和基于字节级 n-grams 的特征提取方法，并通过实验证明 N-grams 方法所具有的优势。

邮件的特征选择在特征工程是非常重要的一个步骤，特征空间向量的维度直接影响到机器学习模型的计算时间。如基于在线支持向量机模型在TREC、CEAS 和 SEWM 数据集表现出极高的性能，但是该模型消耗的计算代价是非常大。虽然 Sculley 在 2007 年提出了松弛在线支持向量机(ROSVM)模型[21]，该模型在一定程度提升了过滤器的运行时间，但与其它模型相比，如贝叶斯模型、逻辑回归模型等，过滤器消耗的时间是过大。从表 3-1 得知在相同的硬件环境下，支持向量机模型测试 TREC06p 数据集消耗时间为 18541 秒，但是逻辑回归模型消耗的时间只有 238 秒，两者相差了 78 倍，但是这两种模型的过滤性能相差较小。所以支持向量机模型即使具有较

高的分类效果，也很难应用在实际系统中。

基于上述问题，本章提出了使用特征选择方法降低特征空间向量的维度，从而提升支持向量机模型过滤器。使用特征选择方法原因有两点：一、支持向量机模型的时间复杂度与特征向量维度的平方成正比，特征向量的维度对模型的计算时间有较大影响。二、邮件中有些特征对过滤器判别邮件几乎是没有影响的，如 body、title、sender 等特征，这些特征甚至会影响过滤器的过滤性能，因此，邮件进行分类和训练之前必须要进行特征选择。本章研究了邮件的特征选择方法，并提出了采用信息增益特征选择方法来提升在线支持向量模型过滤器的性能。本章还对邮件的特征进行了分析，提出了一种更适合垃圾邮件过滤的特征选择方法，即基于贝叶斯统计的特征选择方法。通过实验这种特征选择方法要优于其他方法。

表 3-1 ROSVM 与 LR 模型的性能比较

Table 3-1 The performance comparison of ROSVM and LR model

模型	1-ROCA%	CPU 消耗时间
ROSVM	0.0242	18541
Logistic Regression	0.0305	238

3.2 邮件的特征提取

本节将研究和分析基于词的特征提取方法和基于字节级 n-grams 特征提取方法。

3.2.1 基于词的特征提取方法

基于词的特征提取方法是将一封邮件的内容以词的形式分开，每个词作为一个特征建立特征空间向量。即在建立特征空间之前，需要对邮件进行分词。中英文分词有所不同，不同的系统也有不同的分词形式。对邮件内容为英文的邮件，有多种分词形式，如：一个连续且含非空白字符的字符串为一个特征，即以空格形式分词，也有的是一个连续且只含字母的字符串为一个特征，即以非字母形式分词等。下面举个例子详细说明。

英文句子："My email is last.name@gmail.com."。以空格分词形式，各

个词为："My"，"email"，"is"，"last.name@gmail.com."。以非字母形式分词，分的词为："My"，"email"，"is"，"last"，"name"，"gmail"，"com"。这两种形式的分词并不完全相同。本章实验的英文数据集部分我们使用非字母形式分词。

中文邮件的分词，并不像英文分词那么简单，因为中文每个词与词之间是连续的，并没有用空格分开，需要根据一定的规范将中文句子拆分成每个词。所以中文分词要比英文分词复杂的多，困难得多。目前中文分词方法主要分为三类：基于词典的方法、基于统计的方法和基于理解的方法[31]。这三种方法之间各有特点，之间的分词准确率尚无论断哪个更高，各有所长。

1.基于词典的分词方法

有些文章叫机械分词方法[31]。该方法是将需要分词的句子与词典中词条进行逐一匹配。目前常用基于词典的分词方法为：正向最大匹配算法，逆向最大匹配算法、最少切分算法和双向最大匹配算法。

基于词典的分词方法的优点是容易实现。其缺点是对词典的内容受到限制，对未登录词添加比较困难。

2.基于统计的分词方法

在中文句子的上下文中，如果相邻两个字出现的频率越高，说明这个两个字组成词的概率就越大。因此基于统计的分词算法是通过相邻两个字共同出现的频率或概率来反映这两个字成为词的可信度。目前计算两个字之间可可信度有多种方法，如：互信息的概率统计算法，组合度的决策方法等。

该方法优点是不需要建立词典，而通过训练语料的迭代。其缺点是要想取得较高的精确度，必须有大量的语料支撑。

3.基于理解的分词方法

该方法对待分词的句子进行分析，根据对句子的理解进行分词。

该方法优点是对句子可以自行理解，同时可以补充未登录词，但该方法需要大量语言背景做支撑。而中文的语言结构比较复杂性，将从语言中提取的特征信息很难转为转化为机器直接识别特征，因此该方法目前还不够完善和成熟，常与另两种方法混合使用。

这三类分词方法之间哪种方法准确率更高，目前尚无定论。对目前常用的分词系统来说，大部分都是混合使用这三类分词方法。如，海量科技的分词算法就是采用基于词典和基于统计的分词方法。目前比较成熟额分词系统有：SCWS 分词系统、FudanNLP 分词系统、ICTCLAS 分词系统、CC-CEDICT 分词系统、IKAnalyzer 分词系统、庖丁解牛(Paoding)分词系统和

JE 分词系统等。在本章实验的中文数据集部分使用的 JE 分词系统。

3.2.2 基于字节级 N-grams 的特征提取方法

随着反垃圾邮件技术的发展，发送垃圾邮件技术也在提高，垃圾邮件发送者通过故意拼写错误、字符替换和插入空白等形式对垃圾邮件特征的单词进行变体，从而逃避检测系统的检测。如图 3-1 所示，“Viagra”这个单词在 TREC05p 数据集上出现次数最多的 25 中变体[4]。正常情况下，用户都能知道这些变体和“Viagra”单词是同一个含义。然而在基于词的特征提取方法下，每一个变体的出现都代表一个新的特征，不能识别出是“Viagra”这个单词，从而导致过滤器失效。基于字节的 N-grams 方法能够克服这些问题。

Viagra	VIAGRA	Viiagrra	viagra	visagra
Vi@gra	Viaagrra	Viaggra	Viagraa	Viiaagra
Via-ggra	Viia-gra	V1AAGRRA	Viiagra	Via-gra
Vi graa	V iagra	via gra	Viagrra	V&Igra
VIAgra	V|agra	Viaaggra	vaigra	V'iagra

图 3-1 “Viagra”这个单词在 TREC05p-1 数据集上出现次数最多的 25 中变体

Fig.3-1 The 25 most common variants of the word ‘Viagra’ in the TREC05p-1 corpus

基于 n-grams 的特征提取方法是将邮件按照字节流进行大小为 n 字节进行切分（其中，n 取值为 1，2，3，4...），得到长度为 n 个字节的若干个串，每个串称为 gram。如：information，按照 n=4 时进行滑动窗口切分为：info、nfor、form、orma、rmat、mati、atio 和 tion 这 8 个 4-grams 的特征。

基于字节级 n-grams 特征提取方法使用非常方面，不需要任何词典的支持，不要需要对句子进行分词；在使用之前也不需要语料库进行训练。在对邮件提取特征时，不要对邮件进行预处理，也不要考虑邮件编码问题，直接将邮件作为无差别的字节流。同时该方法能够处理复杂的文档，如：HTML 格式的邮件、邮件中很有的图像文件以及附件等内容。该方法与基于词的特征提取方法相比，能够有效防止信息伪装等问题。如图 3-1 所示的文字变形。基于词的特征提取方法可能就识别不了这些特征，而 n-grams 方法能有效的识别出该特征。例如，Viagra 使用 4-grams 方法提出的特征为：Viag、

iagr、agra；当 Viagra 变形后变成 Viiagra 时提取的特征为：Viia、iiag、iagr、agra；两者共同特征是 iagr 和 agra。过滤器能够识别这两个特征。

中文的一个字用 2 个字节或四个字节，每个字之间都是连续的。对于中文的 n-grams 特征提取方法如图 3-2 所示。如图 3-2，“10 月 29 日”在计算机中存储的形式转化为十六进制为：31 30 D4 C2 32 39 C8 D5。使用基于字节级的 4-grams 特征提取方法提取的特征如图 3-2 所示。

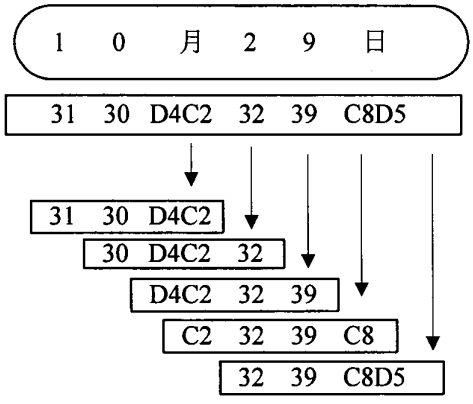

图 3-2 “10 月 29 日”使用 n-grams 方法提取特征

Fig.3-2 The features of “10 月 29 日” is extracted with n-grams method

3.3 邮件的特征选择

特征选择是一项对高维数据进行预处理技术。特征选择的目的是从大量的特征中选择一些有效的，能够描述数据特点的特征，其宗旨是删除无关和冗余的特征，减低模型空间向量维度，从而减少训练时间，提高算法的分类准确度，并提高模型对样本的理解性。

目前已有很多特征选择方法运用在垃圾邮件过滤技术中[32]，包括文档频次 DF(Document Frequency)[33]、互信息 MI(Mutual Information)[34]、信息增益(Information Gain)[32]、统计量[35]、交叉熵[36]和优势率 OR(Odd Ratio)[36]等，它们的特征分值计算形式如表 3-2 所示。其中，M 表示训练邮件集合，c 表示邮件的标注，即垃圾邮件和正常邮件，t 表示一封邮件的一个特征。实验表明信息增益和统计量的方法效果最好，其次是文档频次，互信息的效果相对较差。关于信息增益的特征选择方法将在下面详细介绍。

表 3-2 特征选择方法计算公式

Table 3-2 Term score of feature selection methods

特征选择方法	特征分值计算公式
文档频次(DF)	$\{m_j \mid m_j \in M, f_i \in m_j\}$
互信息(MI)	$MI(t)=\sum_{c\in\{spam,ham\}} p(c)\log\frac{p(t\mid c)}{p(t)}$
信息增益(IG)	$IG(T)=-\sum_{c\in\{spam,ham\}} p(c)\log_2 p(c)+p(t)\sum_{c\in\{spam,ham\}} p(c\mid t)\log_2 p(c\mid t)$ $+p(\bar{t})\sum_{c\in\{spam,ham\}} p(c\mid \bar{t})\log_2 p(c\mid \bar{t})$
统计量(χ^2 Statistic)	$\chi^2(t,c)=\frac{N(AD-BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$ $\chi^2_{avg}=\sum_{c\in\{spam,ham\}} p(c)\chi^2(t,c)\ \ \chi^2_{\max}=\max_{c\in\{spam,ham\}}\{\chi^2(t,c)\}$
交叉熵(CE)	$CE(t)=\sum_{c\in\{spam,ham\}} p(t\mid c)\log\frac{p(c\mid t)}{p(c)}$
优势率(OR)	$OR(t)=\log\frac{p(t\mid c)(1-p(t\mid \bar{c}))}{(1-p(t\mid c))p(t\mid \bar{c})}$

3.3.1 基于信息增益的特征选择方法

在本节我们将介绍信息熵以及信息增益(Information Gain)理论，并介绍信息增益在垃圾邮件中的使用。

1．信息熵

信息熵(又称 Shannon 熵)在随机事件发生之前，它是结果不确定性的量度；在随机事件发生之后，它是人们从该事件中所得到信息的量度。定义一个变量 X，它的可能取值有 n 多种，分别是 $\{x_1,x_2,...,x_n\}$，每一种取到的概率分别是 $\{p_1,p_2,...,p_n\}$，那么 X 的信息熵 $H(X)$ 可表示为公式(3-1)：

$$H(X)=-\sum_{i}^{n} p_i\log_2 p_i \tag{3-1}$$

对邮件来说，邮件有两类：垃圾邮件（spam）和正常邮件（ham），对应的概率为 $p(spam)$ 和 $p(ham)$，则这个过滤器的信息熵 $H(c)$ 为公式(3-2)：

$$\begin{aligned} H(c) &= -\sum_{c\in\{spam,ham\}} p(c)\cdot\log_2 p(c) \\ &= -[p(spam)\cdot\log_2 p(spam)+p(ham)\cdot\log_2 p(ham)] \end{aligned} \tag{3-2}$$

2．信息增益

信息增益的定义：对于某个特征，该特征在系统中的信息量比没在系统中的信息量多出的信息量称为信息增益。特征在系统中时的信息量根据公式(3-1)计算出。而系统没有包含该特征的信息量计算可以转化为把该特征固定不变时的系统信息量。则信息增益的计算形式如公式(3-3)：

$$InforGain(T) = H(c) - H(c|T) \tag{3-3}$$

其中，$H(c|T)$表示类型T的条件熵，T可以表示为t和$\bar{t}$，条件熵$H(c|T)$定义形式如公式(3-4)：

$$H(c|T) = p(t)H(c|t) + p(\bar{t})H(c|\bar{t}) \tag{3-4}$$

其中$p(t)$表示特征t在所有邮件出现的概率，$H(c|t)$表示出现特征的条件下，垃圾邮件和正常邮件的信息熵。$p(\bar{t})$表示特征t在所有邮件不出现的概率，$H(c|\bar{t})$表示没有出现特征的条件下，垃圾邮件和正常邮件的信息熵。因此，根据公式(3-3)可知特征T的信息增益可定义如公式(3-5)形式:

$$\begin{aligned} InforGain(T) = &-\sum_{c\in\{spam,ham\}} p(c)\log_2 p(c) \\ &+p(t)\sum_{c\in\{spam,ham\}} p(c|t)\log_2 p(c|t) \\ &+p(\bar{t})\sum_{c\in\{spam,ham\}} p(c|\bar{t})\log_2 p(c|\bar{t}) \end{aligned} \tag{3-5}$$

我们通过使用特征的信息增益的值来评价每个特征。特征信息增益的值越大，该特征的在整个过滤器系统中越重要。在过滤器系统我们设置阈值θ，当$InforGain(T)\geq\theta$时，则该特征应该被选择，否则，去掉该特征。在过滤器系统，我们根据不同的样本选择不同的θ值。

本文采用信息增益的特征选择方法降低了特征空间向量的维度，降低在线支持向量机模型的训练和分类的时间复杂度。在线支持向量机模型训练的时间复杂度和样本的维度平方成正比。因此信息增益特征选择方法能够从很大长度上解决该模型面临的消耗时间过大问题，从而运用于实际系统中。

3.3.2 基于贝叶斯统计的特征选择方法

在垃圾邮件中，很多特征，如：From、To、subject、Body 等，几乎在每封邮件中都会出现。使用信息增益的方法时，这些特征的信息增益值将非常高，必定会保留，但这些特征不仅对过滤器分类没有影响，同时增加了过滤器的运行时间。所以在本节，基于上述问题，我们提出了一种基于贝叶斯统计的特征选择方法。

1．邮件特征与朴素贝叶斯

一封邮件的特征集合为 $X=\{x_1,x_2,...,x_n\}$，其中每个特征 x_i 是相互独立的。根据第二章介绍的贝叶斯理论可得每个特征是垃圾邮件和正常邮件的概率如公式(3-6)和(3-7)：

$$p(spam\mid x_i)=\frac{p(x_i\mid spam)p(spam)}{p(x_i\mid spam)p(spam)+p(x_i\mid ham)p(ham)} \tag{3-6}$$

$$p(ham\mid x_i)=\frac{p(x_i\mid ham)p(ham)}{p(x_i\mid spam)p(spam)+p(x_i\mid ham)p(ham)} \tag{3-7}$$

在过滤器模型中，$p(spam)$ 和 $p(ham)$ 是根据垃圾邮件和正常邮件出现的频率进行计算的。$p(x_i\mid spam)$ 和 $p(x_i\mid ham)$ 是特征 x_i 分别在垃圾邮件和正常邮件中出现的概率。

2．基于贝叶斯统计的特征选择方法

由于垃圾邮件和正常邮件中含有部分相同的特征，如：From, Subject, To, Body, com 等，这些特征对过滤器分类判断该邮件类型几乎没有影响，但是在邮件中大量出现，所以我们提出一种基于贝叶斯统计的特征选择方法来除去这些特征。根据公式(3-6)和(3-7)得计算公式如下：

$$f(x_i)=\frac{p(spam\mid x_i)}{p(ham\mid x_i)}=\frac{p(x_i\mid spam)}{p(x_i\mid ham)}\frac{p(spam)}{p(ham)} \tag{3-8}$$

其中 $p(spam\mid x_i)$ 和 $p(ham\mid x_i)$ 表示特征 x_i 出现时，该特征属于垃圾邮件和正常邮件的概率。我们用 $f(x_i)$ 来表特征的重要程度，如果值越接近于1，特征对过滤器的作用越低。由于 $p(spam)$ 和 $p(ham)$ 的值是固定，所以公式(3-8)可变成如下形式：

$$f(x_i)=\frac{p(x_i\mid spam)}{p(x_i\mid ham)}\square C \tag{3-9}$$

由于 C 是定值，所以 $f(x_i)$ 可以由公式(3-1)表示特征的重要程度：

$$f(x_i)=\frac{p(x_i\mid spam)}{p(x_i\mid ham)} \tag{3-10}$$

在公式(3-10)中，$p(x_i\mid spam)$ 为特征 x_i 在 spam 中出现的概率，$p(x_i\mid ham)$ 为特征 x_i 在 ham 中出现的概率，其中 $p(x_i\mid spam)$ 计算形式如公式(3-11)：

$$p(x_i\mid spam)=\frac{N(x_i-in-spam)}{N(spam)} \tag{3-11}$$

$p(x_i\mid ham)$ 求法与 $p(x_i\mid spam)$ 类似，其中 $N(x_i-in-spam)$ 表示特征 x_i 在垃圾邮件中出现的次数，$N(spam)$ 表示垃圾邮件的总数。

当特征 x_i 满足 $f(x_i)\in[1/\alpha,\alpha],(\alpha\geq 1)$ 时，即 $f(x_i)$ 的值越接近于 1 时，表示该特征在垃圾邮件和正常邮件出现的概率几乎相当，对邮件过滤几乎没有影响，应该被去掉。当特征满足 $f(x_i)\in(0,1/\alpha)\cup(\alpha,+\infty)$ 时，表明该特征在两类中差别比较大，能够明显区别邮件类型，应该保留。

本文采用基于贝叶斯统计的特征选择方法来降低特征空间向量的维度，从而降低在线支持向量机模型的训练和分类的时间复杂度。

3.4 实验及讨论

本节的实验主要分为两个部分：1. 邮件的特征提取方法试验。在邮件特征提取实验中所使用的数据集是 2.4 节所介绍的四个英文数据集和五个中文数据集，分别为 TREC05p-1、TREC06p、TREC07p 和 CEAS08 英文数据集和 TREC06c、SEWM07、SEWM08、SEWM10 和 SEWM11 中文数据集。2. 采用邮件的特征选择方法减低在线支持向量模型的运行时间，同时还提升了过滤器性能。本部分实验我们选择了具有代表性的 TREC05p-1、TREC06p 和 TREC07p 这三个数据集，并没有选择全部测试数据集，因为支持向量模型在 SEWM 数据集上表现出非常好的性能，准确率几乎接近 100%。所以本实验在 SEWM 数据集测试没有意义，所以我们选择了非常具

有代表性的这个三个数据集。

在信息增益特征选择方法的实验部分，我们选用的参数 θ 取值范围为 0.000001-0.007 之间。基于贝叶斯统计的特征选择方法中，我们选择的参数 α 的取值范围 1.25-3 之间。我们实验所使用的评价指标有三个：过滤性能 1-ROCA%，过滤器消耗时间（CPU Times）和参考指标每封邮件的平均特征数（ANF）

3.4.1 邮件特征提取实验

我们将比较了基于词的特征选择方法和基于 n-grams 的特征提取方法的性能。本实验所使用的过滤器为第二章 2.3 节介绍的基于朴素贝叶斯模型过滤器、基于逻辑回归模型过滤器和基于支持向量机模型过滤器。这些模型所使用参数和第二章实验参数一样。实验中所使用的基于词的特征提取方法在英文数据集上以非字母形式分词，在中文数据集上采用 JE 分词器。试验中所使用的基于 n-grams 特征提取方法在中英文数据集上采用相同的方法。

表 3-3 基于词和 N-grams 的方法在三个过滤器上的 1-ROCA 性能比较

Table 3-3 The performance comparison of three filters in words and n-grams based methods

数据集名称	Naïve Bayes		Logistic Regression		SVM	
	words	n-grams	words	n-grams	words	n-grams
TREC05p-1	0.0451	0.0435	0.0288	0.0124	0.0194	0.0093
TREC06p	0.0837	0.0832	0.0635	0.0295	0.0386	0.0263
TREC07p	0.0251	0.0202	0.0472	0.0058	0.0097	0.0096
CEAS08	0.0183	0.0116	0.0067	0.0021	0.0020	0.0014
TREC06c	0.0068	0.0063	0.0022	0.0010	0.0007	0.0003
SEWM07	0.0544	0.0052	0.0000	0.0000	0.0000	0.0000
SEWM08	0.0183	0.0097	0.0009	0.0000	0.0003	0.0000
SEWM10	0.0168	0.0089	0.0009	0.0000	0.0005	0.0000
SEWM11	0.0260	0.0105	0.0022	0.0000	0.0015	0.0000

三个过滤器在基于词和基于字节级的 n-grams 的特征提取方法上的性能

如表 3-3 所示。从上面的实验结果我们可以明显看出，基于字节级的 n-grams 特征提取方法在过滤性能上要明显优于基于词的特征提取方法。

目前的邮件不仅仅只有文本，还包括图片、邮件附件等内容，这些内容对基于词的特征提取方法来说将无法提取其特征。然而垃圾邮件发送者常常对邮件内容的敏感词进行变形、文字内容变成图片或附件等形式发送，很容易逃避基于词的特征方法提取的特征。而基于字节级的 n-grams 特征提取方法很容易识别这些特征。

虽然基于字节级的 n-grams 特征提取方法取得了较高性能，但是该方法也存在特征数目过大问题。如一个含有 3000 字节的邮件内容，如果基于字节级 n-grams 特征提取方法提取的特征数目为 2997 个，接近 3000 个特征，而使用基于词的特征提取方法提取的特征个数约为 500~600 左右。特征数目越多，增加了特征空间向量的维度，从而加重了模型的计算代价。但总体而言，基于字节级的特征提取方法还是具有一定的优势。

3.4.2 基于信息增益的特征选择方法实验

实验结果如图 3-3 到 3-5。通过图 3-3 和图 3-4 可以看出，随着 θ 的逐渐增大，过滤器的特征平均数(ANF)和过滤器消耗的时间(CPUs)逐渐降低，这是因为随着 θ 的逐渐变大，选择的特征数量变少，所以模型中特征空间向量维度变小，从而导致过滤器消耗的时间变小。但通过图 3-5 可以发现 θ 的范围在 0.000001 到 0.00004 之间，过滤器的分类性能(1-ROCA%)在几乎没有变化，在在 0.00004~0.007 之间，分类性能有一点点减低。所以基于信息增益的特征选择方法提升了在线支持向量机模型的运行速度，并且不影响过滤器性能。

在过滤器系统模型中 θ 的值是根据对系统需求来决定的，通过在 CPU 消耗时间和分类性能之间寻找一种平衡，来确定 θ 的值。

3.4.3 基于贝叶斯统计的特征选择方法实验

实验结果如表 3-4 所示。从实验结果可以发现，随着参数 α 值的增大，每封邮件的平均特征数(ANF)和过滤器消耗的时间(CPUs)大幅减少，但过滤器不但没有变差，反而有一点点提升。原因与上节的实验有点类似，有三点：第一、随着参数 α 值的增大，越来越多的特征被去掉，所以邮件的特征数目减少。第二、在线支持向量机模型的时间复杂度与空间向量的维度平

方成正比，所以随着平均特征数目的减少，空间向量维度也减小，从而过滤器消耗的时间大幅下降。第二、有些特征对模型判别垃圾邮件有一定的影响，这些特征被删除以后，过滤器的性能(1-ROCA%)有了一定的提升。

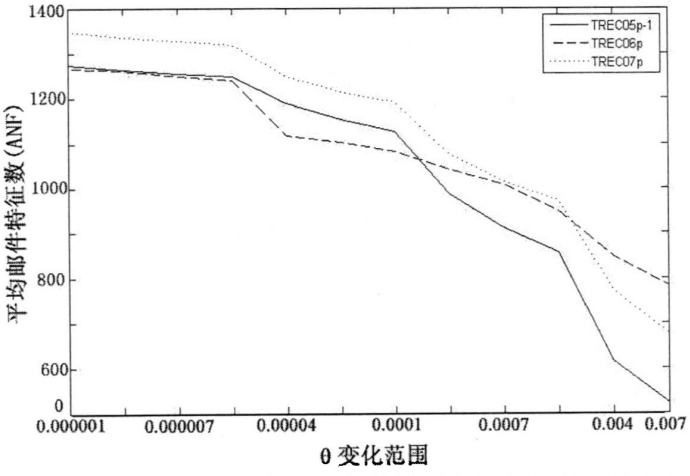

图 3-3 随着参数 θ 的变化，邮件的平均特征数的变化情况

Fig.3-3 Change of the average number of email features with changed values of θ

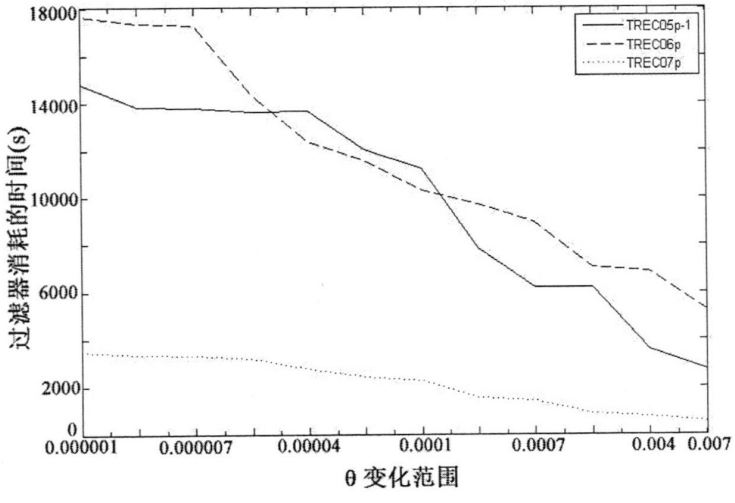

图 3-4 随着参数 θ 的变化，过滤器消耗时间的变化情况

Fig.3-4 Change of the CPU times with changed values of θ

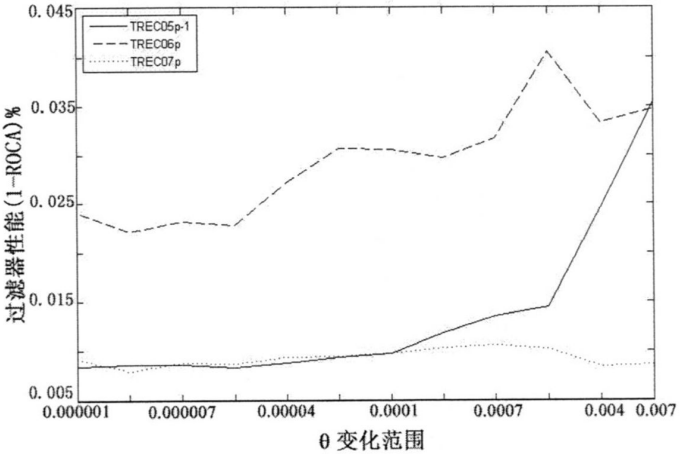

图 3-5 随着参数 θ 的变化，过滤器性能的变化情况

Fig.3-5 Change of the filter performance with changed values of θ

表 3-4 邮件的平均特征数目(ANF)，CPU 的消耗的时间(CPUs)，过滤器性能(1-ROCA)%在三个数据集上的变化情况

Table 3-4 Changes of the ANF, CPU times and filter performance in three datasets

θ	TREC05p-1			TREC06p			TREC07p		
	ANF	CPUs	1-ROCA%	ANF	CPUs	1-ROCA%	ANF	CPUs	1-ROCA%
1.25	1048	14918	0.0087	1056	14226	0.0250	1201	3595	0.0090
1.5	960	13851	0.0084	945	12717	0.0265	1095	3085	0.0096
2	835	12429	0.0096	801	11112	0.0234	976	2367	0.0084
2.5	752	9170	0.0091	714	10273	0.0218	901	2141	0.0083
3	659	7969	0.0099	658	8525	0.0224	839	1932	0.0088

3.4.4 基于信息增益和贝叶斯统计的特征选择方法比较

本文提出的两种特征选择方法在在线支持向量模型的实验结果，以及未使用特征选择方法的实验结果见表 3-5 所示。我们选择合适的参数，在保证过滤器(1-ROCA%)性能没有变差的情况，减少过滤器消耗的时间。

通过表 3-5 可以发现，在 TREC05p-1 和 TREC06p 数据集上，基于贝叶

斯统计的方法有明显的效果。在 TREC05p-1 数据集上，过滤器的性能由 0.0090 变为 0.0091，几乎没有变化，但过滤器消耗的时间有 24720 秒降到 9170 秒，运行时间几乎提升了 2.7 倍。在 TREC06p 数据集上，过滤器性能由 0.0252 提升到 0.0218，同时消耗时间由 17815 秒降到 10273 秒，过滤器过滤性能和运行时间都得到提升。在 TREC07p 数据集上，基于信息增益的特征选择方法效果更明显。过滤器性能由 0.0093 提升到 0.0084，同时过滤器消耗的时间也由 3069 秒降到 759 秒。不仅过滤器性能得到提升，运行速度也提高了 4 倍。

表 3-5 使用两种特征选择方法和未使用特征选择方法之间的实验结果比较

Table 3-5 Results compared between the filter with feature selection methods and without

Corpus	未使用		信息增益		贝叶斯统计	
	CPUs	1-ROCA%	CPUs	1-ROCA%	CPUs	1-ROCA%
TREC05p	24720	0.0090	13633	0.0088	9170	0.0091
TREC06p	17815	0.0252	13660	0.0231	10273	0.0218
TREC07p	3069	0.0093	759	0.0084	2141	0.0083

通过上述数据我们可以发现，基于特征选择方法来提升在线支持向量机模型具有很好的效果。该方法能够克服在线支持向量机模型训练时间过长问题，从而能更好地被运用在实际系统中。

3.5 本章小结

本章我们介绍了邮件特征提取方法，并重点的介绍了基于词的特征提取方法和基于字节级的特征提取方法。这两种方法是目前最常用的两个特征提取方法。实验证明基于字节级的 n-gram 方法在过滤性能上具有更多优势。然而该方法也存在特征数目过大等问题。为了解决此问题，本文对对邮件的特征选择方法进行了深入的研究，并提出基于特征选择方法提升在线支持向量模型，并分别使用基于信息增益和基于贝叶斯统计的特征选择方法对提出的思想进行认证。基于信息增益和基于贝叶斯统计的特征选择方法是从两个不同的角度进行特征选择，基于信息增益的方法是所有邮件中，每个特征所含的信息量的大小，信息量越大，说明该特征被选中的概率越大；反之越

小。该方法是对整个系统进行计算特征信息量，这种方法明显不同于基于贝叶斯统计的特征提取方法。在基于朴素贝叶斯统计的特征选择方法中，如果一个特征在垃圾邮件和正常邮件中出现的概率很大，同时两者概率几乎相等，则该特征将被去掉。这刚好和基于信息增益的方法相反。所以这两种方法是从不同的角度对邮件的特征进行选择。通过实验发现，这两种方法在不同数据集上各有优势。

第4章 基于在线排序逻辑回归学习算法的垃圾邮件过滤技术研究

本章将从过滤器核心评价指标(1-ROCA)%的角度优化过滤器模型，并提出了基于在线排序逻辑回归学习算法的垃圾邮件过滤器。实验表明，该算法比传统的逻辑回归算法相比具有极好的效果。

基于机器学习理论的垃圾邮件过滤技术是对垃圾邮件和正常邮件进行分类，本质上属于二元分类问题。在基于机器学习模型中，分类器优化目标是尽可能的将垃圾邮件和正常邮件分开，如在逻辑回归模型中，训练模型通过梯度下降算法不断优化特征权重向量；在支持向量机中，训练模型通过多次迭代寻找最佳分类面，目的都是为了减低垃圾邮件和正常邮件的误判率。所以基于机器学习理论的垃圾邮件过滤的优化目标是减低过滤系统的错误分类个数。然而过滤器的核心评价指标 1-ROCA 与过滤器的错误分类个数并没有直接的联系。所以降低过滤器分类错误数目并不能保证过滤器的 1-ROCA 值达到最佳。因此，可以从 1-ROCA 评价指标角度优化过滤器性能。

目前 ROC 受到机器学习领域的研究人员关注，ICML(Information Conference on Machine Learning)会议分别与 2004、2005 以及 2006 年专门讨论 ROC 相关的专题研讨会。但在垃圾邮件过滤方面，尚未见到通过 1-ROCA 的角度进行优化。在整个机器学习领域关于这方面的研究也很少。L. Park 等人基于 1-ROCA 的基本理论进行优化[37]。T. Joachims 等人研究了 WRSS 与 1-ROCA 之间的关系[38]。1-ROCA 的计算存在很大困难，L. Yan 等人提出了近似算法来计算 1-ROCA 的值，然而该方法存在较大的误差，所以并不能很好的被采用[39]。T. Joachims 等人从排序的角度改进了支持向量机模型，并同降低错误序对数来优化 1-ROCA 评价指标[38]。然而由于在线支持向量机模型的时间复杂度较大，基于排序优化后的支持向量机模型的时间复杂度更大，因此，该方法很难在实际垃圾邮件过滤系统中使用。

本章针对上述问题，提出了基于对核心评价指标 1-ROCA 直接优化来提升过滤器性能的思想，提出基于排序学习的垃圾邮件过滤建模新方法，以取代传统分类模型的方法，并提出在线顺序逻辑回归算法解决在线排序时邮件得分偏移问题。与支持向量机模型相比，它大大减少模型计算代价。此外，本文通过减少模型问题的大小和减少训练邮件的序对数方法来提升基于

排序模型的运行速度，以达到实际系统的需求。实验结果表明，在中文和英文数据集上，基于排序的逻辑回归模型的性能明显优于经典的逻辑回归模型。

4.1 排序学习

Learning to rank 起源于信息检索，在很多领域都得以应用，例如机器翻译。在信息检索问题中，对于一条查询，检索模型基于一些特性对检索的结果进行排序，这些特性有很多种，如；基于内容、链接以及用户行为等的特性。这种排序方法是通过一些特征构造有效的排序函数（Scoring Function）。目前已有的排序算法有三种：Pointwise、Pairwise 和 Listwise 方法。Pointwise 方法采用一个查询对应的一个文档作为训练样本，其主要思想是将排序问题转化为多分类问题或回归问题，如 Pranking with ranking 算法[40]；Pairwise 方法是采用查询对应的文档对（document pair）作为训练样本，其主要思想是将排序问题形式化为二元分类问题，如 Ranking SVM 算法[41]；Listwise 方法是采用查询对应的文档序列（document lists）作为训练样本，直接对样本的排序结果进行优化，如 ListNet 算法[43]。垃圾邮件其本质是一个二元分类问题，所以我们将排序学习算法的中 Pairwise 方法运用在垃圾邮件过滤问题，将垃圾邮件与正常邮件组成邮件序对数进行训练。

4.2 1-ROCA 与排序学习关系

1-ROCA 为垃圾邮件过滤器的核心评价指标，所以提升过滤器的性能可以以 1-ROCA 的评价指标作为优化目标。本节我们将主要介绍 1-ROCA 以及排序学习。

在目前的垃圾邮件过滤器中，在邮件分类时，过滤器常常会计算这封邮件的分值，通过这个分值来评价这封邮件为垃圾邮件的可能性。ROC 曲线是以 *hm*% 为横坐标，*sm*% 为纵坐标组成的曲线图，即把 *sm*% 计算为 *hm*% 的函数。ROC 曲线下方的面积表示为：在整个垃圾邮件过滤系统中，任意一封垃圾邮件的获得分值要高于任意一封正常邮件的分值的概率。ROC 下方面积越大，表示系统垃圾邮件的分值高于正常邮件的分值越大，过滤器性能越好。由于 *hm*% 和 *sm*% 表示垃圾邮件和正常邮件的误判率，值越小越好，所以垃圾邮件过滤的性能用 ROC 曲线的上方面积来表示，即(1-

ROCA)%。该值越小，过滤器的过滤性能越好。

1-ROCA 值的计算具有较大困难，H. B. Mann 提出了一个 ROC 曲线的阶梯函数以及 1-ROCA 的计算方法[44]。该方法的计算公式如公式(4-1)所示，目前，国内外评测中一直都使用该评价指标。文献[38]也采用该方法。

$$1-ROCA=\frac{SwappedPairs}{mn} \tag{4-1}$$

其中，m 和 n 分别为垃圾邮件和正常邮件的数目。在垃圾邮件过滤系统中，任意一封正常邮件和垃圾邮件组成的序对，一共组成 mn 个序对。在这 mn 个序对中，若正常邮件的得分大于垃圾邮件的得分，该序对被称为不一致序对，即 $SwappedPairs$。

下面我们用一个例子来说明 1-ROCA 的计算方法。

有 5 封邮件，其中垃圾邮件为 2 封，正常邮件为 3 封，即 m=2，n=3，其邮件通过过滤器计算得出的分值和被判成类型的情况如表 4-1 所示。以<ham，spam>形式组成的所有序对如表 4-2 所示。

表 4-1 每封邮件情况（阈值为 0.5）

Table 4-1 Each email state (Threshold is 0.5)

邮件	正确类型	邮件分值	被判成类型
e1	ham	0.2	ham
e2	ham	0.4	ham
e3	ham	0.7	spam
e4	spam	0.6	spam
e5	spam	0.8	spam

根据表 4-2 可以得到不一致序对数 SwappedPairs=1，根据公式(4-1)得：

$$1-ROCA=\frac{SwappedPairs}{mn}=\frac{1}{2\times 3}=0.1667 \tag{4-2}$$

可以算的 1-ROCA 的值为 0.1667。

1-ROCA 的优化目标是尽量使错误序对数最少。因此，基于 1-ROCA 的优化问题不应该从分类的角度优化，更合适从排序的角度优化。在这种情况下，基于 1-ROCA 评价指标优化来提升垃圾邮件过滤性能，可以转化为邮件序对的排序问题来解决，即采用排序学习中 Pairwise 方法解决垃圾邮件过

滤问题。

表 4-2 所有序对数情况

Table 4-2 All pairs state

<h, s>	<s(h), s(p)>	s(h)<s(p)
<e1, e4>	<0.2, 0.6>	True
<e1, e5>	<0.2, 0.8>	True
<e2, e4>	<0.4, 0.6>	True
<e2, e5>	<0.4, 0.8>	True
<e3, e4>	<0.7, 0.6>	False
<e3, e5>	<0.7, 0.8>	True

4.3 基于在线排序的垃圾邮件过滤模型

在本节我们将介绍基于排序策略的垃圾邮件过滤模型，并提出基于在线排序逻辑回归学习算法。逻辑回归模型在第二章已介绍，在此不再重复描述。

4.3.1 基于排序策略的垃圾邮件过滤模型

基于排序策略的垃圾邮件过滤模型的框架：令 x_i 表示正常邮件，x_j 表示垃圾邮件，$\overline{X}_{i,j}=(x_i,x_j)$ 表示两类邮件组成序对，其满足一致的序对，即正常邮件的分值小于垃圾邮件的分值，对应的 $y_{ij}^{'}=1$；$\overline{X}_{j,i}=(x_j,x_i)$ 表示两类邮件不满足不一致的序对，即邮件正常邮件的分值大于垃圾邮件的分值，对应的 $y_{ij}^{'}=-1$。在排序模型中，优化目标是最小化不一致序对 $\overline{X}_{j,i}$：

$$h_w^{'}(\overline{X})=\arg\max\{\sum_i\sum_j y_{ij}^{'}\cdot\Psi(w,x_i,x_j)\} \tag{4-3}$$

公式(4-3)可转化为：

$$h_w^{'}(\overline{X})=\arg\max\{\sum_i\sum_j y_{ij}^{'}\cdot\Psi(w,x_i-x_j)\} \tag{4-4}$$

其中，h 为假设空间 H 中，满足 $h\in H$，根据公式(4-4)得到优化后的 W，对于一封邮件 X，$\Psi^{'}(w,x)$ 为该邮件的分值。

我们令 $\Psi(w,x_i,x_j)=\mathrm{sgn}[\Psi^{'}(w,x_i)-\Psi^{'}(w,x_j)]$，其中 $\mathrm{sgn}(x)$ 为符号函数，

当 $x \geq 0$ 时， $\mathrm{sgn}(x)=1$；否则， $\mathrm{sgn}(x)=-1$。公式(4-4)可以改写成：

$$h_w^{'}(\overline{X})=\arg\max\{\sum_i\sum_j y_{ij}^{'}\cdot\mathrm{sgn}[\Psi^{'}(w,x_i)-\Psi^{'}(w,x_j)]\} \tag{4-5}$$

本文在排序策略的框架基础上深入研究了基于机器学习方法的排序模型。目前，在机器学习领域，关于排序学习算法方面的研究已成为热点问题[43]。在 2005 年的 SIGIR 会议上提出了基于感知器的排序算法，微软公司基于这种排序算法提出了神经网络的排序算法，即 RankNet，该方法已经被成功应用到搜索引擎中。基于支持向量机的排序算法是一种较好的方法，已经被应用在基于排序的信息检索中。但由于该方法的时间复杂度非常高，而无法运用在在线垃圾邮件过滤模型中。我们曾尝试使用 SVM-Light 工具包中的 Ranking-SVM 来训练在线垃圾邮件过滤器。该模型仅训练了 1000 封邮件就耗尽了系统 8GB 的内存。本文基于上述排序学习算法存在的问题和在线垃圾邮件过滤的要求，提出了新的在线排序学习算法。

4.3.2 在线顺序逻辑回归学习算法

在基于机器学习理论的垃圾邮件过滤研究中，逻辑回归模型具有非常高的过滤性能[46]。与性能同样很高的支持向量机模型相比，逻辑回归模型的时间复杂度和空间复杂度明显低于支持向量机模型。因此，我们在逻辑回归模型上进行改进，根据公式(4-5)，$\Psi(w,x_i,x_j)$ 可定义为公式(4-6)所示的形式，使得传统的逻辑回归模型能够解决垃圾邮件过滤的排序问题。

$$\Psi(w,x_i,x_j)=\Psi^{'}(w,x_i-x_j)=\frac{\exp(w\cdot(x_i-x_j))}{1+\exp(w\cdot(x_i-x_j))} \tag{4-6}$$

我们将公式(4-6)的方法称为在线排序逻辑回归学习算法（Pairwise Ranking LR）。其中，(x_i-x_j) 被看成一个样本，参数 W 表的特征的权重向量，更新算法类似于经典的逻辑回归方法，其形式如公式(4-7)：

$$\square w=(y_{ij}^{'}-difference)*TRAIN\_RATE*(x_i-x_j) \tag{4-7}$$

difference 的计算形式如公式(4-6)，*TRAIN_RATE* 训练速率。公式(4-6)是分别将垃圾邮件和正常邮件中的特征组成序对变成新的特征，我们将此方法定义为基于特征的在线排序逻辑回归算法（Features Based Ranking Logistic Regression，F-RLR）。

根据公式(4-6)定义的在线排序逻辑回归算法并不完全适合于垃圾邮件过滤。该算法仅调整邮件序对的特征差值的权重，而不是调整邮件序对对应邮件的分值。在这个方式下，并没有保证垃圾邮件和正常邮件的分值均衡，可能引起两类邮件分值向一侧偏移，从而影响到过滤器的性能。

4.3.3 基于样本的在线排序逻辑回归学习算法

为解决基于特征的在线排序逻辑回归算法存在的问题，提出了基于样本的在线排序逻辑回归算法，即将 $\Psi(w,x_i,x_j)$ 定义为 $\Psi'(w,x_i)-\Psi'(w,x_j)$，即两个类别邮件的得分之差，此时的优化目标函数形式如公式(4-8)：

$$h_w'(\overline{X})=\arg\max\{\sum_i\sum_j \mathrm{sgn}\{y_{ij}'\cdot[\Psi'(w,x_i)-\Psi'(w,x_j)]\}\} \tag{4-8}$$

基于公式(4-8)，结合逻辑回归模型，我们定义 $\Psi(w,x_i,x_j)$ 为：

$$\Psi(w,x_i,x_j)=\frac{\exp(w\cdot x_i)}{1+\exp(w\cdot x_i)}-\frac{\exp(w\cdot x_j)}{1+\exp(w\cdot x_j)} \tag{4-9}$$

这样，就得到了新的排序逻辑回归学习算法，我们将此方法定义为基于样本的在线排序逻辑回归学习算法（Samples Based Ranking Logistic Regression，S-RLR），此算法基于两封邮件的差值为优化目标。

令 $f(w,x)=\frac{\exp(w\cdot x)}{1+\exp(w\cdot x)}$，则有：

$$\begin{aligned}\frac{\partial\Psi}{\partial w}&=\frac{\partial f(w,x_i)}{\partial w}-\frac{\partial f(w,x_j)}{\partial w}\\&=f(w,x_i)\cdot(1-f(w,x_i))\cdot x_i-f(w,x_j)\cdot(1-f(w,x_j))\cdot x_j\end{aligned} \tag{4-10}$$

参数向量权重 W 的更新可以根据公式(4-10)和梯度下降方法解决。

$$\Delta w=(y_{ij}'-difference)*TRAIN\_RATE*\frac{\partial\Psi}{\partial w} \tag{4-11}$$

根据公式(4-10)保证了垃圾邮件和垃圾邮件目标值之间的平衡。从而在理论上保证了两类目标值的对称性，解决了 F-RLR 模型存在的邮件得分偏移问题。

4.3.4 提升在线顺序逻辑回归模型

基于排序策略的逻辑回归模型相比传统模型，如支持向量机模型，具有较低的时间复杂度和空间复杂度，但该模型仍存在一定问题，如，样本的序对数较大，从而导致计算效果变低，因此，该模型很有很大的提升空间。过滤器系统训练全部数据并不是过滤器的性能最佳状态，训练部分数据就能达到性能最佳[47]。本文提出了减少模型问题的大小和减少训练邮件的序对数来降低模型训练的运行效率。下面我们将详细介绍这两种方法。

1．减少模型问题的大小

对于一个包含 n 封邮件的邮件集合，其中 n_1 封正常邮件，n_2 封垃圾邮件，则组成邮件对的个数为 $n_1 n_2$。排序逻辑回归算法要对所有邮件对进行计算，其理论计算复杂度是 $O(n^2)$。垃圾邮件过滤问题中包含了庞大的邮件数以及在线处理的需求，模型不可能支持如此巨大的计算量。考虑到最新出现的垃圾邮件反映出垃圾邮件发送技术的特点，更适用于模型的训练，本文只对最新的 m 封垃圾邮件和 m 封正常邮件作为选择训练样本，即组成邮件对的数目为 m^2，以避免邮件集合中所有邮件参与计算。虽然经过降低采样邮件规模的处理后，模型的时间复杂度仍旧是 $O(m^2)$，但是当 m□ n 时，计算量明显下降。随着 m 值得逐渐增大，至一定值之后，过滤器得性能几乎没有变化。关于 m 的取值我们将在实验中进行讨论。

2．减少训练邮件的序对数

在模型训练中，每出现一封邮件时，对这封邮件组成的所有邮件序对都要进行训练。这种方法在测试时已经取得了很好的效果，但是产生了两个问题。第一，内容相近的邮件对可能被多次训练，增加了资源消耗。第二，邮件对中会出现过度训练，当邮件对中某些特征已经被训练多次，过多的进行训练会导致准备率的下降。在本论文中，我们使用 TONE 方法选择部分邮件对进行训练。TONE 方法是在 TOE 方法的基础上改进的。除了 TOE 中被过滤器错误判别的样本需要参与训练外，过滤器正确判别、但邮件对分值小于设定阈值的样本，即判别比较模糊和容易出错的样本也要参与训练。这就克服了 TOE 方法参与训练的样本数过少带来的问题。例如：当过滤值为 0.5 时，TONE 算法设定阈值为 0.2，则分值为 0.3-0.7 之间的邮件对都被作为训练样本。将 TONE 算法引入排序逻辑回归算法中，仅使用错误或在正确边缘的邮件序对进行训练，使模型的训练速度进一步提高。在 TONE 策略下，对于某一封邮件，首先根据分类器的输出和 TONE 阈值，决定是否对

该邮件进行训练。其次，在对该邮件进行训练时，需要高效的算法找出需要训练的序对。

3．**模型的训练算法**

在降低采样邮件数的同时，我们根据公式(4-10)和公式(4-11)对在线排序逻辑回归算法进行权值更新，设计了新的参数权重更新算法，其中η代表算法学习速率，gap 代表 TONE 的设定阈值。基于 TONE 的参数权重更新算法如下:

Initialize: w=0

Parameters: η, TONE

For each spam-ham pair x_i and x_j

{

Calculate $gap = \Psi(w, x_i, y_j)$

if gap<TONE then

{

$$\Delta w = (1 - gap) * \eta * \frac{\partial \Psi}{\partial w}$$

$w += \Delta w$

}

}

算法具有以下两个优点：一是训练次数的降低带来了模型参数更新速度的提升；二是只选择错误的或者不确定的样本进行训练，有效地增强模型的适应性。

4.4 实验及讨论

在实验部分我们使用第二章 2.3.2 节中介绍的在线逻辑回归模型过滤器作为基准过滤器（Baseline LR），基准过滤器的参数为：TONE=0.45，TRAIN_RATE=0.003，实验结果如表 4-3 所示。

在在线排序逻辑回归模型中，我们采用了利用时间序列对邮件进行采样，仅选择当前邮件及其前 m 邮件形成的邮件序对作为选择训练样本，并非选择所有样本。随着选择样本数 m 值的变化，过滤器的性能变化情况如图 4-1 所示。从图 4-1 可以看出，m 值在 2 到 10 之间逐渐增大时，过滤器性能急剧变好。m 的值大于 10 时，过滤器的性能变化缓慢，m 值达到 100

时，并在此之后，过滤器性能几乎没有变化。然而从图 4-2 我们可以看出，随着 m 的值逐渐增大时，过滤器消耗的时间逐渐增大。所以，随着 m 的增大时，过滤性能趋近于稳定状态，但是过滤器消耗的时间在一直增大，对于过滤器系统来说，并非 m 的值越大越好。在本文的实验中，我们使用的参数 m 为 100。

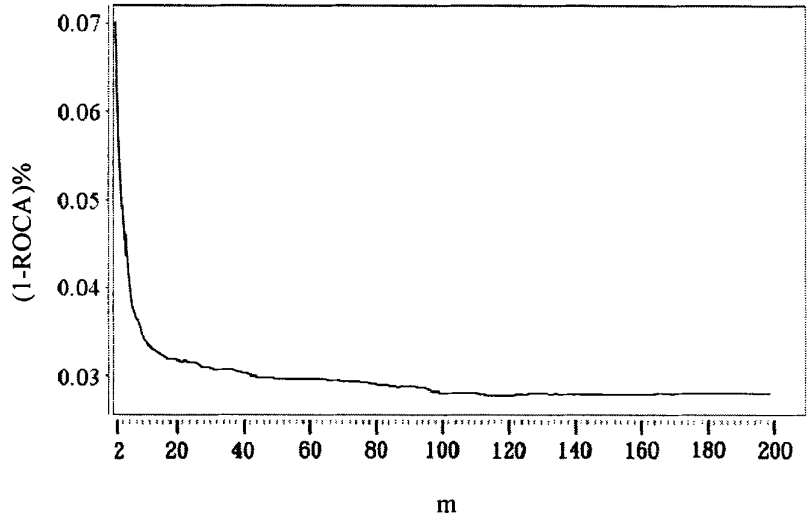

图 4-1 随着 m 值得变化，过滤器性能 1-ROCA%变化情况

Fig.4-1 Change of the filter performance with the changed m

基于特征的在线排序逻辑回归模型（F-RLR）和基于样本的在线排序逻辑回归模型（S-RLR）在 TREC、CEAS 以及 SEWM 数据集上的测试结果如表 4-3 所示，其中，Baseline LR 模型的参数 TONE 和学习速率η，F-RLR 模型和 S-RLR 模型的参数 TONE、学习速率η以及 m 值如表 4-4 所示。

根据 4-3 的结果可以看出，基于特征的在线排序逻辑回归模型（F-RLR）的性能与标准的逻辑回归模型（Baseline LR）相比变得较差，而基于样本的在线排序逻辑回归模型（S-RLR）的性能变得较好。因为 F-RLR 模型在训练时仅调整邮件序对的特征差值的权重，而没有调整邮件序对对应的邮件分值，这样较容易导致邮件得分发生偏移，即偏向其中的一侧，所以 F-RLR 模型性能变得较差。而 S-RLR 模型训练时控制了两封邮件得分均衡机制，并且已邮件序对形式训练也符合 1-ROCA 的优化目标。所以 S-RLR 模型的性能较好，取得了预期的结果。

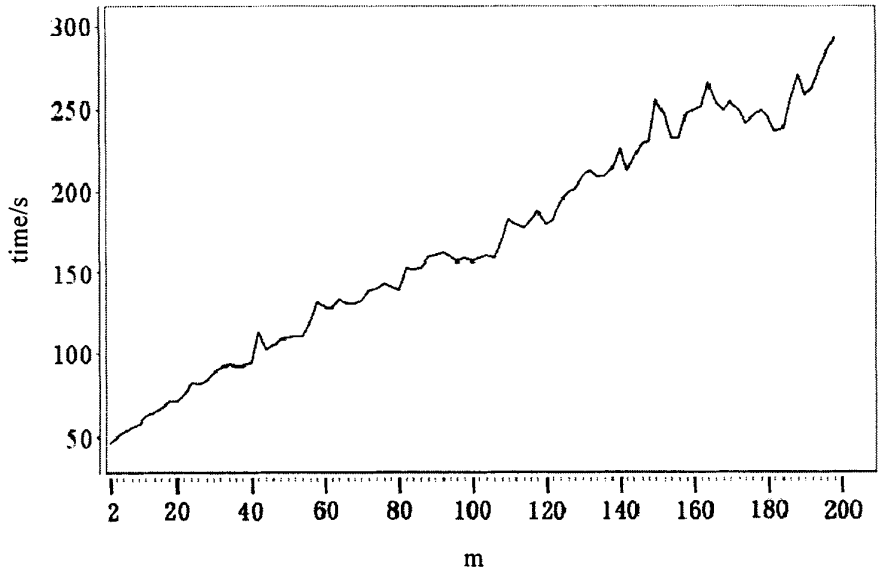

图 4-2 随着 m 的变化，过滤器消耗的时间变化

Fig.4-2 Change of the CPU times with the changed m

表 4-3 F-RLR 模型和 S-RLR 模型的实验结果

Table 4-3 The Results of F-RLR model and S-RLR model

测试集	Baseline LR		F-RLR		S-RLR	
	Lam%	1-ROCA%	Lam%	1-ROCA%	Lam%	1-ROCA%
TREC05p-1	0.42	0.0124	0.45	0.0176	0.38	0.0120
TREC06p	0.55	0.0295	0.57	0.0498	0.51	0.0282
TREC07p	0.14	0.0058	0.29	0.0073	0.12	0.0057
CEAS08	0.11	0.0021	0.07	0.0022	0.08	0.0015
TREC06c	0.07	0.0010	0.09	0.0018	0.05	0.0006
SEWM07	0.00	0.0000	0.00	0.0000	0.00	0.0000
SEWM08	0.01	0.0000	0.06	0.0001	0.02	0.0000
SEWM10	0.00	0.0000	0.00	0.0000	0.00	0.0000
SEWM11	0.01	0.0000	0.00	0.0000	0.01	0.0000

表 4-4 LR、F-RLR 和 S-RLR 模型的参数设置

Table 4-4 The parameters of the F-RLR model and the S-RLR model

模型	TONE	η	m
Baseline LR	0.45	0.003	—
F-RLR	0.99	0.001	100
S-RLR	0.99	0.005	100

4.5 本章小结

垃圾邮件过滤本质上是一个分类问题，优化目标是降低垃圾邮件误判率和正常邮件误判率。然而，通过过滤器的评价指标 1-ROCA 可知，过滤器性能与过滤器判断邮件的错误序对数有关。所以，在本文我们以 1-ROCA 为优化目标来训练过滤器，即提出了基于排序的垃圾邮件过滤模型，并在此基础上提出了基于特征的在线排序逻辑回归模型和基于样本的在线排序逻辑回归模型。由于基于特征的在线排序逻辑回归模型容易导致邮件得分发生偏移，使得过滤器性能变差，而基于样本的在线排序逻辑回归模型能很好的避免此问题。实验结果显示，本文提出了基于样本的在线排序逻辑回归模型可以显著的提高过滤器的性能。

第5章 噪声数据对邮件过滤器的影响研究

从上几章我们可以看出，在标准数据集上，基于机器学习理论的垃圾邮件过滤器具有极高的性能，几乎没有错误，尤其是逻辑回归模型和支持向量机模型。也就是说只要过滤器训练时获取的训练数据的标注完全正确，过滤器的性能就能达到几乎完美的地步。然而，在实际系统中，数据的标注是通过用户反馈得来的，用户反馈的结果有错误，和正确答案不一致的反馈，甚至用户故意返回错误的标注。目前很少有研究者对带有噪声数据集对过滤器性能影响的研究。在本章将研究含有噪声的数据集对过滤器性能的影响。由于研究的局限性，无法获取现实数据，所以本章主要研究数据集中随机噪声对模型性能的影响。

5.1 噪声邮件分析

根据图 2-1 可知，过滤器分类之后，用户通过过滤器分类结果是否正确反馈给过滤中的训练模型。在理想状态下，用户反馈的结果是完全正确的，通过上几章我们可以看出在数据集标注完全正确的情况下，过滤器的性能非常高，(1-ROCA)%的值几乎为 0。在实际系统中，用户的反馈不可能完全正确的，这些错误的反馈导致过滤器性能的降低。

用户错误的反馈产生的原因有很多，通过文献[48]可知，在实际系统中，至少有 3%的用户反馈是错误的，大部分都是把垃圾邮件反馈成正常邮件。错误反馈的原因有很多，有的可能是用户点击错误，即将垃圾邮件点成正常邮件，也有可能是用户不明白系统的反馈机制，甚至有的用户可能是故意将垃圾邮件反馈成正常邮件，从而逃避自己发送的垃圾邮件被检测出来。

有些邮件对不同用户来说可能属于不同类型的邮件。如，我们生活中经常收到一些电子产品广告类邮件，不同的用户可能反馈不同类型的邮件。本文将这类邮件称为临界邮件(有些文章中称灰色邮件，gray email)。这类邮件在邮件系统中占较大比例。通过文献[49]一项据显示：在随机抽取的 418 封邮件中，临界邮件为 163 封，占整个系统邮件的 40%，是占有很大比重。

另外，用户的故意反馈错误是目前大型免费电子邮件系统普遍存在的问题。例如：垃圾邮件发送者可能会申请若干个账号，然后向这些账号发送垃圾邮件，这些账号再对过滤器进行反馈，将该垃圾邮件反馈为正常邮件。过

滤器进行多次训练之后，将此类垃圾邮件判成正常邮件。这种方法是目前垃圾邮件发送者常用的战术。事实上，由于邮件的数量庞大，过滤器训练的数据大部分依靠用户的反馈，所以这类问题也是目前大型垃圾邮件过滤系统面临的问题。

因此，噪声问题是目前实际垃圾邮件过滤系统必须考虑的问题。基于机器技术的垃圾邮件过滤技术必须能够适应各种噪声数据，不能只考虑在理想状态下，数据集无噪声时的过滤器性能。在本章我们将分析基于机器学习的垃圾邮件过滤技术在噪声数据集过滤器的性能变化情况。

5.2 过滤器模型

机器学习模型通常被分类以朴素贝叶斯为代表的生成模型和以逻辑回归模型为代表的判别模型。在本章实验中，我们分别使用朴素贝叶斯模型和逻辑回归模型在噪声数据集进行测试。

朴素贝叶斯模型是目前比较常用的一种机器学习方法，在国际 TREC、CEAS 以及国内 SEWM 评测通常将朴素贝叶斯作为基准(baseline)过滤器，它在实际系统中使用性比较强。朴素贝叶斯模型的理论知识已在第二章做了介绍，在此不再介绍。朴素贝叶斯模型运行速度较快，实现比较简单，但与逻辑回归模型相比，朴素贝叶斯模型的过滤性能较低。

逻辑回归模型是目前比较热门的机器学习模型，在 TREC 评测中取得了较好的成绩，在近几年的 SEWM 评测中也一直保持性能第一的好成绩。逻辑回归模型的理论知识在第二章 2.3.2 节做了详细介绍，在此也不多介绍。逻辑回归模型不仅较高性能，而且运行速度快，实现简单特点。

5.3 噪声数据对过滤器性能影响

本文研究了噪声数据对过滤器性能影响，主要分为两个方面的工作：一、创建了随机噪声数据集；二、研究了数据集中含有随机噪声对过滤器性能影响，并分析了随着噪声数据集中噪声数量，垃圾邮件过滤器的性能变化情况。

1. 随机噪声数据集的设计

邮件产生噪声的原因有很多，如用户点击错误或用户故意反馈错误等，为了模拟真实数据，本文在目前公开的数据集上设计噪声数据集。目前常用

的公开数据集中有 TREC、CEAS 以及 SEWM 提供的九个数据集，分别为：TREC05p-1、TREC06p、TREC07p 和 CEAS08 英文数据集和 TREC06c、SEWM07、SEWM08、SEWM10 和 SEWM11 中文数据集。本文设计的噪声数据集中噪声产生的原因是随机产生，即任意一封邮件为噪声的概率是相等。创建随机噪声数据集的算法如下：

```
if(rand()%100<p*100)
{
  if(classification=spam)
  {
    Classification=ham; }
  else
  {
      Classification=spam;
  }
}
```

其中，p 表示每一封邮件变成噪声的概率，若邮件的总数为 N，则数据集中噪声数量约为 $N\square p$。本文我们不同 p 的值产生不同噪声数量的数据集，p 的取值为 0, 0.005, 0.1, 0.15, 0.20, 0.25, 0.30，其中，当 $p=0$ 时，邮件的数据集为原始数据集，没有任何噪声。

2. 随机噪声数据对过滤器性能的影响

在不同的垃圾邮件过滤系统中，用户反馈错误的比例（用户反馈错误的邮件数目比上用户反馈的总数目）可能是不同的。但是，该比例的大小对过滤器应能影响比较大。在本文分析了不同噪声比例的数据对过滤器性能变化情况。

5.4 实验结果及讨论

实验部分我们分别选择一个生成模型和一个判别模型的垃圾邮件过滤器，即朴素贝叶斯模型和逻辑回归模型。基于朴素贝叶斯模型过滤器使用的是基于词的特征提取方法，其中，在英文数据集上使用的是基于非字母形式进行分词，中文数据集采用的是 JE 分词器进行分词。每封邮件获取邮件的全部内容提取邮件的特征。基于逻辑回归模型中，我们在中英文数据集上都采用基于字节级的 n-grams 特征提取方法提取特征，每封邮件只提取邮件的

前 3000 个特征。逻辑回归模型使用的阈值为 0.5，主动学习方法中 TONE 的值为 0.45，学习速率为 0.003。

表 5-1 随着 p 值的变化，基于朴素贝叶斯模型过滤器在所有数据集上的(1-ROCA)%值

Table 5-1 The (1-ROCA)% of filter based Naïve Bayes model on all datasets with the range of the value of p

	0	0.05	0.1	0.15	0.2	0.25	0.3
TREC05p-1	0.0451	0.0452	3.944	17.915	26.862	26.863	25.522
TREC06p	0.0837	1.9321	1.9321	5.6124	6.1588	19.732	19.013
TREC07p	0.0251	0.055	0.2032	0.4334	3.8508	6.8662	14.580
CEAS08	0.0183	0.1678	0.3057	0.4702	4.0200	11.756	19.796
TREC06c	0.0068	4.7393	5.1851	22.449	42.227	51.038	56.472
SEWM07	0.0554	0.8186	0.5245	1.8691	2.0859	9.3740	12.839
SEWM08	0.1183	0.1770	4.6818	18.991	28.476	28.725	29.135
SEWM10	0.0168	0.0264	6.6742	21.803	30.042	30.645	28.892
SEWM11	0.0260	0.0240	1.3928	11.177	22.309	23.125	21.537

表 5-2 随着 p 值的变化，基于逻辑回归模型过滤器在所有数据集上的(1-ROCA)%值

Table 5-2 The (1-ROCA)% of filter based Logistic Regression model on all datasets with the range of the value of p

	0	0.05	0.1	0.15	0.2	0.25	0.3
TREC05p-1	0.0124	0.0179	3.9235	14.451	23.916	23.918	23.918
TREC06p	0.0295	1.1678	1.1678	3.6533	3.9280	14.187	14.929
TREC07p	0.0058	0.0425	0.4929	2.6871	7.8523	9.4215	16.557
CEAS08	0.0017	0.2235	0.8724	2.7045	10.238	14.900	18.499
TREC06c	0.0010	3.6703	3.7487	13.215	35.902	52.300	64.668
SEWM07	0.0000	0.0587	0.0587	0.3476	0.3556	3.8087	4.3476
SEWM08	0.0000	0.0128	3.2848	11.868	24.552	25.262	25.418
SEWM10	0.0000	0.0031	3.2652	13.215	23.339	23.423	23.428
SEWM11	0.0000	0.0029	3.1250	13.743	23.921	24.005	24.012

基于朴素贝叶斯模型的过滤器在所有噪声数据集上的(1-ROCA)%的值如表 5-1 所示。通过此表我们可以看出，随着数据集中噪声数量的增加，过滤器性能急剧下降。基于逻辑回归模型过滤器在所有噪声数据集上的性能(1-ROCA)%的值如表 5-2 所示。其变化情况与表 5-1 类似，随着数据集中噪声数量的增加，过滤器性能急剧变差。

通过表 5-1 和表 5-2 我们可以看出，当数据集中噪声数量小于 10%时，过滤器性能变化缓慢，噪声数据对过滤器影响较小；但是当数据集中噪声数量大于 10%时，随着数据集中噪声数目变大时，过滤器性能急剧变差。以逻辑回归模型为例，在 TREC07p 数据集上，当噪声数量为 10%时，过滤器的(1-ROCA)%性能由 0.0058 降低到 0.4929，当过噪声数据由 10%变为 30%时，过滤器性能由 0.4929 降低到 16.557。

5.5 本章结论

本章我们研究含有噪声数据的数据集对过滤器的性能的影响。通过数据集中噪声数目的变化来观察过滤性能的变化情况，通过实验表明，根据数据集中噪声数量，可以知道过滤器的性能。本章只对噪声数据集的做了初步研究，下一步的研究主要为两点：第一、在数据集中噪声数目不变的情况下，如何通过改进模型来提高过滤器的性能；第二、本章数据集中的噪声数据是随机产生，我们将创建满足一定条件的邮件为噪声邮件的数据集，并研究该数据集对过滤器模型的影响。

结论

本文研究了国内外基于机器学习的垃圾邮件过滤技术，提出了在线学习的垃圾邮件过滤技术的框架和过滤模式，实现了基于朴素贝叶斯、基于逻辑回归和基于支持向量机等三种模型的垃圾邮件过滤器。本文的主要研究成果包含以几个方面：

1. 针对邮件的特征工程展开研究，重点研究了邮件的特征提取和特征选择方法，详细分析基于词和基于字节级的 n-grams 特征提取方法的优点和缺点，并将这两种方法运用在垃圾邮件过滤系统中。通过对在线支持向量机模型的存在问题进行了讨论，提出了基于信息增益和贝叶斯统计的特征选择方法降低在线支持向量机模型过滤器消耗的时间，取得了很好的效果。该方法通过降低了模型中的空间向量维度来降低计算时间，同时还提升过滤器的过滤性能。

2. 从过滤器核心评价指标(1-ROCA)%的角度优化过滤器模型，分析了评价指标(1-ROCA)与过滤器的错误分类个数并没有直接的联系的原因，提出了基于在线排序逻辑回归学习算法的垃圾邮件过技术。实验结果表明，该算法与传统的逻辑回归算法相比具有很好的效果。

3. 研究了邮件中噪声数据对过滤器性能的影响。实验结果表明数据集中噪声数量对过滤器性能的影响较大，少量的噪声数据对过滤器性能影响较小，当噪声数据变多时，过滤器性能急剧变差。甚至无法使用。

本文的上述研究取得了一些阶段性的成果，为了更好地研究基于机器学习方法的垃圾邮件过滤技术中的各种问题，笔者认为应该从以下几个方面进行进一步的研究：

1. 两类邮件数目不平衡。实际系统中，垃圾邮件和正常邮件在数目上往往存在较大悬殊，这给机器学习模型训练时带来很多问题。可以从这个方面做进一步的研究。

2. 标注数据较少。虽然目前邮件的数目非常庞大，但真正有标注的数据还是很少。可以考虑如何选择少量有效的训练数据，就能达到较好的过滤性能。

3. 用户反馈中的噪声。本文只对噪声数据做了初步的研究，但是噪声仍然是目前实际系统面临的主要问题。可以做进一步的研究。

参考文献

[1] 2011 年第三季度中国反垃圾邮件状况调查报告[OL]. 2011. http://www.anti-spam.cn/pdf/2011_03.pdf.

[2] MARTIN-HERRAN, RUBEL, ZACCOUR. Competing for consumer's attention[J]. Automatica, 2008, 44(2): 361-370.

[3] CARPINTER J, HUNT R. Tightening the net: A review of current and next generation spam filtering tools[J]. Computers and Security, 2006, 25(8): 566–578.

[4] CORMACK G, LYNAM T. TREC 2005 spam track overview[C]. Proceedings of the Fourteenth Text REtrieval Conference Proceedings, 2005.

[5] HOANCA B. How good are our weapons in the spam wars?[J]. IEEE Technology and Society Magazine, 2006, 25(1): 22-30.

[6] HULTEN G, GOODMAN J. Tutorial on junk mail filtering[C]. Proceedings of the Twenty-First International Conference on Machine Learning Tutorial, 2004.

[7] NATHAN T, TIMOTHY S. Deterministic memory-efficient string matching algorithms for intrusion detection[C]. Proceedings of IEEE INFOCOM, 2004: 2628-2639.

[8] CORMACK G. TREC 2006 spam track overview[C]. Proceedings of the Fifteenth Text Retrieval Conference Proceedings, 2006.

[9] CORMACK G, BRATKO A. A batch and on-line spam filter evaluation[C]. Proceedings of the 3rd Conference on Email and Anti-Spam, 2006.

[10] CORMACK G. TREC 2007 spam track overview[C]. Proceedings of the Sixteenth Text Retrieval Conference Proceedings, 2007.

[11] GOODMAN J, YIN W. Online discriminative spam filter training[C]. Proceedings of the Third Conference on Email and Anti-Spam, 2006.

[12] SAHAMI M, DUMAIS S, HECKERMAN D. A bayesian approach to filtering junk e-mail[J]. AAI Press. 1998:Tech. rep. WS-98-05.

[13] GRAHAM P. A plan for spam[OL]. 2002. http://paulgraham.com/spam.html, 2002.

[14] METSIS V, ANDROUTSOPOULOS I, PALIOURAS G. Spam filtering with

naive bayes – which naive bayes?[C] Third Conference on Email and Anti-Spam, 2006.

[15] SEGAL R, MARKOWITZ T, ARNOLD W. Fast uncertainty sampling for labeling large e-mail corpora[C]. In Proc of the third conf on email and anti-spam. 2006.

[16] KIM J., CHUNG K., CHOI K. Spam filtering with dynamically updated URL statistics[C]. IEEE Security and Privacy, 2007, 5(4): 33-39.

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

[18] DRUCKER H, WU D, VAPNIK V. Support vector machines for spam categorization[C]. IEEE Transactions on Neural Networks, 1999, 10(5), 1048-1054.

[19] HAIDER P, BREFELD U, SCHEFFER T. Supervised clustering of streaming data for email batch detection. In Proc of the int conf on mach learn. 2007.

[20] KANARIS I, KANARIS K, HOUVARDAS I, STAMATATOS E. Words versus character N-grams for anti-spam filtering[J]. International Journal of Artificial Intelligence Tools, 2007, 16(6): 1047-1067.

[21] SCULLEY D., WACHMAN G M. Relaxed online SVMs for spam filtering[C]. In Proc of the ann int ACM SIGIR conf on res and devel in inform retrieval. 2007.

[22] CORMACK G, BRATKO A. Batch and online spam filter comparison. Proceedings of CEAS-06, 2006.

[23] WU C. Behavior-based spam detection using a hybrid method of rule based techniques and neural networks[J]. Expert Systems with Applications, 2009, 36(3): 4321-4330.

[24] ODA T, WHITE T. Developing an immunity to spam[J]. Lecture Notes in Computer Science, 2003, 2723: 231-242.

[25] 王斌, 潘文锋. 基于内容的垃圾邮件过滤技术综述[J]. 中文信息学报, 2005, 19(5):1-10.

[26] BAOJUN SU, CONGFU XU. Not so naive online bayesian spam filter[C]. Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence Conference, 2009: 147-152.

[27] 徐从富，王庆幸，彭鹏. 基于 Logistic 回归的中文垃圾邮件过滤方法:发明专利. 申请号: 200810059602.9[P]，申请日: 2008.1.28.

[28] CRISTIANINI N, SHAWE-TAYLOR J. An introduction to support vector machines[J]. Cambridge, UK: Cambridge Univ. Press, 2000.

[29] PLATT J. Sequential minimal optimization: A fast algorithm for training support vector machines[M]. In B. Scholkopf, C. Burges, and A. Smola, editors, Advancesin Kernel Methods - Support Vector Learning. MIT Press, 1998: 158-208.

[30] ANDROU T I, PAL I G M K E. Learning to filter unsolicited commercial e-mail [EB/OL]. 2007-01-16. http://www.aueb.gr/users/ion/docs/TR2004_updated.pdf.

[31] 孙铁利，刘延吉. 中文分词技术的研究现状与困难[J]. 信息技术，2009; 7:187-192.

[32] YANG JAN O. Pedersen. A comparative study on feature selection in text categorization[C]. Proceedings of ICML-97, 1997: 412-420.

[33] STEWART M. Modification of feature selection methods using relative term frequency[C]. Proceedings of ICMLC-2002, 2002: 1432-1436.

[34] PENG H. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy[C]. IEEE Trans. Pattern Anal. Mach. In tell. 2005, 27:1226-1238.

[35] LIU H, SETIONO R. Chi2: Feature selection and discretization of numeric attributes[C]. In Proceedings of the Seventh IEEE International Conference on Tools with Arterial Intelligence, 1995.

[36] MLADENIC D, GROBELNIK M. Feature selection for unbalanced class distribution and naive bayes[C]. In Proceedings of the 16th International Conference on Machine Learning ICML-99. 1999: 258-267.

[37] PARK L, MOON J. A learning method of directly optimizing classifier performance at local operating range[C]. Proceedings of International Conference on Intelligent Computing, 2005.

[38] JOACHIMS T. A support vector method for multivariate performance measures[C]. Proceedings of the 22nd International Conference on Machine Learning, 2005.

[39] YAN L R, DODIER M, MOZER C, WOLNIEWICZ R. Optimizing classifier

performance via an approximation to the wilcoxon-mann-whitney statistic[C]. Proceedings of the 20th Annual International Conference on Machine Learning, 2003.

[40] CRAMMER K, SINGER Y. Pranking with ranking[M]. Advances in neural information processing systems, 14. Cambridge, MA: MIT Press. 2001.

[41] HERBRICH R, GRAEPED T, OBERMAYER K. Large margin boundaries for ordinal regression[J]. Advances in Large Classifiers, 2000:115-132.

[42] JOACHIMS T. Optimizing search engines using click through data[J]. KDD 2002, 2002:133-142.

[43] CAO Z, QIN T, LIN T. Learning to ranking: from pairwise approach to listwise approach[C]. ICML 2007, 2007, 227:129-136.

[44] MANN H B, WHITNEY D R. On a test whether one of two random variables is stochastically larger than the other[C]. Ann. Math. Statist, 1947, 18.

[45] XIA F, LIU T, ZHANG W, LI H. Listwise Approach to learning to rank: theory and algorithm[C]. Proceedings of ICML-2008. 2008: 1192-1199.

[46] SCULLEY D. Online active learning methods for fast label efficient spam filtering[C]. Proceedings of CEAS-07, 2007.

[47] ASSIS F. OSBF-lua - a text classification module for lua the importance of the training method[C]. Proceedings of the Fifteenth Text REtrieval Conference, 2006.

[48] YIN W, GOODMAN J, HULTEN G. Learning at low false positive rates[C]. In Proceedings of the Third Conference on Email and Anti-Spam, 2006.

[49] YIN W, MCCANN R, KOLCZ A. Improving spam filtering by detecting gray mail[C]. In Proceedings of the Fourth Conference on Email and Anti-Spam, 2007.

攻读硕士学位期间发表的学术论文

[1] YUEWU SHEN, GUANGLU SUN, HAOLIANG QI. Using feature selection to speed up online SVM based spam filtering[C]. 2010 International Conference on Asian Language Processing, 2010:142-145.(EI: 20110613644388)

[2] GUANGLU SUN, YUEWU SHEN, HAOLIANG QI. Information gain method to speed up online SVM based spam filtering[J]. Sensor Letters. 2012,3.(SCI)

[3] GUANGLU SUN, YUEWU SHEN, HAOLIANG QI. Speed up information gain based online SVM for spam filtering[C]. 2011 International Conference on Computer Science and Logistics Engineering, 2011:663-666.(EI)

[4] FEI LANG, GUANGLU SUN, YUEWU SHEN. Text categorization in selecting authentic materials on tertiary level[C]. 2011 The 6th International Forum on Strategic Technology, 2011:769-772.(EI: 20114014405653)

[5] 孙广路，沈跃伍. 一种基于信息增益和在线支持向量机的新型分类器:发明专利. 专利申请号:201110458593.2[P]. 申请日期:2011.12.20.

致谢

本论文受到国家自然科学基金(60903083)，国家博士点专项基金(20092303120005)和黑龙江省自然科学基金(F200936)的资助和支持。

两年多的研究生生涯即将结束，回首两年多来，收获真的很多，在此我首先要感谢我的导师孙广路教授。孙老师治学严谨、待人和蔼可亲给我留下深刻的影响。在我攻读研究生期间，给我如何学习和生活点拨迷津。在此我要向孙老师表示我最诚挚的敬意和感谢！

还要感谢黑龙江工程学院齐浩亮教授。在黑龙江工程学院学习期间，给我的生活和学习提供了很大的帮助。齐老师在学术研究、论文写作等方面给予我很多建议，在此表示衷心的感谢。

感谢一直关心和支持我们的同学和朋友们！全丽丽、王莹莹、李震、周三山、乔宗杰、张伟涛、冯均，感谢你们两年多来的帮助。同学之窗，我将终生难忘！感谢我的师弟马英财，感谢他在我论文编写过程中提供的帮助。

特别要感谢我的父母，感谢你们的二十多年来的养育之恩，在我的求学道路上，感谢你们在背后默默的支持。

最后，我要感谢参与我论文评审和答辩的所有老师，谢谢你们的辛苦。祝你们平安幸福！

基于在线学习的垃圾邮件过滤技术研究

作者：沈跃伍
学位授予单位：哈尔滨理工大学