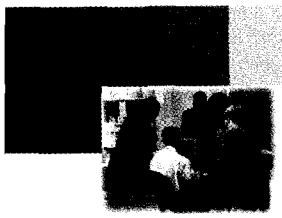

基于主题模型的垃圾邮件过滤系统的设计与实现

寇晓淮，程华
（华东理工大学信息科学与工程学院，上海 200237）

摘　要：垃圾邮件过滤技术在保证信息安全、提高资源利用、分拣信息数据等方面都发挥着重要作用。然而，垃圾邮件的出现影响了用户的体验，并且会造成不必要的经济与时间损失。针对现有的垃圾邮件过滤技术的不足，基于多个主题词理论，构建了基于朴素贝叶斯的垃圾邮件分类方法。在邮件主题获取中，采用主题模型 LDA 得到邮件的相关主题及主题词；并进一步采用 Word2Vec 寻找主题词的同义词和关联词，扩展主题词集合。在邮件分类中，对训练数据集进行统计学习得到词语的先验概率；基于扩展的主题词集合及其概率，通过贝叶斯公式推导得到某个主题和某封邮件的联合概率，以此作为垃圾邮件判定的依据。同时，基于主题模型的垃圾邮件过滤系统具有简洁易应用的特点。通过与其他典型垃圾邮件过滤方法的对比实验，证明基于主题模型的垃圾邮件分类方法及基于 Word2Vec 的改进方法均能有效提高垃圾邮件过滤的准确度。

关键词：文本分类；垃圾邮件；主题模型；贝叶斯原理
中图分类号：TP393　　**文献标识码**：A
doi: 10.11959/j.issn.1000-0801.2017313

Design and implementation of spam filtering system based on topic model

KOU Xiaohuai, CHENG Hua
College of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

Abstract: Spam filtering technology plays a key role in many areas including information security, transmission efficiency, and automatic information classification. However, the emergence of spam affects the user's sense of experience, and can cause unnecessary economic and time loss. The deficiency of spam filtering technology was researched, and a method of spam classification based on naive Bayesian was put forward based on multiple keywords. In the subject of mail, the theme model was used by LDA to get the related subject and keyword of the message, and Word2Vec was further used to search keyword synonyms and related words, extending the keyword collection. In the classification of mails, the transcendental probability of the words in the training dataset was obtained by statistical learning. Based on the extended keyword collection and its probability, the joint probability of a subject and a message was deduced by the Bayesian formula as a basis for the spam judgment. At the same time, the spam filtering system based on topic model was simple and easy to apply. By comparing experiments with other typical spam filtering method, it is proved that the method of spam classification based on theme model and the improved method based on Word2Vec can effectively improve the accuracy of spam filtering.

Key words: text classification, spam, topic model, Bayesian theory

收稿日期：2017-05-12；**修回日期**：2017-09-16

1 引言

伴随着互联网的发展和普及，电子邮件已经成为人们日常工作、生活中通信、交流的重要手段。但由于早期的 SMTP 缺乏发件人认证、大量开放式邮件中转服务器以及互联网分布式管理性质等原因，垃圾邮件已经成为亟待解决的问题。从电子邮件出现以来，研究者就在垃圾邮件拦截方面做出了大量的研究工作。然而，垃圾邮件制造者总会找到更加隐蔽且混淆的手段来躲避相关算法的检测。对于此类研究工作，目前仍然存在两个重要的问题：邮件是一种快速且便捷的通信方式，而大面积的广告推广动机促成了大量为非正当利益而开发的反过滤技术；中文词语的丰富性和特殊性导致垃圾邮件与正常邮件区分难度较大，很多国外的优秀算法在移植过程中将遭遇新的挑战。

针对以上问题，本文深入分析和比较传统垃圾邮件处理方法，指出了现有垃圾邮件过滤方法的不足，对主题模型算法及其在自然语言处理中的应用进行了研究，指出了主题模型算法应用于垃圾邮件过滤的可行性与能够解决的问题；提出了基于主题模型的垃圾邮件过滤算法；设计并实现了一种基于主题模型的垃圾邮件过滤模型，通过与其他方法的对比实验，证明本文基于主题模型的垃圾邮件过滤方法及基于 Word2Vec[1]的改进方法均明显提升了过滤准确度，具有较高的应用价值。

2 基于邮件过滤的相关技术

2.1 面向内容的电子邮件过滤技术

常见的邮箱对于垃圾邮件的过滤策略中，基于内容对邮件过滤的方法有黑白名单、手工建立过滤规则等。手工建立规则的方法通过用户建立一系列规则来判定垃圾邮件。显然，这些方法的主观性会造成大量合法邮件的误判和垃圾邮件的漏判，并且很难做到实时的手工维护，对邮件服务商的人力及经济造成很大压力。因此，垃圾邮件工具逐渐倾向于引入基于内容的机器学习判别方法[2,3]。

基于内容垃圾邮件判别的机器学习方法，一般步骤如下。

步骤 1 获取训练数据集合，通过多种手段渠道获取各类电子邮件，并备注该电子邮件是否是垃圾邮件。

步骤 2 建立模型，使用训练集合训练模型，更新模型中的参数。

步骤 3 使用训练好的模型，对新的电子邮件进行过滤。

总结起来就是通过已有的训练集合（正例、反例）训练出相应的垃圾邮件规则（包括显式规则或隐式规则），然后将规则应用到新的邮件判别中。

最近几年，国内外研究者在此领域已经取得了大量的研究成果。Sheu 等人[4]利用决策树模型构建了三步法垃圾邮件过滤模式。Feng 等人[5]提出了基于朴素贝叶斯分类器的训练集分类方法，提升了数据处理的顽健性，提出的 SVM-NB 方法能够达到较高的垃圾邮件检测精度。而 Bansal 等人[6]构建了基于穿梭判定算法的垃圾词语检测方法，并且在谷歌邮件系统中做了初步的应用。另外，广告产业的发展为垃圾邮件拦截与过滤提出了新的要求，Chan 等人[7]在此方向上做了针对性研究，推出了广告环境下的垃圾邮件过滤方法。除此之外，一些其他的研究成果也引起了学术界和 IT 产业界的广泛关注[8-10]。曹玉东等人[11]基于改进的局部敏感散列算法实现了图像型垃圾邮件过滤，将垃圾邮件过滤方法的应用范围扩大。

2.2 垃圾邮件常用文本分类方法

（1）Decision Tree 方法

决策树利用熵的概念对每次决策产生的结

果进行分类[4]。决策树使用树状结构对目标分类，树中每个节点表示某个对象，每个分叉路径代表某个可能的属性值，而每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。

决策树也可以被称为分类树，它是非常常用的分类方法。从另一个角度来说，决策树是一种监督学习方法。在给定样本机器类别属性后，决策树通过学习能够得到一个固定的分类器，从而给出新进数据的具体类别。

（2）AdaBoost 方法

自适应增强（adaptive boosting，AdaBoost）是加权组合多个弱分类器分类结果，进而得到更好的分类器的方法。Carreras 和 Nicholas[12,13]将 AdaBoost 引入垃圾邮件过滤，获得了很高的性能。AdaBoost 方法的自适应在于：后面的分类器会在那些被之前分类器分错的样本上训练。AdaBoost 方法对于噪声数据和异常数据很敏感。但在一些问题中，相比于大多数学习算法，AdaBoost 方法对于过拟合问题不够敏感。AdaBoost 方法中使用的分类器可能很弱（比如出现很大错误率），但其分类效果只要比随机好一点（比如它的二分类错误率略小于 0.5），就能够改善最终模型。

（3）Rough Sets 方法

Rough Sets 算法是一种比较新颖的算法，粗糙集理论对于数据的挖掘提供了一个新的概念和研究方法。将 Rough Sets 引入垃圾邮件过滤，采用 11 种非文本属性（包括收信人数、中继个数等）来进行邮件分类（正常、广告和反动）。

具体来说，所有属性分为 2 种属性：1 类为条件属性，1 类为决策属性。本文姑且把决策属性设置在数据列的最后一列，算法的步骤依次判断条件属性是否能被约简，如果能被约简，此输出约简属性后的规则，规则的形式大体类似于 IF-THEN 的规则。

（4）*k*NN 方法

k-近邻方法（*k*-nearest neighbour，*k*NN）在线性模型中是最常见的方法，通过选择特征与数据集合中所有特征对比最近的几个样本的标签平均值表示。

2.3 用于垃圾邮件过滤的贝叶斯方法

对于邮件的垃圾分类，一方面邮件就是文本，属于文本分类领域。另一方面，由邮件中的某些关键词来推断是否是垃圾邮件，就是一种贝叶斯条件概率方法的应用。数据挖掘领域主要使用两种贝叶斯方法，即朴素贝叶斯方法和贝叶斯网络方法。贝叶斯方法的一个显著特点，就是在知道结果的情况下了解假设的情况，也就是说，当对某些知识知之甚少，或者毫不知情的时候，贝叶斯方法具有独特优势。

在垃圾邮件检测过程中，其主要依据正常邮件与垃圾邮件的先验概率。而贝叶斯分类模型能够通过适当的独立性假设来简化分布，也就是朴素贝叶斯假设。在这样的假设条件下，能够形成朴素贝叶斯网络。

贝叶斯分类算法是基于概率统计原理的一种分类方法，它具有运算速度快、方法简单、分类精度高等优点，因而被广泛应用在文本分类领域，并表现出非常好的效果。

目前，贝叶斯过滤算法被广泛使用于智能和概率系统中，它具有单词学习的模式和频率，而不需要提前预设任何规则。具体来说，贝叶斯过滤技术能够根据垃圾邮件与正常邮件的联系与特点进行判断。与传统的关键词检测过滤技术相比，贝叶斯过滤算法更加复杂且智能，而反过滤方法不能破解过滤器内部的配置，从而提升了安全性与顽健性。

2.4 基于朴素贝叶斯的文本分类及流程

朴素贝叶斯分类器是垃圾邮件内容过滤中智能应用的分类方法。利用这种方法，可以根据训练集自动训练，训练的结果反映了训练集的性质。

因此训练者可以利用一定数量的垃圾邮件和非垃圾邮件，训练邮件过滤器，从而达到高效、准确过滤垃圾邮件的目的。

朴素贝叶斯分类的流程如图 1 表示。

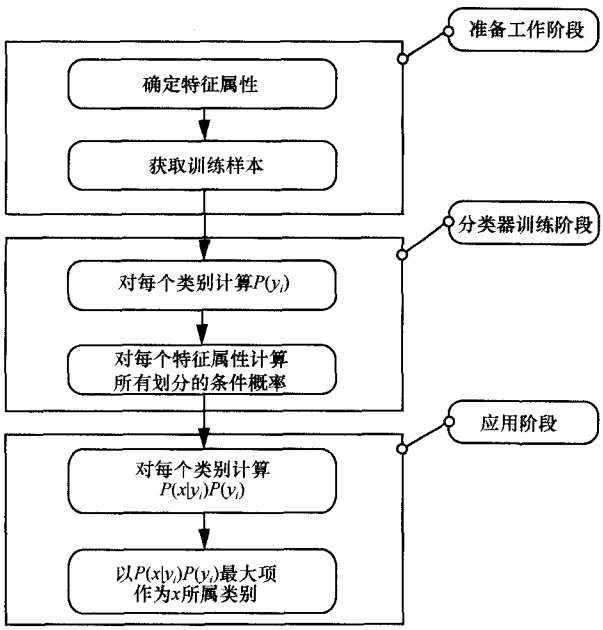

图 1 朴素贝叶斯分类流程

然而，朴素贝叶斯分类也有缺陷，它的假设是基于“各特征项相互条件独立”。在很多的实际问题中，如果此下设表现不够明显，甚至出现不成立时，错误的分类将会出现，从而影响算法的最终表现。在本文中，贝叶斯模型的使用将会被改善，而具体的内容将会在第 3 节中被介绍。

3 主题模型在垃圾邮件过滤中的研究

3.1 基于关键词的垃圾邮件过滤

3.1.1 算法思想

主要算法思想是基于关键词技术，采用朴素贝叶斯分类方法得到关键词，分析邮件内容分类到垃圾邮件的置信概率，进而产生分类结果。这种方法的优势在于复杂度低，且应用范围较广。

3.1.2 基于关键词的邮件过滤算法流程

从内容上看，邮件过滤可以看成一个二值分类问题，即把邮件分为垃圾邮件类和合法邮件类。基于关键词的邮件过滤算法流程简单来讲是朴素贝叶斯方法，贝叶斯过滤算法大致由以下基本步骤组成。

步骤 1 收集大量的垃圾邮件和合法邮件，建立垃圾邮件集和合法邮件集。

步骤 2 提取邮件主题和邮件体中的独立字符串，例如 sale、cash 等作为 token 串并统计提取出的 token 串出现的次数即字频。按照上述的方法分别处理垃圾邮件集和合法邮件集中的所有邮件。采用贝叶斯文本分类法对训练样本学习，得到 $P(S|W)$。

步骤 3 每一个邮件集对应一个散列表，合法邮件集对应表 hashtable_good，垃圾邮件集对应表 hashtable_bad，表中存储 token 串到字频的映射关系。

步骤 4 计算每个散列表中 token 串出现的概率，可以得到 $P_1(t_i)$和 $P_2(t_i)$，$P_1(t_i)$表示 t_i 在 hashtable_good 中的值（也就是 token 串 t_i在合法邮件中的概率）；$P_2(t_i)$表示 t_i在 hashtable_bad 中的值（也就是 token 串 t_i在垃圾邮件中的概率）：

$$P=\frac{\text{某token串的字频}}{\text{对应散列表的长度}} \tag{1}$$

步骤 5 由步骤 2 中贝叶斯文本分类法得到的 $P(S|W)$，综合考虑散列表 hashtable_good 和 hashtable_bad，推断出当新来的邮件中出现某个 token 串时，该新邮件为垃圾邮件的概率。计算式为：

$$P(A|t_i)=\frac{P_2(t_i)}{P_1(t_i)+P_2(t_i)} \tag{2}$$

其中，A 事件表示邮件为垃圾邮件；$t_1,t_2,\cdots,t_n$代表 token 串；$P(A|t_i)$表示当 token 串 t_i出现在所收到的邮件中时，该邮件为垃圾邮件的概率。

假设该邮件共得到 N 个 token 串 $t_1,t_2,\cdots,t_n$，hashtable_probability 中对应的值为 $P_1,P_2,\cdots,P_n$，$P(A|t_1,t_2,\cdots,t_n)$表示在邮件中同时出现多个 token 串 $t_1,t_2,\cdots,t_n$时，该邮件为垃圾邮件的概率。

由联合概率公式可得：

$$P(A|t_1,t_2,\cdots,t_n)=\frac{P_1\times P_2\times\cdots\times P_n}{P_1\times P_2\times\cdots\times P_n+(1-P_1)\times(1-P_2)\times\cdots\times(1-P_n)} \quad (3)$$

当 $P(A|t_1,t_2,\cdots,t_n)$超过预定阈值（例如 0.95）时，就可以判断邮件为垃圾邮件。

3.2 LDA 主题模型

LDA（latent Dirichlet allocation）的产生和发展历经 TF-IDF、LSA、pLSA 等多种主题模型方法，由于 LDA 模型的良好的数学基础和灵活的扩展性，一经提出即得到了来自各个领域研究者的关注，被广泛应用在文本挖掘及信息处理的研究中[14]。

LDA 模型最初是作为一种文本分类和主题聚类方法被提出，它将文档集中每篇文档的主题以概率分布形式给出，从而通过分析便能够得到聚类结果。与此同时，它是一种典型的词袋模型。也就是说，每篇文档将会被分解为一组词，而不用考虑先后顺序。

LDA 是一个三层的贝叶斯概率生成模型，由“主题—词语”和“文档—主题”构成。在 LDA 模型中需要求解“词语—主题”和“主题—文档”两个模型参数。LDA 假设文本集 D 中各文本 w 有如下生成过程，如图 2 所示，T 表示主题的个数，D 表示文档的个数，N_d 表示第 d 篇文档中词语的个数。

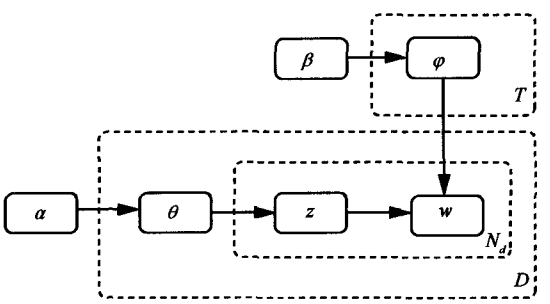

图 2　LDA 模型

步骤 1　确定文档中的词语数 N，使之服从参数为 ξ 的泊松分布。

步骤 2　确定 θ，使之服从参数为 α 的狄利克雷分布。

步骤 3　对于文本中 N 个词中的每一个 w_n：确定一个主题 z_n，使之服从参数为 θ 的多项式分布；依照概率 $p(w_n|z_n,\beta)$ 选择每一个词语 w_n。

3.3 基于主题模型的垃圾邮件过滤方法

基于主题模型抽取垃圾邮件的主题，对已知的垃圾邮件样本进行训练，提取垃圾邮件的特征，采用贝叶斯估计分类算法，构造垃圾邮件的过滤器。利用得到的垃圾邮件过滤器，对新的邮件进行分析、判断，区分垃圾邮件和合法邮件，实现垃圾邮件的过滤。

具体实现步骤如下。

步骤 1　采集一定数量的垃圾邮件与合法邮件，建立相应的垃圾邮件集和合法邮件集，计算词频得到每个词语出现的情况下该邮件是垃圾邮件的概率 $P(S|W)$。

步骤 2　利用 LDA 主题模型对邮件进行主题抽取，分类算法对已知的垃圾邮件样本进行训练，对垃圾邮件集和合法邮件集中的邮件进行解析，并提取邮件的特征，统计相应数据。LDA 是一种文档主体生成模型，也成为一个三层贝叶斯概率模型，包含词、主体、文档这三层结构。生成模型，即一篇文章的每个词都是通过以一定的概率选择了一个主题，并从这个主题中以一定的概率选择这个词语的过程得到的。

步骤 3　由联合概率公式计算每个主题中所有词语的联合概率；得到每个主题出现的情况下该邮件是垃圾邮件的概率；构造邮件分类器。

步骤 4　采用贝叶斯分类器。选取一个判断垃圾邮件适当的阈值，利用所建立的邮件分类器实现对邮件的分类。

3.4 模型的改进

传统判断两个文档相似性的办法是查看两个文档共同出现的单词的多少，如 TF-IDF 等，但这种办法没有考虑到文字背后的语义关联，有可能两个文档说的是相似的内容但并没有词语上的交集。LDA 提取出来的邮件主题关键词能够表达邮

件较高级别的主题内容，能够消除主题关键词之间的歧义。但是此时每个主题关键词并不是使用向量表达，此时本文使用 Word2Vec 方法，将词语转化为向量空间，有利于计算词语之间的相似程度。同时使用主题词向量距离计算方式计算距离主题最近的词语，即用 Word2Vec 生成每个主题中词语的关联词，作为主题词语的扩容，在此基础上再进行垃圾邮件判断。

4 垃圾邮件过滤器的设计及实验分析

4.1 邮件样本集的选取

4.1.1 垃圾邮件过滤的语料库

本文采用的垃圾邮件语料库从网上采集，包含正常邮件和垃圾邮件各 8 000 封。图 3 为比较典型的用于广告的垃圾邮件案例。

> 您好！
> 尊敬的客户：本公司长期代理进出口报关业务。有些发票可以为广大客户优惠代开（税率1.5%左右）以解广大客户财务票据的不足。具体有（增值税专用发票、国税商品销售专用发票、地税运输专用发票、建筑安装专用发票；广告专用发票；还有其他服务发票）等，希望有意者来电详谈，愿合作愉快，成功！可验证后付款!!
> 联系人：……
> 手　机：……
> 电　话：……
> 邮　箱：……
> 地　址：……

图 3　垃圾邮件案例

用这两类邮件建立垃圾邮件过滤器中词的先验概率。过程如下。

首先，解析所有邮件，提取每一个词。然后，计算每个词语在正常邮件和垃圾邮件中的出现频率。例如，假定“发票”这个词，在 8 000 封垃圾邮件中，有 200 封包含这个词，那么它的出现频率就是 2.5%；而在 8 000 封正常邮件中，只有 2 封包含这个词，那么出现频率就是 0.025%。有可能某个词在已有的某一类邮件语料中未出现，为了避免该词的先验概率出现为 0 的情况，设定该词的出现频次为 1。假设某个词只出现在垃圾邮件中，正常邮件中没有，就设定它在正常邮件的出现频率是 0.012 5%(1/8 000)，反之亦然。随着邮件数量的增加，词的先验概率计算结果会更接近于真实情况。

4.1.2 垃圾邮件评价指标

为了对垃圾邮件过滤系统的效果做分析，需要一个评价体系来进行评估，即一个系统可以判定未知文档是否属于某类。假定有 N 个邮件文档通过分类器分别分类，可以用表 1 来表示人工与系统对邮件的评判情况。A 为人工与系统都评判为垃圾的邮件数；B 为人工评判为正常，而系统评判为垃圾的邮件数；C 为系统评判为正常，而人工评判为垃圾的邮件数；D 为人工与系统都评判为正常的邮件数。

表 1　垃圾邮件测评

	实际为垃圾邮件	实际为正常邮件
判定为垃圾邮件	A	B
判定为正常邮件	C	D

定义如下几个指标来检测算法对垃圾邮件的过滤效果。

（1）召回率（recall）

描述收到一封垃圾邮件时，分类器判定为垃圾邮件的概率，召回率越高，表示分类器对邮件分类效果越显著，计算式为：

$$R=A/(A+C)\times 100\% \quad (4)$$

（2）正确率（precision）

描述分类器对正常邮件和垃圾邮件都能正确分辨的概率，将垃圾邮件判为垃圾邮件和将非垃圾邮件判为合法邮件的概率，正确率越高表示分类器的效果越理想，计算式为：

$$P=A/(A+B)\times 100\% \quad (5)$$

（3）误判率（misjudge）

描述正常邮件的误判率，将非垃圾邮件判为垃圾邮件的概率，这是描述一个分类器是否有效

的关键指标，如果误判率很高，则说明分类器没有起到很好的分类效果，误判率越低表示正常邮件被判为垃圾邮件的概率越小，计算式为：

$$M=B/(B+D)\times 100\% \quad (6)$$

（4）精确率（accuracy）

分类器对正常邮件分类的正确性，精确率越高表示邮件对正常邮件的判别越正确，计算式为：

$$A=(A+D)/(A+B+C+D)\times 100\% \quad (7)$$

在对实验结果的评估中将会比较以上数值。准确率 P 是邮件被正确分类的概率，召回率是指实验方法将邮件正确分类的概率，$F1$ 值则是指 β=1 时的 F 值，是最常用的 F 值之一，可以看作模型准确率和召回率的一种加权平均。这 3 个值都是数值越高所代表的分类效果越优秀。

4.2 实验结果与分析

4.2.1 LDA 主题抽取

步骤 1 首先用 jieba 分词算法分词后得到 300 个分词文件，名称如 1-seg.txt、2-seg.txt 等。例如，“合金”“批发”“朋友”“爸妈”等词语。

步骤 2 再用 LDA 主题模型算法解析 300 封邮件，得到 20 个主题词组，如：“0.090*‘交涉’+0.090*‘小白脸’+0.090*‘力阻’+0.090*‘撕破脸’+0.090*‘私事’”。

步骤 3 最后得到 300×20 维的权值矩阵，300 表示 300 封邮件，20 表示 20 个主题，即每封邮件和 20 个主题之间的相关度。

采用 LDA 主题模型算法，从测试集中选取 300 封邮件进行主题抽取 20 个主题，主题词确定为 10 个，选取两个具有代表性的主题，结果见表 2。

4.2.2 LDA 反垃圾邮件过滤实验结果与分析

为了有效地验证该方法的可行性，选用正常邮件和垃圾邮件各 8 000 封，共 16 000 封作为训练集；另取正常邮件和垃圾邮件各 150 封，共 300 封作为测试集。用本文基于 LDA 的垃圾邮件过滤方法进行实验，其中主题数确定为 20 个，主题词为 10 个，共完成 5 组实验，结果见表 3。

表 2　主题模型结果

主题	主题词	词与主题的相关度
主题 1	贵公司	0.090
	运输	0.089
	负责人	0.089
	发票	0.086
	财务	0.086
	代	0.085
	有限公司	0.084
	实业	0.084
	开	0.083
	广告	0.083
主题 2	帐	0.088
	出任	0.081
	总裁	0.081
	恪守	0.081
	微软公司	0.081
	相违	0.081
	北京时间	0.081
	全文	0.081
	新浪	0.081
	微软	0.081
主题 3,4,…,20	…	…

表 3　基于 LDA 的垃圾邮件过滤方法测试结果

	阈值设定	召回率	正确率	$F1$ 值
1	0.40	86%	85%	85%
2	0.44	79%	94%	85%
3	0.45	76%	95%	84%
4	0.42	88%	92%	89%
5	0.44	86%	88%	86%
平均	0.43	83%	90%	86%

在实验中，把垃圾邮件的概率跟合法邮件的概率做比较，需要选择判定垃圾邮件概率的阈值。阈值的控制比较重要，如果太大则会漏掉大量垃圾邮件，通过实验确定最佳阈值为 0.43 左右。

（1）与其他方法比较

为了更好地说明本文设计算法的有效性，本文选取了积累典型的垃圾邮件过滤方法进行比较，包括基于 Naïve Bayes 的邮件过滤方法[15,16]、基于 SVM 的邮件过滤方法[17]、基于 kNN 的邮件过滤方法[18]、基于 MTM（message topic model）的邮件过滤方法[3]、基于决策树的三步邮件过滤方法[4]、基于 SVM-NB 的邮件过滤方法[5]。其中，

前 3 种方法是以简单机器学习为基础的邮件过滤方法，MTM 方法建立了一种有效的方式，用来检测邮件主题词，三步邮件过滤法是以决策树为基础的，而 SVM-NB 方法是基于朴素贝叶斯分类分类的方式，对比见表 4。

表 4　各种不同方法邮件测试结果

方法	召回率	正确率	*F*1 值
Naïve Bayes[15]	66%	90%	76%
SVM[17]	81%	88%	84%
*k*NN[18]	73%	80%	76%
MTM[3]	80%	82%	77%
决策树[4]	79%	81%	79%
SVM-NB[5]	85%	87%	82%
本文提出的 LDA 方法	83%	90%	86%

由表 4 可知，本文基于 LDA 的垃圾邮件过滤方法使垃圾邮件的召回率相比 Naïve Bayes 方法、SVM 方法、*k*NN 方法、MTM 方法、决策树方法有很大提升，分别上升了 17%、2%、10%、3%、4%；识别正确率和 Naïve Bayes 相同，而相比 SVM 方法、*k*NN 方法则分别提高了 2%、10%；*F*1 值相比 Naïve Bayes 方法、SVM 方法、*k*NN 方法、MTM 方法、决策树方法以及 SVM-NB 方法分别提高了 10%、2%、10%、3%、4%、2%。

在基于决策树三步邮件过滤方法中，它利用决策树模型构建了三步法垃圾邮件过滤模式。对于 SVM-NB 算法，它提出了基于朴素贝叶斯分类器的训练集分类方法，提升了数据处理的顽健性，此方法能够达到较高的垃圾邮件检测精度。相比于这两种方法，本文推出的 LDA 算法能够更好地提取文本特征，从而达到更高的分类精度。基于 LDA 的垃圾邮件过滤方法在垃圾邮件正确率方面和 Naïve Bayes 方法相同，在垃圾邮件的召回率方面高于这 3 种方法，并且具有较高的 *F*1 测试值。这说明该方法在性能上要优于 Naïve Bayes、SVM、*k*NN 方法。

（2）采用不同主题数的结果比较

在实验（1）中，选择主题数为 20 个，主题词为 10 个，共完成了 5 组实验，并且与另外 3 种邮件过滤方法进行了比较。在本实验中选择主题词仍为 10 个，分别选择主题数为 10、15、20、25 个进行实验，结果见表 5。

表 5　不同主题数下的测试结果

主题个数	阈值设定	召回率	正确率	*F*1 值
10	0.32%	80%	80%	80%
15	0.47%	85%	84%	84%
20	0.40%	86%	85%	85%
25	0.47%	90%	88%	88%

分析实验结果如图 4、图 5 所示，在选取合适阈值的条件下，系统的召回率和正确率随着主题数的增加而提高。其原因是，随着主题数的增加，对测试集邮件的语义划分更明确，进而使得系统的召回率和正确率明显提升。基于这样的原理，本方法可以取得较好的垃圾邮件过滤结果。

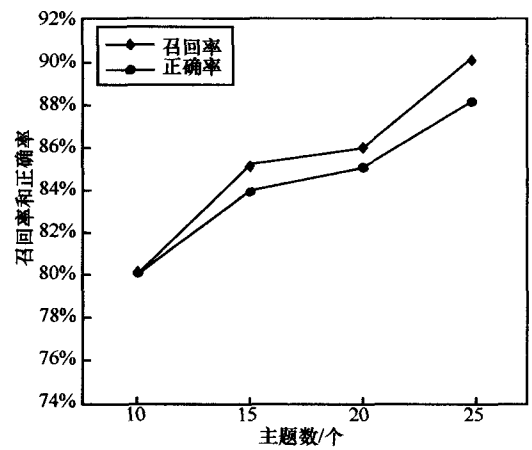

图 4　不同主题数下的召回率和正确率

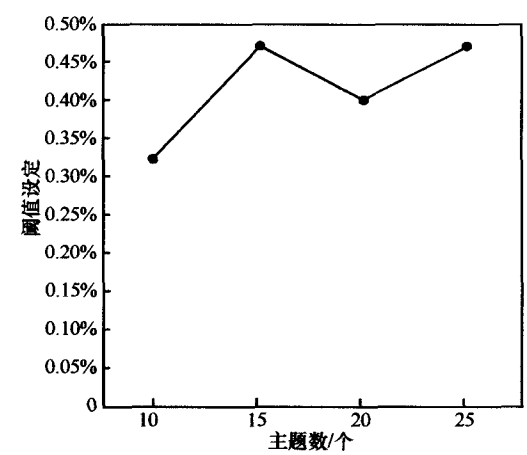

图 5　不同主题数下的阈值

4.2.3　基于改进的主题垃圾邮件过滤方法实验结果与分析

改进的方法主要将获得的主题进行扩展，用 Word2Vec 方法得到每个主题中词的几个相关的词，将获得的词重新构建主题组。

下面列举几个主题词经过 Word2Vec 计算后得到的结果，如图 6 所示。

个性	0.753 134 727 478
性格	0.744 346 022 606
开朗	0.742 993 593 216
坦率	0.740 987 539 291
直爽	0.736 198 425 293
爽朗	0.730 711 936 951
大而化之	0.721 120 536 327
活泼开朗	0.719 886 183 739
天然呆	0.719 811 201 096
轻浮	0.715 863 049 03

(a) 和“随和”最相关的词

致歉	0.737 105 131 149
回应	0.723 619 818 687
此事	0.656 879 663 467
否认	0.652 116 060 257
对此	0.651 973 962 784
表示遗憾	0.624 698 638 916
表示歉意	0.624 465 227 127
声明	0.622 716 546 059
毫无诚意	0.608 891 963 959
质疑	0.608 388 423 92

(b) 和“道歉”最相关的词

交通	0.527 298 688 889
陆上运输	0.521 845 817 566
铁路	0.487 869 590 521
货运	0.486 806 541 681
运输经济	0.486 096 262 932
运输网络	0.478 215 456 009
联合运输	0.476 194 471 121
公共	0.474 913 895 13
运输业	0.473 794 817 924
交通系统	0.472 240 090 37

(c) 和“运输”最相关的词

收据	0.638 935 208 321
现金支票	0.583 482 027 054
银行账号	0.571 007 490 158
持票人	0.570 166 230 202
收款	0.566 872 000 694
退款	0.562 694 907 188
开具	0.558 761 715 889
挂失	0.556 449 472 904
转账	0.547 636 032 104
持卡人	0.545 845 985 413

(d) 和“发票”最相关的词

图 6　主题词经过 Word2Vec 计算后得到的结果

以其中一个主题为例，通过 Word2Vec 扩展原主题词为 12 个词，见表 6。

表 6　原主题与扩展主题

原主题词	词与主题相关度	对应的扩展主题词	词与主题相关度
帐	0.088		
出任	0.081		
总裁	0.081	职位	0.081
恪守	0.081		
微软公司	0.081		
相违	0.081		
北京时间	0.081		
全文	0.081		
新浪	0.081		
微软	0.081	科技	0.081

这里实验过程分两个步骤。

步骤 1　将重建的主题组再次经 LDA 算法获得权值矩阵。

步骤 2　再用测试集进行测试，得到最终的实验结果。

在本实验中选择主题数为 20 个，主题词由 10 个扩展为 12 个，共完成了 5 组实验。

测试结果见表 7。

表 7　基于 Word2Vec 改进的垃圾邮件过滤方法测试结果

	阈值设定	召回率	正确率	*F*1 值	表 3 中 *F*1 值
1	0.48	88%	88%	88%	85%
2	0.49	86%	92%	90%	85%
3	0.47	83%	84%	83%	84%
4	0.49	81%	83%	81%	89%
5	0.48	90%	88%	88%	86%
平均	0.48	85%	87%	85%	86%

分析表 7，改进方法在增加 2 个关联主题词的情况下，在 *F*1 值上比原方法改进明显，在 5 次实验中有 3 次获得了较大的提高，证明了改进方法的有效性。

5　结束语

本文对基于主题模型的垃圾邮件过滤系统的设计与实现进行了分析和验证，与传统的关键词检测过滤技术相比，贝叶斯过滤算法更加有效且智能，从而提升了系统的安全性与顽健性。通过与其他典型垃圾邮件过滤方法的对比及验证，证明基于主题模型的垃圾邮件分类方法及基于 Word2Vec 的改进方法均能有效提高垃圾邮件过滤的准确度。

在未来的研究中，基于语义的文本分类具有非常大的潜力。针对自然语言的具体层次结构，机器学习与深度学习的方式已经在其他领域表现出非常强大的处理能力。在这种背景下，邮件拦截方法的设计可以参考相关研究成果进行深入探索。总之，未来的邮件拦截系统将会

具有非常大的改进空间，因此相关的研究需要被重点关注。

参考文献：

[1] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[2] 祝毅鸣, 张波. 实时黑名单在垃圾邮件过滤系统中的应用[J]. 科技资讯,2012(12):33.
ZHU Y M, ZHANG B. Application of real time blacklist in spam filtering system[J]. Science & Technology Information, 2012(12):33.

[3] MA J, ZHANG Y, WANG Z, et al. A message topic model for multi-grain SMS spam filtering[J]. International Journal of Technology & Human Interaction, 2016, 12(2):83-95.

[4] SHEN J J, CHEN Y K, CHU K T, et al. An intelligent three-phase spam filtering method based on decision tree data mining[J]. Security & Communication Networks, 2016, 9(17): 4013-4026.

[5] FENG W, SUN J, ZHANG L, et al. A support vector machine based naive Bayes algorithm for spam filtering[C]// 2016 Performance Computing and Communications Conference, Dec 9-11, 2016, Las Vegas, NV, USA. New Jersey: IEEE Press, 2017:1-8.

[6] BANSAL R P, HAMILTON I R A. O'CONNELL B M, et al. System and method to control email whitelists: US, US 8676903 B2[P]. 2014.

[7] CHAN P P K, YANG C, YEUNG D S, et al. Spam filtering for short messages in adversarial environment[J]. Neurocomputing, 2015, 155(C):167-176.

[8] DEVI K S, RAVI R. A new feature selection algorithm for Efficient Spam Filtering using Adaboost and Hashing techniques[J]. Indian Journal of Science & Technology, 2015, 8(13).

[9] AFZAL H, MEHMOOD K. Spam filtering of bi-lingual tweets using machine learning[C]// International Conference on Advanced Communication Technology, Jan 31-Feb 3, 2016, Pyeongchang, South Korea. New Jersey: IEEE Press, 2016.

[10] DAS M, BHOMICK A, SINGH Y J, et al. A modular approach towards image spam filtering using multiple classifiers[C]//2014 IEEE International Conference on Computational Intelligence and Computing Research. Dec 20, 2014, Coimbatore, India. New Jersey: IEEE Press, 2015:1-8.

[11] 曹玉东, 刘艳洋, 贾旭, 等. 基于改进的局部敏感散列算法实现图像型垃圾邮件过滤[J]. 计算机应用研究, 2016, 33(6):1693-1696.
CAO Y D, LIU Y Y, JIA X, et al. Image spam filtering with improved LSH algorithm[J]. Application Research of Computers, 2016, 33(6):1693-1696.

[12] 徐凯, 陈平华, 刘双印. 基于 Adaboost-Bayes 算法的中文文本分类系统[J]. 微电子学与计算机, 2016, 33(6):63-67.
XU K, CHEN P H, LIU S Y. A Chinese text classification system based on Adaboost-Bayes algorithm[J]. Microelectronics & Computer, 2016, 33(6):63-67.

[13] 周庆良. 一种基于 Adaboost 和分类回归树的垃圾邮件过滤算法[D]. 武汉: 华中科技大学, 2016.
ZHOU Q L. A spam filtering algorithm based on Adaboost and classification regression tree[D]. Wuhan: Huazhong University of Science and Technology, 2016.

[14] SMITH D A, MCMANIS C. Classification of text to subject using LDA[C]//2015 IEEE International Conference on Semantic Computing (ICSC), Feb 7- Feb 9, 2015, Anaheim, CA, USA. New Jersey: IEEE Press, 2015: 131-135.

[15] 赵治国, 谭敏生, 李志敏. 基于改进贝叶斯的垃圾邮件过滤算法综述[J]. 南华大学学报: 自然科学版, 2006, 20(1): 33-38.
ZHAO Z G, TAN M S, LI Z M. Review of spam filter algorithms based on improved Bayes[J]. Journal of Nanhua University(Science and Technology), 2006, 20(1): 33-38.

[16] 林巧民, 许建真, 许棣华, 等. 基于贝叶斯算法的垃圾邮件过滤技术[J]. 南京师范大学学报: 工程技术版, 2005, 5(4): 61-64.
LIN Q M, XU J Z, XU D H, et al. Research on Bayes-based spam filtering[J]. Journal of Nanjing Normal University(Engineering and Technology), 2005, 5(4): 61-64.

[17] LI L, MAO T, HUANG D. Extracting location names from Chinese texts based on SVM and KNN[C]// 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering(IEEE NLP-KE'05), Oct 30-Nov 1, Wuhan, China. New Jersey: IEEE Press, 2005: 371-375.

[18] 林文香. 改进的KNN算法在过滤垃圾邮件中的应用研究[D]. 长沙: 湖南大学, 2010.
LIN W X. Application of improved KNN algorithm in spam e-mail filtering[D]. Changsha: Hunan University, 2010.

[作者简介]

寇晓淮（1989–），男，华东理工大学信息科学与工程学院硕士生，主要研究方向为信息分析与处理、智能信号处理和网络与信息安全。

程华（1975–），男，博士，华东理工大学信息科学与工程学院副教授，主要研究方向为信息安全、信号处理、网络行为学和流量工程。