文章编号：1003-0077（2014）00-0000-00

基于多层 grams 的在线支持向量机的中文垃圾邮件过滤*

沈元辅[1]，沈跃伍[2]

（1.哈尔滨理工大学 图书馆，黑龙江 哈尔滨 150080；

2. 哈尔滨理工大学 计算机科学与技术学院，黑龙江 哈尔滨 150080）

摘 要：本文提出一种多层 grams 特征抽取方法来提升在线支持向量模型的垃圾邮件过滤器。在线支持向量机模型的垃圾邮件过滤器在大规模性垃圾邮件数据集已取得了很好的过滤效果，但与逻辑回归模型相比，耗时的计算性能是巨大的，很难被工业界所运用。本文提出的多层 grams 特征抽取方法能够有效果减少特征数据，抽取更精准有效的特征，大幅降低模型的运行时间，同时提升过滤器的过滤效果。实验表明，该方法使得在线支持向量机的模型的运行时间从 10337s 减少到 3784s，同时模型（1-ROCA）%提升一倍。

关键词：特征抽取；支持向量机；垃圾邮件过滤。

中图分类号：TP391　　**文献标识码：**A

Using Multi-level-grams to Improve Online SVM Based Chinese Spam Filtering

SHEN Yuanfu[1], SHEN Yuewu[2]

(1. Library, Harbin University of Science and Technology, Harbin 150080, China;

2. School of Computer Science and Technology,

Harbin University of Science and Technology, Harbin 150080, China)

Abstract: In this paper, we propose Mix-grams method to improve online SVM filter for spam filtering. Though online SVM classifier gives high classification performance on online spam filtering on large benchmark data sets, its computational cost turns out to be very expensive for other faster methods such as Logistic Regression. In this paper, we use Mix-Grams method to reduce feature vector dimension of online SVM filter. Experimental results demonstrate that the method greatly improves the filter performance and reduces the computational cost of online SVM filter.

Key words: Feature Extraction; Support Vector Machine (SVM); Spam filtering

1 引言

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。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。

很多研究人员提出了多种解决垃圾邮件的方法。较早的技术是基于黑白名单的过滤技术[1]。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

收稿日期：2014-06-15 定稿日期：2014-07-27

作者简介：沈元辅(1960--)，男，副教授，主要研究方向为信息检索，信息过滤；沈跃伍(1986--)，男，硕士研究生，主要研究方向：机器学习、数据挖掘、特征抽取。

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[2]。

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[3]。

基于内容的垃圾邮件过滤是一个机器学习过程，机器学习技术通常分为生成模型（如朴素贝叶斯模型）和判别模型（如支持向量机模型、逻辑回归模型）。Goodman and Hulten[4]在 PU-1 数据集上测试发现，判别模型的效果要好于生成模型。通过 TRCE 和 CEAS 会议的垃圾邮件评测竞赛发现，取得最好效果的模型都是采用判别模型[5][6][7]。在 2007 年 TREC 垃圾邮件评测中取得最好的成绩是采用在线支持向量模型（online SVM）[7]。然而，在线支持向量模型消耗时间与样本数量成平方关系，即时间复杂度为 $O(n^2)$，逻辑回归模型和朴素贝叶斯模型时间复杂度仅为 O(n)。在 2007 年 Sculley and Watchman[8] 提出了松弛在线向量机（Relaxed Online SVM，简称 ROSVM）算法，大幅的降低了 online SVM 模型的运行时间。尽管 ROSVM 模型的运行时间被大幅降低，但与逻辑回归模型相比，消耗时间仍然是巨大的。如，ROSVM 运行完 TREC06p 数据需要 18541s，而逻辑回归模型仅需 238s，运行效率是 ROSVM 的 78 倍。

文本我们提出多层grams特征抽取方法来提升在线支持向量模型。目前，常用的特征抽取方法是基于词和基于n-grams的特征抽取方法。实验发现基于n-grams的特征抽取方法效果更好，但该方法抽取特征数据较多，如，一封邮件的字符串长度为3000个字符，则抽取的特征接近3000。大量特征数目是导致模型分类器高耗时的关键因素。因此本文提出了多层grams特征抽取方法提升在线支持向量机模型。实验结果表明，该方法不仅能大幅降低系统的运行时间，同时还大幅提升过滤器的过滤效果。

本文的组织结构：第二节介绍在线支持向量机；第三节介绍本文提出的多层grams特征抽取方法；第四节是本文实验和数据分析部分；结论在最后的第五节。

2 在线支持向量机分类器

2．1 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。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：

$$f(x)=w\bullet x+b \tag{2-1}$$

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 $f(x)=\pm 1$。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(2-2)形式：

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

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：

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。

$$minimize_{w,b,\xi}:\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i$$
$$subject\_to:y_i(w\bullet x+b)\geq 1-\xi_i,\forall i,\xi_i\geq 0 \quad (2\text{-}3)$$

2．2 在线支持向量机

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。

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[9]，因为 SMO 方法对线性支持向量机来说是最快的方法。

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，使用三个简化措施。

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 $y_i f(x_i)\leq$ 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 $y_i f(x_i)\leq M$，$(0\leq M\leq 1)$ 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。

3 多层 grams 特征抽取方法

本节我们将介绍基于多层 grams 的特征抽取方法，在介绍该方法之将前介绍基于分词和基于 n-grams 的特征抽取方法，分析存在的问题。

3. 1 基于词和 n-grams 的特征抽取方法

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。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[10]。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。目前比较成熟的分词系统有：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。在本文实验使用的 JE 分词系统。

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。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。

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

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

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…），每次向后滑动一个字符，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。

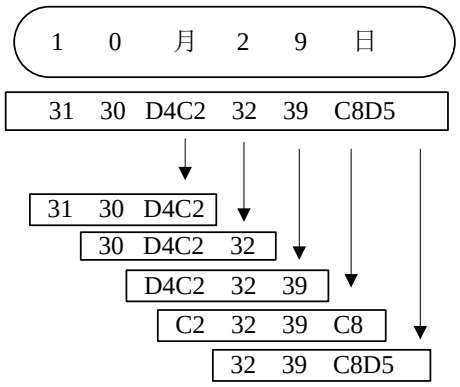
虽然基于 n-grams 的特征抽取方法已经垃圾邮件中取得了非常好的效果，但仍存在以下两个问题：

1）基于 n-grams 的特征抽取方法抽取的特征数目较多，特征数接近邮件内容的字符数，对大规模性系统来说，耗时非常大。

2）基于 n-grams 的特征抽取方法抽取中文邮件特征时，会出现半个汉字情况，半个汉字和其他字符会组成新的汉字，改变特

征原有的含义。

图 3-2：使用 n-grams 方法提取特征实例



3. 2 多层 grams 特征抽取方法

虽然基于 n-grams 特征抽取方法已由于该方法抽取的特征数目较多，导致 Online SVM 消耗的时间仍然过大。因此我们提出了多层 n-grams 特征抽取方法。该方法处理流程如图 2，处理过程如算法 1。

基于多层 grams 特征抽取方法相比 n-grams 方法有两大优点：1、能够大幅降低特征的数据，从而大幅降低 online SVM 模型的运行时间。2、能够抽取更为精准、更为有效的特征，减少无用和具有噪声的特征，从而提升模型过滤效果。

算法 1：多层 n-grams 特征抽取方法处理过程

Input： 一封邮件内容

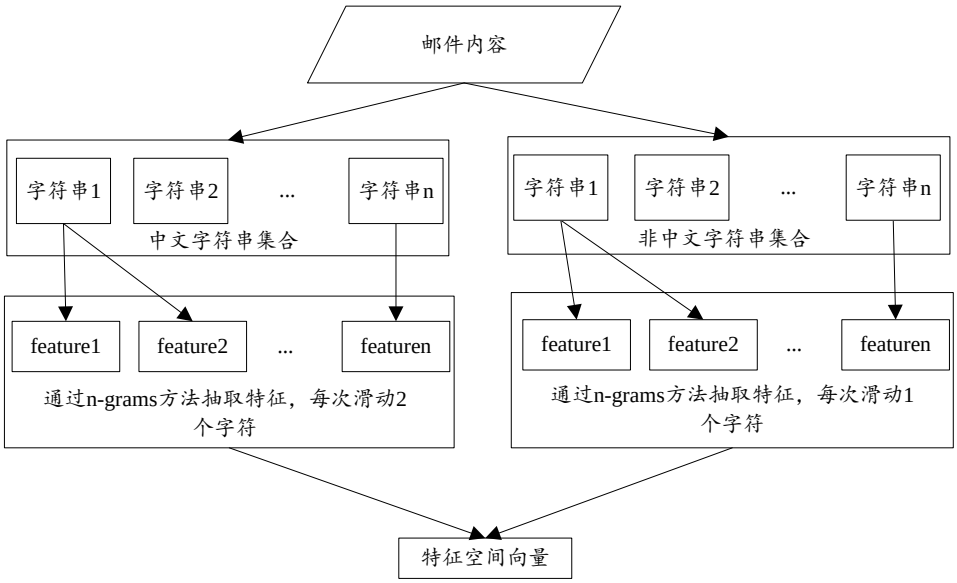
Step1： 将邮件内容按照中英文字符分隔成若干个字符串，每个字符串为全部中文字符或全部非中文字符，并将其分成全部中文字符串集合和全部非中文字符串集合。如邮件内容为："*12 月 12 日 10 个人在 Hilton Hotel 花费 1234567 元人民币*"，被分成若干个字符串，其中，中文字符串集合为："*月,日,个人在,花费,元人民币*"，非中文字符串集合为："*12,12,10, Hilton Hotel,1234567*"。

Step2： 对中文字符串集合中每个字符串使用 n-grams 特征抽取方法抽取特征，但滑动窗口每次往后移动 2 个字符。如：字符串"*元人民币*"一共为 8 个字符，通过 n-grams 方法抽取的特征为："*元人，人民，民币*"。

Step3： 对非中文字符串集合中每个字符串也使用 n-grams 方法抽取特征，但滑动窗口每次往后移动 1 个字段。

Step4： 对两个集合中抽取的特征组成特征空间向量。

图3-3:基于多层grams特征抽取框架



4 实验

在本节，我们将在大规模数据集上测试本文所提出的多层 grams 特征抽取方法，本节的实验结果很好证明了该方法能够提升基于在线支持向量机的中文垃圾邮件过滤器，大幅的降低了模型运行时间，同时提升过滤器的过滤性能。

4．1 实验数据集及评价指标

本文中所用的数据集 TREC06c，该数

据集是 TREC 会议于 2006 年举行了中文垃圾邮件评测数据，其数据集由清华大学提供。TREC06c 数据集的邮件数为 64620 封，其中垃圾邮件为 21766 封，正常邮件为 42854 封。

本文使用 $(1-ROCA)\%$和 $lam\%$ 作为过滤器的评估指标[5]。$hm\%$ 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。

$lam\%$ 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。

4. 2 实验设置

本节所有实验所测试邮件都只提取每封邮件的前 3000 个字符，涉及的 n-grams 特征抽取方法的 n 均为 4。在线支持向量机模型中，正则化参数 C=100，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。

4. 3 基于词和 n-grams 特征抽取方法比较

我们分别采用基于词和基于 n-grams 特征抽取方法在 online SVM 模型上效果，实验如果表 4.1，实验结果表明基于 n-grams 方法在过滤效果要远好于基于分词的方法，（1-ROCA）%由 0.007 提升到 0.004，提升近一倍。

在时间消耗方面，虽然基于分词的方法抽取特征数目明显小于 n-grams，但消耗的时间基本是接近，主要原因是中文分词花费时间较大，导致整体消耗的时间过大。

因此，基于 n-grams 特征抽取方法在 online SVM 模型处理中文垃圾邮件过滤时要明显优于基于分词的特征抽取方法。

表 4.1 基于词和基于 n-grams 特征抽取方法在 Online SVM 模型上指标对比

特征抽取方法	每封邮件的平均特征数	CUP 时间 (s)	lam%	(1-ROCA)%
基于词	527	9841	0.06	0.0007
基于 n-grams	**1625**	**10337**	**0.05**	**0.0004**

4. 4 多层 grams 和 n-grams 特征抽取方法比较

我们将基于多层 grams 的特征抽取方法和最优的 n-grams 方法进行比较，实验结果如表 4.2，通过表可以发现，多层 grams 特征抽取方法相比 n-grams 方法，邮件的平均特征特征由 1625 减少到 835，减少一倍，同时系统运行时间由 10337 减少到 3784，仅为原来的 1/3。但模型的（1-ROCA）%不但没有降低，反而提升了一倍，有原来的 0.004 提升到 0.002，效果显著。

通过表可以发现，多层 grams 特征抽取方法相比 n-grams 方法，邮件的平均特征数由 1625 减少到 835，减少一倍，同时系统运行时间由 10337 减少到 3784，仅为原来的 1/3。但模型的（1-ROCA）%不但没有降低，反而提升了一倍，有原来的 0.004 提升到 0.002，效果显著。

4. 5 与目前最新的垃圾邮件过滤比较

目前研究中比较常用的是基于逻辑回归的垃圾过滤器（LR Model）[11]，该方法不仅过滤效果好，而且运行速度快，普遍被工业届运用。同时，在 2013 年孙广路、齐浩亮等提出了在基于在线排序逻辑回归的垃圾邮件过滤器（Ranking LR Model）[12]，该方法极大提升了过滤器过滤效果。本文将与这两种方法进行比较，实验结果如表 4.3。从实验结果可以看出，本文提出的方法与其它两种方法相比，过滤效果是 LR Model 的

表 4.2 基于多层 grams 和 n-grams 特征抽取方法在 Online SVM 模型上指标对比

特征抽取方法	每封邮件的平均特征数	CUP 时间 (s)	lam%	(1-ROCA)%
N-grams	1625	10337	0.05	0.0004
多层 grams	**835**	**3784**	**0.04**	**0.0002**

5 倍，是 Ranking LR Model 的 3 倍，效果非常显著。上述数据再次证明了本文提出的方法能很好的被广泛使用。

表 4. 3 与目前最新模型的比较

Method	lam%	(1-ROCA)%
LR Model	0.07	0.0010
Ranking LR Model	0.05	0.0006
Online SVM Model (with our Method)	**0.04**	**0.0002**

4. 5 讨论

从表可以看出，文本提出的基于多层 grams 特征抽取方法，不仅降低了模型的特征数目，同时抽取特征能够更精准、更有效的反应垃圾邮件和正常邮件的特性，减少特征集中噪声特征，从而提升模型过滤性能。

5 结论

基于在线支持向量机模型过滤器在 TREC 2007 垃圾邮件评测竞赛中取得非常好的效果，然而与逻辑回归模型过滤器相比，消耗的计算代价是巨大。目前，很多研究人员更倾向于通过改变在线支持向量机模型的训练方法来减少模型的消耗的时间，很少有研究者从特征抽取的角度来提升模型的效率。然而，事实上特征的优化对模型的影响也是至关重要的，影响机器学习模型的计算速度不仅仅是样本的数量，还有特征维度，通过需要提升过滤器的性能，需要抽取能够反映数据特性的特征。因此，我们提出了基于多层 grams 特征抽取方法来提升在线支持向量模型的中文垃圾邮件过滤器。通过实验表明，该方法不仅能大幅减少系统的运行时间，而且还大幅提升系统的过滤效果。

参考文献

[1] G. Hulten and J. Goodman. Tutorial on Junk Mail Filtering[C]. Proceedings of the Twenty-First International Conference on Machine Learning Tutorial, 2004.

[2] T. Nathan, S. Timothy, C. Brad and V. George. Deterministic Memory-efficient String Matching Algorithms for Intrusion Detection[C]. Proceedings of IEEE INFOCOM, 2004: 2628-2639.

[3] Hongyu Gao, Yan Chen, Kathy. Lee Towards Online Spam Filtering in Social Networks. In Proceedings of the 19th Annual Network & Distributed System Security Symposium. 2012

[4] G. Hulten and J. Goodman. Tutorial on junk e-mail filtering. In ICML 2004, 2004.

[5] G. V. Cormack and T. R. Lynam. TREC 2005 spam track overview. In The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings, 2005

[6] G. V. Cormack. TREC 2006 spam track overview. In TREC 2006: Proceedings of the The Fifteenth Text REtrieval Conference, 2006.

[7] G. V. Cormack. TREC 2007 spam track overview. In TREC 2007: Proceedings of the The Sixteenth Text REtrieval Conference, 2007.

[8] D. Sculley and G. Wachman. Relaxed online SVMs for spam filtering. In The Thirtieth Annual ACM SIGIR Conference Proceedings, 2007

[9] J. Platt. 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.

[10] 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.

[11] 沈跃伍. 基于在线学习的垃圾邮件过滤技术研究[D]. 哈尔滨理工大学, 2012.

[12] 孙广路, 齐浩亮. 基于在线排序逻辑回归的垃圾邮件过滤[J]. 清华大学学报：自然科学版, 2013, 5: 734-740.

作者简介：作者一沈元辅（1960——），男，副教授，主要研究领域为信息检索，信息过滤。Email: yuanfu.shen@gmail.com；作者二沈跃伍（1986——），男，硕士研究生，主要研究领域为机器学习，数据挖掘，特征抽取。Email: yuewu.shen@qq.com。