基于类别选择的改进KNN文本分类

刘海峰 张学仁 姚泽清 刘守生
(解放军理工大学理学院 南京 210007)

摘 要 特征高维性以及算法的泛化能力影响了KNN分类器的分类性能。提出了一种降维条件下基于类别的KNN改进模型,解决了k近邻选择时大类别、高密度样本占优问题。首先使用一种改进的优势率方法进行特征选择,随后使用类别向量对文本类别进行初步判定,最后在压缩后的样本集上使用KNN分类器进行分类。试验结果表明,提出的改进分类模型提高了分类效率。

关键词 k-最近邻,特征降维,特征选择,文本分类
中图法分类号 TP391 **文献标识码** A

Improved Sort-based KNN Text Categorization Method

LIU Hai-feng ZHANG Xue-ren YAO Ze-qing LIU Shou-sheng
(Institute of Sciences, PLA University of Science and Technology, Nanjing 210007, China)

Abstract The problem of large feature dimension and the expansibility reduces the KNN function. This paper brougt forward an ameliorative KNN method to solve the problem that big swatch sort with more texts is easy to become the k-nearest neighbors under the feature reduction condition. Firstly, it used an improved odds radio method to select feature. Secondly, it estimated the possible sorts for the text by using the sort vector. Lastly, it used an improved KNN method in the reduced texts to realize text categorization. The experiment shows that this method has improved the precision.

Keywords KNN, Feature reduction, Feature selection, Text categorization

1 引言

文本是最基本的信息载体,文本分类是获取和组织大量文本数据的关键技术。随着信息技术的飞速发展,Internet上各种信息资源呈现指数级的增长趋势。为了有效地管理和利用这些纷繁杂芜的海量信息,基于内容的文本挖掘技术决定着基于Web的信息检索、信息过滤等技术的发展水平。

自动文本分类(text categorization, TC)是指在给定的类别体系下对未知类别的文本根据其内容将其自动划归到一类别或多类别的过程。文本自动分类技术在提高信息利用的有效性和准确性上具有重要的现实意义和广阔的应用前景。20世纪90年代以来得到了长足发展的基于机器学习的文本自动分类方法,在分类模式、算法的性能及其可扩展性、分类效果等方面均有所突破,文本自动分类技术研究已经成为文本挖掘领域的研究热点与核心技术。

文本分类器的计算开销和泛化能力是评价其分类性能的重要指标,也是文本自动分类研究的关键技术之一。目前,在众多的文本分类算法,如朴素Bayes[1]、支持向量机[2]、神经网络[3]、k-最近邻(K-Nearest Neighbor, KNN)[4]、决策树等方法中,KNN方法是一种简单有效的非参数方法,在准确率和召回率方面表现出众,在Reuters语料上是分类效果最好的算法之一。KNN方法是文本分类中最常用的分类器之一,针对其分类性能不足的改进研究成为人们关注的问题[5,6]。

2 KNN算法存在的问题及解决途径

作为一种基于实例的文本分类方算法,KNN算法被认为是向量模型(Vector Space Model)下最好的分类算法之一[7]。利用KNN算法对待分类文本进行分类的过程是:分类器在训练集里查找与待分类文本 d_j 最相似的 k 个文本,根据这些相似训练文本的类别所属情况给文本 d_j 的候选类别逐一评分,将最相似的 k 个文本和 d_j 的相似度作为 k 近邻文本所在类的类别权重,将各类中邻居文本的类权重之和作为该类别与文本 d_j 的相似度,把 d_j 归入相似度最大的类别中。

作为一种有监督机器学习的非参数方法,KNN分类器在文本分类方面效率高、使用简便。但是该算法具有一些明显的弱点[8]:

第一,KNN是一种懒惰的学习方法,计算过程均是在分类时才开始进行,因此体现出训练时快但用于分类时慢的特点,分类时间是非线性的:随着训练样本数的增加,分类用时急剧上升。

第二,确定待分类文本 d_j 的类别,需要计算其与训练样本集里所有样本之间的相似度,然后从中挑选出与其相似度

到稿日期:2008-12-31 返修日期:2009-03-02 本文受国家自然科学基金项目(编号:70571087)资助。
刘海峰(1962-),男,博士,副教授,CCF会员,主要研究方向为数据挖掘、文本分类,E-mail:liuhaifeng19620717@sina.com;张学仁(1955-),男,副教授,主要研究方向为人工智能、信息检索;姚泽清(1960-),男,教授,主要研究方向为统计分析、数据挖掘;刘守生(1965-),男,博士,副教授,主要研究方向为人工智能、遗传算法。

最大的 k 个文本。而文本分类时的训练样本常常是大样本，在数千篇的训练样本集上计算文本相似度，对普通的计算平台而言，其计算开销是难以承受的。随着训练样本数的增加，分类性能下降很快。

第三，KNN 分类器的分类效率受到训练样本分布情况的影响较大。当数据分布倾斜现象严重时，其分类性能可能变得很差。但是在实际使用时，却很难使得训练样本的分布达到均匀的要求，数据分布倾斜现象十分突出。以著名的英文文本分类语料 Reuters-21578 的 ApteMod 版本为例，最大类含有 2877 个文本，最小的类仅含有几个文本。82%的类中，文本数低于 100 个，在这种训练条件中进行分类训练时，大类别样本占有密度上的优势，从而更容易被选中，势必对分类决策函数形成干扰。

第四，在 VSM 下样本相似度的计算量太大。文本向量的特征高维性不仅使得数据稀疏现象严重，也使得算法的效率被严重降低。一篇普通的科技文献，其文本特征向量维数常常是几千维，而训练集所对应的特征维数可能高达上万维，训练过程中建立的高维特征向量含有很多虚假信息，这无疑增加了分类器在类别判断时的难度。因此，合理有效的特征降维是提高 KNN 分类效率的关键因素之一。

针对上述的前 3 个问题，可以考虑采取一种有效的判别方法，首先从训练集 S 里选出待分类文本 d_j 最可能归属的若干类别的训练文本，构成训练集的一个子集 S_1，在 S_1 上使用 KNN 分类器去判定 d_j 的类别归属；对于第 4 个问题，本文考虑使用特征选择方法，先对特征集进行约简，避免特征向量的高维性对分类器性能的干扰。

3 降维模式下的一种基于类别选择的改进 KNN 文本分类模型

文本特征的高维性制约着文本分类器性能。特征降维的效率影响着分类器的分类效率。基于第 1、2 节的讨论，提出一种降维模式下的基于类别选择的改进 KNN 文本分类方法：首先使用一种改进的优势率方法压缩文本特征维数，以节约计算机的存储空间，提高分类器计算性能；然后以训练样本的各个类别向量为依据，对待分类文本 d_j 初步判断其可能属于的几个类别，将这几类可能类别的训练文本构成新的训练文本集 S_1，在 S_1 上使用 KNN 方法寻找 d_j 的 k 个最近邻。

3.1 特征选择及其常用方法

特征选择是两类常用的特征降维方式之一。它一般是利用某个评价函数独立对原始特征项逐一评分，然后从中选出分值最高的前 p 个特征项，其结果是得到原始特征集的一个子集，本质上是对特征集合进行约简。常用的特征选择方法（如词频法、信息增益、互信息、期望交叉熵、文本证据权、优势率等），因为其评分模型构造相对简单、易于理解、使用方便而得到广泛应用。研究结果表明[9]，在样本分布相对均匀条件下优势率方法的性能略为占优。

3.2 基于优势率的特征选择方法及其改进

优势率（Odds Radio，OR）算法体现了特征项作为所属类别 c_i 的类内文本的特征的优势，其常见的计算公式如下：

$$OR(t_k,c_i)=\log\frac{p(t_k|c_i)[1-p(t_k|\bar{c}_i)]}{p(t_k|\bar{c}_i)[1-p(t_k|c_i)]}=\log\frac{N(t_k,c_i)\times N(\bar{t}_k,\bar{c}_i)}{N(t_k,\bar{c}_i)\times N(\bar{t}_k,c_i)} \tag{1}$$

其中，c_i 表示第 i 类文本集，$p(t_k|c_i)$ 表示第 i 类训练集里特征项 t_k 出现的概率，$p(t_k|\bar{c}_i)$ 表示在其它类别的训练集里特征项 t_k 出现的概率。而特征 t_k 在整个训练集里的权重可由下式计算：

$$OR(t_k)=\sum_{i=1}^{r}p(c_i)\log\frac{N(t_k,c_i)\times N(\bar{t}_k,\bar{c}_i)}{N(t_k,\bar{c}_i)\times N(\bar{t}_k,c_i)} \tag{2}$$

其中，$N(t_k,c_i)$ 为含有特征 t_k 的 c_i 类的文本数，$N(\bar{t}_k,\bar{c}_i)$ 为非 c_i 类且不含有特征 t_k 的文本数，$N(t_k,\bar{c}_i)$ 为非 c_i 类且含有特征 t_k 的文本数，$N(\bar{t}_k,c_i)$ 为 c_i 类的不含有特征 t_k 的文本数，r 为文本类别数。

分析式(1)可以看出：一方面 $p(t_k|c_i)[1-p(t_k|\bar{c}_i)]$ 表示特征 t_k 在 c_i 类里出现的次数越多，同时在非 c_i 类里出现次数越少，则对文本类别判定的作用就越大；另一方面，$p(t_k|\bar{c}_i)[1-p(t_k|c_i)]$ 体现了特征 t_k 相对于所有非 c_i 类的特征来说，其对于属于 c_i 类别的文本在类别判定上的优势。也就是说，两者的比率越大，越能体现特征 t_k 相对于非 c_i 类别的特征来说更适合作为 c_i 类别的文本的特征。

这种特征选择方法对仅属于一个类别的类别特征效果很好。然而，却有许多特征对于一些类别都比较重要，是属于这些类别的重要特征：在与其它类别的特征一起比较，它们可以有效地区分与其相应的文本类别。但是，由于这些特征同时属于多个类别，即在 c_i 类和非 c_i 类文本中都出现，其相对于式(1)的优势率却较低而在特征选择时被淘汰[10]。然而在式(2)中由于计算的是诸类别的加权优势率，因此虽然这些特征在个别的类别中优势率较低，但是在整个训练集上其加权优势率得以提高，从而有可能作为文本特征被保留下来。

因此，可以认为优势率作为文本特征选择的一种方法，应该针对不同的需求加以应用：在选择特征用以文本特征降维时，式(2)能够反映特征项 t_k 对文本类别界定上的作用；然而在选择类别特征构造类别特征向量时，式(2)的效果不理想。而式(1)虽然是针对单个类别的特征项的作用的度量准则，然而却难以选出那些对于几个类别都很重要的特征来，因而效率较低。基于此，考虑对式(1)进行改进后用于类别特征选择，以便构造类别特征向量。

分析式(1)可以看出，特征 t_k 相对于确定类别 c_i 的优势率不仅与其在 c_i 类里出现的次数有关，还与其在非 c_i 类文本里出现的次数相关，这是属于多类别的特征其优势率不高的主要原因。而式(1)中分母 $N(t_k,\bar{c}_i)\times N(\bar{t}_k,c_i)$ 既要求在非 c_i 类内特征 t_k 出现的次数越少越好，同时要求在 c_i 类内不含有特征 t_k 的文本数越少越好。这两个要求的“合力”加大了“非 c_i 类内特征 t_k 出现的次数越少越好”这一要求对特征选择的影响力度，过于强调了“非 c_i 类文本不含有特征 t_k”对优势率的作用。

基于此因素考量，对式(1)进行如下的改进：

$$\widetilde{OR}(t_k,c_i)=\log\frac{N(t_k,c_i)\times N(\bar{t}_k,\bar{c}_i)}{N(\bar{t}_k,c_i)} \tag{3}$$

即减少非 c_i 类文本不含有特征 t_k 这一要求对特征选择的干扰。并将上式用于文本类别特征的选择，用以构造类别特征向量。

4 基于优势率的文本向量及类别特征向量构造

4.1 文本向量的一种基于特征频数与优势率的低维表示

设训练文本集的类别数为 r，记为 $C_1,C_2,\cdots,C_r$；文本特

征个数为 N，对每个特征项 t_i，利用式(2)得到相应的优势率值。按照从高到低进行排序，取前 $n<<N$ 个优势率最大的特征 $t_1,t_2,\cdots,t_n$ 构成文本特征集 Q：$Q=\{t_1,t_2,\cdots,t_n\}$，其相应的优势率值记为 $OR(t_1),OR(t_2),\cdots,OR(t_n)$，将文本特征向量的维数从 N 维压缩到 n 维。在此基础上，将文本 d_j 表示成向量形式：$d_j=(w_{1j},w_{2j},\cdots,w_{nj})$，$j=1,2,\cdots,m$，$m$ 为文本数。

w_{ij} 表示文本 d_j 的第 i 个特征 t_i 在文本集里的权重，这里以常用的基于特征频数的 tf-idf 公式计算[11]：

$$w_{ij}=\frac{tf_{ij}\times\log(\frac{m}{df_{ij}}+0.01)}{\sum_{i=1}^{n}\left[tf_{ij}\times\log(\frac{m}{df_{ij}}+0.01)\right]^2} \tag{4}$$

其中，m 为文本数，df_{ij} 为文本集里出现特征 t_i 的文档数，tf_{ij} 为文本 d_j 里含有特征 t_i 的频率。

式(4)主要是基于特征项在文本集里的频数对其关于文本类别判定方面的重要性的度量，没有考虑特征项与文本之间的相关性信息。而式(2)体现了特征项与文本的相互关联的程度，所以可以将这一信息利用到文本的向量表示上去，以提高对文本标引的效率：

$$d_j=(w_{1j}OR(t_1),w_{2j}OR(t_2),\cdots,w_{nj}OR(t_n)) \tag{5}$$

然后将式(4)与式(3)相结合，构造文本 $d_j=(w_{1j},w_{2j},\cdots,w_{nj})$ 的低维的一种向量表示模型：

$$d_j=(OR(d_j,c_1),OR(d_j,c_2),\cdots,OR(d_j,c_r)) \tag{6}$$

这里

$$OR(d_j,c_k)=\sum_{p=1}^{n}w_{pj}\times\tilde{OR}(t_p,c_k),k=1,2,\cdots,r \tag{7}$$

这种表示方法不仅含有特征的频数信息，而且含有特征与文本的相关性信息，同时文本向量的维数仅仅是 $r<<n$ 维，计算开销得到很大的降低。

4.2 基于改进优势率的文本类别特征选择

分析式(3)可以看出，$\tilde{OR}(t_k,c_i)$ 体现了特征项 t_k 与文本类别 c_i 之间的相互关联程度。其优势率值越大，说明两者之间的相关性越强。为方便起见，可以从这 N 个优势率值 $\{\tilde{OR}(t_1,c_i),\tilde{OR}(t_2,c_i),\cdots,\tilde{OR}(t_N,c_i)\}$ 里从大到小取前 r 个值相应的特征为 c_i 的类别特征，其相应优势率值为特征项的权重构造类别 c_i 的类别特征向量；不失一般性，不妨将该类别 C_i 相应的特征向量记为：

$$C_i=(\tilde{OR}(t_1,c_i),\tilde{OR}(t_2,c_i),\cdots,\tilde{OR}(t_r,c_i)),i=1,2,\cdots,r \tag{8}$$

在此基础上，文本向量之间、文本与类别特征向量之间的相似性可以使用常用的向量夹角余弦进行度量：

$$\mathrm{Sim}(d_i,d_j)=\frac{\sum_{k=1}^{r}OR(d_i,c_k)\times OR(d_j,c_k)}{\sqrt{\sum_{k=1}^{r}OR^2(d_i,c_k)}\sqrt{\sum_{k=1}^{r}OR^2(d_j,c_k)}} \tag{9}$$

$$\mathrm{Sim}(C_i,d_j)=\frac{\sum_{k=1}^{r}\tilde{OR}(t_k,c_i)\times OR(d_j,c_k)}{\sqrt{\sum_{k=1}^{r}\tilde{OR}^2(t_k,c_i)}\sqrt{\sum_{k=1}^{r}OR^2(d_j,c_k)}} \tag{10}$$

4.3 基于相似度的待分类文本类别初步选择

在KNN算法使用前，对待分类文本可能的所属类别进行初次判别：试验中发现，直接使用式(9)计算待分类文本 d_j 与类别 C_i 的相似度，取相似度最大值对应的类别为 d_j 的所属类别：

$$C_{d_j\,\mathrm{fairinto}}=\mathrm{argmax}C_i(\mathrm{Sim}(C_i,d_j)) \tag{11}$$

这种方法也有较高的分类效率。但是，由于选择文本类别特征时不同的类别对应的类别特征并不相同，这种单类别的决策由于特征向量维数较少、诸类别特征子集之间的特征又不完全相同，影响了分类效率，使得分类效果不太满意。但是，如果仅将式(10)用于对待分类文本 d_j 可能所属类别的初步判定，然后结合KNN算法进一步确定 d_j 的所属类别，效果却很好。具体地说对于待分类文本 d_j 首先使用式(10)计算其与诸类别之间的相似度 $\mathrm{Sim}(C_i,d_j)$，$i=1,2,\cdots,r$，得到 r 个相似度值，然后取其最大的、占类别总数约三分之一的相似度对应的类别为 d_j 的候选类别。而试验中发现，绝大多数待分类文本的人工分类结果都在这些候选类别中。也就是说，这种基于可能类别的对待分类文本 d_j 类别归属的初步判定效率是相当的高。

当 d_j 相应的可能类别确定后，从训练文本集 S 里选出这些类别相应的文本，构成新的训练文本子集 S_1，在子集 S_1 上使用KNN分类器进行分类。

这种做法与经典的KNN算法需要计算 d_j 和整个训练样本集 S 里所有样本之间的相似度的模式相比，在 S_1 上使用KNN算法其计算量减少约三分之二。从计算的速度、存储空间的减少等方面看，新的方法大大减少了计算开销。

4.4 特征选择模式下基于训练子集的KNN算法实现

使用KNN分类器需要计算待分类文本与训练文本之间的相似度。一般说来，训练文本的类别如果不再细分的话，类别数不是很多，对于常用的一些文本的类别更是如此。而如果使用式(9)进行文本相似度计算，因文本向量的维数偏低，丢失的特征信息可能过多，从而影响分类效果。基于此种因素考虑，在文本之间进行相似度计算时，文本的表示使用式(5)。此时文本向量维数适中，且含有文本的绝大多数信息，有利于对文本的准确标引。

综上分析，提出改进的KNN算法步骤如下：

1)根据式(2)计算文本集诸特征的优势率值共 N 个，取其前 n 个最大值相应的特征构成文本特征集，实现基于选择的文本特征降维；

2) 使用式(4)进行文本的向量表示，使用夹角余弦计算文本之间的相似度；

3)使用式(3)进行文本的类别特征选择，在诸类别里分别选择其前 n 个最大值相应的特征构成类别特征集，进而使用式(8)表示诸类别向量；

4)使用式(6)表示文本向量；

5)使用式(10)计算文本与类别之间的相似度，按照相似度大小取其类别总数三分之一的类别为待分类文本 d_j 的可能所属类别，实现 d_j 的初次类别归属判定；

6)文本之间的相似度计算使用向量夹角余弦，文本的向量表示使用式(5)，最终在文本子集 S_1 上使用KNN算法确定 d_j 的所属类别。

5 实验及其结果分析

本文对上述方法的分类效果进行了试验。试验数据为网上随机下载的6683篇文本，共15个类别。文本类别与文本

数如表1所列。

表1 文本数与类别分布统计

经济	体育	股票	文学	军事	卫生	娱乐	房产	教育	电脑	科技	生物	艺术	汽车	法律
312	487	269	416	613	515	472	338	616	357	525	447	421	382	513

试验时采用4分交叉试验法，将6683篇文本分为4份，3份为训练集，1份为测试集；每份轮流作为测试集循环测试共4次，取平均值为测试结果。使用东北大学自然语言处理实验室提供的分词软件进行分词，用禁用词表过滤停用词，人工删除冷僻的特低频词，并剔除了虚词、助词、人称代词、特高频词等预处理后建立候选特征项集合，得到特征个数6218个，先用式(2)进行特征选择，选择权值最大的 $n=225$ 个特征构造文本向量。类别数 $r=15$，取待分类文本 d_j 初次可能所属的类别判断中的可能类别数为5，使用提出的改进KNN分类算法进行分类。经过反复试验，最终确定KNN算法中的参数 $k=7$。分类效果评估指标使用常用的查准率、查全率以及 F_1 测试值(相应的值以下标new进行标注)：

查准率＝分类的正确文本数/实际分类文本数

查全率＝分类的正确文本数/应有文本数

$$F_1=\frac{\text{查准率}\times\text{查全率}\times 2}{\text{查准率}+\text{查全率}} \tag{15}$$

并将传统KNN分类器(即不进行初次类属判定这一预处理过程，其它试验条件不变)分类效果与之比较(相应的值以下标old标注)，结果如表2所列。

表2 文本分类试验结果统计

	查准率$_{new}$	查准率$_{old}$	查全率$_{new}$	查全率$_{old}$	F_1 值$_{new}$	F_1 值$_{old}$
经济	0.833	0.722	0.817	0.743	0.825	0.732
体育	0.876	0.812	0.886	0.791	0.881	0.801
股票	0.943	0.871	0.932	0.841	0.937	0.856
文学	0.916	0.864	0.923	0.847	0.919	0.855
军事	0.896	0.837	0.919	0.853	0.907	0.845
卫生	0.952	0.886	0.895	0.816	0.923	0.827
娱乐	0.862	0.836	0.873	0.817	0.867	0.826
房产	0.823	0.782	0.801	0.776	0.812	0.779
教育	0.886	0.827	0.893	0.912	0.889	0.867
电脑	0.836	0.848	0.895	0.849	0.864	0.848
科技	0.915	0.927	0.931	0.879	0.923	0.902
生物	0.938	0.874	0.884	0.828	0.910	0.850
艺术	0.927	0.908	0.891	0.906	0.909	0.907
汽车	0.847	0.852	0.928	0.873	0.886	0.862
法律	0.926	0.873	0.919	0.846	0.922	0.859
平均值	0.892	0.848	0.893	0.839	0.892	0.841

从表2可以看出，改进的KNN分类算法在股票、科技、卫生、法律与文学类别的文本分类效果较好，经济与房产两类准确率略低一点，可能是由于这两类文本共用词组相对多一些，文本数也略少一些。改进后的模型分类的查准率、查全率、F_1 测试值均有所提高，其中 F_1 测试值平均提高约6.1%，改进的效果还是令人满意的。

结束语 特征降维的合理性和分类器的性能是文本分类技术研究的两个核心问题，本文在特征选择和分类器改造两个方面进行了研究，提出了一种特征降维模式下的改进KNN文本分类算法。特征降维问题是文本处理所必须面对的主要问题之一，是制约提高文本分类效率的瓶颈；根据训练集训练文本的不同分布状况构造合适的分类器成为影响分类效果的因素，提高分类器在大样本集、样本分布不均等环境下的健壮性是文本分类领域需要研究的关键问题。这也是今后应该继续进行的工作。

参考文献

[1] Ghosh A K, Chaudhuri P, Murthy C A. Multiscale classification using nearest neighbor density estimates[J]. IEEE Transactions on Systems, man, and Cybernetics-part b: cybernetics, 2006, 36(5): 1139-1148

[2] Debole F, Sebastiani F. An analysis of the relative hardness of reuters-21578 subsets[J]. Journal of the American Society for Information Science and Technology, 2004, 56(6): 584-596

[3] Jacobs D W, Weinshall D. Classification with nonmetric distance: image retrieval and class representation [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2000, 22(6): 583-600

[4] Yang Y. An evaluation of statistical approaches to text categorization[J]. Information Retrieval, 1999, 1(1): 76-88

[5] 余小鹏，马费成. 一种基于k-最近邻的无监督文本分类算法[J]. 情报学报，2008，27(4)：550-555

[6] Vries A D, Mamoulis N, Nes N, et al. Efficient KNN search on vertically decomposed data[C]// Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, Madison, Wisconsin. Madison: ACM Press, 2002: 322-333

[7] 王煜，王正欧，白石. 用于文本分类的改进KNN算法[J]. 中文信息学报，2007，21(3)：76-81

[8] 印鉴，谭焕云. 基于 χ^2 统计量的KNN文本分类算法[J]. 小型微型计算机系统，2007，28(6)：1094-1097

[9] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]// Proceedings of ICML-97, 14th International Conference on Machine Learning (Nashville, US). 1997: 412-420

[10] 李刚，夏晨曦，郑重. 局部文本特征选取算法的比较和研究[J]. 情报学报，2008，27(4)：506-511

[11] 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17(9)：1848-1859

(上接第139页)

[5] 曾庆田，吴哲辉，马炳先. Petri网的进程表达式与语言表达式[J]. 小型微型计算机系统，2004，25(4)：654-658

[6] Garg V K, Ragunath M T. Concurrent Regular Expressions and Their Relationship to Petri Nets[J]. Theoretical Computer Science, 1992, 96(2): 258-304

[7] 张继军，吴哲辉，董卫. Petri网的状态转换图[J]. 小型微型计算机系统，2008，29(9)：1714-1718

[8] Murata T. Petri nets: Properties Analysis and Applications[J]. Proc. of The IEEE, 1989, 77(4)

[9] 张继军，吴哲辉. 下推自动机的状态转换图与下推自动机的化简[J]. 计算机科学，2006，33(3)：271-274

[10] 张继军，吴哲辉. 有界Petri网的最小化化简[J]. 计算机科学，2007，34(11)：26-28

[11] 吴哲辉. Petri网导论[M]. 北京：机械工业出版社，2006

[12] 吴哲辉. Pumping引理的Petri网描述——Petri网语言属型的一组判定条件[J]. 计算机学报，1994，17(11)：853-858