基于标题类别语义识别的文本分类算法研究

王 强　关 毅　王晓龙
(哈尔滨工业大学计算机科学与技术学院　哈尔滨　150001)

摘　要： 本文提出了一种基于标题类别语义识别的文本分类算法。算法利用基于类别信息的特征选择策略构造分类的特征空间，通过识别文本标题中的特征词的类别语义来预测文本的候选类别，最后在候选类别空间中用分类器执行分类操作。实验表明该算法在有效降低分类候选数目的基础上可显著提高文本分类的精度，通过对类别空间表示效率指标的验证，进一步表明该算法有效地提高了文本表示空间的性能。

关键词： 标题类别语义识别；候选类别；类别空间表示效率
中图分类号：TP391　　**文献标识码：**A　　**文章编号：**1009-5896(2007)12-2885-06

Applying Title Category Semantic Recognition for Text Categorization

Wang Qiang　Guan Yi　Wang Xiao-long
(*School of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150001, *China*)

Abstract: This paper presents a new algorithm using title category semantic recognition for text categorization. The algorithm generates feature space based on its category, picks up category semantic words of the title to produce candidate category and finally classifies it under these candidate categories. The experimental results firmly prove that the new algorithm performs better with fewer candidates and higher precision. Further research introduces category space representation efficiency to verify the validity of the new algorithm and proves that it can achieve great improvement in text representation.

Key words: Title Category Semantic Recognition (TCSR); Candidate category; Category space representation efficiency

1　引言

自动文本分类是将自然语言文本自动指定至一个或几个预定义的文本类别中的方法，它是处理和组织大量文本数据的一项关键技术。自动文本分类技术利用训练样本进行特征选择和分类器训练，并根据选择的特征对测试样本进行形式化表示，然后输入到训练好的分类器中进行类别判定。目前的文本分类研究主要围绕着提高文本表示质量和开发高性能分类器这两个方面展开。本文的研究以向量空间模型(Vector Space Model, VSM)为基础，着力提高文本向量的表示方法。传统的VSM是以基于词的特征选择算法进行文本降维，利用词袋法(Bag Of Word, BOW)表示文本，以相似性计算或判别函数来最终确定文本的类别。但在VSM中，传统的特征选择算法，如信息增益(Information Gain, IG)与卡方分布(χ^2 - test , CHI)[1]，是在整个文本空间对特征项进行评价，而BOW方法也是将一篇文本中出现的特征向整个特征空间中映射，它忽略了文本特征的类别属性对分类的影响以及文本中的关键语句对判别文本类别的指示作用。目前，针对BOW方法的相关改进研究受到人们的广泛重视。在信息检索的特征权重计算研究中，Jin[2]引入了监督学习的思想，提出了一种应用类别信息优化特征词权重的方法，实验证明类别信息有效改善了文本特征的权重计算精度，提高了信息检索的效果，但Jin的研究主要应用了文本的类别信息，并未涉及对特征类别信息的探讨；Youngjoong Ko[3]提出了一种使用文本中重要性高的句子进行文本分类的方法，他利用自动文摘的相关技术对句子进行评价，使用重要性较高的句子进行文本分类模型的训练，实验表明这种方法可显著提高分类精度；李伟[4]提出了一种利用标题词、高频词和专有名词计算句子权重的关键句矢量模型，用于提高文本相似度的计算精度。但这两种方法在计算文本的关键句时都使用了较复杂的句子评价算法，会增加文本分类的代价。战学刚[5]及林鸿飞[6]关注标题与文章主题的密切关系，分别引入语义分析技术及基于示例的潜在语义分析技术构造标题分类函数来指导分类，但这两种方法使用时需引入相应的专业领域词表，并且分类易受语义分析结果的影响。

本文提出了一种应用标题类别语义识别技术(Title Category Semantic Recognition, TCSR)进行文本分类的算法。算法中采用基于类别的文本特征选择策略，将生成的文本特征按文本类别进行划分；并改进了传统的 BOW 文本表示方法，引入标题类别语义识别方法，即通过识别文本标题

2006-04-17 收到，2006-09-26 改回
国家自然科学基金(60435020，60504021)资助课题

中的特征的类别语义，预测该文本的候选类别，在此基础上执行最终的分类操作。实验表明本文提出的算法能有效提高文本表示的性能，同时可在降低分类候选数目的基础上提高文本分类精度，具有较好的性能。

2 文本标题对文本类别主题的指向分析

标题对文本类别主题的指向作用, 是指从读者的角度看标题与正文之间的关系，即标题对正文的类别主题具有预示性[7]。这是因为标题和正文是一个有机的整体, 标题统领整篇文章，是正文内容的最高概括和抽象, 正文内容是对标题的解构与阐释。

篇章语言学家们认为每个句子都是一个“话题-述题”结构，如果把这种认识进一步拓展，把整个文本看作是一个“话题-述题”结构，那么文本的标题就是整个篇章的话题，而文本内容则是述题。话题是叙述的出发点，通常代表被描述的事物或现象；述题是叙述的核心，通常表示对事物或现象的描写、叙述。标题作为文章的话题，引导读者以某种特定的方式解读文章，左右读者对整个文章的理解。

文本的标题与文本的类别主题有密切关系，其对文本类别主题的指向作用大致表现为以下几种情况：(1)标题直接揭示类别主题。如《反对自由主义》、《椭圆曲线密码学》、《鱼雷战》等，这一类标题直接反映了正文要论述的主要内容，通过标题中的类别主题词“自由主义”、“椭圆”,“曲线”,“鱼雷战”等可直接确定文本的类别主题；(2)标题揭示类别主题的范围、性质。如《中国雷达五十年》、《论油气分布的有序性》、《糖尿病患者运动的好处》等，这一类标题对正文所要论述内容的性质、范围进行了限定，限定成份中的中心词如“雷达”、“油气”、“糖尿病”等也具有对类别主题明确的指向意义；(3)标题表明类别主题的线索。如《中国经济学百年回顾》、《新世纪报告文学的走势》、《胶片和前卫电影》等，这一类标题，单从标题并不能了解正文所要论述的具体内容，但这并不妨碍对标题文本类别主题的指向作用，“经济学”、“报告文学”、“电影”都具有明确的类别主题意义。

实际上，绝大多数文本内容也确实是围绕标题展开的，是对标题的进一步阐述，这在科技论文中表现尤为明显。文献[8]中对国内中文期刊进行的抽样统计表明，自然科学论文的标题反映文章主题的概率高达 99%；而我们对 3,600 篇新闻类文章的统计表明，标题可预示文章类别主题的概率也可达到 95%以上。这说明文本标题对文本类别主题具有重要的指向作用，大多数文本的类别可以通过其标题进行准确的判定。

3 TCSR 分类算法

TCSR 分类算法是通过识别文本标题中的特征词，利用特征词所属的语义类信息预测文本的主题类别，然后通过训练好的分类器对预测类别进行类别确认的一种算法。它既不需要使用专业词典，也不需要复杂的语义分析过程，并且具有较好的分类性能。实现该算法的前提条件有两个：(1)基于类别的特征选择算法；(2)文本标题的识别算法。

3.1 基于类别信息的特征选择算法(Category information Analysis based Feature Selection, CAFS)

在文本分类系统中，由词、字串或概念来表达的特征集维数一般都很高，文本特征选择的目的就是去除那些不能表示信息或表示信息较弱的特征，以提高分类准确度，减少计算复杂度。

定义1 特征空间与特征向量 文本的特征是用于描述文本模式的一组抽象的元素，可记作$T=\{t_1, t_2, \cdots, t_{n-1}, t_n\}$，将$T$ 中每个特征看成一个维度， 这样T 就构成了一个特征空间，该特征空间中的每个点代表一个样本，由特征向量$\boldsymbol{X}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_N)$来表示。

定义2 特征评分函数 特征评分函数是衡量特征重要性的量度，可形式化表示为一个映射：$F_s:T\rightarrow S_H$，其中T为特征空间，S_H为特征值集合，F_s为特征评分函数。F_s函数一般从分散度和集中度两个方面对特征进行评估，分散度描述了特征在某类内部的分布情况，集中度则描述特征出现在不同类别之间的差异性。一个具有分类意义的特征在某个类别中应尽量分布均匀，同时它应该集中出现在某一类或某几类文本中，而不是平均分布在各个类别中。在本文的特征选择算法中，引入特征的类别贡献函数Sw_{ij}及方差机制来衡量特征的重要程度。Sw_{ij}的定义如下：(设特征为 $w_i \in T, i=1, \cdots, n$ ， 类别为 $j, j=1,\cdots,m$)

$$Sw_{ij}=fw_{ij}\cdot\log(dw_{ij}+1.0)\Big/\sqrt{\sum_{i=1}^{n}[fw_{ij}\cdot\log(dw_{ij}+1.0)]^2} \quad (1)$$

Sw_{ij}正比于w_i在C_j中出现的词频数，正比于w_i在C_j中分布的均匀度，用于衡量特征对类别的重要性。其中，$fw_{ij}=T_{ij}/L_j$, T_{ij}是w_i在类别C_j中的出现的词频数，L_j是类别C_j中所有词出现的总次数；$dw_{ij}=d_{ij}/D_j$，d_j是类别C_j中出现w_i的文档数，D_j是类别C_j中的文档个数。在此基础上的特征评分函数$\mathrm{Im}p(w_i)$定义如下：

$$\mathrm{Im}\,p(w_i)=\sqrt{\sum\nolimits_j(Sw_{ij}-\overline{S}_i)^2\Big/\sum\nolimits_j Sw_{ij}} \quad (2)$$

$\mathrm{Im}p(w_i)$通过计算Sw_{ij}的方差来评价w_i的重要性，$\mathrm{Im}p(w_i)$越大，表明w_i在不同类之间的贡献差异性越大，其分类价值也就越大。公式中，$\overline{S}_i=\sum_j Sw_{ij}/m$ 。

定义3 基于类别的特征代表子集 特征代表子集是由经$\mathrm{Im}p(w_i)$函数筛选出的特征构成的集合，为实现对标题语义类别的应用，还需要对特征所属的类别进行划分，这可通过Sw_{ij}来实现。定义基于类别的特征代表子集为$H_C=\{(t_i,j)\mid t_i\in T\wedge \mathrm{Im}p(w_i)>\varphi\wedge j=\mathrm{Cid}(\max(Sw_{ij}))\}$，其中$\varphi$ 为特征阈值，$\mathrm{Cid}(\max(Sw_{ij}))$ 表示w_i具有最大贡献量的类别的类别号，集合H_C中的元素称为代表元素。

CAFS 算法可有效提取文本特征，并且可以根据特征词所属的类别语义对其进行细致划分，使得文本特征的表示功能得到了进一步扩展。

3.2 文本标题识别算法(Title Recognition Algorithm)

应用文本标题中的特征的类别语义指导文本分类操作，就要实现对文本标题的正确识别。文本标题从语言特点上来说具有一定的标记性[9]，即可以通过文本标题中的词语、结构、形式或位置等信息进行识别。标题的词语标记指那些通常只出现在标题中而几乎不出现在正文中的一些词语，结构标记是指整个结构一般只在标题中使用的一些句法格式，而形式标记则指标点符号在标题中的特殊用法。本文基于以上特征，通过构造一个标题识别知识库，基于有限状态自动机原理实现了一个通用的文本标题识别算法，描述如下：定义 $M=(Q,T,\delta,q_0,F)$，其中 $Q=\{q_0,q_1,q_2,q_3,q_4,e\}$；$T=\{0,1\}$，1/0 表示识别算法判断相应状态成功或失败；$F=\{e\}$。

状态转换函数 $\delta: Q\times T\to Q$ 定义为：$\delta(q_0,1)=q_1$，$\delta(q_1,1)=q_2$，$\delta(q_1,0)=e$，$\delta(q_2,1)=q_4$，$\delta(q_2,0)=q_3$，$\delta(q_3,1)=e$，$\delta(q_3,0)=q_4$，$\delta(q_4,1)=e$，$\delta(q_4,0)=e$。

图 1 中描绘了文本标题识别的动态行为框架。系统应用词语标记、结构标记及形式标记等多种策略来实现标题识别。状态转换模式表示了识别过程和系统根据实际情况采取的相应处理措施。其中各个状态的含义为:e为结束状态；q_0为初始状态；q_1为标题识别状态；q_2为词语标记检测状态；q_3为结构标记检测状态；q_4为形式标记检测状态。

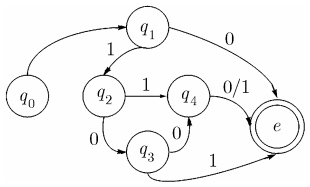

图 1　文本标题识别自动机 M 状态转换图

标题识别算法以每一行文本为处理单元，当开始工作时，自动机M 从q_0状态转换为q_1状态。在q_1状态,当发现该文本的标题已经识别时，M直接转换到e结束状态，若文本标题尚未识别，则M 转换为q_2状态；在q_2状态，M检测处理对象中是否包含文本标题特有的词语标记，如纪行、拾零、随笔、面面观、纪要、访谈录等标记性词语是否出现，是的，话系统就进入q_3状态， 否则M 转换到q_4状态；在q_3 状态, M检测处理对象是否满足文本标题的结构标记形式，如从X说(谈、想)开去(起)，X话(道)Y，X看Y等，满足结构标记M就输出 1 进入e结束状态，表示当前文本内容为文本标题，否则M转换到q_4状态；在q_4状态, M检测处理对象是否满足文本标题的形式标记，如中文标题中一般不出现逗号、句号等标点，英文标题要求全部使用大写字母或实词的首字母大写等特征，如果条件满足，则M输出 1 进入e结束状态，当前文本内容为文本标题，否则输出 0 进入e结束状态，表示当前文本非文本标题。

该标题识别算法对中、英文进行了一体化处理，以标题的形式标记为首要识别标志，并且进一步引入了结构标记与词语标记来完善识别过程，获得了较准确的标题识别率。

3.3 基于标题类别语义识别的文本分类算法

TCSR 算法的处理流程是：首先对标题中出现的类别特征词进行识别，从而确定文本的候选类别，并进一步通过分类器进行类别确认；当标题分析无法判定类别时，则回退到传统的 BOW 方法进行文本表示和分类。这个分类算法的模型可用一个四元组 $M=<D,C,R,T>$ 来表示，其中 D 表示文本，C 表示类别，R 是标题类别语义分析函数，T 是分类器函数。用函数映射关系可表示为：$(T\cdot R): D\to C$ 或 $C=(T\cdot R)(D)$。函数 R 定义如下：

$$R=\begin{cases} R_T(D) & \forall c_j \quad W_T\cap F_{c_j}\neq\phi \\ R_B(D) & \forall c_j \quad W_T\cap F_{c_j}=\phi \end{cases} \tag{3}$$

公式(3)中 F_{c_j} 表示含有类别属性C_j的所有特征组成的集合，W_T 表示文本标题中的特征词集合，$R_T(D)$ 和 $R_B(D)$ 分别表示应用标题语义类识别与回退到传统BOW的文本表示方法。经过 $R_T(D)$ 表示的文本，可有效降低文本中噪声特征对分类结果的影响，并且分类器函数T只在候选类别范围内执行分类即可，因此效率较高；而经 $R_B(D)$ 表示的文本，T函数将在所有类别范围内执行分类操作。以下通过一个实例来说明 $R_T(D)$ 与 $R_B(D)$ 之间的差异。

例　中华人民共和国 中国 人民 银行法 总则

为了 确立 中国 人民 银行 的 地位 和 职责，保证 国家 货币 政策 的 正确 制定 和 执行，建立 和 完善 中央 银行 宏观 调控 体系，加强 对 金融业 的 监督 管理，制定 本法。中国 人民 银行 是 中华人民共和国 的 中央 银行。中国 人民 银行 在 国务院 领导 下，制定 和 实施 货币 政策，对 金融业 实施 监督 管理。货币 政策 目标 是 保持 货币 币值 的 稳定，并 以此 促进 经济 增长。

上文属于政治、法律类文本，由于文本中存在着大量的如银行、货币、金融业等经济方面的特征词，导致使用传统 BOW 方法表示文本时引入了大量噪声特征，文本容易被错分到经济类别中。应用标题语义识别则可改善分类情况。

表 1　示例文本经 $R_T(D)$ 与 $R_B(D)$ 的词频向量表示情况对比

特征空间		$TF(R_B(D))$	$TF(R_T(D))$
政治、法律	银行法	1	1
	国家	1	1
	政策	3	3
	国务院	1	1
经济	银行	5	0
	货币	4	0
	金融业	2	0
	币值	1	0
	经济	1	0

从表 1 可以看出，应用 $R_B(D)$ 方法，会在文本向量中引入诸如“银行”、“货币”等经济类的特征词，导致文本容易被错分；而应用 $R_T(D)$ 方法后，由于示例文本中的标题类别特征词“银行法”的语义类别为政治、法律，其类别指向作用会把与经济类别相关的特征词排除在文本向量之外，生成了一个更为简捷有效的文本特征空间，有效去除了噪声特征对分类器的干扰，提高了分类精度。

模型中的分类器T函数可以为任何一种文本分类器，本文采用经过Sigmoid[10]后处理的SVM线性分类器。SVM是近年来在统计学习理论(Statistical Learning Theory，SLT)的VC维理论和结构风险最小原理基础上发展起来的一种通用学习方法。它可以根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度(accuracy)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷，获得最好的推广能力(generalization ability)。但由于SVM是通过文本向量与最优分类面之间的距离来判定文本类别，导致分类器在多类别分类问题中不能在全局输出最优判别，因此需要引入Sigmoid函数来对SVM的输出进行概率转换。Sigmoid函数中的参数可在训练集中通过交叉验证(m-fold cross-validation)的方式获得。实验结果表明，Sigmoid函数的引入使得SVM分类器在多分类问题中可选择更为合理的分类结果。

4 实验结果与分析

本文在中英文数据集上对TCSR算法进行了测试。中文测试应用中图法分类体系[1)]，训练集采用 2003 年 863 文本分类评测的 3,600 篇文本，测试集为 2004 年 863 文本分类评测的 3,600 篇文本。英文分类测试采用Reuters-21578 数据集[2)]，并且去除其中TOPIC为空或者<BODY></BODY>标签不包含文本内容的所有文本，其中训练集包含 6,574 篇文本，测试集包含 2,315 篇文本。实验包括 3 个部分：(1)以传统的特征选择算法如IG与CHI为参照对象，比较CAFS算法在中、英文数据集上的分类性能，验证CAFS算法的有效性，在此基础上，分别使用CAFS＋BOW与CAFS+TCSR算法执行训练和分类，比较两种算法的分类性能和运行效率，验证TCSR算法的有效性；(2)测试TCSR算法对文本主题类别的覆盖精度以及在保证分类精度的前提下，使分类器使用频率降低的情况；(3)应用类别空间表示效率指标，验证使用标题类别语义识别后文本向量表示效率的提高情况，并分析算法效率改进的原因。

4.1 实验设置

实验中，中文文本的分词使用实验室的ICSU系统完成，利用CAFS算法提取特征，φ 值设为 0.65，使用一个变型的Okapi权值计算公式计算特征词权重[11]，分类器使用一对多策略的线性核函数SVM算法，其结果经过Sigmoid方法转换为概率值。

本文引入类别空间表示效率指标(Ceff)来验证算法的有效性。根据 Fisher 判别准则，当文本类别间的距离尽量大，文本类别内的距离尽量小的时候，可获得最佳的分类效果。类别间与类别内的距离分别用离散性(Discreteness)和内聚性(Cohesion)两个指标描述，定义 Ceff 为内聚性与离散性的比值，公式如下：

$$\boldsymbol{C}_k = \frac{1}{|L_k|}\sum_{\boldsymbol{d}\in L_k}\boldsymbol{d} \tag{4}$$

$$\mathrm{Co}_k = \frac{1}{|L_k|}\sum_{\boldsymbol{d}\in L_k}\boldsymbol{d}\boldsymbol{C}_k \tag{5}$$

$$\mathrm{Dis}_k = \frac{1}{|C|}\sum_{i=1}^{|C|}\boldsymbol{C}_i\boldsymbol{C}_k \tag{6}$$

$$\mathrm{Ceff}_k = \mathrm{Co}_k / \mathrm{Dis}_k \tag{7}$$

其中 L_k 表示训练文档集中的第k个类别，$\boldsymbol{C}_k$ 表示第k个类别的中心向量，Co_k 和 Dis_k 分别表示类别k的类别内聚性与类间离散性，向量间的相似度应用向量余弦距离计算。类别的内聚性 Co_k 越大(相似度高)，类间离散性 Dis_k 越小(相似度低)，则Ceff_k越大，则可获得更好的分类性能。

实验中中文文档用 863 文本分类评测标准进行评测[3)]，即设定每一篇文本至多可以分到两个类别中去，评测只以分类器第一候选输出的类别为准。英文文本评测采用Reuters-21578 提供的评测软件[4)]，允许multi-label输出。评测中均采用宏、微平均(macro & micro average)策略，利用准确率(P)，召回率(R)及$F1$ 值进行对比实验。

4.2 结果分析

图 2 显示的是在中、英文数据集上随着特征数目的变化，CAFS 与 IG，CHI 算法的分类性能对比情况。

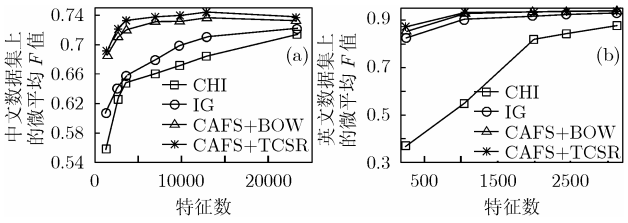

图 2 中英文数据集上 CAFS 算法与 IG 及 GHI 算法随特征数变化的分类性能比较

从图中可以看出，随着特征数目的变化，CAFS算法保持了相当稳定的分类性能，当特征的数目比较多的时候，其分类效果与IG，CHI相当，而当特征数目比较少时，在中文数据集上其MicroF值比IG与CHI分别高出约 9 和 14 个百分点(如图 2(a))，而在英文数据集上则高出 2.6 和 48 个百分点

1) 第四版，二级以上类目录，由于“T 工业技术”和“Z 综合性图书”这两类难以判定主题类别，因此不予考虑，故实际总的分类数为 36 类。

2) 数据集用ModApte方式切分，取前 10 个包含文档数最多的类别。

3) http://www.863data.org.cn/cursyllabus.php

4) http://www.lins.fju.edu.tw/~tseng/Collections/Reuters-21578.html

(如图 2(b))，CAFS算法表现出了更优的分类效果。同时，CAFS+BOW分类方法的系统在 2004 年 863 中文文本分类评测中获得了第一名，其在Reuters-21578 数据集上的MicroF与MacroF值分别达到 0.935 与 0.867，也获得了与目前最好的分类结果(Franca[12], 2005)相当的分类性能，这充分证明了CAFS算法获取文本特征的有效性。另外，图 2 也表明在CAFS基础上引入TCSR算法后，在中英文数据集其分类性能(-*-型线所示)分别提高了 1−2 个百分点，证实了通过识别标题类别语义获取文本主题类别算法的合理性。

图 3 进一步比较 CAFS+TCSR 与 CAFS+BOW 算法在中英文数据集的各个类别上的 MacroF 值的情况。图中的一个点表示相应分类体系下的一个类别，其横、纵坐标分别为在不同算法下(CAFS+TCSR 和 CAFS+BOW 算法)该类别的 MacroF 值。如图 3(a)，在中图法数据集上，TB 出现在对角线的上方，其在 TCSR 和 BOW 算法下的 MacroF 值分类为 0.173(横坐标)与 0.357(纵坐标)，说明针对 TB 类，TCSR 方法的性能优于 BOW 方法；图 3(b)中，在 Reuters-21578 数据集上，Corn 点出现在对角线下方，其 MacroF 在两种表示算法下的值分别为 0.79 与 0.667，表示针对 Corn 类，BOW 算法优于 TCSR。可见，通过比较出现在对角线上方与下方的类别数目，就可以看出两种文本表示算法在不同类别上的分类性能。图 3 表明，在中英文数据集上，TCSR 算法在多数的类别上都取得了较 BOW 方法更为理想的分类效果，其对总体分类性能的改善效果是比较稳定的。

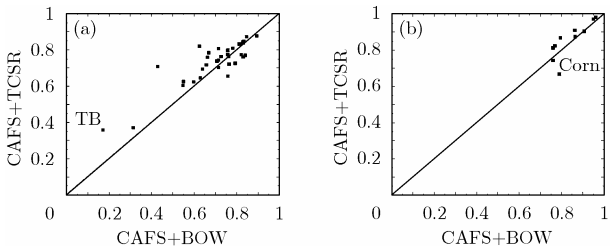

图 3　中英文数据集上应用 TCSR 算法前后分类性能比较

表 2 统计了 TCSR 算法的文本主题类别覆盖精度，并分析了 TCSR 算法对分类速度的影响。

结果表明，在中英文数据集上，TCSR 算法判断分类器与文本主题相关的正确率达到 90.2%与 98.4%，大大超过了分类器的实际分类准确率，表明该算法可以正确判断文本的主题类别，算法是有效的。同时，表 2 显示 TCSR 算法会对分类速度产生较大的影响。在传统的一对多 SVM 分类模型中，每处理一个文本都要调用所有的分类器进行分类判别，其分类器的使用频率为类别数与文本数目的乘积；而在 TCSR 方法中，受标题类别语义限制，只有与标题类别语义相关的分类器才真正参与到分类决策中，使分类器的使用频率大幅下降，分类速度得到提高，而且这种提高的趋势随着类别数目增加而变得更加明显，如在中图分类数据集上，分类器使用频率的下降幅度高达 73%，在 Reuters-21578 数据集上，分类器使用频率的下降幅度也可达 37%。

分析 TCSR 算法提高文本分类性能的原因，主要是算法改善了文本特征空间的表示效率。本文通过 Ceff 指标来验证这一结论。图 4 是在中、英文数据集上应用 TCSR 算法前后的每一个类别 Ceff 情况对比。

图 4 显示应用基于标题类别语义识别算法后，在所有的中、英文类别上，文本的类别内聚性都大幅提高，而文本类别间的离散性(相似度)则降低，即类别的 Ceff 值明显增大，类别特征空间的表示效率得到提高。文本表示性能的改进可以带来分类性能的提高，而实验结果也证明了这一点。

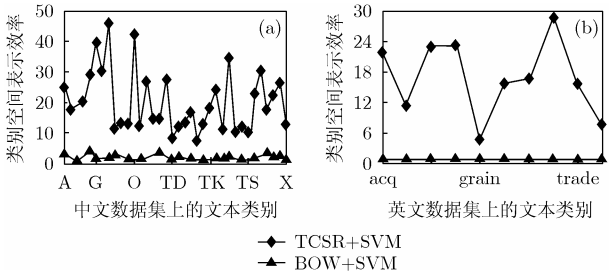

图 4　中英文数据集上应用 TCSR 算法前后类别 CeffX 比较

5　结束语

TCSR 算法能够有效地处理文本分类问题，该算法通过基于类别信息的特征选择算法，将文本特征进行类别语义划分，通过识别文本标题中的特征词的类别语义来构造候选类别集，有效指导分类操作，提高文本分类的精度。此外，该算法具有较好的扩展性，通过基于错误实例的反馈学习或手工添加类别语义词的方法，可进一步提高文本分类的精度。

下一步的工作，一是在基于类别的特征选择算法中加入专业领域词(主要是新词)识别的算法，以提高文本特征类别语义的质量；二是研究在保证分类性能与分类效率的前提下，进一步提高相关分类器判别正确率的方法。

表 2　中英文测试集上基于标题类别语义识别算法与传统分类算法分类速度对比

	中图分类法数据集			Reuters-21578 数据集		
	CAFS+TCSR	CAFS+BOW	下降幅度	CAFS+TCSR	CAFS+ BOW	下降幅度
分类器用次数	34，957	129，600	73.03%	14，499	23，150	37.37%
	TCSR 判断正确	TCSR 判断错误	正确率	TCSR 判断正确	TCSR 判断错误	正确率
文本数	2770	302	90.2%	1257	21	98.4%

参 考 文 献

[1] Yiming Yang and Jan O P. A comparative study on feature selection in text categorization. In Proceedings of the 14th International Conference on Machine Learning (ICML97), San Francisco, USA, 1997: 412–420.

[2] Rong Jin, Joyce Y C, and Luo Si. Learn to weight terms in information retrieval using category information. In Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005: 353–360.

[3] Young joong Ko,Park Jinwoo, and Seo Jungyun. Automatic text categorization using the importance of sentences. In Proceedings of the 19th International Conference on Computational Linguistics, Taipei, Taiwan, 2002: 474–480.

[4] Li Wei, Yuan Chunfa, Wong Kam-Fai, and Li Wenjie. Text similarity calculating based on critical sentence vector model. In Proceedings of the 20th International Conference on Computer Processing of Oriental Languages (ICCPOL2003), Shenyang, China, 2003: 424–430.

[5] Zhan Xuegang, Yao Tianshun. The classification method for Chinese document title based on Chinese semantic analysis. In Proc of the Int'l Conf Chinese Information Processing, Beijing, Tsinghua University Press, 1998, 321–324.

[6] 林鸿飞. 基于示例的文本标题分类机制. 计算机研究与发展, 2001, 38(9): 1132–1136.
Lin Hong-Fei. The mechanism of text title classification based on examples. *Journal of Computer Research & Development*, 2001, 38(9):1132–1136.

[7] 张加民. 标题预示性的元功能视角. 外语教学, 2004, 25(6): 36–39.
Zhang Jia-ming. The meta-function research on title's prediction. *Foreign Language Education*, 2004, 25(6): 36–39.

[8] 麻志毅，姚天顺. 基于情境的文本主题求解. 计算机研究与发展, 1998, 35(4): 344–348.
Ma Zhi-yi, Yao Tian-shun. Calculating texts' topics based on situations. *Journal of Computer Research & Development*, 1998, 35(4): 344–348.

[9] 刘云. 论篇名语言的标记性. 云梦学刊, 2003, 4: 104-107.
Liu Yun. On the markedness of title language. *Journal of Yun Meng*, 2003, 4: 104–107.

[10] John C P. Probabilistic outputs for support vector machines and comparisons to regularized likelihood, methods. Advances in Large Margin Classifiers, 1999: 61–73.

[11] Tom Ault and Yang Yiming. KNN at TREC-9. In Proceedings of the Ninth Text REtrieval Conference (TREC-9). Maryland, USA, 1999: 127–134.

[12] Franca Debole and Fabrizio Sebastiani. A analysis of the relative hardness of Reuters-21578 subsets: research articles. *Journal of the American Society for Information Science and Technology*, 2005, 56(6): 584–596

王 强： 男，1973 年生，博士生，研究方向包括文本挖掘、机器学习算法等.

关 毅： 男，1970 年生，硕士生导师，副教授，主要研究领域包括问答系统、统计语言处理及文本挖掘等.

王晓龙： 男，1955 年生，博士生导师，教授，代表性研究成果是语句级智能汉字输入技术.