● 王 昊[1,2]，邓三鸿[1,2]，苏新宁[1,2]
（1. 南京大学 信息管理学院，江苏 南京 210023；2. 南京大学 江苏省数据工程与知识服务重点实验室，江苏 南京 210023）

中文短文本自动分类中的汉字特征优化研究*

摘 要：采用含语义的词语或篇幅更长的语言片段作为中文短文本的特征描述存在明显的可操作性问题。文章综合探讨了汉字特征在中文短文本分类计算中的可行性以及影响规律，比较了关键词、词语和汉字的类目区分能力，认为后者的分类效果略低于篇幅大的语言片段，但其具有可计算性强和文本覆盖率高的优点；基于类现频次和信息增益复合方法对汉字特征进行了筛选，总结了汉字特征数量减少对分类效果的影响规律；分析了不同特征权重设置对汉字特征分类效果的影响及其原因，认为汉字在词语中的位置参数及其频次参数的有效结合可以在一定程度上提高汉字特征的分类效果。

关键词：短文本；文本分类；汉字特征；自动分类；优化

Abstract: It has significant operability problems that using words with clear meanings or language fragments much longer as descriptive features for Chinese short-text. This paper comprehensively discussesfeasibility and influence rules of character features applied to classification calculation for Chinese short-texts, compares the category distinguishing ability of keywords, terms and Chinese character. The paper indicates that character features, which performance slightly worse than longer language fragment, have the advantage of stronger calculability and higher text coverage. The paper screens character features by a method with composition of occurrence frequency in categories and information gain, and summarizes the influence rules of the decrease in the number of Chinese character features for classification effect. Based on the analysis of the impact on classification effectiveness of character features with different setting schemes for features weight and its reasons, the paper believes that the approach of Chinese character position in words effectively integrated frequencies factors could improve classification performance of Chinese character features to some extent.

Keywords: short-text; text classification; Chinese character features; automatic classification; optimization

文本分类（Text-Classification，TC）是指利用自然语言处理技术，由计算机自动标记文本类目的方法和过程[1]。TC的实质是预测文本离散化属性的取值，其应用包括图书论文等资源类目的生成[2-3]、网络评论情感等的划分[4-5]、临床医疗的分级[6-7]等，均可采用TC方式进行建模处理。中文文本分类（Chinese TC，CTC）则以汉语文本作为处理对象，由于语言差异，直接采用字符语言处理技术实现CTC存在诸多问题。

传统的TC方法将类目和待分类文本均表示为向量形式，进而计算两者之间的相似度，其中相似度最大的类目即为文本的所属类目[8-9]。由于类目的特征向量构建复杂且可操作性弱，该方法很快被机器学习（Machine Learning，ML）所取代，即采用ML模型对已分类文本的特征分布进行学习，然后将获得的分类器应用于待分类文本以自动生成其类目。ML方法中最重要的是构建文本×特征矩阵（Text-Feature Matrix，TFM），即将所有文本用一个特征集合进行统一描述。在英文文本中，一般以单词（Word）作为分类属性，常用Word数量不会太多，是文本特征的理想选择；但是在CT特别是篇幅较短的CT中却很难确定理想属性，词汇特征数量巨大，可操作性差，汉字特征又被主观认为缺乏语义，分类效果欠佳。

本文探讨以汉字作为中文短文本分类特征的可行性以及可改进的策略，重点在于描绘出汉字特征在进行特征选择和权重设置时对分类效果的影响规律，为特定领域分类计算选择合适的特征类型和优化方案提供可参考的理论依据和事实支持。

* 本文为国家社会科学基金重大招标项目“面向突发事件应急决策的快速响应情报体系研究”（项目编号：13&ZD174）和江苏省自然科学基金青年项目“面向专利预警的中文本体学习研究”（项目编号：BK20130587）的成果。

1 相关研究

基于 ML 的 CTC 方法主要包括 ML 算法研究和 TR（Text Representation，文本表示）方法优化等两部分。由于可靠数学模型的支持，ML 算法相对成熟，常用的有 NB[10]，ANN[11]，SVM[12]等，事实表明这些算法均具有很强的适应能力。于是，文本作为一种非结构化数据，其分类的难点多集中在了 TR 上，具体包括特征类型（Feature Type，FT）确定、特征集合（Features Set，FS）选择以及特征权重（Feature Weight，FW）设置等。①FT 确定是解决以什么作为特征项的问题。CT 最典型的特征来源是词语或短语[13-14]，这继承了英语中以非停用单词作为特征项的传统；基于语义角度，以更具内涵的关键/主题词或本体等作为文本特征能够更有效地表达文本的语义[15-17]；基于结构角度，构成汉语言片段的最小单元汉字也可作为文本特征的来源[18-19]，但由于汉字语义的不定性和繁杂性导致人们对其具有较强的不信任心理。②FS 选择解决遴选出最佳特征项的问题，目前多采用计算特征与类目之间关联度的方法，如信息增益（Information Gain，IG）[20]、基尼指数（Gini Index，GI）[21]、互信息（Mutual Information，MI）[22]以及 X2 统计（CHI）[23]等，过滤掉含分类信息量少或意义重复的特征项，达到消除冗余、降低特征维度、减少计算量的目的。③FW 设置则是解决特征项对文本重要程度度量的问题。TFM 是一个高维稀疏的信息空间，采用具有较强可操作性和较低计算复杂度的布尔型（Presence）权重[24]是最直接的思维；若文本篇幅较长或者特征项语义较含糊，可采用频次型（Frequency）或频率型权重[25]；进一步考虑特征项在各文本中的分布，则可采用 TF-IDF 值[26]来修正特征在文本中的多现性；此外，逆类目频次、MI[27]等能够描述类目和特征间普适关系的度量值也可作为特征权重因子。

然而上述研究在对大规模中文短文本进行实际操作中却出现了瓶颈。①主题词、本体等的获取需要相应语言资源的支持；中文关键词抽取的正确率较低，特别是在上下文特征较少的短文本环境下。②对文本分词可以得到粒度较小的词语特征，但是中文分词的不准确性使词语特征存在较大的语义缺失和噪声引入，而且词语较低的文本覆盖可能导致待分类短文本中不一定存在特征词语。③低粒度的汉字特征较易获取，可操作性强，但从理论上讲汉字语义不明确可能会造成分类效果的严重下降。④目前特征筛选研究中多忽略“特征的减少不可能提高分类效果”和“特征的减少将导致其文本覆盖范围变小”的事实，前者指出特征筛选的主要目的是降维以减少计算量，对分类效果多起消极作用；后者认为在短文本分类应用中需要容忍一定程度的特征冗余，以减少待分类短文本由于特征缺失而无法召回的现象[28]。⑤CTC 中，FW 多以 Frequency 为基础进行计算，甚少考虑特征在文中的位置因素（Position Factor，PF）的影响。

2 研究方法

2.1 研究框架

本文的研究框架如图 1 所示，从总体上可以分为数据预处理、汉字特征分布规律探测、汉字特征在 TC 中的表现及其 FS 选择对分类的影响规律和 FW 设置对分类的影响规律 5 个部分。①本文以 CSSCI 期刊论文的学科划分作为短文本分类的模拟环境。先从数据集中随机选择指定学科条目作为实验数据，再利用 ICTCLAS 分词软件和停用词表对论文题名进行分词并筛选，然后对获得的词语做切分以获得汉字特征，最后统计关键词、词语以及汉字等特征在文本中的出现次数，构建 < 文本，特征，频次 > 3 元组，同时抽取论文类目信息，完成实验数据的预处理。②图 1 中左上角所示，对不同类型特征的 3 元组进行数理统计，一方面比较关键词、词语和汉字等特征的分类计算量，以及经过特征初步筛选后数据分布的变化情况；另一方面分析汉字特征随文献量增长的变化趋势。综合两方面的内容，从理论上归纳汉字特征的优点。③从预处理数据中抽取出不同类型特征的 3 元组及类目数据，构建 TFM 和 TCM（Text-Category Matrix，文本—类目矩阵）进行 SVM 计算，比较不同类型特征的分类结果，分析汉字特征的表现及其可行性。④如图 1 右上角所示，在获得汉字特征 3 元组及文本类目的基础上，计算汉字特征的 IG 值，并据此筛选特征及相关数据，重新组建文本—汉字特征矩阵（Text-Character Matrix，TCFM）和 TCM 进行 SVM 计算，对不同 FS 的分类结果进行比较分析，探索汉字特征选择对分类效果的影响规律。⑤根据事先制定的权重设置方案对汉字特征值进行修正并进行 SVM 计算，分析分类结果总结 FW 设置对汉字特征分类效果的影响规律。

2.2 FS 选择的方法

FS 选择的主要目的是实现降维，排除分类贡献小或语义重复的特征项，从而降低分类计算量。一般认为，在不影响分类效果的基础上，特征的数量越少越好；然而，上述论断在 CTC 中却存在较大局限，因为 CTC 中的大量特征项均仅与部分样本存在关联，即每个特征项都仅仅是少数样本的有效属性，这也是造成 TFM 极其稀疏的主要原因。在这种情况下，对特征进行筛选就要比较慎重。一方面，过滤特征就意味着削弱文本的属性描述，文本之间的区别度降低了，因此可能带来分类效果的下降；另一方面，缩小 FS 能够降低计算复杂度，但在实际应用中，新

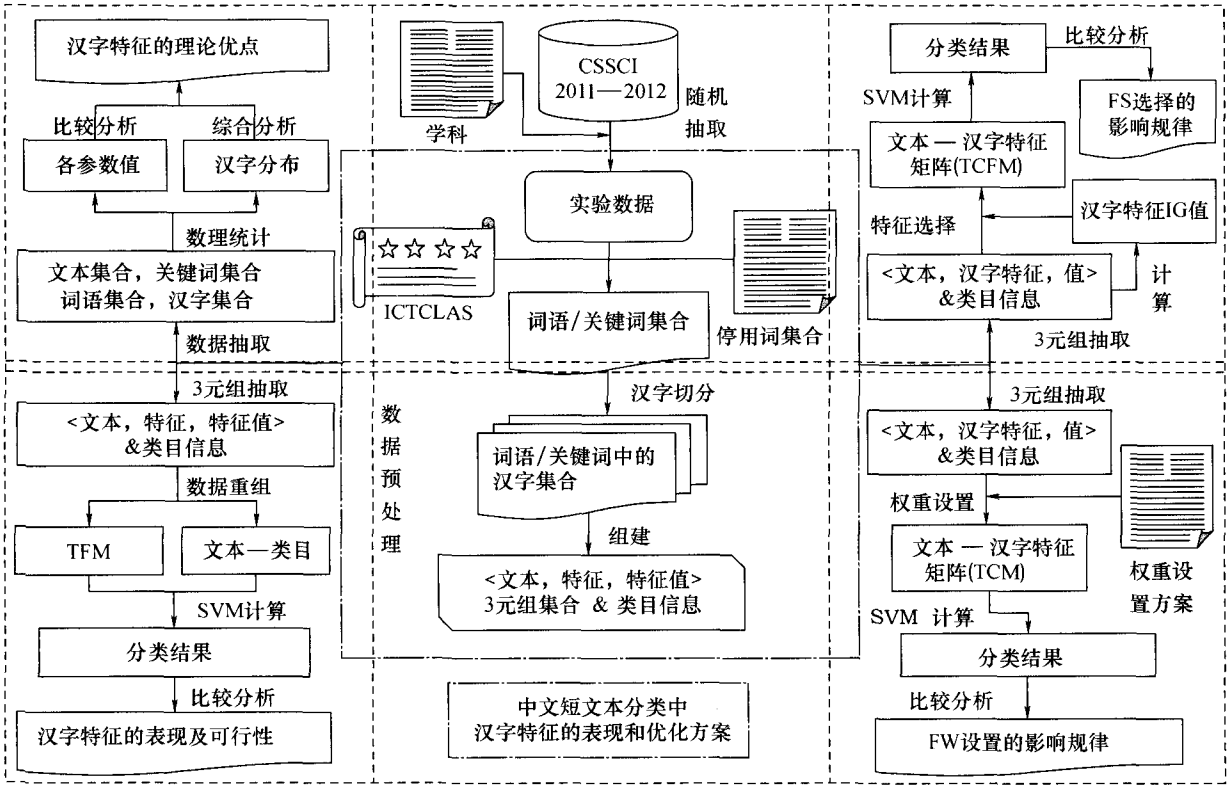

图1 中文短文本分类中汉字特征的表现和优化方案的研究思路

的待分类文本可能因为不包含特征项而无法归类，特别是短文本中。FS越小，这种文本覆盖不足的现象就越容易发生。因此在CTC中，有些冗余特征也有存在的意义，它们减少了待分类文本的特征向特征集合中的特征转换的计算。因此，笔者认为在短文本分类应用中，尽可能维持FS的规模具有更好的实用性。

基于事实理解和数学理论，笔者提出了类现频次（Frequency in Category，FIC）和IG相结合的复合筛选方法。面对中文文本空间中的庞大FS，直接采用IG方法不仅存在计算复杂度高的问题，而且不符合事实规律，因为大部分出现频率低的特征即便具有较高的IG值，也不适宜作为文本特征，这种现象在组词复杂的中文文本中普遍存在。因此，针对中文的特殊性，笔者认为可以先根据FIC过滤掉偶然出现的特征，之后再采用IG方法进一步优化。前者的依据是“至少在一个类目中存在N次及以上”，而后者的数学描述如图2所示。A为未分类状态，B为已分类状态，C则为经过特征T分类后的半分类状态，从状态A到B共需要注入I(B)的信息量；若A经过C到达B，而若在C的状态下出发到B需要I(B|C)的信息量，那么从A到C需要的信息量I(B)-I(B|C)，即为特征T的IG值。

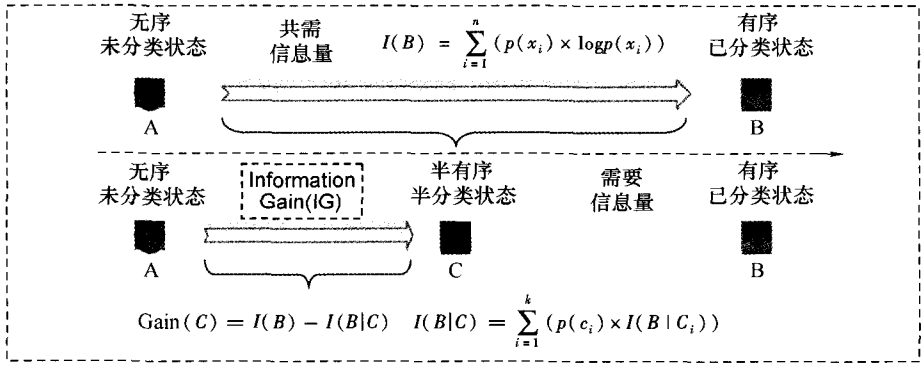

图2 IG算法的理论描述

2.3 FW设置的方法

FS筛选只能减少冗余特征对TC的干扰，不可能同时提高对象类目的识别度；因此在指定特征类型及不扩大特征数量的基础上，只有通过改变FW设置来提高分类正确率。前文对分类特征的权重设置进行了总结，认为目前主要采用以Frequency为基础衍生的数值，例如频率、TF-IDF、MI等，作为特征的权重。这是因为在当前CTC中，

基本上默认 FT 为词语，而词语在文本中表现最为明显的性质就是其出现频次。

其实词语在文本中的位置也能反映出词语特征的某种性质；位置参数对于语义不明确的汉字特征而言，其意义更为明显，即汉字在其所处词语中的位置往往能够强化其真实含义。因此，笔者引入汉字特征的位置参数，采用汉字特征的 Presence、Frequency、TF-IDF 以及在文本中所处的平均位置（Position）等 4 类权重，再加上 Frequency 与 Position 的混合（Mix_FP）和结合（Com_FP），以及 TF-IDF 与 Position 的混合（Mix_TP）和结合（Com_TP）共 8 种 FW 进行实验比较和分析，以明确各种类型权重对汉字特征类目区分能力的影响。

上述 FW 中，特征值的计算如下：①Presence 采用 0/1 标记，若特征在文本中出现记为 1，否则记为 0；②Frequency 记录特征在当前题名中出现的频次；③TF-IDF 描述了特征与所在文本之间的关系紧密程度，由特征 i 在当前文本 j 中的出现率 $TF_{i,j}$ 以及特征在整个样本空间中的普遍率 IDF_i 决定；④Position 以汉字特征在词语（可分词获得）中的位置平均值（多次出现）作为权重值；⑤Mix_FP 则将 F（requency）和 P（osition）的权重值归一化后以 $n:m$ 加权相加获得，笔者分别以 3∶7、5∶5 和 7∶3 的比例进行了权重复合；⑥Com_FP 则是将 Frequency 和 Position 的权重值分列，使得单一汉字特征被拆分为两个特征项；⑦Mix_TP 和 Com_TP，分别将⑤和⑥中的 F（requency）替换成 T（F-IDF）即可。

3 汉字特征的分布规律及其在 CTC 中的表现

3.1 汉字特征的统计和分布

本文随机选择了 CSSCI（2011—2012）中管理学、经济学、政治学、法学、图情学和教育学等学科共 36000 篇（每学科 6000 篇）期刊论文作为实验数据。以其中第 6 期论文作为测试样本共 3927 篇，其余 32073 篇作为训练样本；以题名为文本，学科类型为类目。笔者对关键词、题名词以及汉字等分类特征分别进行了统计，获得关键词特征 64526 个，非停用词语 15025 个，汉字（包括非汉字字符串）4105 个。不难发现，越是语义明确的中文语言片段，其数量也越大。因此，若仅从计算量角度而言，汉字是分类特征的最佳选择。

事实证明，IG 值能够有效压缩特征数量。然而，计算上万甚至几万个特征的 IG 值本身就存在计算复杂的问题；而且特征数量大主要是由于中文组词繁杂，表达方式多样化等原因造成的，即使根据 IG 值减少了特征项，大量低频词的存在同样会造成 TFM 极度稀疏，不利于未分类文本的类目召回。在本文数据中，以关键词作为分类特征，TFM 中仅有 132660 个单元具有有效值，所占比例不足 0.006%；类似的，以词语作为特征，也仅有约 0.05% 的单元具有有效值；而汉字特征，约有 0.34% 个单元具有非零值。可见特征的语义越丰富，能覆盖的中文短文本就越少，待分类短文本不含有特征的概率也越高。因此，以最小的汉语单元作为分类特征具有文本中特征呈现率高的优点。

图 3 列出了汉字特征量（FA）随文本量（DA）增大的变化趋势。①图 3 中实线表示 FA 的实际变化，随着 DA 的增大，FA 也随之增大；虚线则为其高度拟合的趋势线，用于描述 FA 和 DA 二元关系的发展状况。②根据趋势线公式，可计算当 DA 大概达到 55000 左右时，汉字特征达到最多，约 7200 左右。虽然这只是根据拟合曲线计算的估计值，但在一定程度上说明在本文的分类环境中，汉字特征的总量大概在 7000 ~ 8000 之间，远远小于词语或关键词等大粒度特征的可能总量。这个数量级具有较强的可计算性，在对特征进行有效筛选的条件下，可适当提高学习样本量，使之达到充分学习。

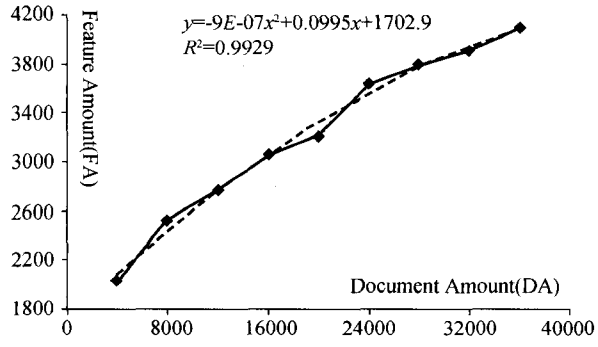

图 3 汉字特征量（FA）随文本量（DA）增大的变化趋势

3.2 三种类型特征分类效果的比较分析

笔者选择关键词（Keywords）、词语（Words）以及汉字（Characters）3 种语言片段作为分类特征，分别构建 TFM 和 TCM 进行分类计算。其中 Keywords 来自关键词文本，Words 和 Characters 均来自对题名的切分；由于 Keywords 和 Words 数量巨大，笔者采用 FIC 对特征进行筛选，令 Keywords 的 FIC = 3，即要求 Keywords “至少在一个类中出现 3 次及以上”，同理令 Words 的 FIC = 2。如此，可将特征数量控制在 10000 以下；同时其所能覆盖的文本集合也随之缩小，部分文本因为不含有 FS 中的任一项而被排除在实验之外。

经过特征筛选，可获得 Keywords 6086 个，覆盖训练样本 27501 个，测试样本 3366；Words 6973 个，被包含在 32052 个训练样本和 3926 个测试样本中；Characters 没有

被筛选，其训练/测试样本为32073/3927。特征权重则均采用最基本的Frequency，其中Keywords的出现次数被记为1。以此构建学习和测试矩阵，经过SVM计算，最终获得的分类正确率 P 如图4所示。①大粒度Keywords的 P 值最高，达到了81.40%，而粒度最小的Characters的P=79.73%，略低于Words 1.22个百分点。可见，特征语义是否明确在分类中具有重要意义；而Character由于语义缺失使得其类目区分度稍显不足，但获得的分类模型也具有一定实用价值（接近80%），劣势并不明显，设想若完善其权重设置策略或增加学习样本量，还可进一步提高其分类效果。②从总体上看，学科分类的 P 值均低于85%，可能包括两个方面原因，一是Keywords和Words的过滤可能导致类目特征的失真，影响了 P 值。但是FIC方法虽然不如IG，但是对分类效果影响甚微（参见4.1节）；二则是待分类的学科均为社会科学，相互之间存在较大的主题交叉，类目之间区别不明显。例如关键词“知识管理”，在图情学和管理学中均有大量出现。③Characters的分类效果虽然有所不足，但是Keywords和Words由于计算量问题需要筛选特征，而此举又造成了样本缺失，说明后两者的文本覆盖率较低，这在实际应用中存在较大问题。鉴于Characters在计算复杂度、文本覆盖率等方面的明显优势，而且在分类效果上还存在较大的提升空间，笔者认为以Characters作为CTC的特征具有更好的可行性和实用性。

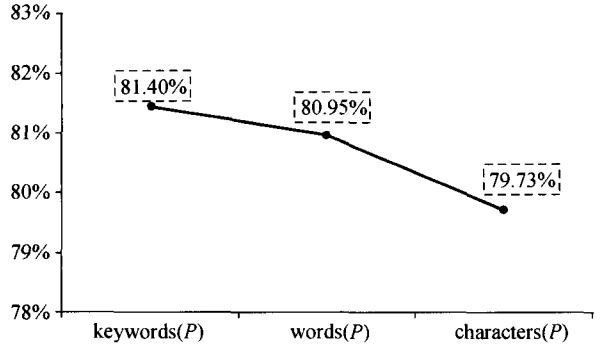

图4 三种类型特征的分类效果比较

本文重点考察Characters作为CTC特征的可行性及其随参数变化的影响规律。对于不同粒度中文语言片段作为分类特征的更深入的比较分析，笔者将另行撰文探讨。

4 汉字特征对CTC的影响规律

4.1 汉字特征的FS选择对分类效果的影响

笔者根据TCFM和TCM计算出了所有汉字特征的IG值。在4105个汉字特征中，IG值最大的为“教”0.195815，最小的为“耐”仅有 10^{-6}；在 $IG>10^{-3}$ 的670个特征中，出现次数最少的有15次，而其中出现频次超过100的有531个，约占总数的80%，可见IG值较高的特征项其出现频次也较高。笔者以 $IG>10^{-4}$ 到 $IG>10^{-3}$ 筛选特征，一共进行了10次实验；均以Frequency作为特征权重；训练/测试样本为32073/3927，不排除特征数量较小时样本量发生变化的可能性；以SVM作为机器学习算法。

于是，连同IG>0，11次实验的分类效果随特征数量的变化趋势如图5所示。①从总体上来看，随着特征数量的持续减少，分类效果呈现减弱趋势，这符合“特征量与类目区分能力呈正比”的基本规律。②P 值在出现大幅下滑前，经历了一段稳定期，甚至还出现了小幅上升，说明FS中存在大量的冗余特征。③当特征数量不少于1138时，分类效果较好，此后虽然期间有所震荡，但 P 值总体下降趋势已经形成，可见在本文4105个汉字特征中大约有近3000个冗余特征。④随着特征数量持续减少，FS所涵盖的文本量也不断减少，部分样本因不含任何特征项而无法参加实验。解决该问题主要有两种方法：一是尽量维护FS的规模，容忍特征冗余来保证分类效果和文本覆盖率，这会增大样本训练的计算量；二是将文本原有特征项通过相似度计算转化为FS中的特征项，以保证文本有特征，很明显可操作性较差。笔者认为，具体采用何种方法要视实际应用而定。在采用汉字特征的情况下，特征数量一般有限，可在适当排除冗余特征的基础上，尽量保持FS的覆盖面，增强分类系统的可操作性和实用性，寻找分类正确率和类目召回率之间的平衡点。

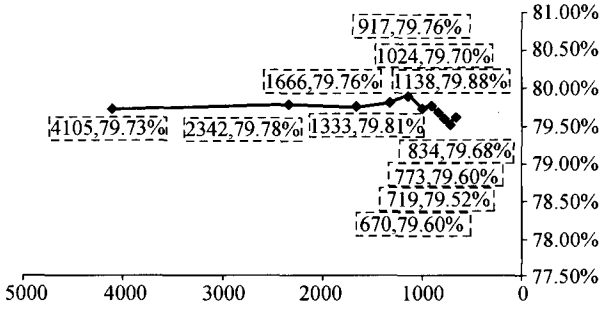

图5 汉字特征的分类效果随特征数量（根据IG值筛选）的变化规律

然而，直接使用IG方法筛选特征并不合理。例如3.2节中的Keywords和Words特征，直接计算上万个特征项的IG值存在计算复杂的问题，而且在类目中偶然出现的语言片段也并不适合作为分类特征。因此，笔者追加了两个实验：在采用IG方法筛选特征之前，先基于“特征至少在一个类目中出现2/3次”对FS进行选择，获得的 P

值为：79.73%（特征项分别为2471/2025个），即FIC方法在阈值较小的情况下是合理的，3.2节中对Keywords及Words的筛选不会大幅改变其分类效果。因此在实际应用中，当特征数量较大时，可以采用FIC和IG相结合的特征选择方法。

4.2 汉字特征的FW设置对分类效果的影响

汉字特征的FW设置主要围绕Frequency和Position两个参数进行。Frequency衍生出Presence、TF以及TF-IDF等计算值；而Position则有文本中位置（Position_Text）、语句中位置（Position_Sentence）以及所在词语中位置（Position_Word）3种类型，且若汉字在文本中多次出现，那么其位置又可以用位置和（Position_Sum）和平均位置（Position_Avg）两种方式来表示。本文采用了Position_Word和Position_Avg的使用策略。前文论述的8种权重设置方案均是从这两个参数及两者的结合中派生出来的，笔者以此进行了12次实验，分类正确率P的变化情况如图6所示。

如图6所示，①Presence的P值太小，仅为17.65%，

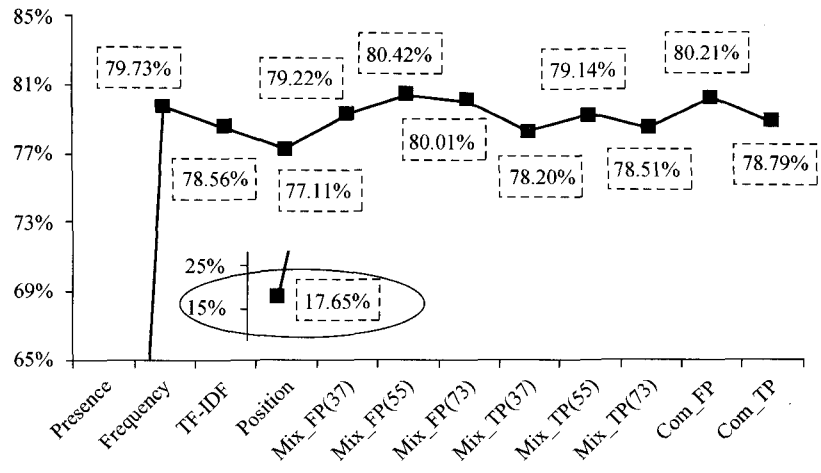

图6 汉字特征在不同FW设置方案下的分类效果比较

影响到了其他权重下分类效果的对比，在椭圆内单独标注。很明显，若仅考虑特征是否存在，不具明确语义的Characters基本上没有类目区分能力，远落后于Keywords（81.40%）。②在其他单权重中，Frequency效果最好达到了79.73%，而Position效果最差，仅为77.11%，可见位置因素对于汉字具有一定的指示作用，汉字在词语中位置不同，往往也表现出不同的语义。值得探讨的是TF-IDF，从理论上讲该值比Frequency更加合理，更能描述特征对文本的重要程度。但是TF-IDF值有个假设前提，就是认为样本中出现的相同语言片段具有相同语义，而这一假设对于汉字并不成立，因此TF-IDF权重在Characters中表现不佳。③Position与其他权重复合时，表现出了一定的强化作用。当其与Frequency复合且后者权重较大时，均可获得比单独使用Frequency更好的P值，并在两者各占一半权重时达到最佳80.42%，基本上接近了Words，可见Characters在适当改进FW后，在分类效果上具有一定的提升空间；当其与TF-IDF值复合且两者比例各占50%时，分类正确率也超过了79%。④当Frequency以及TF-IDF和Position分列为两个特征值时，P值均出现了较单独使用Frequency和TF-IDF略微的提升，但这种以牺牲计算复杂度为代价的权重设置方法至少在本文的数据环境中并不可取。

对汉字FW设置方案的改进，可以在一定程度上提高分类效果；虽然最终没有超过Words的分类效果，但已经与其非常接近，而且综合计算复杂度、文本覆盖率等实用性因素，在短文本环境下，使用汉字特征可能更具有合理性和有效性。

5 结束语

TC的效果取决于TR的特征，包括FT以及FS选择和FW设置。CT通常采用具有明确含义的Words或篇幅更长的语言片段作为其特征，而中文构词的复杂性往往导致此类型特征数量庞大，需要降维来降低其计算量，却因此又带来了分类效果下降和短文本特征缺失等问题。因此，本文对Characters进行了全面剖析，探讨了其在CTC中的表现及其对分类的影响规律，笔者认为：Characters的分类效果略低于篇幅大的语言片段，但其具有可计算性强和文本覆盖率高的优点；合理的FS选择需要在尽量去除冗余特征和保证特征较大文本覆盖率之间寻找最佳平衡点；Position和Frequency参数的有效结合可以在一定程度上提高Characters的分类效果。

本文研究表明，Characters因其粒度小具有可操作性强和实用性好等优点，经过特征的筛选和权重设置的综合平衡，可最大限度地提高其分类效果，在某些应用环境下可以作为Words或其他语言片段的替代。□

参考文献

[1] SEBASTIANI F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34: 1-47.

[2] 王昊，严明，苏新宁. 基于机器学习的中文书目自动分类研究 [J]. 中国图书馆学报，2010（6）：28-39.

[3] WANG Jun. An extensive study on automated dewey decimal

classification [J]. Journal of American Society for Information Science and Technology (JASIST), 2009, 60 (11): 2269-2286

[4] WANG Hongwei, YIN Pei, YAO Jiani, et al. Text feature selection for sentiment classification of Chinese online reviews [J]. Journal of Experimental & Theoretical Artificial Intelligence, 2013, 25 (4): 425-439.

[5] MAKS I, VOSSEN P. A lexicon model for deep sentiment analysis and opinion mining applications [J]. Decision Support Systems, 2012 (53): 680-688.

[6] FIGUEROA R L, et al. Active learning for clinical text classification: is it better than random sampling? [J]. Journal of the American Medical Informatics Association, 2012, 21 (6): 651-658.

[7] BOTSIS T, NGUYEN M D, WOO E J, et al. Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection [J]. Journal of the American Medical Informatics Association, 2011, 18 (5): 631-638.

[8] PENG F C, HUANG X J. Machine learning for Asian language text classification [J]. Journal of Documentation, 2007, 63 (3): 378-397.

[9] ZHANG Yongkui, LI Hongjuan. Text classification of accident news based on category keyword [J]. Journal of Computer Applications, 2008, 28 (S1): 139-140, 143.

[10] CHEN Z G, SHI G, WANG X J. Text classification based on Naive Bayes algorithm with feature selection [J]. Information-An International Interdisciplinary Journal, 2012, 15 (10): 4255-4260.

[11] GHIASSI M, OLSCHIMKE M, MOON B, et al. Autom-ated text classification using a dynamic artificial neural network model [J]. Expert Systems with Applications, 2012, 39 (12): 10967-10976.

[12] JOACHIMS T. Text categorization with support vector machines: learning with many relevant features [C]//Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, 1998: 137-142.

[13] YEN Show-Jane, LEE Yue-Shi, YING Jia-Ching, et al. A logistic regression-based smoothing method for Chinese text categorization [J]. Expert Systems with Applications, 2011, 38 (9): 11581-11590.

[14] ZHANG B, MARIN A, HUTCHINSON B, et al. Learning phrase patterns for text classification [J]. IEEE Transactions on Audio Speech And Language Processing, 2013, 21 (6): 1180-1189.

[15] 林伟，孟凡荣，王志晓．基于概念特征的语义文本分类[J]．计算机工程与应用，2011，47（28）：139-142.

[16] 宁亚辉，樊兴华，吴渝．基于领域词语本体的短文本分类[J]．计算机科学，2009，36（3）：142-145.

[17] ZHONG J, SUN Q G, LI X, et al. A novel feature selection method based on probability latent semantic analysis for chinese text classification [J]. Chinese Journal of Electronics, 2011, 20 (2): 228-232.

[18] 王梦云，王素格．一个基于字特征的文本分类模型[J]．计算机工程与应用，2004（13）：64-65，191.

[19] ZHOU X Z, WU Z H. Distributional character clustering for Chinese text categorization [C]//Proceeding in PRICAI 2004: Trends in Artificial Intelligence, Lecture Notes in Computer Science 2004: 575-584.

[20] LEE C, LEE GG. Information gain and divergence-based feature selection for machine learning-based text categorization [J]. Information Processing & Management, 2006, 42 (1): 155-165.

[21] HEUM P, HYUK-CHUI K. Improved Gini-Index Algorithm to Correct Feature-Selection Bias in Text Classification [J]. IEICE Transactions on Information and Systems, 2011 (4): 855-865.

[22] LIU H, SUN J, LIU L, et al. Feature selection with dynamic mutual information [J]. Pattern Recognition, 2009, 42 (7): 1330-1339.

[23] CHEN Y T, CHEN M C. Using chi-square statistics to measure similarities for text categorization [J]. Expert Systems with Applications, 2011, 38 (4): 3085-3090.

[24] WERI Zhihua, MIAO Duoqian, et al. N-grams based feature selection and text representation for Chinese Text Classification [J]. International Journal of Computational Intelligence Systems, 2009, 2 (4): 365-374.

[25] SALTON G, BUCKLEY C. Term weighting approaches in automatic text retrieval [J]. Information Processing & Management, 1988, 24 (5): 513-523.

[26] ZAHEDI M, SORKHI A G. Improving text classification performance using PCA and Recall-Precision Criteria [J]. Arabian Journal for Science and Engineering, 2013, 38 (8): 2095-2102.

[27] REN Fuji, SOHRAB M G. Class-indexing-based term weighting for automatic text classification [J]. Information Sciences, 2013, 236: 109-125.

[28] AIZAWA A. Linguistic techniques to improve the performance of automatic text categorization [C]//proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS), 2001: 307-314.

作者简介：王昊，男，1981年生，博士，副教授。
邓三鸿，男，1975年生，博士，副教授。
苏新宁，男，1955年生，教授，博士生导师。

收稿日期：2014-11-24