基于改进TF-IDF-CHI算法的农业科技文献文本特征抽取*

杜若鹏 鲜国建 寇远涛
（中国农业科学院农业信息研究所/农业农村部农业大数据重点实验室，北京 100081）

摘要：针对相近农业科研领域文献的文本特征信息高度重合的特点，以及传统的文本特征抽取方法存在的不足，对TF-IDF算法进行优化并加以应用验证。通过引入卡方检验值与特征词频修正因子等方式，对特征词加权函数进行重构，形成改进的ImpTF-IDF-CHI方法。将该方法与文档频率法、信息增益法及TF-IDF 3种传统的文本特征抽取结果应用于朴素贝叶斯分类实验，根据实验结果判定方法的优劣性。通过4种方法的58组特征抽取与文本分类实验，发现与前述的3种特征抽取方法相比，ImpTF-IDF-CHI方法抽取的特征词，应用于文本分类的正确率最高，平均准确率达94%，F1值为0.844，证明该方法在对相近农业科研领域文本进行特征抽取方面，具有准确率高、稳定性好、主题词代表性强等优点，可以有效地应用于此类文献文本分类、特征表达、主题抽取等场景。

关键词：特征抽取；TF-IDF；卡方统计；文本分类；农业科技文献
中图分类号：TP391；G250
DOI：10.3772/j.issn.1673-2286.2019.08.003

在海量的科技信息中，文本文献是最重要的部分[1]，文本自动分类技术是组织和管理海量科技信息的重要手段[2]。文本自动分类研究中，对内容相似类目（用词上非常接近的不同类别）的处理是其中一个重要课题[3]。

在农业科技文献中，相近的研究领域的文献，其文本特征信息是高度重合的，在很多情况下，虽然研究的对象不同，但研究方向相同或相近时，其研究手段、分析方法往往都是相同或相似的。如番茄、辣椒和茄子，虽然是不同的作物，但是产品器官均为果实，其育种目标、育种途径及应用的主要技术方法基本相同，因此这3种作物在遗传育种方面的文献，其关键词、高频词以及全文用词的相似度非常高。如何对这种内容相似类目进行精准分类，是农业数字图书馆进行专题文献分类以及开展个性化检索服务时需要解决的重要问题。

文本自动分类中较为关键的环节是文本特征抽取，特征抽取准确与否，直接影响文本分类的最终效果。基于信息测度的特征选择算法是目前最常用的，包括文档频率（Document Frequency，DF）、信息增益（Information Gain，IG）、词频-逆文档频率（Term Frequency-Inverse Document Frequency，TF-IDF）、卡方检验（Chi-squared，CHI）、互信息（Mutual Information，MI）以及期望交叉熵（Expected Cross Entropy，ECE）等[4]。上述方法在实践应用中各自表现出优点和不足，所以一直处于不断改进和完善中。

笔者从探索适合农业科研领域内容相似类目文献精准分类方法的目的出发，在前人对TF-IDF改进的基础上，结合实际应用情况作进一步改进，形成改进的TF-IDF-CHI（ImpTF-IDF-CHI）方法。本实验中运用该方法以及传统的文档频率法、信息增益法、TF-IDF法，对从农业专业知识服务系统中文科技期刊论文数据中选取的近十年有关番茄、辣椒、茄子、黄瓜、马铃薯和模式植物拟南芥等遗传育种主题文献4 000多篇进行文本特征抽取，并应用于朴素贝叶斯分类实验，比较其效果。结果表明，ImpTF-IDF-CHI方法效果最好，抽

*本研究得到国家社会科学基金项目“科技论文全景式摘要知识图谱构建与应用研究”（编号：19BTQ61）、中国农业科学院科技创新工程项目（编号：CAAS-ASTIP-2016-AII）和中国工程科技知识中心建设项目（编号：CKCEST-2018-1-15）资助。

取的主题词代表性强，主题分类平均准确率达94%，且稳定性好，为今后进一步开展农业科技文献主题词扩展、专题文献分类、检索，以及个性化专题情报服务打下一定基础。

1 实验中采用的3种传统方法

1.1 文档频率法

文档频率法是统计文档集中包含每一个词的文档个数，设置阈值，保留高于阈值的文档频数所对应的词作为特征词，过滤掉低于阈值的低频词。文档频率法简单易行，但是较为粗糙，而且词条的文档频率阈值不好确定，阈值过大易导致具有代表性的词条丢失，过小又会导致入选词条包含大量无贡献的低频词，影响分类效果[5]。

1.2 信息增益法

信息增益法是根据词条能为整个分类系统提供的信息量的多少来决定其重要程度。信息增益用特征词在文本中出现时与不出现时的信息熵之差表示，依据差值的大小决定其作为特征词的取舍[6]。信息增益算法相对简单。但是由于考虑特征词出现与不出现两种情况，对于小数据集，或在类别分布不平衡的情况下，不出现的特征权值将产生主导作用，因此很难提取小样本集特征，或者增益比较大的特征词的实际词频较低[7-8]。由于在实际应用中很多增益比较大的特征词的实际词频较低，当选择的特征数量偏小时容易陷入数据稀疏的问题，所以本实验中的信息增益法采用IG方法与DF文档频率法结合（IG-DF）的方式进行。

1.3 TF-IDF

TF-IDF由Salton在1988年提出[9-10]，利用词频和逆文档频率（出现某词条的文档的倒数）的乘积来衡量词条作为特征词的分类能力，该算法在文本分类领域得到广泛应用。但是传统的TF-IDF法存在不足[10-18]，如同一个特征词在长文档中往往比在短文档中出现的频数更大，会影响到分类效果；还有就是忽略了数据集偏斜（数据集中各类文档数不均衡，可能存在数量级的差距）和特征词在类间和类内的分布（在一类中经常出现而在其他类中很少出现的特征词权重会被低估）等问题。众多学者针对TF-IDF算法存在的不足进行过多次改进。如使用特征词在文档中的频率代替其在文档中的频数[10]，来减弱文档长短带来的影响；How等[11]提出CTD（Category Term Descriptor）法，以弥补类别数据集偏斜带来的困扰；沈志斌等[12]提出BOR-TF-IxDF法，张瑜等[13]提出由“权重调整因子–类内离散度–类间偏斜度”组成的WA-DI-SI算法，赵小华等[14]将TF-IDF与CHI相结合的算法，路永和等[15]将特征权重（Term Weight，TW）与TF-IDF结合的算法等，都是重新修正各个特征词的权重，以减弱特征词在类间、类内分布带来的影响。

2 改进的TF-IDF-CHI（ImpTF-IDF-CHI）

笔者改进的TF-IDF-CHI（ImpTF-IDF-CHI）方法包括重新构造卡方值加权函数、特征词词频加权函数以及逆文档频率加权函数，并对特征词加权函数引入修正因子。

（1）重新构造卡方值加权函数。通过重新构造卡方值加权函数对特征词权重计算进行优化。构造函数见公式（1）。

$$W_{CHI}(t_i)=\begin{cases}\log CHI(t_i) & A_{t_i}D_{\bar{t}_i}-B_{t_i}C_{\bar{t}_i}>0\\ -\log CHI(t_i) & A_{t_i}D_{\bar{t}_i}-B_{t_i}C_{\bar{t}_i}<0\end{cases} \quad (1)$$

其中，t_i代表第i个特征词；$CHI(t_i)$表示该特征词的卡方值，计算方式见公式（2）。

$$CHI(t_i)=\frac{N(A_{t_i}D_{\bar{t}_i}-B_{t_i}C_{\bar{t}_i})^2}{(A_{t_i}+C_{\bar{t}_i})(B_{t_i}+D_{\bar{t}_i})(A_{t_i}+B_{t_i})(C_{\bar{t}_i}+D_{\bar{t}_i})} \quad (2)$$

N代表总文档数，A_{t_i}表示包含特征词t_i且属于假设目标分类的文档数，B_{t_i}表示包含特征词t_i且不属于假设目标分类的文档数，C_{t_i}表示不包含特征词t_i且属于假设目标分类的文档数，D_{t_i}表示不包含特征词t_i且不属于假设目标分类的文档数。构造函数中，对卡方值进行取对数运算是为避免直接使用卡方值导致加权值波动剧烈，出现过大或过小的情况。同时，根据$A_{t_i}D_{\bar{t}_i}$与$B_{t_i}C_{\bar{t}_i}$差值，对构造函数进行了分段处理。对于$B_{t_i}C_{\bar{t}_i}$大于$A_{t_i}D_{\bar{t}_i}$这种情况，通过对特征加权值判定为负值的方式，对其进行特征过滤。

（2）重新构造特征词词频加权函数。对TF-IDF特征词加权函数进行优化处理。首先是对特征词词频加权部分进行函数重构，见公式（3）。

$$TW(t_i, d_u) = \left(\sum_u^n \frac{freq_{(t_i, d_u)}}{\sum_j freq_{(t_i, d_u)}}\right) \times \frac{freq\_C_{(t_i)}}{\sum_k freq_{(t_{i_k})}} \times \frac{freq\_C_{(t_i)}}{\sum_m freq\_C_{(t_m)}} \times \lambda \quad (3)$$

其中，第u篇文档用d_u表示，$freq_{(t_i, d_u)}$表示文档d_u中的特征词t_i的绝对词频，$\sum_j freq_{(t_i, d_u)}$表示文档d_u中所有特征词的总词频。二者相除等于特征词t_i在文档d_u中的相对词频，然后再计算每篇文档中的特征词t_i的相对词频并且相加求和，得到特征词t_i的总相对词频。所以使用相对词频代替传统方法中绝对词频，是为避免特征加权计算时词频计算偏向长文本。同时，引入2个修正因子，对特征词t_i的加权计算进行优化。$freq_C_{(t_i)}$表示目标分类C所包含的全部文档中特征词t_i的绝对词频之和，$\sum_k freq_{(t_{i_k})}$表示特征词t_i在所有文档中的绝对词频之和。二者之比，表示特征词t_i在目标分类C中的词频占自身总词频的多少。该比值取值范围在[0，1][19]。比值越大、越接近1，表示特征词t_i主要分布在目标分类C的所属文档，从侧面反应特征词t_i的词频分布与目标分类C文档的相关性。同理，$\sum_m freq_C_{(t_m)}$表示目标分类C文档中所有特征词的词频之和，$freq_C_{(t_i)}$与其比值越大、越接近1，表示特征词t_i在目标分类C文档中的词频比重大。该比值可以一定程度上反应特征词t_i在分类C中的重要程度。在实际应用中，多个小数相乘的结果往往会很小，甚至接近0。因此，引入常数λ对运算结果进行放大。本试验中常数λ设置为10 000，取得了良好效果。引入上述2个修正因子，理论上可以有效地筛选与目标分类C契合度高且重要的特征词。

（3）重新构造逆文档频率加权函数。对逆文档频率加权计算部分进行优化改造，见公式（4）。

$$W_{IDF}(t_i) = \log \frac{N}{n_{(t_i)} + 1} \quad (4)$$

N表示文本集中的总文档数，$n_{(t_i)}$代表包含特征词t_i的文档数，即t_i的文档频率。采用对数运算是为了使逆文档频率更加平滑。为应对特征词的文档频率为0的情况，在分母增加常数项，从而避免导致无法计算的局面。

综上所述，改进的文本特征词加权函数见公式（5）。

$$\begin{aligned} W(t_i, d_u) &= ImpTF\_IDF\_CHI(t_i, d_u) \\ &= TW(t_i, d_u) \times W_{IDF}(t_i) \times W_{CHI}(t_i) \end{aligned} \quad (5)$$

最终的特征词加权计算公式，分别从特征词词频的相关性及贡献度、文档频率普遍性、特征词与目标分类间的独立性等多维度对特征词的加权进行综合衡量。

3 实验的方法与步骤

3.1 实验方法

为了检验ImpTF-IDF-CHI(t_i, d_u)对文本特征词抽取加权算法的有效性，进行文本分类实验。首先，对农业专业领域的相近主题文献进行特征词抽取，然后通过文本分类验证所抽取的特征词对所在分类文献的区分效果，证明其具有很强的文本特征代表性。同时，为证明ImpTF-IDF-CHI(t_i, d_u)方法的优越性，将其分类效果与前述的3种特征抽取方法进行比较。实验平台的硬件、软件环境，如表1所示。

表1 实验平台

硬件环境	软件环境
CPU：Intel i7 2600 内存：32GB 硬盘：500GB	操作系统：64位Windows 10 专业版 开发语言：JDK 1.8.0 & Python 3.7.2 开发工具：Netbeans8 Pycharm2018 检索工具：Lucene 7.3 分词器：IKAnalyzer

3.2 实验步骤

采用农业专业知识服务系统的近十年的部分中文期刊论文数据，从中选取番茄遗传育种、辣椒遗传育种、黄瓜遗传育种、茄子遗传育种、马铃薯遗传育种以及以模式植物拟南芥为研究主体的6类文献（其主要特点是高频词、关键词重合，全文用词相似度高），采用文档频率（DF）法、信息增益与文档频率相结合（IG-DF）法、TF-IDF法和ImpTF-IDF-CHI 4种特征抽取方法进行二元分类。实验流程如图1所示。

首先，选取6类主题文献共计4 297篇作为本实验的文本集，其中番茄遗传育种主题为1 387篇，其余5类主题文献为2 910篇。设定番茄遗传育种主题为二元分类的目标分类，其余5类分类为干扰分类。从文本集中，各自选取目标分类与干扰分类主题文献各1 000篇作为训练集，其余2 297篇作为测试集，6类主题比例接近1∶1，平均每类主题测试集文档近390篇。

对文本集的预处理主要包括剔除非文本标点符号与中文分词处理。由于所涉及文献包含大量农业专业领域词汇，如果使用日常生活词库进行中文切词，则会破坏专业词汇的完整性，导致分词错误。因此，需要构

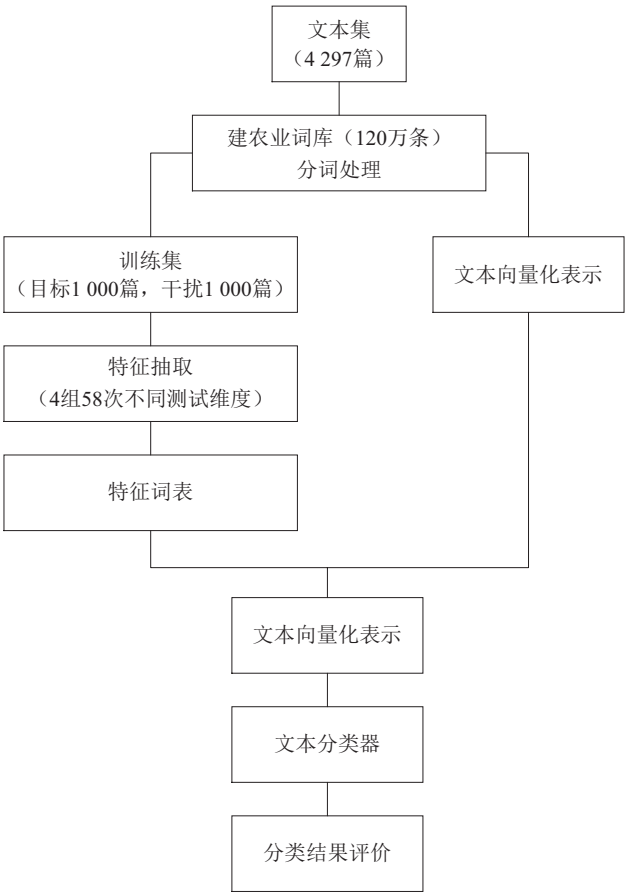

图1 特征抽取及文本分类实验流程

建农业专业领域词表作为专业分词词库。农业专业分词词库主要包括中国农业叙词表（CAT）、搜狗农业专业词表以及农业专业知识服务系统论文词表。其中，农业专业知识服务系统论文词表中的词汇主要来源于服务系统收录的近十五年的350万篇农业专业中文核心期刊论文的关键词。最终通过去重、过滤低频词等，将上述专业词表与日常用语词表进行整合，形成包含120万词条的中文分词词库。

对分词后的训练集数据进行DF、IG-DF、TF-IDF和ImpTF-IDF-CHI 4类特征抽取方法测试，生成各自对应的特征词表。通过特征词表，实现对文本集数据的降维文本表示。将原本一篇数千字的科技论文，使用特征词表中出现的词汇进行表达，过滤特征词表中未出现的非特征词，从而达到降维、优化文本分类计算的效果。特征表中的词汇数量就是特征维度。特征维度的选取对文本分类效果具有决定性影响。为更好地对比分析4类特征抽取方法的效果，通过分组选取不同特征维度进行测试对比的方式，找出最优的特征抽取方法。每种方法为一组，分别进行特征维度为100、150、200、250、3 000、3 200等16次特征抽取实验。通过向量空间模型（VSM）将降维后的文本集进行文本向量化表示，供分类器进行文本分类运算。分类算法采用朴素贝叶斯分类算法[20]。

对分类结果的评价指标主要有准确率（accuracy）、精确率（precision）、召回率（recall）、F1值（F1-measure）及上述指标的宏观平均值[21]。

4 实验结果与分析

DF、IG-DF、TF-IDF和ImpTF-IDF-CHI 4种特征抽取方法在朴素贝叶斯分类下的应用评价结果如图2和表2所示。

（1）ImpTF-IDF-CHI方法相较TF-IDF方法，改进效果明显，无论从准确率与精确率，还是从召回率与F1值而言，均存在较大优势。在采用相同特征数量进行对比实验中，ImpTF-IDF-CHI方法的4个评价指标均大幅优于TF-IDF方法。ImpTF-IDF-CHI方法的16组实验的平均准确率为94.07%，平均F1值为0.844，而TF-IDF方法的平均准确率只有78.00%，平均F1值为0.566%。

（2）ImpTF-IDF-CHI方法相较IG-DF方法，结果更加稳定。由于进行信息增益计算的特征词必须满足文档词频超过预设的阈值，从而避免数据稀疏性问题。通过反复实验，在全部文本集中，满足超过阈值的全部特征词数为1 000词左右，所以在实验中IG-DF方法的对比实验截止到特征词数为1 000。由实验结果可以看出，IG-DF方法的不足在于特征抽取词的分类正确率波动太大，在本实验中，在特征词数为1 000时效果最好，但特征词数较少时，则效果较差，正确率方差为6.2%。IG-DF方法还面临数据稀疏性问题。在本实验中通过反复迭代设置阈值测试，最终确定阈值为190文档频率时，才能完全避免稀疏性问题。同时，阈值的大小与避免稀疏性存在一定正相关性，但绝非严格递进关系，如当阈值取200文档频率时，反而有些文档无法用IG-DF方法抽取出的特征词来表示，导致该文本为空的现象。ImpTF-IDF-CHI方法与之相比，稳定性优势较为突出，本实验中在特征词数量为250时效果最好，而整体上，在实际应用中任取某一组特征数量进行文本分类任务，其效果都较好。

（3）ImpTF-IDF-CHI方法相较DF方法，准确率、精确率和F1值都较高且稳定，但平均召回率（95.27%）略低于DF（96.27%），说明DF方法与朴素贝叶斯文本

分类方法的组合，在应对长文本分类任务时有不错的表现。但从表3可以看出，DF方法仅从统计维度进行考量，提取的前N个特征词往往不具有该分类的主题代表性，如“研究”“表”“分析”“采用”“图”等。相比之下，ImpTF-IDF-CHI方法抽取的Top N个词就具有一定的主题代表性。

（4）ImpTF-IDF-CHI方法与3种传统方法整体比较，4单位评价指标中，3单位均为第一，且3单位指标的方差均为最小。在较为重要的平均准确率对比中，ImpTF-IDF-CHI方法高出第二名4个百分点，同时16次测试的准确率方差只有0.667%（方差最小说明最为稳定）。从图2中也可以看出，ImpTF-IDF-CHI方法曲线较为平滑。综合加权统计指标F1值高且接近于1，意味着精确率与召回率双高且较为平衡。ImpTF-IDF-CHI方法的F1值领先第二名0.074，近8个百分点。另外，ImpTF-IDF-CHI方法抽取的主题词在4种方法中代表性最强（见表3）。

因此，综合比较可以得出：ImpTF-IDF-CHI方法在4种特征抽取方法中正确率最高、稳定性最好，与其他3种方法相比具有明显的优越性。

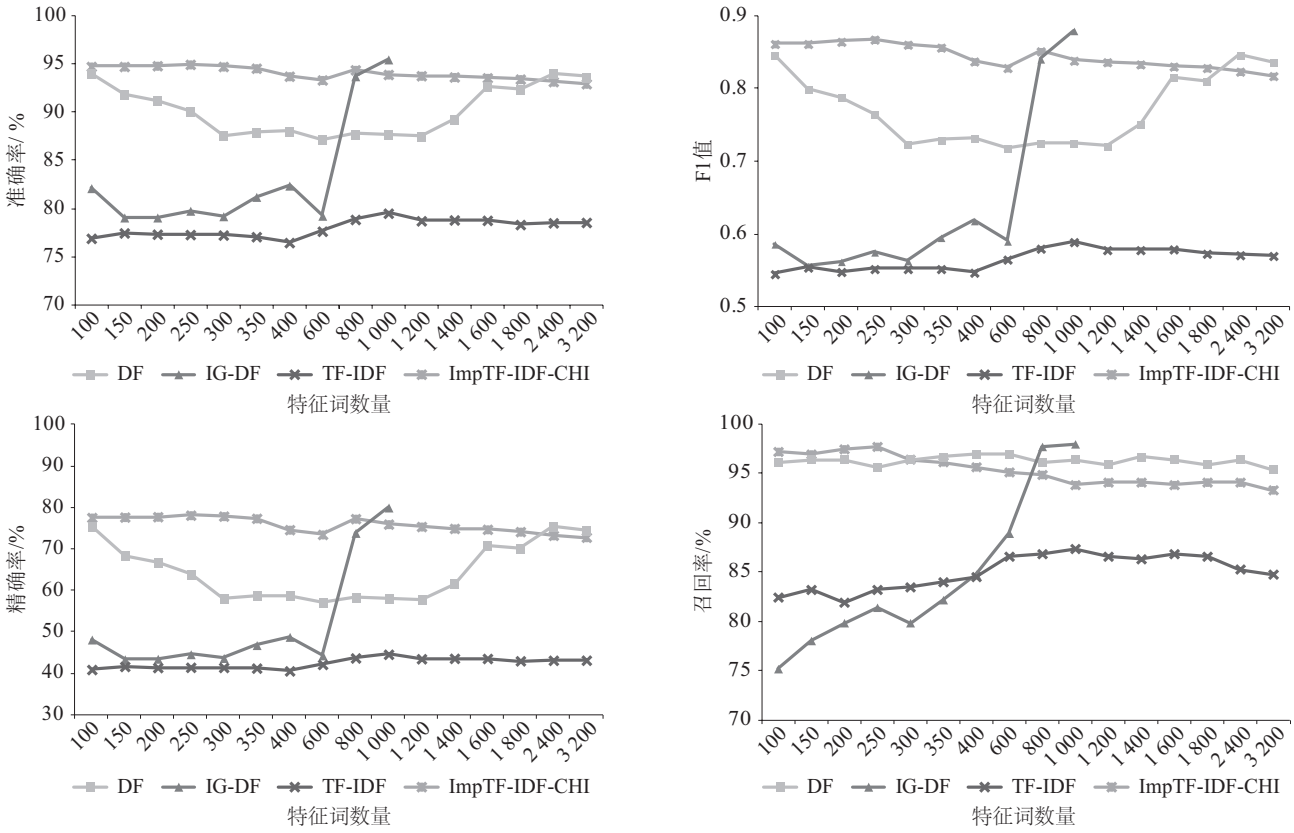

图2 DF、IG-DF、TF-IDF和ImpTF-IDF-CHI方法的准确率、F1值、精确率和召回率对比

表2 DF、IG-DF、TF-IDF和ImpTF-IDF-CHI的平均准确率、F1值、精确率和召回率对比

方 法	平均准确率/%	平均F1值	平均精确率/%	平均召回率/%
DF	90.20	0.770	64.68	96.27
IG-DF	83.20	0.640	51.49	84.55
TF-IDF	78.00	0.566	42.38	85.01
ImpTF-IDF-CHI	94.07	0.844	75.79	95.27

5 结束语

本文针对相近农业科研领域文献的文本特征信息高度重合的特点，以及传统的文本特征抽取方法存在的不足，对TF-IDF算法进行优化并加以应用验证，提出了ImpTF-IDF-CHI方法，通过对照实验证明其与其他3种传统特征抽取方法相比具有较高的正确性与可靠性。ImpTF-IDF-CHI方法主要从特征词的重要性、代表性等因素考虑，通过使用相对词频与引入贡献因

表3 4种特征抽取方法的Top 40特征词对比

序 号	DF	IG-DF	TF-IDF	ImpTF-IDF-CHI
1	番茄	番茄基因组	基因	番茄
2	***研究***	***穗***	果实	基因
3	***表***	可溶性固形物	***植株***	tomato
4	***分析***	单果质量	***对照***	果实
5	***采用***	***各处***	引物	引物
6	***利用***	***恢复***	品种	番茄品种
7	***材料***	病毒病	***材料***	转基因番茄
8	***结果表明***	番茄叶片	*plant*	抗性
9	***结***	外源基因	产量	品种
10	***植株***	随机区组	表达	表达
11	***中的***	***株距***	tomato	番茄果实
12	王	***河南***	***接种***	***材料***
13	基因	***茄科***	***培养基***	转基因植株
14	***李***	多态性	***愈伤组织***	番茄红素
15	***图***	序列分析	***叶片***	DNA
16	***讨论***	***后用***	外植体	gene
17	***果与***	***白色***	抗性	单果质量
18	***影响***	***正向***	诱导	PCR
19	***高***	***消毒***	突变体	扩增
20	tomato	遗传分析	DNA	***园艺***
21	***植物***	***得的***	***嫁接***	外植体
22	***发现***	一代杂种	***野生型***	抗病
23	***方法***	***群***	扩增	鉴定
24	***差异***	***吕***	抗病	***对照***
25	果实	*levels*	增产	克隆
26	***材料与方法***	*isolation*	***浓度***	转化
27	***对照***	***部***	***子叶***	***花序***
28	*plant*	氨基酸序列	转基因植株	***利用***
29	***进一步***	hm^2	亲本	产量
30	***提供***	***预期***	***比对照***	序列
31	品种	基因对	*CMV*	***乙烯***
32	***张***	*pepper*	转化	可溶性固形物
33	***北京***	*map*	单果质量	选育
34	***叶***	互作	一代杂种	Fruit
35	***园艺***	*tion*	***病情指数***	番茄植株
36	***试验***	***比对***	***提高***	高抗
37	***提高***	*horticulturae*	***蛋白***	转基因
38	鉴定	***存在着***	***花序***	片段
39	***重复***	***症状***	***含量***	新品种
40	***叶片***	高产	自交系	母本

注：粗斜体字为明显不符合“番茄遗传育种”主题

子等手段，提升特征词词频加权的代表性；同时，在构造函数中加入了卡方检验因子，强化特征词的分类特征一致性，使得同一分类下的特征词，在加权得分上更加聚合。

首先本文用于实验的文本语料仅限于中文，尚未证明加权改进方法对英文等语种文献是否同样有效；其次，本文的文本分类任务是二元分类，尽管目前需要对科技文献进行二元分类的应用场合依然很多，但要满足需求的多元化，需要继续进一步优化方法及实现，以便适应多元分类需求，拓展ImpTF-IDF-CHI方法的应用范围；最后，词语在文档中位置不同，对文本特征的贡献度是不一样的，这也是需要在今后对该项技术进一步修改时要考虑的。

参考文献

[1] 隗中杰. 文本分类中TF-IDF权重计算方法改进 [J]. 软件导刊，2018，17（12）：39-42.

[2] 牛永洁，田成龙. 融合多因素的TFIDF关键词提取算法研究 [J]. 计算机技术与发展，2019，29（7）：80-83.

[3] 李湘东，阮涛. 互信息特征选择法在《中图法》内容相似类目中的运用及改进 [J]. 数字图书馆论坛，2018（1）：46-52.

[4] 徐冠华，赵景秀，杨红亚，等. 文本特征提取方法研究综述 [J]. 软件导刊，2018，17（5）：13-18.

[5] 杨凯峰，张毅坤，李燕. 基于文档频率的特征选择方法 [J]. 计算机工程，2010，36（17）：33-35，38.

[6] 唐康，汪海涛，姜瑛，等. 混合CHI与IG的特征选择方法研究 [J]. 信息技术，2019（2）：53-57.

[7] 任永功，杨荣杰，尹明飞，等. 基于信息增益的文本特征选择方法 [J]. 计算机科学，2012，39（11）：127-130.

[8] 王理冬. 基于信息增益的文本特征选择方法 [J]. 电脑知识与技术，2017，13（25）：242-244.

[9] 郑霖，徐德华. 基于改进TFIDF算法的文本分类研究 [J]. 计算机与现代化，2014，229（9）：6-9，14.

[10] 施聪莺，徐朝军，杨晓江. TFIDF算法研究综述 [J]. 计算机应用，2009，29（6）：167-170，180.

[11] HOW B C，NARAYANAN K. An empirical study of feature selection for text categorization based on term weightage [C] // Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence. Washington：IEEE Computer Society，2004：599-602.

[12] 沈志斌，白清源. 文本分类中特征权重算法的改进 [J]. 南京师范大学学报（工程技术版），2008，8（4）：95-98.

[13] 张瑜，张德贤. 一种改进的特征权重算法 [J]. 计算机工程，

2011，37（5）：210-212.

［14］赵小华，马建芬. 文本分类算法中词语权重计算方法的改进［J］. 电脑知识与技术，2009，5（36）：10626-10628.

［15］路永和，李焰锋. 改进TF-IDF算法的文本特征项权值计算方法［J］. 图书情报工作，2013，57（3）：90-95.

［16］赵迎光，范少萍，安新颖. 学科背景知识在医学文本特征抽取中的应用［J］. 医学信息学杂志，2017，38（4）：51-54.

［17］黄珊珊，廖闻剑. 基于混合特征的文本分类研究［J］. 电子设计工程，2019，27（7）：61-65.

［18］叶雪梅，毛雪岷，夏锦春，等. 文本分类TF-IDF算法的改进研究［J］. 计算机工程与应用，2019，55（2）：104-109.

［19］赵鸿山，范贵生，虞慧群. 基于归一化文档频率的文本分类特征选择方法［J/OL］. 华东理工大学学报（自然科学版）：1-8［2019-05-06］. https://doi.org/10.14135/j.cnki.1006-3080.20180914005.

［20］李航. 统计学习方法［M］. 北京：清华大学出版社，2012：47.

［21］陶峰，汤鲲，程光. 基于改进TFIDF算法的邮件分类技术［J］. 计算机技术与发展，2018，28（8）：27-31.

作者简介

杜若鹏，男，1983年生，硕士，助理研究员，研究方向：数字图书馆理论与技术、信息管理与信息系统。
鲜国建，男，1982年生，博士，副研究馆员，研究方向：数字图书馆理论与技术、信息管理与信息系统。
寇远涛，男，1982年生，博士，副研究馆员，通信作者，研究方向：数字图书馆理论与技术、信息管理与信息系统，E-mail：kouyuantao@caas.cn。

Improvement and Application of TF-IDF-CHI in Agricultural Science Text Feature Extraction

DU RuoPeng XIAN GuoJian KOU YuanTao
(Agricultural Information Institute of CAAS/Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, China)

Abstract: This paper is aimed at improving the lack of traditional TF-IDF method and verifying its effectiveness through text classification tests in the agricultural field. The improved method is called ImpTF_IDF_CHI which is to reconstruct the feature word weighting function by adding chi-square test values and weight correction factors. First, we use the ImpTF-IDF-CHI method, document frequency method, information gain method and the TF-IDF to perform the feature word extraction test. Then we use feature extraction words for test of text classification and judge the pros and cons based on the test. In all the test results, the best results were obtained using the ImpTF-IDF-CHI method. The Accuracy of naive Bayesian text classification using the ImpTF-IDF-CHI method is 94% and F1 value is 0.844. The experiment fully proves the effectiveness and advancement of the ImpTF-IDF-CHI method. The ImpTF-IDF-CHI method has the characteristics of high accuracy, good stability, strong subject representative in text feature extraction. This method can be applied to fields such as text categorization, feature expression and theme extraction.

Keywords: Feature Extraction; TF-IDF; Chi-Square Statistics; Text Categorization; Agricultural Science

（收稿日期：2019-07-06）