基于多元特征加权改进的TextRank关键词提取方法*

余本功 张宏梅 曹雨蒙
(合肥工业大学管理学院，合肥 230009)

摘要：现有的关键词提取方法从文档集或者单文档方面考虑词语的特征，很少考虑词语在单文档和文档集中的综合特征对关键词提取效果产生的影响，因此，本文提出多元特征加权的关键词提取方法。该方法通过Word2vec模型提取出词语在文档集中的语义关系特征与词语在单文档中的重要性特征，通过线性加权的方式计算出词语的综合影响力，用于改进TextRank模型中的概率转移矩阵，最后迭代计算选取排名靠前的词语作为文档的关键词。实验结果表明，从单文档和文档集两方面综合考虑词语的影响力，可以有效地改善关键词的提取效果。

关键词：关键词提取；TextRank ；Word2vec；多元特征加权
中图分类号：TP391
DOI：10.3772/j.issn.1673-2286.2020.03.006

随着互联网技术的发展和移动互联网的普及，以论坛、博客、头条和知乎社区为主流的媒介平台成为人们共享知识及发表言论的重要场所。这些平台上存储大量有用的非结构化文本信息，如何从承载这些信息的文本中提取对用户有用的信息成为了一个亟需解决的难题。自然语言处理技术为解决这一难题提供了强有力的技术支撑。

关键词提取作为自然语言处理的核心技术之一，对自然语言处理技术的应用有重要的作用。一方面，它为自然语言处理中的文本聚类分类、热点识别、创新评价研究，以及知识图谱和领域知识网络的构建打下了基础[1-5]；另一方面，关键词提取技术可以提高用户检索信息的效率和准确性[6]，帮助用户获得有用信息。如在中国知网上搜索学术论文时，用户一般会通过输入的关键词检索论文，而网页是通过与用户输入的关键词进行匹配，返回给用户相似度最高的文章。因此，对关键词提取进行研究是十分必要的，特别是在文本信息应用和信息检索等方面具有极其重要的现实意义和应用价值。

当前，关键词提取方法主要分为有监督方法和无监督方法。有监督方法是将关键词提取问题转化为分类问题或标注问题[7-8]，借助分类算法来判断候选词是否为关键词，由于语料集难以获取，有监督方法受到了制约，无监督方法因不需训练语料而受到了学者的广泛关注。因此，研究者们围绕无监督方法进行了大量的研究工作，来改进关键词提取的效果。

1 相关工作

随着自然语言处理技术的发展，研究者在关键词提取方法上不断创新，使得关键词提取方法更加成熟。关键词提取无监督方法主要包括3种，基于主题模型的关键词提取方法[9]、基于统计特征的关键词提取方法[10]和基于图模型的关键词提取方法[11-12]。在这3种方法中，基于主题模型的关键词提取方法仅考虑了主题信息，丢失了关键词本身的统计特征信息；基于统计特征的关键词提取方法容易忽略词语的语义信息；基于图模型的关键词提取方法没有考虑统计特征对词语节点权重的影响。因此，在对无监督方法进行研究的过

*本研究得到国家自然科学基金资助项目“基于制造大数据的产品研发知识集成与服务机制研究”（编号：71671057）资助。

程中，如何扬长避短是研究者思考的重点。

基于主题模型的关键词提取方法是通过主题模型中主题分布的性质对关键词进行提取。LDA是主题模型中应用最广的模型[13]，其核心思想是文档由多个主题构成，而主题是由词语的概率分布表示，只要找到文档的主题，然后选择主题中概率最大的词语，就可以将其作为文档的关键词。为进一步提高关键词提取效果，研究者在LDA模型上做了许多改进工作。朱泽德等[14]将LDA模型与TFIDF相融合，提出一种基于文档隐含主题的关键词提取新算法TFITF；李湘东等[15]在抽取粗粒度特征时，将词性、词语位置等权重扩展到LDA的生成模型中，增强了特征的表意性；邱明涛等[16]利用扩展的LDA模型调整词语的权值，弥补了LDA模型在话题解释性上的不足；杨春艳等[17]引入引用内容，建立Labeled-LDA模型，从语义层面分析了文档中词汇之间的关系，提高了主题提取的质量与准确率。

基于统计特征的关键词提取方法主要是利用词语在文档中的词权重、词语位置，以及词语的关联信息衡量词语是否能够作为文章的关键词。词权重主要包括词性、词频、词长等，而词语位置是指文档中词语的分布信息，如标题、段首、段尾；词语的关联信息涵盖互信息、均值、方差、TFIDF[18]等。在基于统计的关键词提取方法中，有学者对这些统计特征进行线性组合，通过计算得分来选取关键词，如著名的YAKE方法[19]综合影响词语得分的词频、长度、位置、首字母状态等信息对关键词重要性进行评分。但大多数学者是以TFIDF为核心，将常见的统计特征引入TFIDF中来改进关键词提取方法。罗燕等[20]通过齐普夫定律推导出同频词数统计规律，提出结合同频词数统计规律的TFIDF关键词提取方法；余本功等[21]使用词性和调节函数对TFIDF进行优化，并结合问答社区中多个用户特征综合计算词语的权重，获得更加精准的关键词；陈列蕾等[22]提出结合词语位置分布特征与基于Scopus数据库检索的TFIDF从英文摘要中提取关键词的方法。除此之外，为使TFIDF方法能够适合不同长度的语料，Florescu等[23]提出使用单词的算数平均值来代替IDF的对数取值计算方式，其效果优于传统的TFIDF方法。

基于图模型的关键词提取方法以TextRank[12]模型为代表，是目前应用最广泛的方法。该方法受PageRank的启发，通过词语间的共现关系建立网络图，然后进行迭代排序，抽取前N个词语作为关键词。由于该模型具有很强的适应性和扩展能力，因此，研究者在此基础上进行了改进，主要分为两个方面。一是在TextRank中引入统计特征属性。李航等[24]使用神经网络对词语平均信息熵、词性、位置进行加权计算，将得到的综合权重融合到TextRank中，以改进词语节点的初始权重及概率转移矩阵；Yan[25]将词语的上下文信息、词语位置、词语中心等特征引入图模型中，用于改进节点的初始权重；Biswas等[26]提出用于提取Twitter的KECNW模型，着重强调了图模型的集体节点权重取决于频率、中心性、邻居节点位置等参数；张莉婧等[27]通过引入G1赋权法对TFIDF、词语位置、词语长度和词性赋予不同权重并计算综合权重，对TextRank中的重启概率和概率转移矩阵进行改进；夏天[28]将覆盖影响力、位置影响力和频度影响力引入TextRank中，通过计算词语间的影响力，从而实现对概率转移矩阵的改进；刘竹辰等[29]在学者夏天的基础上对词语位置进行修改，提高了关键词提取的准确率。二是模型间的相互融合。在模型相互融合方面，主要是利用LDA模型和Word2vec[30]对TextRank进行改进。在LDA与TextRank结合方面，一些学者选择先对候选关键词进行聚类，然后将其作为图中的节点进行迭代计算，从而获得关键词，如TopicRank[31]模型与Multipartiterank[32]模型，后者是在前者的基础上进行改进，更加强调主题的多样性；然而，通过主题模型获得主题影响力或用词语相似性来改进TextRank中的概率转移矩阵和节点初始值占据了该方面研究的主流地位[33-35]。在Word2vec与TextRank结合方面，夏天[36]利用Word2vec生成词向量，对词向量进行聚类以获取聚类影响力，并与位置影响力、覆盖影响力进行加权，改进词语节点间的概率转移矩阵，提高了关键词提取的准确率；宁建飞等[37]利用Word2vec将文档集中的词语生成词向量，构建词语相似度矩阵，改进TextRank中节点的初始权重以及概率转移矩阵。

综上所述，利用多特征融合或模型结合的无监督方法在一定程度上提升了关键词提取的效果。如在基于图模型的关键词提取方法中，将Word2vec计算出的词语相似性引入图模型中，取得了一定的效果，但尚未考虑到词语在文档内的重要性特征。因此，本文在已有研究的基础上将文档内词语重要性与词语在文档集上的语义关系进行线性加权，将计算的词语综合影响力用于改进TextRank中的概率转移矩阵，通过强化词语节点的权值，达到改善节点间影响力的相互传递目的，从而提高关键词提取的效果。

2　多元特征加权的关键词提取方法

2.1　模型框架及概述

在现有关键词提取方法的基础上，本文提出一种多元特征加权改进的TextRank关键词提取方法（Improved TextRank Keyword Extraction Method Based on Multivariate Features Weighted，MFW-ITKEM），基本流程如图1所示。词语语义关系特征会在一定程度上影响词语节点间的关系；而文档内词语的重要性有利于反映词语是否为文档的核心部分，其权值越大，说明该词语越有可能是文档的关键词，文档内词语的重要性体现在词语节点出度特征、词语节点位置特征以及词语节点频次特征3个方面。本文通过线性加权的方式将词语语义关系、词语节点出度、词语节点位置和词语节点频次4个特征进行综合度量，计算词语的综合影响力，并将其用于改进候选关键词图中的概率转移矩阵，优化图中词语节点的迭代计算过程，获取文档内词语节点的权值，实现关键词的抽取。

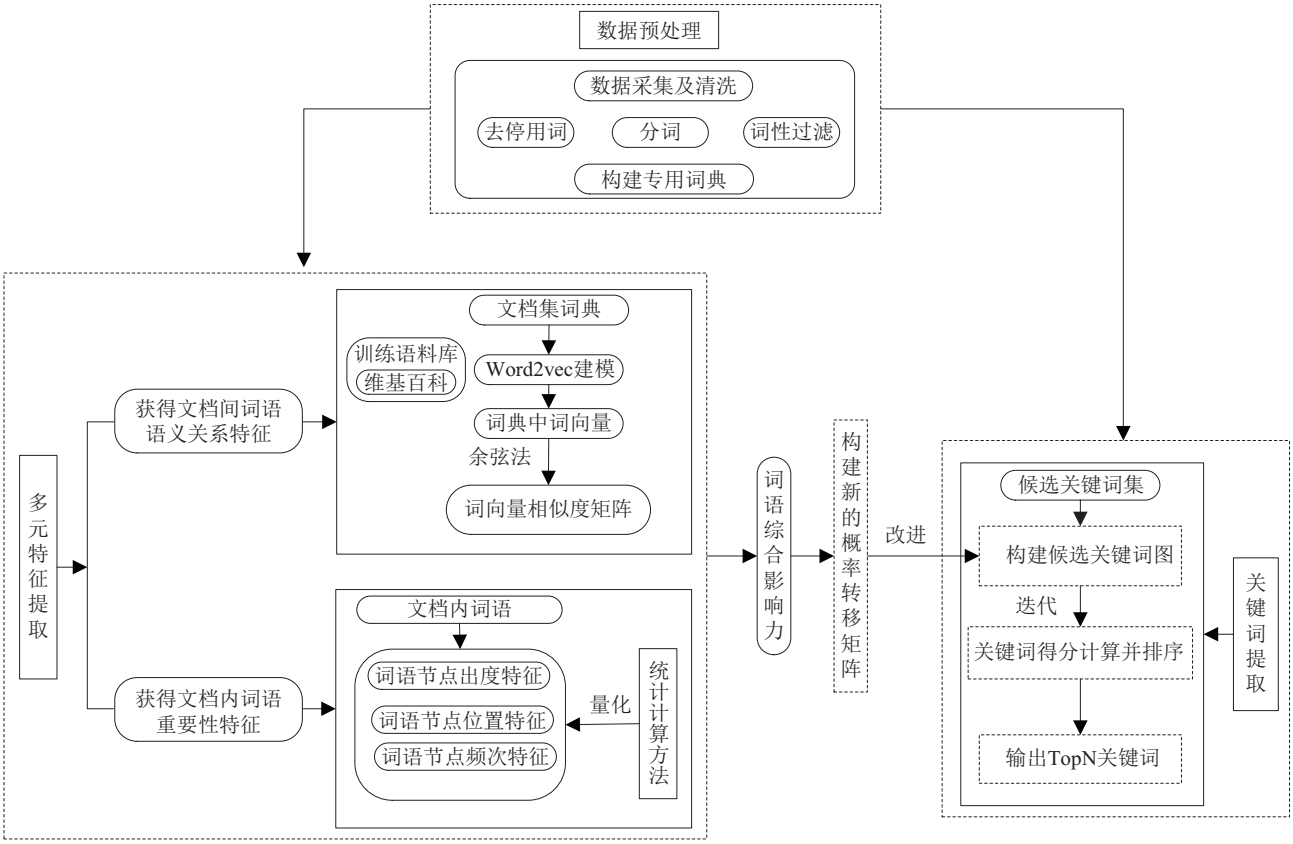

图1　多元特征加权改进的TextRank关键词提取方法流程

2.2　初始候选关键词图构建

根据TextRank原理，中文文档的候选关键词图的构建分为两个环节。对文档进行分句、分词，去停用词，保留词性为名词、动词、形容词、副词的词语，获得候选关键词集合T=[w_1，w_2，…w_m]。根据T中词语的相邻关系构建候选关键词图G=（V，E），V是图中的节点集合，由T中的候选关键词组成，E是相邻候选关键词之间的边集合。对于图中任意的两个相邻的节点，添加$v_i \to v_j$和$v_j \to v_i$两条边，将TextRank构建为一个有向图。节点v_i的TextRank值见公式（1）。

$$R(v_i) = d\sum_{j:v_j \to v_i} \frac{1}{OD(v_j)} R(v_j) + (1-d)\frac{1}{|V|} \quad (1)$$

其中，$OD(v_j)$表示节点v_j的出度，d是阻尼系数，默认取值为0.85，V是节点集合数。通过公式（1）进行迭代至收敛，即可获得文档中每个词的权重。

2.3 多元特征提取

多元特征用于计算节点的综合影响力，即词语在单文档中的重要性以及词语在文档集中存在的语义关系，多元特征导向见图2。词语在单文档中的重要性由词语节点的出度特征、词语节点的频次特征、词语节点在文档内的位置特征构成；而词语在文档集中的语义关系是通过Word2vec将词典表征为词向量，计算向量间的相似度来获得词语在语义方面的关系。因此，提出词语的综合影响力计算公式（2）。

$$CI(v_i, v_j)=\theta W(v_i, v_j)+\pi M(Sim(v_i, v_j)) \quad (2)$$

其中，θ和π是词语在文档内及文档集上特征的系数，$M(Sim(v_i, v_j))$为在文档集中词汇之间的相似度，θ和π在实验中取值都为0.5。

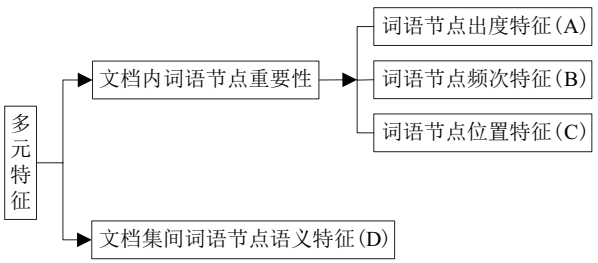

图2 多元特征导向图

2.3.1 文档内词语节点重要性的计算

在关键词图中，一个词语节点对其邻居节点的影响力是由该词语节点的重要性决定的，本文在已有研究基础上将词语节点在单文档中的特征分为词语节点出度、词语节点频次以及词语节点位置。令W表示节点的重要性，α、β、γ表示3个要素所占的比例，通过公式（3）计算词语节点的重要性。根据经验，参数设为α=0.34、β=0.33、γ=0.33。

$$W(v_i, v_j)=\alpha w_\alpha(v_i, v_j)+\beta w_\beta(v_i, v_j)+\gamma w_\gamma(v_i, v_j) \quad (3)$$

（1）词语节点出度特征。指词语节点v_i将其出度影响力均匀地分配给其他词语节点，旨在说明词语节点v_i与其相邻词语节点之间的关系。

（2）词语节点频次特征。指词语在文本中出现的次数，频次越高的词语其获得的影响权重越大。

（3）词语节点位置特征。指词语在文本中所处的位置，一般词语在标题中的重要性高于其他位置。如果词语在标题中出现，则取值为一个参数，参数取值范围$\phi\in[20, 30]$；如果在其他位置，则赋值为1。

2.3.2 文档集词语间的关系特征

（1）Word2vec模型。Word2vec是Google团队开源的将词表征成向量的工具[30]，主要包含跳字模型（skip-gram）和连续词袋模型（Continuous Bag-Of-Words Model，CBOW），如图3所示。CBOW模型和skip-gram模型都是由输入层、投影层和输出层组成，两个模型不同之处在于CBOW模型是利用上下文来预测中心词出现的概率，skip-gram模型是用中心词预测上下文出现在中心词附近的概率。与统计语言模型相比，Word2vec模型生成的词向量不仅解决了维度灾难问题，而且通过相似性的计算强化了词语之间的语义关系。因此，可以利用Word2vec训练得到的词向量计算相似性，来获得词语之间的语义关系。

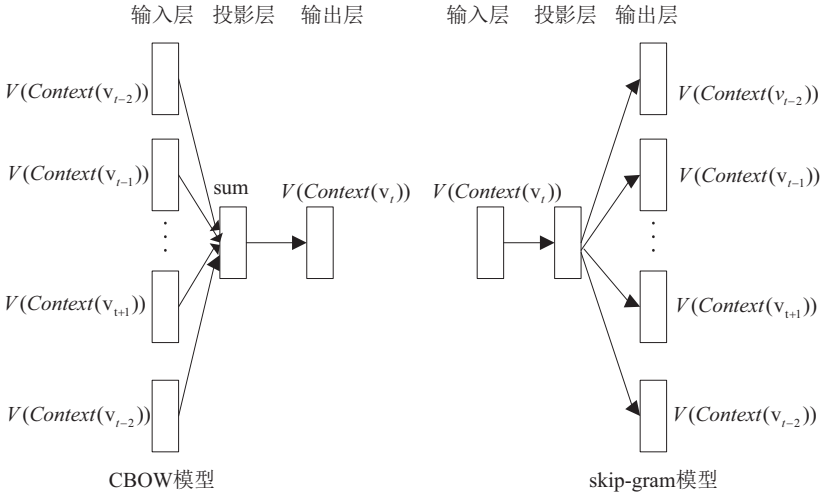

图3 Word2vec模型示意图

（2）文档集中词语节点语义特征的计算。为进一步研究文档集合中词语节点存在的语义关系对单文档中词语节点的影响力，需要对文档集合中词语节点的语义关系进行量化。本文利用Word2vec对文档集中的词语节点进行词向量表征，通过余弦公式计算词向量的相似性，获得词语节点在文档集中的语义关系特征。词语节点在文档集中语义关系的计算需要在构建候选关键词图前完成，一般分为：①对给定的文档集进行分句、分词，获得词汇集S_1，S_1由N个子词汇集组成，每组子词汇集对应一篇文档；②对词汇集S_1去停用词，保留词性为名词、动词、形容词及副词的词语，进行合并生成词典D=[w_1，w_2，…w_n]，该词典是关键词图中所有候选关键词的全集；③利用训练好的Word2vec对词典D进行词向量表达，得到D的词向量。

通过词典中词语的词向量，利用余弦公式计算词典D中词语的相似度，获得词语在文档集中所存在的语法关系，故词典中词语的相似度计算见公式（4）。

$$Sim(u_i, c_j) = \cos\theta = \frac{u_i \cdot c_j}{\|u_i\| \|c_j\|} \tag{4}$$

其中，c_j是目标文档句中的第j个词，u_i是源文档句中第i个词，u_i与c_j均为词向量。

假设词典的大小为n，则可以得到一个n×n的词语相似度矩阵，见公式（5）。

$$M(Sim(v_i, v_j)) = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix} \tag{5}$$

其中，$M(Sim(v_i, v_j))$表示词典的相似度矩阵，w_{ij}表示词典中词语节点v_i与v_j的相似度。

2.4 概率转移矩阵的计算与关键词提取

传统的词图中，词语节点的权重依赖于相邻词语节点的贡献度。为了对TextRank进行改进，本文引入词语综合影响力对概率转移矩阵进行优化，提高关键词提取的准确性。词语节点的权重由两个因素所决定：一是词语节点本身的重要性，代表词语在文档内部结构中的作用，一般设定为1，在迭代过程中由相邻词语节点的分值进行调整，记为$TR(v_i)$；二是由词语在单文档中重要性和词语在文档集中语义关系所构成的分值，表示词语的综合影响力。因此，定义新的节点重要性迭代计算公式（6）。

$$TR(v_i) = d\left(\delta \sum_{j:v_j \to v_i} \frac{CI(v_j, v_i)}{OCI(v_j, v_i)} TR(v_j) + \sigma \sum_{j:v_j \to v_i} \frac{1}{OD(v_j)} R(v_j)\right) + (1-d)\frac{1}{|V|} \tag{6}$$

其中，δ+σ=1，分别表示综合影响力和投票贡献影响力，取值均为0.5，$OCI(v_j, v_i) = \sum_{j:v_i \to v_j} CI(v_i, v_j)\ R(v_j)$为节点$v_j$的权重，V为词汇节点集合，这里$CI(v_i, v_j)$由公式（2）计算得到。

在迭代计算前，构建词语节点间的概率转移矩阵M，见公式（7）。

$$M = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix} \tag{7}$$

其中，w_{ij}表示节点v_j的影响力转移到其他节点的概率，每列概率之和为1。w_{ij}的权重可以通过公式（8）计算得到。

$$w_{ij} = \delta \sum_{j:v_j \to v_i} \frac{CI(v_j, v_i)}{OCI(v_j, v_i)} + \sigma \frac{1}{O(w_j)} \tag{8}$$

在引入概率转移矩阵M之后，令B_i表示一次迭代的结果，则迭代公式可以转化为公式（9）。

$$B_i = d \times M \times B_{i-1} + (1-d) \times \frac{e}{k} \tag{9}$$

其中，e为维数为k的单位向量。通过公式（9）进行迭代计算，当两次的计算结果差异小于0.001时，表明计算结果达到收敛状态。最后对所有的词语节点权重降序排列，将排名靠前的N个词作为关键词。

3 实验及分析

为了对提出的基于多元特征加权改进的TextRank关键词提取方法进行验证，本文选取专利文本摘要进行实证分析，并与其他学者提出的方法进行对比，分析关键词提取效果。专利文本是一种特殊的文本，它以精简的方式存储着最新的科学技术，通过对专利文本进行挖掘，能够快速地捕捉到技术前沿，为企业提供一定的参考价值，激发企业的创新能力。本文选取制造行业的汽车专利文本，提取汽车技术前沿的关键词，为人们快速了解最新技术提供便捷。

3.1 数据集获取及处理

本文数据来自国内文献检索平台中国知网，选择高级检索方式，以“申请人=安徽江淮汽车股份有限公司”为检索条件，选择公开日期为2016年4月20日—2017年2

月15日共1 038条文本，剔除文本摘要篇幅小于150字的专利文本，共得到843条文本，对得到的843条专利文本进行数据清洗。剔除申请号、专利号、申请日、公开号等结构化信息，保留专利文本的标题和摘要文本，将每条专利摘要和标题看作一个文档存储在xlsx文件中，为解决专利文本摘要中没有标准的关键词问题，笔者采用人工标注的方式在每条专利摘要中标注10个关键词作为标准关键词，与算法自动提取出的关键词进行对比分析。

本实验使用Python自带的结巴分词工具对数据进行分词，通过停用词词典将通用词以及标点符号过滤掉，进行词性标注，在团队所构建的2万条汽车专用词典的基础上加入未收录的汽车专用术语，共引入41 891个汽车术语，以此来提高分词效果。

本文使用维基百科语料作为Word2vec训练集，完成词向量的训练，利用训练好的参数对专利文本进行词向量的表达。

3.2 对比实验设置及结果分析

本文采用的数据语料是江淮专利文本摘要和标题，为了对关键词的提取效果进行评估，本文选择准确率（P值）、召回率（R值）和F值3个指标。

本文提出的多元特征的关键词提取方法是将文档外部信息与文档内部信息相结合，对专利文本摘要进行关键词提取研究，提取的关键词取值范围为[3-10]。本文设置了两类对比实验，第一类是特征组合实验，通过对不同特征的融合，说明特征的叠加能够有效提升关键词的提取效果；第二类是不同关键词算法之间的比较，旨在表明本文提出的算法优于其他算法。在各性能对比图表中仅显示关键词个数为3、5、7、10的准确率、召回率以及F值。

3.2.1 特征组合

通过单个特征进行分析，以TextRank模型为基准，分别加入表示词语在文档内的重要性特征，即词语节点出度特征（A）、词语节点位置特征（B）、词语节点频次特征（C）和词语在文档集间的语义关系特征（D），依次对模型中的初始概率转移矩阵进行改进。从图4可以看出，在单个特征中，B的准确率、召回率和F值均高于其他特征，而D是单个特征中提取效果最差的，原因在于仅考虑文档集间的语义关系，忽略了单文档的词语节点出度、词语在文本中的位置以及频次产生的影响，所以对于提取单文档关键词来说，准确率、召回率、F值均较差。

为更好地说明特征对实验结果的影响，本文在单个特征的基础上将不同特征进行组合，如图5所示。

图5的实验结果显示，将词语在文档集上的语义特征与词语在文档内的重要性相融合，其准确率、召回率和F值均大于词语位置特征与其他单个特征相组合的效果，即A+B+C+D的关键词提取性能要胜于其他特征的组合性能。

3.2.2 算法比较

不同算法的对比在这里分为两组，第一组是将本文提出的MFW-ITKEM算法与传统的TextRank和TFIDF算法作对比，第二组是将本文提出的MFW-ITKEM算法与其他研究者提出的算法作比较。

第一组实验包括以下3种算法。

（1）TextRank。通过滑动窗口构建共现网络，迭代

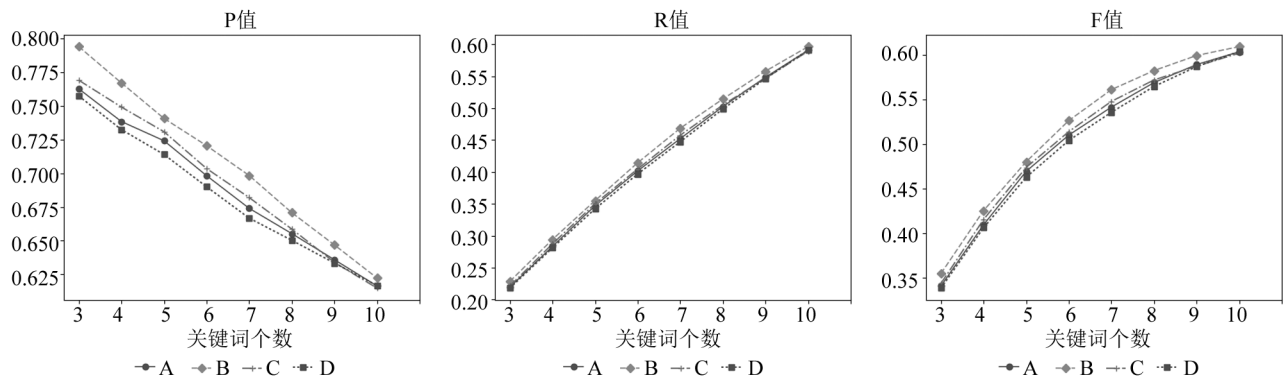

图4 单个特征性能对比

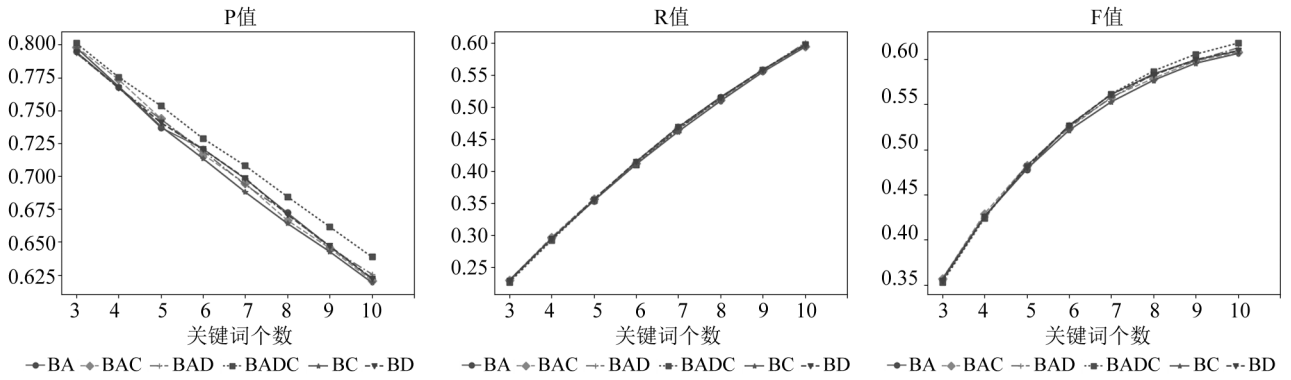

图5　组合特征性能对比

计算词语重要性，输出排名靠前的词语作为关键词[12]。

（2）TFIDF。词频逆文档算法，在基于词频的关键词提取算法中，既考虑了词语在单篇文档中词频的大小，也将词语对整个文档集的区分能力纳入计算中，这是一种经典算法。

（3）MFW-ITKEM。本文提出的方法在已有研究的基础上，将词语在文档内的特征与词语在文档集上的语法特征相融合，改进TextRank中的概率转移矩阵，进行关键词的抽取。

表1和图6是本文提出的MFW-ITKEM算法与经典的TextRank和TFIDF算法的对比，可以看到，本文提出的方法在提取效果上均优于经典方法，其优势在于将节点在文档内的特征与节点在文档集上的语义特征引入TextRank方法中。

表1　与经典算法的效果对比

特　征	准确率（P值）				召回率（R值）				F值			
	3	5	7	10	3	5	7	10	3	5	7	10
TextRank	0.542 5	0.455 1	0.391 7	0.320 6	0.156 0	0.218 2	0.262 9	0.307 0	0.242 4	0.294 9	0.314 6	0.313 7
TFIDF	0.719 7	0.709 2	0.669 7	0.593 0	0.207 0	0.340 0	0.449 5	0.568 5	0.321 6	0.459 6	0.537 9	0.580 5
MFW-ITKEM	0.786 1	0.738 5	0.692 9	0.623 6	0.226 1	0.354 0	0.465 0	0.597 9	0.351 2	0.478 6	0.556 6	0.610 4

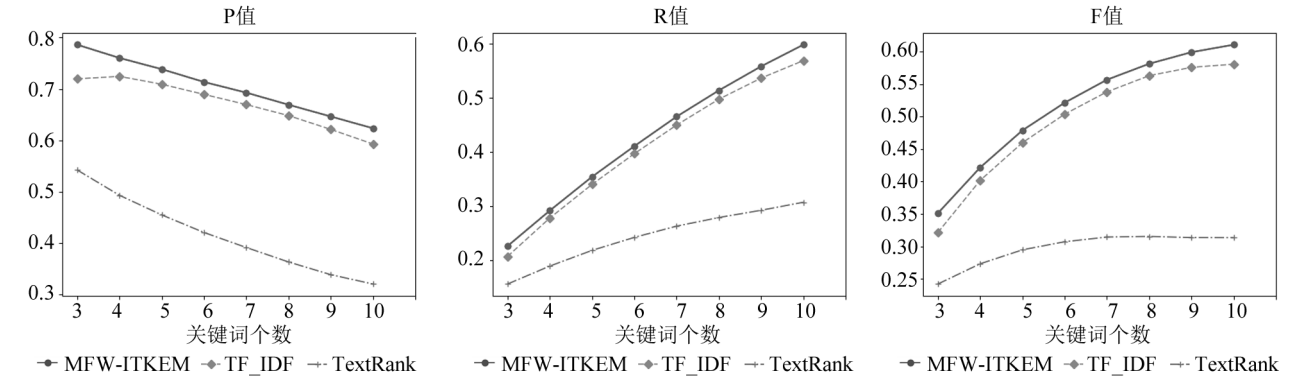

图6　3种方法性能对比

第二组实验包括以下5种算法。

（1）T1。Word2vec算法，通过词向量计算词语的相似性，然后聚类得到关键词[38]。

（2）T2。将词向量进行聚类，将外部知识融入TextRank的计算中[36]。

（3）T3。将词语的位置信息和词距融入词图模型中，来提升单文档的关键词提取效果[29]。

（4）T4。将Word2vec与TextRank相结合，将提取出的词向量作为TextRank的输入，采用了模型结合的方式[37]。

（5）MFW-ITKEM。本文提出的算法。

表2和图7的统计结果显示，横向上比较来看，在关

键词个数较小时，5种方法的准确率和F值基本相等，但是随着关键词个数的增加，MFW-ITKEM的准确率和F值都有所提高，且高于其他研究者的方法，表明MFW-ITKEM方法在关键词提取方面有更明显的优势。

表2 等个数关键词下不同方法性能对比

特征	准确率（P值）				召回率（R值）				F值			
	3	5	7	10	3	5	7	10	3	5	7	10
T1	0.768 0	0.734 1	0.682 0	0.616 1	0.220 9	0.351 9	0.457 7	0.590 7	0.343 1	0.475 7	0.547 8	0.603 1
T2	0.788 7	0.731 5	0.689 9	0.618 3	0.226 9	0.350 7	0.463 1	0.592 8	0.352 4	0.474 1	0.554 2	0.605 3
T3	0.784 0	0.729 0	0.674 9	0.593 3	0.225 5	0.349 5	0.453 0	0.568 9	0.350 2	0.472 4	0.542 1	0.580 8
T4	0.769 6	0.715 9	0.651 0	0.564 2	0.221 4	0.343 2	0.456 9	0.540 9	0.343 8	0.464 0	0.523 0	0.552 3
MFW-ITKEM	0.786 1	0.738 5	0.692 9	0.623 6	0.226 1	0.354 0	0.465 0	0.597 9	0.351 2	0.478 6	0.556 6	0.610 4

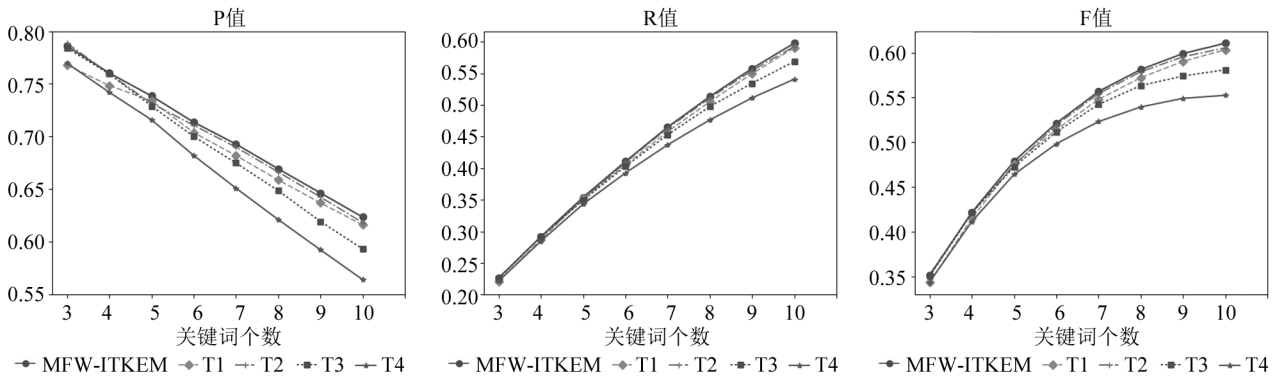

图7 5种方法性能对比

根据F值从纵向上分析，在关键词数量为5、7、10的情况下，MFW-ITKEM在效果上均优于其他4种算法。具体来说，关键词数目为5时，5种算法的排序为T4<T3<T2<T1<MFW-ITKEM；而在关键词为7和10时，5种算法的排序基本稳定为T4<T3<T1<T2<MFW-ITKEM。将T2方法同T1、T4方法进行对比，这3种方法的共同之处在于都是将word2vec模型与TextRank模型相融合，不同之处在于T2在考虑了词语在文档集上语义关系的基础上加入了词语在单文档中的统计特征，克服了仅使用词向量计算相似度或是通过聚类获得关键词的不足；T3提出的词位置分布加权方法是在李航等[24]研究基础上提出的，而效果比T2差的原因是考虑了词语在单文档中的重要性，忽略了外部信息对词语的影响力。本文提出的MFW-ITKEM同T2及T4的算法相比，说明通过Word2vec计算词语之间的相似性来获得的文档间的语义信息与单文档的词语的本身特征相结合改进图模型中的节点跳转概率，能够更好地提高关键词的抽取效果，比使用单一的模型或者特征以及聚类更有优势。

通过两类实验结果分析，验证了本文提出的方法在使用词向量获取文档集间的词语关系的基础上引入文档内的相邻词语的出度特征、频率特征和词语位置特征，能够有效地提高关键词的提取效果，比其他学者提出的仅考虑单文档的统计特征或是通过词向量聚类的算法更有优势。

4 结语

本文在基于图模型的关键词提取方法的基础上，综合考虑词语在单文档中的重要性和其在文档集中的语义关系，提出将这两部分通过线性加权的方式融合来计算词语的综合影响力，并以此来改进TextRank方法的概率转移矩阵，实现图中节点的权重计算并获得关键词，经过实验验证，该算法提高了关键词的提取效果。

本文所提出的算法也存在一些不足。训练Word2vec的语料均来自维基百科，尚未涵盖汽车专业领域的一些术语，造成在使用单特征提取关键词的实验中效果并不理想。后续研究将考虑使用汽车领域的语料集来训练Word2vec模型，并进一步扩大关键词提取的文本，

且将该方法与具体的应用领域相结合，如热点分析、创新评价以及主题演化方面，为用户提供更有价值的参考。

参考文献

[1] 毛太田，蒋冠文，李勇，等. 新媒体时代下网络热点事件情感传播特征研究 [J]. 情报科学，2019，37（4）：29-35，96.

[2] 王健，张俊妮. 统计模型在中文文本挖掘中的应用 [J]. 数理统计与管理，2017，36（4）：609-619.

[3] 马宗国，尹圆圆. 我国研究联合体研究的知识图谱分析——基于1992—2017年中国知网期刊文献 [J]. 科技管理研究，2019，39（5）：246-250.

[4] 余本功，陈杨楠，杨颖. 基于主题模型和专利数据的技术创新评价研究 [J]. 现代情报，2019，39（1）：111-117，168.

[5] 赵汝南，常志远，姜博，等. 基于网络演化的领域知识发展趋势研究 [J]. 数字图书馆论坛，2016（3）：24-29.

[6] 温有奎. 信息检索系统的关联关键词推荐研究 [J]. 数字图书馆论坛，2016（4）：11-14.

[7] 赵京胜，朱巧明，周国栋，等. 自动关键词抽取研究综述 [J]. 软件学报，2017，28（9）：2431-2449.

[8] 常耀成，张宇翔，王红，等. 特征驱动的关键词提取算法综述 [J]. 软件学报，2018，29（7）：2046-2070.

[9] WEI H X，GAO G L，SU X D. LDA-Based Word Image Representation for Keyword Spotting on Historical Mongolian Documents [C] //Neural Information Processing（ICONIP）. Springer，2016：432-441.

[10] 傅柱，王曰芬，陈必坤. 国内外知识流研究热点：基于词频的统计分析 [J]. 图书馆学研究，2016（14）：2-12.

[11] BOUDIN F. A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction [C] //Proceedings of the 6th International Joint Conference on Natural Language Processing. Nagoya：Asian Federation of Natural Language Processing，2013：834-838

[12] MIHALCEA R，TARAU P. TextRank：Bringing Order into Texts [C] //Proceedings of Conference on Empirical Methods in Natural Language Processing，Stroudsburg：ACL，Barcelona. 2004：404-411.

[13] BLEI D M，NGA Y，JODAN M I.Latentdirichlet allocation [J]. The Journal of Machine Learning Research，2003，3：993-1022.

[14] 朱泽德，李淼，张健，等. 一种基于LDA模型的关键词抽取方法 [J]. 中南大学学报（自然科学版），2015，46（6）：2142-2148

[15] 李湘东，巴志超，黄莉. 一种基于加权LDA模型和多粒度的文本特征选择方法 [J]. 现代图书情报技术，2015（5）：42-49.

[16] 邱明涛，马静，张磊，等. 基于可扩展LDA模型的微博话题特征抽取研究 [J]. 情报科学，2017，35（4）：22-26，31.

[17] 杨春艳，潘有能，赵莉. 基于语义和引用加权的文献主题提取研究 [J]. 图书情报工作，2016，60（9）：131-138，146.

[18] PAIK J H. A novel TF-IDF weighting scheme for effective ranking [C] //Proceedings of the 36th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM，2013：343-352.

[19] CAMPOS R，VÍTOR M，PASQUALI A，et al. YAKE! Collection-Independent Automatic Keyword Extractor [C] //In Advances in Information Retrieval-40th European Conference on Information Retrieval. Springer ECIR 2018，Lecture Notes in Computer Science，Grenoble，France. Cham，2018：806-810.

[20] 罗燕，赵书良，李晓超，等. 基于词频统计的文本关键词提取方法 [J]. 计算机应用，2016，36（3）：718-725.

[21] 余本功，李婷，杨颖. 基于多属性加权的社会化问答社区关键词提取方法 [J]. 图书情报工作，2018，62（5）：132-139.

[22] 陈列蕾，方晖. 基于Scopus检索和TFIDF的论文关键词自动提取方法 [J]. 南京大学学报（自然科学），2018，54（3）：604-611.

[23] FLORESCU C，CARAGEA C. A New Scheme for Scoring Phrases in Unsupervised Keyphrase Extraction [C] // Proceedings of the Advances in Information Retrieval-39th European Conference on Information Retrieval.ECIR 2017，Lecture Notes in Computer Science Aberdeen，UK，2017.

[24] 李航，唐超兰，杨贤，等. 融合多特征的TextRank关键词抽取方法 [J]. 情报杂志，2017，36（8）：183-187.

[25] YAN Y. A Graph-based approach of automatic key phraseextraction [J]. Procedia Computer Science，2017，107：248-255.

[26] BISWAS S K，BORDOLOI M，SHREYA J. A graph based keyword extraction model using collective node weight [J]. Expert Systems with Applications，2018，97：51-59.

[27] 张莉婧，李业丽，曾庆涛，等. 基于改进TextRank的关键词抽取算法 [J]. 北京印刷学院学报，2016，24（4）：51-55.

[28] 夏天. 词语位置加权Text Rank的关键词抽取研究 [J]. 现代图书情报技术，2013（9）：30-34.

[29] 刘竹辰，陈浩，于艳华，等. 词位置分布加权TextRank的关键词

提取 [J]. 数据分析与知识发现，2018，2（9）：74-79.

[30] MIKOLOVT，CHEN K，CORRADO G，et al. Efficient Estimation of Word Representations in Vector Space [C] // Proceedings of the 2013 International Conference on Learning Representations，ICLR 2013，Workshop Track，Scottsdale，Arizona，USA. 2013：1-12.

[31] BOUGOUIN A，BOUDINF，BÉATRICE D. TopicRank：Graph-Based Topic Ranking for Keyphrase Extraction [C] // Proceedings of the 6th International Joint Conference on Natural Language Processing，IJCNLP 2013，Nagoya，Japan2013：543-551.

[32] BOUDINF. Unsupervised key phraseextraction with multipartite graphs [C] //Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，NAACL HLT，Association for Computational Linguistics，NewOrleans：June 1-6，2018，2：667-672.

[33] STERCKX L，DEMEESTER T，DELEU J，et al. Creation and evaluation of large keyphrase extraction collections with multiple opinions [J]. Language Resources and Evaluation，2017，52：503-532.

[34] 顾益军，夏天. 融合LDA与TextRank的关键词抽取研究 [J]. 现代图书情报技术，2014（7/8）：41-47.

[35] 刘啸剑，谢飞，吴信东. 基于图和LDA主题模型的关键词抽取算法 [J]. 情报学报，2016，35（6）：664-672.

[36] 夏天. 词向量聚类加权TextRank的关键词抽取 [J]. 数据分析与知识发现，2017，1（2）：28-34.

[37] 宁建飞，刘降珍. 融合Word2vec与TextRank的关键词抽取研究 [J]. 现代图书情报技术，2016（6）：20-27.

[38] 李跃鹏，金翠，及俊川. 基于word2vec的关键词提取算法 [J]. 科研信息化技术与应用，2015，6（4）：54-59.

作者简介

余本功，男，1971年生，博士，教授，研究方向：信息系统、机器学习。
张宏梅，女，1994年生，硕士，通信作者，研究方向：数据挖掘、自然语言处理，E-mail：18856002708@163.com。
曹雨蒙，女，1994年生，硕士，研究方向：机器学习、自然语言处理。

Improved TextRank Keyword Extraction Method Based on Multivariate Features Weighted

YU BenGong ZHANG HongMei CAO YuMeng
(School of Management, Hefei University of Technology, Hefei 230009, China)

Abstract: Existing keyword extraction methods take into account the characteristics of words from the document set or single document, and rarely comprehensively considered the impact of the comprehensive features of words in single document and document set on the keyword extraction effect. This paper proposed a multi-feature weighted keyword extraction method. This method used the Word2vec model to extract the semantic relationship characteristics of words in the document set, and the importance characteristics of words in a single document to calculate the comprehensive influence of the words in a linear weighting manner, which was used to improve the probability transition matrix in the TextRank model. Finally, iterative calculation selected the top-ranked words as the keywords of the document. Experimental results show that comprehensive consideration of the influence of words from both a single document and a document set can effectively improve the effect of keyword extraction.

Keywords: Keyword Extraction; TextRank; Word2vec; Multivariate Feature Weighting

（收稿日期：2020-02-28）