计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700041

融合语义特征的关键词提取方法

高 楠 李利娟 李 伟 祝建明
浙江工业大学计算机科学与技术学院 杭州 310023

摘 要 关键词提取被广泛应用于文本挖掘领域,是文本自动摘要、自动分类、自动聚类等研究的基础。因此,提取高质量的关键词具有十分重要的研究意义。已有关键词提取方法研究中大多仅考虑了部分文本的统计特征,没有考虑词语的隐式语义特征,导致提取结果的准确率不高,且关键词缺乏语义信息。针对这一问题,文中设计了一种针对词语与文本主题之间的特征进行量化的算法。该算法首先用词向量的方法挖掘文本中词语的上下文语义关系,然后通过聚类方法抽取文本中主要的语义特征,最后用相似距离的方式计算词语与文本主题之间的距离并将其作为该词语的语义特征。此外,通过将语义特征与多种描述词语的词频、长度、位置和语言等特征结合,文中还提出了一种融合语义特征的短文本关键词提取方法,简称 SFKE 方法。该方法从统计信息和语义层面分析了词语的重要性,从而可以综合多方面因素提取出最相关的关键词集合。实验结果表明,相比 TFIDF,TextRank,Yake,KEA 和 AE 等方法,融合多种特征的关键词提取方法的性能有了明显的提升。该方法与基于有监督的 AE 方法相比,*F-Score* 提升了 9.3%。最后,用信息增益的方法对特征的重要性进行评估,结果表明,添加语义特征后模型的 *F-Score* 提升了 7.2%。

关键词:文本挖掘;统计特征;语义特征;支持向量机;分类模型
中图法分类号 TP391

Keywords Extraction Method Based on Semantic Feature Fusion

GAO Nan,LI Li-juan,Wei-william LEE and ZHU Jian-ming
School of Computer Science and Technology,Zhejiang University of Technology,Hangzhou 310023,China

Abstract Keyword extraction is widely used in the field of text mining,which is the prerequisite technology of text automatic summarization,classification and clustering. Therefore,it is very important to extract high quality keywords. At present,most researches on keyword extraction methods only consider some statistical features,but not the implicit semantic features of words,which leads to the low accuracy of extraction results and the lack of semantic information of keywords. To solve this problem,this paper designed a quantification method of the features between words and text themes. First,the word vector method is used to mine the context semantic relations of words. Then the main semantic features of the text is extracted by clustering. Finally,the distance between the words and the topic with the similar distance method is calculated. It is regarded as the semantic features of word. In addition,by combining the semantic features of word with the features of word frequency,length,location,language and other various description of words,a keywords extraction method of short text with semantic features was proposed,namely SFKE method. This method analyzes the importance of words from the statistical and semantic aspects,thus can extract the most relevant keyword set by integrating many factors. Experimental results show that the keyword extraction method integrating multiple features has significant improvement compared with TFIDF,TextRank,Yake,KEA,AE methods. The *F-Score* of this method has improved by 9.3% compared with AE. In addition,this paper used the method of information gain to evaluate the importance of features. The experimental results show that the *F-Score* of the model is increased by 7.2% after adding semantic feature.

Keywords Text mining,Statistical features,Semantic features,Support vector machine,Classification model

1 引言

文档关键词抽取,也称为关键词提取或关键词标注,是从文本中把与该文本所表达的意义最相关的一些词或短语抽取出来[1]。学术关键词向读者提供了对文本最简洁的描述,能基本反映文本主题。读者可以通过关键词在短时间内获得文本主要的研究内容,从而节省信息检索的时间,提升获得目标文本的效率。关键词提取技术被广泛应用于文本检索、摘要

到稿日期:2019-06-04 返修日期:2019-08-19 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61702456);浙江省科技厅公益科项目(2017C33108)
This work was supported by the National Natural Science Foundation of China (61702456) and Zhejiang Public Welfare Technology Research Program(2017C33108).
通信作者:高楠(gaonan@zjut.edu.cn)

生成、文本分类[2-3]等领域，具有十分重要的研究价值。

现有的研究多基于词语的统计特征[4]对关键词进行提取，少数的研究基于词语的语义信息挖掘词语与文本主题之间的语义联系。而基于词语语义的研究也多集中于将基于图的方法与语义信息结合[5]，或是将关键词提取视为一个序列标注问题[6]，利用神经网络的方法提取词语的上下文本关系并结合条件随机场的方法对关键词进行标注。而将词语的隐式语义关系与词语的多种统计信息进行结合的方法，还需要更深入的研究。

针对以上问题，本文提出了一种融合语义特征的关键词提取算法。该算法对词语和文本主题之间的语义相关性进行量化并将其作为词语的语义特征；通过结合语义特征和词语的其他特征创建候选关键词的特征集合，并用该集合组成的样本数据训练关键词分类模型，实现对文本关键词的预测。

本文的主要贡献有以下几点：1）设计了一种词语与主题之间相关性的量化方法，深层次挖掘了词语的语义信息；2）综合考虑候选关键词的词频、长度、位置和语言特征等多种因素，筛选了 14 个特征作为关键词的属性，设计了一种融合语义特征的关键词提取方法，并从多方面对词语的重要性进行评估；3）利用信息增益的方法量化并验证了语义特征的重要性，并通过多组特征组合对比实验进一步验证了结论的正确性；4）在学术文本数据集合上将本文方法与 5 种关键词提取方法进行对比。实验结果表明，当系统提取前 5 个关键词并且聚类个数设为 4 时，本文方法相比其他方法有更好的表现。

2　相关工作

根据数据的标注情况，可以将关键词提取技术分为无监督学习方法和有监督学习方法。基于无监督学习的方法[4]将关键词提取看作一个排序问题，这类方法主要利用词语的统计特征或语法规则对其进行重要性排序。经典的基于无监督的关键词提取算法包括 TFIDF[4]，TextRank[5]，YAKE[7]等。TFIDF 算法利用候选关键词的词频和逆文档频率信息对词语的重要性进行排序；TextRank 算法则是利用词语共现情况来评估候选关键词的重要性；而 YAKE 是近期提出的综合考虑了词语的词频、长度、位置、首字母状态等信息的关键词重要性评分算法。这类利用统计信息设计评分公式的方法较为简单，不需要标注语料的训练，且具有很好的适用性。然而，评估算法的设计主要依靠专家的经验总结，存在很强的主观性，且缺乏对语义信息的考量。

基于有监督学习的方法将关键词提取看作一个二分类问题。算法首先从文本中选择出候选关键词集合，然后从文本中提取候选关键词的多种特征，最后用该特征集合训练关键词分类模型。这类算法的研究重点集中在特征的提取和分类模型的训练两个方面。常用的特征主要包含描述词语长度、频率、位置等相关的统计特征，以及描述词语词性、句法结构等相关的语言特征。常用的分类算法主要有决策树[8]、朴素贝叶斯分类器[9]、支持向量机[10]、最大熵模型[11]、隐马尔可夫模型[12]、条件随机场模型[13]以及神经网络模型[14]等。不同的特征集和分类算法会产生不同的模型。例如，Frank[9]首次将关键词提取任务看作二分类问题，用 TFIDF 和词语首次出现的属性作为特征，结合朴素贝叶斯模型提出了 KEA 算法。Aquino[15]提取了词语的频率、首次出现位置、最后一次出现位置、是否为专有名词等 20 个属性值，训练出了新的关键词提取算法，简称 AE 算法。

为进一步提升关键词提取的准确率，研究者利用共现矩阵或语言模型等方法从文本的主题信息中挖掘词语的语义特征。Mikolov[16]提出词嵌入模型是一种新的词向量表示方式，为词语之间的关系建立了语义联系。借助词向量，Chen[10]提出了改进的 TextRank 方法，该方法基于预先训练的词向量，从科学出版物中提取和生成关键词。Zhang[14]针对社交短文本信息，提出利用深度递归神经网络（Recurrent Neural Network，RNN）学习关键词及其上下文信息，从而提取推文中的关键词。这类利用词向量提取关键词的方法都取得了显著的成果。

综上，本文借助词向量工具设计了一种对词语和文本主题间语义相关性进行量化的方法，并结合词语的统计特征从多方面对词语的重要性进行评估，以提升系统的准确率。

3　融合语义特征的关键词提取

语义信息在一定程度上描述了词语所在文本的上下文信息，能够更加容易地挖掘词语间的语义相关性并从语境中理解词语所表达的主题，更加符合人的主观意识。为了提升关键词提取的准确率，并使提取的关键词与文本的主题信息更加贴合，本文通过提取词语的语义特征并将其与多种统计特征结合，提出了一种融合语义特征的关键词提取方法，简称 SFKE 方法。该方法主要由 3 部分组成：1）文本预处理；2）特征提取；3）构建分类模型。主要工作集中在语义特征提取部分。图 1 显示了本文关键词提取的总体流程。

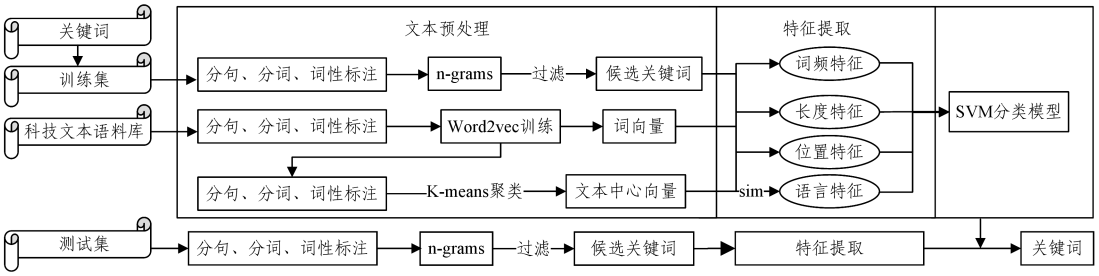


图 1　关键词提取框架图

Fig. 1　Framework of keyword extraction

3.1　文本预处理

文本预处理的目的是得到候选关键词集合。首先需要对句子进行分割，然后对句子中的词语进行分词，并标注其词性信息。关键词可能是一个词语，也可能是多个词语。本文参

考文献[17]中指出的词组中包含的词语一般不会超过3个，利用n-grams($n=1,2,3$)模型抽取词组，得到词组集合。为了减少候选关键词的数量，本文通过对关键词词性、词频信息进行统计分析，设计了对词语集合进行筛选的3条规则。

1)删除词性为代词、介词、数词及拟声词的词语。

2)删除词组中不是以名词或形容词开头或结尾的词语。

3)删除包含在停用词表[18]中的词语。

3.2 特征提取

关键词既要包含文本的主题相关性，又要反映词语的重要性[17]。本文选择了一些已被证明有效的统计特征和语言特征[15,19]，从词频、长度、位置和语言等维度衡量词语的重要性。

此外，文中还提出了一种新的语义特征，通过该特征衡量词语与文本主题之间的相关性挖掘词语的隐式语义信息。本文共提取了词语的14个特征作为候选关键词的特征属性，包括4个词频特征、2个长度特征、5个位置特征和3个语言特征。

相关特征的类型及描述如表1所列。

表1 特征类型及描述

Table 1 Feature types and descriptions

Type	Id	Name(Abbreviation)	Description
词频特征	1	词频(TF)	$tf_{i,j}=\frac{n_{i,j}}{\lvert j\rvert}$，$n_{i,j}$表示词语$i$在文档$j$中出现的次数，$\lvert j\rvert$表示文档$j$中包含的数目
	2	逆文档频率(IDF)	$idf_i=\log_2\frac{\lvert D\rvert}{df_i}$，$\lvert D\rvert$表示语料$D$中包含的文本数量，$dfi$表示语料集中包含词语$i$的文本数量
	3	术语频率逆文档频率(TFIDF)	$tfidf_{i,j}=\frac{n_{i,j}}{\lvert j\rvert}\times\log_2\frac{\lvert D\rvert}{df_i}$，参数说明如上
	4	标题词频(TTF)	$ttf_{i,j}=\frac{n_{i,j\_t}}{\lvert j\_t\rvert}$，$n_{i,j_t}$表示词语$i$在文本$j$的标题中出现的次数，$\lvert j_t\rvert$表示标题中包含的词语数目
长度特征	5	词长(WL)	$wl_i=\frac{length(i)-u}{\partial}$，$length(i)$表示词语的长度，$u$表示文本中所有词语长度的均值，$\partial$表示文本中所有词语长度的方差
	6	句子长度(SL)	$sl_s=\frac{length(s)-shortest(s)}{longest(s)-shortest(s)}$，$length(s)$表示句子$s$包含的词语数，$shortest(s)$表示文本中最短的句子包含的词语数，$longest(s)$表示文本中最长的句子包含的词语数
位置特征	7	是否在标题中出现(OIT)	如果词语在标题中出现过则取值为1，否则取值为0
	8	首次出现位置(FP)	$fp_{i,j}=\frac{position(i,j)}{\lvert j\rvert}$，$position(i,j)$表示词$i$在文本$j$中首次出现的位置
	9	最后出现位置(LP)	$lp_{i,j}=\frac{positon_{-1}(i,j)}{\lvert j\rvert}$，$positon_{-1}(i,j)$为词语$i$在文本$j$中最后一次出现的位置
	10	词跨度(Span)	$span_{i,j}=lo_{i,j}-fo_{i,j}$，表示词语$i$在文本$j$中最后一次出现和首次出现位置的差值
	11	句中平均位置(SP)	$sp_i=average(\sum_{s\in S}\frac{position(i,s)}{\lvert s\rvert})$，表示词语$i$在文本所有句子$s$中出现位置的平均值，$position(i,s)$为词语$i$在句子中的位置，$\lvert s\rvert$为句子$s$中包含词语的个数
语言特征	12	词性特征(POS)	记录词语的词性信息，并用one-hot编码对词性进行编码
	13	是否为专有名词(IPN)	利用哈尔滨工业大学的LTP分词工具对词语进行标注，如果是专有名词则取值为1，否则取值为0
	14	语义特征(SF)	用语义特征来描述词语和文本主题之间的相关性，参考3.2节

词频特征主要描述词语在文本或语料中出现的频率。词在文本中出现的频率越高则越重要。但也有部分词无法表现出文本之间的差别，不能作为文章的关键词。长度特征是指候选关键词本身及其所在句子的长度特征，关键词一般为2～6个。此外，通常认为句子越长其包含的信息越丰富，句中包含关键词的概率就越大。位置特征常用候选关键词在文中出现的位置分布、跨度等信息进行描述。词语的位置一定程度上反映了文本的结构信息，所以位置信息对关键词的提取有很大的参考价值。语言特征主要从候选关键词的词性、构成以及语义信息等方面提取。

关键词提取的目的就是要提取出文本中最能够反映文本主题的词语。本文借助词向量模型量化了词语与文本主题之间的相似度，并将其作为描述候选关键词的语义特征。

本文采用Mikolov[16]提出的Skip-gram模型训练词语的分布式表示，训练之后的词语被表示为指定维度的词向量。词向量可以看作单词在句法和语义空间中的投影，利用词语的分布式词向量可以计算出词语间的语义相似度。假设文本集合中包含n个词语，将其表示为$w_1,w_2,\cdots,w_n$，其对应上下文表示为$c_1,c_2,\cdots,c_n$。如果某篇文本中包含k个词语，单词上下文依赖关系表示为条件概率p，则利用Skip-gram模型训练词向量的目标就是使所有词语的平均概率值达到最大，目标函数如式(1)所示。

$$\max\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{-k\leqslant j\leqslant k,j\neq 0}\log(p(c_{i+j}\mid w_i))\right) \tag{1}$$

用softmax函数处理概率p：

$$p(c\mid w;\theta)=\frac{e^{\boldsymbol{v}_c\cdot\boldsymbol{v}_w}}{\sum_{c'\in C}e^{\boldsymbol{v}_{c'}\cdot\boldsymbol{v}_w}} \tag{2}$$

其中，C表示文本对应的词典，$W=\{w_1,w_2,\cdots,w_n\}$表示需要训练的词语集合，D表示C和W组成的集合，$\boldsymbol{v}_w$表示训练得到的词向量，$\boldsymbol{v}_{c'}$表示上下文的词向量。将式(2)代入式(1)得到：

$$\max\left(\sum_{(w,c)\in D}\log(p(c\mid w))\right)=\max\left(\sum_{(w,c)\in D}\left(\log e^{\boldsymbol{v}_c\cdot\boldsymbol{v}_w}-\log\sum_{c'}e^{\boldsymbol{v}_{c'}\cdot\boldsymbol{v}_w}\right)\right) \tag{3}$$

经过上述训练过程，可以将词语及其语义信息表示成一个低维、稠密的词向量。为了能使词语的向量表示更加准确，算法首先加载了Li[20]提供的词向量模型，然后用预处理后的

学术文本数据对词向量进行再训练，得到更加全面且更符合语言情景的词向量。对文本进行 3.1 节的预处理之后，得到候选关键词集合 $W=\{w_1,w_2,\cdots,w_m\}$，集合中包含了 unigram，bigram 和 trigram 词组。但是，对文本词向量进行再训练的过程中只考虑了单个词语的上下文语义信息，因此只对单个词语的向量进行了更新，忽略了词组的向量。为了能够方便地对词组的语义主题相似度进行计算，需要对词组的表示方式进行设计。词语的向量表示为高维空间中的一个坐标点，词语间的语义相似度即为空间中坐标点之间的距离。根据这种思路，本文将词组也表示为空间坐标中的一个点，而坐标的位置则由组成词组的多个词语决定，用多个词语向量的平均值表示词组的向量坐标，计算公式如式(4)所示：

$$\boldsymbol{v}_i=\frac{\sum_{j=1}^{n}\boldsymbol{v}_j}{n} \tag{4}$$

其中，$\boldsymbol{v}_i$ 表示词组 i 的词向量，$\boldsymbol{v}_j$ 表示词语 j 的词向量，n 表示组成词组的词语个数。

用词向量表和词组向量计算公式把候选关键词序列转化为词向量序列 $V=\{v_1,v_2,\cdots,v_m\}$，然后利用 K-means 算法生成当前文本的主题中心向量。聚类结果的好坏与聚类中心的初始化有关，初始值不同可能导致得到不同的聚类划分。为了使聚类结果更加合理，本文根据候选关键词的重要程度设置了聚类中心 $C=\{c_1,c_2,\cdots,c_k\}$ 的初始值。选择词语权重较大的 k 个词作为中心的初始值。词语权重的计算式(5)如下：

$$w_{i,j}=\gamma\times tfidf_{i,j}+\alpha\times ttf_{i,j}+\beta\times span_{i,j} \tag{5}$$

其中，$w_{i,j}$ 表示文本 j 中词语 i 的权重；$tfidf_{i,j}$，$ttf_{i,j}$ 和 $span_{i,j}$ 分别表示词语的文档-逆文档频率、标题中出现的频率以及词语的跨度信息：

$$tfidf_{i,j}=\frac{n_{i,j}}{|j|}\times\log_2\frac{|D|}{df_i} \tag{6}$$

$$tf_{i,j}=\frac{n_{i,j}}{|j|} \tag{7}$$

$$span_{i,j}=\frac{positon_{-1}(i,j)}{|j|}-\frac{position(i,j)}{|j|} \tag{8}$$

其中，$n_{i,j}$ 表示词语 i 在文本 j 中出现的次数，$|j|$ 表示文本中词语的个数，$|D|$ 表示语料库中文本的总数，df_i 表示语料库中包含词语 i 的文本数，$position_{-1}(i,j)$ 表示词语 i 在文本 j 中最后一次出现的位置，$position(i,j)$ 表示词语 i 在文本 j 中首次出现的位置。参数 γ,α,β 作为调节参数，本文将其分别设置为 0.3，0.5，0.2。

得到初始化的聚类中心后，利用式(9)计算每个词向量到每个聚类中心的欧氏距离。然后将每个词向量分配给距离最近的簇，从而得到 k 个簇 $\{S_1,S_2,\cdots,S_k\}$。

$$\begin{aligned}dis(\boldsymbol{v}_i,c_j)&=\sqrt{(v_{i,1}-c_{j,1})^2+\cdots+(v_{i,d}-c_{j,d})^2}\\&=\sqrt{\sum_{t=1}^{d}(v_{i,t}-c_{j,t})^2}\end{aligned} \tag{9}$$

其中，$\boldsymbol{v}_i$ 表示第 i 个词向量，$1\leqslant i\leqslant m$；c_j 表示第 j 个簇的聚类中心，$1\leqslant j\leqslant k$；$v_{i,t}$ 表示第 i 个词向量的第 t 维属性。

聚类的目标是得到能使各个簇的中心到该簇中各个对象之间距离之和最小的簇的划分。聚类目标如式(10)所示：

$$J=\operatorname{argmin}\sum_{j=1}^{k}\sum_{i=1}^{m}dis(\boldsymbol{v}_i,c_j) \tag{10}$$

如果目标函数不满足收敛准则，则重复更新簇中成员，直到达到稳定状态时得到最终的聚类划分。聚类的中心描述了文本的主题信息，我们通过式(11)得到文本主题向量。

$$c_j=\frac{\sum_{\boldsymbol{v}_i\in S_j}\boldsymbol{v}_i}{|S_j|} \tag{11}$$

最后，通过计算候选关键词与文本主题向量间的距离来量化词语与文本主题之间的相关度。

$$ss_{i,j}=sim(\boldsymbol{v}_i,\boldsymbol{v}_c)=\frac{\vec{\boldsymbol{v}_i}\cdot\vec{\boldsymbol{v}_c}}{\|\boldsymbol{v}_i\|\cdot\|\boldsymbol{v}_c\|} \tag{12}$$

其中，文本 j 中词语 i 的词向量用 $\boldsymbol{v}_i$ 表示，$\boldsymbol{v}_c$ 表示聚类中心的词向量。

3.3　构建分类器

关键词提取可以看作是一个二分类问题，关键词的词语标签被标注为 1，否则被标注为 0。参照 3.2 节的特征提取步骤，提取出描述候选关键词的词频特征、位置特征、长度特征、语言特征共 14 个。将每个词语转化成 n 维向量 $\boldsymbol{x}_i$，并为训练集中的候选关键词添加标签 y_i。

$$\begin{aligned}\boldsymbol{x}_i=[&TF,IDF,TFIDF,TTF,WL,SL,OIT,FP,LP,\\&Span,SP,POS,INP,SF_1,\cdots,SF_k]\end{aligned} \tag{13}$$

$$y_i=-1\text{ or }+1 \tag{14}$$

把上述训练集合中的数据输入模型中，采用支持向量机(Support Vector Machine，SVM)算法训练关键词分类模型。

4　实验设计与结果分析

4.1　数据及评估标准

本文利用网络爬虫技术从中国知网上抓取数据，分别以“生物学”“化学”“数学”“地质学”“地理学”为搜索关键词，从结果集中选取文本长度大于 200 字符的 2016－2017 年的学术期刊文本的标题、摘要、关键词作为实验数据。数据集包含了 4713 篇学术文本，以及作者提供的 16987 个关键词。实验将文本按照4∶1拆分成了训练集和测试集，用 3.3 节设计的特征向量构建样本集合，并用其训练分类模型。模型搭建过程中用到的参数如表 2 所列。

表 2　模型参数

Table 2　Model parameters

	Description	Symbol	Value
svm 参数	核函数	kernerl	RBF
	交叉验证次数	cv	5
	并行数	n_jobs	8
Skip-gram 参数	学习率	alpha	0.02
	窗口大小	window	5
	最小词频	min_count	3

本文采用关键词提取结果评估中较常用的准确率检测方法，即把分类结果的关键词集合与科技论文中作者的关键词集合进行对比，以分类问题中较为流行的准确率 P(Precision)、召回率 R(Recall)以及 F-Score 为指标对模型的性能进行评估。

$$P=\text{抽取正确的关键词数量}/\text{抽取的总的关键词数量} \tag{15}$$

$$R=\text{抽取正确的关键词数量}/\text{标注的关键词数量} \tag{16}$$

$$F\text{-}Score=\frac{2\times P\times R}{P+R} \tag{17}$$

4.2 实验结果

为了选择合适的算法构建关键词分类模型，我们选择了几种常用的分类算法做了多组对比实验，结果如表 3 所列。可以看出，SVM 分类算法的准确率更高。虽然 SVM 模型的召回率没有逻辑回归(Logistic Regression，LR)的结果高，但是针对分类问题，综合准确率和召回率的评估标准 F-Score 的结果更有说服力，而在训练集和测试集上，SVM 模型的 *F-Score* 比其他分类模型都高。

表 3 分类算法的对比

Table 3 Comparison of different classification algorithms

	训练集			测试集		
	Precision	*Recall*	*F-Score*	*Precision*	*Recall*	*F-Score*
KNN	0.75	0.74	0.75	0.72	0.70	0.71
DT	0.86	**0.83**	0.85	0.83	0.78	0.81
SVM	**0.93**	0.80	**0.86**	0.87	0.77	**0.82**
LR	0.78	0.87	0.82	0.70	0.79	0.74
NB	0.88	0.86	0.87	0.87	0.79	0.83

在特征提取过程中，本文除了提取了一些统计特征外，还提取了描述词语与文本主题相关度的语义特征。由于一篇文本中包含的主题数目不定，因此聚类的个数 k 无法确定。为了选择合适的聚类个数，我们在提取 5 个关键词的情况下对比了不同聚类个数时实验结果的准确率，如图 2 所示。可以看出，当聚类个数为 4 时，模型表现最好。

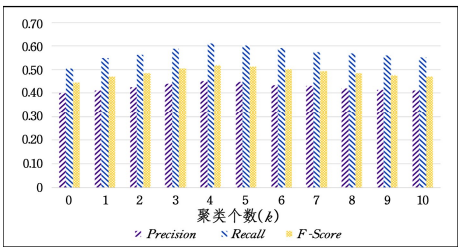


图 2 聚类个数对结果的影响

Fig. 2 Influence of clustering number on results

为了验证语义特征的重要性，我们用信息增益(Information Gain，IG)的方法检验了每个特征的重要性，评估结果如表 4 所列。可以看出，对结果影响最大的因素是词语是否出现在标题中；其次是本文提出的语义特征，说明引入语义特征对提升准确率有很大的帮助。

表 4 特征重要性排序

Table 4 Ranking of feature importance

Rank	Feature	Description	IG Score
1	TTF	标题中频率	0.0314
2	SF	语义特征	0.0251
3	TFIDF	词频-逆文档频率	0.0238
4	WL	词长度	0.0206
5	POS	词性	0.0185
6	SL	句子长度	0.0177
7	SP	句中平均位置	0.0154
8	IPW	是否是专有名词	0.0148
9	OIT	是否出现在标题中	0.0129
10	FP	首次出现位置	0.0098
11	LP	最后出现位置	0.0085
12	Span	词跨度	0.0078
13	TF	词频	0.0570
14	IDF	逆文档频率	0.0049

为了度量每个特征对整个系统性能的贡献，我们对比了从本文提到的属性集合中删除某个属性后的结果，如表 5 所列。可以看出，在学术文本中，候选关键词在标题中出现的频率有很大的参考价值。使用标题词频(TTF)特征时，系统的 *F-Score* 提升了 0.107。此外，本文提出的语义特征在提升模型准确率方面有很大的贡献。使用 SF 特征时，系统的 *F-Score* 提升了 0.072。

表 5 删除某一特征后系统的性能

Table 5 System performances after removing a feature or a family of features

Feature	*Precision*	*Recall*	*F-Score*
TFIDF	0.3967	0.539753	0.4573
TTF	0.3740	0.4551	0.4106
SF	0.3986	0.5047	0.4456
WL	0.4075	0.6007	0.4856
POS	0.4249	0.5975	0.4966
SL	0.4352	0.5836	0.4986
SP	0.4295	0.6099	0.5040
IPW	0.4476	0.5809	0.5056
OIT	0.4386	0.6044	0.5083
FP	0.4416	0.6060	0.5109
LP	0.4431	0.6085	0.5128
Span	0.4492	0.6032	0.5149
TF	0.4478	0.6073	0.5155
IDF	0.4484	0.6087	0.5164

为了验证算法的有效性，我们对比了多种关键词提取算法，实验结果如表 6 所列。可以看出，SFKE 方法明显优于其他对比方法。

表 6 不同方法的结果对比

Table 6 Comparison of results with different methods

Algorithm		*Precision*	*Recall*	*F-Score*
无监督方法	TFIDF	0.2879	0.3417	0.3125
	TextRank	0.2184	0.2917	0.2498
	YAKE	0.3382	0.4139	0.3722
有监督方法	KEA	0.3597	0.3693	0.3645
	AE	0.3618	0.5148	0.4249
	SFKE	0.4494	0.6102	0.5176

为了验证本文所提方法的性能，在关键词个数 k 为 1～10 的情况下，将本文提出的关键词提取方法与其他方法进行实验对比。图 3－图 5 展示了多种提取结果随关键词个数的变化情况。由图 3 和图 4 可以看出，提取的关键词越多，结果的准确率越高，而召回率越低。为了综合考虑这两方面的因素，我们选用 *F-Score* 作为综合评价指标。图 5 显示，随着关键词提取个数的增加，SFKE 算法的 *F-Score* 值呈先增大后减小的趋势，当 $k=5$ 时其 *F-Score* 取最大值，相比于 AE 算法的 *F-Score* 提升了 0.092。

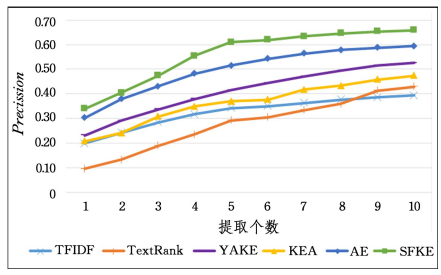


图 3 准确率随关键词个数的变化

Fig. 3 Results of Precision changing with number of keywords

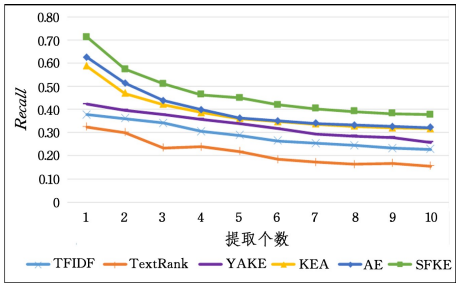


图 4　召回率随关键词个数的变化

Fig. 4　Results of Recall changing with number of keywords

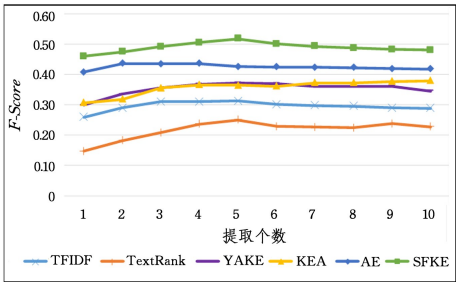


图 5　F-Score 随关键词个数的变化

Fig. 5　Results of F-Score changing with number of keywords

结束语　本文提出了量化语义特征的方法，并将该特征与词语的多种特征结合，设计了一种融合语义特征的短文本关键词提取(SFKE)方法。该方法综合考虑了关键词的多种特征，使提取的关键词既考虑了词语的词频、长度、位置等统计特征，又考虑了词语词性、语义等语言特征，从多方面衡量了词语的重要性。最后，实验验证了本文提出的关键词提取方法在关键词提取个数为 5、聚类个数为 4 的情况下，较其他方法有更好的表现。此外，通过对多组特征信息增益结果以及多组特征组合情况的对比，验证了本文提出的语义特征对结果的提升有显著的贡献。目前，本文提出的方法只考虑了有标注数据集合情况下的关键词提取问题，不适用于无标注语料情况下对关键词的提取。而半监督学习能够很好地学习有标签数据的规律，并深入挖掘无标签数据中包含的信息。因此，将半监督学习的方法与本文提出的方法结合，从大量无标签数据中挖掘数据信息，将是本研究的后续工作之一。

参 考 文 献

[1] ZHAO J S, ZHU Q M, ZHOU G D, et al. Review of Research in Automatic Keyword Extraction[J]. Journal of Software, 2017, 28(9): 2431-2449.

[2] BABAR S A, PATIL P D. Improving Performance of Text Summarization[J]. Procedia Computer Science, 2015, 46: 354-363.

[3] ONAN A, KORUKGLU S, BULUT H. Ensemble of Keyword Extraction Methods and Classifiers in Text Classification[J]. Expert Systems with Applications, 2016, 57(C): 232-247.

[4] LUHN H P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information [J]. IBM Journal of Research and Development 1957, 1(4): 309-317.

[5] MIHALCEA R, TARAU P. TextRank: Bringing Order into Texts[C] // Proceeding Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: 2004: 404-411.

[6] CHEN W, WU Y Z, CHEN W L, et al. Automatic keyword extraction Based on BiLSTM-CRF[J]. Computer Science, 2018, 45(S1): 104-109.

[7] CAMPPOS R, MANGARAVITE V, PASQUALI A, et al. A Text Feature Based Automatic Keyword Extraction Method for Single Documents [C] // Advances in Information Retrieval (EDS). Cham: Springer, 2018: 10772.

[8] ARDIANSYAH S, MAJID M A, ZAIN J M. Knowledge of extraction from trained neural network by using decision tree [C] // International Conference on Science in Information Technology. IEEE, 2017.

[9] FRANK E, PAYNTER G W, et al. Domain-Specic Keyphrase Extraction [C] // International Joint Conference on Artificial Intelligence. 1999: 668-673.

[10] CHEN Y, YIN J, ZHU W, et al. Novel Word Features for Keyword Extraction [M] // Web-Age Information Management. Springer International Publishing, 2015: 148-160.

[11] KANIS J. Digging Language Model-Maximum Entropy Phrase Extraction[C] // International Conference on Text. Speech: Brno, Czech, 2016: 46-53.

[12] ZHOU C, LI S. Research of Information Extraction Algorithm based on Hidden Markov Model[C] // International Conference on Information Science and Engineering. Springer, 2010: 1-4.

[13] ZHANG C. Automatic Keyword Extraction from Documents Using Conditional Random Fields[J]. Journal of Computational Information Systems, 2008, 4(3): 1169-1180.

[14] ZHANG Q, WANG Y, GONG Y, et al. Keyphrase Extraction Using Deep Recurrent Neural Networks on Twitter[C] // Empirical Methods in Nnatural Language Processing. 2016: 836-845.

[15] AQUINO, GERMAN O, LANZARINI L C. Keyword Identification in Spanish Documents using Neural Networks[J]. Journal of Computer Science & Technology, 2015, 15(2): 55-60.

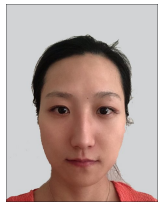
[16] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C] // International Conference on Learning Representations (ICLR). 2013: 1301-3781.

[17] LIU Z Y. Research on Keyword Extraction Method Based on Document Topic Structure[D]. Beijing: Tsinghua University, 2011.

[18] GitHub[OL]. https://github.com/uk9921/StopWords.

[19] CHEN Y C, ZHANG Y X, WANG H, et al. Features Oriented Survey of State-of-the-Art Keyphrase Extraction Algorithms [J]. Journal of Software, 2018, 29(7): 2046-2070.

[20] LI S, ZHAO Z, HU R, et al. Analogical Reasoning on Chinese Morphological and Semantic Relations[J]. Meeting of the Association for Computational Linguistics, 2018, 2: 138-143.



GAO Nan, born in 1983, Ph.D, is member of China Computer Federation. Her main research interests include data mining, machine learning and intelligent transportation system.