第47卷 第10期 计 算 机 工 程 2021年10月
Vol.47 No.10 Computer Engineering October 2021

·人工智能与模式识别· 文章编号：1000-3428(2021)10-0082-07 文献标志码：A 中图分类号：TP391.1

融合语义特征的TextRank关键词抽取方法

杨延娇，赵国涛，袁振强，韩家臣

（西北师范大学 计算机科学与工程学院，兰州 730070）

摘 要： TextRank使用共现窗口代替PageRank网页超链接以判断词语关系，但共现窗口机制下的词汇图是无向图，且实际中文文本中词语与其共现窗口内的词语之间在多数情况下没有认知上的指向性链接关系，导致共现窗口机制下的词语关系与PageRank网页超链接关系存在较大差别。为此，提出一种融合语义特征的关键词抽取方法S-TextRank。在TextRank方法的基础上以依存关系代替共现窗口判断词语关系，以模拟PageRank网页指向性超链接。对不同词性词语赋予相应的权重系数，从而模拟不同性质网页的重要程度。在此基础上，使用IDF方法结合汉语语法规则构建非关键词表，排除无关词语以降低其对抽取结果的影响。实验结果表明，S-TextRank方法在测试集上的准确率达到74%，比TextRank方法高19.4个百分点。

关键词： TextRank方法；关键词抽取；依存关系；词性重要度；IDF方法；PageRank方法

开放科学（资源服务）标志码（OSID）：

中文引用格式： 杨延娇，赵国涛，袁振强，等. 融合语义特征的TextRank关键词抽取方法[J]. 计算机工程，2021，47(10)：82-88.

英文引用格式： YANG Y J，ZHAO G T，YUAN Z Q，et al. TextRank-based keyword extraction method integrating semantic features[J]. Computer Engineering，2021，47(10)：82-88.

TextRank-based Keyword Extraction Method Integrating Semantic Features

YANG Yanjiao，ZHAO Guotao，YUAN Zhenqiang，HAN Jiachen

（College of Computer Science and Engineering，Northwest Normal University，Lanzhou 730070，China）

【Abstract】 TextRank uses a co-occurrence window instead of PageRank Web hyperlinks to determine the relationships between words. However，the vocabulary graph under the co-occurrence window mechanism is an undirected graph，and in most cases，there is no cognitive directional link between the words in the actual Chinese texts and the words in the co-occurrence window. Under this mechanism，the relationship between the words is sharply different from the hyperlink relationship of PageRank. To address the problem，a keyword extraction method，S-TextRank，is proposed integrating semantic features. Based on TextRank，S-TextRank employs dependency relationships instead of co-occurrence windows to determine the relationships between words to simulate directional PageRank hyperlinks. In addition，different part-of-speech words are assigned with corresponding weight coefficients to simulate the importance of different types of Web pages. Finally，a non-keyword list is constructed by using the IDF method and Chinese grammar rules to exclude the influence of irrelevant words on the extraction results. Experimental results show that the accuracy of the S-TextRank method achieves 74% on the test set，19.4 percentage points higher than that of the TextRank method.

【Key words】 TextRank method；keyword extraction；dependency relationship；part-of-speech importance；IDF method；PageRank method

DOI： 10. 19678/j. issn. 1000-3428. 0059116

0 概述

关键词抽取作为自然语言处理中的一项基础性研究，在信息检索、文本归类、自动摘要等领域得到广泛应用[1]，关键词抽取分为有监督的抽取方法与无监督的抽取方法[2]两类。

有监督的关键词抽取方法通过人工标注的方式得到标注集，使用机器学习方法训练语料得到分类器，使用分类器判断文档中的词语是否为关键词[3]，典型代表有SVM、Bytes方法。有监督的关键词抽取

基金项目： 国家自然科学基金（61662068）；甘肃省高等学校创新能力提升项目（2019A-006）。

作者简介： 杨延娇（1976—），女，副教授、硕士，主研方向为数据挖掘；赵国涛、袁振强、韩家臣，硕士研究生。

收稿日期： 2020-07-31 **修回日期：** 2020-09-11 **E-mail：** yangyanjiao@nwnu.edu.cn

方法准确率较高,但需要大量人工参与,难以适用于信息量巨大的现代应用场景。无监督的关键词抽取方法使用某种方式将文档中的词语按重要性进行排序,将排名靠前的词语输出为关键词。无监督的关键词抽取方法基于统计,通过词频、主题特征、文档信息等统计信息筛选出关键词,代表方法包括TF-IDF、LDA、TextRank[4]等。

TF-IDF方法基于词频,通过计算词语在单文档中的文档频率(Term Frequence,TF)与词语在文档间的逆文档频率(Inverse Document Frequence,IDF)得到词语的综合重要度。TF-IDF方法计算过程简单,准确率较高,但是纯基于统计,没有考虑句中词语的其他特征,在中文文本关键词抽取领域效率不高[5]。LDA方法基于隐含主题模型,是一种包括词、主题、文档的三层贝叶斯概率模型[6],其找出文档主题,以主题中出现概率最大的词语作为关键词。LDA应用广泛,但需要对语料进行预训练,关键词抽取效率在很大程度上取决于训练集文档的主题分布。

TextRank方法基于词汇图,是Google著名排序方法PageRank的衍生算法[7]。TextRank方法在2004年由R.MIHALCEA提出[8],通过词性标记提取名词、形容词、动词等候选词,以候选词之间的共现关系构建词汇图,迭代计算词汇图节点权重,将排序靠前的词语作为关键词。TextRank方法仅仅通过分析文档自身就可实现关键词抽取,具有快速反馈、弱语言相关等优点,但该方法没有考虑词语本身的重要性,词语重要度受词频影响较大,无法从文档整体角度进行考量[9]。许多学者在TextRank的基础上进行改进:文献[10]结合TextRank与LDA主题模型抽取关键词,发现当数据集有较强的主题分布时关键词抽取效果能够得到显著改善;文献[11]将世界知识以Word2vec词向量的方式融入TextRank模型,改进了TextRank单文档关键词抽取效果;文献[12]通过应用类似于反向传播概念的错误反馈机制增强了TextRank方法的性能;文献[13]在TextRank的转移概率计算中融合词图边和点的信息来提升关键词抽取结果;文献[14]提出用于提取Twitter的KECNW模型,着重强调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;文献[15]将权重公式应用到TextRank候选关键词得分公式中,提升了关键词抽取的准确率。文献[16-18]融合词频、词长、词性、位置等关键词提取因素,综合多种因子改进TextRank的关键词抽取效果,在特定类型文档关键词抽取中取得了较好的效果。

综上,现有TextRank改进方法在传统方法的基础上引入了词长、词性、位置等因子,或融合LDA、TFIDF、Word2vec等模型,在相关领域关键词抽取任务中取得了较好的效果。上述改进方法虽然综合各种影响因子提升了TextRank方法的性能,但并未在TextRank本身机制尤其是共现窗口和词性过滤机制上进行改进。

本文提出一种融合语义特征的关键词抽取方法S-TextRank,以指向性的依存关系代替共现窗口构建有向词汇图,对不同词性词语赋予相应的权重系数,并对不同类型的文本使用IDF方法结合汉语语法规则挖掘非关键词表,通过非关键词表提升S-TextRank的关键词抽取效果。

1 TextRank简介

TextRank是PageRank的衍生方法,用于为文本生成关键词和摘要。TextRank将分词后词语看作PageRank中的网页,将词语与其共现窗口范围内词语的共现关系看作PageRank网页间的超链接关系,构建类似PageRank网络模型的词图模型,迭代多次得到词语的TR值,将排名靠前的词语输出为关键词。TextRank词语TR值的计算公式如(1)所示:

$$T_{\text{TextRank}}(V_i)=(1-d)+d\times\sum_{V_j\in \text{In}(V_i)}\frac{w_{ji}}{\sum_{V_k\in \text{Out}(V_j)}w_{jk}}T_{\text{TextRank}}(V_j) \quad (1)$$

其中:d为阻尼系数,一般取0.85;V_i、V_j为词语节点;$\text{In}(V_i)$为指向词语节点V_i的词语节点集合;$\text{Out}(V_j)$为词语节点V_j指向的词语节点集合;w_{ji}、w_{jk}为词语节点V_j到V_i、V_j到V_k的边权重[19]。由式(1)可知,TextRank计算词语重要度TR值的方法与PageRank计算网页重要度PR值的方法基本一致,均通过多次迭代的方式得到结果。

2 S-TextRank方法

2.1 TextRank的不足

TextRank方法应用于中文文档关键词抽取时效果不太理想,其本源方法PageRank简单高效,经过大量分析验证,本文总结TextRank方法主要存在以下不足:

1)TextRank方法将词语当作PageRank方法中的随机网页,将词语与其共现窗口内词语判断为相互联系。在实际中文语料中,词语不一定与其共现窗口内词语有修饰性或联合性的链接关系。在中文文本中,除短句文本外,多数情况下词语与其共现窗口内词语(近位词语)并没有语义上的相互联系,因此,以共现窗口判断词语关系不合理。

2)PageRank方法并未过滤特定性质的网页,基于互联网全局信息计算页面重要度。TextRank方法在构建词汇图之前按照词性过滤词语,只选取名词、

形容词、动词等词语构建词汇图，由此得到的结果基于文档局部信息，降低了关键词抽取的可靠性。

3）TextRank方法将词语与词语之间的共现频次作为方法的边权值，词语重要度受词频影响较大，词语出现次数越多，越容易被筛选为关键词。方法抽取的部分关键词对于此类型文档没有任何表征意义，比如新闻类语料关键词抽取结果中出现的“报道”“表示”等词语。

本文提出一种S-TextRank方法，其通过以下方式弥补传统TextRank方法的不足：

1）使用句法依存关系代替共现窗口判断词语链接关系。句法依存关系是词语与词语之间语义上的指向性修饰关系，相比共现窗口而言，句法依存关系更贴近PageRank网页间指向性超链接关系。

2）使用所有词语构建词汇图，并对不同词性词语赋于相应的重要度权重系数，经过大量训练语料拟合出最佳的词性重要度权重系数。

3）使用IDF方法结合汉语语法规则在特定类别文档中挖掘非关键词表，用以过滤关键词抽取结果中的无关项。

2.2 S-TextRank词汇图构建

哈工大LTP（Language Technology Platform）平台具备分词、词性标记、依存关系分析、语义角色标记等一系列中文信息处理功能[20]。S-TextRank使用哈工大LTP平台对文档进行依存关系分析，使用依存关系判别词语间的相互联系，在TextRank方法的基础上融入语义特征。

在词汇图构建之前，S-TextRank对文档依次进行分句、分词、词性标注、依存关系分析，对单个词语标注词语汉字、词语词性、依存关系指向位置、依存关系、词语在句中位置等5个特征，并将其归为一个词语元祖，形式如（Word，$Word_P$，$Word_{RD}$，$Word_R$，$Word_D$）。其中：Word为词语本身汉字；$Word_P$为词语词性；$Word_{RD}$为词语在句中的依存关系指向位置；$Word_R$为依存关系类型；$Word_D$为词语在句中的位置。

在正常的文档中，词语间的依存关系不可能跨句存在，因此，本文对分句后句子进行句法依存分析以得到句中词语间的依存关系。因为标点符号对词语重要度没有实质性贡献，所以本文将标点符号部分从处理结果中删除。以“十九大”报告的主题为例，列举出本文方法对原文进行分句、分词、词性标注、依存关系分析之后得到的词语元组，结果如表1所示。

表1 例文及处理结果

Table 1 Examples and treatment results

“十九大”主题	处理结果
不忘初心，牢记使命，高举中国特色社会主义伟大旗帜，决胜全面建成小康社会，夺取新时代中国特色社会主义伟大胜利，为实现中华民族伟大复兴的中国梦不懈奋斗。	[('不','d',2,'ADV',1) ('忘','v',0,'HED',2) ('初','nt',4,'ATT',3) ('心','n',2,'VOB',4) ('牢记','v',2,'COO',6) ('使命','n',6,'VOB',7) ('高举','v',2,'COO',9) ('中国','ns',11,'ATT',10) ('特色','n',9,'VOB',11) ('社会主义','n',14,'ATT',12) ('伟大','a',14,'ATT',13) ('旗帜','n',9,'VOB',14) ('决胜','v',18,'SBV',16) ('全面','a',18,'ADV',17) ('建成','v',9,'COO',18) ('小康','n',20,'ATT',19) ('社会','n',18,'VOB',20) (',','wp',9,'WP',21) ('夺取','v',9,'COO',22) ('新','a',24,'ATT',23) ('时代','n',26,'ATT',24) ('中国','ns',26,'ATT',25) ('特色','n',22,'VOB',26) ('社会主义','n',29,'ATT',27) ('伟大','a',29,'ATT',28) ('胜利','v',22,'VOB',29) ('为','p',40,'ADV',31) ('实现','v',31,'POB',32) ('中华民族','n',34,'SBV',33) ('伟大','a',38,'ATT',34) ('复兴','v',34,'COO',35) ('的','u',35,'RAD',36) ('中国','ns',38,'ATT',37) ('梦','n',32,'VOB',38) ('不懈','z',40,'ADV',39) ('奋斗','v',22,'COO',40)]

以词语“高举”为例，当TextRank方法使用共现窗口判别词语链接关系且窗口大小为默认值5时，得到的词语链接关系如图1所示，使用句法依存关系时得到的词语链接关系如图2所示。

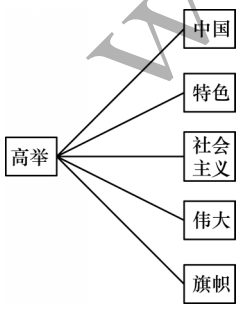

图1 共现窗口下的词语链接关系

Fig.1 Word link relations under co-occurrence window

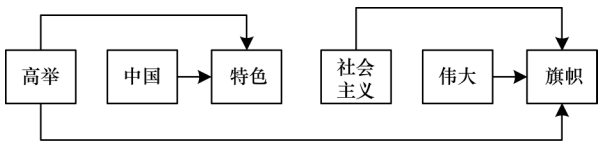

图2 依存关系下的词语链接关系

Fig.2 Word link relations under dependence relations

共现窗口机制下的词汇图是无向图，表示词语与其他词语在位置关系上具有泛化联系。由图1可知，词语“高举”与其共现窗口内多数词语没有认知上的相互联系。

本文验证大量不同题材的文档，通过句法分析判断词语与其后续词语间的语义关系，发现词语与其共现窗口内词语的关联程度与语句长度成反比。句中大部分词语一般只与其他一个词语存在语义链

接关系，词语与其后续词语之间存在语义链接的平均比率一般等于或略大于1/n（n为句子分词数）。

综上，在短句文本（每句分词数不大于5）中，词语容易与其共现窗口内词语有认知上的修饰性或联合性语义关系。在非短句文本中，词语与其共现窗口内多数词语没有认知上的语义链接关系。绝大多数中文文本都是非短句文本，因此，使用共现窗口机制判断词语关系时词语与其共现窗口内多数词语没有认知上的语义链接关系，且共现窗口机制判断词语与其共现窗口内词语为相互联系（无指向性），作为一种位置上的近位关系，与PageRank网页间自由灵活的指向性超链接关系相差较大。

由图2可知，指向性的依存关系构建了类似于PageRank网页链接图的有向词汇图，词与词之间的依存关系本质上接近网页的超链接关系，几乎完全正确地还原了词语间的语义链接，证实了S-TextRank以依存关系代替共现窗口判断词语关系具有可靠性。

S-TextRank使用依存关系构建有向词汇图，建立“词语-依存关系指向词语”的边，此边中出链节点为该词语，入链节点为依存关系指向词语。需要注意的是，依存标记为“VOB”的词语具有被动属性，会指向依存标记为“HED”（谓语）或“COO”（谓语同位语）的词语，作为一种“宾语-谓语”的关系。S-TextRank在构建依存标记为“VOB”词语的词汇图时，构建“依存关系指向词语-词语”的边，即将“谓宾关系”加入词汇图中。

在表1中，虽然依存关系是由第14项（“旗帜”，“n”，9，“VOB”，14）指向第9项（“高举”，“v”，2，“COO”，9），但“旗帜”作为“VOB”（宾语）指向“高举”（COO，谓语同位语，等同于谓语），是一种被动的指向关系。在正常情况下，宾语作为谓语动词的动作目标，在词汇图中应该建立“谓语-宾语”的链接关系而非“宾语-谓语”的链接关系。此外，句法依存标记为“HED”的部分指向0，此为句法依存关系中的句根虚拟节点，并无实际意义。

2.3　词性权重系数

TextRank方法使用词性过滤表过滤文档中副词、介词、连词等词语，只保留名词、动词、形容词等词性词语，使用这些词语构建词汇图。这种方式在一定程度上减少了用于计算的词汇量，简化了计算复杂性。

PageRank方法并未按网页的某种性质剔除不符合该性质的网页，作为PageRank的衍生方法，TextRank只选取特定词性词语构建词汇图，基于文档局部信息抽取关键词，其关键词抽取结果并不全面。

S-TextRank并未对文档进行词性过滤，而使用文档中所有词语构建词汇图。在词汇图构建过程中，使用词性判别不同词性词语对词汇图边权重的贡献程度。词性类似网页标签，最简单的网页标签是域名后缀，例如，“gov”表示政府官方，“int”表示国际组织，“com”“cn”表示普通互联网页面。从认知上而言，域名后缀为“gov”“int”的网页相较于域名后缀为“com”“cn”的网页具有更高的权威性，其所链接的网页理应具有更高的PR值。

鉴于上述思想，S-TextRank通过词性区分词语对边权重的贡献系数。从认知上而言，各词性词语对词汇图边权重的贡献系数应该有所差异，名词等实词性词语的贡献系数应大于副词等虚词性词语。本文使用已标注3~10个关键词的500篇篇幅大于3 000字的多领域文档做训练集，通过大量实验得到最优词性权重系数，具体过程如下：

1）因词性类别过多，为便于计算，本文将相似词性划为一个大类，根据语言学知识为每个词性设置一个初始权重λ，具体取值详见表2。其中，连词对句意影响较大，作为实词计算，数量词对句意影响较小，作为虚词计算。

2）每次迭代中为每个词性设置一个随机化参数γ，取值范围为−0.05~0.05，使词性权重随机变化（$\lambda=\lambda+\gamma$）。此过程中优先确定实词的权重最小值（权重下限），若虚词变化后权重越过实词权重下限，重新随机化γ直至虚词变化后权重不超出实词权重下限，得到一组新的词性权重系数。

3）对训练集每一篇文档使用此组词性权重系数结合依存关系构建有向加权词汇图，计算词语STR值，将排名靠前的n个词语输出为关键词（n为此文档已标注关键词数目）。参照已标注关键词得到此组权重系数下的抽取准确率。

4）迭代上述第2步、第3步100次，将抽取准确率最高的一组作为最优词性权重系数。

表2　初始词性权重系数

Table 2　Initial part of speech weight coefficient

词性	重要度
n、ns、nh、nz(普通与专有名词)	1.0
nl、nd(地点方位名词)	0.9
v(动词)	0.6
a(形容词)	0.5
m、q(数词与量词)	0.4
u(助词)	0.2
p(介词)	0.1
d(副词)	0.3
c(连词)	0.7
r(代词)	0.8

最优词性权重系数如表3所示。表2与表3中省略词性为叹词等虚词，这些词出现频率极低，对关键词抽取的影响可忽略不计。

表3 最优词性权重系数

Table 3 Optimal part of speech weight coefficient

词性	重要度
n、ns、nh、nz(普通与专有名词)	1.03
nl、nd(地点方位名词)	0.92
v(动词)	0.59
a(形容词)	0.47
m、q(数词与量词)	0.36
u(助词)	0.33
p(介词)	0.24
d(副词)	0.37
c(连词)	0.77
r(代词)	0.91

在得到各词性权重系数之后,S-TextRank根据“词语-依存关系指向词语”中出链词语节点词性,参考表3对此链接边赋予权重系数,累加得到此链接边的边权重。S-TextRank方法词语STR值计算公式如(2)所示,最大迭代次数为100次,迭代停止阈值为0.000 1。

$$S_{\text{S-TextRank}}(V_i)=(1-d)+d\times\sum_{V_j\in \text{In}(V_i)}\frac{w_{ji}}{\sum_{V_k\in \text{Out}(V_j)}w_{jk}}S_{\text{S-TextRank}}(V_j) \quad (2)$$

其中:d的经验值依旧取0.85;$\text{In}(V_i)$为指向词语节点V_i的词语节点集合;$\text{Out}(V_j)$为词语节点V_j指向的词语节点集合;w_{ji}、w_{jk}分别为V_j到V_i、V_j到V_k的边权重。

2.4 非关键词表

因TextRank方法本身特性,在初步的关键词抽取结果中,经常含有此类型语料环境中没有表征意义的“关键词”,比如在新闻类语料中,关键词抽取结果中有“报道”“表示”“认为”等泛类词语。此类词语在相关语料中出现频率极高,容易作为关键词被抽取出来,但此类词语在相关语境下没有任何特殊表征含义,作为关键词是不合理的。

IDF用于计算词语在相关文档集中的出现频率,词语的IDF值越大,说明该词语的泛化性越强。为了过滤泛类词语,S-TextRank将文档划分为生活、教育、娱乐、国际、国内等五大类型,引入IDF方法训练相关类型语料,训练集为同类型的500篇语料,设定阈值为0.7,将IDF值大于阈值的词语视为泛类词语。

本文为此设置一个非关键词表,认为非关键词表中的词语不可能作为关键词。将IDF训练得到的泛类词语加入非关键词表后,可以有效过滤关键词抽取结果中的泛类词语,排除泛类词语对关键词抽取结果的影响。

此外,S-TextRank根据词语类型参考汉语语法规则扩充非关键词表。除IDF值较大的泛类词语外,情态动作类词语如“知道”“了解”“发现”等、指代类词语如“人们”“这么”“那么”等、动作过程类词语如“继续”“有着”“之后”等,经常出现在TextRank关键词抽取结果中,此类词语无法表示文档特性,作为无关词项,不能作为关键词,使用此类词语扩充非关键表可以过滤更多的非关键词语。

有些关键词抽取方法通过构建语料的关键词库,将语料词语与关键词库重复词语作为关键词语,这种方法准确率较高,但其效率在极大程度上取决于所构建关键词库的可靠性,需要庞大的数据量及人工分析且适用性不高。本文方法构建的非关键词表计算简单,不需要人工分析,具有良好的适用性。

S-TextRank方法的总体流程如图3所示。

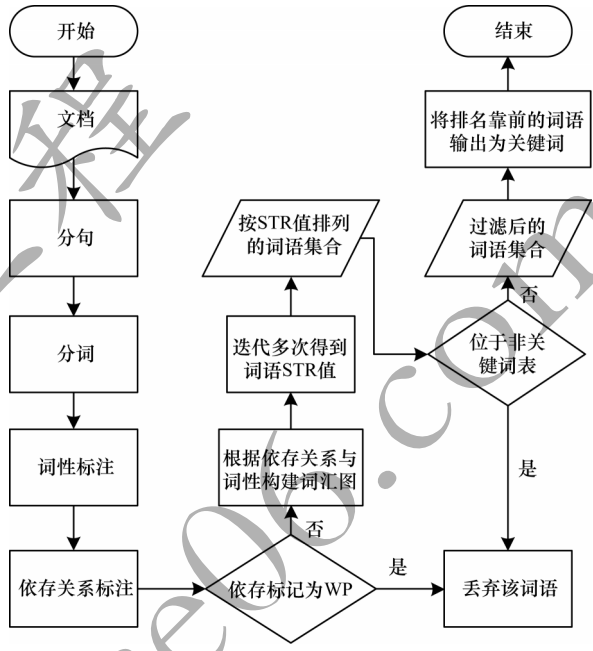

图3 S-TextRank方法流程

Fig.3 Procedure of S-TextRank method

3 实验与分析

3.1 实验环境与过程

实验的硬件环境:Intel® Core™ 4510-U,内存8 GB。软件环境:Win10 64位旗舰版,Python3.6.8+Pycharm2019.3.4。LTP库:ltp_data_v3.4.0。

目前中文文本关键词抽取方面没有公开的数据集,因此,本文抓取新浪网2020年5月—6月生活、教育、娱乐、国际、国内等5个频道共4 500多篇新闻语料构建语料库。每个频道随机选择100篇字数大于3 000字的语料共500篇作为测试集。为了保证关键词抽取结果的可靠性,针对测试集,采用多人人工交叉标注的形式提取新闻关键词,10名语言学相关研究人员参与标注,针对每一篇语料,对比每位研究人员的标注结果使用投票机制筛选出得票最高的10个词条,将其作为人工标注的关键词参考结果(通常10个关键词足以概括一篇新闻的主要内容)。

关键词抽取方法都是将重要度靠前的K个词语输出为关键词,为了使实验结果更加清晰,本文所有方法将重要度由大到小进行排序,前10个词语作为关键词抽取结果,通过准确率来反映方法效果。实验中的对比方法如表4所示,各方法在测试集上的实验结果如表5所示。

表4 实验方法

Table 4 Experimental methods

实验方法	方法简介
TextRank	jieba包实现的TextRank方法,使用词性过滤,共现窗口大小为默认值5
TF-IDF	jieba包实现的TF-IDF方法,使用词性过滤
Word2vec	Gensim包实现的Word2vec方法,词向量模型训练语料为800万篇多领域微信公众号的文章,总词数达650亿,词向量维度为256
TFIDF-TextRank	TFIDF改进初始权重的TextRank方法
FT-TextRank	文献[21]提出的基于词向量与TextRank的关键词提取方法(TextRank+Word2vec+FastText)
EPRank	文献[22]提出的融合词向量与位置信息的关键词提取算法(TextRank+Word2vec+位置信息)
S-TextRank	融合语义特征的TextRank(TextRank+依存关系+词性权重系数+非关键词表)

表5 实验结果

Table 5 Experimental results %

实验方法	准确率
TextRank	54.6
TF-IDF	52.7
Word2vec	33.6
TFIDF-TextRank	55.2
FT-TextRank	56.3
EPRank	57.5
S-TextRank	74.0

由表5可知,S-TextRank在测试集上的关键词抽取准确率达到了74%,高出TextRank方法19.4个百分点,其关键词抽取效果较对比方法具有大幅提升。

3.2 性能分析

在上述所有对比方法中,Word2vec方法准确率最低,文献[11]提到,对单文档直接使用Word2vec词向量聚类方法时,选择聚类中心作为文本的关键词本身就是不准确的,因此,与其距离最近的N个词语也不一定是关键词。

TFIDF-TextRank方法结合TF-IDF与TextRank,FT-TextRank结合TextRank、Word2vec、FastText,EPRank结合TextRank、Word2vec、位置信息,这些改进方法在准确率上相较原方法提升不大。作为PageRank的衍生方法,TextRank使用的共现窗口与词性过滤机制是不合理的,如果不改进本身缺陷,即使综合再多因子,也无法对TextRank进行本质上的改造,难以提升关键词抽取效果。

S-TextRank根据PageRank原理对TextRank方法进行本体改造,使用指向性的依存关系构建有向词汇图,使用全词性词语加权计算词汇图边权值,使TextRank在模型和计算上更加贴近PageRank方法,因此,其关键词抽取效果对比原方法得到大幅提升。

TextRank与TF-IDF是纯基于统计的关键词抽取方法,词语的词频越大,其作为关键词的概率越大。S-TextRank在TextRank的基础上融入了语义特征,依存关系类似于网页超链接,根据依存关系计算词图边权值,从而弥补了TextRank方法本身词频相关的缺陷。PageRank核心思想为:一个网页被许多网页链接,其PR值因此而提高[23]。在S-TextRank中,一个词语的STR值只与依存关系指向该词语的其他词语(出链词语节点)的数量多少和词性类别有关,与该词语(入链词语节点)在文档中出现次数以及位置无关。S-TextRank方法在词语关系判定机制上无限接近于PageRank指向性网页超链接关系。

实验发现,TextRank与TF-IDF方法容易出现主语丢失问题。在绝大多数文档中,因为文章表述的简洁性,文档主语一般出现一两次,之后通过代词指代或直接省略,因为TextRank与TF-IDF词频相关的特性,文档主语一般被赋于较小的重要度值,难以作为关键词被抽取出来,而失去了主语的关键词抽取结果难以代表全文信息。

S-TextRank方法使用依存关系替代共现窗口判断词语关系,文档主语虽为低频词语,但有许多依存关系修饰主语的部分,例如依存标记为“VOB”“ATT”的词语。S-TextRank方法能够有效抽取文档主语作为关键词,避免了纯基于统计的关键词抽取方法出现的主语丢失问题。

TextRank方法侧重于高频词语的抽取,因使用共现窗口判断词语关系,高频词语更容易出现在其他词语的共现窗口范围内,TR值一般偏大。TF-IDF方法侧重于名词性质关键词的抽取,因为动词等其他性质词语容易在相关语料中大范围出现,IDF值较低,而专有名词几乎只在相关文档中出现,IDF值较高。相较于以上方法而言,本文方法抽取语义上的关键词语,并没有如上述方法一般侧重于某类词语的抽取,因此,S-TextRank的关键词抽取结果较为中肯。

此外,因现有分词技术的限制,关键词抽取容易出现词语不完整的问题。例如,“美丽中国”作为一个关键词语,现有的分词技术大概率将其分为“美丽”“中国”,在得到的关键词中只能得到“中国”,缺少限定修饰词“美丽”之后的“中国”,其表征意义是不完整的。S-TextRank使用了句法依存关系,得到关键词后根据依存指向关系找到此关键词的修饰词,例如修饰名词类关键词的“ATT”部分,将修饰词与关键词作为一个词串输出,可以解决关键词抽取中词语不完整的问题,进一步提升S-TextRank方法的关键词抽取效果。

PageRank方法将网页排名的准确率由20%~30%提升到70%以上,由实验结果可知,S-TextRank方法在准确率上逼近原始PageRank方法,关键词抽取效果得到大幅提升。

4 结束语

本文基于PageRank思想对TextRank进行改进,使用依存关系代替共现窗口模拟PageRank的网页超链接关系,利用词性权重系数模拟不同性质网页的重要

程度，结合非关键词表得到最终的关键词抽取结果。实验结果表明，S-TextRank通过融入语义特征的方式，使关键词抽取效果相对TextRank得到大幅提升。下一步将融入命名实体等语义信息，综合考虑实体标注等因素提升TextRank方法的关键词抽取效果。

参考文献

[1] 李俊，吕学强. 融合BERT语义加权与网络图的关键词抽取方法[J]. 计算机工程，2020，46(9)：89-94.
LI J，LÜ X Q. Keyword extraction method based on BERT semantic weighting and network graph [J]. Computer Engineering，2020，46(9)：89-94. (in Chinese)

[2] SIDDIQI S，SHARAN A. Keyword and keyphrase extraction techniques：a literature review[J]. International Journal of Computer Applications，2015，109(2)：18-23.

[3] XIE F，WU X，ZHU X. Document-spcific keyphrase extraction using sequential patterns with wildscards[C]// Proceedings of IEEE International Conference on Data Mining. Washington D. C. ，USA：IEEE Press，2015：1055-1060.

[4] 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.
NING J F，LIU J Z. Using Word2vec with TextRank to extract keywords[J]. New Technology of Library and Information Service，2016(6)：20-27. (in Chinese)

[5] YAN Y，LIANG H，MENG Q. Exploration and improvoment in keyword extraction for news based on TFIDF[J]. Energy Procedia，2011(13)：3551-3556.

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

[7] BRIN S，PAGE L. Reprint of the anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks，2012，56(18)：3825-3833.

[8] MIHALCEA R，TARAU P. TextRank：bringing order into texts[C]//Proceedings of Empirical Methods on Natural Language Processing(EMNLP). Barcelona，Spain：Association for Computation Linguistics，2004：404-411.

[9] 罗有志，陈征明，陈明，等. 一种基于自适应关联熵的关键字提取算法[J]. 计算机与现代化，2020(4)：67-71.
LUO Y Z，CHEN Z M，CHEN M，et al. A keyword extraction algorithm based on adaptive association entropy [J]. Computer and Modernization，2020(4)：67-71. (in Chinese)

[10] 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(Z1)：41-47.
GU Y J，XIA T. Study on keyword extraction with LDA and TextRank combination[J]. New Technology of Library and Information Service，2014(Z1)：41-47. (in Chinese)

[11] 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.
XIA T. Extracting keywords with modified TextRank model[J]. Data Analysis and Knowledge Discovery，2017，1(2)：28-34. (in Chinese)

[12] FIGUEROA G，CHEN P C，CHEN Y S. RankUp：enhancing graph-based keyphrase extraction methods with error-feedback propagation[J]. Computer Speech and Language，2017，47：112-131.

[13] ZHANG Y，CHANG Y，LIU X，et al. Mike：keyphrase extraction by intergrating muitidimensional information[C]// Proceedings of 2017 ACM Conference on Information and Knowledge Management. New York，USA：ACM Press，2017：1349-1358.

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

[15] 徐立. 基于加权TextRank的文本关键词提取方法[J]. 计算机科学，2019，46(S1)：142-145.
XU L. Text keyword extraction method based on weighted TextRank[J]. Computer Science，2019，46(S1)：142-145. (in Chinese)

[16] 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.
LI H，TANG C L，YANG X，et al. TextRank keyword extraction based on multi feature fusion[J]. Information Magazine，2017，36(8)：183-187. (in Chinese)

[17] 张建娥. 基于多特征融合的中文文本关键词提取方法[J]. 情报理论与实践，2013，36(10)：105-108.
ZHANG J E. Method for the extraction of Chinese text keywords based on multi-feature fusion[J]. Information Studies：Theory&Application，2013，36(10)：105-108. (in Chinese)

[18] 艾金勇. 融合多特征的TextRank藏文文本关键词抽取方法研究[J]. 情报探索，2020(7)：1-6.
AI J Y. Research on the keyword extract method of Tibetan text based on TextRank integrated multiple features[J]. Information Research，2020(7)：1-6. (in Chinese)

[19] 刘治国，宋广跃，蔡文珠，等. 基于TextRank的未知协议帧定位方法研究[J]. 计算机工程，2020，46(7)：179-184.
LIU Z G，SONG G Y，CAI W Z，et al. Research on unknown protocol frame location method based on TextRank[J]. Computer Engineering，2020，46(7)：179-184. (in Chinese)

[20] 王明文，徐雄飞，徐凡，等. 基于word2vec的大中华区词对齐库的构建[J]. 中文信息学报，2015，29(5)：76-83.
WANG M W，XU X F，XU F，et al. word2vec based word alignment corpus for the greater China region[J]. Journal of Chinese Information Processing，2015，29(5)：76-83. (in Chinese)

[21] 周锦章，崔晓晖. 基于词向量与TextRank的关键词提取方法[J]. 计算机应用研究，2019，36(4)：1051-1054.
ZHOU J Z，CUI X H. Keyword extraction method based on word vector and TextRank[J]. Application Research of Computers，2019，36(4)：1051-1054. (in Chinese)

[22] 樊玮，刘欢，张宇翔. 融合词向量与位置信息的关键词提取算法[J]. 计算机工程与应用，2020，56(5)：179-185.
FAN W，LIU H，ZHANG Y X. Keyphrase extraction algorithm integrating word embeddings and position information[J]. Computer Engineering and Applications，2020，56(5)：179-185. (in Chinese)

[23] PAGE L. The page rank citation ranking：bringing orderto the Web[J]. Stanford Digital Libraries Working Paper，1998，9(1)：1-14.

编辑 吴云芳