现代计算机

XIANDAI JISUANJI

第29卷第2期（总第770期）

半月刊（1984年创刊）

2023年1月25日出版

主管单位 中山大学

主办单位 广州中山大学出版社有限公司

出版单位 广东现代计算机杂志社有限公司

发 行 广东省报刊发行局（全国公开发行）

印 刷 广州一龙印刷有限公司

社 长 黄少伟

主 编 石玉珍

编 委 邹岚萍 熊锡源 李 文 石玉珍 梁嘉璐

地 址 广州市海珠区新港西路135号
中山大学内（510275）

电 话 020-84112089（编辑部）

网 址 www.moderncomputer.cn

电子邮箱 tougao@moderncomputer.cn

ISSN 1007-1423
CN 44-1415/TP

邮发代码：46-121

定价：30.00元

邮局订刊二维码

现代计算机
官方网站二维码

ISSN 1007-1423
02
9 771007 142239

现代计算机 第29卷 第2期（总第770期） 2023年1月

2023年1月 第29卷

第2期 （总第770期）

ISSN 1007-1423
CN 44-1415/TP

MODERN COMPUTER

现代计算机

中山大学出版社 主办

中国期刊数据库CNKI全文收录期刊
中国学术期刊（光盘版）收录期刊
中文科技期刊数据库全文收录期刊
中国核心期刊（遴选）数据库收录期刊
中国学术期刊综合评价数据库收录期刊

- 研究与开发：计算机发展和软、硬件开发的理论研究
- 图形图像：重点为与图形图像相关的理论及实践研究
- 开发案例：基于某方面的计算机开发案例研究与分析
- 实践与经验：计算机应用的实例及心得

现代计算机
XIANDAI JISUANJI
第29卷 第2期（总第770期）2023年1月25日出版

目 次

研究与开发

实践与经验

开发案例

期刊基本参数：CN44-1415/TP*1984*S*16*120*zh*P*¥30.00*1200*18*2023-01

文章编号: 1007-1423(2023)02-0068-05 DOI: 10.3969/j.issn.1007-1423.2023.02.009

基于词向量与TextRank的政策文本关键词汇抽取方法研究

李 晨[1]，赵燕清[1]，于俊凤[1]，张铭君[1]，DMYTRO LANDE[1,2]

（1. 齐鲁工业大学(山东省科学院)情报研究所，济南 250014；
2. 乌克兰国立技术大学信息与计算机科学学院，基辅 03056）

摘要：通过对政策文本进行分析，设计了一种基于机器学习的关键词汇抽取方法，该方法可以自动从政策中提取关键性词语或短语。首先，从互联网采集相关政策文本并与维基百科数据融合，利用fastText构建词向量；其次，综合考虑词语的位置信息和词语之间的语义相似度，共同构建转移矩阵；最后选择得分最高的K个词语作为政策关键词。结果表明提出的方法抽取效果较好，实用性较高。

关键词：政策；词向量；关键词提取；TextRank

0 引言

政策通常是指政府、机构、组织为实现目标而订立的计划。政策文献是政策的物化载体，是政府处理公共事务的真实反映和行为印迹，是对政策系统与政策过程客观的、可获取的、可追溯的文字记录[1]。关键词是对文本的高度概括和抽象，能够帮助人们快速了解政策全文信息。因为政策本身的特殊性，原文中并不会设置关键词字段，如果可以采用自动的方式提取出与主题相关的词语或短语，则可以更好地辅助政策解读。

目前政策文本关键词提取算法大都以开源分词工具为基础，结合词频统计和人工辅助来实现。如吴宾等[2]利用开源工具人工提取海洋工程装备制造业政策主题词；吴爱萍等[3]运用扎根理论从政策样本中提取高频关键词。如果可以利用机器学习的方式自动抽取政策关键词，那么就可以进一步提高政策分析效率。现有的基于机器学习的关键词提取算法大体可以分为两类，分别是有监督提取方法和无监督提取方法。有监督的提取方法是一种分类方法，需要人工提前设置好训练集，然后训练出分类模型，最后通过分类模型完成关键词提取。如果训练集质量较高，用此方法可以得到比较好的结果，但是这种方式需要人工的参与，总体来说效率低、代价大。无监督的关键词抽取方式目前主要有三种：基于统计方法(TF-IDF)的抽取方式、基于主题模型(LDA[4])的抽取方式和基于图模型(TextRank)的抽取方式。

1 相关工作

基于TF-IDF的抽取算法是较为简单的一种实现。该方法以词频统计为基础，按照某个词在单篇文档中出现的次数和在所有文档中出现的文档频率进行计算。该方法可以过滤一些常见的无关紧要的词语，同时还能保留区分度较

收稿日期：2022-09-14 **修稿日期：**2022-10-21

基金项目：山东省重点研发计划(软科学项目)(2021RZA01017)：山东省科技型企业发展现状及对策研究；齐鲁工业大学(山东省科学院)科教产融合创新试点工程项目(2022GH015)：基于“双过程模型架构”的认知图谱关键技术及应用研究

作者简介：李晨(1988—)，男，山东济南人，硕士，图书馆员，研究方向为大数据与数据挖掘；赵燕清(1971—)，女，山东济南人，硕士，研究馆员，研究方向为智能决策与情报分析；于俊凤(1979—)，女，山东济南人，硕士，副研究员，研究方向为情报分析；张铭君(1990—)，女，山东济南人，硕士，图书馆员，研究方向为情报分析；Dmytro Lande(1959—)，男，乌克兰基辅人，博士，教授，研究方向为智能信息系统

高的重要词语。张骁等[5]就利用TF-IDF算法结合实际对科技服务业政策文本的关键词进行了提取。方法虽然易于实现但是缺点也很明显，单纯以词频衡量词的重要性，不够全面，同时也不能反映词的位置信息。为了克服这些缺点，很多人对TF-IDF算法进行了改进。张瑾[6]在原有的算法基础上加入位置权值及词跨度权值，避免了单纯采用TF-IDF算法产生的偏差。

LDA模型在自然语言领域被大规模应用，该技术同样也适用于文本关键词抽取。基于LDA的实现方式需要对数据集进行训练得到主题模型，选取能够反映主题的词语作为候选关键词，这种方式抽取的关键词很大程度上依赖训练数据的主题分布情况。

基于图模型的关键词提取算法近年来研究较多，借鉴PageRank[7]算法思想进行改进与扩展。该方法是将文本转化为相关词的词语网络图，该图的节点是词，边是词语之间的共现关系，该类算法无需引入外部语料进行训练，只需对图进行随机游走即可实现词语排序和关键词抽取。TextRank[8]借鉴了PageRank算法思想，首次实现了对词图上的关键词评分并根据评分结果完成关键词提取，成为了无监督关键词抽取方法的典型代表。为了进一步改进TextRank算法的提取效果，很多人对该算法进行了改进。夏天[9]以TextRank为基础，引入词语位置信息加权计算邻接词语的影响力转移矩阵，有效提高了抽取效果；李航等[10]提出一种综合考虑词性、词语位置信息、词语对文档集重要程度的改进TextRank方法；刘啸剑等[11]利用LDA构建主题模型，计算词语相似度并以此相似度为权重构建图的边，以短语作为图的节点，选择top-k个词作为文章的关键词。

随着词向量技术的产生，越来越多的人开始研究将词向量与TextRank结合进行关键词提取。词向量技术可以挖掘出词与词之间的语义关系，然后将这种语义关系引入到TextRank算法的计算过程当中，从而解决TextRank只考虑词共现的缺陷。周锦章等[12]通过构建词向量，基于隐含主题分布思想和词汇的语义差异构建转移矩阵，将词向量与TextRank融合。

本文在已有研究基础之上以维基百科作为外部知识库结合互联网获取的政策文本构建词向量，根据《国务院公文主题词表》为词语初始化权重，再利用词向量计算词语之间语义相似度，结合政策文本位置权重共同构建TextRank转移矩阵，最终选择K个关键词。

2　方法实现

2.1　基于fastText的词向量构建

词向量是指用来表示词语的向量，如比较简单的One-hot representation。由Mikolov提出的word2vec是至今比较有名的词向量表示方式。word2vec的出现解决了传统词袋模型的缺点，而word2vec再生成词向量的时候会把每个词当成原子，忽略词内部的形态特征。相对于word2vec，fastText添加了subwords特性，使用字符级的n-grams来表示单词，这样每个单词除保留了本身外还被表示成多个n-grams。对于每一个单词，fastText在拆分成n-grams表示的时候，还在单词前后端加入“<”和“>”，用于区分前缀和后缀，如单词hello采用3-grams可以表示为：<he，hel，ell，llo，lo>和<hello>。在训练模型的时候，当前词的词向量就是n-grams的向量和：

$$\boldsymbol{v} = \sum_{g \in \Phi} z_g^{\mathrm{T}} \boldsymbol{v}_c \tag{1}$$

本文利用开源的fastText工具，将维基百科和政策内容融合共同构建词向量。

2.2　TextRank转移矩阵构建

利用TextRank进行关键词抽取的思想比较简单：首先根据词共现关系构建无向带权图，然后利用PageRank循环迭代计算节点权值，排序权值即可得到最终关键词。TextRank算法的核心计算公式如式(2)所示：

$$WS(V_i) = (1 - d) + d * \sum_{V_j \in \ln(V_i)} \frac{w_{ji}}{\sum_{V_k \in \mathrm{Out}(V_j)} w_{jk}} WS(V_j) \tag{2}$$

其中：$WS(V_i)$表示节点V_i的权值；$\ln(V_i)$表示指向V_i的节点集合；$\mathrm{Out}(V_j)$表示节点V_j指向的节点集合；w_{ji}表示两个节点之间边的权重；$WS(V_j)$表示节点V_j的权值；d为阻尼系数，一般取值0.85。基于TextRank的关键词抽取步骤如下：

（1）文本预处理。包括按句子进行文本分割、分词、词性标注、去停用词。

（2）构建词图。文本预处理之后的词语构成节点集合，根据词语的共现关系构建边集。边的构建采用滑动窗口机制，即当两个节点在长度为K的窗口中共现，它们之间才会存在边。

（3）根据公式（2）迭代各节点的权重，直到结果收敛。

（4）对结果进行排序，得到top-k关键词。

（5）对所得到的关键词进行组合，如果组合的词汇在政策全文中出现，则选择该组合词作为一个关键短语。

本文通过引入词向量等方式对上述步骤进行修改，从而达到面向政策文献的关键词抽取。

对于步骤（1），在进行分词的时候引入《国务院公文主题词表》作为词库，同时剔除此表中无区分度的词语，如：章程、条例、办法、细则、规定、命令、决定、决议、公告、通告、通知、通报、报告、请示、批复、函、会议纪要、答复等。

当关键词出现在词表中的时候，在原有权重的基础上再乘1.5。

对于步骤（2），在构建此图边集的时候综合考虑词语的位置信息和词语之间的语义相似度，共同构建转移矩阵。图中任意两个节点v_i和v_j之间的权重转移是通过边w_{ij}来完成的，w_{ij}的构建如公式（3）所示：

$$w_{ij} = ft(i,j) + pos(i,j) \tag{3}$$

其中，$ft(i, j)$计算方式如下：

$$ft(i,j) = vsim(i,j)*coc(i,j) \tag{4}$$

$vsim(i,j)$表示v_i和v_j的的语义相似度，将两者的词向量取出，按照余弦相似度进行计算，$coc(i,j)$表示二者共现次数。$pos(i, j)$表示政策文本位置信息的重要影响程度，其计算方式如下所示，p_j表示词语出现的位置权重：

$$pos(i,j) = \frac{p_j}{\sum_{k \in \mathrm{Out}(v_j)} p_k} \tag{5}$$

$$p_j = \begin{cases} 1.0, & j\text{出现在标题或段首} \\ 0.8, & \text{其他位置} \end{cases} \tag{6}$$

3 结果分析

本文以我国“双创”政策为例，验证提出方法的有效性和实用性。首先从互联网相关网站搜索以国务院以及各部委为发文主体，以“双创”为内容的政策文件共计163篇（如表1所示）。

表1 “双创”政策（部分）

编号	政策名称	发文时间	发文字号
1	国务院办公厅关于进一步支持大学生创新创业的指导意见	2021-10-12	国办发〔2021〕35号
2	国务院办公厅关于建设第三批大众创业万众创新示范基地的通知	2020-12-24	国办发〔2020〕51号
3	国务院办公厅关于支持多渠道灵活就业的意见	2020-07-31	国办发〔2020〕27号
4	国务院办公厅关于提升大众创业万众创新示范基地带动作用进一步促改革稳就业强动能的实施意见	2020-07-30	国办发〔2020〕26号
5	国务院关于促进国家高新技术产业开发区高质量发展的若干意见	2020-07-17	国发〔2020〕7号
6	国务院办公厅关于应对新冠肺炎疫情影响强化稳就业举措的实施意见	2020-03-20	国办发〔2020〕6号
…	…	…	…
56	国务院关于进一步做好新形势下就业创业工作的意见	2015-05-01	国发〔2015〕23号
57	国务院办公厅关于创新投资管理方式建立协同监管机制的若干意见	2015-03-19	国办发〔2015〕12号
58	国务院关于取消和调整一批行政审批项目等事项的决定	2015-03-13	国发〔2015〕11号
59	国务院办公厅关于发展众创空间推进大众创新创业的指导意见	2015-03-11	国办发〔2015〕9号
…	…	…	…
161	关于支持和促进重点群体创业就业税收政策有关问题的补充通知	2015-01-27	财税〔2015〕18号
162	人力资源社会保障部办公厅关于做好留学回国人员自主创业工作有关问题的通知	2015-01-16	人社厅函〔2015〕19号
163	科技部关于进一步推动科技型中小企业创新发展的若干意见	2015-01-10	国科发高〔2015〕3号

根据本文提出的算法，对上述政策列表进行关键词提取，剔除权重小于1.0的关键词，最终得到关键词192个(见表2)。从关键词列表和词云可以看出，国家以创新创业为核心制定多项保障措施，如政府工作改革、提供支持政策、加强科技支持、培养创新孵化企业、优化税收政策、提供贷款资金支持等。

表2　“双创”政策关键词列表(部分)

关键词	权重	关键词	权重	关键词	权重
创新创业	20.0	工作改革	12.22	科技支持	9.43
就业创业	9.0	创业服务	9.0	孵化	9.0
发展改革	9.0	建设	8.28	创新示范	6.84
…	…	…	…	…	…
税收优惠政策	5.0	创业投资	4.26	贷款	4.10
创新发展	4.0	创业基地	4.0	众创空间	4.0
…	…	…	…	…	…
技术人员创新	1.0	投资引导	1.0	拉动重点	1.0
农村产业融合	1.0	就业服务工作	1.0	奖励股权	1.0

精确率方面，本文使用维基百科中文语料和从互联网采集到的163篇“双创”政策作为词向量训练文本(参数维度设置为100，窗口大小为5)，选择其中50条政策作为测试集，采用多人交叉标注的方式为每篇政策选择10个关键词。实验指标采用精准率P、召回率R和$F1$值进行评测，其中N_1表示人工标注合集，N_2表示算法抽取合集。

$$P = \frac{N_1 \cap N_2}{N_2} \tag{7}$$

$$R = \frac{N_1 \cap N_2}{N_1} \tag{8}$$

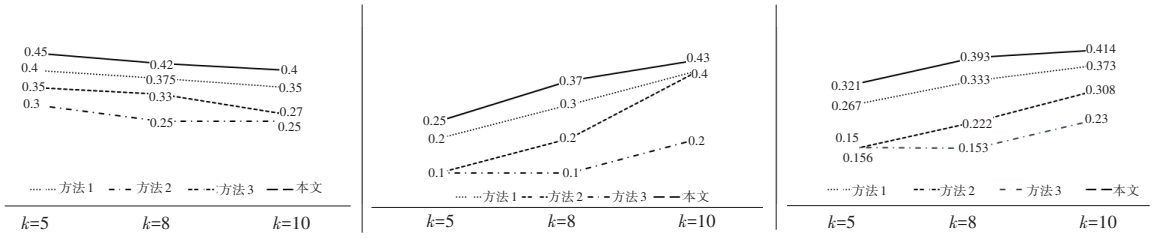
$$F1 = \frac{2*P*R}{P + R} \tag{9}$$

实验选择提取5、8、10个关键词对不同的方法进行对比，结果如表3所示。具体包括引入政策文本位置信息的TF-IDF方法(方法1)、TextRank方法(方法2)、基于Word2Vector的关键词抽取方法(方法3)。

表3　不同提取方法的比较结果

	K=5			K=8			K=10		
	P	R	$F1$	P	R	$F1$	P	R	$F1$
方法1	0.40	0.2	0.267	0.375	0.30	0.333	0.35	0.40	0.373
方法2	0.30	0.10	0.15	0.250	0.20	0.222	0.25	0.40	0.308
方法3	0.35	0.10	0.156	0.330	0.10	0.153	0.27	0.20	0.230
本文方法	0.45	0.25	0.321	0.420	0.37	0.393	0.40	0.43	0.414

为了能够更直观地查看对比结果，采用折线图的方式将准确率P、召回率R和$F1$值进行展示，如图1所示。



(a) 准确率　　(b) 召回率　　(c) $F1$值

图1　对比结果

从图1可以看出，实现方式最简单的是方法1，其实验结果要优于方法2和方法3。对于方法2，它的实现虽然不依赖于语料环境，但是在没有任何改进的情况下也不能取得较好的结果。在未加入外部语料和只考虑词语相似度的情况下，方法3虽然引入了词向量技术，但是实验结果却是最差。本文提出的方法在关键词抽取个数不同的情况下相对其他几种算法效果都有明显的提升，准确率、召回率和$F1$值均高于其他三种方法，验证了提出方法的有效性和实用性。

4　结语

当对政策文本进行主题分析时，往往需要提取政策文本的关键词汇，而由于其本身的特

殊性，并不会直接提供关键词字段，所以就要对政策关键词进行提取。本文提出了一种将外部知识库和政策库融合共同构建词向量，利用《国务院公文主题词表》修改词语权重，综合考虑位置信息和词语相似度构建TextRank转移矩阵的政策文本关键词抽取方法。以“双创”政策为例，提取政策关键词，结果表明本文提出的方法具有较好的效果，可用于政策文本主题分析，为政策研究人员提供辅助支持。

参考文献：

[1] 黄萃,任弢,张剑.政策文献量化研究:公共政策研究的新方向[J].公共管理学报,2015,12(2):129-137.

[2] 吴宾,杨一民,娄成武.基于文献计量与内容分析的政策文献综合量化研究:以中国海洋工程装备制造业政策为例[J].情报杂志,2017,36(8):131-137.

[3] 吴爱萍,董明,李华.“互联网+”与“大众创业、万众创新”政策结构分析:基于扎根理论和共词分析法[J].科技管理研究,2018,38(10):44-52.

[4] 朱泽德,李森,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.

[5] 张骁,周霞,王亚丹.中国科技服务业政策的量化与演变:基于扎根理论和文本挖掘分析[J].中国科技论坛,2018(6):6-13.

[6] 张瑾.基于改进TF-IDF算法的情报关键词提取方法[J].情报杂志,2014(4):153-155.

[7] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the web [R]. Stanford InfoLab, 1999.

[8] MIHALCEA R, TARAU P. Text rank: bringing order into texts [C] //Proceedings of Conference on Empirical Methods in Natural Language Processing. Barcelona, 2004: 404-411.

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

[10] 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.

[11] 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.

[12] 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.

Research on keyword extraction of policy using word vector and TextRank

Li chen[1], Zhao Yanqing[1], Yu Junfeng[1], Zhang Mingjun[1], DMYTRO LANDE[1,2]

(1.Information Research Institute, Qilu University of Technology (Shandong Academy of Sciences), Ji'nan 250014;
2.Faculty of Information and Computer Science, National Technical University of Ukraine, Kyiv 03056)

Abstract: By analyzing policy texts, a key word extraction method based on machine learning is designed, which can extract key words or phrases automatically from policies. Firstly, it used fastText to construct word vectors for policies obtained from the Internet; then, integrated word vector and position information into TextRank transfer matrix; finally, selected the Top K words with the highest score as the policy keywords. The results show that the method proposed has good extraction effect and high practicability.

Keywords: policy; word vector; keywords extraction; TextRank