文章编号：2095-2163(2024)01-0008-08 中图分类号：TP391.1 文献标志码：A

基于先验知识的纺织论文关键词自动抽取系统研究

李启正[1]，戴 豪[2]，胡崴琳[2]，祝成炎[2]
(1 浙江理工大学 杂志社，杭州 310018；2 浙江理工大学 纺织科学与工程学院(国际丝绸学院)，杭州 310018)

摘 要：为解决文章关键词数量过少、词义泛化、选词生僻、一义多词等问题，在搜集整理大量纺织领域论文和专业名词的基础上，遵循“避免泛化词”和“作者习惯”的原则，提出一种基于先验知识的论文关键词抽取新算法。首先统计候选关键词在概要库和关键词集中的出现频次，计算其先验概率；再借鉴“影响因子百分位”的思想，计算每个候选关键词的词频百分位，得到候选关键词的排序指标用于关键词抽取系统的排序推荐。经测试，该算法平均准确率(*F*1 值) 是无监督关键词抽取算法 TextRank 的 1.7 倍，并高于计算机领域同类型的半监督主流算法，证明了先验知识用于关键词排序推荐的有效性。

关键词：影响因子百分位；自动抽取；先验知识；先验概率；纺织论文

Automatic keyword generation system in textile field based on prior knowledge

LI Qizheng[1], DAI Hao[2], HU Weilin[2], ZHU Chengyan[2]
(1 Periodicals Agency of Zhejiang Sci-Tech University, Hangzhou 310018, China; 2 College of Textile Science and Engineering(International Institute of Silk), Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In order to solve the problems of too small number of keywords, generalization of words, remote selection of words, and multiple words with one meaning, based on collecting and organizing a large number of theses and professional terms in the field of textile, and following the principles of "avoiding generalization of words" and "author′s habit", a new algorithm for extracting keywords from theses based on a priori knowledge is proposed. Based on the principle of "avoiding generalized words" and "author′s habit", a new algorithm based on a priori knowledge is proposed to extract keywords from papers. Firstly, the frequency of the candidate keywords in the summary library and keyword set is counted, and its priori probability is calculated; then, the idea of "impact factor percentile" is used to calculate the word frequency percentile of each candidate keyword, and the ranking index of the candidate keywords is obtained, which is used for the ranking recommendation of the keyword extraction system. After testing, the average accuracy (*F*1 value) of the algorithm is 1.7 times higher than that of unsupervised keyword extraction algorithm TextRank, and higher than that of the same type of semi-supervised mainstream algorithms in the field of computers, which proves the effectiveness of priori knowledge in keyword ranking recommendation.

Key words: percentile of influencing factors; automatic extraction; prior knowledge; prior probability; textile thesis

0 引 言

在中国知网发布的中国学术期刊影响因子年报(2021 年版)中，纺织科学技术学科共有 37 本期刊，收录了 197 079 篇期刊论文(统计时间为 2022 年 7 月 10 日)。在这些论文中，有 10 万余篇未标关键词。论文关键词缺失会导致论文检索困难[1]，此外很多论文还存在关键词少、关键词生僻[2]、关键词泛化[3]和一义多词[4]等问题，通过人工对这些论文进行关键词标注存在较大困难，而且标注质量参差不齐。论文的标题是对论文高度浓缩的、精确的概括，论文摘要是一篇论文的核心与精华[5]，其能够反映论文核心内容和要点。通过研究发现，80%纺织期刊论文的作者关键词能够从论文的标题和摘要中抽取出来。因此，根据计算机程序算法，从论文的标题和摘要等核心信息中，自动抽取论文关键词是切实可行的方案，但技术方案的核心和关键在于需建立“避免泛化词”和“遵循作者习惯”的模型算法。

目前，关键词自动抽取系统的研究主要在计算机[6]和医学[7]等领域有见报道。关键词自动抽取

作者简介：李启正(1981-)，男，博士，副教授，主要研究方向：大数据分析；戴 豪(1998-)，男，硕士研究生，主要研究方向：自然语言处理；胡崴琳(1998-)，男，硕士研究生，主要研究方向：大数据分析。
通讯作者：祝成炎(1962-)，男，硕士，教授，主要研究方向：织物设计 CAD。Email：cyzhu@ zstu.edu.cn

收稿日期：2023-01-19

主要包括 2 个步骤，即确定候选词集和对候选词集合进行合理排序[8]。其中，第一步是基础工程，需要系统收集大量本领域的专业词汇和关键词合集；第二步需要考虑到不同专业领域的特殊性，根据关键词的重要性进行排序，这也是本文的研究重点。关键词重要性排序主要有词频[9]和词共现[10]两类方法，但这两类方法都未考虑到词语在领域中的使用习惯和问题状况。例如，当“研究”“应用”“方法”等词在标题、摘要等运算文本中反复出现时，基于词频的方法就会将其优先推荐为关键词，但根据科技论文关键词标引原则[11]，此类泛化词不适合作为论文的关键词。

为建立合理的关键词推荐算法，本研究引入先验知识[12]的思想，由于先验概率区分精度有限，在充分考虑领域特征对关键词价值影响的基础上，借鉴“影响因子百分位”的思想，再引入词频百分位，构建基于先验概率和词频百分位的关键词排序算法，并进行了软件系统的研发。通过在纺织领域 37 种期刊 99 683 篇论文中的运算验证结果表明，相较于传统的关键词排序方法，结合先验概率和词频百分位的关键词排序算法，可以较好地实现关键词的抽取。

1　研究现状

1.1　关键词抽取算法研究

按照是否需要对数据进行标注，关键词抽取算法可分为有监督和无监督两种方式[13]。有监督的方式需要在人工标注数据的基础上进行训练，相对成本较高；无监督的方式不需要对数据进行人工标注，主要根据词语在上下文中的自然分布特征。无监督的关键词抽取算法主要分为 3 类：第一类是基于简单的统计方法，根据对候选词的一些特征指标，如词频[14]和逆文本概率[15]进行统计，并基于统计指标的排序实现候选词的选用抽取关键词，这类方法简单易用，但是准确率不高；第二类是基于语义主题的方法，通过已有的“词语-文档”矩阵和训练得到的“文档-主题”矩阵优化了关键词抽取的效果，可以抽取出语料库中主题鲜明的词语，但是计算复杂度较高，如 Blei[16]利用 LDA 主题模型进行关键词抽取等；第三类是基于图模型的方法[17]，该方法的主要思想是将文档中候选词视为一个个节点，根据词和词之间的共现关系构建“词共现网络”，根据节点在网络中的重要程度作为词的权重，根据权重排序实现关键词抽取，如基于 PageRank 算法思想的 TextRank 算法[18]。此外，还有一些学者后续通过更改边权重[19]、更改词语影响力[20]、构建新的概率转移矩阵[21]、进行向量表征[22]等方法对关键词抽取算法进行了改进，取得了不同程度的效果提升。但是，这 3 类方法对于词权重的处理均不够理想。

1.2　先验知识相关研究

先验知识（priori knowledge）是指先于经验，与一切具体经验无关的知识。在本研究中，主要体现在先验概率和词频百分位两个算法指标的引入。先验概率是指根据以往经验和分析，在实验或采样前就可以得到的概率。后验概率是指某件事已经发生，想要计算这件事发生的原因是由某个因素引起的概率。因为，在关键词抽取任务中，需要借助词表和“作者习惯”进行关键词抽取，并且需要在实验前得到词的概率，因此可以引用先验概率的概念，并且先验概率已经在深度学习、实体识别以及纹样抽取等领域都有相关应用，也有学者将这一指标用于关键词抽取的相关研究工作。由于通过先验概率得到的候选关键词排序指标区分精度不够，本文借鉴“影响因子百分位”[23]的思想，增加了词频百分位指标，用于候选关键词的排序推荐。

2　研究方法

本研究将关键词抽取分为候选关键词建立和关键词排序两个阶段，在候选词集合建立阶段，首先收集纺织领域的学术论文，提取论文作者关键词，建立受控词表，计算每个关键词出现的频次；为遵循“作者习惯”原则，减少推荐冷僻词，本研究将出现频次小于等于 1 的关键词进行排除，将剩下的作者关键词作为本研究的候选关键词，并构建关键词集合。

在关键词排序阶段，首先建立由所有论文标题和摘要构成的概要库集，再统计候选关键词在概要库集中的出现频次，计算每个候选关键词的先验概率，作为主要排序指标；由于先验概率指标区分精度不够，本研究借鉴“影响因子百分位”思想，增加了词频百分位作为候选关键词的另一排序指标。流程框架如图 1 所示。

2.1　实验数据准备

2.1.1　数据集采集

本研究通过 Python 收集了中国学术期刊影响因子年报（2021 年版）中被分类到纺织科学技术学科的 37 本期刊共 197 079 篇论文的标题、作者关键词和摘要等数据，在删除标题、关键词缺失的论文后，得到 99 683 篇论文数据，数据采集时间截止到

2022 年 7 月 10 日。将收集到的数据集进行随机打乱,并按照训练集与测试集为 9 ∶ 1 的比例进行分配,训练集用于模型构建,测试集用于模型性能评估。

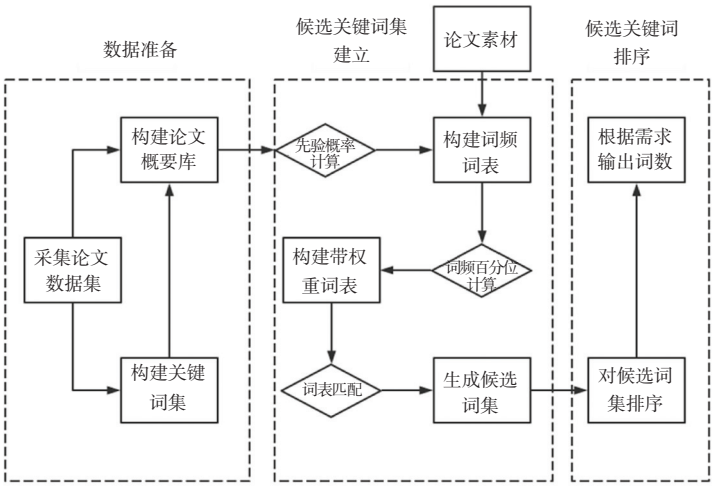

图 1 基于先验概率和词频百分位的论文关键词自动抽取流程框架

Fig. 1 Framework diagram of automatic keyword extraction process for papers based on prior probability and word frequency percentiles

2.1.2 构建关键词集

从训练集中抽取出作者标注的关键词,在去重后得到 103 877 个关键词。然后,计算这些作者关键词在论文的标题、摘要以及作者关键词中出现的频次,去除成为有效关键词的概率很低的词频为 1 的作者关键词,最终得到具有 29 146 个关键词的纺织关键词集。

2.1.3 构建论文概要库

本文提及的概要由“标题”和“摘要”两部分组成,在收集到的论文数据中,将标题的前缀标为 TI,摘要的前缀标为 AB,通过正则表达式,抽取标题和摘要的数据并进行合并,组合成论文的概要库 TA,供软件系统调用。

2.2 候选关键词集合建立

2.2.1 先验概率计算

在完成数据准备之后,关键词排序将成为关键问题。根据得到的关键词集合,将词表中的词放入论文的概要库中进行运算,统计出关键词在概要中出现的次数以及关键词在概要和作者关键词中同时出现的次数,通过这几个指标进行计算,得出对应关键词 W_i 的先验概率 P,如式(1) 所示:

$$P_{W_i} = \frac{A_{i1} + A_{i2}}{A_{i1} + A_{i3}} \tag{1}$$

注:当 $A_{i1} + A_{i3} = 0$ 时,$P_{Wi} = 0$。

式中:A_{i1} 为关键词 W_i 在训练集每篇论文概要中出现且同时被作者标注为关键词的次数之和,A_{i2} 表示关键词 W_i 在训练集每篇论文概要中未出现但被作者标注为关键词的次数之和,A_{i3} 表示关键词 W_i 在训练集每篇论文概要中出现但未被标注为关键词的次数之和。在 Jupyter Notebook 软件开发环境下,采用 Python 3.7 编程语言,A_{i1}、A_{i2}、A_{i3} 抽取的核心代码如下:

```
if word in part.keywords or word in ad:
if word not in dic.keys():
    dicword = Dword()
    dicword.name = word
    wordlis.append(dicword)
    dic[word] = dicword
    #将对象取出
    dicword = dic[word]
#如果在概要和关键词中同时出现
    if word in ad and word in part.keywords:
        dicword.key_count += 1
    #如果在关键词中出现,概要中未出现
    if word in part.keywords and word not in ad:
        dicword.ta_count += 1
    #如果在概要中出现,关键词中未出现
    if word in ad and word not in part.keywords:
        dicword.tk_count += 1
```

通过系统测试发现,由于纺织领域关键词分布的特殊性,仅根据公式(1)得出的先验概率区分精度不够,先验概率相同的关键词个数有 24 531,占总词数的 84.16%。为了增加排序的区分度,本研究借鉴“影响因子百分位”的思想,计算每个候选关键词 W_i 的词频百分位 Z,如式(2) 所示:

$$Z_{W_i} = \frac{(n - r + 0.5)}{n} P_{W_i} \quad (2)$$

式中：n 为关键词集中的关键词总数 29 146，r 为候选关键词在概要库中出现的频次排名，可由 $A_{i1} + A_{i3}$ 计算所得，范围为 1～29 146，最终排序指标为先验概率 Q 乘以词频百分位 Z。验证计算发现，如此得到的排序指标更为合理，并且具有较高区分度，对比结果如图 2、图 3 所示。

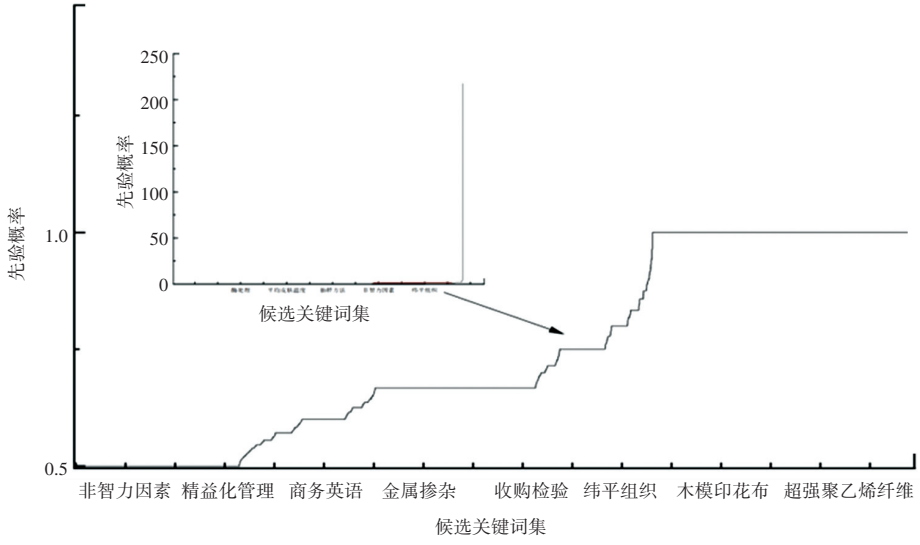

图 2　先验概率排序

Fig. 2　Priori probability ranking

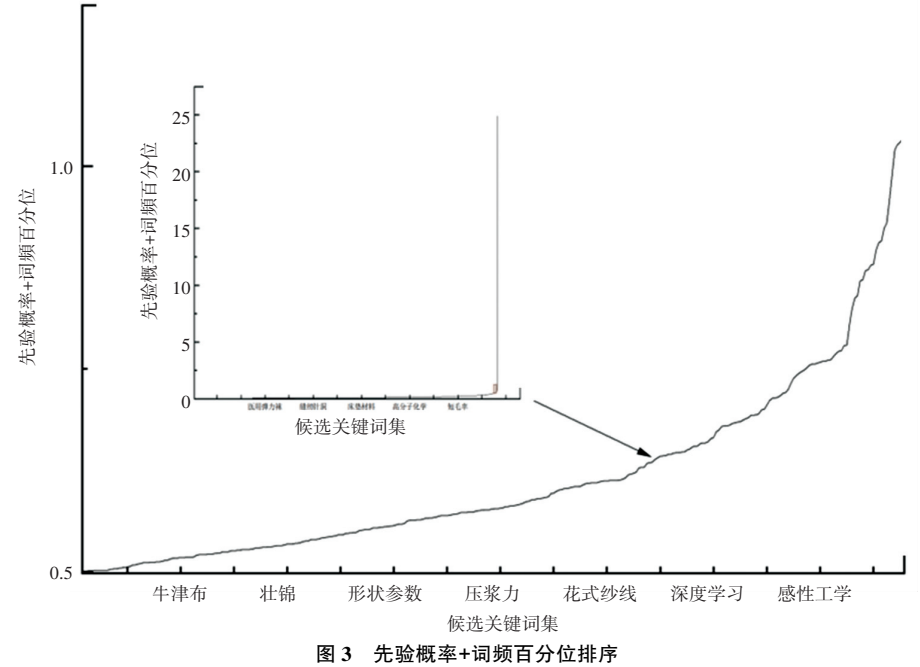

图 3　先验概率+词频百分位排序

Fig. 3　A priori probability+word frequency percentile ranking

2.2.2　词表匹配

根据先验概率和词频百分位得到候选关键词表排序，外部论文素材进入系统后，根据词表的最大匹配法，匹配出论文素材中所有纳入候选关键词集的词，这些词将保存在推荐关键词集合中。

2.3　推荐关键词排序

在得到了推荐关键词列表之后，根据候选关键词所带的排序指标，排名较高的关键词将优先输出。用户可以根据需求，输出论文素材中所有的推荐关键词或设定数量的关键词。在软件系统中，推荐关键词排序并以词云形式输出的核心代码如下：

```
#将有权重的关键词存入候选词库
weightword = Read_txt('* * * *.txt')
weight_dic = {}
```

```
    for part in weightword:
        temp =part.split('\t')
        weight_dic[temp[0]] = round(float(temp
[1]),5)
    #推荐关键词排序并以词云形式输出
    def inside_extract(ts,n):
      temp = {}
      temp_score = 0
        for k,v in weight_dic.items():
            if k in ts:
                temp[k] = v
        temp =Sort_value(temp)
        print(temp)
        k = []
        v = []
        for k1,v1 in temp.items():
            k.append(k1)
            v.append(v1)
        kn = k[:n]
        vn = v[:n]
        words=[(i,j) for i ,j in zip(kn,vn)]
        worldcloud=(
            WordCloud()
            .add("", words, word_size_range =
[12, 55])
                .set_global_opts(title_opts = opts.
TitleOpts(title="关键词词云"))
                )
        worldcloud.render('***.html')
        worldcloud.render_notebook()
```

3 测试结果分析

3.1 对比模型

为了评估本研究提出的关键词排序算法的效果以及设定的候选关键词的权重是否对关键词抽取起到增益效果,引入传统关键词抽取算法作为参照。首先,为了研究文档外部特征的作用,选取了文档内特征算法 TextRank 进行对比,抽取参数的设置参考了 Hasan[24] 等学者的研究成果。同时,本研究也将基于先验概率和词频百分位的纺织论文关键词抽取算法(Prior Knowledge-Extract,PK-Extract)与当前较为成熟的关键词排序算法(PK-TextRank)进行了对比。

3.2 评价指标

设候选关键词集合为 WP,作者标注的原文关键词集合为 TP,算法抽取的推荐关键词集合为 FP,抽取准确率评价指标 P 的计算方法如式(3)所示:

$$P = \frac{FP \cap WP}{TP \cap WP} \tag{3}$$

3.3 结果分析

考虑到对比算法只能对论文中已有的关键词进行抽取,为了公平起见,评价时只考虑论文中已有的关键词。首先,让中心关键词在规模为 11 076 数量的概要库中进行抽取,对这 11 076 篇概要中抽取的准确率计算综合平均值 $F1$,计算过程如式(4)所示,计算结果如图 4 所示。

$$F1 = \frac{\sum_{i}^{G} P_i}{G} \tag{4}$$

其中,$F1$ 为综合平均值;G 为测试集数量;P 为准确率。

由图 4 可以看出,与基准算法的两种无监督算法 TextRank 相比,本研究提出的 PK-Extract 算法在抽取关键词的平均准确率较基准的无监督算法 TextRank 相比有明显提升。PK-Extract 算法相对于 TextRank 算法提升了 171.13%,相对于 PK-TextRank 也有 8.5%的提升,说明本系统有效改进了前人的算法,有望在纺织领域得到应用。

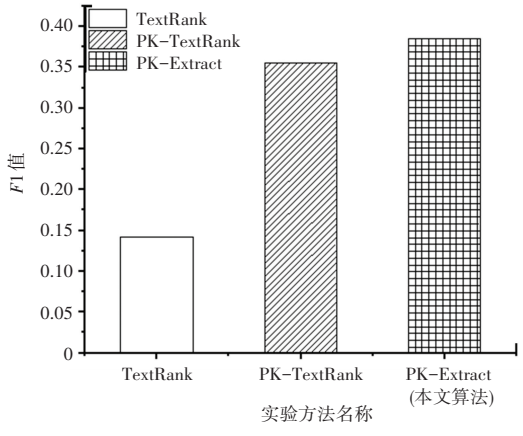

图 4 各算法抽取关键词的平均准确率(*F1* 值)

Fig. 4 Average accuracy rate of keywords extracted by each algorithm (*F1* value)

3.4 抽取效果展示

研究中,作者从纺织领域染整、织造、服装 3 个方向各随机抽取了 1 篇论文,在构建的软件系统中进行关键词抽取。不同关键词抽取算法的抽取效果见表 1。

表 1　抽取效果展示

Table 1　Extraction effect display

案例	原文关键词	不同抽取算法得到的 6 个推荐关键词		
		TextRank（无监督）	PK-TextRank（半监督）	PK-Extract（本文算法）
棉织物冷轧堆前处理氧漂活化剂的制备及其应用。针对传统棉织物冷轧堆前处理过程中存在碱和双氧水用量大，生产周期长，效率低的问题，以自制过渡金属盐混合物、羧甲基纤维素钠（CMC）、焦磷酸钠为原料制备了一种新型氧漂活化剂，并通过单因子及正交试验确定了其在纯棉织物冷轧堆新工艺中最佳应用条件：H_2O_2 质量浓度 30 g/L，NaOH 质量浓度 30 g/L，活化剂质量浓度 6 g/L，渗透剂质量浓度 3 g/L，堆置时间 7 h，轧液率 100%，温度为室温。该工艺条件下，织物白度为 84.2%，强力为 682.7 N，30 min 毛效为 8 cm。结果表明，新工艺不仅可使织物获得较好的处理效果，还能在一定程度上减少碱和双氧水用量，缩短堆置时间，提高效率，降低生产成本。	活化剂，棉织物，冷轧堆，前处理	用量，应用，双氧水，新工艺，棉织物，质量浓度	棉织物，活化剂，前处理，羧甲基纤维素钠，冷轧堆，双氧水	棉织物，活化剂，冷轧堆，前处理，正交试验，羧甲基纤维素钠
静电纺聚氨酯纳米纤维非织造布的制备，研究了聚氨酯在几种常见有机溶剂中的溶解性能，寻求静电纺丝最佳溶剂及配比，并采用静电纺丝法制备纳米级聚氨酯纤维膜。通过改变共混溶剂的质量比、纺丝液的浓度、纺丝电压、挤出速度和接收距离，借助扫描电子显微镜测量纤维的直径，分析了各因素对纤维形貌结构的影响。结果表明：DMF/THF 共混溶剂配比为 1∶3 时，聚氨酯纺丝液静电纺丝效果佳；在纺丝液浓度 8%～12%、纺丝电压 12～30 kV、接收距离 10～30 cm 范围内，能纺制出纤维直径分布在 800～1 500 nm 之间的聚氨酯纳米纤维非织造布。	静电纺丝，聚氨酯，纳米纤维，非织造布	溶剂，纤维，配比，结构，借助，浓度	聚氨酯纤维，静电纺丝，非织造布，纳米纤维，聚氨酯，有机溶剂	静电纺丝，非织造布，聚氨酯，纳米纤维，有机溶剂，纤维直径
基于深度学习的化妆品塑料瓶缺陷检测。提出一种基于深度卷积神经网络的化妆品塑料瓶表面缺陷检测算法。采用百万像素级别的工业相机采集大量的塑料瓶图像样本，并通过 HSV（Hue，Saturation，Value）颜色空间变换和 Otsu 阈值分割等方法对图像进行预处理。采用随机图像变换法对数据集进行增强，并对图像进行标准归一化处理。在卷积神经网络模型中应用深度可分离卷积和 Dropout 技术以减少参数量，从而避免过度拟合。使用训练样本集训练该模型，并在测试集中将结果与几种经典图像识别算法进行比较分析，结果显示，本文算法的识别准确率高达约 0.97。由此表明本文算法的效果优于其他经典算法，有望将其应用于化妆品塑料瓶缺陷检测的工业自动化系统，以提升缺陷识别效果，从而提高生产效率。	深度学习，缺陷检测，卷积神经网络，化妆品塑料瓶	算法，图像，进行，背景，采用，缺陷	缺陷识别，图像识别，卷积神经网络，深度学习，深度可分离卷积，缺陷检测	深度学习，卷积神经网络，缺陷检测，图像识别，阈值分割，化妆品

案例 1 染整数据中，TextRank 偏向于给一些泛化词赋予更高的权重，导致将应用等排在较高的位置抽取出来；PK-TextRank 提取的词中，主要是冷轧堆未抽取出，羧甲基纤维素钠在训练集中的关键词词频为 7，概要词频为 26，冷轧堆在训练集中的关键词词频为 80，概要词频为 328，这样两者在 PK-TextRank 的处理下是羧甲基纤维素钠的权重高于冷轧堆的，但是冷轧堆的 $F1$ 值是经过 80 次验证的 0.245，羧甲基纤维素钠的 $F1$ 值只经过 7 次验证的 0.269，综合下来冷轧堆的权重应该高于羧甲基纤维素钠。

案例 2 织造数据中，TextRank 仍然将纤维、结构等泛化词抽取了出来，在算法的底层数据中，这类词所占的比重过高，导致虽然限定在词表中抽取，在排序输出和作者给出相同数量的关键词时，过低权重的关键词不能输出来；PK-TextRank 抽取的词中，虽然聚氨酯纤维和聚氨酯是两个相同的概念，但是聚氨酯在论文概要中出现的次数为 883，关键词出现次数为 263，聚氨酯纤维在论文概要中出现的次数为 9，关键词出现次数为 10，因此在 PK-TextRank 的处理下，聚氨酯纤维的权重高于聚氨酯，但是聚氨酯的 $F1$ 值是经过 883 次验证的 0.298，聚氨酯纤维的 $F1$ 值是经过 9 次验证的 1.11，因此聚氨酯更适合作为关键词。

案例 3 服装数据中,塑料瓶在概要中出现了 4 次,在训练集中出现 238 次;算法这个词在给定概要中出现了 5 次,在训练集中出现1 398 次,出现频率都很高,因此 TextRank 就将塑料瓶权重设定过高,但是塑料瓶等都属于比较泛化的词;缺陷识别在训练集中关键词词频为 3、概要词频为 2,缺陷检测在训练集中关键词词频为 10、概要词频为 23,缺陷检测的在论文中出现的次数是高于缺陷识别的,社会认可度较高,但是未经过词频百分位处理时,缺陷识别先验概率(1.5)高于缺陷检测先验概率(0.435),这不符合公众认知。

3.5 关键词抽取系统的扩展应用

本文提出的基于先验概率的纺织论文关键词自动抽取算法,不仅可以对纺织论文摘要进行关键词抽取,还可以对纺织行业的新闻报道进行关键词抽取,测试效果如图 5 所示。

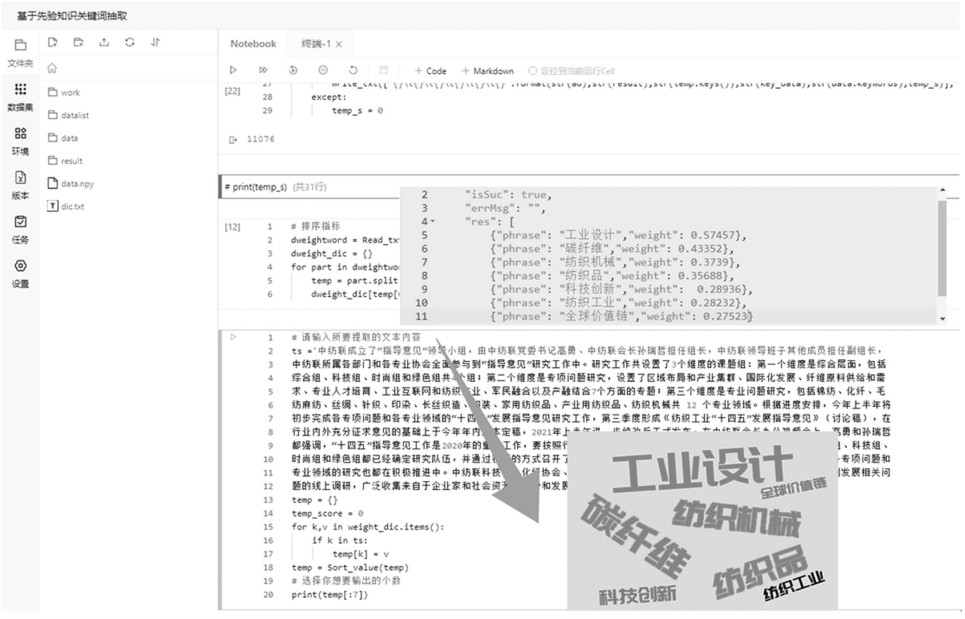

图 5 关键词抽取系统界面展示

Fig. 5 Keyword extraction system interface display

选用十四五规划大纲中的部分文字进行测试试验,输出关键词数量为 7,输出的关键词结果较为合理。

4 结束语

本文提出了一种基于先验知识的纺织论文关键词自动抽取算法,并开发了软件系统进行大规模数据训练和测试。算法的主要特点是使用选择概率量化了候选关键词在特定领域的使用情况,将得到的概率作为先验概率,与 PK-TextRank 不同,本文算法借鉴了“影响因子百分位”的思想。词频百分位的加入,使每个候选词的先验概率都不相同,使相同先验概率词的排名有了依据,并改善候选词权重排名具有合理性。实验结果表明,先验概率和词频百分位的加入,可以有效地提升关键词的排序性能,相较于基准算法有 170% 左右的提升,对于 PK-TextRank 算法也有 8.5%的提升。经过完善和优化,基于先验概率和词频百分位算法的软件系统有望在更多学科领域得到应用。

虽然 PK-Extract 算法有效地提升了系统候选词的排序性能,但开发的软件系统仍然存在一些不足。仅使用了文档的标题和摘要进行抽取会导致抽取不够完全,如在测试概要集中,有 20% 的概要未在标题或摘要中包含作者关键词,导致无法抽取,影响了关键词的抽取效果,这也是本研究未来的研究重点。除了探索采用全文中能覆盖作者关键词的内容作为关键词提取的源文本外,后续还可研究更多能够表达领域特征的论文信息,并重点解决关键词“一词多义”的问题。

参考文献

[1] 于超宁. 大数据视域下期刊关键词的应用分析[J]. 采.写.编, 2022,186(1):126-127.
[2] 谢晓红, 王淑华, 肖骏. 地学科技论文关键词选取存在的主要问题探讨[J]. 编辑学报, 2013,25(S1): 23-24.
[3] 索传军, 葛倩, 魏长青. 基于论题视角的图情中文期刊论文关键词标注探析——以“基于”类论文为例[J]. 图书情报工作, 2022, 66(12): 117.
[4] 刘庆颖, 陈庄. 水产学术论文的中英文关键词标引[J]. 农业图书情报学刊, 2005, 17(5): 139-142.

[5] 柏晶瑜. 学术论文中文摘要撰写常见问题浅析[J]. 中国科技期刊研究，2008，19(4)：687.
[6] 罗婉丽，张磊. 结合拓扑势与 TextRank 算法的关键词提取方法[J]. 计算机应用与软件，2022，39(1)：334-338.
[7] 杨兵，聂铁铮，申德荣，等. 一种面向医学文本数据的结构化信息抽取方法[J]. 小型微型计算机系统，2019，40(7)：1479-1485.
[8] WU Y B，LI Q，BOT R S，et al. Domain-specific keyphrase extraction [C]//Proceedings of the 14th ACM International Conference on Information and Knowledge Management，2005：283-284.
[9] SALTON G，YANG C S. On the specification of term values in automatic indexing[J]. Journal of Documentation，1973，29(4)：351-372.
[10] MATSUO Y，ISHIZUKA M. Keyword extraction from a single document using word co-occurrence statistical information[J]. International Journal on Artificial Intelligence Tools，2004，13(1)：157-169.
[11] 张柏秋，吴晓鐨. 科技查新检索中的关键词选择[J]. 情报科学，2008，205(9)：1344-1348.
[12] 方俊伟，崔浩冉，贺国秀，等. 基于先验知识 TextRank 的学术文本关键词抽取[J]. 情报科学，2019，37(3)：75-80.
[13] 胡少虎，张颖怡，章成志. 关键词提取研究综述[J]. 数据分析与知识发现，2020，5(3)：45-59.
[14] 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.
[15] SALTON G，YANG C S，YU C T. A theory of term importance in automatic text analysis[J]. Journal of the American Society for Information Science，1975，26(1)：33-44.
[16] BLEI D M，NG A Y，JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research，2003，3(1)：993-1022.
[17] 杨洁，季铎，蔡东风，等. 基于 TextRank 的多文档关键词抽取技术[C]//第四届全国信息检索与内容安全学术会议论文集(上). 2008：397-404.
[18] MIHALCEA R，TARAU P. Textrank：Bringing order into text[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing，2004：404-411.
[19] 李鹏，王斌，石志伟，等. Tag-TextRank：一种基于 Tag 的网页关键词抽取方法[J]. 计算机研究与发展，2012，49(11)：2344-2351.
[20] 梦彤，谷晓燕，刘甜甜. 基于改进 TextRank 的关键句提取方法[J]. 郑州大学学报(理学版)，2023，55(1)：15-20.
[21] 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，2014，30(7)：41-47.
[22] 李晨庚，谢四江. 基于改进的 TextRank 算法的计算机辅助定密研究[J]. 计算机应用与软件，2022，39(3)：336-340，345.
[23] 俞立平. “影响因子百分位”指标的特点研究[J]. 图书情报工作，2016，60(10)：103.
[24] HASAN K S，NG V. Conundrums in unsupervised keyphrase extraction：making sense of the state-of-the-art [C]// International Conference on Computational Linguistics：Posters. Association for Computational Linguistics，2010：365-373.

(上接第 7 页)

[10] CUBUK E D，ZOPH B，SHLENS J，et al. Randaugment：Practical automated data augmentation with a reduced search space[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2020：702-703.
[11] GIRSHICK R. Fast R-CNN[C]// Proceedings of the IEEE Conference on International Conference on Computer Vision. Piscataway，NJ：IEEE，2015：1440-1448.