新时代人民日报分词语料库下关键词抽取及分析研究

周　好[1, 2]　王东波[1, 2]　黄水清[1, 2]
（1. 南京农业大学信息管理学院，南京　210095；
2. 南京农业大学人文与社会计算研究中心，南京　210095）

摘　要：[**目的 / 意义**] 面对海量的新闻文本，通过提取少量能表征其内容的关键词，来帮助用户快速掌握新闻内容，是关键词提取的首要任务。[**方法 / 过程**] 本文以新时代人民日报分词语料库中部分语料作为研究对象，主要对比 TF-IDF、TextRank、LDA、LSI、Rake、Yake 六种无监督关键词抽取方法的抽取效果，并对抽取结果进行分析。[**结果 / 结论**] 结果显示：在 Pooling 评价方法下，TF-IDF 算法以及 Yake 算法在大规模人民日报关键词提取任务中表现最优，TextRank 算法性能尚可。另外，通过对政治、经济、社会类别下的关键词进行分析，可快速发现、梳理当月的重要事件。本文的研究可为新闻报刊语料的关键词提取分析提供参考。

关键词：关键词抽取　新时代人民日报分词语料　无监督抽取方法
分类号：G255
DOI：10.31193/SSAP.J.ISSN.2096-6695.2022.01.02

0　引　言

关键词对于新闻文本至关重要。面对海量、庞杂的新闻文本，逐一阅读所有的文字信息显然是不现实的。通过阅读和查找关键词，能帮助用户快速了解、定位新闻的相关信息，促进信息的快速处理，节省大量时间。通过对新闻文本进行关键词抽取，可以自动从文本中提取最常用和最重要的词汇，帮助总结整个新闻文本的主题思想，也可进一步应用于新闻文本的推荐和搜索。

鉴于此，本研究从黄水清团队自行构建的新时代人民日报分词语料库（New Era People's Daily Segmented Corpus，简称 NEPD）[1] 中选取 2015 年 1 月、2015 年 6 月、2016 年 1 月、2018

[作者简介] 周好（ORCID：0000-0003-1721-3718），女，博士研究生，研究方向为自然语言处理，E-mail:2019214003@njau.edu.cn；王东波（ORCID：0000-0002-9894-9550），男，教授，博士生导师，研究方向为自然语言处理与知识挖掘、信息计量、数字人文，Email：db.wang@njau.edu.cn；黄水清（ORCID：0000-0002-1646-9300），男，教授，博士生导师，研究方向为数字人文、文本挖掘、信息计量，Email:sqhuang@njau.edu.cn（通讯作者）。

年 1 月共计四个月的精加工语料为研究对象，主要比较了六种无监督关键词抽取算法的性能，并对得到的部分新闻文本的关键词进行分析。这一探究不仅能通过关键词从词汇角度对过去的热点事件进行梳理与比较，还能找到一种有效的新闻文本无监督关键词提取办法。

1 研究现状

随着自然语言处理技术的发展，关键词抽取成为继分词之后文本分析的另一重要基础，在诸多的任务中发挥了重要作用。关键词抽取算法主要分为基于有监督和基于无监督两类。其中，有监督学习需要提供已经标注好的训练语料，虽然其抽取效果良好，但对于大规模语料来说，需要消耗大量的人工劳动，其适用性不强。而无监督学习因无需对数据进行训练，仅需要文本自身的信息就能进行关键词的抽取，在实际任务中被广泛采用。有监督的关键词抽取方法需人工预先确定关键词特征，随后对数据进行训练。鉴于语料规模较大，人工标注语料显然不现实。本文重点关注仅需文本自身信息的无监督关键词抽取方法。现有无监督关键词抽取方法根据其抽取原理的不同，主要可分为三类：第一类是基于统计特征的关键词抽取方法，以 TF-IDF 和 Yake 为代表；第二类是基于词图模型的关键词抽取方法，以 TextRank 和 Rake 为代表；第三类是基于主题模型的关键词抽取方法，以 LDA 和 LSI 为代表。

国内许多学者已经用基于无监督的关键词抽取方法在中文文本中做了大量的研究工作。在以往的研究中，以 TF-IDF、TextRank、LDA 三者的应用最广泛。为了进一步提升关键词抽取的效果，诸多学者在上述基本算法的基础上进行了不同层面的优化。优化方式主要可分为两种：一是将词语内外部特征融入抽取算法中，二是将多种算法的优点进行融合。牛永洁[2]等在 TF-IDF 算法中对词语的位置、词性、关联性、词长和词跨度五个因素进行综合考虑，赋予不同因素不同的权重，经对比，抽取性能均优于经典方法。杨凯艳[3]从文本内外部同时着手，对传统的 TF-IDF 算法进行了改进。针对文本外部，增加信息增益和离散度二者的考量。针对文本内部，综合考虑词频、词性、词长、词位置、词跨度五种属性。顾益军[4]等、刘啸剑[5]等均将 TextRank 与 LDA 融合在一起，前者构建了融入主题影响力的迁移矩阵，后者将短语作为节点构建戴荃无向图，二者均不同程度的提高了关键词抽取的效果。朱泽德[6]等基于 LDA 模型，以文档隐含主题分析为基础，提出一种新的关键词抽取算法 TF-ITF，即词频 – 逆主题频率。宁珊[7]等基于 TextRank 并融合基于 word2vec 模型和 LSTM 模型的语义相关性影响度、基于 LDA 模型的主题差异性影响度，得到最终关键词排序。夏天通过分别将词语位置加权[8]、词向量聚类加权[9]融入到 TextRank 的计算过程中，实现了较好的关键词抽取效果。杨延娇[10]等针对 TextRank 方法在词语关系判断不合理、词性覆盖不全、无关词语过多等不足，提出了具有针对性的 S-TextRank 方法，改进后的关键词抽取准确率达到 74%。除上述三种主流算法之外，Rake 算法在中文文本关键词抽取中也有少量应用，而 Yake 和 LSI 算法鲜有出现。徐明明[11]等在 Rake 算法中将词语位置作为特征加入，用以提取亚马逊商品信息中的关键词。陈可嘉[12]等依托 Rake 核心算法，从文本预处理、共现矩阵构造以及关键词过滤三方面将词语特征融入算法流程中，改进后的 Rake 算法较之原算法的性能有所提高。

而关于《人民日报》语料的研究，则主要涉及新闻学、语言学、图书情报学、政治学等领域。其主要研究视角为采用不同的方法，从词汇角度出发进行历时定量分析。郑成郎[13]通过对不同时期《人民日报社论全集》逐年的主题词进行历时统计、分类、比较、分析，来考察社会历时的发展变化轨迹，以此促进《汉语主题词表》的修订。刘晓丽[14]以《人民日报社论全集》的语料为研究对象，通过分词、计量等手段，对不同年份的高频词、敏感词进行统计分析。董志成[15]基于计量语言学、统计学等原理，从词长特征、词汇丰富度程度、主题集中度、高频词四方面对 1946 年至 2016 年间《人民日报》语料样本进行了历时性分析。李琪[16]以 1946 年至 2016 年《人民日报》语料为研究对象，通过对关键词进行提取并进行分析，寻找适合报刊语料的关键词抽取办法和量化的研究方法。黄水清[17]等基于其自行构建的新时代人民日报分词语料库，从历时角度，比较 1998 年和 2018 年《人民日报》的字与词单位上的句子长度分布，以及词汇静态分布情况，并通过对比统计 2020 年中央一号文件中相关的关键词在各年度 1 月份《人民日报》中出现的频次，证明用人民日报语料对政策文件做历时性词频分析的可行性[18]。饶高琦[19]等基于 70 年跨度的《人民日报》语料，使用 TF-IDF、互信息、联合熵等共计九种计算方法抽取历时稳态词候选词集，并对该词集进行特征统计。

前期研究无疑取得了丰硕的成果，但前期研究中的语料大多没有经过人工逐一校对分词，并且，针对新闻文本的关键词抽取算法对比与分析研究相对较少。因此，在上述研究的基础上，本文主要采用三类主流无监督思想，基于 TF-IDF、Yake、TextRank、Rake、LDA 和 LSI 六种具体算法，对 NEPD 语料进行关键词抽取对比实验。不仅对模型的关键词抽取结果进行评价，同时，也从多角度对抽取结果进行系统的统计与分析。

2　各类算法对比介绍

无监督的关键词抽取方法一般遵循如下步骤[20]：（1）文本预处理。该步骤主要包含分句、分词、去停用词、词性标注等操作。（2）确定候选词集。该步骤主要是在文本预处理的基础上，添加一些特征对候选词进行筛选，通常是根据一些指标，比如词频、特定词性、词长、词汇所在位置等。（3）对候选词进行排序。在确定了候选词集后，根据不同的方法，设定一系列的指标量化候选词集，对其重要性进行排序，选取 Top K 个候选词作为文档的关键词。（4）效果评估。在获得关键词集后，需要对关键词抽取的效果进行评估。最广泛的方法有计算 P、R、F 值以及 Pooling 评价方法等。但各类算法所依托的关键词抽取原理不同，其实现流程也有所区别。

2.1　基于统计思想的关键词抽取算法

基于统计思想的关键词抽取方法是最简单有效的，其中，TF-IDF 是最基本的方法。通常来说，词频是判断一个词在一篇文档中是否重要的标准。但根据“齐普夫定律”[21]，并不是出现次数越多的词就越重要。TF-IDF 综合考虑二者，出于统计思想来评估一个字词对于整个文档的重要程度。简单来说，TF-IDF= 词频（TF）* 逆向文档频率（IDF）。其中，TF 是指这个词在文档中出现的频率，IDF 是指该词在整个语料库中出现的频率。如果某个词的 TF 高、IDF 低，则认为此词具有很好的类别区分能力。计算公式如公式 1 所示，其中，IDF 的分母加 1 是为了防止分

母为 0 的情况。

$$TF-IDF = TF \times IDF = \frac{该词频次}{文章的总词数} * \log \frac{语料库的文档总数}{包含该词的文档数+1} \quad (1)$$

该算法的关键词抽取步骤如下：

（1）对文本进行预处理，得到候选关键词的集合。

（2）根据公式，计算每个词汇在每个文档中的词频和在所有文档的逆文档频率，从而计算得到每个词汇的 TF–IDF，并重复上述步骤，直至得到所有词汇的 TF–IDF 值。

（3）对所有词汇的 IF–IDF 值进行倒序排序，得到排名靠前的 Top K 个词汇作为指定文档的关键词。

同样作为基于统计特征的抽取算法，Yake 算法[22]不依赖词典或同义词库，也不依赖任何语料库进行训练。作为一种从文本中提取特征的无监督的方法而言，经开发者测试，在与其它十余种关键词提取办法的对比测试中，展现出优良的性能，是目前无监督关键词提取中效果较好的。相比部分图模型，Yake 算法速度快，而且可以提出关键短语。与 TF–IDF 相比，它可以在单个文档的基础上提取关键字，不需要大型语料库。

该算法的关键词抽取步骤如下：

（1）对文本进行预处理，去停用词。

（2）特征提取。Yake 算法计算文档中词语的五个统计特征：大小写（中文无效，不做考虑）、词语位置（越靠前的词语越重要）、TF 值、上下文关联性、词频。其中，Yake 的源码并不能直接作用于中文的关键词提取，需要对其进行调整使其适用于中文文本。比如，源码中将长度小于 3 的字符直接定义为停用词，在英文中是适用的，但对于中文来说，单字都有可能是有实义的，因此，需将字符长度设置为小于等于 1。

（3）计算术语分数，生成 n–gram 并计算关键词分数。

（4）删除相似的关键词，并对关键词列表进行排序。

2.2　基于词图方法的关键词抽取算法

基于词图方法的关键词抽取主要通过考虑图的结构来对顶点重要性进行评分排序。最著名的基于词图思想的方法之一是 TextRank。TextRank 算法的思想源于 PageRank，主要是通过把文本切分为若干组成单元（单词、短语或者句子）并建立图模型，在构建图模型的时候将节点由原来的网页改成了句子，并将文档看作一个词的网络，该网络中的链接表示词与词之间的语义关系，最后利用投票机制对文本中的重要程度进行排序，仅利用单篇文档本身的信息即可实现关键词提取。

该算法的关键词抽取步骤如下：

（1）对文本进行预处理，得到句子和候选关键词的集合。

（2）构建词共现图。将所有单词排成一个序列，在指定窗口内的所有单词都是图中的节点，采用共现关系构造任意两节点之间的边。两个节点之间存在边仅当它们对应的词汇在长度为 K 的窗口中共现，K 表示窗口大小，即最多共现 K 个单词。

（3）关键词图排名并计算关键词得分。依照连接节点的多少，给每个节点赋予一个初始的权

重数值，经过不断迭代、传递，直到收敛。计算每个单词的得分，对节点权重进行倒序排序，从而得到最重要的 K 个单词作为指定文档的关键词。

$$WS(V_i)=(1-\mathrm{d})+\mathrm{d}*\sum_{V_i\in(V_i)}\frac{W_{ij}}{\sum_{V_k\in Out(V_j)}W_{jk}}W_{ij} \quad (2)$$

其中，V_i、V_j 表示任意两点，W_{ij} 表示两点之间边的权重，Out（V_i）是指向点 V_i 的集合，d 为阻尼系数，一般取值为 0.85。

Rake 是另一种基于词图的关键词提取算法，相对于 TextRank 而言，Rake 考虑关键词内部的共现而不是固定窗口，且评分程序更简单、更具统计性。Rake 算法[23] 自 2010 年被提出以来，在关键词提取上展现出优良的性能。所提取的关键词并不是单一的词汇，也有可能是短语，还能提取一些较长的专业术语。

Rake 算法提取关键词的思路主要由以下几步骤构成：

（1）文本预处理，构建候选关键词集。对于给定的一篇文档，以标点符号及停用词作为分隔符，将其作为候选的关键词。

（2）构建关键词共现图，并且给每次词汇评分。对于每个词汇，根据其字符长度、词频、词汇之间的共现关系，计算该词汇的得分。

（3）将每个候选词的得分累加并进行排序输出，得到最终关键词。值得注意的是，Rake 算法倾向于较长的词汇。

2.3 基于主题模型的关键词抽取算法

基于主题模型的关键词抽取方法认为，词和文档之间并没有直接的联系，每个文档中存在若干主题，在主题把词串联起来的时候，就能得到了每个文档的关键词分布。LSI 是最早出现的主题模型，它的算法原理很简单，一次奇异值分解就可以得到主题模型，同时解决词义的问题。LDA 算法的理论基础是贝叶斯理论，主要利用文档中词汇的共现关系，对词汇按照主题进行聚类，得到“文档-主题”和“主题-单词”两个概率分布，进而将词和文本映射到同一语义空间。通俗来说，LDA 模型的主要工作就是根据给定的一篇文档来反推其主题分布。

本文使用 Gensim 库进行 LDA 和 LSI 主题模型关键词提取实验。主要的步骤如下：

（1）对文本进行分词、去停用词等处理。

（2）构建语料库，对文本中的每个词赋予一个主题。对语料库中的每个词，逐词重新按照吉布斯采样公式重新采样其主题，直至收敛。

（3）统计文档中主题的分布。选择 LDA 或是 LSI 进行模型训练，得到每个文档的主题分布。通过计算余弦距离，得到权重最大的主题下的关键词。

3 实验流程与语料信息

3.1 语料基本信息

本文以 NEPD 语料库中 2015 年 1 月、2015 年 6 月、2016 年 1 月、2018 年 1 月共计 4 个月经过人工分词标注加工的精语料为研究对象，数据真实、可靠、准确。经统计，NEPD 语料库中

2015 年 1 月、2015 年 6 月、2016 年 1 月和 2018 年 1 月四个月的语料规模共计 48923KB，总体字符数 8580560 个，总计 314963 个句子，去除停用词后共计 7192461 个词汇。从表 1 可以看出，各月份之间新闻的总词数差别不大。

表 1　NEPD 语料基本信息

语料	文件大小	总字符数	总句数	总词数
2015 年 1 月	12462KB	2189308 个	84393 句	1837389 个
2015 年 6 月	11960KB	2093036 个	79209 句	1755150 个
2016 年 1 月	12629KB	2205230 个	84576 句	1848703 个
2018 年 1 月	11872KB	2092986 个	66785 句	1751219 个

根据人工分词后的语料，统计出每个词汇的长度以及它们的频次和占比。从表 2 可以看出，NEPD 语料中的词汇仍以单字词（词长为 1）和二字词（词长为 2）居多，二者之和占比约为 93%。词长为 5 或者以上的词多为固定搭配词汇或是英文字符。

表 2　NEPD 语料词长信息

词长	2015 年 1 月		2015 年 6 月		2016 年 1 月		2018 年 1 月	
	频次	占比	频次	占比	频次	占比	频次	占比
1	655968	35.7%	608617	34.68%	640168	34.63%	616692	35.22%
2	1059427	57.66%	1027760	58.56%	1079976	58.42%	1010868	57.72%
3	85285	4.64%	84472	4.81%	90736	4.91%	87173	4.98%
4	31201	1.7%	28232	1.61%	31918	1.72%	30659	1.75%
5	3320	0.18%	3526	0.20%	3722	0.20%	3316	0.19%
6	1069	0.06%	1330	0.08%	1050	0.06%	1108	0.06%
7	576	0.03%	573	0.03%	610	0.03%	534	0.03%
8	249	0.01%	210	0.01%	302	0.02%	221	0.01%
9 及以上	294	0.02%	430	0.02%	221	0.01%	648	0.04%

3.2　实验流程

首先，在本研究中，将四个月的精分词语料按照句号（。）、感叹号（！）、问号（？）、省略号（……）、引号（“”）、分号（；）、冒号（：）、其他［如：括号（）］八类进行分句处理。

其次，候选词集的好坏会直接影响关键词抽取的质量。停用词的选取能在很大程度上过滤掉非相关词汇。本研究选取的基础停用词表是综合哈工大停用词库、四川大学机器学习智能实验室停用词库、百度停用词表以及网上各种资源，其中包括数字、符号、标点和无实际意义的词汇。对其进行去重和删除其中的英文部分以及明显无关的字符，最终得到包含 1764 个词汇的现代汉语停用词表。

考虑到本研究的文本为新闻语料，在现代汉语停用词表的基础上，根据“齐普夫定律”，对NEPD 四个月的语料分别进行词频统计。为更准确的筛选出候选词集，在此将各自频次出现在500 次以上的词汇认定为高频词。由于高频词并非均为停用词，因此本研究在词频统计基础上，对得到的 2297 个词汇进行人工去重，最终对得到的 688 个词汇逐一确认其是否可以被界定为停用词。新增停用词包括记者、本报、新华社、摄、电、报道、版、新闻、发布、年、月、日等。

并且，考虑到本语料对人名采用姓和名分开标注、数词与计量单位采用分开标注、地名采用分开标注处理。比如，在“广东省罗定市（县级市）市委书记万木林说”这一表述中，具体的标注结果为“广东 / 省 / 罗定 / 市 /（/ 县级 / 市 /）/ 市委 / 书记 / 万 / 木林 / 说 /”；在“全面推进依法治国是一个系统工程”这一表述中，具体标注结果为“全面 / 推进 / 依法 / 治国 / 是 / 一 / 个 / 系统 / 工程 /”。因此，在考虑词汇的频次外，还将高频、常见的姓氏以及数词、计量单位、“省”“市”“县”等列入停用词表。经过最终筛查，结合基础的现代汉语停用词表，本研究的最终停用词表共包含 2053 个词汇。去除停用词后，余下的词汇组合成候选词集。

最后，依据各类算法对四个月的关键词分别进行抽取实验。

3.3 语料整体分析

词频是语料中一个极为重要的信息，它能直观的展示不同的词在语料中的使用次数。为初步了解 NEPD 语料中的词汇信息，本文从词频角度出发，分别统计四个月语料中的高频词。由于所含词汇较多，仅将频次大于 1000 的词汇选作高频词，在去除停用词后对其进行统计分析。据统计，2015 年 1 月、2015 年 6 月、2016 年 1 月、2018 年 1 月频次超过 1000 的词汇分别为 84 个、81 个、89 个、89 个。由于《人民日报》多是通用词汇，为方便统计与分析，特整理出四个月的共有词汇。如表 3 所示，四个月通用的共有词共计 66 个，各个月份之间的高频词也多有重合。此外，2015 年至 2018 年四个月份独有高频词（深色粗体字部分），分别为 4 个、3 个、5 个、9 个。

表 3　高频词展示

语料	高频词汇
四个月共有词	中国 发展 国家 工作 经济 社会 改革 建设 企业 文化 服务 政府 市场 制度 管理 北京 合作 干部 人民 创新 部门 中央 世界 群众 国际 全国 环境 党 推进 组织 我国 项目 生活 技术 公司 地区 活动 政策 历史 关系 人员 领导 建立 城市 精神 信息 研究 保护 政治 传统 教育 机制 产业 资源 时间 会议 机构 基础 中心 解决 领域 美国 保障 农村 能力 水平
2015 年 1 月余下词	责任 法律 单位 重点 投资 金融 地方 **机关** 体系 **法治** 生产 生态 资金 农民 网络 **收入** 支持 **行政**
2015 年 6 月余下词	投资 生态 法律 战略 支持 **交流** **创业** 责任 全球 资金 **银行** 产品 地方 单位 标准
2016 年 1 月余下词	质量 战略 **代表** **主席** 责任 生产 农业 体系 支持 地方 落实 全球 投资 治理 科技 **监督** **完善** 产品 网络 标准 金融 **规划** 重点
2018 年 1 月余下词	生态 质量 体系 乡村 **十九大** **特色** **平台** **社会主义** 治理 **时代** **工程** 全球 科技 **新时代** 网络 战略 农业 金融 **基层** 落实 **学习** 生产 农民

就共有词而言，除去“中国”“国家”“全国”“我国”等表示主体的词之外，《人民日报》这四个月的主要内容围绕经济、社会、国际、文化等方面展开，且国家的发展与建设问题是其核

心内容。首先，“改革”“市场”“投资”“金融”“资金”等词语反映出我国重视经济发展。其次，“美国”“全球”“世界”“国际”等字眼的出现，说明《人民日报》对国际形势以及其他国家的状况也十分关注。另外，“农村”“农民”“农业”等词陆续的出现，可见《人民日报》对“三农”问题给予的高度重视，这也显示出这个问题对国家发展的重要性。另外，阶段性的重大历史事件也能通过高频词来体现。比如，2014 年反腐工作取得良好效果，“机关”“法治”等词在 2015 年 1 月新闻中高频出现，不仅是对 2014 年取得成绩的总结，更意味着 2015 年反腐工作仍将继续进行；2015 年被称为中国创业元年，关于创业发生了很多事，2015 年 6 月的高频词也正是体现了这一点；2017 年 10 月 18 日至 10 月 24 日在北京召开了十九大，“十九大”“新时代”“社会主义”“特色”等与十九大相关的词汇出现在 2018 年 1 月份的语料中。

图 1　2015 年 1 月总词频词云图

图 2　2015 年 6 月总词频词云图

图 3　2016 年 1 月总词频词图

图 4　2018 年 1 月总词频词图

4　实验结果分析

4.1　抽取效果对比分析

对于本文的关键词提取任务而言，要想做到十分客观去评价提取的性能十分困难。一是因为一个月的 NEPD 语料数量较大，且范围广；二是关键词没有统一的标准，关键词的评定很主观。因此，为了能比较各算法的性能，本文选取一种综合来看较为客观合理的办法——Pooling 评价

方法。具体而言，本文通过人工判定，对每月六种算法下提取的排名前 500 的关键词进行整合，最终得到总体排名前 500 的关键词。通过对照各算法抽取结果的重合率，以判别其各自的关键词抽取准确情况。通过人工对各个算法下关键词提取结果的筛选和计算，图 5 ~ 6 为六种算法在四个月 NEPD 语料下的具体性能以及平均性能，在本文的关键词提取实验中，表现最佳的为 TF-IDF 算法、Yake 算法，其次是 TextRank 算法，主题模型 LSI 和 LDA 算法表现最为不佳，Rake 算法准确率平均在 27.70%。

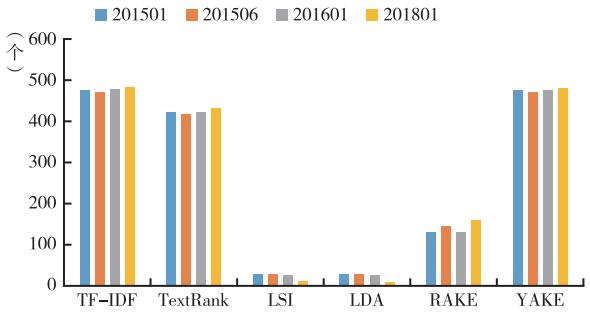

图 5　各算法关键词提取性能对比

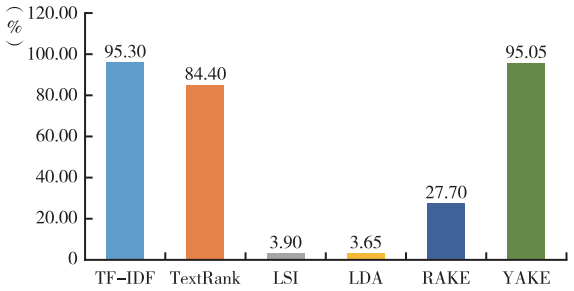

图 6　各算法平均性能对比

在人工浏览算法提取出的关键词的过程中，发现表现效果最佳的三种模型提取出来的关键词与高频词存在大量重复。图 7 显示了最终选取的 500 关键词与各月排名前 500 的高频词的重复数。在构建跟语料库相关的停用词表时，选取的是频次在 500 次以上的高频词，但由于诸多词汇虽然在 NEPD 语料中属于常用词，但其仍具有实体意义，因此本文并没有将其划分至停用词范畴，比如“中国”“党”“人民”“国家”等。另外，通过计算，如图 8 所示，高频词在整个文本中平均占比 20% 左右，而四个语料的高频词经去重之后，得到的 688 个词汇在语料中占比极大。

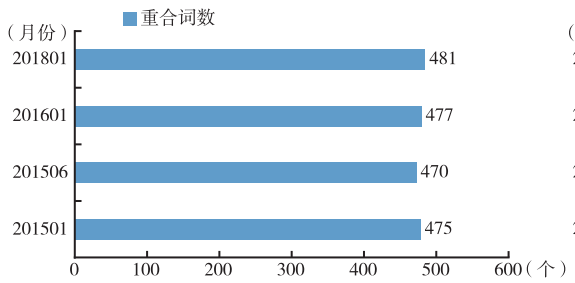

图 7　关键词与高频词重复数

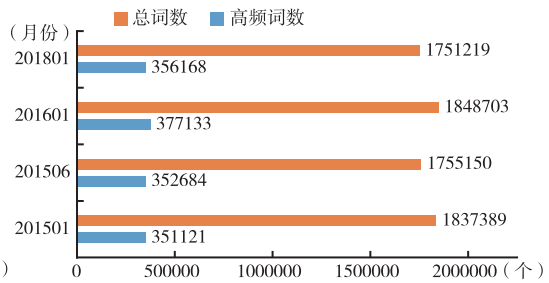

图 8　高频词与总词数比例

考虑到 TF-IDF 算法的主要思想是通过词频来衡量一个词汇的重要程度，因此虽然它没有考虑到词汇之间的关联性，也无法处理一词多义与一义多词的情况，很多情况下不够全面，但有些情况下仍能反映出文档的主题。Yake 算法一开始表现并不佳，在经过对源码进行调参后，调整出适用中文文本的模型，在多个特征的加持下以及停用词表的作用下，最终得到的结果跟高频

词类似。TextRank 算法认为文档或句子中相邻的词语重要性是相互影响的，所以引入了词语的顺序信息。虽然用到了词汇之间的关联性，能将相邻的词汇链接起来，但是仍然倾向于将频繁出现的词汇作为关键词。且涉及网络构建和随机游走的迭代算法，效率不高，计算量大。所抽取的关键词虽能反映整个文档，但代表整个文档的词并不一定是该文档所独有的。由于 Rake 算法主要解决的问题是找包含高频词的多词短语，但在本文的实验中，将其分词符设定为间隔符，对于 Rake 算法而言，这改变了其源码的运算与输出规则，虽然它的强大之处在于它的易用性，但由于后期并没有根据前期调整对参数进行对应调整，导致效果不佳。主题模型认为文章是由主题组成，文章中的词是以一定概率从主题中选取的。不同的主题下，词语出现的概率分布是不同的。根据词的共现信息的分析，拟合出词 – 文档 – 主题的分布，进而将词、文本都映射到一个语义空间中。算法利用文档的隐含语义信息来提取关键词，但是主题模型提取的关键词比较宽泛，不能很好的反映文档主题。

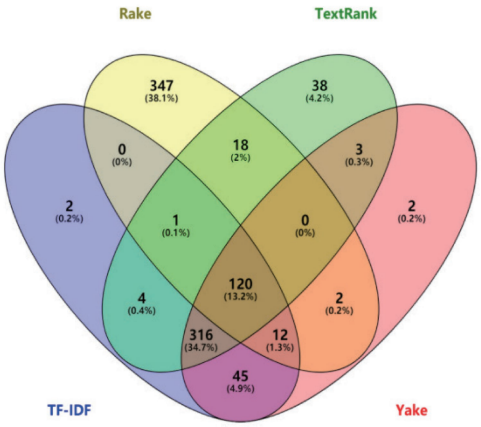

图 9　2015 年 1 月四种算法下关键词韦恩图

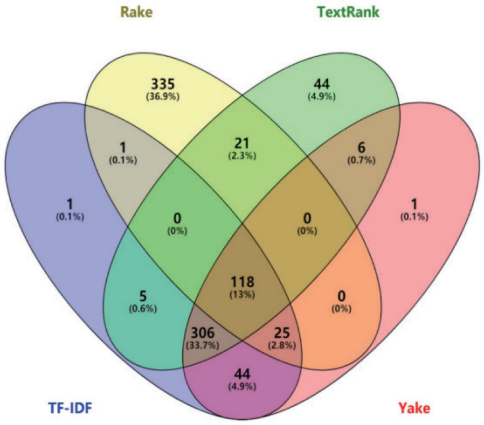

图 10　2015 年 6 月四种算法下关键词韦恩图

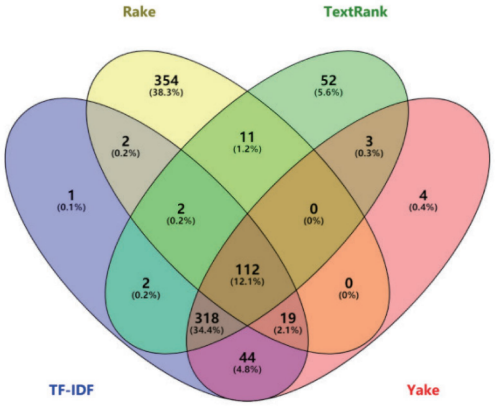

图 11　2016 年 1 月四种算法下关键词韦恩图

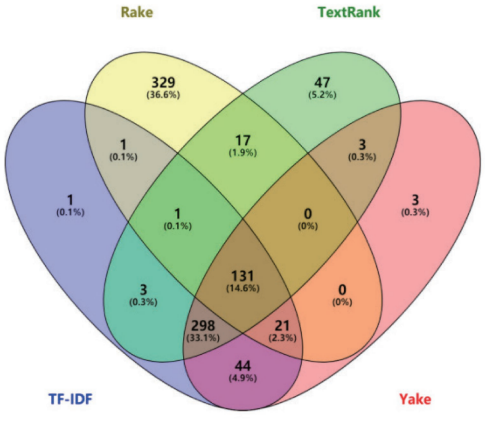

图 12　2018 年 1 月四种算法下关键词韦恩图

除去表现最差的两种主题模型，将其余四种算法下排名前 500 的所有关键词集合进行交叉统计，得到图 9 ~ 12 四张韦恩图。通过观察发现，除去四者的公共部分之后，Rake 算法跟其余三种算法的交叉集最小，TF-IDF 算法和 Yake 算法与其它算法的交叉度最高。从算法原理而言，TF-IDF 算法和 Yake 算法均倾向于输出高频词汇，TextRank 相较二者而言，由于涉及网络构建和随机游走的迭代算法，关键词提取性能不稳定。

4.2 不同类别下关键词抽取对比分析

在上述关键词抽取实验中，TF-IDF 算法和 Yake 算法的表现不相上下，本文进一步利用二者，分别针对 2015 年 1 月和 2018 年 1 月 NEPD 语料中的政治、经济和社会三个板块的语料再次进行了抽取实验。由于分类别的语料规模相对较小，仅挑选两个算法中排名前 50 个关键词，并在去除重复关键词后，经人工筛选得出综合排名前 50 个关键词，以此验证其抽取性能是否能帮助用户梳理出当月相关领域所发生的事件。

表 4　2015 年 1 月和 2018 年 1 月政治板块关键词

政治板块	关键词
两个月份共有词	工作　机关　干部　案件　行政　部门　群众　政府　法院　领导　书记　服务　纪委 建设　制度　活动　社会　法律　基层　发展　北京　改革　单位　责任　企业　组织 信息　监督　解决　管理　意见　社区
2015 年 1 月余下词	犯罪　中央　司法　会议　依法　检察院　律师　国家　检察　调研　县委　专项　法官 机构　职务　腐败　省委　青年
2018 年 1 月余下词	办理　平台　党员　业务　执法　涉嫌　武汉　调查　审计　落实　项目　教育　公司 党委　考核　负责人　网站　线索

表 5　2015 年 1 月和 2018 年 1 月经济板块关键词

经济板块	关键词
两个月份共有词	发展　企业　农村　产业　服务　建设　金融　推进　农业　工作　公司　经济　技术 市场　管理　改革　资金　创新　国家　银行　投资　监管　产品　信息　领域　项目 政府　业务　政策　行业　机构　制度　部门
2015 年 1 月余下词	土地　价格　出口　试点　贷款　铁路　审批　体制　北京　深化　美元　成本　经营 互联网　社会　存款　机制
2018 年 1 月余下词	乡村　人工智能　农民　生产　振兴　科技　质量　旅游　体系　消费　食盐　规模 基础　风险　网络　建立　综合

表 6　2015 年 1 月和 2018 年 1 月社会板块关键词

社会板块	关键词
两个月份共有词	服务　医院　工作　社会　市场　企业　管理　养老　政府　项目　北京　部门　制度 标准　保障　建设　发展　平台　国家　信息　城市　扶贫　政策　建立　就业　医疗 机构
2015 年 1 月余下词	改革　慈善　出租车　单位　网络　公司　资金　公益　农民工　事业　组织　工资 机关　缴费　春运　司机　器官　捐献　交通　专车　收入　老人　募捐
2018 年 1 月余下词	租赁　贫困　住房　脱贫　厕所　群众　社区　疫苗　解决　旅游　公寓　长租　集体 居民　用地　食品　产业　生活　小区　培训　贫困户　需求　HPV

从共有的关键词来看，根据表 4，在政治板块，“案件”“法院”“法律”等词反映了《人民日报》通过实际案例来传达教育、警示作用的叙事风格；“建设”“改革”“监督”“责任”“管理”等词反映出我国在政策制定和执行方面所做出的努力；而在 2015 年 1 月，“犯罪”“检察院”“律师”“法官”“腐败”等关键词，说明贪污、腐败案件的出现得到了国家的高度重视。“执法”“调查”“审计”“考核”等 2018 年 1 月的关键词显示出我国在预防违法犯罪上所采取的相应措施。

根据表 5，从共有词的高频关键词“发展”“企业”“农村”等词，可以看出《人民日报》持续关注企业、农村等的经济发展状况，而“投资”“创新”“信息”“领域”“项目”说明我国持续关注经济发展的新兴领域。同时，从 2015 年初的“土地”“铁路”“互联网”，到 2018 年初的“人工智能”“科技”“旅游”等词，均可以看出我国经济领域在不同年份的工作重心。

在社会板块，通过“服务”“医院”“养老”“保障”“扶贫”“就业”等词，可以看出《人民日报》对社会的关注主要围绕基本的民生问题，如医疗服务质量的提升、养老制度的完善、扶贫政策的制定以及就业问题等。2015 年初的“农民工”“春运”“交通”等词，关注点在于一年一度的春运问题，比如农民工返乡难等。而“慈善”“器官”“捐献”“募捐”等词则体现了我国人民群众的社会责任感与大义。在 2018 年 1 月的关键词中，“疫苗”“HPV”等词可反映出群众对自身健康的关注，“租赁”“贫困”“住房”“脱贫”“长租”“贫困户”等关键词说明住房和贫困两个问题依然严峻。

总之，通过对以上关键词的梳理与分析，可以大致判断当月的重要事件。除去在两月中重复的关键词，可以看到，在这两个不同时期，国家各方面的工作都各有偏重。

5 总 结

本文主要基于六种算法对 NEPD 语料进行了关键词提取，结果具有一定合理性，但各个算法提取性能不一，差距较大。总体而言，TF-IDF、Yake 和 TextRank 三种算法表现较好。其中，前二者性能表现十分接近。就本次关键词提取实验而言，所提取的关键词与高频词存在大量重合。主要是本研究所选取停用词存在一定缺陷，本研究除去人工精分词语料外，并未添加譬如词性、词长、词语位置等诸多特征入内，势必在运算过程中影响了关键词提取效果。在今后的工作中，重点考虑将各类特征与深度学习思想相融合，提升各类算法的关键词提取性能。同时，通过对各类别语料下所抽取的关键词进行分析，在一定程度上，能快速掌握新闻的大致内容。虽然提取关键词的算法有很多，但有时在阅读新闻时，关键词并不会一直出现，而且对于中文来说，蕴涵的中心思想也往往不是简单的一些关键词能说明的。因此，在未来关键词提取的研究中，可借助多方资源扩充语料内容，与主题模型相融合，提高关键词提取效果。此外，也应该关注怎样界定关键词的问题，以便更好的评估算法的性能。

［参考文献］

［1］黄水清，王东波 . 新时代人民日报分词语料库构建、性能及应用（一）——语料库构建及测评［J］. 图书情报工作 ,2019,63(22):5-12.

［2］牛永洁，田成龙 . 融合多因素的 TFIDF 关键词提取算法研究［J］. 计算机技术与发展 ,2019,29(7):80-83.

［3］杨凯艳 . 基于改进的 TFIDF 关键词自动提取算法研究［D］. 湘潭大学 ,2015.

［4］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.

［5］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.

［6］朱泽德，李淼，张健，曾伟辉，曾新华 . 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.

［7］宁珊，严馨，周枫，王红斌，张金鹏 . 融合 LSTM 和 LDA 差异的新闻文本关键词抽取方法［J］. 计算机工程与科学 ,2020,42(1):153-160.

［8］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.

［9］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.

［10］杨延娇，赵国涛，袁振强，韩家臣 . 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.

［11］徐明明，杨文璐，夏斌，谢宏 . 基于改进 RAKE 算法的商品关键词提取方法［J］. 现代计算机 (专业版),2018(21):7-11.

［12］陈可嘉，黄思翌 . 中文短文本自动关键词提取的改进 RAKE 算法［J］. 小型微型计算机系统 ,2021,42(6):1171-1175.

［13］郑成朗 .《人民日报社论全集》主题词研究［D］. 广西师范学院 ,2017.

［14］刘晓丽 .《人民日报》社论词汇统计与分析［D］. 广西师范学院 ,2015.

［15］董志成 .《人民日报》文本的历时计量研究［D］. 黑龙江大学 ,2020.

［16］李琪 . 时代精神：历时文本关键词提取与解读——基于《人民日报》文本的实践［J］. 数字人文 ,2020(3):125-150.

［17］黄水清，王东波 . 新时代人民日报分词语料库构建、性能及应用 (三)——句长与词的分析比较［J］. 图书情报工作 ,2019,63(24):5-15.

［18］黄水清，王东波 . 基于人民日报语料的中央一号文件词频历时分析［J］. 农业图书情报学报 ,2020,32(3):4-9.

［19］饶高琦，李宇明 . 基于 70 年报刊语料的现代汉语历时稳态词抽取与考察［J］. 中文信息学报 ,2016,30(6):49-58.

［20］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 ,2021,5(3):45-59.

［21］Zipf G K. Human behaviour and the principle of least-effort［M］. Cambridge: Addison-Wesley,1949.

［22］Campos R, Mangaravite V, Pasquali A, et al. YAKE! Keyword Extraction from Single Documents using Multiple Local Features［J］. Information Sciences, 2020,509: 257-289.

［23］Dalal R. RAKE［EB/OL］.［2021-06-14］. http://github.com/rahuldalal/RAKE#rake.

Research on Keyword Extraction and Analysis of New Era People's Daily Segmented Corpus

Zhou Hao[1,2] Wang Dongbo[1,2] Huang Shuiqing[1,2]

(1. School of Information Management, Nanjing Agricultural University, Nanjing 210095, China;
2.Research Center for Humanities and Social computing, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: [**Purpose/significance**] In the face of massive news text, the primary task of keyword extraction is to help users quickly master news content by extracting a small number of keywords that represent its content. [**Method/process**] This paper takes part of the corpus from the New Era People's Daily Segmented Corpus (NEPD) as the research object, compares the extraction effects of six unsupervised keyword extraction methods, TF-IDF, TextRank, LDA, LSI, Rake, and Yake, and analyzes the extraction results. [**Result/conclusion**] The results show that under the Pooling evaluation method, the TF-IDF algorithm and the Yake algorithm perform best in the large-scale People's Daily keyword extraction task, and the performance of the TextRank algorithm is acceptable. In addition, through the analysis of the overall high-frequency words and keywords under the political, economic and social categories, we can quickly find and sort out the important events of the month. This article can provide a reference for the keyword extraction and analysis of news newspapers and periodicals.

Keywords: Keyword extraction; NEPD; Unsupervised extraction method

（本文责编：孔青青）