裁判文书关键词提取的改进方法研究

白凤波[1],常 林[2],王世凡[2],李 彬[3],王颖洁[4],周 红[5],刘 耀[5]
1.中国政法大学 证据科学研究院,北京 100088
2.浙江迪安鉴定科学研究院,杭州 310000
3.中国科学技术大学 软件学院,江苏 苏州 215000
4.大连大学 信息工程学院,辽宁 大连 116622
5.公安部物证鉴定中心,北京 100038

摘 要:在国家加强依法治国的方针指引下,自然语言处理(NLP)和信息检索(IR)等领域与法治社会的深入结合是必然趋势。为司法工作者提供正确、全面的智能化辅助以提高工作效率,对裁判文书的关键词提取方法进行了研究。针对传统关键词提取方法的劣势,结合词语的词性、长度、词跨度、位置以及文档所属类别等多重因素,并基于图模型的TextRank算法,提出了一种改进的TF-IDF算法(IAKEF),引入信息熵、离散度、融合特征的概念,主要解决了传统算法对于词语在语义上的忽略和类间、类内信息分布上的问题,使其能够更有效地从文本中选择特征。通过对比实验,对改进算法的效果进行分析和评价,实验结果表明改进的算法与传统的算法相比在准确率、召回率及*F*1-Measure上均有显著的提高。

关键词:改进TF-IDF;关键词抽取;信息熵;离散度;特征融合
文献标志码:A **中图分类号:**TP391.1 doi:10.3778/j.issn.1002-8331.2004-0097

白凤波,常林,王世凡,等.裁判文书关键词提取的改进方法研究.计算机工程与应用,2020,56(23):153-160.
BAI Fengbo, CHANG Lin, WANG Shifan, et al. Improved method study on extracting keywords in Chinese judgment documents. Computer Engineering and Applications, 2020, 56(23):153-160.

Improved Method Study on Extracting Keywords in Chinese Judgment Documents

BAI Fengbo[1], CHANG Lin[2], WANG Shifan[2], LI Bin[3], WANG Yingjie[4], ZHOU Hong[5], LIU Yao[5]
1.Institute of Evidence Law and Forensic Science, China University of Political Science and Law, Beijing 100088, China
2.Di'an Institute of Forensic Sciences in Zhejiang, Hangzhou 310000, China
3.School of Software Engineering, University of Science and Technology of China, Suzhou, Jiangsu 215000, China
4.College of Information Engineering, Dalian University, Dalian, Liaoning 116622, China
5.Institute of Forensic Sciences, Ministry of Public Security, Beijing 100038, China

Abstract: Under the national policy the guidance to rule the country by law, it is an inevitable trend to combine the field of artificial intelligence, such as NLP(Natural Language Processing) and IR(Information Retrieve), with the need to rule of law. In this paper, through the research of keyword extraction method for judicial documents, the purpose is to provide accurate and comprehensive intelligent assistance for judicial service workers to improve work efficiency. This paper proposes an improved TF-IDF algorithm, named Improved Algorithm for Keyword Extraction in Forensics(IAKEF),

基金项目:中国工程院中长期咨询研究项目(No.2017-ZCQ-10)。
作者简介:白凤波(1978—),通信作者,男,博士研究生,高级软件工程师,CCF会员,主要研究方向为人工智能、数据科学和证据科学,E-mail:windbai@hongstech.com;常林(1963—),教授,博士生导师,主要研究方向为法医学、法庭科学和证据法学;王世凡(1964—),高级法官,主要研究方向为证据法学、法医学和法庭科学文化与历史;李彬(1996—),硕士研究生,主要研究方向为人工智能;王颖洁(1977—),副教授,CCF会员,主要研究方向包括软件工程、可信软件;周红(1969—),研究员,主要研究方向为理化检验;刘耀(1937—),教授,博士生导师,中国工程院院士,主要研究方向为法医毒物分析。
收稿日期:2020-04-08 **修回日期:**2020-07-23 **文章编号:**1002-8331(2020)23-0153-08
CNKI **网络出版:**2020-07-27, https://kns.cnki.net/kcms/detail/11.2127.TP.20200727.1243.012.html

targeting to the disadvantages of traditional keyword extraction methods, according to the multiple factors such as part of speech, length, word span, position and document category, based on the TextRank algorithm of graph model, introducing the concepts of information entropy, dispersion degree and fusion features. The algorithm mainly solves the problems of traditional algorithms for semantic neglect of words and distribution of information among classes or a class inner, so that the features from text can be selected more effectively. With the experiments and the comparison of algorithms, the improvement effect is analyzed and verified, the experimental results show that the improved algorithm has a significant improvement in accuracy, recalling-rate and F1-Measure compared with the traditional algorithm.

Key words: improved TF-IDF; keyword extraction; information entropy; dispersion; feature fusion

1 引言

关键词是反映文章中心或主旨思想的一组词或短语[1],关键词提取作为文本聚类或自然语言处理的关键步骤之一,是指采用自动化的技术在文本中提取关键词的过程。关键词提取可以分为有监督提取、半监督提取和无监督提取三类[2],目前普遍采用的是适应性较强的无监督的关键词抽取算法。近年来,研究者们对关键词提取算法开展了大量的研究工作[3]。其中,TF-IDF[4]算法作为一种基于统计学的提取方法,主要通过词频计算提取文章中的关键词,但由于对词频的过度依赖往往会降低提取的效果。有些研究者在传统的TF-IDF算法上加入语义、词频、词长、位置等多方面的信息对其进行改进[5-7]。针对TF-IDF算法没有考虑特征词在文本集上的分布特征,文献[8-10]将信息熵、互信息、信息增益等与TF-IDF进行结合。为弥补数据集偏斜带来的问题,文献[11]提出用Category Term Descriptor(CTD)来改进TF-IDF。还有研究者用特征选择函数来代替IDF,将传统的TF-IDF算法改进成TF乘以特征选择函数的形式[12]。基于传统算法的复杂中文的专业短语抽取的准确程度问题,本文提出一种改进的加权算法(IAKEF)。在我国加强依法治国的方针指引下,人工智能研究特别是自然语言处理与法治社会的结合必将进一步加深。然而,通过对作为最终法庭审判结论的裁判文书进行关键词提取,为审判人员、司法鉴定、律师等司法服务工作者提供正确、全面的参考案例,有效提高工作效率显得尤为关键。

2 模型描述

判决文书虽然在词频与逆文档频率方面与其他文本处理差异很小,其关键词抽取与日常用语中的关键词有所不同。其一,停用词有所不同。例如,像“本院认为”“裁定”“一审”等在文书中频繁出现但又无法表达文书主旨内容的词语,将会成为关键词提取的干扰项。其二,关键词的词长较长。裁判文书中词长为四个字以上的词汇超过四成比例。如图1所示,见Word2Vec抽取的关键词列举。其三,文本内容分类较细。由于裁判文书的类型有七大类,各类型的文书主题的侧重不同,关键词倾向也不同。

key
到庭 行政处罚 初犯 从犯 决定书 简易程序 具结 抢劫罪 共同犯罪 上诉状
到庭 数额较大 转账 复检 退赔 上诉状 本院认为 输光 没人管 借条
举证 行政处罚 诉讼费用 正当理由 证据确凿 情节严重 司法解释 到庭 决定书 调查取证
到庭 正当理由 撤诉 传票 诉讼法 传唤 裁 借贷 审 民事
行政处罚 决定书 复议 权属 强制执行 行政诉讼法 履行义务 催告 本院认为 处罚权
行政处罚 决定书 实施细则 行政诉讼法 裁量权 履行义务 强制执行 催告 本院认为 期限内
撤诉 本院认为 财产保险 书记员 支公司 审判员 诉讼法 法定代表 浙商 减半
诉讼费用 撤诉 诉讼费 免交 书记员 审判员 民事诉讼 诉讼法 事务所律师 分子筛
决定书 行政处罚 行政诉讼法 本院认为 强制执行 事实清楚 国土资源 准予 法定代表 学兵
到庭 批准逮捕 勘验 事故责任 取保候审 上诉状 回执 供述 本院认为 认罪态度

图1 裁判文书的Word2Vec关键词抽取举例

2.1 传统的TF-IDF算法

词频(Term Frequency,TF)是指词语在整个文本中出现的频率,计算方法是用该词语在文本中总共出现的次数除以文本中的单词总数;逆文档频率(Inverse Document Frequency,IDF)是用来衡量词语是否具有文章代表性的评价方法,计算方法为用语料库中文档的总数除以出现词语的文档数目并取对数。倘若一个词在某一类中出现的频率较高,而在其他类中出现的频率较低,则说明该词语在此类中具有很强的代表性,其IDF值就高;相反,若一个词在每个文档中均有出现,则其文章代表性较低,IDF值也就低。IDF因子在一定程度上避免了文章中常用非关键词语成为关键词的可能。

TF-IDF的主要思想是,用TF相乘IDF的结果作为计算词语的权重,词语在文章中出现的频率越高,而包含该词语的文档数越少,则认为该词语的重要性越高。

计算公式如式(1)~式(3)所示:

$$词频(TF)=\frac{词w在文档中出现的次数}{文档的总词数} \tag{1}$$

$$逆文档频率(IDF)=\log\frac{语料库的文档总数}{包含词w的文档数+1} \tag{2}$$

$$TF\text{-}IDF=TF\times IDF \tag{3}$$

TF-IDF是一种基于统计特征的传统的关键词提取算法,算法的性能较好,运算速度快,提取的结果比较符合实际。但是该算法仅仅考虑词频方面的因素,没有考虑词语出现的位置、词性、词长等信息,具有一定的局限性。

2.2 改进的关键词提取算法

针对TF-IDF算法的缺陷,目前已有一些学者提出

了改进的算法,例如TF-IDF′[7]算法、CTD[11]算法、TF-IDF-IGD[10]算法以及融合多特征的TF-IDF-MTF[13]算法,这些算法都在一定程度上提高了权重计算的准确率,但对于裁判文书专业性较强的数据文本来说提取效果不是太好。本文根据裁判文书的特点,结合已改进的TF-IDF算法,提出了一套新的改进方法,对裁判文书的关键词提取收到了良好的效果。

2.2.1 改进的文本预处理

通过Jieba[14]分词和Ltp[15]分词技术对文书的文本进行分词,两种分词工具的结合可以进行互补,克服各自的缺点,从而提高了分词的准确率。同时考虑到本文所用到的数据为判决文书,具有一定的法律专业性。所以,又以"搜狗"这一在线工具为例对其细胞词库中关于文书、法律和专有名词方面的词库进行解析,作为自定义字典,加入到Jieba和Ltp分词工具中,进一步提高了分词的专业性。停用词字典在第一轮分词的基础上进行更新,将出现频率较高的无效词逐一加入停用词词典,以提高提取关键词的有效率。文本中的同义词在一定程度上也会影响关键词提取的准确率,例如"审判"和"审讯"作为近义词,若在提取时单独对待,可能因为两者的权重均较低而被忽略,若作为同一个词进行处理,则能避免这种情况的出现。在预处理阶段进行同义词处理,根据同义词库合并文本相似度高的词语成为了预处理阶段重要的一步。改进的文本预处理阶段流程,见图2。

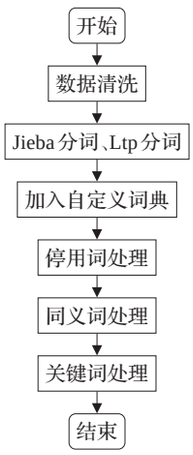

图2 预处理阶段流程

2.2.2 基于语义的改进

传统的TF-IDF算法仅仅统计词频信息,倾向于频率较高的词语,没有考虑词语的语义、位置、词长、词性等方面的信息,而这些因素都会影响词语在文章中的表示意义,使得提取了一些频率较高但与主题无关的关键词。裁判文书作为一种特殊的法律文本,文书的标题在一定程度上可以很好地概述全文的内容,这也说明了不同位置的词语代表文本内容的重要性程度不同。

针对以上提起的几方面因素,杨凯艳在文献[16]中已有探索,本文在其研究的基础上加以改进,使其在语义上更能符合裁判文书的特点。本文提出将多个影响因素进行特征融合,把得到的融合特征MTF(Multi-Term-Feature)作为乘数因子加入权重计算公式中去,以改善传统算法在语义方面的不足。融合特征MTF的计算公式如式(4)所示:

$$MTF = WL \times (Pos + TL + WS) \tag{4}$$

(1)词性因子POS(Part of Speech)

在一个句子中,不同词性的词语对句子修饰性不同,所蕴含的信息量不同,也会导致句子的语义不同。本文通过对人工标注的1 000条判决文书的5 000个关键词词性进行分析,结果如图3所示。

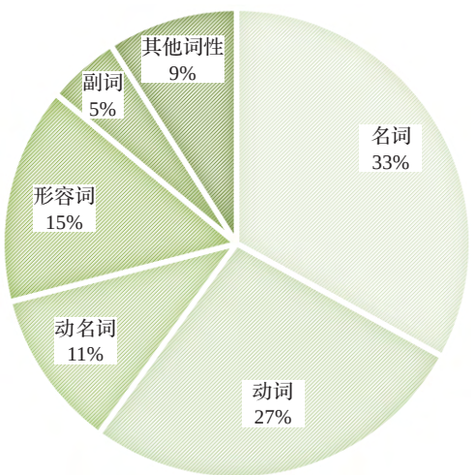

图3 关键词词性分布

通过图2可知关键词中大约70%的词语词性为名词、动词、动名词,词性为形容词、副词的所占比例约为20%,其他词性所占比例为10%。所以根据词性的分布不同,应给予词语不同的权重。本文的词性因子计算公式如式(5)所示:

$$Pos(T)=\begin{cases}0.7, & \text{若}T\text{为名词、动词、动名词}\\0.2, & \text{若}T\text{为形容词、副词}\\0.1, & \text{若}T\text{为其他词性}\end{cases} \tag{5}$$

(2)词长因子TL(Term Length)

针对词长因子,目前常用的两种评价方法,分别为词长比例权重法和设置常数法。前者是将词语的长度与全文中最长词语长度的比值作为词长因子,后者则是通过对特定长度的词语进行人工设置系数来为词长因子赋值。考虑到裁判文书中词长作为一项重要因子,对文档主题的贡献度较大,本文评价方法采用后者。通过对人工标注的5 000个关键词词长进行分析,其分布如图4所示。

根据实验数据的百分比,得出的词长权重系数计算公式如式(6)所示:

图4 关键词词长分布

$$Wl(T)=\begin{cases}4, & 4\leqslant l_i\leqslant 6\\ 2, & \text{其他}\end{cases} \tag{6}$$

其中 l_i 为词语长度。

(3)词位置因子WL(Word Location)

判决文书中标题与内容的基本符合率为95%,在文书的首段和结尾位置也往往富含大量有效信息,这些特定的位置关键词出现的概率比较大。通过对词语的位置信息进行有效标识可以提高关键词的提取效果。本文通过将词语第一次和最后一次出现位置作为特征对提取算法进行改进。

① 预先的标题处理

根据标题与文章内容的长度,通过增加标题在总文本中出现的次数加以改进。计算公式如式(7)所示:

$$Count(title)=ceil(text\_sentences\_len\times 0.4) \tag{7}$$

其中,$Count(title)$ 为标题重复次数,$ceil$ 代表向下取整。

② 词位置因子的处理

能够总结全文的词语往往出现在文章的末尾,对整篇文章进行概括;而能够开门见山、指明文意的词语出现在开头的概率更大,起到统领全文的作用;因此词位置因子的计算公式如式(8)~(10)所示:

$$WL(T)=\frac{1}{FP(T,d)+LP(T,d)} \tag{8}$$

$$FP(T,d)=\frac{FirstPos(T)}{SumWords(d)-FirstPos(T)} \tag{9}$$

$$LP(T,d)=\frac{LastPos(T)}{SumWords(d)-LastPos(T)} \tag{10}$$

其中,$FP(T,d)$ 代表词语的首位置,$LP(T,d)$ 代表词语的末位置,$FirstPos(T)$ 是词语 T 在文本 d 中首次出现时已出现的词语数,$LastPos(T)$ 代表词语 T 最后一次出现时文章还未出现的词语数,$SumWords(d)$ 是文本 d 的词语总数。

(4)词跨度因子WS(Word Span)

词跨度代表词在文本中首次出现的位置与最后一次出现的位置之间的距离,反映了词在文中的出现范围。通常来讲,词在文章中出现的范围越广,即词跨度越大,说明该词越能反映文章的主题[17];相反,词跨度越小,说明词语集中在局部出现,不能概括全文的主旨。倘若某词在局部出现的频率很大,则会一定程度上影响全文关键词的提取,所以引入词跨度因子可以较好地避免这一问题。

词跨度因子主要是过滤某个局部范围内频率较高的词语,计算方法如式(11)所示:

$$WS(T,d)=\frac{las(T)-fir(T)+1}{SumWords(d)} \tag{11}$$

其中,$las(T)$ 为词 T 在文本 d 中最后一次出现的序号,$fir(T)$ 为词 T 在文本 d 中首次出现的序号,$SumWords(d)$ 为文本 d 分词后的总词数。

2.2.3 基于分类内分布的改进

由于裁判文书的类型有刑事判决、民事裁定、行政判决、行政赔偿、执行裁定、民事判决、其他类七大类,文书类型的不同会导致文书主题的侧重不同,即各种类型文书之间的关键词倾向不同。本文区分了裁判文书的分类,假设文本集合 C 中共有 N 种分类,$C=\{C_1,C_2,\cdots,C_n\}$,类 C_i 的文本数为 M_i,$\overline{tf_i(t_k)}$ 为词语 t_k 在 C_i 这一分类出现的频率[18]。计算公式如式(12)和式(13)所示:

$$\overline{tf_i(t_k)}=\frac{1}{M_i}\sum_{j=1}^{M_i}tf_{ij}(t_k) \tag{12}$$

此时分类内离散度 D_{ic} 如式(13)所示:

$$D_{ic}=\frac{\sqrt{D_{ii}}}{\frac{M}{\sqrt{M-1}}\overline{tf_i(t_k)}} \tag{13}$$

其中 D_{ii} 为词语 t_k 在类 C_i 中无偏估计的样本的方差,如式(14)所示:

$$D_{ii}=\frac{1}{M-1}\sum_{j=1}^{M}(tf_{ij}(t_k)-\overline{tf_i(t_k)})^2 \tag{14}$$

对于两个词语 $T1$、$T2$,假设计算时得出的 IDF 值相等,说明包含两词的文档数是一样的。而在同一类 C_i 中,假设词语 $T1$ 普遍出现在类 C_i 的各项文档中,而词语 $T2$ 仅出现在类 C_i 的某几篇文档中,此时说明词语 $T1$ 对类 C_i 更显著,其成为类 C_i 文档中关键词的可能性较大;计算它们的离散度,得到 $T1$ 的类内离散度值比 $T2$ 的类内离散度值要小,说明类内离散度越低,词语对应的权重就会越高[5]。

词语 T 在各个类中分布不均匀,其在各个类中代表文档主题的影响力就不同。考虑裁判文书有民事、刑事等七大类,且各类别的裁判文书特点鲜明,本文引入类内离散度,将词语最能代表那个类别的类内离散度作为该词语的调整因子,解决类内分布差异问题。离散度 D 计算方法如式(15)所示:

$$D = 1 - \min\{D_{ii}\} \tag{15}$$

2.2.4 基于分类间分布的改进

本文中引入信息增益来解决文书分类间的词语分布的问题。信息增益是一种基于信息论的特征选择方法[19-20]。信息熵是由美国数学家克劳德·艾尔伍德·香农(Claude Elwood Shannon)提出的对信息的一种度量单位,表示所蕴含信息量的多少;在信息论中,信息熵用于描述信息空间的突发性和不确定性。熵的值越小,表示信息空间概率分布越均匀;条件熵的定义是在给定 X 的条件下,Y 的条件概率分布的熵对 X 的数学期望,它描述的是观测某个变量之后信息空间的不确定性程度;信息量被用来度量不确定性的减少程度,因此信息增益代表了所观测的变量携带的信息量。其量化思想为:当词语在各个类中分布越均匀,说明它对类别的区分能力越弱,即所含的信息量越少,应给予较低的权重,反之亦然。通过把信息增益公式引入到文本集合的类别间,依靠数据集合中类别信息熵和文本类别中词语条件熵之间信息量的增益关系,来确定该词语在文本分类中所能提供的信息量,并把这个信息量反映到词语的权重中[20]。

信息增益计算公式如式(16)所示:

$$IG(C,T) = E(C) - E(C|T) \tag{16}$$

假设文档结合共有 n 种类别,$E(C)$ 为文档集合类别 C 的信息熵,$E(C|T)$ 为词语 T 对文本集类别的条件熵;$P(c_i)$ 表示类别 c_i 的概率,$P(t)$ 表示词语 T 在文档集合中出现的概率,$P(\overline{t})$ 表示词语 T 不出现的概率,$P(\overline{t}) = 1 - P(t)$;每个样本子集的熵,可以转化为子集与文本集合类别 c_i 的条件熵,$E(C|t)$ 表示词语 T 出现时类别集合的条件熵,$E(C|\overline{t})$ 表示词条 T 不出现时类别集合的条件熵;$P(c_i|t)$ 表示 c_i 类中含有词语 T 的文档数,$P(c_i|\overline{t})$ 表示 c_i 类中不含词语 T 的文档数。信息熵、条件熵的计算公式如式(17)~式(20)所示:

$$E(C) = -\sum_{i=1}^{n} P(c_i) \times \mathrm{lb} P(c_i) \tag{17}$$

$$E(C|T) = P(t)E(C|t) + P(\overline{t})E(C|\overline{t}) \tag{18}$$

$$E(C|t) = -\sum_{i=1}^{n} P(c_i|t) \times \mathrm{lb} P(c_i|t) \tag{19}$$

$$E(C|\overline{t}) = -\sum_{i=1}^{n} P(c_i|\overline{t}) \times \mathrm{lb} P(c_i|\overline{t}) \tag{20}$$

2.2.5 基于TextRank的改进

TextRank算法是利用局部词汇之间关系(共现窗口)对后续关键词进行排序,用到了词之间的关联性,这是其优于TF-IDF的地方,可以弥补传统的TD-IDF算法仅考虑词频的问题,因此本文提出的改进的计算公式如式(21)所示:

$$\text{权重} = (\partial \times TFIDF \times IG \times D + (1-\partial) \times TextRank) \times MTF \tag{21}$$

本文中 ∂ 为加权因子。

3 模型实现

基于改进的关键词抽取算法,其流程见图5。

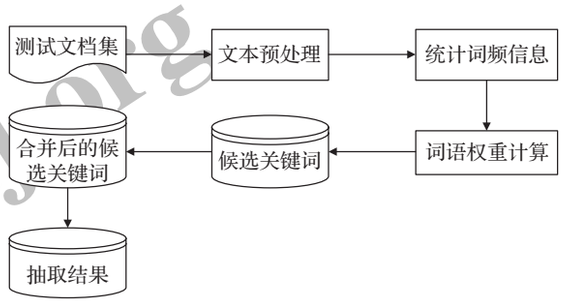

图5 改进的关键词抽取算法流程图

步骤1 文本预处理,数据清洗、格式标记的去除、中文分词技术、词性标注、以及停用词过滤。

步骤2 统计词语信息,主要包括词语的词频、词性、长度、出现的文档数和位置。

步骤3 结合信息熵、离散度、融合特征以及TextRank,根据改进的公式计算候选词的权重。

步骤4 将候选词权重由大到小的顺序排序,取前一个词语作为关键词。

4 实验验证

4.1 实验环境

4.1.1 软件环境

基于Windows 10操作系统,采用Python 3.5编译环境对上述实验进行验证实现。主要采用Python语言中第三方工具对系统数据结构进行设计,以实现大数据存储和各种操作。其中,第三方工具包含Pyltp0.2.1、Ltp3.4.0、Pandas0.24.2、Jieba0.39、Gensim 3.7.2等。

4.1.2 硬件环境

计算机型号:神舟战神Z7-i78172S2。

处理器:Intel® Core™ i7-4720HQ CPU @ 2.60 GHz。

安装内存:32.00 GB。

4.2 实验数据集和评价方法

本文实验数据集是采用由网络获取的裁判文书文本数据。数据集共包含1 000条训练集,已经人工标注关键词;包含240 000条记录作为测试集进行关键词提取。语料集中主要包括刑事判决、民事裁定、行政判决、行政赔偿、执行裁定、民事判决、其他类七大类。本文选取已有人工标注的1 000条文件数据,每个文书选取5个词语作为关键词(其中包含不少于5 000个的关键词)用于实验仿真。因为人力和时间有限,尽量较好地保证

质量,由专业司法工作专家依照案件的缘由和类别完成少量标注数据。

实验所用测试集是用来进行关键词提取的验证与评价。因为测试集包含文本数量较大没有全部人工标注和评价,实验抽样选取其中大约1 000条文本进行人工标注对比评价和分析,验证这部分数据的准确率。实验设计验证关键词提取准确程度,主要是基于人工标注的1 000条文书进行的,每篇文书人工标注5个关键词,算法提取10个关键词,用*F*1-Measure作为评价标准。

4.3　实验步骤与结果分析

为衡量关键词的有效性,本文将算法提取的关键词与人工标注作比较,来判断提取算法效果的优劣。

4.3.1　实验数据处理

本文选取人工标注的1 000条数据作为实验数据集,其中刑事判决类200条、民事裁定类200条、行政判决类200条、行政赔偿类50条、执行裁定类200条、民事判决类50条、其他类100条。每条数据已经有人工标注出5个关键词作为实验结果比较。

4.3.2　实验步骤

第一步进行文本的预处理:通过对数据进行标题内容合并、分词、停用词和同义词过滤处理,构建候选词集合。

第二步关键词提取:分别采用传统TF-IDF算法、TextRank[21]算法、Word2Vec[22]及改进的算法进行关键词提取,每篇文章标准10个关键词用于与人工标准的关键词对比。

4.3.3　实验评价指标

实验选用准确率 P 、召回率 R 、*F*1-Measure($F1$)作为评价指标;准确率 P 是人工标注的关键词与计算机提取关键词的交集和计算机提取关键词的比率,是用于评价查找准确程度的指标;召回率是用人工标注的关键词与计算机提取关键词的交集和人工标注的关键词的比率,是用于评价查找完全程度的指标; $F1$ 因子是两者的综合指标,当 $F1$ 因子较高时,则能说明实验方法比较有效。设算法提取关键词集合为 T ,人工标注关键词集合为 H 。 P、R、$F1$ 的计算公式如式(22)~(24)所示:

$$P=\frac{|T\cap H|}{|T|} \tag{22}$$

$$R=\frac{|T\cap H|}{H} \tag{23}$$

$$F1=\frac{2P\times R}{P+R} \tag{24}$$

4.3.4　实验结果分析

按照以上步骤进行关键词提取,将不同算法的各项指标存入表1中。为了分析单一改进策略对算法评价结果的影响,将逐一排除改进策略的生成的评价指标记入表2中。不同算法准确率 P 、召回率 R 、*F*1-Measure($F1$)评价指标比较如图6,加权因子与 P、R、$F1$ 值对应图如图7。

表1　不同算法在判决文书中评价指标

算法	*P*	*R*	*F*1
TF-IDF	0.268 7	0.537 4	0.358 27
TextRank	0.222 1	0.444 2	0.296 13
Word2Vec	0.006 9	0.013 8	0.009 20
IAKEF (∂=0)	0.312 3	0.624 6	0.416 40
IAKEF (∂=0.1)	0.320 0	0.640 0	0.426 67
IAKEF (∂=0.2)	0.324 6	0.649 2	0.432 80
IAKEF (∂=0.3)	0.329 1	0.658 2	0.438 79
IAKEF (∂=0.4)	0.332 4	0.664 8	0.443 20
IAKEF (∂=0.5)	**0.334 7**	**0.669 4**	**0.446 27**
IAKEF (∂=0.6)	0.332 5	0.664 3	0.443 19
IAKEF (∂=0.7)	0.323 6	0.647 2	0.431 46
IAKEF (∂=0.8)	0.304 4	0.608 8	0.405 86
IAKEF (∂=0.9)	0.265 0	0.530 0	0.355 33
IAKEF (∂=1.0)	0.223 1	0.446 2	0.297 46

注:“IAKEF”表示本文所讨论的法庭科学关键词抽取改进算法。

表2　各改进策略单独排除的评价指标

算法	*P*	*R*	*F*1
No Semantic	0.313 3	0.626 6	0.417 73
No C-Inner	0.333 2	0.666 4	0.444 26
NoC-Inter	0.328 4	0.649 6	0.433 07
IAKEF (∂=0.5)	**0.334 7**	**0.669 4**	**0.446 27**

注:“No C-Inner”和“No C-Inter”分别表示分类内和分类间。

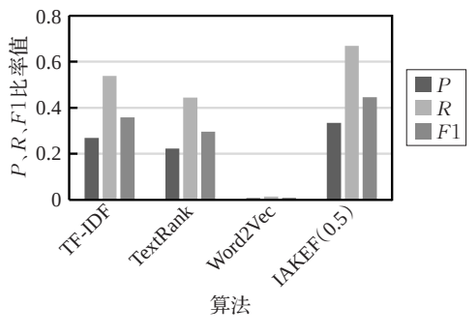

图6　不同算法评价指标比较

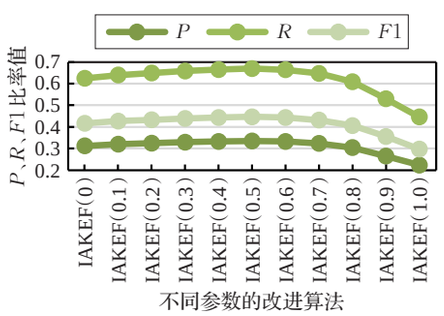

图7　加权因子 ∂ 与 P、R、$F1$ 值对应图

通过图6可知,不同算法对各类判决文书的 $F1$ 评测进行比较,TextRank算法相比其他算法来说,在各类文书中关键词提取的效果稳定性较高,但其准确率要低于传统的TF-IDF算法和改进的算法。Word2Vec提取效果最差,不适用于裁判文书关键词的提取。

通过图7可知,改进的算法与传统的TF-IDF、TextRank、Word2Vec相比,在准确率 P 、召回率 R 、$F1$-Measure上都有了很大的提高,$F1$ 值分别提高了0.088 00、0.150 14、0.437 07;实验结果表明该方法可以很好地提高对裁判文书进行关键词提取的准确率;并进一步对加权因子 ∂ 研究比较,当 $\partial=0.5$ 时,改进的TF-IDF算法关键词准确率达到最高。

通过图8~11可见(其中title字段较长,故只截取了前三个字),虽然改进的算法比传统的TF-IDF、TextRank、Word2Vec评估值较好,与真正的司法常用关键词仍有较大差距。例如,文中"国土资源局""土地资源""劳动合同""危险驾驶罪""民事诉讼法",这些词汇并没有符合专家标注意图。因此,司法相关命名实体库的缺失可能是导致这一结果的原因。

id	title	key
199d9a	石家庄	国土资源 行政处罚 决定书 土地 土地管理法 占用 作出 审查 处罚 永华
19bb3e	赵洪伟	驾驶 现经 依法逮捕 刑诉 找到 刑事诉讼法 移送 起诉 交付 涉嫌
19db17	保定市	莲池 保定市 国土资源 分局 资莲 占用 行政处罚 罚字 国土 土地
19f83c	华强方	动漫 有限公司 超市 分公司 综合 著作权 股份 站区 准许 纠纷
1acd2c	孟克道	道尔 扰乱 反恐 行政处罚 汽车站 行罚 违法 牧民 决字 海牛
1b59cf	杨建明	人力资源 保障局 权洲 职责 社会 法定 新光 晨光 征缴 物流
1b5a80	李亚停	农村土地 号码 承包合同 身份 公民 个体 现住 停于 撤回 纠纷
1bde9f	浦小庆	驾驶 缓刑 宣告 供述 附加刑 拘役 犯罪 如实 判处 公诉
1be7c9	张道明	协助执行 信用风险 名下 登记 惩戒 申报表 下落 单位 劳务 车辆
1bf92a	杨永君	行政 强制 镇长 传票 正当理由 传唤 违法 撤诉 减半 行政诉讼法
1c05c4	陈铖与	文情 大地 支公司 财产保险 准许 撤诉 交通事故 官桥 民政 机动车
1c4ff8	武汉市	构件 联合 土地 规划局 占用 土资 规罚 建筑物 国土资源 原状

TF-IDF | TextRank | Word2Vec | IAKEF-0.5

图8 TF-IDF算法的裁判文书关键字抽取样本

id	title	key
199d9a	石家庄	决定书 行政处罚 处罚 土地 实施 审查 国土资源 占用 强制执行 作出
19bb3e	赵洪伟	驾驶 传唤 起诉 案件 生于 职业 汉族 小学文化 羁押于 移送
19db17	保定市	莲池 分局 强制执行 保定市 国土资源 占用 行政处罚 补正 资莲 国土
19f83c	华强方	有限公司 分公司 内容 法律 股份 动漫 著作权 综合 准许 超市
1acd2c	孟克道	行政处罚 证明 违法 证据 事实 行政 处罚 道尔 合法 作出
1b59cf	杨建明	保障局 撤诉 社会 履行 准许 法定代表 受理费 起诉 撤回 法定
1b5a80	李亚停	现住 起诉 受理费 撤回 准予 身份 公民 减半 纠纷 诉讼
1bde9f	浦小庆	缓刑 驾驶 供述 如实 公诉 上诉 宣告 判处 机动车 拘役
1be7c9	张道明	登记 查询 财产 惩戒 名下 提示 单位 申报表 信用风险 信息
1bf92a	杨永君	正当理由 行政 受理费 案件 减半 镇长 法定代表 传票 传唤 强制
1c05c4	陈铖与	法律 起诉 准许 中心 撤诉 支公司 受理费 有限公司 机动车 股份
1c4ff8	武汉市	土地 有限公司 占用 构件 联合 罚款 建筑物 违反 局长 行政诉讼

TF-IDF | TextRank | Word2Vec | IAKEF-0.5

图9 TextRank算法的裁判文书关键字抽取样本

id	title	key
199d9a	石家庄	我院 法律效力 结案 扛 立案 民事 市中区 执 履行 义务
19bb3e	赵洪伟	支公司 退休工人 简易程序 财产保险 事务所律师 诉讼法 毕力 裁 代理人 机动车
19db17	保定市	举证 累犯 情节严重 追究其 违禁品 盗窃罪 犯罪分子 到庭 决定书 上缴国库
19f83c	华强方	银行存款 责任保险 复议 保单 本院认为 财产保险 保函 书记员 审判员 诉讼法
1acd2c	孟克道	撤诉 本院认为 书记员 先路 市政工程 审判员 事务所律师 民事诉讼 诉讼法 法定代表
1b59cf	杨建明	诉讼费用 书记员 范金 审判员 诉讼法 现住 法定代表 交纳 送达 副行长
1b5a80	李亚停	法律依据 履行义务 经查明 书记员 审判员 诉讼法 邮政储蓄 自本 法律效力 邮寄
1bde9f	浦小庆	撤诉 本院认为 事务所律师 诉讼法 网络科技 劳务 裁 代理人 审 民事
1be7c9	张道明	撤诉 本院认为 诉讼法 减半 水泥厂 分厂 裁 悦 艳 审
1bf92a	杨永君	到庭 批准逮捕 故意伤害 上诉状 简易程序 作案工具 本院认为 询问笔录 可予 对驾
1c05c4	陈铖与	强制执行 诉讼法 人向 法律效力 恢 送达 裁 给付 审 当事人
1c4ff8	武汉市	清偿债务 债务人 限令 书记员 审判员 现住 联社 诉讼法 法定代表 督促

TF-IDF | TextRank | Word2Vec | IAKEF-0.5

图10 Word2Vec算法的裁判文书关键字抽取样本

id	title	key
199d9a	石家庄	国土资源 领土 内政 行政 非诉 资源局 土地管理法 作出 处罚 强制执行
19bb3e	赵洪伟	决定书 驾驶 危险 自诉 支配 完全小学 取保候审 拘留 检察院 强制措施
19db17	保定市	国土资源 分局 强制执行 领土 保定市 莲池 内政 资源局 处罚 行政
19f83c	华强方	股份公司 有限公司 超市 著作权 综合 动漫 分公司 股份 公民权 杂货店
1acd2c	孟克道	治安管理 内政 行政处罚 公安局 处罚 道尔 治校 罚款 扰乱 行政
1b59cf	杨建明	人力资源 保障局 人力 泉源 社会 商店 行政 职责 内政 律师
1b5a80	李亚停	承包合同 农村土地 承揽 乡村 纠纷 合同 领土 身份 号码 公民
1bde9f	浦小庆	危险 驾驶 缓刑 不法之徒 违法 供述 主刑 处罚 驾驶罪 同意
1be7c9	张道明	协助执行 信用风险 归于 近似 登记 劳务 合同 劳动 名下 车辆
1bf92a	杨永君	行政 内政 强制 挨制 行政诉讼法 审判员 传唤 省长 呼唤 理由
1c05c4	陈铖与	交通事故 机动车 畅通 法律 责任 文情 纠纷 准许 事故 股份公司
1c4ff8	武汉市	国土资源 有限公司 股份公司 领土 土地 构件 联合 省长 规划局 占用

TF-IDF | TextRank | Word2Vec | IAKEF-0.5

图11 IAKEF(0.5)算法的裁判文书关键字抽取样本

5 结束语

传统的TF-IDF关键词提取算法仅考虑词语出现的词频及逆词频,具有一定的局限性。本文将融合特征、信息熵、离散度、TextRank引入词语的权重计算公式中,考虑词语的词性、词长、词位置和词跨度等多重因素,提出了一种改进的关键词提取算法,解决了传统算法在语义、类内外分布上的存在的不足的问题。最后通过实验证明本文提出算法的有效性。分别对TF-IDF、TextRank、Word2Vec和改进的关键词提取算法进行实验,对实验结果进行对比分析,实验结果表明改进的关键词抽取算法(IAKEF)实验效果要优于传统的算法,且当加权因子 $\partial=0.5$ 时准确率达到最高。提出的算法分别在语义、类内外分布上的改进策略独立影响虽不显著,但每项影响对计算结果都有积极效果。

本文还有很多不够完善的地方,在以后的学习研究中有以下方面可以加以改进。首先,同义词处理可以加强。由于裁判文书中涉及一些专业性较强的词汇,本文所用的同义词词库来源于网络,对裁判文书中进行同义词识别时效果不是很好。采用专业性更为适当的司法相关命名实体词库和同义词词库是下一步的研究内容。其次,增加处理命名实体识别过程以解决未登录词问题。使用现有的工具词典不能识别的某些司法领域专有词语等未登录词,下一步将通过大规模语料标注和训练以减少未登录词的影响。另外,词语权重计算问题可以通过增强特征采集方法改善。在特征设计上,通过结合更多的特征性提高关键词提取效果,是下一步的研究内容。

致谢 感谢迪安鉴定科学研究院院长常林教授的支持及其实验室的同仁们的帮助。

参考文献:

[1] Ratzek W.International encyclopedia of information and library science[J].BIBLIOTHEK Forschung und Praxis, 2004,28(3):378-379.

[2] 赵京胜,朱巧明,周国栋,等.自动关键词抽取研究综述[J].软件学报,2017,28(9):2431-2449.

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

[4] Ramos J.Using tf-idf to determine word relevance in

document queries[C]//Proceedings of the First Instructional Conference on Machine Learning,2003,242:133-142.
[5] 龚静,周经野.一种基于多重因子加权的文本特征项权值计算方法[J].计算技术与自动化,2007,26(1):81-83.
[6] Rajaraman A,Ullman J D.Data mining[M]//Mining of massive datasets.Cambridge:Cambridge University,2011:1-17.
[7] 侯泽民,巨筱.一种改进的基于潜在语义索引的文本聚类算法[J].计算机与现代化,2014(7):24-27.
[8] Clim A,Zota R D,TinicÁ G.The Kullback-Leibler divergence used in machine learning algorithms for health care applications and hypertension prediction:a literature review[J].Procedia Computer Science,2018,141:448-453.
[9] Weinberg G V.Kullback-Leibler divergence and the pareto-exponential approximation[J].Springer Plus,2016,5(1):604.
[10] 陈科文,张祖平,龙军.文本分类中基于熵的词权重计算方法研究[J].计算机科学与探索,2016,10(9):1299-1309.
[11] How B C,Narayanan K.An empirical study of feature selection for text categorization based on term weightage[C]//Proc of IEEE/WIC/ACM International Conference on Web Intelligence.Washington DC,USA:IEEE Computer Society,2004.
[12] 陈白雪,宋培彦.基于用户自然标注的TF-IDF辅助标引算法及实证研究[J].图书情报工作,2018(1):132-139.
[13] Moulin C,Barat C,Ducottet C.Fusion of tf.idf weighted bag of visual features for image classification[C]//2010 International Workshop on Content-Based Multimedia Indexing(CBMI),2010.
[14] Github.Jieba Tokenizer[EB/OL].[2019-11-15].https://github.com/fxsjy/jieba.
[15] 语言云.LTP分词[EB/OL].[2019-11-15].https://www.ltp-cloud.com/.
[16] 杨凯艳.基于改进的TF-IDF关键词自动提取算法研究[D].湖南湘潭:湘潭大学,2015.
[17] 王良芳.文本挖掘关键词提取算法的研究[D].杭州:浙江工业大学,2013.
[18] 孙建凯.数据挖掘中半监督K均值聚类算法的研究[D].杭州:浙江理工大学,2013.
[19] 任卫杰.基于信息论的特征选择算法研究[D].辽宁大连:大连理工大学,2018.
[20] 李学明,李海瑞,薛亮,等.基于信息增益与信息熵的TF-IDF算法[J].计算机工程,2012,38(8):37-40.
[21] Mihalcea R,Tarau P.TextRank:bringing order into texts[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,2004:404-411.
[22] Mikolov T,Sutskever I,Chen K,et al.Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems,2013:3111-3119.