doi:10.3772/j.issn.2095-915x.2022.01.008

文献摘要结构功能识别在关键词抽取中的应用

孟旭阳 白海燕

中国科学技术信息研究所 北京 100038

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 传统的关键词自动抽取将摘要看成一个整体，常以候选词的出现频次等非语义信息构建特征，并未考虑学术文献摘要中目的、方法、结论等各个结构功能语义蕴含的差异性。本文以中文文献为研究对象，探讨候选词所在的结构功能域对关键词抽取的影响和作用。[方法 / 过程] 本文将文献标题和摘要文本共分为 4 个结构功能域，在传统的词频、词长、词跨度等基准特征上，融合了基于 BERT 的语义特征和结构功能特征，并以不同的特征组合方式，使用图书情报领域的中文学术文献，基于分类模型进行关键词自动抽取实验。[结果 / 结论] 实验结果表明，融合结构功能特征后，关键词抽取效果整体提升了 6.82%，证明了学术文献摘要结构功能的识别形成的结构功能特征对关键词抽取效果的提升有良好作用。

关键词：学术文献；关键词抽取；结构功能；分类模型

中图分类号：G35；G25

Structure-Function Recognition of Literature Abstract and Application in Keyword Extraction

MENG Xuyang BAI Haiyan

Institute of Scientific and Technical of Information of China, Beijing 100038, China

Abstract: [Purpose/significance] Traditional automatic keyword extraction takes the abstract text as a whole, and often constructs features based on the frequency of alternative words, which are ignored the differences in semantic meaning of each structure-function in the academic abstract, such as the purpose, method and conclusion. This paper takes Chinese literature as the research object to research the influence of the structure-function of candidate words on keyword extraction. [Method/process] This paper divided the document title and abstract into 4 structure-function domains. On the traditional benchmark features such as word frequency, word length and word span, we proposed a mixed feature method with academic text semantic features based on BERT and structure features, and conducted automatic keyword extraction experiments based on the classification model using different feature combination methods and Chinese academic literature in the field of library and information. [Result/conclusion] The

作者简介 孟旭阳（1992-），硕士，研究实习员，研究方向为自然语言处理、数字图书馆，E-mail: mengxy@istic.ac.cn；白海燕 (1973-)，硕士，研究馆员，研究方向为信息组织、数字图书馆、关联数据、知识组织系统。

引用格式 孟旭阳 , 白海燕 . 文献摘要结构功能识别在关键词抽取中的应用 [J]. 情报工程 , 2022, 8(1): 79-89.

doi:10.3772/j.issn.2095-915x.2022.01.008

experimental results show that the keyword extraction effect is improved by 6.82% after fusing the structure-function features, which proves that the structure-function recognition of academic abstracts plays a positive role in automatic keyword extraction.
Keywords: Academic literature; keyword extraction; structure-function; classification model

引言

关键词自动抽取是指从文本中自动抽取能够反映文本主题内容的词语，在多个应用领域有着广泛应用，如文本检索、文本聚类、文本分类、文本摘要、文本分析挖掘、推荐领域等[1]，是一直备受关注的研究课题。从学术论文中抽取准确、有效的关键词，可以方便学者根据关键词查找文献，掌握领域最新研究成果，有助于对文献做分类、聚类分析，方便数据的管理和使用，同时在学术推荐和主题发现等应用中都具有重要的基础作用。

关键词抽取方法的现有研究使用的特征主要有：①词频、词长等统计特征；②词间关系、中心度量等图结构特征；③主题特征；④词嵌入向量特征。上述几类特征更多地考虑词汇本身的统计信息和分布特点，忽略了词汇所在结构功能语义上的差异。有部分研究探讨了候选词的位置信息特征，但常指词汇首次出现的索引位置，未曾深入探讨候选词所处结构功能位置对结果的影响。对于学术文献这类特殊的文本内容来说，摘要是文献内容的浓缩，可以让读者方便快捷地了解论文的关键内容。学术文献摘要具有鲜明的逻辑性、目的性、功能性等特点，不同结构功能的语句体现不同的语义，不同结构功能的词汇蕴含的信息量和语义也是不同的。

基于此，本文以中文文献标题和摘要为研究对象，尝试将学术文献摘要的结构功能作为特征，融入关键词自动抽取的特征组合中，同时采用基于有监督的方法，将关键词抽取看作机器学习分类任务，训练分类模型，对关键词自动抽取的效果进行验证和分析，探讨学术文献摘要结构功能特征对关键词自动抽取影响和作用。

1 相关研究

已有许多研究对关键词自动抽取任务进行了探讨，提出了不同的算法和模型，并取得了较好的效果。胡少虎等[2]对关键词抽取的相关研究进行了系统的梳理、分析与总结。根据是否需要提供已经标记好的语料，一般分为无监督和有监督两种方法。

1.1 基于无监督的关键词抽取方法

基于无监督的方法，不需要提前准备标注好的语料，通过利用文本中词语的统计特征和文本语言特点，规定关键词权重的量化指标，计算权重进行排序，最终评估选取出重要的词作为关键词，常见方法如下。

（1）基于简单统计的方法：这种方法侧重于从文本中获得非语言的统计特征，例如词频、词长、单词位置等，进行特征项和权重计算，

最终评估遴选最终结果。Luhn 等 [3] 最先提出了基于词频的简单统计方法，Salton 等 [4] 提出 TF-IDF 算法，综合词汇的词频和文档频率对候选词的重要性进行评分。Matsuo 等 [5] 通过词共现统计信息从文本中提取关键词。这类方法易于理解和实现、简单易用，但是准确率较低。

（2）基于图模型的方法：这种方法基于构建的网络图进行分析寻找关键词。Mihalcea 等 [6] 通过词间的共现关系特征，构建了网络图，并使用 PageRank 算法为每个词打分排序实现关键词的抽取。又有许多学者通过对 TextRank 算法进行优化改进来提升抽取的准确度。李鹏等 [7] 提出了一种 Tag-TextRank 算法，该算法利用 Tag 值优化了图模型节点和边的权重计算。顾益军等 [8] 融合了 LDA 和 TextRank 两种算法进行关键词提取，算法的结合实现两者的优势互补。

（3）基于主题模型的方法。这种方法主要利用主题的分布特性进行关键词抽取。LDA（Latent Dirichlet Allocation）[9]、Lda2Vec[10] 以及 PLDA[11] 等，都是使用主题模型实现关键词抽取的。

1.2 基于有监督的关键词抽取方法

基于有监督的方法，需要提供已标记好的语料训练模型，利用训练好的模型实现文本的关键词自动抽取。根据对关键词抽取任务理解的不同，有监督的关键词抽取方法可以分为基于分类和基于序列标注两种方法。

（1）基于分类的方法：该方法将关键词抽取任务视为二分类问题，即候选词是关键词或不是关键词。根据文本内容信息构建特征，进而提取候选词的特征信息，基于提取的特征信息对模型进行训练实现候选词的分类。Witten 等 [12] 提出了经典 KEA 关键词抽取算法，使用 TFIDF 和词汇首次出现的位置等特征训练朴素贝叶斯模型，实现关键词抽取。Caragea 等 [13] 除上述特征外，还利用引文上下文构造新特征，提出了朴素贝叶斯二分类模型 CeKE，提升了抽取效果。Turney[14] 基于 C4.5 决策树提出了 GenEx 模型。Zhang 等 [15] 基于支持向量机（SVM）算法实现关键词抽取。总的来说，基于分类的方法在抽取关键词的质量上较无监督的方法有了大幅的提高。姜艺等 [16] 考虑了关键词承担的特定角色，即词汇功能，通过实验证明了词汇功能特征在关键词提取中有重要作用。这类算法的相关研究及改进集中在两方面：①特征的改进；②分类模型的改进。

（2）基于序列标注的方法：该方法将关键词抽取视为文本的序列标注问题，即利用序列标注模型学习已标注关键词的句子序列中单词之间的关系，进而为未标注的句子序列进行标注，抽取文本中的关键词。Gollapalli 等 [17] 使用单个特征或组合特征训练 CRFs 模型抽取关键词。Patel 等 [18] 在 TFIDF、相对位置等特征的基础上将词嵌入向量作为特征之一，训练 CRFs 模型实现关键词抽取。同时，随着神经网络的兴起，有学者也开展了相关的研究，Sahrawat 等 [19] 利用 BERT（bidirectional encoder representation from transformers）等预训练模型获得上下文嵌入向量，提出了 BiLSTM-CRF 抽取模型。Martinc 等 [20] 使用 Transformer 模型，提出了 TNT-KID 模型，这些模型都取得了较好的效果。

doi:10.3772/j.issn.2095-915x.2022.01.008

综上所述，基于不同的模式和任务类型，关键词自动抽取都有了较多的研究成果。从抽取的准确率上来看，基于有监督的方法抽取效果更好。模型使用的文本特征，主要考虑词汇在文本中的词频、词长、位置、与其他词的共现关系以及词嵌入向量、上下文信息等对关键词自动抽取的有效性。但是，几乎没有深入讨论研究文本结构功能在关键词自动抽取中的应用，特别是面向学术文献摘要这样文本结构功能明确、不同结构功能语义蕴含和信息量差异较大的对象。因此，本文拟采取有监督的基于分类的机器学习算法，融合学术文献摘要结构功能特点构造相关特征参与关键词抽取，从而探讨学术文献摘要结构功能在关键词自动抽取上的有效性，优化关键词抽取的效果。

2 研究方法

2.1 学术文献摘要结构功能及其自动识别

学术文献摘要的各个结构部分反映了特定的语义功能 [21]。每个结构功能中的语句具有鲜明的逻辑性、功能性和目的性，这些鲜明的特征使文献摘要更加的结构化、语义化。

在摘要结构功能的类型上，一些学者进行了总结分析，张智雄等 [22] 通过收集整理大量具有结构功能标记的论文摘要数据，进行了结构功能类型的统计，结果显示数量最多的结构功能类型分别为：目的、方法、结果、结论。沈思 [21] 通过对情报领域大量期刊的摘要文本结构进行调研，整理得出主要结构功能包括：目的、方法、结果和局限。

在摘要结构功能的自动识别研究上，一般将该识别任务转化为分类或者序列标注问题。王立非等 [23] 构建了基于条件随机场的摘要语步结构自动识别模型。张智雄等 [22] 对比了各类深度学习模型在文献摘要语步识别研究中效果，并剖析了原因。沈思 [21] 基于 LSTM-CRF 的深度学习模型，面向期刊论文摘要，构建了摘要的结构功能自动识别模型，取得了较好的识别效果。

本文重点探讨和研究摘要结构功能对关键词抽取的影响和应用，因此，直接抽取已具有规范摘要结构功能标注的文献数据进行相关的统计分析和在关键词抽取中的应用实验。摘要结构功能标记类型则根据 3.1 节抽取的文献数据调研统计情况进行划分。

2.2 融合文献摘要结构功能特征的关键词抽取

本文采取有监督的基于分类的方法，将关键词自动抽取视为二分类问题，构建机器学习分类模型。关键词抽取的主要流程包括 4 个部分：①领域关键词集构建；②获取候选词集；③特征构建与计算；④分类模型训练；⑤结果评估。整体流程如图 1 所示。

2.2.1 领域关键词集构建及获取候选词集

对于中文学术文献来说，关键词抽取的结果在很大程度上取决于切词的质量。由于学术文献的领域性、专业性极强，一般的切词工具针对大量的专业术语并不能够正确切分。此外，针对一些领域关键短语，会被切词工具切分为单个的词语，例如“朴素贝叶斯网络”会被切分为“朴素”“贝叶斯”和“网络”，几个分

开的词语所各自表达的含义与“朴素贝叶斯网络”作为一个整体所表达的含义具有显著差异，进而影响最终关键词抽取的效果。因此，有必要构建外部领域词库辅助中文文献分词。

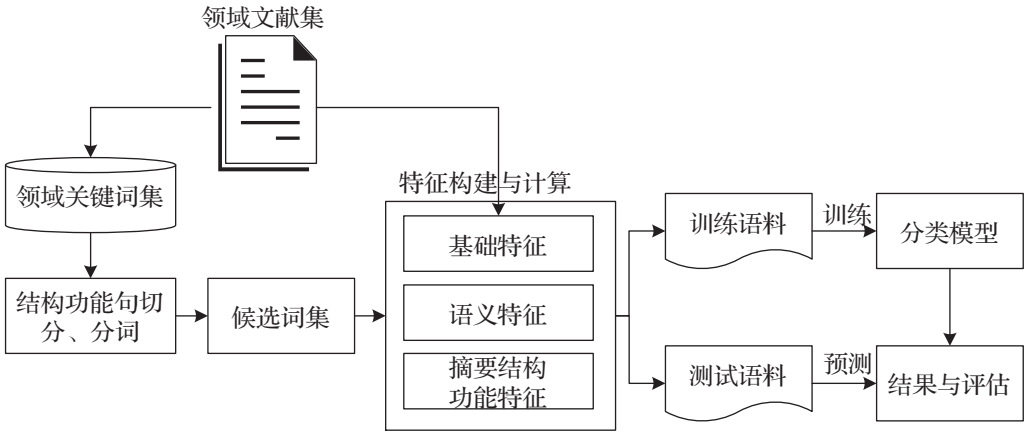

图 **1** 融合文献摘要结构功能特征的关键词抽取

本文首先构建领域文献集 $D=(d_1,d_2,d_2,\cdots, d_N)$，针对文献集 D 抽取全部作者关键词，去除重复词后形成领域关键词集。通过向分词工具添加自定义词典，即构建好的领域关键词集，辅助预处理后的文献分词，分词结果去除停用词、通用词，得到候选词集 $W=(w_1,w_2,w_3,\cdots, w_m)$。为了后续的分类实验，对每个候选词进行标注，若候选词是文献 d 的作者关键词则标记为 1，否则，标记为 0。

2.2.2 特征构建与计算

特征构建是关键词自动抽取的关键，大多算法都是通过对原有特征的改进来优化提取效果的。本研究选取常用的词频、词长、词跨度作为候选词的基准特征。利用 BERT 预训练语言模型获取候选词和文本的语义向量，计算候选词与文本的语义相关性作为候选词的语义特征，通过该特征挖掘隐式的语义信息。根据识别算法或者切分得到的“标题”“目的”“方法”和“结论”4 类结构功能语义段，获得候选词的结构功能特征。

（1）基准特征

基准特征包含词频、词长、词跨度三个特征。词频，即词在文本中出现的次数，是信息检索和数据挖掘中常使用的一种统计指标。词长，即候选词汇的长度，一般认为候选词长度越长，它所能表达的信息就更全面 [24]。词跨度，表示词语在文本中首次出现和最后一次出现位置间的距离。一般认为，距离越大说明其在文中的影响范围越广，越能反映文章的主题 [25]。

（2）语义特征

关键词提取的目的就是要提取出能够反映文本主题语义的词语。BERT[26] 是 Google 在 2018 年提出的预训练语言模型，它基于双向 Transformer 网络结构作为编码器，语义表达能力优势显著。本文借助 Google 公开的中文版 BERT 预训练语言模型生成候选词与文本内容的语义向量，有效获取词汇和文本的语义信息，

doi:10.3772/j.issn.2095-915x.2022.01.008

然后根据生成的语义向量计算候选词与文本内容语义之间的相似度，并将其作为候选词的语义特征。

（3）结构功能特征

结构功能特征的引入是为了弥补基准特征在细粒度层面特征表示的不足，以提升关键词抽取的效果。本文面对中文文献标题和摘要进行关键词抽取，根据上节对学术文献摘要的结构功能划分的描述，将文献数据（标题+摘要）表示为 S=（s1,s2,s3,s4），分别表示“标题”“目的”“方法”“结论”。在对结构功能特征的表示方式上，采用直接增加新的特征维度且采用布尔值来表示候选词是否在文献的特定结构功能片段中出现。

2.2.3 模型训练

将关键词自动抽取视为二分类问题，即对候选词进行二值判断（是关键词或不是关键词）。本文基于 Python 机器学习工具 scikit-learn，实现 SVM 分类模型算法，并利用不同特征组合的训练数据（具体设置见 3.3 节）分别训练关键词分类模型，从而对比分析融合结构功能特征的关键词抽取作用的效果。

2.2.4 结果评价

对于关键词抽取结果的评价，我们分两个方面进行评估。第一，以候选关键词为单位，对于二分类模型的效果进行评价，采用准确率 P、召回率 R、F1 值为评价指标，评估 SVM 模型对关键词的判别能力。第二，以文献为单位，本文直接采用文献的关键词作为关键词抽取任务的抽取目标，并作为评价关键词抽取的依据。采用准确率 P、召回率 R、F1 值为评价指标，评估针对文献的关键词抽取能力。

3 试验与分析

3.1 实验数据说明

本文选定情报领域的《图书情报工作》《情报杂志》《情报科学》《情报理论与实践》这 4 种摘要具有结构化标记的期刊进行数据抽取。绝大多数文献具体被标记为【目的 / 意义】、【方法 / 过程】和【结果 / 结论】，少数文献除上述标记之外多一个【局限】标记。为了获取统一规范化摘要结构功能标注的期刊文献数据，本文从 NSTL 抽取了上述 4 种期刊的文献数据共 4 万篇，并依据【目的 / 意义】、【方法 / 过程】和【结果 / 结论】三个标记规则进行过滤，抽取摘要中包含且仅包含此三种标记的文献数据，经过滤共得到 6360 篇统一规范化标注的文献数据。此外，根据上述 6360 篇文献数据构建领域关键词集，即文献中作者关键词集合，共 27602 个，用于辅助分词。

3.2 数据统计分析

本研究对获取的具有统一规范化摘要结构功能标注的 6360 篇文献数据中的作者关键词（共 27602 个）进行了相应的统计分析。

其中，文献关键词个数分布情况如图 2 所示。由图 2 可看出，文献关键词数在 3 ∽ 5 个的居多。

词语长度不同，成为关键词的概率也是不一样的。因此，对文献关键词长度分布进行统计，结果如表 1 所示。

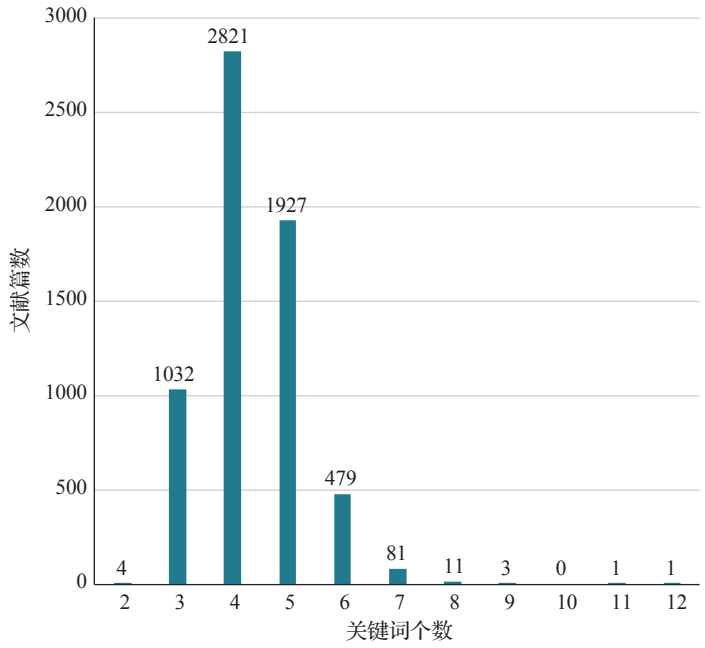

图 **2** 文献作者关键词个数分布情况

表 **1** 作者关键词（**27602** 个）长度统计

词长	1	2~3	4	5~6	>6
词数	10	3737	14695	6997	2163
占比/%	0.03	13.54	53.24	25.35	7.84

由表 1 可看出，关键词长度多集中在 2 ∽ 6 之间，总占比超过 92%，长度为 4 的关键词最多，占总数的一半以上。

分别统计作者关键词在各结构功能区域中的分布情况。图 3 描述了作者关键词在各结构功能区域中的分布情况。由于一些关键词会未出现在文本中，或者同时出现在多个结构功能区域中，表 2 分别统计了作者关键词未出现在文本中、出现在 1 个、2 个、3 个和 4 个结构功能区域中的关键词数和占比。

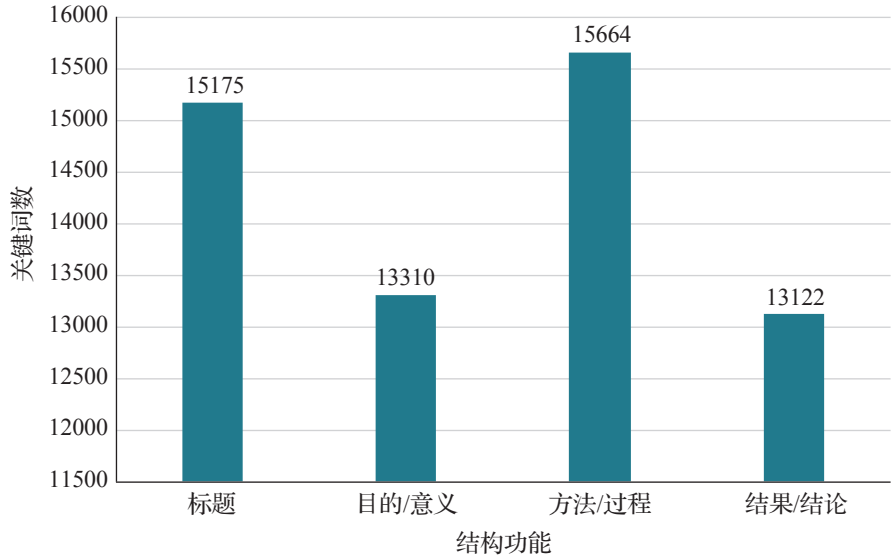

图 **3** 作者关键词在各结构功能中的分布

doi:10.3772/j.issn.2095-915x.2022.01.008

由图 3 可看出，在方法 / 过程和标题结构功能区域中出现的关键词最多，结果 / 结论中最少。

表 2 作者关键词在各结构功能中的分布

	未出现	1个	2个	3个	4个
词数	5426	5877	4572	4658	7069
占比/%	19.66	21.29	16.56	16.88	25.61

由表 2 可看出，超过 80% 的关键词都是出现在文本中（标题 + 摘要）的，这从侧面说明了大部分的作者关键词来自文本中，从文本中抽词是合适的。同时，可看出有超过 25% 的关键词同时出现在了 4 个结构功能区域。

3.3 实验设置

对于 6360 篇具有规范化摘要结构功能标注的文献，实验按照 9:1 的比例划分为训练集和测试集。对文献标题和摘要进行预处理、分词（添加自定义领域关键词集辅助分词）、去除停用词、通用词，得到候选词集。对于二分类模型，获取每个候选词的特征数据，并以是否为作者关键词为依据为每个候选词打上标记 1/0。经统计，标记为 1 的关键词数据有 20316 个，标记为 0 的非关键词数据有 327062 个。为了解决数据不平衡对分类器的影响，实验选择训练文献集中全部标记为 1 的候选词特征数据，并随机抽取等量标记为 0 的候选词特征数据作为训练集。抽取每篇测试文献中所有标记为 1 的候选词特征数据和等量标记为 0 的候选词特征数据作为测试集，以此训练 SVM 分类器。

另外，实验分别对基准特征、语义特征和结构特征进行归一化处理。为方便记录，后面统一将基准特征记为①，语义特征记为②，结构特征记为③。实验设置了 4 组不同的特征组合进行二分类实验，分别为：①，仅使用基准特征；①＋②，使用基准特征和语义特征组合；①＋③，使用基准特征和结构特征组合；①＋②＋③，使用基准特征、语义特征和结构特征组合。

3.4 实验结果与分析

以关键词为单位评估 SVM 二分类模型识别关键词的性能，4 组不同特征组合下的分类效果如表 3 所示。

表 3 以关键词为单位 SVM 分类实验结果

	①	①+②	①+③	①+②+③
P	0.8662	0.8662	0.8733	0.8726
R	0.8728	0.8715	0.8735	0.8729
F1	0.8656	0.8657	0.8733	0.8726

由表 3 可看出，融合了结构功能特征①＋③的模型比用基准特征①的整体性能提升了 1% 左右，略有效果，但是效果不显著。同时，也可看出，语义特征对模型性能的提升效果不佳。

以文献为单位，对测试集中文献关键词抽取的能力进行统计评估，4 组不同特征组合下的关键词抽取的平均准确率、召回率和 F1 值统计结果如表 4 所示。

表 4 以文献为单位关键词抽取实验结果

	①	①+②	①+③	①+②+③
P	0.9419	0.9350	0.8986	0.8992
R	0.2496	0.2538	0.3163	0.3164
F1	0.3823	0.3867	0.4505	0.4507

从表 4 可看出，整体上，关键词抽取的准确率高，召回率低，F1 值是综合准确率与召

回率的指标。表 6 中可看出，在关键词提取的 F1 值上，融合了结构功能特征① + ③相较于仅用基准特征①提升了 6.82 %，这充分说明，在基于二分类的关键词自动提取上，摘要结构功能特征具有显著的积极作用。语义特征对关键词抽取的效果不是很显著。由于本文直接采用 Google 官方的 BERT 中文预训练模型，不是很适用于我们特定领域的文献语义计算，后续需要根据特定领域的文献数据训练合适的词向量模型，再次进行评估。

本文主要将关键词自动抽取视为二分类问题，以上均是基于 SVM 进行了结构功能特征作用的相关实验论证。同时，也与目前经典且流行的序列标注任务类型的取模型 Bert-Bi-LSTM-CRF 进行效果对比分析。为保证模型效果对比在其他维度上的统一性，此处采用与 SVM 同样的训练集和测试集，根据作者关键词为文本序列做自动标注，模型抽取效果对比分析结果如表 5 所示。

表 **5** 不同模型抽取效果对比分析

	P	R	F1
本文方法	0. 8992	0. 3164	0. 4507
Bert-Bi-LSTM-CRF	0.3341	0.6620	0.4441

如表 5 所示，在准确率上本文方法结果较优，在召回率上，Bert-Bi-LSTM-CRF 模型的效果较优，从 F1 值整体上来看，本文方法的抽取效果较优。Bert-Bi-LSTM-CRF 基于谷歌预训练的 Bert 模型，在上下文特征的获取及语义理解能力上具有较大的优势，但该模型依赖大量的有标注的训练语料，同时对硬件的要求比较高。在训练语料有限的情况下，本文方法的结果较优。

为了针对模型的进一步改进和优化提供合理的建议，此处选取了一些抽取效果不好几个代表性示例进行深入分析，结果如表 6 所示。

表 **6** 关键词抽取示例分析

示例	标题	摘要	作者关键词	本文方法抽取结果
1	基于句法规则和社会网络分析的网络舆情热点主题可视化及演化研究	[目的 / 意义] 为直观、深入地认知网络舆情热点主题演化提供新的方法和视角。[方法 / 过程] 基于句法规则和社会网络聚类提出了舆情热点主题发现和可视化分析框架，依据五条句法规则将舆情热点词关联，从而构建舆情生命周期内不同阶段的加权无向关系网络 , 然后进行主题聚类和可视化展示，最后从宏观和微观两个角度对热点主题的演化规律进行量化分析。[结果 / 结论] 所提出的方法能够直观、全面、深入地揭示各个阶段舆情的热点主题及其演化规律。	网络舆情； 主题发现； 句法规则； 可视化；	热点主题； 句法规则； 社会网络分析； 网络舆情； 社会网络； 演化规律； 可视化分析； 可视化； 量化分析； 主题聚类；
2	基于 SOM 神经网络和排序因子分解机的图书资源精准推荐	[目的 / 意义] 传统基于协同过滤的图书资源推荐算法难以处理数据稀疏问题，而传统基于矩阵分解的推荐算法在处理高维数据时可扩展性差，且它们的推荐结果仅依据预测评分大小确定，导致推荐准确度不高 . 鉴于此，文章提出基于 SOM 神经网络和排序因子分解机的图书资源推荐方法 . [方法 / 过程] 该方法首先利用 SOM 神经网络，基于用户学术背景信息对用户进行聚类，然后利用用户对图书资源的显式和隐式 Web 访问行为构建图书资源偏序关系，最后利用因子分解机 (FM) 作为排序函数对用户学术背景、Web 访问行为和借阅图书简介文本等多种特征信息进行建模，并使用对级 (Pairwise) 排序学习算法实现图书资源的精准推荐 . [结果 / 结论] 实验结果表明，文章所提出的方法能有效缓解数据稀疏问题，提高推荐的准确率和效率 .	SOM 神经网络； 排序因子分解机； 排序学习； 图书推荐； 个性化服务；	精准推荐； 排序学习算法； SOM 神经网络； 排序因子分解机； 图书； 推荐算法； 资源；

doi:10.3772/j.issn.2095-915x.2022.01.008

由表 5 中的示例中可看出，本文方法能够较好的抽取文中的关键词，虽然有些结果并未出现在作者关键词中，但经分析确实能够反映文章内容的主题内容。但抽取的结果也有不尽如人意的地方，如示例 1 的抽取结果中出现了“可视化分析”和“可视化”两个结果，明显是重复的，如何进一步的进行语义去重是后续需要进一步优化的。示例 2 结果中的“图书”“推荐算法”“资源”各自的语义并不完整丰富，期望的语义完整的结果为“图书资源推荐 (算法)”。因此，后续可在语义完整度上进一步的优化抽取结果。

4 结束语

本文为探讨候选词所在的结构功能对关键词抽取的影响，将文献（标题 + 摘要）标记为 4 个结构功能域，采用基于分类的关键词抽取方法，构建了候选词的基准特征、语义特征和结构特征，并采用不同的特征组合方式，以 SVM 二分类模型实现文献的关键词自动抽取。实验结果表明，候选词的结构功能特征对关键词抽取的提升起到了积极作用，在一定程度上提升了关键词的抽取效果。

本文提出的融合结构功能特征的关键词自动抽取方法具有较好的实验结果，但仍存在一些问题需要进一步探索研究：首先，本文直接采用具有规范结构功能标记的数据探讨结构功能特征对关键词抽取的影响，在实际应用中，对无标记数据需要进一步研究摘要结构功能的自动识别算法。其次，在结果评价上使用作者关键词作为正确依据，在节省人力标注成本下，从一定层面客观反映了抽取效果，但从统计数据显示 19.66% 的作者关键词未出现在文本中，因此该评价方法并不能够完全反映抽取实际效果，需要进一步增加人工标注，更全面、精准的进行评价。最后，本文采用图书情报领域的部分文献进行实验探究，相关结论具有一定的领域局限性，下一步将进一步扩大领域范围和数据规模进一步探索研究。

参 考 文 献

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

[2] 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.

[3] 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.

[4] Salton G, Buckley C. Term-Weighting Approaches in Automatic Text Retrieval[J]. Information Processing & Management, 1988, 24(5):513-523.

[5] Matsuo Y, Ishizuka M. Keyword Extraction from a Document using Word Co-occurrence Statistical Information [J]. Transactions of the Japanese Society for Artificial Intelligence, 2002, 17(3):217-223.

[6] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[C]. Proceeding Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain. 2004:404-411.

[7] 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.

[8] 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.

[9] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[10] Moody C E. Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec[J/OL]. arXiv

Preprint, arXiv:1605. 02019.

[11] Liu Z, Chen X, Zheng Y, et al. Automatic Keyphrase Extraction by Bridging Vocabulary Gap[C]. Proceedings of the 15th Conference on Computational Natural Language Learning. 2011:135-144.

[12] Witten I H, Paynter G W, Frank E, et al. KEA: Practical Automatic Keyphrase Extraction[C]. Proceedings of the Fourth ACM Conference on Digital Libraries. New York: ACM Press, 1999: 254-255.

[13] Caragea C, Bulgarov F, Godea A, et al. Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1435-1446.

[14] Turney P D. Learning Algorithms for Keyphrase Extraction[J]. Information Retrieval, 2000, 2(4):303-336.

[15] Zhang K, Xu H, Tang J, et al. Keyword Extraction Using Support Vector Machine[C]. Proceedings of the International Conference on Web-Age Information Management. Heidelberg: Springer, 2006: 85-96.

[16] 姜艺 , 黄永 , 夏义堃 , 等 . 学术文本词汇功能识别——在关键词自动抽取中的应用 [J]. 情报学报 , 2021, 40(2):152-162.

[17] Gollapalli S D, Li X L. Keyphrase Extraction using Sequential Labeling[J/OL]. https://arxiv.org/pdf/1608.00329v1.pdf

[18] Patel K, Caragea C. Exploring Word Embeddings in CRF- based Keyphrase Extraction from Research Papers[C]. Proceedings of the 10th International Conference on Knowledge Capture. New York: ACM Press, 2019: 37-44.

[19] Sahrawat D, Mahata D, Zhang H, et al. Keyphrase Extraction as Sequence Labeling Using Contextualized Embeddings[C]. Proceedings of the European Conference on Information Retrieval. Cham: Springer, 2020: 328-335.

[20] Martinc M, Krlj B, Pollak S. TNT-KID: Transformer-based Neural Tagger for Keyword Identification [J/OL]. https://arxiv.org/pdf/2003.09166.pdf.

[21] 沈思 , 胡昊天 , 叶文豪 , 等 . 基于全字语义的摘要结构功能自动识别研究 [J]. 情报学报 , 2019, 38(1):79-88.

[22] 张智雄 , 刘欢 , 丁良萍 , 等 . 不同深度学习模型的科技论文摘要语步识别效果对比研究 [J]. 数据分析与知识发现 , 2020, 3(12):1-9.

[23] 王立非 , 刘霞 . 英语学术论文摘要语步结构自动识别模型的构建 [J]. 外语电化教学 , 2017(02):45-50.

[24] 陈伟鹤 , 刘云 . 基于词或词组长度和频数的短中文文本关键词提取算法 [J]. 计算机科学 , 2016(12):57-64.

[25] 谢晋 . 基于词跨度的中文文本关键词自动提取方法 [J]. 现代经济：现代物业中旬刊 , 2012(4):108-111.

[26] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies. Minneapolis, USA: IOA Press, 2019: 4171-4186.