软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(7):2046−2070 [doi: 10.13328/j.cnki.jos.005538]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

特征驱动的关键词提取算法综述*

常耀成[1], 张宇翔[1], 王 红[1], 万怀宇[2], 肖春景[1]

[1](中国民航大学 计算机科学与技术学院,天津 300300)

[2](北京交通大学 计算机与信息技术学院,北京 100044)

通讯作者: 张宇翔, E-mail: yxzhang@cauc.edu.cn

摘 要: 面向文本的关键词自动提取一直以来是自然语言处理领域的一个关键基础问题和研究热点.特别是,随着当前对文本数据应用需求的不断增加,使得关键词提取技术进一步得到研究者的广泛关注.尽管近年来关键词提取技术得到长足的发展,但提取结果目前还远未取得令人满意的效果.为了促进关键词提取问题的解决,对近年来国内、外学者在该研究领域取得的成果进行了系统总结,具体包括候选关键词生成、特征工程和关键词提取 3 个主要步骤,并对未来可能的研究方向进行了探讨和展望.不同于围绕提取方法进行总结的综述文献,主要围绕着各种方法使用的特征信息归纳总结现有成果,这种从特征驱动的视角考察现有研究成果的方式有助于综合利用现有特征或提出新特征,进而提出更有效的关键词提取方法.

关键词: 关键词提取;候选关键词生成;特征;有监督方法;图方法

中图法分类号: TP391

中文引用格式: 常耀成,张宇翔,王红,万怀宇,肖春景.特征驱动的关键词提取算法综述.软件学报,2018,29(7):2046−2070. http://www.jos.org.cn/1000-9825/5538.htm

英文引用格式: Chang YC, Zhang YX, Wang H, Wan HY, Xiao CJ. Features oriented survey of state-of-the-art keyphrase extraction algorithms. Ruan Jian Xue Bao/Journal of Software, 2018,29(7):2046−2070 (in Chinese). http://www.jos.org.cn/1000-9825/5538.htm

Features Oriented Survey of State-of-the-Art Keyphrase Extraction Algorithms

CHANG Yao-Cheng[1], ZHANG Yu-Xiang[1], WANG Hong[1], WAN Huai-Yu[2], XIAO Chun-Jing[1]

[1](School of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)

[2](School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Keyphrases that efficiently represent the main topics discussed in a document are widely used in various document processing tasks, and automatic keyphrase extraction has been one of fundamental problems and hot research issues in the field of natural language processing (NLP). Although automatic keyphrase extraction has received a lot of attention and the extraction technologies have developed quickly, the state-of-the-art performance on this task is far from satisfactory. In order to help to solve the keyphrase extraction problem, this paper presents a survey of the latest development in keyphrase extraction, mainly including candidate keyphrase generation, feature engineering and keyphrase extraction models. In addition, some published datasets are listed, the evaluation approaches are analyzed, and the challenges and trends of automatic keyword extraction techniques are also discussed. Different from the existing surveys that mainly focus on the models of keyphrase extraction, this paper provides a features oriented survey of automatic keyphrase extraction. This perspective may help to utilize the existing features and propose the new effective extraction approaches.

Key words: keyphrase extraction; candidate keyphrase generation; feature; supervised approach; graph-based approach

* 基金项目: 国家自然科学基金(U1533104, U1633110, 61603028); 中央高校基本科研业务费(ZXH2012P009)

Foundation item: National Natural Science Foundation of China (U1533104, U1633110, 61603028); Fundamental Research Funds for the Central Universities (ZXH2012P009)

收稿时间: 2017-07-19; 修改时间: 2017-11-02; 采用时间: 2018-01-04; jos 在线出版时间: 2018-02-08

CNKI 网络优先出版: 2018-02-08 11:55:30, http://kns.cnki.net/kcms/detail/11.2560.TP.20180208.1155.001.html

随着文本数据(如科技文献、Web 页面、社交推文等)的快速增长,对文本数据的分析和挖掘成为当前备受关注的重要研究领域.其中,如何从文本文档中提取反映文档主题的关键词(keyphrases,包括单词或词组)一直以来都是自然语言处理领域亟待解决的关键基础问题和研究热点,其研究成果可广泛用于文档检索[1]、文本摘要[2]、文本分类[3]、话题检测[4,5]、问答系统[6−8]等具体应用领域.特别是,随着近年来对非结构化文本大数据研究的兴起,关键词提取问题进一步得到广泛关注和深入研究,一些研究成果及时出现在近几年人工智能、自然语言处理等相关领域的顶级国际会议上,如 IJCAI 2013[9]、AAAI 2014[10]、ACL 2014[11]、ACL 2015[12]、WWW 2015[13]、EMNLP 2016[14]、AAAI 2017[15,16]、ACL 2017[17]等.

早在 20 世纪 70~80 年代,研究者直接利用词频-逆文档频率(term frequency-inverse document frequency,简称 TF-IDF)提取关键词[18,19],简单地将 TF-IDF 大于设定阈值的词分为关键词,这是早期提出的朴素无监督关键词提取方法.随着研究的不断推进,1999 年,Turney[20]、Frank[21]等研究者相继尝试利用有监督分类学习方法自动提取关键词.随后,沿着有监督和无监督两个方向不断提出新的关键词提取方法,直到最近,如在 2017 年,Florescu 等人[16]在 AAAI 发表文章提出基于随机游走的无监督关键词提取方法 PositionRank,同年,Gollapalli 等人[15]也在 AAAI 发表文章提出基于条件随机场的有监督关键词提取方法.

有监督关键词提取方法通常是将关键词提取问题视作二分类问题处理.具体来说,首先对训练集进行标注,即将关键词标注为正样本和将非关键词标注为负样本,然后根据正、负样本的特征集学习一个分类函数,最后用学习得到的分类函数去判别新的候选关键词是否为关键词.而无监督提取方法通常是采取各种评分指标(如 TF-IDF、基于词图的度中心性等)对候选关键词进行排序,然后选取排名最高的几个作为关键词.通常情况下,对于给定的任务集,有监督方法提取效果略优于无监督方法,可是多数情况下训练数据并不易获得,甚至标注代价很大,且学习得到的分类函数受限于训练数据的特征,还可能存在过拟合的问题.随着无监督方法的不断改进,其提取性能越来越接近有监督方法[16,22].此外,因无监督方法不需要事先标注数据,从而一直得到研究者的广泛关注,特别是近年来基于图的无监督关键词提取方法取得了长足的进步.

学术界普遍认为提取的关键词应该满足以下几个基本准则.清华大学的刘知远在文献[23]和复旦大学的丁卓冶在文献[24]中都各自提出一些基本准则,本文对这些准则进行了归纳总结并将“重要性”列为准则之一.

(1) 可读性(readability)[23].关键词本身应该是有意义的词或者短语.例如,“关键词提取”是一个有意义的短语,而“关键提取”则不是.

(2) 相关性(relevance)[23,24].关键词必须与文档主题相关.例如,本文综述关键词提取,其中第 1 段顺带提到“话题检测”,显然不希望这个短语被选作本文的关键词.

(3) 重要性(importance).关键词应该是文档中最重要的词语,可进一步认为是文档主题分布上的最重要的词语.尽管已有文献没有将其列出,然而这显然是一个必须要满足的最重要的特性.

(4) 覆盖度(coverage)[23,24].关键词要能够对文档的主题有较好的覆盖,不能只集中在文档某个主题而忽略了文档其他主题.

(5) 一致性(coherence)[24].文档的几个关键词在语义和逻辑上应有关联,所表述的意思构成一个逻辑统一体.例如,一篇主要介绍图方法关键词提取的学术论文,关键词集合为{“关键词提取”、“图方法”}较{“关键词提取”、“应用价值”}更合适.

要满足上述准则中的任何一个都面临重大的技术挑战,如关键词可读性的提升要依赖于分词等关键词生成技术(详见第 2 节)的进一步发展,关键词与文档主题的相关性要依赖于现有概率主题模型的进一步改进.再如在重要性方面,现有的关键词提取技术很大程度上依赖于候选关键词的出现频率,或者说,出现频率低的词,其最终的评分较低,这样在短文本中会漏掉出现频率低的关键词[25].此外,现有关键词提取方法大都针对上述 1~2 个基本准则展开设计[26],而一种方法中同时兼顾更多的上述基本准则会非常困难.因此,目前的关键词提取方法远未取得令人满意的效果.

由此可见,一方面,在当前对文本数据分析应用的急迫需求的推动下,关键词提取研究在近些年来取得了许多成果;另一方面,在相关技术不能突破的情况下,自动提取的关键词远未达到令人满意的效果.因此,有必要对

现有研究成果进行系统的分析、对比和总结,以便于研究者未来在此基础上提出更好的解决方法.

为了能够系统综述相关研究成果,我们查阅了近年来绝大部分的相关研究工作,包括人工智能、自然语言处理、机器学习等相关领域的国际会议和学术期刊.值得一提的是,2014 年 Hasan 等人[11]在国际计算语言学年会(ACL)和 2017 年赵京胜等人[27]在《软件学报》上分别发表关键词提取研究综述论文,前者主要围绕“提取方法”对当时的研究成果进行了总结(即围绕图 1 中的左侧第(4)步),后者更多的是从相对宏观的角度(如原文所述“从语言学、认知科学、复杂性科学、心理学和社会科学等多个方面研究了自动关键词抽取的理论基础[27]”)和围绕“提取方法”总结现有研究成果.而本文主要围绕“文本数据特征”对近期的研究成果进行详细总结(即围绕图 1 中的左侧第(3)步),这样做的好处是可以从更基础的(或更底层的)角度对现有解决方法进行考察,这将有助于综合利用现有的特征和提出新的特征,进而提出更有效的关键词提取方法.相对于上述两篇综述文献,以“文本数据特征”作为出发点使得本文的内容更加微观和具体,能够深入到关键词提取的每个细节,这将有助于提取方法的具体实现.总之,本文的出发点不同于上述两篇综述,内容很少与上述两篇综述重复,而是与上述两篇综述相互补充.

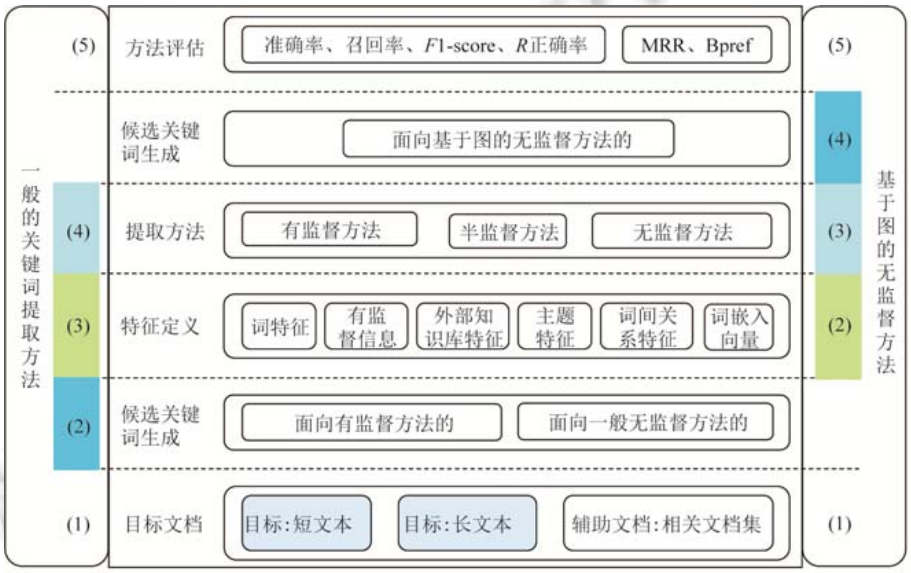

Fig.1 The framework of keyphrase extraction approaches

图 1 关键词提取的具体步骤及解决框架

除本节之外,第 1 节给出问题的形式化定义以及通用解决框架.第 2 节对候选关键词生成(即针对文本的预处理)进行详细总结.第 3 节详细总结目前文献中提出的文本数据特征,因提取方法依赖于其所用的特征,故对其进行详细归纳.第 4 节以特征为驱动详细介绍关键词提取的主要方法.第 5 节归纳常用数据集、评估方法及指标,方便研究者开展评估实验.最后总结全文,并对未来值得关注的研究方向进行初步探讨.

1 问题定义及解决框架

1.1 问题定义

设 $D=\{d_1,d_2,\ldots,d_c\}$为包含 c 篇文档的集合.每篇文档 $d_i\in D$ 中包含一个由 n_i 个单词组成的集合 $W=\{w_{i,1},w_{i,2},\ldots,w_{i,n_i}\}$.关键词提取的目标是:

(1) 生成候选关键词(包括单词或词组)集合 $P_i=\{p_{i,1},p_{i,2},\ldots,p_{i,m_i}\}$.

(2) 找到一个函数,其将候选关键词 $p_{i,j}\in P_i$ 映射至某个类别或者分值,然后根据候选关键词的类别或者分值

从候选关键词集合中抽取出最能概括目标文档 d_i 的一组候选关键词.

上述两个步骤的次序被多数关键词提取方法所遵循,然而在基于图的无监督提取方法(详见第 4.2 节)中,这两个步骤的次序恰好相反,通常是先从单词集合 W 中选出可能成为组成关键词的单词(通常保留名词和形容词),然后利用图方法对这些单词打分,最后在此基础上生成候选关键词集合 P 并取分数排名靠前的作为关键词.这种差异也可见图 1 左右两侧的编号,其代表不同方法所包含的具体步骤的先后次序.

1.2 解决方案框架

解决关键词提取问题的框架和具体步骤如图 1 所示.

(1) 确定目标文档.确定要提取的文档是长文本还是短文本,是学术文献还是其他文档,如 Web 页面、医学诊断报告等.此外,还可考虑是否借助辅助文档集或外部知识库来丰富文本信息进而提高提取效果.

(2) 生成候选关键词.候选关键词生成方法详见第 2 节,其在整个解决框架中的位置次序(如第 1.1 节所述)视具体提取方法而定.

(3) 定义特征.在关键词提取中目前常用的特征主要是图 1 中列举的 6 类特征,不同的方法采用不同的特征.如前言所述,本文主要围绕提取方法中用到的文本数据特征展开,故将对已有文献中提出和使用的文本特征进行详细的归纳总结,具体见第 3 节.

(4) 确定提取方法.如前言所述,关键词提取大致可分为有监督和无监督两类方法以及少数的半监督方法.结合拟采用的文本特征最终确定具体的关键词提取方法.详细介绍见第 4 节.

(5) 评估提取方法的性能.最后对实验结果进行评估,评估方法、指标及常用的公开数据集详见第 5 节.

上述步骤紧密联系,接下来详述每一个具体步骤.

2 候选关键词生成

因几乎所有的关键词提取方法均会涉及候选关键词生成,故候选关键词生成的准确程度直接影响到关键词提取效果[28].本节从两个维度对现有候选关键词生成方法进行归纳和总结,并概括了这些不同候选关键词生成方法对不同关键词提取方法的影响.如图 2 所示,从横向视角来看,每种方法大致包括 4 步:切分文本串、过滤处理、生成多元词组以及取词干,每一步又包括若干具体操作.从纵向视角来看,现有候选关键词筛选方法大致可分为 4 种.

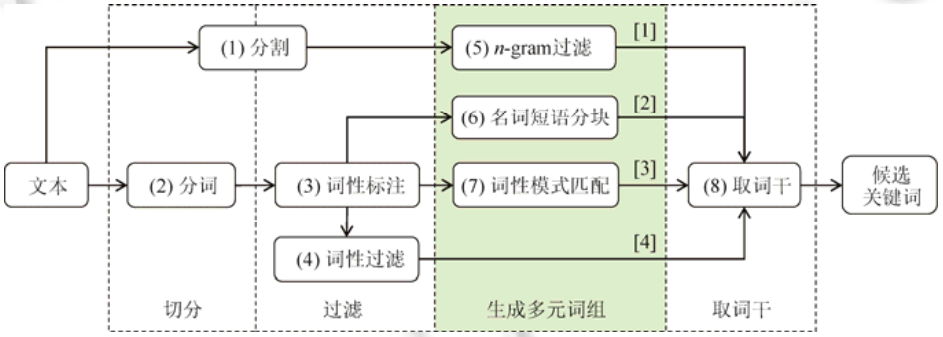

Fig.2 The basic process of the candidate keyphrase generation

图 2 候选关键词生成基本流程

涉及的 8 个具体操作介绍如下.

(1) 分割(spliting).以停用词和符号作为分界符分割文本串,得到去除了停用词与符号的文本串片段.

(2) 分词(tokenizing).区别于分割操作,分词仅将文本串切分为单词.

(3) 词性标注(part-of-speech tagging,简称 POST)[29].使用 POS 标签标注单词的词性.

(4) 词性过滤(POS filtering).根据单词的词性标签,滤掉不符合词性要求的单词,通常仅保留形容词和名词.

(5) *n*-gram 过滤(*n*-gram filtering)[28].针对文本串分割后的每个文本片段,对其进行遍历得到所有的 *n* 元单词序列(通常 $n\leqslant 3$),接着按照一定的规则筛选出符合要求的多元词组,如保留词频较高的词组[28]、保留内部紧密度(可用特征 F4.1 计算)较高的词组[23,30].过滤规则对多元词组生成的效果影响较大[31].

(6) 名词短语分块(NP-chunking)[32].因关键词通常是名词短语,故需从词性标注后的文本序列中识别出名词短语.名词短语的词性构成通常遵循特定的排列模式,如"形容词+名词"模式,因此,通常采用模式匹配结合句法规则来识别名词短语[33,34].

(7) 词性模式匹配(POS tag patterns)[32].该操作的基本思想与名词短语分块相同,均默认关键词的词性序列遵循特定排列模式.不同之处在于,词性模式匹配既包括名词短语又包括非名词短语.为了得到更全面的候选关键词,显然相对于名词短语分块操作,该操作需要定义更复杂的匹配规则.如文献[35]结合句法分析树来生成的多元词组,除了名词短语外,还可包括一些非名词短语.更进一步地,文献[36,37]使用带间隔约束的序列模式挖掘算法来生成多元词组,利用间隔约束实现对模式的自动定义.

(8) 取词干(stemming)[38,39].抽取词的词干,以便消除不同词态形式的影响.

现有候选关键词生成方法由上述基本操作组合而成,这些方法之间的区别主要体现在多元词组生成环节上,具体生成方法如下.

[1] 分割+*n*-gram 过滤+取词干.该方法使用 *n*-gram 过滤来生成多元词组,优点是实现简单且可以灵活设置多元词组长度.

[2] 分词+词性标记+名词短语分块+取词干.该方法只生成名词短语,但往往容易漏掉非名词短语的关键词[32].

[3] 分词+词性标注+词性模式匹配+取词干.该方法需要根据待提取关键词的文档类别定义相应的匹配规则[32],实现难度较大,且针对不同的目标文档集通常需要定义不同的匹配规则,其优点是既可筛选出名词短语也可筛选出非名词短语.

[4] 分词+词性标注+词性过滤+取词干.该方法多用于基于图的关键词提取.首先通过词性过滤仅保留形容词和名词,然后构成词图并应用图算法对节点单词进行评分,最后对排名靠前的单词生成多元词组,常用 *n*-gram 过滤和名词短语分块方法,该过程常被称为后期处理(post-process)[40,41].

近年来一些文献专门分析了候选关键词生成对最终关键词提取结果的影响.文献[42]对比了不同的多元词组生成方式对无监督关键词提取方法的影响,实验结果表明,名词短语块分块方法更适合无监督方法.文献[28]对比了不同候选关键词生成方法对有监督和无监督方法的影响,实验结果表明,候选关键词生成方法对无监督方法的影响更大,而对有监督方法的影响较小.

对于上述基本操作有成熟工具包可直接使用,针对英文有宾夕法尼亚大学的综合工具包 NLTK(Python)(http://www.nltk.org/)和斯坦福大学的词性标注工具 POS Tagger(Java)(https://nlp.stanford.edu/software/tagger.shtml);针对中文有清华大学的 THULAC 中文词法分析工具(Python)(https://github.com/thunlp/THULAC-Python)和复旦大学的 FudanNLP(Java)(https://github.com/FudanNLP/fnlp)、北京理工大学的 NLPIR(Java)(https://github.com/NLPIR-team/NLPIR)、哈尔滨工业大学的 LTP(C++)(https://github.com/HIT-SCIR/ltp)等综合工具包.

3 特征定义

特征定义(有的文献称其为特征提取或特征工程)是关键词自动提取的关键步骤之一.所谓特征定义就是定义一些尽可能较好地区分关键词和非关键词的特征.本文将现有文献提出的有效特征归结为四大类:第 1 类是候选关键词特征,针对每个候选关键词而定义,通常仅用于有监督关键词提取方法中;第 2 类是候选关键词图结构特征,通常仅用于基于图的无监督关键词提取方法中,具体包括候选关键词之间关系的特征以及词图中的中心度量特征;第 3 类是候选关键词的主题特征,是指候选关键词在文档主题上的分布,在有监督和无监督提取方法中都会用到;第 4 类是词嵌入向量(embedding)特征,利用深度学习技术学习从显式特征中学得的隐式向量.上述分类不同于已有综述文献的局部分类[11,33,43],它是基于整体视角的,具体来说,可将第 1 类认为是词图中节点

的特征;第 2 类中的词间关系特征可认为是词图中边的特征,其中的中心度量特征可认为是词图中节点的特征;第 3 类可认为是词图中节点的分类特征;第 4 类是从多显式特征中学得的隐式特征.这样分类将有助于考察不同特征间的关联关系,进而提出新的有效特征.

此外,尽管已有文献提出许多特征,然而这些特征的提出大多是基于对目标文档中局部文本信息的分析,多数情况下并没有直接指出或分析所定义的特征是从哪些方面并在多大程度上支持作为关键词需要满足的基本准则.尽管本文无法做到通过总结和归纳准确回答上述问题,但还是对此问题尝试着进行初步分析,主观结论详见表 1,以便引起研究者思考进而提出更有效的特征,并在此基础上改进现有关键词提取方法或提出新的提取方法.

Table 1 The support of different types of features for criterias that a keyphrase needs to satisfied

表 1 各类特征对关键词需满足的基本准则的支持程度

大类	细类	可读性	相关性	重要度	覆盖度	一致性
候选关键词特征	词频特征	*		***		
	长度特征	**		**		
	位置特征			**		
	语言特征	***				*
	外部知识库特征		**	**		
词图结构特征 +TextRank 方法	词间关系特征		**	***	***	
	中心度量特征			***		
主题特征	主题特征		**		***	
词嵌入向量	Embedding 向量	视学习对象而定				

注: ***为最大支持度

3.1 候选关键词特征

3.1.1 词频特征

词频相关的具体特征及其分类详见表 2,涉及每个特征的类别、编号、特征名称和计算公式.除此表列出的结构化信息之外,接下来对每个特征及特征间提出的关联关系进行简要描述.

Table 2 The features related to term frequency

表 2 词频相关特征

类别	编号	特征名称	计算公式
TF (F1)	F1.1	词频[9,14,32,45,46,49,53,54]	$TF(p,d)=f(p,d)/\|d\|$,$f(p,d)$为词 p 在文档 d 中出现的次数,$\|d\|$为文档 d 中包含的单词数目
	F1.2	头词频[44]	$HF(p,d/4)=f(p,d/4)/\|d/4\|$,$d/4$ 表示文档 d 的前四分之一,$\|d/4\|$为其包含的单词数目
	F1.3	对数词频[45]	$\log TF(p,d)=\log(TF(p,d))$
	F1.4	子词频之和[46,49,55]	$SFS(p,d)=\sum_{s\in \mathrm{sub}(p)} f(s,d)/\|d\|$,sub($p$)为词 p 中所包含的单词或短语的集合
DF (F2)	F2.1	逆文档频率[9,45,49,54]	$IDF(p,D)=\log(\|D\|/\mathrm{df}(p,D))$,df($p$,$D$)为语料库 D 中包含 p 的文档数,即文档频率
	F2.2	文档频率邻接变化度[48,56]	$DF\text{-}AV(p,n,d,D)=\sum_{t\in T_L(p,n,d)}\log(\mathrm{df}(t,D))\times\sum_{t\in T_R(p,n,d)}\log(\mathrm{df}(t,d))$, 其中,$T_L(p,n,d)$和 $T_R(p,n,d)$分别表示文档 d 中词 p 的左部分和右部分长度为 n 的内容
TF-IDF (F3)	F3.1	词频-逆文档频率[9,14,21,45–47,49,50,53,54,57–59]	$TF\text{-}IDF(p,d,D)=TF(p,d)\times IDF(p,D)$
	F3.2	对数 TF-IDF[49]	$\log TFIDF(p,d,D)=\log TF(p,d)\times\max[0,\log((\|D\|-\mathrm{df}(p,D))/\mathrm{df}(p,D))]$
	F3.3	布尔 TF-IDF[50]	$TDIDF\text{-}over(p,d,D,\theta)=\begin{cases}0, & \text{if } TFIDF(p,d,D)<\theta;\\ 1, & \text{if } TFIDF(p,d,D)\geqslant\theta\end{cases}$
	F3.4	引文 TF-IDF[50]	$CitationTFIDF(p,d,D)=\sum_{d'\in C_d} TFIDF(p,d',D)$, C_d 为相关引用文献的集合
	F3.5	上下文 TF-IDF[51]	$ContextTFIDF(p,d,D)=\sum_{p'\in s_p} TFIDF(p',d,D)$, s_p 表示词 p 所在的句子
	F3.6	TF-IDF 比值[49]	$TFIDFRatio(p,d,D)=\dfrac{TFIDF(p,d,D)}{\max\limits_{c\in comp(p)} TFIDF(c,d,D)}$, $comp(p)$为组成 p 的单词的集合
	F3.7	文档最大短语指数[49]	$DPM\text{-}index(p,d)=1-\max\limits_{s\in\sup(p,d)}\dfrac{f(s,d)}{f(p,d)}$, sup($p$,$d$)为文档 d 中包含 p 的词组的集合
	F3.8	DPM-TFIDF[49]	$DPM\text{-}TFIDF(p,d,D)=DPM\text{-}index(p,d)\times TFIDF(p,d,D)$

Table 2 The features related to term frequency (Continued)
表 2 词频相关特征(续)

类别	编号	特征名称	计算公式
其他词频相关特征(F4)	F4.1	总体 Dice 系数[30,49]	$GDC(p,d)=\lvert p\rvert\log\big(f(p,d)\big)f(p,d)\big/\sum_{c\in comp(p)}f(c,d)$, $\lvert p\rvert$为 p 所含单词数
	F4.2	最大似然估计[43,49]	$MLE(p,d)=\mathrm{pr}(p,d)=f(p,d)\big/\sum_{p'\in P_d}f(p',d)$, P_d 为文档 d 的候选关键词集合
	F4.3	KL 散度[43,49,52]	$KLD(p,d,D)=\mathrm{pr}(p,d)\log\big(\mathrm{pr}(p,d)/\mathrm{pr}(p,D)\big)$, 其中, $\mathrm{pr}(p,D)=f(p,D)\big/\sum_{p'\in P}f(p',D)$, P 为语料库 D 中所有文档的候选关键词的集合

词频(term frequency,简称 TF)是指词或短语在给定文档中出现的频率,通常认为词频越高,其在文档中的重要度越高,成为关键词的可能性越大.由词频(TF)衍生出许多相关具体特征.文档开头部分常常更有可能包含关键词,故头词频(head frequency,简称 HF)[44]这一概念被提了出来.为了增加值接近于 0 的词频的区分度,对数词频(log frequency,简称 logTF)[45]这一概念被提了出来.对于一些多元词组,常常存在子词频偏大而整体词频偏小的现象,于是文献[46]提出了子词频和(substrings frequencies sum,简称 SFS)来辅助衡量多元词组的重要性.

通常上述各词频特征仅与候选关键词所在文档有关,不能反映其与语料库的关系,为此,文献[47]引入了逆文档频率(inverse document frequency,简称 IDF)来衡量词或词组所在的文档在整个语料库中的频率.逆文档频率越大表明该词越重要.针对关键词周围多为文档频率(document frequency,简称 DF)偏高的常用词,文献[48]提出了文档频率邻接变化度(document frequency accessor variety,简称 DF-AV),认为 DF-AV 值高的词语更有可能成为关键词.

词频-逆文档频率(TF-IDF)[19]结合词频和逆文档频率来衡量候选关键词的重要度,是所有特征中最有效、最常用的特征之一.针对不同的应用场景,研究者在 TF-IDF 的基础上又提出了其他相关特征.为了提高 TF-IDF 的区分度,对数 TF-IDF(logTFIDF)[49]这一概念被提了出来.相反地,为了得到粗粒度的 TF-IDF,文献[50]通过指示函数将 TF-IDF 转化为布尔值得到布尔 TF-IDF(TFIDF-over).此外,针对科技文献的关键词提取应用,文献[50,51]分别将 TF-IDF 值的统计范围扩展到与目标科技文献有引用关系的引用文献集和引用文献中候选关键词所在的语句,为此提出引文 TF-IDF(citation TFIDF)和上下文 TF-IDF(context TFIDF).为了衡量多元候选关键词与组成其子词组之间的相对重要性,文献[49]提出了 TF-IDF 比值(TFIDF ratio)和最大短语指数(DPM-index),前者是从 TF-IDF 的角度考量,而后者是从 TF 的角度考量.此外,文献[49]还将 TF-IDF 与 DPM-index 相结合提出了 DPM-TFIDF,该特征可同时包含 TF-IDF 和 DPM-index 的实际含义.

基于这些基本特征,一些其他的词频相关特征被提了出来.Dice 系数(generalized dice coefficient,简称 GDC)[30,49]用来衡量多元词组的内部紧密度,通常关键词的 Dice 系数较高.最大似然估计(maximum likelihood estimate,简称 MLE)[43,49]用目标文档中的词频占比来估计候选关键词的相对重要性.KL 散度(Kullback-Leibler divergence,简称 KLE)[43,49,52]用来衡量候选关键词相对于整个语料集的信息量(infomativeness)大小.

3.1.2 长度特征

长度特征(length feature)是指候选关键词本身及其所在句子的长度.因关键词的长度通常小于等于 3,故候选关键词的长度具有较好的区分性.另外,通常认为句子越长,包含的信息越丰富,也即候选关键词所在句子的长度越长,其成为关键词的可能性也就越大.具体特征包括:

(1) 词长(length,表示为 L1)[9,14,32,45,46,49,53,54,58,60]指的是候选关键词所含单词的数目.

(2) 句子长度(sentence length,表示为 L2)[45,49]指的是所有包含候选关键词的句子所含的单词数目.该特征进一步派生出 3 个特征:平均句长(average sentence length,表示为 L2.1)[45,49]、最长句长(longest sentence length,表示为 L2.2)[45]和最短句长(shortest sentence length,表示为 L2.3)[45].

3.1.3 位置特征

位置特征(position feature)常用候选关键词在目标文档中出现位置的分布、跨度等指标来度量.因为关键词经常出现在文档中的一些特定的重要位置,如出现在文档的开头、段落的开头等位置的词,相对于出现在其他位置的词,更有可能成为关键词.位置特征是一类非常有效的且被广泛使用的特征.表 3 列出了候选关键词在文

档和句子中出现的位置统计特征.位置特征是基于相对位置,而非绝对位置.

除上述位置特征之外,还有一些基于上述特征的其他位置特征,具体如下.

(1) 2-means 位置(2-means position,表示为 P3.1)[49].将候选关键词在文档中每次出现的位置($pos_i(p,d)$)作为输入,使用 *k*-means 算法对候选词进行聚类,将聚类结果作为新的位置特征.

(2) 是否出现在标题中(occurrence in title,表示为 P3.2)[15,45,46].出现在标题中的候选关键词通常更有可能成为关键词.类似的特征还有是否出现在摘要中、是否出现在引言中等.

(3) 是否部分出现在标题中(occurrence of members in title,表示为 P3.3)[45,46].与前一特征类似,统计由多元词组构成的候选关键词的部分词是否出现在文档标题中.

(4) 章节位置向量(section occurrence vector,表示为P3.4)[57].该特征常针对科技文献,使用定长0/1向量表示候选关键词是否出现在科技文献的特定章节.

Table 3 The position features of candidate keyphrase in a document or sentence

表 3 候选关键词在文档和句子中的位置特征

类别	编号	特征名称	计算公式
文档位置(P1)	P1.1	首次出现位置[9,21,32,44,45,49,50,54,57–59]	$FP(p,d)=pos(p,d)/\|d\|$,$pos(p,d)$为词 p 在文档 d 中首次出现的位置(即 p 前的单词数)
	P1.2	首现句子位置[49]	$FS(p,d)=\text{sent}(p,d)/\|S_d\|$,$\text{sent}(p,d)$为第 1 个包含词 p 的句子在文档 d 中的位置(即该句前的句子数),$\|S_d\|$为文档 d 句子总数
	P1.3	平均位置[49]	$AP(p,d)=\frac{1}{n}\sum_{i=0}^{n} pos_i(p,d)\big/\|d\|$, $pos_i(p,d)$为词 p 在文档 d 中第 i 次出现的位置,假设词 p 在文档 d 中共出现 n 次
	P1.4	最后出现位置[45,46,58]	$LP(p,d)=pos_{-1}(p,d)/\|d\|$,$pos_{-1}(p,d)$为词 p 在文档 d 中最后出现的位置
	P1.5	位置跨度[14,45]	$SPAN(p,d)=LP(p,d)-FP(p,d)$
	P1.6	段落分布[61]	$PDF(p,d)=m_p\big/\max_{i\in d} m_i$, p 出现的段落数 m_p 与文中所有单词中最大出现段落数的比值
句中位置(P2)	P2.1	句中平均位置[45]	$SP(p,d)=\left(\sum_{s\in S(p,d)} pos(p,s)\big/\|s\|\right)\Big/\|S(p,d)\|$, $S(p,d)$表示文档 d 中所有包含 p 的句子的集合,$\|S(p,d)\|$表示其中句子总数
	P2.2	句中最远位置[45]	$MSP(p,d)=\max_{s\in S(p,d)}\left(pos(p,s)\big/\|s\|\right)$
	P2.3	句中最近位置[45]	$NSP(p,d)=\min_{s\in S(p,d)}\left(pos(p,s)\big/\|s\|\right)$

3.1.4 语言特征

语言特征(linguistic feature)主要是指从候选关键词的构成(如词性等)和候选关键词所在的句子的句法等方面提取的相关特征.针对词语粒度的语言特征主要包括:

(1) 词性序列(POS sequence,表示为 Li1.1)[32,57].该特征是指组成候选关键词(包括单词和词组)的单词的词性序列.因关键词的词性序列常常会遵循一定的规律(如通常为名词短语),故该特征也是非常重要的常用特征,特别是常用于基于条件随机场的关键词提取方法[15,62,63]中.

(2) 后缀序列(suffix sequence,表示为 Li1.2)[57].该特征类似于词性序列,只是将词性换为单词的后缀(如-ion、-ics、-ment),进而形成后缀序列,同样,关键词的后缀序列常常会遵循一定的规律.

(3) 是否为专有名词(proper noun,表示为 Li1.3)[45,59].有的应用场景,如人名、地名等专有名词常为关键词.

(4) 是否为特殊格式(special format,表示为 Li1.4)[15,46,53,57].特殊格式具体包括粗体、大小写、包含特殊符号、缩写等.在有的应用场景中,关键词常常标有特定的格式.

针对句子粒度的语言特征主要包括:

(1) 依存关系特征(dependency parsing,表示为 Li2.1)[63].依存关系是指句子中词语之间的主谓宾关系,根据具体应用场景可增加句子中某部分(如主语或者谓语)的权重.文献[63]将多元词组中所包含的依存关系作为特征应用在了条件随机场方法中.

(2) 修辞手法数目(number of rhetorical device,表示为 Li2.2)[64].在一些应用场景的文档中,常用一些修辞手

法(如对比、转折等)来强调或突出一些内容,通常关键词所在的句子所包含的修辞手法数较多,故该特征在一些场景中具有较好的区分性.

3.1.5 外部知识库特征

为了弥补因目标文档自身的信息不足而制约关键词提取效果这一缺陷,研究者常借助外部知识库来获取候选关键词的目标文档之外的有效特征,这些特征被称作外部知识库特征(external knowledge-based feature).外部知识主要来自维基百科、搜索引擎查询结果、WordNet[65]等.其中,基于维基百科的特征主要包括:

(1) 是否为维基词条(Wikipedia term,表示为 E1.1)[53,59].通常作为维基百科词条的词语是关键词的可能性较大.

(2) 维基百科关键词度(维基百科关键词度数据下载地址:http://www.ntu.edu.sg/home/axsun/datasets.html)(Wikipedia keyphraseness,表示为 E1.2)[58].维基页面中的链接词语可视为该页面的关键词,该特征定义为包含候选关键词且作为链接词的页面数目除以包含该候选关键词的页面总数目.

(3) 逆维基百科链接频率(inverse Wikipedia linkage,表示为 E1.3)[58].类似于 IDF 特征,用来衡量候选关键词所在的维基百科页面数目占所有页面数目的比例,其计算公式为 $IWL=-\log_2(linksTo(A_p)/N)$,其中,$linksTo(A_p)$为维基百科中指向候选关键词 p 所对应的维基百科页面 A_p 的链接数,N 为维基百科中总链接数.

除此之外,文献[60]提出了 15 个基于维基百科的特征,这里不再一一列举.基于搜索引擎查询结果、WordNet 等其他外部知识源的特征主要包括:

(1) 搜索引擎评分(score by search engine,表示为 E2.1)[53].该特征是指候选关键词作为被搜索词在搜索引擎中搜索得到的 Web 页面的数目.

(2) 词汇链特征(feature of lexical chain,表示为 E2.2)[66,67].一些特定的语义网络知识库(如 WordNet[66]、知网[67])常被用来生成表示词语间同义和上下位关系的词汇链,文献[67]将候选关键词所在的词汇链长度作为特征,其值越大,越有可能成为关键词;文献[66]则定义评分规则直接对词汇链中的候选关键词进行评分.

(3) 关键词度(keyphraseness,表示为 E2.3)[58,68].该特征为候选关键词在训练集或外部知识库中作为关键词出现的频率,即候选关键词在外部知识库中作为关键词出现的次数除以其总共的出现次数.

(4) 是否在领域字典中(dictionary lookup,表示为 E2.4)[69,70].领域词典是相关领域内的常用专业词汇的集合,而关键词很有可能是对应领域的专业词汇,故在领域词典中可查到的候选关键词更为重要.清华大学推出的开放中文词库(THUOCL(http://thuocl.thunlp.org/))包含了 IT、财经等多个领域的字典.

3.2 词图结构特征

3.2.1 词间关系特征

在候选关键单词构成的词图中,词间的关系特征主要来自于词共现和词间相似度两个方面,其主要用于无监督关键词提取方法中,特别是用于基于图的方法中.词共现特征主要包括如下两个特征.

(1) 共现次数(co-occurrence frequency,表示为 C1.1)[40]是目标文档中的候选单词对在固定长度的滑动窗口中同时出现的次数,其值越大,说明两个候选单词之间的语义关系越近.在图方法中,是最重要的特征.

(2) 引文共现次数(citation co-occurrence frequency,表示为 C1.2)[10].文献[10]将原来仅统计目标文档内候选单词对的共现关系扩展到统计与目标文档有引用关系的文档集中,进而提出该特征.

此外,词间相似度也是非常重要的关系特征,可以直观认为是对共现关系在不同评价标准下的度量.在计算词之间的相似度时,首先要给出词的向量表示,可以借助外部知识库(如维基百科[9,23,58]、搜索引擎查询结果[68])、LDA 模型抽取词的主题向量[9,13]、根据具体的词空间模型(如 Word2vec 方法[71])学习得到的词嵌入向量(embedding)[22]等;然后利用各种距离度量指标计算词间的相似度,具体度量指标包括余弦相似度[9,13]、欧氏距离[22]、点互信息(point-wise mutual information,简称 PMI)[53]、规范化谷歌距离(normalized Google distance,简称 NGD)[72]等.常用的具体词间相似度特征如下.

(1) 主题相似度(topic similarity,表示为 C2.1)[9].候选单词在主题向量下的相似度,常用余弦相似度来计算.

(2) 语义相似度(semantic relatedness,表示为 C2.2)[9,58,68].该特征可利用维基百科[9,58]或者搜索引擎查询结

果[68]来计算候选单词对在语义上的相近程度.

3.2.2 中心度量特征

针对候选关键词图,文献[73]利用图的中心度量指标(centrality measure)来提取关键词,直接根据指标值取top-10 的候选关键词作为关键词,并给出了度量指标的计算公式(见表 4)和代码(https://github.com/boudinfl/centrality_measures_ijcnlp13).

(1) 度中心性(degree).该指标是网络中刻画节点重要性最简单的指标.在关键词提取中,认为一个候选单词节点的邻居数目越多,其影响力就越大.

(2) 接近中心性(closeness).该指标通过计算节点与网络中其他所有节点的距离的平均值来衡量节点的重要性,其值越大说明该节点对信息的流动具有最佳的观察视野.在关键词提取中,一个候选单词节点与词图中其他节点的平均距离越小,说明该词节点的接近中心性就越大,其重要度越大.

(3) 介数中心性(betweenness).用来衡量网络中节点作为"桥梁"的重要程度.通常认为词图中所有节点对的最短路径中经过一个节点的最短路径数越多,这个词节点就越重要.

(4) 特征向量中心性(eigenvector).该指标认为网络中一个节点的重要性取决于其邻居节点的数量和每个邻居节点的重要性.关键词提取过程中,该值越大表明候选单词的长期影响力越大.

Table 4 The features of centrality measures

表 4 中心度量特征

编号	特征名称	计算公式
C3.1	度中心性	$C_D(w_i)=\|N(w_i)\|/(\|V\|-1)$,$V=\{w_i,w_j,\ldots\}$为图中点的集合,$\|V\|$为其中节点数,$N(w_i)$为和点 w_i 相连的点的集合,$\|N(w_i)\|$表示集合中元素个数
C3.2	接近中心性	$C_C(w_i)=(\|V\|-1)\Big/\sum_{w_j\in W} d_{\min}(w_i,w_j)$, $d_{\min}(w_i,w_j)$表示点 w_i 到 w_j 的最短路径的长度
C3.3	介数中心性	$C_B(w_i)=2\cdot\sum_{w_i\neq w_j\neq w_k\in V}\left(\sigma(w_j,w_k\mid w_i)/\sigma(w_j,w_k)\right)\Big/\left((\|V\|-1)(\|V\|-2)\right)$, $\sigma(w_j,w_k)$表示点 w_j 到 w_k 的最短路径数目,$\sigma(w_j,w_k\|w_i)$表示 w_j 到 w_k 的最短路径中经过点 w_i 的路径的数目
C3.4	特征向量中心性	$C_E(w_i)=\frac{1}{\lambda}\sum_{w_j\in N(w_i)} e(w_j,w_i)\times C_E(w_j)$, $e(w_j,w_i)$为 w_j 到 w_i 的边的权重,λ为常数

3.3 主题特征

如本文最开始部分所述,提取的关键词要与文档主题(topic,类型编号记为 T)相关且尽可能覆盖文档要表达的所有主题,由此可知,文档的主题对关键词的提取起着重要的作用,为此,刘知远博士对基于文档主题结构的关键词提取进行了专门研究(详见其博士论文[23]).在关键词提取中,常用的文档主题提取模型主要是 LDA[74](成熟的工具包括由 Python 实现的 Gensim(https://radimrehurek.com/gensim/)、Java 的 JGibbLDA(http://jgibblda.sourceforge.net/)和 C++的 PLDA(http://openbigdatagroup.github.io/plda/)),利用该模型可计算单词 w 在主题 z 下的分布 $\Pr(w|z)$以及主题 z 在文档 d 中的分布 $\Pr(z|d)$,具体应用详见文献[75].

3.4 词嵌入向量

上述的候选关键词特征、词图结构特征以及主题特征都是显式特征,随着深度学习的兴起,特别是 Word2vec[71,76]方法的提出,使得可以用深度学习技术从多个显式特征中学习得到融合统一的词嵌入向量(embeddings)[77].此外,特别是可以将词图中的候选关键单词之间的关系特征转换为每个候选关键单词的特征.2014 年 Wang 等人[22]首次将词向量引入到关键词提取中用来增强候选单词之间的语义关系;2016 年张奇等人[78]用深度递归神经网络(recurrent neural network,简称 RNN)学习关键词及其上下文信息,从而提取推文中的关键词.同年,Wang 等人[79]将候选单词的词向量输入深度信念网络(deep belief network,简称 DBN),进而计算候选单词之间的语义层级关系,然后利用词向量计算每个候选单词与其他单词之间的平均相似度,最后结合单词的语义层级度量和平均相似度进行评分.2017 年 Papagiannopoulou 等人[80]将文档中所有候选单词的词向量的平均值作为参考向量,然后计算候选单词的词向量与参考向量之间的相似度,并将相似度作为候选单词的评分.

这方面的研究才刚刚展开,随着面向文本的深度学习技术研究的深入,利用词嵌入向量特征的关键词提取方法将会得到更多的关注.

4 关键词提取方法

关键词提取方法主要可分为有监督和无监督两大类,有关利用半监督方法提取关键词的研究较少.本节以方法中提出和使用的特征以及方法在时间轴上的演进为线索,详细归纳和总结了现有的研究工作,具体包括方法所属类型、使用的特征(因多数工作使用的特征较多,故仅列出了特征类别)、对比方法以及实验用的数据集.

4.1 有监督方法

有监督方法是被广泛应用的一类提取方法,其具有提取效果良好的优点,但也存在过拟合及标注成本高的明显缺点.该类方法的基本步骤是:通过文本预处理生成候选关键词(单词和词组);定义并计算特征;使用特定的分类器来提取关键词.有监督提取方法通常将关键词提取作为二元分类问题利用传统分类器来处理.具体步骤如下.

- 将待提取关键词的文档集分为训练集和测试集,将训练集中的候选关键词标注为关键词(即正样本,用 1 来表示)或非关键词(即负样本,用 0 来表示),每个候选关键词的类型标签记为 $Y_i(Y_i\in\{0,1\})$.
- 计算每个候选关键词的特征集 $X_i(X_i\in R^N)$;然后在训练集$\{(X'_i,Y'_i)\}$上学习一个分类函数 $F:X\to Y$.
- 最后用学习得到的函数 F 对测试集中的候选关键词进行分类,即将其分为关键词和非关键词.

为了便于读者阅读,人为地将有监督提取方法研究工作分为两个阶段:初期阶段和扩展阶段.在研究初期阶段,一些较为直观的特征被使用,且以不断地尝试各种经典分类器为主(主要采用的分类工具为 Weka(http://www.cs.waikato.ac.nz/ml/weka/));在研究扩展阶段,以细化已有特征和提出新的特征为主,并尝试少数其他分类方法.表 5 具体列举了大部分利用有监督方法提取关键词的文献,这些工作之间的差别主要集中于使用的特征和分类方法上.

Table 5 The list of summaries of supervised keyphrase extraction approaches

表 5 有监督关键词提取方法总结列表

分类	#	提取方法	特征数目	主要特征						对比方法	数据集
				Fre.(F)	Len.(L)	Pos.(P)	Lin.(Li)	Ext.(E)	Others		
朴素贝叶斯	1	KEA,1999[21]	2	**1:F3.1**		1:P1.1				#9	J,W
	2	Turney,2003[68]	13	1:F3.1		1:P1.1		**1**:**E2.3**	**10**:-	#1	J
	3	KEA++,2006[83]	4	1:F3.1	1:L1	1:P1.1			**1**:**C2.2**	#1	W1
	4	Nguyen,2007[57]	6	1:F3.1		**2**:**P3.4**	**3**:**Li2.2**			#1	J2
	5	Maui,2009[58]	9	1:F3.1	1:L1	2:P1.5		**3**:**E2.3**	**2**:**C2.2**	#1	J7
	6	CeKE,2014[50]	9	3:F3.1		**5**:**P1.1**	1:Li1.1			#1,12	J9
	7	CeKE,2015[84]	10	3:F3.1		5:P1.1	1:Li1.1	1:E2.3		#1,5	J9
	8	KeyEx,2017[36]	7	1:F3.1		1:P1.1	**5**:**Li2.4**			#1	N4,J
遗传算法	9	Turney,1999[20]	12	**3**:**F1.1**	2:L1	**2**:**P1.1**	5:LI1.1			DT	J,E
	10	Joorabchi,2013[60]	20	1:F1.1	1:L1	3:P1.1		**15**:-		#3,5	J8
	11	Liu,2015[69]	–	F1.1	L1		LI1.1	E2.4	C1.1		M
决策树	12	Hulth,2003[52]	4	2:F1.1		1:P1.1	**1**:**Li2.1**				J1
	13	Hulth,2004[82]	4	2:F1.1		1:P1.1	1:Li2.1			#12	J1
	14	Ercan,2007[66]	7	1:F1.1		2:P1.1		**4**:**E2.2**		#1	J
	15	Krapivin,2010[85]	20	2:F1.1	1:-	2:P3.2	**15**:**Li1.1**			NB	J3
	16	John,2016[54]	9	2:F3.1	1:L1	2:P1.1	1:Li1.2	**1:-**	**2**:**T**		J5
	17	Sterckx,2016[14]	9	2:F3.1	2:L1	2:P1.1	2:Li1.1		**1**:**T**		N
支持向量机	18	Wang,2005[61]	4	2:F1.1		**2**:**P1.6**					W
	19	Zhang,2006[51]	8	**2**:**F3.5**		3:P3.2	1:Li1.1		**2**:-	#1	J
	20	Jiang,2009[86]	–	2:F1.1	1:L1	3:P1.1				#1	J,W
	21	KeyWE,2010[53]	8	2:F3.1	1:L1	1:P3.2	1:Li1.3	2:E1.1	**1**:**C2.2**		J5
	22	Chen,2016[87]	10	1:F3.1	1:L1	1:P1.3	2:Li1.3	**5**:-		#5	J5
逻辑回归	23	Haddoud,2014[49]	18	**11**:**F3.1**	2:L1	**5**:**P3.1**				DT	J5
	24	Haddoud,2015[44]	18	11:F3.1	2:L1	5:P3.1					J5
神经网络	25	Sarkar,2012[88]	6	**3**:**F1.1**	**2**:**L1**	**1**:-				#1	J
	26	AE*,2015[45]	20	6:F3.1	**4**:**L1**	**8**:**P1.1**	1:Li1.2	1:E2.3		#5	J
	27	Zhang,2016[78]	1					**Embeddings**			T

Table 5 The list of summaries of supervised keyphrase extraction approaches (Continued)

表 5 有监督关键词提取方法总结列表(续)

分类	#	提取方法	特征数目	主要特征						对比方法	数据集
				Fre.(F)	Len.(L)	Pos.(P)	Lin.(Li)	Ext.(E)	Others		
条件随机场	28	Zhang,2008[62]	22	1:F3.1	1:L1	10:P3.2	1:Li1.1		9:-	SVM	J
	29	Bhaskr,2012[63]	11	1:F1.1		4:P3.2	**5**:**Li2.5**		1:-		J5
	30	Gollapalli,2017[15]	–	-:-	-:-	-:P3.2	5:Li1.1		**1**:-	#1,5,6	J9
整数规划	31	Ding,2011[24]	–	-:F3.1		-:P1.1			**2**:**C2.2**		N
	32	Boudin,2015[89]	–	-:F3.1		-:P1.1			-:C1.1		J5

注:(1) 数据集包括论文(Jounal,J)、邮件(Email,E)、页面(Web,W)、新闻(News,N)、推特(Twitter,T)、医学诊断报告(Medical report,M);(2) 加粗表示该类有新特征提出;(3) 对比方法中的编号对应提取方法的编号或具体的分类方法;(4) 数据集中的编号对应(1)中的类型或表 6 中的编号;(5) 主要特征优先填入常用或新特征,“–”表示不明确

4.1.1 初期阶段

1999 年,Turney[20]最先采用有监督分类方法提取关键词,分别将词频、相对位置等特征输入到遗传算法(genetic algorithm,简称 GA[81])和 C4.5 决策树两个分类器进行关键词抽取,实验结果表明,遗传算法取得较好的效果.同年,Frank 等人[21]针对 GA 分类器训练时间长的问题尝试采用贝叶斯分类器(Naïve Bayes,简称 NB)提取关键词,其方法 KEA(http://www.nzdl.org/Kea/)仅用了 TF-IDF 和相对位置两个特征,提取效果与 GA 分类器在不同参数设置下互有高低.2003 年 Turney[68]为了使所提取的关键词具有语义一致性,增加了统计关联度特征(如 PMI),使用了 NB 分类器,并对比了 4 个不同类型的数据集.同年,Hulth[32]将语言特征(见第 3.1.4 节)输入 Bagged 决策树分类器来提取关键词,取得了不错的提取效果.2004 年,同一作者 Hulth[82]对上述工作进行了优化,主要是通过尝试不同的特征组合来过滤掉非关键词,从而提高提取效果.2005 年 Wang 等人[61]尝试采用支持向量机(support vector machine,简称 SVM)来抽取 Web 页面中的关键词,主要使用了频率和位置两类 4 个特征.2006 年 Zhang 等人[51]同样利用 SVM 分类器来抽取关键词,较前一篇文献使用了更多的特征,如 POS、ContextTFIDF 等.

4.1.2 扩展阶段

(1) 朴素贝叶斯

贝叶斯方法是使用最广泛的有监督提取方法,因其具有实现简单且提取效果良好的特性[83].自 1999 年 KEA[21]被提出以来,研究者不断地尝试使用不同的特征来提升 NB 方法的提取效果.KEA++[83]利用特定领域的同义词语料集(Agrovoc(www.fao.org/agrovoc))计算候选关键词的语义关系特征,然后结合 KEA 的特征集利用 NB 方法提取关键词.文献[57]添加了语言特征,如单词的 POS 标签(Li1.1)、英文词语的后缀(Li1.2)特征等.Maui[58]在文献[57]的基础上进一步扩展特征,特别是整合一些从维基百科提取的外部特征,然后采用了 NB 方法和 Bagged 决策树,实验结果表明,未引入外部特征的情况下,NB 方法更优;而在加入新的外部数据特征后,Bagged 决策树方法取得了更好的结果.CeKE[50]引入了引用网络文献作为扩展的特征提取源,共提取了 3 类 9 个特征.2015 年,文献[84]在 CeKE 的基础上引入了关键词度(E2.3)特征.2017 年,文献[36]首先使用序列模式挖掘算法提升候选关键词生成效果,然后用不同的分类算法提取关键词,NB 方法的提取效果最佳.

(2) 遗传算法

自 1999 年利用遗传算法提取关键词以来,2013 年,文献[60]从维基百科中抽取一些外部知识库特征,结合其他共计 20 个特征用 GA 方法提取关键词.2015 年,文献[69]提取医学诊断报告中的关键词,首先用 3 个无监督方法 PrefixSpan[90]、C-Value[91]以及 TextRank[40]分别从词序列模式、统计语言和词共现图中节点的重要性 3 个方面对候选关键词进行打分,然后利用 GA 算法学习上述 3 种无监督方法得到的评分的线性权重.

(3) 决策树(decision tree,简称 DT)

自 2003 年利用决策树提取关键词以来,2007 年,文献[66]利用 WordNet 生成词汇链并定义链中候选关键词的分数特征,然后结合位置特征,采用 Bagged 决策树提取关键词.2016 年,文献[14]为了使关键词提取方法适用于不同用户标注的数据集,使用 Boosted 决策树提取关键词.2010 年,文献[85]利用随机森林分类器(random forest,简称 RF)来提取科技论文中的关键词,相对于 SVM 等方法取得了不错的效果.2016 年,文献[54]同样使用 RF 分类器来提取关键词.

(4) 支持向量机(support vector machine,简称 SVM)

该方法自 2005 年被文献[61]使用以来,2009 年,文献[86]认为二分类方法中依赖于候选关键词的特征将其归为关键词或非关键词具有一定的局限性,而排序方法更适合该问题,鉴于此,作者提出线性排序 SVM 方法[92].2010 年,文献[53]使用 SVM 分类器对 SemEval 2010/Task-5 评测任务提取关键词.2016 年,陈忆群等人[87]提出了针对专利文档的关键词提取方法,利用专利数据集构建了 5 个背景知识库,并在每个知识库上定义一个相关特征,结合另外已提出的 5 个常用特征利用 SVM 分类器提取关键词.

(5) 逻辑回归(logistic regression,简称 LR)

文献[44,49]采用逻辑回归分类器来提取关键词,在分类之前对特征进行离散化和二值化预处理,实验结果表明,相对于其他分类器取得了不错的提取效果.相对其他方法,使用 LR 分类器提取关键词的工作较少.

(6) 神经网络(neural network,简称 NN)

2012 年,文献[88]从候选关键词出现的词频、位置、长度等方面定义了 6 个特征,实验对比了多层神经网络算法(MPL)、C4.5 和 NB 算法,MPL 取得较好的实验效果.2015 年,文献[45]直接利用 NN 对西班牙文献进行关键词提取.2016 年,文献[78]针对社交短文本信息提出一个深度递归神经网络(RNN)来提取关键词,该方法包括两个隐藏层:第 1 层用来提取关键词信息;第 2 层基于关键词信息采用序列标注方法来提取关键词.

(7) 条件随机场(conditional random field,简称 CRF)

2008 年,文献[62]首先用 CRF 针对中文文档提取关键词,使用到词频、位置、POS 词性标签等特征,并与 SVM、多变量线性回归(MLR)等方法进行比对,实验结果表明,CRF 优于其他方法.2012 年,文献[63]除了使用词频、位置特征、词性等特征外,还加入描述词语的主谓宾关系的语法成分特征(Li2.1)和描述文本序列是否为词组的分块(chunking)特征,尝试了多种特征组合,得到了适用于 SemEval 2010 数据集的最佳特征组合.2017 年,文献[15]进一步拓展了特征范围,引入了专家知识特征.专家知识是指其他成熟关键词提取算法的结果以及引文特征,利用这些信息作为弱监督信息设置标签分布偏好,进而提高 CRF 的标注结果.

(8) 二元整数规划(binary integer programming,简称 BIP)

文献[24,89]采用二元整数规划方法提取关键词,作者认为,一些现有方法是根据排名分数直接将分数高的候选关键词作为关键词,如根据 TF-IDF 指标将排名靠前的作为关键词,这些方法的局限性在于没有从全局出发进行推断,为此,提出基于全局推断的整数规划关键词提取方法.该方法的基本思想是首先利用有监督或无监督关键词提取方法计算候选关键词的权重,然后在此基础上进行整数规划.文献[24]将 TF-IDF 值作为权重,文献[89]中的权重是 TF-IDF、TextRank 和 LR 方法求得的排名指标的均值.该方法的不足也很明确,即权重的可扩展性差,如文献[89]将平均值作为权重,故有待进一步完善.

4.2 基于图的提取方法

基于图的关键词提取方法是目前最有效的、被广泛研究的一类无监督提取方法.因该方法考虑了文档中词和词的共现关系(两个词存在共现关系,表明它们在该文档中存在语义关联)且可以融合更多的其他特征信息,因此取得了较好的提取效果,通常优于其他无监督方法,且在一些情况下接近有监督方法[16,22].鉴于此,该方法一直以来得到了研究者的广泛关注,从 2004 年 Mihalcea 等人[40]最初提出的 TextRank 方法至 2017 年 Florescu 等人[16]提出的 PositionRank,十几年来研究者相继提出了许多具体相关方法.接下来,我们对这些方法进行详细总结,便于未来研究者在此基础上提出更有效的方法.

图方法的关键词提取主要步骤如下.

- 对文本进行预处理生成单词.通常是仅保留名词和形容词,构成词集 $V=\{w_1,w_2,\ldots,w_N\}$.
- 根据文档中单词在滑动窗口中的共现关系,构建文档对应的单词图 $G=(V,E)$,其中,E 是单词间的共现关系集合,节点 w_i 到 w_j 的边常表示为(w_i,w_j)或 $w_i\rightarrow w_j$.除用共现关系直接建词图之外,还有其他建图方式.如 2017 年 Tang 等人[8]根据单词间的关联度大小对最初的共现关系进行筛选,只保留大于给定阈值的共现关系.同年,Shi 等人[93]结合知识图谱(knowledge graph)构建候选单词之间的语义关联词图.
- 使用各种排序方法对单词进行评分,也即计算分数 $R(w_i)$.

• 根据评分生成多元词组(*n*-grams)取其中 top-*n* 作为关键词.

在基于图的提取方法中,通常采用的两类评分指标为:PageRank 分数(基于 PageRank 算法[94])和其他中心度量指标(详见第 3.2.2 节).PageRank 算法最早提出是为了利用图链接结构来递归地计算图中各节点的重要性,即模拟用户通过点击链接随机访问图中节点的行为(随机游走模型)计算稳定状态下各节点得到的随机访问概率.对于图中的任一节点 w_i,其 PageRank 分数 $R(w_i)$计算如下:

$$R(w_i)=\lambda\sum_{j:w_j\to w_i}\frac{1}{O(w_j)}\cdot R(w_j)+(1-\lambda)\cdot r(w_i) \tag{1}$$

其中,λ为阻尼系数,通常取值为 0.85;$O(w_j)$表示节点 w_j 的出度;$r(w_i)$为重启概率,通常取值为 $1/N$,N 为图中节点的总数.到目前为止,研究者基于 PageRank 算法提出了一系列关键词提取方法,下面我们对相关的主要方法进行逐一的总结,围绕具体问题的发现及其改进展开介绍,这样有助于深入了解现有方法乃至提出全新的方法.

4.2.1 TextRank 方法

受 PageRank 算法的启发,2004 年基于 PageRank 算法的 TextRank 方法[40]首先被提了出来.相对于 PageRank 算法,该方法修改了节点间边的转移概率,具体而言,用边的共现次数取代了出度.

在目标文档 d 对应的词图中,设任意一条边的权重 $e(w_i,w_j)=freq_d(w_i,w_j)$,其中,$freq_d(w_i,w_j)$表示边的$(w_i,w_j)$在 d 中的共现次数,与此相对应,$O(w_j)=\sum_{w'\to w_j}e(w_j,w')$,于是节点 w_i 的 PageRank 分数 $R(w_i)$计算如下:

$$R(w_i)=\lambda\sum_{j:w_j\to w_i}\frac{e(w_j,w_i)}{O(w_j)}\cdot R(w_j)+(1-\lambda)\cdot r(w_i) \tag{2}$$

4.2.2 ExpandRank 方法

在 TextRank 方法中,词图仅利用目标文档自身而建立,针对单目标文档对应的词图信息不够丰富这一局限,2008 年,北京大学的万小军等人[95]提出了 ExpandRank 方法.该方法利用与目标文档相近的文档集来辅助构建词图,显然,这种方式扩充了词图的信息量,从而使得提取效果得到显著提升.相对于 TextRank 方法,ExpandRank 方法主要对候选关键词之间的权重 $e(w_i,w_j)$进行了调整.

给定目标文档 d 和文档集 D,首先通过相似计算从文档集 D 中找出与目标文档 d 相近的子文档集 D_d,然后利用目标文档 d 与其相近文档集合 D_d 建立词图.于是,边权重 $e(w_i,w_j)$的计算公式如下:

$$e(w_i,w_j)=\sum_{d_l\in D_d}\cos(d,d_l)\cdot freq_{d_l}(w_i,w_j) \tag{3}$$

其中,$\cos(d,d_l)$为文档 d 和 d_l 之间的余弦相似度,$freq_{d_l}(w_i,w_j)$表示边(w_i,w_j)在文档 d_l 中的共现次数.

4.2.3 CiteTextRank 方法

采用与 ExpandRank 方法相似的技术路线,2014 年 Gollapalli 等人[10]提出了 CiteTextRank 方法.前者是用与目标文档相近的文档集来增强词图的信息量,而后者是用科技文献中的引用上下文(包括引用文献和被引用文献)来增强词图的信息量.边权重 $e(w_i,w_j)$的计算公式修改为

$$e(w_i,w_j)=\sum_{k\in T}\sum_{c\in C_k}\gamma_k\cdot\cos(d,d_c)\cdot freq_c(w_i,w_j) \tag{4}$$

其中,T 表示关系类型集合,具体包括引用、被引用和自身;C_k 表示属于某个 k 类型的文档的集合,如若 k 赋值为引用类型,则 C_k 表示目标文档 d 引用的文档集;参数γ_k用来调节不同类型文档的权重.

4.2.4 WordAttractionRank 方法

为了增强候选单词之间的语义关系,2014 年 Wang 等人[22]首次将词向量引入到关键词提取中,使用 SENNA[96]方法(http://ml.nec-labs.com/senna/)在维基百科数据集上计算词向量.于是,反映候选单词间语义相关性(原文中称其为词间吸引力分数,attraction score)的边权重 $e(w_i,w_j)$定义如下:

$$e(w_i,w_j)=\frac{2\cdot freq(w_i,w_j)}{freq(w_i)+freq(w_j)}\cdot\frac{freq(w_i)\cdot freq(w_j)}{\left(d(w_i,w_j)\right)^2} \tag{5}$$

其中,$freq(w_i)$表示 w_i 在文档中出现的次数,$d(w_i,w_j)$为 w_i 和 w_j 的词向量间的欧氏距离.

4.2.5 TopicalPageRank(TPR)方法

正如文献[75]所述,关键词应该反映文档的主题.为此,清华大学的刘知远将主题敏感的 PageRank(topic-sensitive PageRank)[97]方法用于关键词提取中,于 2010 年提出了 TopicalPageRank 方法.TPR 方法的基本思想是为每个隐含主题单独地运行带偏好的 PageRank 算法.每个主题相关的 PageRank 都会偏好那些与该主题有较大语义相关度的单词.具体算法如下.

- 首先,采用 LDA 主题模型计算单词 w 在主题 z 下的分布 $\Pr(w|z)$以及主题 z 在文档 d 中的分布 $\Pr(z|d)$.
- 然后,计算每个单词 w_i 在主题 z 下的 PageRank 分数 $R_z(w_i)$,具体计算公式如下:

$$R_z(w_i)=\lambda\sum_{j:w_j\to w_i}\frac{e(w_j,w_i)}{O(w_j)}\cdot R_z(w_j)+(1-\lambda)\cdot \Pr_z(w_i) \tag{6}$$

其中,$\Pr_z(w)$表示单词 w 在主题 z 下的重启概率,文献[75]具体定义了 3 种计算方法.

(1) $\Pr_z(w)=\Pr(w|z)$;

(2) $\Pr_z(w)=\Pr(z|w)$;

(3) $\Pr_z(w)=\Pr(w|z)\times\Pr(z|w)$.

另外,边权重 $e(w_j,w_i)$与 TextRank 中的相同.

- 最后,计算每个单词 w_i 在所有主题下的 PageRank 分数 $R(w_i)$,计算公式如下:

$$R(w_i)=\sum_{z=1}^{k}R_z(w_i)\cdot\Pr(z\mid d) \tag{7}$$

4.2.6 Single-TPR 方法

2015 年 Sterckx 等人[13]针对 TPR 方法计算量大的不足,提出了 single-TPR 方法.该方法将 TPR 方法的每个主题对每个单词的个性化影响合并为所有主题对单词的影响.于是,相对于 TPR 方法,该方法极大地降低了计算量且关键词的提取效果与 TPR 方法接近.具体而言,将所有主题对每个单词的影响体现到重启概率中,设单词 w 在主题集 $Z(|Z|=K)$下的分布向量为 $\mathbf{Pr}(w|Z)=(\Pr(w|z_1),\ldots,\Pr(w|z_K))$,文档 d 中主题的分布向量为 $\mathbf{Pr}(Z|d)=(\Pr(z_1|d),\ldots,\Pr(z_K|d))$,于是,重启概率 $r(w_i)$按照如下公式计算:

$$r(w_i)=\frac{P_z(w_i)}{\sum_{w\in V}P(w)},\quad P_z(w_i)=\frac{\mathbf{Pr}(w_i\mid Z)\cdot\mathbf{Pr}(Z\mid d)}{\left\|\mathbf{Pr}(w_i\mid Z)\right\|\cdot\left\|\mathbf{Pr}(Z\mid d)\right\|} \tag{8}$$

最后将 $r(w_i)$代入 TextRank 方法中的公式(2),计算单词 w_i 的 PageRank 分数 $R(w_i)$.

4.2.7 SalienceRank 方法

2017 年 Teneva 等人[98]提出了降低 TPR 方法所需大量计算的新方法,该方法与 single-TPR 相同,只在主题上进行一次 PageRank 运算,二者的不同之处在于每个候选单词在所有主题下的重启概率有所不同,该方法的主要意图是增大候选单词在不同主题下的区分度,其重启概率计算公式如下:

$$\left.\begin{aligned}&r(w_i)=(1-\alpha)\cdot CF(w_i)+\alpha\cdot TS(w_i),\\&CF(w_i)=\Pr(w_i\mid D),TS(w_i)=KL(\Pr(z\mid w_i)\|\Pr(z))=\sum_{z\in Z}\Pr(z\mid w_i)\cdot\log\frac{\Pr(z\mid w_i)}{\Pr(z)}\end{aligned}\right\} \tag{9}$$

其中,$CF(w_i)$为单词 w_i 在语料集 D 中出现的频率,$TS(w_i)$为单词 w_i 在所有主题上相对区分度之和,$\Pr(z)$为随机选取一词在主题 z 下的概率.最后将 $r(w_i)$代入 TextRank 方法中的公式(2),计算单词 w_i 的 PageRank 分数 $R(w_i)$.

4.2.8 TopicRank 方法

为了利用主题信息,2013 年 TopicRank 方法[99]被提了出来,该方法没有采用 TPR 和 single-TPR 方法用到的 LDA 主题模型对候选单词进行主题划分,而是采用层次聚类方法对候选单词进行分类(类似主题划分).随后,将划分后的每一类候选单词集作为一个词图中的一个顶点,构成完全图,边权重 $e(w_j,w_i)$的具体计算公式如下:

$$e(w_i,w_j)=\sum_{w_i\in c_i}\sum_{w_j\in c_j}d(w_i,w_j),\quad d(w_i,w_j)=\sum_{p_i\in pos(w_i)}\sum_{p_j\in pos(w_j)}\frac{1}{\left|p_i-p_j\right|} \tag{10}$$

其中,c 表示类别集,$pos(w)$表示候选词 w 的位置.之后使用 PageRank 算法对每个类别进行评分.最后从排名靠前

的类别中选出关键单词,选择策略分别为:选出现位置靠前的单词;选出现频率高的单词;选出聚类中心.

4.2.9 TSAKE 方法

同样,为了利用主题信息,2017 年 Rafiei 等人[26]提出了新的词图和主题相结合的方法,首先在候选单词共现图上建立不同主题的主题图,图中的边权重计算如下:

$$e_z(w_i,w_j)=\alpha\cdot \Pr_z(w_j\mid w_i)+(1-\alpha)\cdot e(w_i,w_j) \tag{11}$$

然后利用社交网络中的社区发现技术获得主题下每个子图(子社区),并用中心度量指标计算子图的中心词,接着计算主题下每个候选单词的分数 $topic\_socre(w_i,z)$,最后计算候选单词在所有主题下的分数,计算公式为 $R(w_i)=\sum_{z\in Z}\Pr(z|d)\cdot topic\_score(w_i,z)$,并据此分数对候选单词进行排名.该方法与 TPR 方法一样,需要计算单词在每个主题下的分数,因此其计算量也很大.

4.2.10 PositionRank 方法

2017 年 PositionRank 方法[16]被提了出来,该方法将候选单词在文档中每次出现的位置加入到 PageRank 模型中,与 single-TPR 方法类似,均是对重启概率 $r(w_i)$进行了修改,计算公式如下:

$$r(w_i)=\sum_k \frac{1}{pos_k(w_i,d)} \tag{12}$$

其中,$pos_k(w_i,d)$为候选单词 w_i 在文档 d 中第 k 次出现的位置.

4.2.11 SEAFARER 方法

上述基于图的提取方法,既有修改转移概率的,如 TextRank、CiteTextRank 等方法根据实际情况修改了边权重 $e(w_j,w_i)$;也有修改重启概率 $r(w_i)$的,如 single-TPR、PositionRank 等方法.特别是在 CiteTextRank 方法中,边权重 $e(w_j,w_i)$针对具体的 3 个权重进行了线性组合,权重系数参数 γ_k 的值是事先给定的经验值.鉴于此,2013 年 Zhang 等人[9]将词图中边的多个权重(如在主题下的语义相似度、在维基百科知识库下的语义相似度等)和候选词节点的多个权重(如 TF-IDF、出现位置、关键词长等)融合到 PageRank 框架中,然后以最小化 AUC 为目标函数学习边和节点权重的线性组合的系数参数.具体而言,设词图中边(w_i,w_j)对应的边特征向量为 x_{ij},相应的系数向量为ω;词图中节点 w_i 对应的边特征向量为 y_i,相应的系数向量为ψ_i,于是边权重 $e(w_j,w_i)$和重启概率 $r(w_i)$分别按如下公式计算:

$$e(w_i,w_j)=\frac{1}{1+exp(-\omega^T x_{ij})},\ r(w_i)==\frac{1}{1+exp(-\psi^T y_i)} \tag{13}$$

最小化目标函数 AUC 定义为

$$AUC=\frac{\sum_{k\in T_p}\sum_{l\in T_n} I\left(R(w_k)-R(w_l)\right)}{\left|T_p\right|\left|T_n\right|} \tag{14}$$

其中,T_p 和 T_n 分别为正、负样本集.$I(.)$是指示函数,如果 $R(w_k)>R(w_l)$,则 $I(R(w_k)-R(w_l))=1$;否则,相反.

4.2.12 MIKE 方法

2017 年 Zhang 等人[100]提出了更加泛化的多维异质信息融合的提取方法.较 SEAFARER 方法而言,MIKE 方法除了可以将词图中的多个边权重和节点权重融合到 PageRank 框架中以外,还可以融合候选单词的主题信息以及候选单词之间的重要性关系信息,对应的边权重 $e(w_j,w_i)$和重启概率 $r_z(w_i)$分别按如下公式计算:

$$e(w_i,w_j)=\frac{\sum_k \omega_k a_{ijk}^E}{\sum_j\sum_k \omega_k a_{ijk}^E},\ r_z(w_i)=\frac{P_z(w_i)\sum_k \varphi_k a_{ik}^V}{\sum_i\left(P(w_i)\sum_k \varphi_k a_{ik}^V\right)} \tag{15}$$

最小化目标函数定义为

$$\min_{\boldsymbol{\omega}\geqslant 0;\boldsymbol{\varphi}\geqslant 0;\boldsymbol{R}\geqslant 0} \alpha\left\|\lambda \boldsymbol{MR}+(1-\lambda)\boldsymbol{r}_z\right\|^2+(1-\alpha)\boldsymbol{\mu}\left(\boldsymbol{1}-\boldsymbol{BR}\right) \tag{16}$$

其中,$P_z(w_i)$为主题权重,见公式(8);$\boldsymbol{M}$ 为词图的转移矩阵,由 $e(w_j,w_i)$计算得到;$\boldsymbol{R}$ 为候选单词评分向量;$\boldsymbol{r}_z$ 为主题下的重启概率向量;$\boldsymbol{B}$ 为候选单词之间相对重要性关系矩阵;向量 $\boldsymbol{\mu}$ 调节 $\boldsymbol{B}$ 矩阵的权重;$\boldsymbol{1}$ 是元素值全为 1 的向量.

4.2.13 中心度量方法

2013年Boudin[73]提出了中心度量方法.该方法以TextRank方法所提出的词图为基础,采用了多种中心度量指标(详见第3.2.2节)来计算单词节点的分数,然后直接生成多元词组,取其中top-*n*作为关键词.

4.2.14 基于多图的方法

上述方法均是基于单一的单词图的,还有一些方法是基于多图的方法.2007年万小军等人[101]分别建立了候选单词图、单词所在的句图、词和句之间的联系图,然后使用异构网络上的相互增强的PageRank算法[102]对候选单词进行评分.2017年Yan等人[103]在上述工作的基础上使用聚类方法对候选单词进行分类,其目的类似于主题模型希望提取的关键词兼顾到每个类别或主题.同年,Shi等人[93]引入知识图谱并将词图中的候选单词与知识图谱中的实体(entity)进行了关联,通过这种方式扩展了候选单词图,从而提升了关键词的提取效果.

4.2.15 基于超图(hypergraph)的方法

2014年Bellaachia等人[104]对社交网络中的推文提取关键词,因推文的主题随时间而演化,作者认为,静态文档(主题不随时间而演进)设计的传统提取方法很难胜任提取推文中的关键词,为了考虑时间演化信息和社会属性信息(如流行度)对关键词的影响,提出了基于超图的关键词提取方法.在超图中,节点依然是候选单词,超边是短文本包含的候选单词.超图建立后,在其上具体定义转移概率和重启概率,然后使用PageRank算法框架计算每个候选单词的分数.因该方法与前面的方法共用的模型细节较少且模型细节较为繁杂,鉴于此,不再介绍具体模型,模型详见原文[104,105].

以上基本总结了近年来提出的基于图的关键词提取方法,上述方法基本上是从以下几个方面展开的:(1) 通过新补充语料集从而扩展候选单词图,如补充相似文档[95]、引用文档[10];(2) 赋予词图中的边更多的语义信息,如WordAttractionRank方法[22]首先在维基百科数据集上学习候选单词的表示向量,然后计算单词间的语义相关性,又如DBRank方法[93]利用知识图谱辅助建立语义关联图,从而使单词之间的边携带语义信息.基于以上思路的方法大都要修改PageRank计算框架中的转移概率;(3) 增加候选单词的主题信息和位置信息,分别如TPR方法[75]增加了主题信息、PositionRank方法[16]增加了位置信息.基于该思路的方法大都要修改PageRank计算框架中的重启概率;(4) 融合多维异质信息到一个统一的模型,如方法SEAFARER[9]和MIKE[100].基于该思路的方法大都要同时修改PageRank计算框架中的转移概率和重启概率,并且需要提出相应的参数学习模型学习相关的参数;(5) 在多个异质图上利用多PageRank算法进行相互增强计算,从而提升候选单词评分的准确度,如同时包括词图、句图[101].此外,针对一些关键词可能会是词图中出现次数和连接边少的单词节点,Figueroa等人[25]提出了基于错误反馈传播(error-feedback propagation)的图提取方法.

4.3 其他无监督方法

4.3.1 TF-IDF提取方法

TF-IDF方法是最朴素的无监督提取方法.尽管方法简单,但其提取效果较好[43],因此成为后来提出的大多数关键词提取方法的对比实验的基准,尤其是在无监督提取方法中.该方法通常的提取步骤是:(1) 计算候选关键词的TF-IDF值;(2) 生成*n*-grams词组并计算TF-IDF值的加和;(3) 根据TF-IDF的加和取top-*n*作为关键词.

4.3.2 基于主题的聚类提取方法

2009年Grineva等人[106]提出了CommunityCluster方法,该方法只从最重要的一个聚类主题中选取关键词.实验结果表明,该方法与TextRank、TF-IDF等方法相比,在保证准确率的同时,具有更好的召回率.

同年,清华大学的刘知远等人[72]提出了KeyCluster方法.该方法不同于前者仅从最重要的一个主题中选取关键词,而是从所有聚类主题中选取关键词.首先使用词共现次数和词间语义相似度对候选单词进行聚类;然后选取每个聚类的中心词作为种子词;最后生成包含种子单词的名词短语,并将其作为关键词.因该方法是从所有的聚类主题中选取关键词,故其提取结果能够较好地满足相关性和覆盖度准则.

5 数据集与实验评估

5.1 数据集

对于待提取关键词的文档集,需要关注以下几个统计指标,因这些指标会影响到各种提取方法的参数设置.

(1) 文档集中文档数目(#documents);

(2) 每个文档中关键词的数目(#gold keyphrases/doc);

(3) 文档集中关键词总数(#gold keyphrases);

(4) 文档集中组成关键词的 *n*-grams 的数目(#uni-grams,#bi-grams,#tri-grams,#>tri-grams).

此外,其他一些统计指标也可影响各种提取方法的参数设置.

(1) 每个文档中 Tokens 数目(#tokens/doc);

(2) 每个文档中候选关键单词数目(#candidate words/doc);

(3) 每个候选关键词中 Tokens 数目(#tokens/candidate phrase);

(4) 每个关键词中 Tokens 数目(#tokens/gold keyphrase).

为了方便研究者获取数据集,本文整理了现有研究工作常用到的公开数据集及其 URL 链接,详见表 6.

Table 6 The datasets

表 6 数据集列表

分类	编号	数据集	简述	URL
新闻网页	N1	Wan 2008[11,46]	308 篇 DUC-2001	https://github.com/tapilab/is-karthikbmk/tree/master/data/DUC2001
	N2	Marujo 2012[64]	450 篇分类新闻	https://github.com/snkim/AutomaticKeyphraseExtraction/blob/master/500N-KPCrowd-v1.1.zip
	N3	WikiNews 2012[99]	100 篇 WikiNews	https://github.com/adrien-bougouin/WikinewsKeyphraseCorpus
	N4	Reuters-21578	21 578 篇 Newswire	http://www.daviddlewis.com/resources/testcollections/reuters21578/
	W1	FAO[83,107]	联合国粮、农文档	https://github.com/zelandiya/keyword-extraction-datasets
科技文献	J1	Hulth 2003[32]	2 000 篇论文摘要	https://github.com/snkim/AutomaticKeyphraseExtraction/
	J2	Nguyen 2007[57]	120 篇科技论文	https://github.com/snkim/AutomaticKeyphraseExtraction/
	J3	Krapivin 2009[85,108]	2 304 篇科技论文	https://github.com/snkim/AutomaticKeyphraseExtraction/
	J4	Nguyen 2007[46,57]	211 篇科技论文	http://wing.comp.nus.edu.sg/downloads/keyphraseCorpus/
	J5	SemEval-2010	SemEval 2010	https://github.com/snkim/AutomaticKeyphraseExtraction/
	J6	SemEval-2017[109]	SemEval 2017	https://scienceie.github.io/
	J7	CiteUlike 180[58,107]	180 篇科技论文	https://github.com/zelandiya/keyword-extraction-datasets/
	J8	Wiki 20[60,107]	20 篇维基百科文章	https://github.com/zelandiya/keyword-extraction-datasets/
	J9	Caragea 2014[10,50]	论文摘要及引文	http://www.cse.unt.edu/~ccaragea/keyphrases.html

5.2 实验评估指标

关键词提取效果评估指标大致可分为针对无序结果的评价指标和针对有序结果的评价指标[110]两类.前者主要包括正确率(precision,*P*)、召回率(recall,*R*)和 *F* 值(*F*1-score,*F*),计算公式具体如下:

$$P=|C|/|E|,\ R=|C|/|S|,\ F=2PR/(P+R) \tag{1}$$

其中,*C* 为正确提取的关键词的集合,*E* 为提取的关键词集合,*S* 为标注关键词的集合.需要特别注意的是,对于这 3 个指标的统计方法又可分为微观(micro)和宏观(macro)两种[110]:微观统计方法先对每篇文档分别计算评价指标再取平均值;而宏观则是先统计好所有结果中的正确关键词数目和,再一次性计算评价指标.一般来说,相同的实验结果,采用宏观统计方法得到的正确率、召回率和 *F* 值会略低于微观统计方法.

此外,由于提取关键词时通常需要事先设定一个统一的提取数目,这样会导致提取的关键词数目和实际标注的数目不匹配.为此,文献[41]提出了 *R* 正确率(*R*-precision,简称 *R*-*p*),要求每篇文档的关键词提取数目和实际数目一致,在此基础上再计算正确率 *P*.当 *R* 正确率的值为 1.0 时,表示获得最理想的提取结果.

针对有序结果的评价指标包括平均倒数等级(mean reciprocal rank,简称 MRR)[111]和二元偏好度量(binary preference measure,简称 Bpref)[112].这两个指标考虑了关键词的排名信息,使用时要求关键词有序序列事先给定.其中,MRR 用来度量每个文档第 1 个被准确提取的关键词的排名情况,而 Bpref 则用来度量提取结果中错误

提取的词语的排名情况,具体计算公式如下:

$$MRR=\frac{1}{|D|}\sum_{d\in D}\frac{1}{r_d},\quad Bpref=\frac{1}{|C|}\sum_{r\in C}1-\frac{|n \text{ ranked higher than } r|}{|E|} \tag{2}$$

其中,r_d为文档 d 第 1 个正确提取结果的排序,$|n$ ranked higher than $r|$表示排列在正确提取词 $r\in C$ 之前的错误提取的词 $n\in(E-C)$的数目.

5.3 现有方法的实验结果对比

近期有几篇文献[49,84,113,114]给出了更多的不同方法的实验结果对比,其中,文献[49,84,113]主要是针对有监督方法,而文献[114]既包括有监督方法也包括无监督方法.从总体上来说,有监督方法的提取效果优于无监督方法的提取效果,但差距不是特别显著,详细结果可参见文献[114],如其中针对体育新闻文档的关键词提取,无监督的 TPR 方法[75]的 F 值为 0.271,有监督方法 SVM 方法(包括了 TF-IDF 和 LDA 主题两个特征,这两个特征与无监督提取方法相关)的 F 值为 0.339.尽管相对于无监督提取方法,有监督提取方法的效果略占优,然而它的不足也很明显,如训练数据并不易获得,标注代价大,且学习得到的分类函数受限于训练数据的特征,还可能存在过拟合的问题.

有监督提取方法的实验结果会受到目标文档数据集、提取的特征以及分类模型这 3 方面的影响,现有方法中实验结果的对比都是在其中两个方面确定的前提下对比第 3 个方面对提取效果的影响,如在数据集和特征确定的前提下,分析不同分类模型的提取效果.在实际应用中,需要根据待提取文档的情况,提取相应的区分度大的特征,进而选用适当的分类模型提取关键词.而对于无监督关键词提取方法,基于图的提取方法优于其他无监督方法,且候选单词图的信息量越丰富,其提取的效果就越好.

此外,值得一提的是,2016 年 Boudin[115]提供了经典关键词提取方法(包括有监督和无监督)的 Python 代码(https://github.com/boudinfl/pke)实现,为后续研究提供了便利.

6 未来研究方向探讨和总结

通过对现有相关研究工作的总结,未来可从以下几个方面展开相关研究.

(1) 继续挖掘一些有助于关键词提取的背景知识.

对于有监督提取方法,其中一条清晰的研究线索是不断地挖掘与目标文档相关的外部知识库中的相关信息特征,如借助维基百科、WordNet 知识库、搜索引擎等外部知识库提取有助于关键词提取的相关特征.而对于基于图的关键词提取方法,同样一条清晰的研究线索是不断添加与目标文档相关的文档以便创建信息量更加丰富的候选关键单词图,如 ExpandRank[95]方法使用与目标文档相似的文档集来辅助构建词图、CiteTextRank[10]方法添加了引用文档集来辅助构建单词图.据此可得,不论是有监督还是无监督关键词提取方法,不断挖掘并整合与目标文档相关的背景知识可以提升关键词提取的效果,故一直以来是重要的研究方向,相信未来也是如此.

(2) 融合现有的多维异质信息从而提升提取效果.

现有的有监督提取方法主要是利用候选关键词的特征,也即词图中节点的特征;而基于图的关键词提取方法主要是利用候选关键词之间的关系特征,也即词图中边的特征.此外,如表 1 所示,不同类型的特征对所提取的关键词需要满足的基本准则的支持程度不同,因此需要对现有的多维异质特征进行融合,以便提高提取效果.尽管 SEAFARER 提取方法[9]首次尝试将节点特征和边特征两种不同类型的特征融合到同一种方法中,然而主题特征是以边在主题下的相似度特征融合到 PageRank 框架模型中.事实上,主题是对节点本身的类别划分,所以将其转化为边特征未必是很好的选择.此外,SEAFARER 方法依然用到有监督信息,然而有监督信息的标注代价非常高且存在过拟合问题.尽管如此,SEAFARER 方法仍然指出了一条很好的研究路线,即尝试把现有的多维异质信息(特别是主题信息)融合到一个统一的模型中,且这些信息的权重系数是通过无监督学习得到的,而不是借助有监督信息通过有监督学习得到.

(3) 尝试通过学习隐式特征提升关键词提取效果.

近年来,表示学习技术得到了长足进步和广泛应用.特别是 Word2Vec[71,76]的提出极大地促进了表示学习技术在自然语言处理领域的应用,尽管个别文献(如 WordAttractionRank 方法[22])尝试在维基百科数据集上通过 SENNA 方法[96]学习词向量特征,然后利用词向量特征提升关键词提取效果,然而在该文献中,词向量仅仅是用来计算词间的相似度特征的.该文献显然存在几方面的不足:词向量的学习应该优先尝试从目标文档中学习,而不是直接从维基百科中学习;在该文献发表时表示学习技术还没有得到长足的发展,现在在表示学习技术得到长足发展的情况下,可以尝试用不同的表示学习技术来学习词嵌入向量,进而提升提取效果;对词向量的使用过于简单,仅用来计算词间的相似度.由此可见,在这个方向上还有很大的研究空间,未来可能会成为一个重要的研究方向,且非常有可能出现突破性的研究成果.

(4) 针对主题动态演进的短文本提出有针对性的提取方法.

目前提出的关键词提取方法大都是针对主题静态的文本,即文档的主题是不变的.这些方法不能很好地适用于主题动态演进的短文本,为此,HG-RANK[104]方法将超图用于推文的关键词中,将时间变化特征和社交网络中的社会属性添加到超图模型中.目前这方面的研究成果非常有限,随着对推文等短文本数据分析需求的不断增加以及关键词提取技术的不断发展,相信未来针对主题动态演进的短文本的关键词提取问题研究将成为一个重要的研究方向.

本文详细综述了面向文本的自动关键词提取技术,围绕各关键词提取方法中使用的特征,展开对现有研究成果进行系统的分析、对比和总结.这种以特征为驱动综述现有关键词提取方法,可使研究着力点由以往重点关注提取方法前移至重点关注特征的定义和对其的综合利用,这种视角的变化将有助于研究者在此基础上提出更好的解决方法.此外,还对一些可能的研究方向进行了简要的探讨,期望能借此推动国内对关键词提取技术的关注和深入研究.

致谢 作者感谢对本文提出宝贵修改意见的评审专家.

References:

[1] Gutwin C, Paynter G, Witten I, Nevill-Manning C, Frank E. Improving browsing in digital libraries with keyphrase indexes. Decision Support Systems, 1999,27(1):81–104.

[2] Kim SN, Medelyan O, Kan MY, Baldwin T. Automatic keyphrase extraction from scientific articles. Language Resources and Evaluation, 2013.47(3):723–742.

[3] Hassaine A, Mecheter S, Jaoua A. Text categorization using hyper rectangular keyword extraction: Application to news articles classification. In: Proc. of the ARAMiCS. Cham: Springer-Verlag, 2015. 312–325. [doi: 10.1007/978-3-319-24704-5_19]

[4] Zhao WX, Jiang J, He J, Song Y, Achananuparp P, Lim EP, Li X. Topical keyphrase extraction from twitter. In: Proc. of the ACL. Stroudsburg PA: ACL, 2011. 379–388.

[5] He WM. Chinese social topic's keywords extraction algorithm [MS. Thesis]. Beijing: Beijing Jiaotong University, 2017 (in Chinese with English abstract).

[6] Zhang WN, Ming ZY, Zhang Y, Liu TS, Chua TS. Exploring key concept paraphrasing based on pivot language translation for question retrieval. In: Proc. of the AAAI. Palo Alto: AAAI Press, 2015. 410–416.

[7] Wu HC, Tian ZH, Wu W, Chen EH. An unsupervised approach for low-quality answer detection in community question-answering. In: Proc. of the DASFAA. Cham: Springer-Verlag, 2017. 85–101. [doi: 10.1007/978-3-319-55699-4_6]

[8] Tang YX, Huang WL, Liu Q, Tung AKH, Wang XL, Yang JS, Zhang BB. QALink: Enriching text documents with relevant Q&A site contents. In: Proc. of the CIKM. New York: ACM, 2017. 1359–1368. [doi: 10.1145/3132847.3132934]

[9] Zhang W, Feng W, Wang JY. Integrating semantic relatedness and words' intrinsic features for keyword extraction. In: Proc. of the IJCAI. San Francisco: Morgan Kaufmann Publishers Inc., 2013. 2225–2231.

[10] Gollapalli SD, Caragea C. Extracting keyphrases from research papers using citation networks. In: Proc. of the AAAI. Palo Alto: AAAI Press, 2014. 1629–1635.

[11] Hasan KS, Ng V. Automatic keyphrase extraction: A survey of the state of the art. In: Proc. of the ACL. Stroudsburg: ACL, 2014. 1262–1273.

[12] Marujo L, Ling W, Trancoso I, Dyer C, Black AW, Gershman A, Matos DMD, Neto JP, Carbonell J. Automatic keyword extraction on twitter. In: Proc. of the ACL and IJCNLP. Stroudsburg: ACL, 2015. 637–643. [doi: 10.3115/v1/P15-2105]

[13] Sterckx L, Demeester T, Deleu J, Develder C. Topical word importance for fast keyphrase extraction. In: Proc. of the WWW. New York: ACM, 2015. 121–122. [doi: 10.1145/2740908.2742730]

[14] Sterckx L, Caragea C, Demeester T, Develder C. Supervised keyphrase extraction as positive unlabeled learning. In: Proc. of the EMNLP. Stroudsburg: ACL, 2016. 1924–1929.

[15] Gollapalli SD, Li XL, Yang P. Incorporating expert knowledge into keyphrase extraction. In: Proc. of the AAAI. 2017. Palo Alto: AAAI Press, 3180–3187.

[16] Florescu C, Caragea C. A position-biased pagerank algorithm for keyphrase extraction. In: Proc. of the AAAI. Palo Alto: AAAI Press, 2017. 4923–4924.

[17] Meng R, Zhao SQ, Han SG, He DQ, Brusilovsky P, Chi Y. Deep keyphrase generation. In: Proc. of the ACL. Stroudsburg: ACL, 2017. 582–592.

[18] Sparck-Jones K. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 1972, 28(1):11–21.

[19] Salton G, Buckley C. Term-Weighting approaches in automatic text retrieval. Information Processing & Management, 1988,24(5): 513–523.

[20] Turney PD. Learning algorithms for keyphrase extraction. Information Retrieval, 1999,2(4):303–336.

[21] Frank E, Paynter GW, Witten IH, Gutwin C, Nevill-Manning CG. Domain-Specific keyphrase extraction. In: Proc. of the IJCAI. San Francisco: Morgan Kaufmann Publishers Inc., 1999. 668–673.

[22] Wang R, Liu W, McDonald C. Corpus-Independent generic keyphrase extraction using word embedding vectors. In: Proc. of the Software Engineering Research Conf. 2014. 39.

[23] Liu ZY, Research on keyword extraction using document topical structure [Ph.D. Thesis]. Beijing: Tsinghua University, 2011 (in Chinese with English abstract).

[24] Ding ZY, Zhang Q, Huang XJ. Keyphrase extraction from online news using binary integer programming. In: Proc. of the IJCNLP. Stroudsburg: ACL, 2011. 165–173.

[25] Figueroa G, Chen PC, Chen YS. RankUp: Enhancing graph-based keyphrase extraction methods with error-feedback propagation. Computer Speech & Language, 2018,47:112–131. [doi: 10.1016/j.csl.2017.07.004]

[26] Rafiei-Asl J, Nickabadi A. TSAKE: A topical and structural automatic keyphrase extractor. Applied Soft Computing, 2017,58: 620–630. [doi: 10.1016/j.asoc.2017.05.014]

[27] Zhao JS, Zhu QM, Zhou GD, Zhang L. Review of the research in automatic keyword extraction. Ruan Jian Xue Bao/Journal of Software, 2017,28(9):2431−2449 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5301.htm [doi: 10.13328/j.cnki.jos.005301]

[28] Boudin F, Mougard H, Cram D. How document pre-processing affects keyphrase extraction performance. In: Proc. of the COLING Workshop on Noisy User-Generated Text. Osaka: The COLING 2016 Organizing Committee, 2016. 121–128.

[29] Toutanova K, Klein D, Manning CD, Singer Y. Feature-Rich part-of-speech tagging with a cyclic dependency network. In: Proc. of the ACL. Stroudsburg: ACL, 2003. 173–180.

[30] Park Y, Byrd RJ, Boguraev BK. Automatic glossary extraction: Beyond terminology identification. In: Proc. of the ACL. Stroudsburg: ACL, 2002. 1–7.

[31] Kumar N, Srinathan K. Automatic keyphrase extraction from scientific documents using *n*-gram filtration technique. In: Proc. of the 8th ACM Symp. on Document Engineering. New York: ACM, 2008. 199–208. [doi: 10.1145/1410140.1410180]

[32] Hulth A. Improved automatic keyword extraction given more linguistic knowledge. In: Proc. of the ACL. Stroudsburg: ACL, 2003. 216–223.

[33] Kim SN, Kan MY. Re-Examining automatic keyphrase extraction approaches in scientific articles. In: Proc. of the ACL Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications. Stroudsburg: ACL, 2009. 9–16.

[34] Wang LT, Li F. SJTULTLAB: Chunk based method for keyphrase extraction. In: Proc. of the ACL Workshop on Semantic Evaluation. Stroudsburg: ACL, 2010. 158–161.

[35] Le TTN, Nguyen ML, Shimazu A. Unsupervised keyphrase extraction: introducing new kinds of words to keyphrases. In: Proc. of the AJCAI. Cham: Springer-Verlag, 2016. 665–671. [doi: 10.1007/978-3-319-50127-7_58]

[36] Xie F, Wu XD, Zhu XQ. Efficient sequential pattern mining with wildcards for keyphrase extraction. Knowledge-Based Systems, 2017,115:27–39. [doi: 10.1016/j.knosys.2016.10.011]

[37] Wang QR, Sheng VS, Wu XD. Keyphrase extraction with sequential pattern mining. In: Proc. of the AAAI. Palo Alto: AAAI Press, 2017. 5003–5004.

[38] Lovins JB. Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 1968,11:22–31.

[39] Bird S. NLTK: The natural language toolkit. In: Proc. of the COLING/ACL on Interactive Presentation Sessions. Stroudsburg: ACL, 2006. 69–72.

[40] Mihalcea R, Tarau P. TextRank: Bringing order into texts. In: Proc. of the EMNLP. Stroudsburg: ACL, 2004. 404–411.

[41] Zesch T, Gurevych I. Approximate matching for evaluating keyphrase extraction. In: Proc. of the RANLP. Stroudsburg: ACL, 2009. 484–489.

[42] Wang R, Liu W, Mcdonald C. How preprocessing affects unsupervised keyphrase extraction. In: Proc. of the CICLing. Berlin, Heidelberg: Springer-Verlag, 2014. 163–176.

[43] Hofmann K, Tsagkias M, Meij E, De Rijke M. A comparative study of features for keyphrase extraction in scientific literature. 2009. http://edgar.meij.pro/comparative-study-features-keyphrase-extraction

[44] Haddoud M, Mokhtari A, Lecroq T, Abdeddaïm S. Accurate keyphrase extraction from scientific papers by mining linguistic information. In: Proc. of the CLBib. 2015. 12–17.

[45] Aquino GO, Lanzarini LC. Keyword identification in Spanish documents using neural networks. Journal of Computer Science & Technology, 2015,15.

[46] Nguyen TD, Luong MT. WINGNUS: Keyphrase extraction utilizing document logical structure. In: Proc. of the ACL Workshop on Semantic Evaluation. Stroudsburg: ACL, 2010. 166–169.

[47] Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG. KEA: Practical automatic keyphrase extraction. In: Proc. of the JCDL. New York: ACM, 1999. 254–255. [doi: 10.1145/313238.313437]

[48] Li GY, Wang HF. Improved automatic keyword extraction based on textrank using domain knowledge. Communications in Computer & Information Science, 2014,(496):403–413.

[49] Haddoud M, Abdeddaïm S. Accurate keyphrase extraction by discriminating overlapping phrases. Journal of Information Science, 2014,40(4):488–500. [doi: 10.1177/0165551514530210]

[50] Caragea C, Bulgarov F, Godea A, Gollapalli SD. Citation-Enhanced keyphrase extraction from research papers: A supervised approach. In: Proc. of the EMNLP. Stroudsburg: ACL, 2014. 1435–1446. [doi: 10.3115/v1/D14-1150]

[51] Zhang K, Xu H, Tang J, Li JZ. Keyword extraction using support vector machine. In: Proc. of the WAIM. Berlin, Heidelberg: Springer-Verlag, 2006. 85–96. [doi: 10.1007/11775300_8]

[52] Tomokiyo T, Hurst M. A language model approach to keyphrase extraction. In: Proc. of the ACL Workshop on Multiword Expressions. Stroudsburg: ACL, 2003. 33–40. [doi: 10.3115/1119282.1119287]

[53] Eichler K, Neumann G. DFKI KeyWE: Ranking keyphrases extracted from scientific articles. In: Proc. of the ACL Workshop on Semantic Evaluation. Stroudsburg: ACL, 2010. 150–153.

[54] John AK, Di Caro L, Boella G. A supervised keyphrase extraction system. In: Proc. of the SEMANTiCS. New York: ACM, 2016. 57–62. [doi: 10.1145/2993318.2993323]

[55] Sarkar K. Automatic keyphrase extraction from medical documents. Pattern Recognition and Machine Intelligence, 2009, 273–278. [doi: 10.1007/978-3-642-11164-8_44]

[56] Feng H, Chen K, Deng XT, Zheng WM. Accessor variety criteria for chinese word extraction. Computational Linguistics, 2004,30(1):75–93. [doi: 10.1162/089120104773633394]

[57] Nguyen TD, Kan MY. Keyphrase extraction in scientific publications. In: Proc. of the ICADL. Berlin, Heidelberg: Springer-Verlag, 2007. 317–326. [doi: 10.1007/978-3-540-77094-7_41]

[58] Medelyan O, Frank E, Witten IH. Human-Competitive tagging using automatic keyphrase extraction. In: Proc. of the EMNLP. Stroudsburg: ACL, 2009. 1318–1327.

[59] Berend G. Exploiting extra-textual and linguistic information in keyphrase extraction. Natural Language Engineering, 2014,22(1): 73–95. [doi: 10.1017/S1351324914000126]

[60] Joorabchi A, Mahdi AE. Automatic keyphrase annotation of scientific documents using Wikipedia and genetic algorithms. Journal of Information Science, 2013,39(3):410–426. [doi: 10.1177/0165551512472138]

[61] Wang JB, Peng H. Keyphrases extraction from Web document by the least squares support vector machine. In: Proc. of the IEEE/WIC/ACM Int'l Conf. on Web Intelligence. Washington: IEEE, 2005. 293–296. [doi: 10.1109/WI.2005.87]

[62] Zhang CZ, Wang HL, Liu Y, Wu D, Liao Y, Wang B. Automatic keyword extraction from documents using conditional random fields. Journal of Computational Information Systems, 2008,4(3):1169–1180.

[63] Bhaskar P, Nongmeikapam K, Bandyopadhyay S. Keyphrase extraction in scientific articles: A supervised approach. In: Proc. of the COLING. Mumbai: The COLING 2012 Organizing Committee, 2012. 17–24.

[64] Marujo L, Gershman A, Carbonell J, Frederking R, Neto JP. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. In: Proc. of the LREC. European Language Resources Association, 2012. 1385–1389.

[65] Fellbaum C. Wordnet: An electronic lexical database. Computational Linguistics, 1998,25(2):292–296.

[66] Ercan G, Cicekli I. Using lexical chains for keyword extraction. Information Processing & Management, 2007,43(6):1705–1714. [doi: 10.1016/j.ipm.2007.01.015]

[67] Suo HG, Liu YS, Cao SY, A keyword selection method based on lexical chains. Zhong Wen Xin Xi Xue Bao/Journal of Chinese Information Processing, 2006,20(6):25–30 (in Chinese with English abstract).

[68] Turney PD. Coherent keyphrase extraction via Web mining. In: Proc. of the IJCAI. San Francisco: Morgan Kaufmann Publishers Inc., 2003. 434–439.

[69] Liu W, Chung BC, Wang R, Ng J, Morlet N. A genetic algorithm enabled ensemble for unsupervised medical term extraction from clinical letters. Health Information Science and Systems, 2015,3(5):1–14. [doi: 10.1186/s13755-015-0013-y]

[70] Zhou XH, Zhang XD, Hu XH. Maxmatcher: Biological concept extraction using approximate dictionary lookup. PRICAI: Trends in Artificial Intelligence, 2006, 1145–1149. [doi: 10.1007/978-3-540-36668-3_150]

[71] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proc. of the Workshop at ICLR. 2013. 1–12.

[72] Liu ZY, Li P, Zheng YB, Sun MS. Clustering to find exemplar terms for keyphrase extraction. In: Proc. of the EMNLP. Stroudsburg: ACL, 2009. 257–266.

[73] Boudin F. A comparison of centrality measures for graph-based keyphrase extraction. In: Proc. of the IJCNLP. Nagoya: Asian Federation of Natural Language Processing, 2013. 834–838.

[74] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003,3:993–1022.

[75] Liu ZY, Huang WY, Zheng YB, Sun MS. Automatic keyphrase extraction via topic decomposition. In: Proc. of the EMNLP. Stroudsburg: ACL, 2010. 366–376.

[76] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proc. of the NIPS. New York: Curran Associates Inc., 2013. 3111–3119.

[77] Lai SW, Word and document embeddings based on neural network approaches [Ph.D. Thesis]. Beijing: The University of Chinese Academy of Sciences, 2016 (in Chinese with English abstract).

[78] Zhang Q, Wang Y, Gong YY, Huang XJ. Keyphrase extraction using deep recurrent neural networks on twitter. In: Proc. of the EMNLP. Stroudsburg: ACL, 2016. 836–845. [doi: 10.18653/v1/D16-1080]

[79] Wang YL, Jin Y, Zhu XD, Goutte C. Extracting discriminative keyphrases with learned semantic hierarchies. In: Proc. of the COLING. Osaka: The COLING 2016 Organizing Committee, 2016. 932–942.

[80] Papagiannopoulou E, Tsoumakas G. Local word vectors guide keyphrase extraction. arXiv Preprint arXiv:1710.07503, 2017.

[81] Whitley D. The genitor algorithm and selection pressure: Why rank-based allocation of reproductive trials is best. In: Proc. of the 3rd Int'l Conf. on Genetic Algorithms. San Francisco: Morgan Kaufmann Publishers Inc., 1989. 116–121.

[82] Hulth A. Reducing false positives by expert combination in automatic keyword indexing. In: Proc. of the RANLP. John Benjamins Publishing Company, 2003. 367–376. [doi: 10.1075/cilt.260.41hul]

[83] Medelyan O, Witten IH. Thesaurus based automatic keyphrase indexing. In: Proc. of the JCDL. New York: ACM, 2006. 296–297. [doi: 10.1145/1141753.1141819]

[84] Bulgarov F, Caragea C. A comparison of supervised keyphrase extraction models. In: Proc. of the WWW. New York: ACM, 2015. 13–14. [doi: 10.1145/2740908.2742776]

[85] Krapivin M, Autayeu M, Marchese M, Blanzieri E, Segata N. Improving machine learning approaches for keyphrases extraction from scientific documents with natural language knowledge. In: Proc. of the JCDL. Berlin, Heidelberg: Springer-Verlag, 2010. 102–111. [doi: 10.1007/978-3-642-13654-2_12]

[86] Jiang X, Hu YH, Li H. A ranking approach to keyphrase extraction. In: Proc. of the SIGIR. New York: ACM, 2009. 756–757. [doi: 10.1145/1571941. 1572113]

[87] Chen YQ, Zhou RQ, Zhu WH, Li MT, Yin J. Ming patent knowledge for automatic keyword extraction. Journal of Computer Research and Development, 2016,53(8):1740–1752 (in Chinese with English abstract).

[88] Sarkar K, Nasipuri M, Ghose S. Machine learning based keyphrase extraction: Comparing decision trees, Naïve Bayes, and artificial neural networks. Journal of Information Processing Systems, 2012,8(4):693–712. [doi: 10.3745/JIPS.2012.8.4.693]

[89] Boudin F. Reducing over-generation errors for automatic keyphrase extraction using integer linear programming. In: Proc. of the ACL Workshop on Novel Computational Approaches to Keyphrase Extraction. Stroudsburg: ACL, 2015. 19–24.

[90] Pei J, Han JW, Mortazavi-Asl B, Wang JY, Pinto H, Chen QM, Dayal U, Hsu MC. Mining sequential patterns by pattern-growth: the prefixspan approach. IEEE Trans. on Knowledge and Data Engineering, 2004,16(11):1424–1440. [doi: 10.1109/TKDE.2004.77]

[91] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: The *c*-value/*nc*-value method. Int'l Journal on Digital Libraries, 2000,3(2):115–130. [doi: 10.1007/s007999900023]

[92] Herbrich R, Graepel T, Obermayer K. Large margin rank boundaries for ordinal regression. In: Advances in Large Margin Classifiers. 2000. 115–132.

[93] Shi W, Zheng WG, Yu JX, Cheng H, Zou L. Keyphrase extraction using knowledge graphs. In: Proc. of the APWeb and WAIM Joint Conf. on Web and Big Data. Cham: Springer-Verlag, 2017. 132–148. [doi: 10.1007/978-3-319-63579-8_11]

[94] Page L, Brin S, Motwani R. The pagerank citation ranking: Bringing order to the Web. Technical Report, Stanford InfoLab, 1999.

[95] Wan XJ, Xiao JG. Single document keyphrase extraction using neighborhood knowledge. In: Proc. of the AAAI. Palo Alto: AAAI Press, 2008. 855–860.

[96] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011,12:2493–2537.

[97] Haveliwala TH. Topic-Sensitive PageRank: A context-sensitive ranking algorithm for Web search. IEEE Trans. on Knowledge and Data Engineering, 2003,15(4):784–796. [doi: 10.1109/TKDE.2003.1208999]

[98] Teneva N, Cheng WW. Salience rank: Efficient keyphrase extraction with topic modeling. In: Proc. of the ACL. Stroudsburg: ACL, 2017,2:530–535. [doi: 10.18653/v1/P17-2084]

[99] Bougouin A, Boudin F, Daille B. TopicRank: Graph-Based topic ranking for keyphrase extraction. In: Proc. of the IJCNLP. Stroudsburg: ACL, 2013. 543–551.

[100] Zhang YX, Chang YC, Liu XQ, Gollapalli SD, Li XL, Xiao CJ. MIKE: Keyphrase extraction by integrating multidimensional information. In: Proc. of the CIKM. New York: ACM, 2017. 1349–1358. [doi: 10.1145/3132847.3132956]

[101] Wan XJ, Yang JW, Xiao JG. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In: Proc. of the ACL. Stroudsburg: ACL, 2007. 552–559.

[102] Ching WK, Fung ES, Ng MK. A multivariate markov chain model for categorical data sequences and its applications in demand predictions. IMA Journal of Management Mathematics, 2002,13(3):187–199. [doi: 10.1093/imaman/13.3.187]

[103] Yan Y, Tan QP, Xie QZ, Zeng P, Li PP. A graph-based approach of automatic keyphrase extraction. Procedia Computer Science, 2017,107:248–255. [doi: 10.1016/j.procs.2017.03.087]

[104] Bellaachia A, Al-Dhelaan M. HG-Rank: A hypergraph-based keyphrase extraction for short documents in dynamic genre. In: Proc. of the MSM. 2014. 42–49.

[105] Bellaachia A, Al-Dhelaan M. Short text keyphrase extraction with hypergraphs. Progress in Artificial Intelligence, 2015,3(2):73–87. [doi: 10.1007/s13748-014-0058-1]

[106] Grineva M, Grinev M, Lizorkin D. Extracting key terms from noisy and multitheme documents. In: Proc. of the WWW. New York: ACM, 2009. 661–670. [doi: 10.1145/1526709.1526798]

[107] Medelyan O. Human-Competitive automatic topic indexing [Ph.D. Thesis]. The University of Waikato, 2009.

[108] Krapivin M, Autaeu A, Marchese M. Large dataset for keyphrases extraction. Technical Report, University of Trento, 2008.

[109] Augenstein I, Das M, Riedel S, Vikraman L, McCallum A. Semeval 2017 task 10: Scienceie-Extracting keyphrases and relations from scientific publications. arXiv Preprint arXiv:1704.02853, 2017.

[110] Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008.

[111] Voorhees EM. The trec-8 question answering track report. In: Proc. of the TREC-8. 1999. 77–82.

[112] Buckley C, Voorhees EM. Retrieval evaluation with incomplete information. In: Proc. of the SIGIR. 2004. 25–32. [doi: 10.1145/1008992.1009000]

[113] Camacho JEP, Ledeneva Y, Hernández RAG. Comparison of automatic keyphrase extraction systems in scientific papers. Research in Computing Science, 2016,115:181–191.

[114] Sterckx L, Demeester T, Deleu J, Develder C. Creation and evaluation of large keyphrase extraction collections with multiple opinions. Language Resources and Evaluation, 2017,52:503–532. [doi: 10.1007/s10579-017-9395-6]

[115] Boudin F. PKE: An open source python-based keyphrase extraction toolkit. In: Proc. of the COLING. Osaka: The COLING 2016 Organizing Committee, 2016. 69–73.

附中文参考文献:

[5] 何伟名.中文社交媒体话题关键词抽取算法[硕士学位论文].北京:北京交通大学,2017.

[23] 刘知远.基于文档主题结构的关键词抽取方法研究[博士学位论文].北京:清华大学,2011.

[27] 赵京胜,朱巧明,周国栋,张丽.自动关键词抽取研究综述.软件学报,2017,28(9):2431–2449. http://www.jos.org.cn/1000-9825/5301.htm [doi: 10.13328/j.cnki.jos.005301]

[67] 索红光,刘玉树,曹淑英.一种基于词汇链的关键词抽取方法.中文信息学报,2006,30(6):25–30.

[77] 来斯惟.基于神经网络的词和文档语义向量表示方法研究[博士学位论文].北京:中国科学院大学,2016.

[87] 陈忆群,周如旗,朱蔚恒,李梦婷,印鉴.挖掘专利知识实现关键词自动抽取.计算机研究与发展,2016,53(8):1740–1752.

常耀成(1992－),男,江苏淮安人,硕士,主要研究领域为自然语言处理.

万怀宇(1981－),男,博士,副教授,CCF 专业会员,主要研究领域为社交网络挖掘,用户画像.

张宇翔(1975－),男,博士,副教授,CCF 专业会员,主要研究领域为自然语言处理,网络数据分析.

肖春景(1978－),女,讲师,CCF 专业会员,主要研究领域为推荐系统,数据挖掘,人工智能.

王红(1963－),女,教授,CCF 专业会员,主要研究领域为智能信息处理,大数据挖掘.