

等,边的权重用词汇间语义的紧密程度来表示.

自动关键词抽取过程中,文档的语言网络图构建需先将文档进行预处理,然后以特征词为节点,以特征词之间的关系作为边,可以是无权的,如果是有权语言网络图,关系程度作为边的权重,可以是有向,也可以无向的.可以不借助外部知识,仅通过文档自身的信息,也可以利用外部知识库,如 HowNet, WordNet 等,确定特征词之间关联及关联程度.

(1) 综合特征值法

在复杂网络的各种基础研究工作中,对网络中节点的重要性进行评估,发掘网络中的重要节点,具有重要的实用价值^[43,44].可以综合利用复杂网络中的统计指标进行节点的重要度计算,从而实现特征提取,也称社会网络中心性分析方法,核心思想是重要性等价于显著性,对网络中重要节点的发现以不破坏网络的整体性为基础.

语言网络图的中心性分析就是定量的分析语言网络的结构拓扑性质,包括网络的局部属性和全局属性.测量这些基本属性的常用统计指标见表 2.

Table 2 Common statistical indicators in language network diagrams (normalized processing)

表 2 语言网络图常用的统计指标(归一化处理)

指标	计算方法	定义或说明
度	$C_{D_i} = \frac{k_i}{N-1}$	节点 i 的度是指与该节点直接相连的节点数目 (有向图中有入度和出度), 表达节点的局部影响力
强度	$C_{SC_i} = \frac{\sum_j w_{ij}}{N-1}$	节点 i 的强度是加权网络中度的表示
中介性	$C_{B_i} = \frac{2}{(N-1)(N-2)} \sum_{s \neq i \neq t} \frac{g_{st}(i)}{g_{st}}$	包括节点介数和边介数, 节点 i 的介数指网络中任意两点间的最短路径通过该节点的比例. 反映节点控制性
接近性	$C_{C_i} = \frac{N-1}{\sum_j d_{ij}}$	节点 i 到其他节点最短路径之和的倒数. 反映了信息传播的紧密度
特征向量	$C_{EC_i} = \frac{1}{\lambda} \sum_j a_{ij} x_{ij}$	节点的分值是与其相邻的邻居分值之和, 对特征方程 $A\lambda = \lambda X$ (λ 是常数) 求特征值
K -core & K -shell	递归去除节点及连边	去除度为 1 的节点以其连边称为 1-shell, 以此类推得到 K -shell. 节点核数的最大值为核数 k -core
离心性	$ECC_i = \max_j(d_{ij})$	节点 i 与其他所有节点距离的最大值
集聚系统	$CC_i = \frac{2 e_{ik} }{k_i(k_i-1)}$	节点 i 的集聚系数是它的相邻节点之间的连接数与它们所有可能存在连接的数量的比值
平均最短路径	$AST_i = \frac{\sum_j d_{ij}}{N}$	节点 i 与图中其他节点最短路径之和的平均值

研究者通常综合利用这些中心性指标来检测语言网络图中的中心节点,并对中心节点进行排序^[45].一般根据研究问题的背景来确定选择哪种中心性进行分析,比如关注的是网络中节点之间的交互行为,可以使用度中心性进行测量;关注的是网络中节点对信息的控制能力,可以使用介数中心性进行测量;关注的是网络中节点传递信息的有效性和独立性,则可以利用接近度中心性进行测量.Lahiri^[46]构建了名词的词搭配网络,采用有向/无向、加权/无权网络,以 TFIDF 为基线,采用 11 种复杂网络统计指标,对英文语言网络图进行了比较研究.Boudin^[47]针对英语和法语构建共现和语法语言网络图,在 3 种标准数据集上,分别通过度中心性、中介中心性、接近中性、特征向量中心性和 TextRank 进行自动关键词抽取实验,比较了各种中心性指标在关键词抽取效果上的差异.Schluter^[48]对单文档进行网络图表示,分别采用 7 种中心性指标对不同表示下的文档语言网络图进行自动关键词抽取,分析不同指标对不同文档语言网络图表示的自动关键词抽取效果.Rousseau^[49]利用 k -core 分解机制选择节点的凝聚子集,将这种方法抽取出来的关键词与单纯通过中心性指标抽取出来的关键词,通过两个标准数据集进行比较,该方法的 F -score 有明显提升.Antoine^[50]选择 k -core, k -truss, k -shell 等对语言网络图进行图收缩,抽取影响力大的节点作为文档关键词.

李军锋^[51]采用 K -最邻近耦合图将专利文献映射成语言网络图,结合平均路径变化量、平均聚类系数变化

量以及当前节点对整个图模型流动性的影响,提出平均连通权重评价指标.分析关键词位置信息、关键词跨度信息以及关键词逆文档频率信息,提出专利综合相关特征衡量关键词的重要性.在传感器领域专利文献的实验结果中,Top-8 级别上的准确率达到 60.9%,Top-10 级别上的召回率达到 73.4%.马力^[52]分析了图的小世界特性,认为文档词语网络图中聚类性强的词语能为关键词的提取提供更为合理的依据.为此引入了一个新的用于度量词语聚类特性变化的变量,测量词语的重要程度,提高对词语的重要性判断能力.左晓飞^[53]提出一种综合考虑网络节点介数和节点加权中心度的综合权值公式,设计并实现了一个基于复杂网络的关键词抽取的原型系统,实验验证,算法获得了较好的抽取效果.

(2) 系统科学法

可以采用系统科学的中心性分析方法,即,节点删除法实现自动关键词抽取.节点删除法主要基于节点(集)的删除,核心思想是重要性等价于该节点(集)被删除后对网络的破坏性.对网络中重要节点的挖掘是通过节点(集)删除前后网络连通性、性能的变化来反映的.系统中,节点(集)的删除除了对系统连通性可能造成破坏外,还会影响到系统的一些其他指标,包括平均度、平均节点强度、直径、平均最短路径、孤立节点数、最大强(弱)连通分支节点数等.可以通过计算这些指标的变化程度来度量节点的重要性^[54].

(3) 随机游走法

网络图上的随机游走(random walk)是指给定图和出发点,随机地选择邻居节点,移动到邻居节点上,然后把当前节点作为出发点,重复以上过程.那些被随机选出的节点序列就构成了一个在图上的随机游走过程.著名的 PageRank 算法^[55]采用了随机游走思想进行互联网网页的重要性计算,该算法不仅在搜索引擎中扮演重要的作用,在文本处理领域也具有很广泛的应用.其实现思想是:在有向图中,当一个节点指向另一个节点时,相当于由弧的起点给弧的终点投上一票,节点得票越多,说明该节点在有向图中越重要,但同时,还跟给它投票的节点的重要程度相关,即,一个节点的重要程度取决于它的票数和给它投票的节点本身的重要程度.计算方法见公式(6):

$$S(v_i) = (1-d) + d \times \sum_{v_j \in \text{In}(v_i)} \frac{1}{|\text{out}(v_j)|} s(v_j) \quad (6)$$

Mihalcea 等人^[56]将 PageRank 应用于关键词抽取领域,并命名为 TextRank. Erkan 等人^[57]也在文本总结中采用了 PageRank 模型. TextRank 的计算方法见公式(7):

$$S(v_i) = (1-d) + d \times \sum_{v_j \in \text{Adj}(v_i)} \frac{w_{ji}}{\sum_{k \in \text{Adj}(v_j)} w_{jk}} s(v_j) \quad (7)$$

其中, $S(v_i)$ 表示节点 v_i 的 PageRank 值或 TextRank 值, d 为阻尼系数.

TextRank 作为模型框架,具有以下几个优点:(1) 无需训练数据,节省了大量成本;(2) 适应性强,TextRank 是一种无监督学习方法,具有很强的适应和扩展能力,对文本没有主题方面的限制;(3) 速度快,TextRank 虽然是矩阵计算,但由于收敛速度快,加之近年来矩阵计算软硬件支持越来越多,使得算法计算速度较快.原始的 TextRank 模型是一个无向加权图,词语之间的语义链接是对等的,仅通过词和词之间逻辑上的分布特性,一般是共现频率构建网络,忽略了词汇的语义相关性,也未考虑上下文及辅助信息.针对这些问题,国内外学者提出了一系列改进的算法.

Habibi^[58]比较了文档语言网络图利用度中心性、接近中心性和 TextRank 进行自动关键词抽取的性能,发现度中心性和 TextRank 在一般性文档中的抽取性能相近,但接近中心性对短文本的抽取性能比前两者都好. Huang^[59,60]构造了文档的两种语言的并行文本网络图,节点为词汇,按语言进行统计,边包括两种语言对应词汇的交叉边,开发了 BiKEA 系统,性能优于 PageRank,将系统应用于语言学习中的自动关键词抽取,在教育领域的阅读理解测试中取得了良好的效果. Yang^[61]针对 TextRank 只关注词的逻辑分布关系的问题,改进了该算法,考虑了句子的重要性,构建了 WS-Rank 系统,与 TextRank 算法做了对比性实验,效果有所提升. Rose^[62]提出的 RAKE 算法比 TextRank 算法效果更好,算法主要抽取关键短语,倾向于较长的短语.

梁伟明^[63]通过修改马尔可夫转移矩阵对应的列,将网络强化为有向加权图,成功实现了将伪语义信息、长

度信息、位置信息和背景信息等融入到 TextRank 模型中.文献[64]把关键词自动抽取问题看做是构成文档的词语的重要性排序问题,基于 TextRank 思想构建候选关键词图,引入覆盖影响力、位置影响力和频度影响力等用于计算词语之间的影响力概率转移矩阵,通过迭代法实现候选关键词的分值计算,并挑选 Top K 个词作为抽取结果.刘通^[65]借鉴 PageRank 思想构造了基于词汇共现关系的词汇概念复杂网络,对词汇的重要性指标进行计算分析,综合考虑目标词汇的频率以及其相邻节点的贡献度,证实了该网络节点评价指标与基于加权度和加权集聚系数的综合指标相比具有优越性.此外,通过复杂网络社区合并的手段,发现了关键节点之间的网络拓扑关系,即核心网络,通过分析核心网络,获得关键词和文本主题的对应关系.张莉婧^[66]设计并实现了自动关键词抽取的 TextRank-CM 算法.将 TextRank-CM 算法、TextRank+TF-IDF 算法和 TextRank 算法分别应用于中文关键词的抽取,结果表明,TextRank-CM 算法在中文关键词抽取中的准确率和召回率明显优于另两种算法.

3.3 其他方法及方法融合

机器学习是人工智能的重点研究领域之一,其核心思想是利用经验来改善计算机系统的性能.所谓的经验,在计算机系统中主要以数据的形式存在,因此,机器学习需要设法对数据进行分析^[67].数据挖掘有时也称为知识发现,目的是识别出海量数据中有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程.简单来讲,数据挖掘就是试图从海量数据中找出有用的知识^[68].而文本就是一种广泛存在的数据,因此,机器学习和数据挖掘在自然语言处理领域中必然发挥重大作用.事实上,机器学习和数据挖掘,还包括众包、机器翻译、查询日志等多种方法,已经应用在自动关键词抽取的许多系统中了.

Onan^[69]分析了 5 种机器学习方法(朴素贝叶斯、支持向量机、逻辑回归、随机森林、组合法)和 5 种统计方法(词频、逆文档频率、共现信息、离心中心性、TextRank 等)对自动关键词抽取性能的影响,给出了文本表示、方法组合等进行自动关键词抽取的建议.Yang^[70]针对 TextRank 中语义缺失的问题,考虑词义信息,构建了基于主题的 TextRank.Li^[71]将背景知识库、Lahiri^[46]将网络指标,与 TextRank 算法结合起来,实现自动关键词抽取.Xu^[72]将聚类算法用于自动关键词抽取,同时考虑词长、聚类中心的窗口大小,实验表明, F -score 提升了 7.5%.Bougouil^[73]提出了 TopicRank 方法,首先对文本中候选短语进行聚类,形成主题,然后构造主题的完全图,采用打分算法获取关键词.Habibi^[74]针对对话语料、Marujo^[75]针对 Twitter 语料,实行众包方式(crowdsourcing methods)进行自动关键词抽取,取得了不错的效果.Liu^[76]对在线评论文本中的评论对象和评论词联合建模,构建异构网络图并进行协作评价、协同打分,实现了自动关键词抽取.Xu^[77]基于词和语言模型,借鉴自动摘要抽取思想,联合建模实现了自动关键词抽取.Chahine^[78]在自动关键词抽取中利用了领域本体库.Wang^[79]综合应用了外部知识库、PageRank、Hits 算法,构建了两层概念文本网络图,实现了自动文本摘要和关键词抽取.

周志^[80]、刘啸剑^[81]发现,基于网络的关键词抽取方法对文档自身的依赖过强:当文档信息足够丰富时,以此建立的词汇网络相对合理,关键词抽取能够取得较好的效果;但是当文本长度较短,文档信息不足以支持建立一个合理的词汇网络,通过该词汇网络的关键词抽取的准确率下降明显.为此,提出一种综合本体知识和复杂网络理论的关键词抽取算法.张建娥^[82]针对 TFIDF 算法数据集偏斜,类间、类内分布偏差的问题,以及复杂网络仅仅依靠词语之间的相互关系作为基本特征,忽略单词的频率特征的弱点,提出一种基于 TFIDF 和词语关联度的关键词提取方法,显著提高了关键词抽取的准确率.翟周伟^[83]应用 K 最邻近耦合图构造文档的图模型,将文档映射为一个语义结构图;然后,结合聚类系数变化量、平均路径变化量(删除节点前后的聚类系数之差、平均路径之差)、TF-IDF 以及区域位置因子来衡量词语节点的重要性,根据重要性得分选择候选关键词集;最后,根据短语合并规则形成最终的关键词.陈忆群^[84]针对某一知识领域构造背景知识库,在此基础上进行目标文本的关键词自动抽取,在专利数据集和开放数据集的实验结果证明,明显优于现有算法.

一般来说,复杂问题解决的多种方法中,单一的方法往往都有其局限性,是在限定条件或一定约束下的实现.经验证明,可以将多种方法有机地结合.自动关键词抽取的宏观方法——匹配法和抽取法可以相互补充;中观方法中的有监督、半监督和无监督机制也可以统筹考虑;在各种微观方法中,各种方案结合实现自动关键词抽取,其有效性是显而易见的.

自然语言处理中的各种文本表示,包括布尔模型、VSM 模型、概率模型、网络图模型等,由于文本建模时

的角度不同,各因素在模型中有所取舍.后续的自动关键词抽取算法中,每个特定的算法通常都有一定的局限性,抽取参数估计时的难易不同,可能会舍弃一部分次要因素.因此,使用单一的自动关键词抽取方案,最终给出的结果未必是最优的.可以考虑方法的融合,比如对多种方案进行叠加,并对各种方案赋予一定的权值;可以对文档分层次或分布实施自动关键词抽取,上一层或前一步的结果作为下一层或下一步的输入;针对同一文档,采用不同的关键词抽取方法,对抽取结果进行交集、并集等集合处理等.

4 自动关键词抽取的评价

关键词抽取的目标是选择一组词语,概括文档的主题.好的关键词除了要文档相关外,还要满足一些约束,包括关键词的数量、话题的覆盖、语义一致性等.这些约束是定义在关键词之间的全局特征,无法通过优化个体关键词而实现.

目前,自动关键词抽取的评价主要有两种形式:一种是单纯借助人工的评价方式,由领域专家进行评价,这种方式可操作性强但缺点也明显,比如认识分歧、词或短语的组合歧义等;另一种是借鉴信息检索模型中的评价指标,包括准确率 P (precision)、召回率 R (recall)、综合指标 F (F -measure)或 $F1$ (F -score)来评价算法的准确性,计算方法见公式(8)~公式(11):

$$P = \frac{\text{抽取出的正确关键词条数}}{\text{抽取出的关键词条数}} \quad (8)$$

$$R = \frac{\text{抽取出的正确关键词条数}}{\text{文档中的关键词条数}} \quad (9)$$

$$F = \frac{(\delta^2 + 1)P \times R}{\delta^2(P + R)} \quad (10)$$

$$F = \frac{2PR}{P + R} \quad (11)$$

P 和 R 的取值都在0和1之间,数值越接近1,查准率或查全率就越高. F 值为正确率和召回率的调和平均值, δ 为调节参数,当参数 $\delta=1$ 时,就是最常见的 $F1$.

从关键词定义和内涵出发,评价自动关键词抽取质量优劣的最佳标准是其符合文档的实际语义.从学术和索引需求出发,衡量关键词(短语)的标准是结构稳定、语义完整单一、统计指标上流通度大,非临时组合等. Saga^[85]、丁卓冶^[34]经过对实际标注的关键词进行深入分析后,提出高质量关键词应该满足以下准则.

- (1) 可读性.关键词本身应该是有意义的词或者短语;
- (2) 相关性.关键词应该和文档的主题密切相关,这是关键词最为本质的要求;
- (3) 覆盖性.关键词应该覆盖文档的各个主题和每个主题的主要方面;
- (4) 连贯性.关键词之间应该是语义相关的,逻辑一致的,如果一个关键词和其他关键词关联很小,那么这个关键词很可能不是好的关键词;
- (5) 简洁性.关键词集合中不应该包含冗余的关键词,因为关键词的数量是有限的,所以关键词之间不应该出现冗余的情况.

5 自动关键词抽取的挑战和发展趋势

目前,自动关键词抽取尽管在某些领域已经满足基本的需要,但作为自然语言处理的一项基础性工作,由于缺乏广泛认可的理论基础、明确有效的数学模型以及切实可行的验证标准,特别是自动关键词抽取的研究尚处于开始阶段,所以,一般意义上的自动关键词抽取的质量和效率仍面临许多挑战.

- (1) 文本预处理不够准确.对中文自动关键词抽取来说,由于分词、新词发现以及短语识别等问题,对系统的准确率、召回率等产生极大的影响;
- (2) 效率低下,复杂度高(特别是融合方法).并不是文章中所有的词语都可以作为候选,如何有效降低候选

词的数量、提高候选词的质量,是必须面对的问题;

- (3) 语义上关联的去重、歧义消解等,怎么计算候选词和文章之间的相关性、如何覆盖文章的各个主题等问题,一直是困扰自动关键词抽取乃至其他自然语言处理的难题;
- (4) 关键词的组合问题,关键词一般具有稳定的结构、较完整的语义、统计特点明显,短语识别的排歧和关键词抽取的质量直接影响自动关键词抽取的质量,关键词以单个词的形式还是以关键词短语的形式进行抽取,尽管已经有一些方法讨论,但缺乏普遍认可的理论基础。

宏观上,关键词分配算法需要预先定义一个关键词词库,这就限定了关键词候选范围,算法的可扩展性较差,且耗时耗力,而关键词抽取算法是从文章的内容中抽取一些词语作为关键词,对于大多数应用来说,现有的方法能够满足基本需求,从长期来看,自动关键词抽取的发展有如下几个方向。

- (1) 多种方法的有效融合,基于规则的方法和基于统计的方法几乎是所有自然语言处理的基本方法,自动关键词抽取也不例外,目前,主流的自动关键词抽取系统多采用这种方式,深入研究语言的特点,确立规则,结合机器学习、数据挖掘等方法,有效组合关键词短语,实现自动关键词的抽取,将规则机制和统计方法有机结合,是自动关键词抽取的必由之路;
- (2) 结合语义的方法,伴随着自然语言处理多年来的持续研究,特别是人工智能、认知科学和脑科学的新发现,由浅层的语言知识向基于语义的深层理解发展,给自动关键词抽取提供了新的研究思路;
- (3) 借助外部知识库,随着互联网的普及,诸如维基百科、百度百科、各种词典、搜索日志、网络评论、专业或领域知识库等可用的资源越来越多,关键词自动抽取时可以借助这些资源,甚至可以把海量的网络资源当作外部知识库,借助外部资源来扩充文本内容,提高关键词自动抽取的效果,将是一种主流的方法;
- (4) 新型模型和思想探索,无论是单文档还是多文档的自动关键词抽取,由于模型和机器效率的考虑,多少限制了技术的实现,伴随着新一代信息技术的出现和广泛应用,比如大数据、物联网、云计算、知识图谱、深度学习、空间关键词^[86]等,必将催生新的自动关键词抽取的模型和思想,如何将趋于成熟的自动关键词抽取方法和这些新技术融合,是未来自动关键词抽取的创新性途径;
- (5) 相关的技术,自动关键词抽取的评价如何克服现有方法的主观性和高成本,特殊文本,比如超长文本、短文本、不规范文本等的处理;有监督中语料的权威性、标准化、标注等;文章中关键词和主题的关系,文档的内部信息、外部信息、主题信息的充分应用等。

6 结束语

科学研究的驱动力来自于人类对宇宙、物质、生命、自我等本质的好奇,而“人类如何表达思想”无疑是一个重大哲学问题,20世纪初的“哲学的语言学转向”,引起了语言学研究的浪潮,使自然语言处理不仅在语言学科中占有重要地位,而且融合了统计学、计算机等学科,有了一定的自然科学色彩,随着计算机、互联网、大数据技术的发展,促使自然语言处理成为一个热门的交叉学科,当前,自然语言处理在理论、技术和应用方面取得了巨大的成果,自动关键词抽取作为自然语言处理的基础性工作,是语言学、图书馆学、情报科学等多学科的研究热点,本文从宏观、中观和微观上对自动关键词抽取技术进行了分类,对主流技术和方法进行了回顾和分析,对自动关键词抽取方法的评价进行了讨论,对自动关键词抽取面临的挑战和未来的发展方向做出了预测和展望,可以预见,自动关键词抽取必然随着技术的发展产生新的思想、模型和方法,其应用也会越来越广泛,为了提高抽取效率和质量,该领域将长期成为自然语言处理的研究内容之一。

References:

- [1] Feather JSP. Int'l Encyclopedia of Information and Library Science. 2nd ed., London & New York: Routledge, 2004. 38-96.
- [2] de Saussure F. Wrote; Liu L, Trans. Course in General Linguistics. In: Liu L, ed. Beijing: the Social Science Press, 2009. 37-49 (in Chinese).
- [3] Liu Y. Computational Linguistics. Beijing: Tsinghua University Press, 2014. 121-132 (in Chinese).

- [4] Baetens J. Conversations on cognitive cultural studies: Literature, language, and aesthetics. *Leonardo*, 2015,48(1):93–94. [doi: 10.1162/LEON_r_00944]
- [5] Wang C, Zhang Q, Gan JP. Study on efficient complex network model. In: Yang Y, Ma M, eds. *Proc. of the 2nd Int'l Conf. on Green Communications and Networks*. Berlin, Heidelberg: Springer-Verlag, 2012. 159–164. [doi: 10.1007/978-3-642-35398-7_20]
- [6] Lai YY, Li C, Goldwasser D, Neville J. Better together: Combining language and social interactions into a shared representation. In: КоврунН H, ed. *Proc. of the Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-10)*. San Diego: Association for Computational Linguistics, 2016. 29–33. [doi: 10.18653/v1/W16-1405]
- [7] Even S. *Graph Algorithms*. 2nd ed., London: Cambridge University Press, 2011. 100–112.
- [8] Aronson E, Wilson TD, Akert RM, Wrote; Hou YB, Zhu Y, Trans. *Social Psychology*. 8th ed., Beijing: Mechanical Industry Press, 2014. 97–103 (in Chinese).
- [9] Luhn HP. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research & Development*, 1957,1(4):309–317. [doi: 10.1147/rd.14.0309]
- [10] Edmundson HP, Oswald VA. Automatic indexing and abstracting of the contents of documents. *Planing Research Corp, Document PRC R-126, ASTLA AD No.231606*. Los Angeles: Planning Research Corp, 1959. 1–142.
- [11] Lois LE. Experiments in automatic indexing and extracting. *Information Storage and Retrieval*, 1970,6(4):313–330. [doi: 10.1016/0020-0271(70)90025-2]
- [12] Turney PD. Learning algorithms for keyphrase extraction. *Information Retrieval Journal*, 2000,2(4):303–336. [doi: 10.1023/A:1009976227802]
- [13] Frank E, Paynter GW, Witten IH, Gutwin C, Nevill-Manning CG. Domain-Specific keyphrase extraction. In: Dean T, ed. *Proc. of the 16th Int'l Joint Conf. on Artificial Intelligence, ACM CIKM Int'l Conf. on Information & Knowledge Management*. 1999. 668–673.
- [14] Hulth A. Improved automatic keyword extraction given more linguistic knowledge. In: Collins M, ed. *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. Sapporo, 2003. 216–223. [doi: 10.3115/1119355.1119383]
- [15] Song M, Song IY, Hu X. KPSpotter: A flexible information gain-based keyphrase extraction system. In: Chiang R, Laender AHF, Lim EP, eds. *Proc. of the 5th ACM Int'l Workshop on Web Information and Data Management*. New Orleans, 2003. 50–53. [doi: 10.1145/956699.956710]
- [16] Hong B, Zhen D. An extended keyword extraction method. In: Yang DH, ed. *Proc. of the Int'l Conf. on Applied Physics and Industrial Engineering*. 2012. 1120–1127. [doi: 10.1016/j.phpro.2012.02.167]
- [17] Zhang C. Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, 2008,3(4):1169–1180.
- [18] Suzuki S, Takatsuka H. Extraction of keywords of novelties from patent claims. In: *Proc. of the 26th Int'l Conf. on Computational Linguistics: Technical Papers*. 2016. 1192–1200.
- [19] Li D, Li S, Li W, Wang W, Qu W. A semi-supervised key phrase extraction approach: Learning from title phrases through a document semantic network. In: Kleber H, ed. *Proc. of the ACL 2010 Conf. on Short Papers*. Stroudsburg: Association for Computational Linguistics, 2010. 296–300.
- [20] Li DC, Li SJ. Hypergraph-Based inductive learning for generating implicit key phrases. In: Simpson S, ed. *Proc. of the Int'l Conf. on Companion on World Wide Web*. New York: ACM Press, 2011. 77–78. [doi: 10.1145/1963192.1963232]
- [21] Lynn HM, Choi C, Choi J, Shin J, Kim P. The method of semi-supervised automatic keyword extraction for Web documents using transition probability distribution generator. In: Kim J, ed. *Proc. of the Int'l Conf. on Research in Adaptive and Convergent Systems*. Odense: ACM Press, 2016. 1–6. [doi: 10.1145/2987386.2987399]
- [22] Siddiqi S, Sharan A. Keyword and keyphrase extraction techniques: A literature review. *Int'l Journal of Computer Applications*, 2015,109(2):18–23. [doi: 10.5120/19161-0607]
- [23] Salton G, Buckley C. Term-Weighting approaches in automatic text retrieval. *Information Processing & Management*, 1988,24(5): 513–523. [doi: 10.1016/0306-4573(88)90021-0]
- [24] Yang KY. Research on automatic extraction algorithm based on improved TFIDF keywords [MS. Thesis]. Xiangtan: Xiangtan University, 2015 (in Chinese with English abstract).

- [25] Huang L, Wu YP, Zhu QF. Research and improvement of keyword automatic extraction method. *Computer Science*, 2014,41(6): 204–207 (in Chinese with English abstract).
- [26] Besils R, Moschitti A, Paziienza M. A text classifier based on linguistic processing. In: Teresa PM, ed. *Proc. of the Int'l Joint Conf. on Artificial Intelligence. UCAI*, 1999. 36–40.
- [27] How BC, Narayanan K. An empirical study of feature selection for text categorization based on term weightage. In: Zhong J, ed. *Proc. of the Int'l Conf. on Web Intelligence. Los Alamitos: IEEE Computer Society*, 2004. 599–602. [doi: 10.1109/WI.2004.10060]
- [28] Suzuki Y, Mitsukawa M, Kawagoe K. A image retrieval method using TFIDF based weighting scheme. In: *Proc. of the Int'l Workshop on Database & Expert System Application. IEEE*, 2008. 112–116. [doi: 10.1109/DEXA.2008.106]
- [29] Liu L, Peng T. Clustering-Based method for positive and unlabeled text categorization enhanced by improved TFIDF. *Journal of Information Science & Engineering*, 2014,30(5):1463–1481.
- [30] Qin P, Xu W, Guo J. A novel negative sampling based on TFIDF for learning word representation. *Neurocomputing*, 2016,177: 257–265. [doi: 10.1016/j.neucom.2015.11.028]
- [31] Hofmann T. Probabilistic latent semantic indexing. In: Gey F, Hearst M, Tong R, eds. *Proc. of the 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press*, 1999. 50–57. [doi: 10.1145/312624.312649]
- [32] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003,3:993–1022.
- [33] Liu Z, Huang W, Zheng Y, Sun M. Automatic keyphrase extraction via topic decomposition. In: Fosler-Lussier E, ed. *Proc. of the Conf. on Empirical Methods in Natural Language Processing. Mit Stata Center*, 2010. 366–376.
- [34] Ding Z, Qiu X, Zhang Q, Huang X. Learning topical translation model for microblog hashtag suggestion. In: Rossi F, ed. *Proc. of the 23rd Int'l Joint Conf. on Artificial Intelligence*. 2013. 2078–2084.
- [35] Pu X, Jin R, Wu G, Han D, Xue GR. Topic modeling in semantic space with keywords. In: Koh YS. *Proc. of the 24th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press*, 2015. 1141–1150. [doi: 10.1145/2806416.2806584]
- [36] Siu MH, Gish H, Chan A, Belfield W, Lowe S. Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. *Computer Speech & Language*, 2014,28(1):210–223. [doi: 10.1016/j.csl.2013.05.002]
- [37] Song Y, Pan S, Liu S, Zhou MX, Qian W. Topic and keyword re-ranking for LDA-based topic modeling. In: Cheung D, Song IY, Chu W, Hu XH, Lin J, eds. *Proc. of the Conf. on Information and Knowledge Management (CIKM 2009). Hong Kong: ACM Press*, 2009. 1757–1760. [doi: 10.1145/1645953.1646223]
- [38] Wei HX, Gao GL, Su XD. LDA-Based word image representation for keyword spotting on historical mongolian documents. In: Hirose A, Ozawa S, Doya K, Ikeda K, Lee M, Liu D, eds. *Proc. of the Neural Information Processing (ICONIP). LNCS*, 2016. 432–441. [doi: 10.1007/978-3-319-46681-1_52]
- [39] Watts D, Strogatz S. Collective dynamics of 'small world' network. *Nature*, 1998,393(6684):440–442. [doi: 10.1038/30918]
- [40] Barabási AL, Albert R. Emergence of scaling in random networks. *Science*, 1999,286(5439):509–512. [doi: 10.1126/science.286.5439.509]
- [41] Cancho RFI, Sole RV. The small world of human language. *The Royal Society of London Series B—Biological Sciences*, 2001, 268(1482):2261–2265. [doi: 10.1098/rspb.2001.1800]
- [42] Quillian MR. Semantic networks. *Approaches to Knowledge Representation Research Studies*, 1968,23(92):1–50.
- [43] Ohara K, Saito K, Kimura M, Motoda H. Resampling-Based gap analysis for detecting nodes with high centrality on large social network. In: Cao T, Lim EP, Zhou ZH, Ho TB, Cheung D, Motoda H, eds. *Proc. of the Advances in Knowledge Discovery and Data Mining (PAKDD). LNCS*, 2015. 135–147. [doi: 10.1007/978-3-319-18038-0_11]
- [44] Santos EE, Korah J, Murugappan V, Subramanian S. Effectively handling new relationship formations in closeness centrality analysis of social networks using anytime anywhere methodology. In: Cai ZP, ed. *Proc. of the IEEE Int'l Conf. on Big Data and Cloud Computing. IEEE Computer Society*, 2016. 354–361. [doi: 10.1109/BDCLOUD-SOCIALCOM-SUSTAINCOM.2016.60]
- [45] Ma HY, Lu P, Zhan ZQ, Huang XX, Wang RB. Research on complex network characteristics of micro-blog language. *Computer Engineering and Applications*, 2015,51(19):119–124 (in Chinese with English abstract).
- [46] Lahiri S, Choudhury SR, Caragea C. Keyword and keyphrase extraction using centrality measures on collocation networks. *Computer Science*, 2014,26(1):1–16.

- [47] Boudin F. A comparison of centrality measures for graph-based keyphrase extraction. In: Convention N, Bureauthe V, eds. Proc. of the 6th Int'l Joint Conf. on Natural Language Processing (IJCNLP). 2013. 834–838.
- [48] Schluter N. Centrality measures for non-contextual graph-based unsupervised single document keyword extraction. Proc. of the Traitement Automatique des Langues Naturelles, 2014, 92(2):455–460.
- [49] Rousseau F, Vazirgiannis M. Main core retention on graph-of-words for single-document keyword extraction. In: Hanbury A, Kazai G, Rauber A, Fuhr N, eds. Proc. of the Advances in Information Retrieval (ECIR). LNCS. Springer Int'l Publishing, 2015. 382–393. [doi: 10.1007/978-3-319-16354-3_42]
- [50] Tixier AJP, Malliaros FD, Vazirgiannis M. A graph degeneracy-based approach to keyword extraction. In: Patwardhan S, Pighinthe D, eds. Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Austin: Association for Computational Linguistics, 2016. 1860–1870. [doi: 10.18653/v1/D16-1191]
- [51] Li JF, Lu XQ, Zhou SJ. Patent keyword indexing based on weighted complex graph model. New Technology of Library and Information Service, 2015,31(3):26–32 (in Chinese with English abstract).
- [52] Ma L, Jiao LC, Bai L, Zhou YF, Dong LB. Research on a compound keywords detection method based on small world model. Journal of Chinese Information Processing, 2009,23(3):121–128 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2009.03.016]
- [53] Zuo XF, Liu HL, Fan YJ, Zhao H. Research of text clustering algorithm based on conceptual semantic field. Journal of Intelligence, 2012,31(5):184–188+195 (in Chinese with English abstract).
- [54] Li P. Study on center nodes of co-occurrence networks of six different languages [MS. Thesis]. Ji'nan: Shandong University, 2014 (in Chinese with English abstract).
- [55] <https://en.wikipedia.org/wiki/PageRank> 2017.
- [56] Mihalcea R, Tarau P. TextRank: Bringing order into text. In: Proc. of the EMNLP 2004. Unt Scholarly Works, 2004. 404–411.
- [57] Erkan G, Radev DR. LexRank: Graph-based lexical centrality as salience in text summarization. Artificial Intelligence Res. (JAIR), 2004,22(1):457–479.
- [58] Habibi M, Popescu-Belis A. Diverse keyword extraction from conversations. In: Vinogradova OI, ed. Proc. of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia: Newdesign, 2013. 651–657.
- [59] Huang CC, Eskenazi M, Carbonell J, Ku LW, Yang PC. Cross-Lingual information to the rescue in keyword extraction. In: Koller A, Yusuke M, eds. Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore: Association for Computational Linguistics, 2014. 1–6.
- [60] Huang CC, Chen MH, Yang PC. Bilingual keyword extraction and its educational application. In: Zervanou K, van Erp M, Alex B, eds. Proc. of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications. Beijing: Association for Computational Linguistics and Asian Federation of Natural Language Processing, 2015. 43–48. [doi: 10.18653/v1/W15-4407]
- [61] Yang F, Zhu YS, Ma YJ. WS-Rank: Bringing sentences into graph for keyword extraction. In: Li F, *et al.*, eds. Proc. of the APWeb. Switzerland, 2016. 474–477. [doi: 10.1007/978-3-319-45817-5_49]
- [62] Rose SJ, Engel D, Cramer N, Cowley W. Automatic keyword extraction from individual documents. In: Berry MW, Kogan J, eds. Proc. of the Text Mining: Applications and Theory. 2010. 1–20. [doi: 10.1002/9780470689646.ch1]
- [63] Liang WM, Huang CN, Li M, Lu BL. Extracting keyphrases from Chinese news articles using TextRank and query log knowledge. In: T'sou B, ed. Proc. of the 23rd Pacific Asia Conf. on Language, Information and Computation. 2009. 733–740.
- [64] Xia T. Study on keyword extraction using word position weighted TextRanl. New Technology of Library and Information Service, 2013,29(9):30–34 (in Chinese with English abstract).
- [65] Liu T. Algorithm research of text keyword extraction based on complex network. Computer Application Research, 2016,33(2): 365–370 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-3695.2016.02.010]
- [66] Zhang LJ, Li YL, Zeng QT, Lei JL, Yang P. Keyword extraction algorithm based on improved text rank. Journal of Beijing Institute of Graphic Communication, 2016,24(4):51–55 (in Chinese with English abstract). [doi: 10.3969/j.issn.1004-8626.2016.04013]
- [67] Zhou ZH. Maching Learning. Beijing: Tsinghua University Press, 2016. 123–145 (in Chinese).

- [68] Bandaru S, Ng AHC, Deb K. Data mining methods for knowledge discovery in multi-objective optimization: Part A—Survey. *Expert Systems with Applications*, 2017,70:139–159. [doi: 10.1016/j.eswa.2016.10.015]
- [69] Onan A, Korukoglu S, Bulut H. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 2016,57:232–247. [doi: 10.1016/j.eswa.2016.03.045]
- [70] Yang K, Chen ZH, Cai Y. Improved automatic keyword extraction given more semantic knowledge. In: Du XY, ed. *Proc. of the 21th Int'l Conf. on Database Systems for Advanced Applications*. Springer-Verlag, 2016. 112–125. [doi: 10.1007/978-3-319-32055-7_10]
- [71] Li GY, Wang HF. Improved automatic keyword extraction based on TextRank using domain knowledge. In: Zong C, ed. *Proc. of the NLPCC*. New York: Springer-Verlag, 2014. 403–413. [doi: 10.1007/978-3-662-45924-9_36]
- [72] Xu SH, Kong F. Toward better keywords extraction. In: Zhou MQ, ed. *Proc. of the Int'l Conf. on Asian Language Processing*. IEEE, 2015. 181–184. [doi: 10.1109/IALP.2015.7451561]
- [73] Bougouin A, Boudin F, Daille B. TopicRank: Graph-Based topic ranking for keyphrase extraction. In: Jiang J, Ku LW, eds. *Proc. of the Int'l Joint Conf. on Natural Language Processing*. ACL, 2013. 543–551.
- [74] Habibi M, Popescu-Belis A. Diverse keyword extraction from conversations. In: Navigli R, Chang JS, eds. *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*. Red Hook: Curran Associates, Inc., 2013. 651–657.
- [75] Marujo L, Ling W, Trancoso I, Dyer C, Black AW, Gershman A, de Matos DM, Neto JP, Carbonell J. Automatic keyword extraction on Twitter. In: Che WX, Zhou GD, eds. *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing (Short Papers)*. Sweden: Taberg Media Group AB, 2015. 637–643. [doi: 10.3115/v1/P15-2105]
- [76] Liu K, Xu LH, Zhao J. Extracting opinion targets and opinion words from online reviews. In: Kristina T, Hua W, eds. *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore: Association for Computational Linguistics, 2014. 314–324.
- [77] Xu H, Martin E, Mahidadia A. Extractive summarization based on keyword profile and language model. In: Mohammad SM, ed. *Proc. of the NAACL-HLT 2015*. Denver: Association for Computational Linguistics, 2015. 123–132. [doi: 10.3115/v1/N15-1013]
- [78] Chahine CA, Chaignaud N, Kotowicz JP, Pécuchet JP. Context and keyword extraction in plain text using a graph representation. In: *Proc. of the 4th Int'l Conf. on Signal Image Technology and Internet Bases Systems Bali (SITIS. INDONESIA.)*. Los Alamitos: IEEE Computer Society, 2008. 692–696. [doi: 10.1109/SITIS.2008.47]
- [79] Wang X, Wang L, Li JW, Li SJ. Exploring simultaneous keyword and key sentence extraction: Improve graph-based ranking using Wikipedia. In: He Q, Melville P, Yin YL. *Proc. of the ACM Int'l Conf. on Information & Knowledge Management*. Maui: CIKM, 2012. 2619–2622. [doi: 10.1145/2396761.2398706]
- [80] Zhou Z, Zou X, Lv X, Hu J. Research on weighted complex network based keywords extraction. In: Liu P, Su Q, eds. *Proc. of the CLSW*. Springer, Berlin, Heidelberg: Chinese Lexical Semantics, 2013. 442–452. [doi: 10.1007/978-3-642-45185-0_47]
- [81] Liu XJ, Xie F, Wu XD. Graph based keyphrase extraction using LDA topic model. *Journal of the China Society for Scientific and Technical Information*, 2016,35(6):664–672 (in Chinese with English abstract). [doi: 10.3772/j.issn.1000-0135.2016.006.010]
- [82] Zhang JE. A Chinese keywords extraction approach based on TFIDF and word correlation. *Journal of the China Society for Scientific and Technical Information*, 2012,10:110–112,123 (in Chinese with English abstract). [doi: 10.13833/j.cnki.is.2012.10.010]
- [83] Zhai ZW, Liu G, Liu YQ. Keywords mining method based on graph model. *Software*, 2012,33(8):9–13 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-6970.2012.08.002]
- [84] Chen YQ, Zhou RQ, Zhu HW, Li MT, Yin J. Mining patent knowledge for automatic keyword extraction. *Journal of Computer Research and Development*, 2016,53(8):1740–1752 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20160195]
- [85] Saga R, Kobayashi H, Miyamoto T, Tsuji H. Measurement evaluation of keyword extraction based on topic coverage. In: Stephanidis C, ed. *Proc. of the HCI*. Switzerland: Springer Int'l Publishing, 2014. 224–227. [doi: 10.7544/issn1000-1239.2016.20160195]

- [86] Liu XP, Wan CX, Liu DX, Liao GQ. Survey on spatial keyword search. Ruan Jian Xue Bao/Journal of Software, 2016,27(2): 329-347 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/4934.htm> [doi: 10.13328/j.cnki.jos.004934]

附中文参考文献:

- [2] de Saussure F, 著;刘丽,译.普通语言学教程.北京:商务印书馆,2009.37-49.
- [3] 刘颖.计算语言学(修订版).北京:清华大学出版社,2009.37-49.
- [8] Aronson E, Wilson TD, Akert RM, 著;侯玉波,朱颖,译.社会心理学:阿伦森眼中的社会性动物.第8版.北京:机械工业出版社,2014. 97-103.
- [24] 杨凯艳.基于改进的 TFIDF 关键词自动提取算法研究[硕士学位论文].湘潭:湘潭大学,2015.
- [25] 黄磊,伍雁鹏,朱群峰.关键词自动提取方法的研究与改进.计算机科学,2014,41(6):204-207.
- [45] 马宏伟,陆蓓,谌志群,黄孝喜,王荣波.微博语言的复杂网络特征研究.计算机工程与应用,2015,51(19):119-124.
- [51] 李军锋,吕学强,周绍钧.带权复杂图模型的专利关键词标引研究.现代图书情报技术,2015,31(3):26-32.
- [52] 马力,焦李成,白琳,周雅夫,董洛兵.基于小世界模型的复合关键词提取方法研究.中文信息学报,2009,23(3):121-128. [doi: 10.3969/j.issn.1003-0077.2009.03.016]
- [53] 左晓飞,刘怀亮,范云杰,赵辉.基于概念语义场的文本聚类算法研究.情报杂志,2012,31(5):184-188+195.
- [54] 李萍.6种语言词同现网络中心节点研究[硕士学位论文].济南:山东大学,2014.
- [64] 夏天.词语位置加权 TextRank 的关键词抽取研究.现代图书情报技术,2013,29(9):30-34.
- [65] 刘通.基于复杂网络的文本关键词抽取算法研究.计算机应用研究,2016,33(2):365-370. [doi: 10.3969/j.issn.1001-3695.2016.02.010]
- [66] 张莉婧,李业丽,曾庆涛,雷嘉丽,杨鹏.基于改进 TextRank 的关键词抽取算法.北京印刷学院学报,2016,24(4):51-55. [doi: 10.3969/j.issn.1004-8626.2016.04.013]
- [67] 周志华.机器学习.北京:清华大学出版社,2016.
- [81] 刘啸剑,谢飞,吴信东.基于图和 LDA 主题模型的关键词抽取算法.情报学报,2016,35(6):664-672. [doi: 10.3772/j.issn.1000-0135.2016.006.010]
- [82] 张建娥.基于 TFIDF 和词语关联度的中文关键词提取方法.情报科学,2012,10:110-112,123. [doi: 10.13833/j.cnki.is.2012.10.010]
- [83] 翟周伟,刘刚,吕玉琴.基于图模型的关键词挖掘算法.软件,2012,33(8):9-13. [doi: 10.3969/j.issn.1003-6970.2012.08.002]
- [84] 陈忆群,周如旗,朱蔚恒.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. [doi: 10.7544/issn1000-1239.2016.20160195]
- [86] 刘喜平,万常选,刘德喜.空间关键词搜索研究综述.软件学报,2016,27(2):329-347. <http://www.jos.org.cn/1000-9825/4934.htm> [doi: 10.13328/j.cnki.jos.004934]



赵京胜(1969 -),男,山东新泰人,副教授,主要研究领域为中文信息处理,自然语言处理,信息抽取.



周国栋(1967 -),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,篇章理解.



朱巧明(1963 -),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为中文信息处理,自然语言处理,信息抽取.



张丽(1991 -),女,学士,主要研究领域为自然语言处理,复杂网路,数据挖掘.