软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(9):2431−2449 [doi: 10.13328/j.cnki.jos.005301]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

自动关键词抽取研究综述*

赵京胜[1,2], 朱巧明[1], 周国栋[1], 张 丽[2]

[1](苏州大学 计算机科学与技术学院,江苏 苏州 215006)

[2](青岛理工大学 通信与电子工程学院,山东 青岛 266520)

通讯作者: 赵京胜, E-mail: zhao5199@163.com

摘 要: 自动关键词抽取是从文本或文本集合中自动抽取主题性或重要性的词或短语,是文本检索、文本摘要等许多文本挖掘任务的基础性和必要性的工作.探讨了关键词和自动关键词抽取的内涵,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.从宏观、中观和微观角度,回顾和分析了自动关键词抽取的发展、技术和方法.针对目前广泛应用的自动关键词抽取方法,包括统计法、基于主题的方法、基于网络图的方法等,总结了其关键技术和研究进展.对自动关键词抽取的评价方式进行了分析,对自动关键词抽取面临的挑战和研究趋势进行了预测.

关键词: 自动关键词抽取;机器学习;统计;主题;语言网络图

中图法分类号: TP181

中文引用格式: 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

英文引用格式: Zhao JS, Zhu QM, Zhou GD, Zhang L. Review of research in automatic keyword extraction. Ruan Jian Xue Bao/Journal of Software, 2017,28(9):2431−2449 (in Chinese). http://www.jos.org.cn/1000-9825/5301.htm

Review of Research in Automatic Keyword Extraction

ZHAO Jing-Sheng[1,2], ZHU Qiao-Ming[1], ZHOU Guo-Dong[1], ZHANG Li[2]

[1](School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

[2](School of Communication and Electronics Engineering, Qingdao Technological University, Qingdao 266520, China)

Abstract: Automatic keyword extraction is to extract topical and important words or phrases form document or document set. It is a basic and necessary work in text mining tasks such as text retrieval and text summarization. This paper discusses the connotation of keyword extraction and automatic keyword extraction. In the light of linguistics, cognitive science, complexity science, psychology and social science, this paper studies the theoretical basis of automatic keyword extraction. From macro, meso and micro perspectives, the development, techniques and methods of automatic keyword extraction are reviewed and analyzed. This paper summarizes the current key technologies and research progress of automatic keyword extraction methods, including statistical methods, topic based methods, and network based methods. The evaluation approach of automatic keyword extraction is analyzed, and the challenges and trends of automatic keyword extraction are also predicted.

Key words: automatic keyword extraction; machine learning; statistics; topic; language network graph

文档关键词表征了文档主题性和关键性的内容,是文档内容理解的最小单位.文档关键词抽取,也称关键词提取或关键词标注,是从文本中把与该文本所表达的意义最相关的一些词或短语抽取出来,文档的自动关键词

* 基金项目: 国家自然科学基金(61272260, 61273320)
Foundation item: National Natural Science Foundation of China (61272260, 61273320)
收稿时间: 2017-01-26; 修改时间: 2017-04-05; 采用时间: 2017-04-13; jos 在线出版时间: 2017-06-05
CNKI 网络优先出版: 2017-06-05 16:32:41, http://kns.cnki.net/kcms/detail/11.2560.TP.20170605.1632.002.html

抽取是识别或标注文档中具有这种功能的代表性的词或短语的自动化技术.

在不支持全文搜索的文献检索初期,关键词作为搜索论文的词语,必须在论文中安排关键词这一项.直到现在,仍然在许多文档中设置该项.但随着互联网的发展和大数据时代的到来,文本信息大量涌现,这些文本不再限于提供了关键词规范的论文,许多文本并没有提供关键词,这就需要人工或计算机程序去抽取这些关键词.同时,在自然语言处理中,关键词在文本聚类、文本分类、文本摘要等领域中有着重要的作用.比如:从某天的所有新闻中提取出这些新闻的关键词,就可以大致知道那天发生了什么事情;从某一学术领域最近一段时间范围内的文献中抽取出关键词,就可以了解当前该领域的学术研究热点.

关键词抽取在图书馆学、情报学、自然语言处理等领域一直受到极大的关注,早期的关键词抽取是通过手工标注,借助人类专家知识完成,这是一项十分繁重的工作.伴随着计算机技术的发展,自动关键词抽取越来越受到关注,大量的自动关键词抽取技术、框架和工具涌现,这些方法取得了一定的成绩和较好的效果.但是,自动关键词抽取的性能依然低下,距离任务的真正解决还有很长的距离,亟待进一步提升抽取或标注的效率和质量.

本文以自动关键词抽取为研究对象,对主流自动关键词抽取方法进行回顾和分析,对自动关键词抽取的质量进行评价,对自动关键词抽取技术面临的挑战和未来发展趋势进行大胆的预测.本文首先分析自动关键词抽取的内涵和理论基础.接下来回顾和分析自动关键词抽取的主要技术和方法.最后对自动关键词抽取的评价标准进行了讨论,并对自动关键词抽取面临的挑战和未来的发展趋势进行了预测.

1 自动关键词抽取的内涵

若要获取大文本中的信息,可能没有办法从文本集合或整段句子集合中找到结果,所以要从大文本中挖掘出信息的核心意思,再从核心意思转换成若干个单词或短语,这个单词或短语就是关键词.因此,它是一个大文本数据下的精炼体,下面给出关键词的几个典型定义.

定义 **1**. 国际信息与图书馆学百科全书[1]:A word that succinctly and accurately describes the subject, or an aspect of the subject, discussed in a document. Both single words (keywords) and phrases (key phrases) may be referred to as ―key terms.

定义 **2**. WordNet: (1) A word that is used as a pattern to decode an encrypted message; (2) A significant word used in indexing or cataloging.

定义 **3**. 现代汉语词典:(1) 指能体现一篇文章或一部著作中心概念的词语;(2) 指检索资料时所查内容中必须有的词语.

定义 **4**. 学术论文写作国家标准(GB7713-87):关键词是为了文献标引工作从报告、论文中选取出来的用以表示全文主题内容条目的单词或术语.单词是指能包含一个词素(语言中最小的有意义的单位)的词或语言里最小的可以自由运用的单位,术语则是指某个学科中的专业用语.

定义 **5**. 维基百科:关键词是一种获取信息的精炼的词汇.

定义 **6**. 百度百科:关键词特指单个媒体在制作使用索引时所用到的词汇.

归纳起来,关键词是表达文档主题意义的最小单位,关键词抽取是一种识别有意义且具有代表性文本片段或词汇的技术,自动关键词抽取是通过计算机程序从文档中自动抽取具有重要性和主题性的词或短语的自动化技术.

自动关键词抽取在文本挖掘领域被称为自动关键词抽取(automatic keyword extraction),在信息检索领域通常被称为自动标引(automatic indexing).从抽取结果上看,关键词抽取可以仅仅抽取词语,这种实现较多,比如FudanNLP,jieba,SnowNLP 等;另一种则是词和短语一起抽取出来,这种方法需要增加短语抽取的功能,这一类的实现包括 ICTCLAS,Stanford Word Segmenter,ansj_seg 等.

从宏观上划分,关键词获取方法分为两种:一种是关键词分配(keyword assignment),给定预定义的语料或关键词库,对待抽取关键词的一篇文档,在语料或词库里面找到几个词语作为这篇文章的关键词;另一种是关键词抽取(keyword extraction),对一篇文档,从文档中抽取一些词语作为这篇文档的关键词.目前,大多数独立于语言

和领域的关键词抽取算法是基于后者的,这种方法在实际使用中更有意义.宏观划分中,还有一种划分思想,就是将自动关键词抽取系统分为有外部知识的支持和没有外部知识的支持.前者在自动关键词抽取时除了文档自身的内容和结构信息外,还需要文档词法、句法、篇章或语义相关的其他知识,比如词典(如 WordNet,Cilin 等)、百科(如 Baidu,Wikipedia 等)、通用知识库(如 HowNet)、特定领域的知识库或者与文档主题相关的其他文档等;后者只应用文档自身的信息,不需要外部知识的辅助.

自动关键词抽取系统的实现主要包括 3 个步骤:首先是文本预处理,涉及文本(集合)的组织、格式处理、分词、去停用词等;其次是候选词选择,统计发现,大多数的关键词往往是名词、动词、形容词等实词(短语),而连词、助词、叹词等虚词(短语)几乎不作为关键词,所以可以按照语言规则筛选文本中的词(短语),形成候选关键词集合;最后是评价和确定关键词,对每个候选词,利用特定方法或方法组合计算它是否为关键词的特性,可以进行判断、分级或打分,将大于概率阈值、评定级别或得分高的候选词按数量要求确定为关键词.

2 自动关键词抽取的理论基础

长期以来,自动关键词抽取一直作为语言学、计算机科学和统计学等学科研究的一项工程课题,是自然语言处理和文本挖掘中的基础性工作,较少探讨其理论依据.但细究起来,关键词抽取还是具有较强的理论依据的.关键词在学术领域称为术语,是特定学科领域的概念称谓.术语是索绪尔意义的语言符号,其语言学表达称为关键词,为所指和能指的统一体,其研究已经有几十年的历史和广泛的应用[2].

2.1 语言学基础

人类语言中的句子总是由有限的语法元素(词、词组等)以某种非随机的方式组织而成的,有限的语法元素可通过不同的排列组合构成表达任何语义的句子.所谓“某种非随机的方式”是指在符合语法规范下将相关的语法元素组织成有明确表达意图的句子的形式.语言中句子的构造过程强调语法元素间的相关性.关键词提取方法应该反映语言组织结构中深层次的元素间相关性[3].

计算语言学,也称自然语言处理,是建立形式化的数学模型,利用计算机自动实现分析和处理自然语言的过程.经过 50 多年的探索和研究,在词法、语法、篇章、语义和语用等方面建立了一些模型和系统,取得了一些进展和成就.

2.2 认知科学基础

按照认知科学的观点,人类必须首先识别、学习和理解文本中的实体或者概念,才能理解自然语言文本,而这些实体和概念大都是由文本句子中的名词或名词短语所描述的,因此,计算语言学对关键词的研究焦点集中在名词或名词短语上,特别是基本名词短语的识别和结构分析,这是自然语言处理和信息检索领域的基础研究课题[4].

一个词在交流过程中出现的频率越高,其语言产生的能力越强,即,人脑能够更容易使用这个词表达思想.另外,同一作者表达同一种观点时往往用意义相近的词语来描述,这些词语就是认知意义下的关键词.

2.3 复杂性基础

可以将文档表示成反映元素间某种关系的语言网络图(节点表示离散的词语,边代表词语间的关联)来研究语言的内容或组织结构,研究发现,文档词语网络中存在小世界特性.分析原因可以认为,当作者写文章时,总是会逐个描述主题,使词语聚集起来形成多个簇,然后把这些主题结合起来表达一种观点,即:在多个主题间建立关联,其结果是文章的词语网络图呈现小世界特性[5].

依据复杂网络的观点,词语网络图中聚类性强的节点,就代表文档的一个关键词,在文档的词语网络图中寻找聚类性强的词语,为准确提取关键词提供了深层次的依据.一篇文档词语网络图呈现小世界特性时,则该图中存在一些这样的词语,它们或者多次出现在较短的连接路径上或者有较高的聚类系数,若去掉某个词后,则该文档词语网络图的小世界特性的程度就会减弱[6,7].

2.4 心理学和社会学基础

心理学的邻近联系法则是指曾经在一起感受过的对象往往在想象中也联系在一起,以至于在想起他们中的某一个的时候,其他的对象也会以曾经同时出现时的顺序想起.根据该法则,两个词的联系可用同时感知到两个词的相对频率来衡量;同时,词之间的联系强度决定了用词过程中词汇的选择,只有存在关联的词汇才会被想起、说出或写下.如果文本中的两个词频繁出现在一起,可以按其相似性或相关性确定其是否为关键词.

知识结构和映射原则是指热衷或从事某项研究的科学家,无论其社会和知识背景如何,很大程度上,对同一研究课题和概念,其所使用的词汇基本上是一样的.这对实现领域相关性自动关键词抽取提供了依据[8].

3 自动关键词抽取的技术和方法

美国 IBM 公司的 Luhn[9]提出了基于词频统计的文献自动标引方法,标志着关键词抽取研究和应用的开始.Edmundson[10]开启了用计算机程序进行文献自动检索的实用化时代,设计并实现了最早的自动关键词抽取系统.从那时起,许多学科的研究人员投入到自动关键词抽取的研究中,形成了统计分析法、语言分析法、人工智能法、混合法等几大技术体系[11].下面从多个角度对主流的技术和方法进行梳理和分析.

3.1 中观方法分析

从中观上划分,自动关键词抽取方法分为 3 种.

- 第 1 种是无监督方法,将关键词抽取看做是二元分类问题,判断文档中的词或短语是或不是关键词.这种方法必须提供已经标注关键词的训练语料:首先,利用训练语料训练关键词抽取模型;然后,利用得到的模型对需要抽取关键词的文档进行自动关键词抽取;
- 第 2 种是半监督方法.这种方法不像前者需要大量的训练数据,只需要少量的训练语料,利用这些语料训练抽取模型,利用模型进行未标注文本的关键词抽取,人工对抽取结果进行甄别,将正确的标注加到训练语料中再训练模型.之所以称为半监督方法,是因为有人工的参与,而非全自动的实现;
- 第 3 种是无监督方法.这种方法不需要训练语料,也不需人工参与,利用抽取系统完成文档或文档集合的自动关键词抽取.

3.1.1 有监督方法及语料库技术

广义上的有监督机器学习是从给定的训练集中训练出一个模型,对新数据,利用这个模型来预测结果.在关键词抽取领域中,可以把关键词抽取任务转化为分类问题或标注问题,即:把文档中的词看成是候选的关键词,通过分类学习算法或序列标注方法来判断这些候选词是否为关键词.基于有监督学习的关键词抽取的一般步骤是:首先,建立一个包含大量文本并标出关键词的训练集合;然后,利用训练集合对分类或标注算法进行训练得到一个模型;最后,应用训练好的模型对新文本进行关键词抽取.

有监督机器学习的分类或标注方法常借助决策树(DT)、朴素贝叶斯(NB)、支持向量机(SVM)、最大熵模型(ME)、隐 Markov 模型(HMM)、条件随机场(CRF)模型等.

在有监督方法中,主要有两个研究方向.

- 一个方向是将自动关键词抽取看做是二分类任务,即,判断文档中的一个词是关键词或不是关键词.这个方向的研究主要是基于一些特征建立抽取关键词的分类器.Turney[12]实现了一个自动关键词抽取系统 GenEx,该系统通过使用词性的频率和词性信息作为特征,使用决策树算法作为分类器; Frank[13]实现了自动关键词抽取系统 KEA,使用朴素贝叶斯方法构造分类器;
- 另外一个方向是基于语言模型的,研究人员从训练集中建立语言模型,并选择出符合关键词特征的模型.Hulth[14]利用词性标注、名词短语块等作为特征进行自动关键词抽取;Song[15]的 KPSpotter 系统利用词性标注、信息增益、词位置等作为特征进行自动关键词抽取.

这几个系统,特别是 GenEx 和 KEA,奠定了自动关键词抽取的有监督方法,已经成为后续改进方法和其他自动关键词抽取系统的参照基准系统[16−18].

一般情况下,基于有监督学习的算法往往需要建立训练集合,称为语料库(corpus),主要指由大量实际使用的语言信息组成,专供分析、描述和研究的语言资料库.语料库一般存储在计算机中,并能利用计算机进行检索、查询和分析,是在随机采样的基础上收集人们实际运用的、具有代表性的真实语言材料,并针对通用或特定需求加以标注.20 世纪 60 年代的布朗语料库(Brown corpus)是第一个具有代表性的通用平衡语料库.80 年代后,英国制作了标注语料库(LOB corpus)与语音语料库(birmingham corpus).20 世纪 90 年代出现了一批实用的语料库,包括日本的 EDR 语料、NHK 新闻稿语、英国国家语料库(BNC)等.典型的中文语料包括国际语言资源联盟 LDC 的 Chinese Gigaword 新闻语料、搜狗实验室(Sogo labs)和数据堂(datatang)的网络语料.

事实上,进入 21 世纪以来,大型语料库建设的势头已经放缓,取而代之的是大批小型的针对特定应用的语料库构建.在自动关键词抽取领域,大型的语料库几乎不作为主要的服务目的.近年来,出现了许多特定应用领域的小型语料库,如军事语料库、海事语料库、商务语料库、计算机专业语料库等.另外,还有文学类、法律类、经贸类、旅游类、交通类、餐饮类等语料库.

训练语料的质量往往会直接影响到模型的准确性,从而影响着关键词抽取的结果.现已标注关键词的文本有限,训练集要自己去标注,人工标注带有一定的主观因素,会造成实验数据具有不真实性.因此,如何获取一个高质量的训练集合,是这类算法的一个瓶颈问题.

3.1.2 无监督方法和半监督方法

自动关键词抽取作为一种无监督的学习任务,不需要人工标注语料,可以通过对关键词排名的手段来实现.首先,设置一些权重量化指标;然后,根据这些权重排序;最后,选择出具有最高权重的词或短语.无监督方法是近年来研究和应用的重点,经常采用的技术手段包括统计法、基于主题的方法、基于网络图法等,第 3.2 节中将详细分析.

自动关键词抽取的半监督方法研究和应用较少,Li[19,20]构建了文本语义网、超图,利用文本标题中出现的词汇进行迭代,利用维基百科知识进行推理实现自动关键词抽取.Lynn[21]构建了 TPDG(transition probability distribution generator)基准系统,借助马尔可夫条件转移矩阵进行自动关键词抽取.

基于有监督的自动关键词抽取方法通过对大量文本进行分析,从而得到一种关键词抽取的规则,这种规则是通过大量文本训练得到的,相对于基于无监督的自动关键词抽取方法所得到的规则更加科学、有效,抽取的关键词的质量有大幅度的提高.但是,有监督的自动关键词抽取方法不能直接对文本进行关键词的抽取,过程较为复杂,使用起来不是很方便.方法需要通过大量文本训练,大规模人工标注的训练语料难以获取,抽取效果受训练语料的规模和领域性影响较大,只要训练集不同,构造的分类模型也会有所差异,最终也会影响关键词抽取的准确性,相对于基于无监督的自动关键词抽取方法更加复杂.因此,有监督的自动关键词抽取方法的应用不是很广泛.

3.2 微观方法分析

从微观上划分,近年来提出的自动关键词抽取方法有上百种,有些方法在特定领域的关键词抽取正确率在 90%以上,基本达到实用的程度,但通用的独立于语言和领域的自动关键词抽取方法还需进一步研究.目前,主流的自动关键词抽取方法,比如基于语言分析的方法、统计法、机器学习法等有 10 余种.有些自动关键词抽取系统不只是用单一的方法,可能是两种或多种方法的组合,有的系统可能是多个简单系统的融合.

与大多数自然语言处理任务一样,自动关键词抽取过程中对文档进行分析分为两个层次:一种是浅层的语言分析,包括文档的组成元素和结构信息分析,组成元素主要涉及语言中的词性、词频等基础语言信息,结构信息主要指共现、语法等组合语言信息,通过分析获取候选关键词,进而得到文档关键词;另一种是深层的语义分析,因为现实中的实体和概念通常表达成词汇,如果能让系统理解词义,分析出其表达的现实实体和概念,则自动关键词抽取就非常简单了.目前,深层语义分析不仅是自动关键词抽取的难点,也是其他自然语言处理研究的难点.从微观角度对主流自动关键词抽取方法的分类见表 1[22].

目前,有的自动关键词抽取系统可能是单一的实现方法,有的可能是多种方法的结合,按其采用的核心方法将其归到某一类,本文对典型的最具代表性的几种方法归纳总结.

Table 1 Micro category of the method in automatic keyword extraction

表 **1** 自动关键词抽取微观方法分类

方法	技术特点	代表性成果	发展趋势
语言模型法	浅层词法、语法分析、深层语义分析	H.P.Luhn	应用较多,长期研究方向
简单统计法	词性、词频等概率统计,易于实现	TFIDF	当前主流方法
主题法	文档-主题-词汇生成模型,较复杂	PLSA,LDA	当前主流方法,多领域应用
网络图法	文本复杂网络、图论、矩阵理论应用	TextRank	当前主流方法,多领域应用
机器学习法	利用学习算法训练模型、测试、应用	GenEx,KEY	应用较多,长期研究方向
数据挖掘法	利用数据挖掘算法,如聚类、关联规则等		领域性、小范围应用
神经网络	深度学习网络模型	word2vec	研究热点
在线众包法	众包、验证		领域性、小范围应用
查询日志法	结合信息检索的日志(关键词)		领域性、小范围应用
机器翻译法	不同语言之间的对应		领域性、小范围应用
超网络法	文本网络的网络		领域性、小范围应用
协同功能法	与其他文本挖掘任务协同,如文本摘要		长期研究方向
本体法	结合现实世界的实体、概念(本体)		长期研究方向
注:有的系统易于归于某种方法,许多系统涉及的实现方法是几种方法的组合			

3.2.1 统计法

利用文档中词语的统计信息抽取文档的关键词,这种方法相对来说较简单,不需要训练数据,一般也不需要外部知识库.文档经过预处理得到词向量后,可以利用简单的统计规则,比如词性过滤、词频等,如果涉及短语的抽取,可利用互信息、均值、方差、共现等进行筛选,形成候选词(短语)集合.对候选词的关键词判定采用特征值量化的方式,比较常用的统计量化指标主要有 3 种.

(1) 基于词权重(term weight),包括词性、词计数、词频、文档频率、逆文档频率、x^1 测度、平均词频、相对词频、词长等;

(2) 基于词的文档位置(term location),包括文档前 N 个词、文档后 N 个词、段首 N 个词、段尾 N 个词、标题词、文档特殊位置词(摘要、引言、结论)等;

(3) 基于词的关联信息(term collocation),包括互信息、均值、方差、共现度、依存度、中心性测度、聚类系数、TFIDF 值、PageRank 值、Hits 值、短概念缩减、长概念提升、词跨度等.

很明显,这类方法具有模型可泛化性,易于实现,具有语言和领域的独立性,关注统计特性.主流的简单统计方法是 TFIDF(term frequency-inverse document frequency)及其改进方法.

TFIDF 是由 Salton[23]在 1988 年提出的,用于评估一个词对一个文档集或一个语料库中的其中一份文档的重要程度,其中:TF 称为词频,用于计算该词描述文档内容的能力;IDF 称为逆文档频率,用于计算该词区分文档的能力.TFIDF 方法的指导思想建立在这样基本假设之上:在一个文档中出现很多次的单词,在另一个同类文档中出现次数也会很多;反之亦然.另外还要考虑单词区别不同类别的能力,一个词出现的文本频率越小,它区别不同类别的能力就越大.以 TF 和 IDF 的乘积作为特征空间坐标系的取值测度.各项目的计算方法见公式(1)~公式(4).

$$tf_i = \frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

$$idf_i = \log \frac{|D|}{|D_i|+1} \tag{2}$$

$$w_{ij}=tfidf_{ij}=tf_{ij}\times idf_i \tag{3}$$

$$w_{ij} = \frac{w_{ij}}{\sqrt{\sum_k w_{kj}}} \tag{4}$$

其中,$n_{i,j}$ 是词 t_i 在文档 d_j 中的出现次数,$|D|$表示语料库中的文档总数,$|D_i|$表示语料库中包含词 t_i 的文档总数,w_{ij} 表示词 t_i 在文档 d_j 中的权重(归一化).

用 TFIDF 算法计算特征词的权重时,当一个词在一篇文档中出现的频率越高,同时在其他文档中出现的次数越少,表明该词对于表示这篇文档的区分能力越强,其权重值就越大.将所有词的权重值排序,根据需要可以有两种选择方式:选择权值最大的 Top K 个关键词或选择权值大于某一阈值的关键词.

TFIDF 算法的优点是简单快速,结果比较符合实际情况.但传统 TFIDF 算法也有明显的缺点,单纯以词频衡量一个词的重要性,不够全面,有时重要的词可能出现次数并不多.而且,这种算法无法体现词的位置、词性和关联信息等特征,更无法反映词汇的语义信息.本质上,IDF 是一种试图抑制噪音的加权,并且单纯地认为文档频率小的单词就越重要,文档频率大的单词就越无用.这对于大部分文本信息,并不是完全正确的.事实上,有些不能代表文本的低频词,IDF 的值反而很高;有些能够很好代表文本的高频词,IDF 值却很低.主要的原因是,TFIDF 中 IDF 的算法没有考虑特征项在文档集合类间和类内的分布情况[24,25].

(1) IDF 没有考虑特征项在类间的分布信息.

假设某一类 C_i 包含词条 t_i 的文档数为 n,而其他类包含 t_i 的文档数为 m,包含词条 t_i 的文档数为 w,$w=n+m$,n 增大时 w 也变大,可是其 IDF 值却变小.实际上,n 增大表示词条 t_i 在 C_i 类中出现次数增多,t_i 作为特征词代表这一类权重应该增大,但由公式(2)计算出的 IDF 值却变小,这就导致其 *TFIDF* 值变小.另一方面,如果包含词条 t_i 的文档总数 w 小,而词条 t_i 均匀地分布在各个类间,则该词条 t_i 不具有区分能力,不应该作为特征词,应该赋予较低的权重.但是按照传统的 TFIDF 算法,得出的 IDF 值却很大,与分析的结果相反.

(2) IDF 没有考虑特征项在类内的分布信息.

假设某一类 C_i 包含的文档数为 n,词条 t_i 共 w 个,较均匀地分布在 n 个文档中,表明词条 t_i 对该类的特征表示作用明显,可以计算出这种情况下的 TDIDF 值.假设这 w 个词条 t_i 只分布在一个或 $n_1(n_1<<n)$个文档内,可能是偶然因素导致的特殊词条,其 *TF* 值未变,但 IDF 值却非常大,也就是其 TFIDF 值将非常大.应该是同一类文档中不同的特征项,类内分布均匀的比分布不均匀的特征权重要高,但个别特征项在类内的异常分布,会使其 TFIDF 值很大.

国外对 TFIDF 方法的研究较早,提出的改进方法侧重 IDF 量化指标的优化.Besils[26]提出了 TF*IWF*IWF 算法,IWF(inverse word frequency)表示语料库词语总数与待查文本中该词出现的次数之比的对数,方法有效提高了特征词在语料库的权重,一定程度上克服了 TFIDF 算法中 IDF 带来的问题.Bong[27]根据不同类别的文档数可能存在数量级的差距,提出利用 CTD(category term descriptor)来改进 TFIDF,以改善类别数据集偏斜所引起的误差.文献[28]将 TFIDF 方法应用于图像处理领域,克服传统图像处理重点处理考虑图像的颜色图、形状特征的弱点,利用信息检索思想计算其 TFIDF 特征权重.

国内很多学者对 TFIDF 算法进行了研究,特别是应用于汉语处理,取得了显著的成果.文献[25]对这些成果进行了较全面的总结:修改 IDF 的计算方法,增加那些在一个类中频繁出现的特征项权重,考虑了 IDF 在类别中的分布情况;把信息论中的信息增益应用到文本集合的类别层次上,重新考虑特征词对类别的区分度;针对特定领域的 TFIDF 算法改进等.刘露[29]增加了 TFIDF 方法的计算能力,提出了 TFIPNDF(term frequency inverse positive-negative document frequency),通过对“支持”、“反对”的情感判断,提高传统 TFIDF 的准确性.秦鹏达[30]进行深度学习训练时,提出了 NEG-TFIDF(negative sampling-TFIDF)方法,利用反例的特征权重来优化模型.杨凯艳[24]提出了 M_TF-IDF-IGD 算法,一方面考虑词语在文本集上的分布情况对权重的影响,通过信息增益量化类间分布信息,抑制 IDF 值对文档频数的高度依赖性;另一方面,综合考虑文本中词语与权重联系密切的多个特征,并将这些特征进行融合,然后用融合后的新特征替代传统 TFIDF 算法中的 TF 因子.

3.2.2 基于主题的方法

作者在日常生活中是如何构思文章的?简单来看,假如要写一篇文章,首先需要确定几个主题,然后选择对主题描述较好的词汇,按语法规则组成句子、段落、篇章等.实际上,统计意义下该主题使用概率较高的词,就是该主题的关键词.

主题模型就是模拟人类作者写文章的概率语言模型,其主要思想包括两点:① 文档是若干主题的混合分布;② 每个主题又是词语的概率分布.最早的想法是 Hofniann[31]给出的 PLSA(probability latent semantic

analysis)模型,模型认为:一篇文档可以由多个主题混合而成,而每个主题都是词汇上的概率分布.每篇文档中的每个词都是这样一个过程得到的:“通过概率选取出某个主题,并从该主题中通过概率选取某个词语”.一篇文档的生成,其中每个词语出现的概率计算见公式(5):

$$P(\text{词语}/\text{文档})=\sum_{\text{主题}}P(\text{词语}/\text{主题})\times P(\text{主题}/\text{文档}) \tag{5}$$

当前,最主要的主题模型是 LDA(latent dirichlet allocation)模型,其创始者 Blei[32]在 PLSA 的基础上加入了 Dirichlet 先验分布,是 PLSA 的突破性的延伸.PLSA 在文档对应主题的概率分布的计算上没有使用统一的概率主题模型,参数较多会导致过拟合现象,对训练数据集以外的文档的概率分布计算会比较难.在这个基础上,Blei 在 LDA 中引入了超参数,无论外部文档数量怎么变化,都只有一个超参数,然后,通过概率方法对模型进行推导,寻找文档集的语义结构,挖掘文档的主题,其实现过程如图 1 所示.

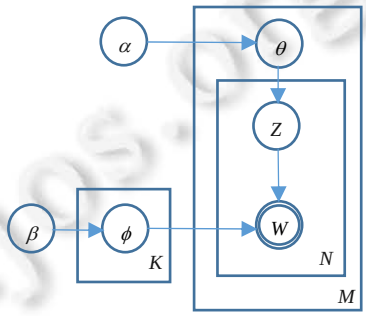

Fig.1 LDA topic model

图 1 LDA 主题模型

LDA 模型中,文档-主题和主题-词都服从多项分布,其先验概率是 Dirichlet 分布.文档中的词不考虑顺序,符合 BOW(bag of word)词袋模型.假设文档集总词汇数为 V,一篇由 n 个词组成的文档,每个词的生成都服从多项分布,就像上帝抛一个有 V 面的骰子(每面对应一个词),抛 n 次就可以生成一篇文档了,对应图 1 中的ϕ分布(β为超参);类似地,不同文档之间的骰子不是同一个,每次为文档选一个主题骰子,这个过程也服从多项分布,对应图 1 中的θ分布(α为超参).归纳起来,LDA 有两个阶段.

(1) $\alpha\rightarrow\theta_m\rightarrow Z_{m,n}$,选取一个参数为$\theta_m$的文档-主题分布,然后对第 m 篇文档的第 n 个词的主题,生成 $Z_{m,n}$ 的编号.从一个参数为α的 Dirichlet 分布中采样出一个多项分布θ_m,作为该文档在 k 个主题上的分布;

(2) $\beta\rightarrow\phi_k\rightarrow W_{m,n}|k=Z_{m,n}$,生成第 m 篇文档的第 n 个词.对该文档中的每个词,根据上步中的θ_m分布,采样出一个主题编号,然后根据主题-词分布对应参数为β的 Dirichlet 分布中采样出一个多项分布,作为该主题下词的分布.

参数α和β是由经验给出的,$Z_{m,n}$,θ_m,ϕ_k是隐含变量,需要推导.LDA 的推导方法有两种:一种是精确推导,例如 Blei 在论文里阐述的用 EM(expected maximum)计算;另一种是近似推导,实际工程中通常用这种方法,其中最简单的是 Gibbs(Gibbs sampling)采样法.

很显然,文档中词的主题特征越高,其代表某一主题的能力就越强.利用主题模型来计算词的主题权重,首先选择训练语料,然后进行文本预处理,最后训练 LDA 模型.得到 LDA 模型之后,使用模型中各主题下词的分布来计算词的主题特征值:一种方法是假定每个词只能代表一个主题,取模型中各主题下权重高的 TOP K 个词作为该主题的词;另一种方法假定主题区分度大的词应该是那些在某个主题下的权重高、而在其他主题中出现频率少的词语,每个词都只是代表其最能代表的那个主题.计算出词的主题特征值之后,即可以作为关键词,也可以将词的主题特征值作为其他自动关键词抽取方法中词的一个特征使用.

主题模型是语义挖掘的核心,LDA 主题模型是一种比较有代表性的模型.在主题模型中,主题表示一个方面、一个概念,表现为相关词的集合,是这些词的条件概率.主题模型的优点是[33,34]:

(1) 可以获得文本语义相似性的关系.根据主题模型可以得到主题的概率分布,可以通过概率分布计算文本之间的相似度;

(2) 可以解决多义词的问题.例如,“苹果”是一种水果,但也可能是手机或公司名.通过主题模型的 P(词语|主题),得出苹果与哪些主题相关,从生成模型层次看,可以得出其所属的主题下的其他词语之间的相似度;

(3) 可以去除文档中噪音的影响.在主题分布中,一般文本的噪声往往在次要的主题中,因此,可以对主题进行排序去噪;

(4) 无监督、完全自动化.主题模型无需人工标注,可以直接通过模型得到概率分布;

(5) 语言无关.主题模型对语言没有限制,任何语言都可以通过主题模型获得主题分布.

主题模型诞生的时间不长,但基于主体模型的应用越来越广泛,在自动关键词抽取过程中,主题模型也发挥了巨大的作用.Pu[35]提出了 TDCS(topic distilling with compressive sensing)模型,利用无监督方法和迭代法对少量文档关键词中隐含的主题进行建模和分析,取得了较好的效果.Siu[36]通过训练 HMM 语音模型发现主题信息和关键词,系统在测试语料上取得了理想的效果.Song[37]利用主题模型训练的结果作为关键词量化的主要准则,获取文档关键词.Wei[38]将 LDA 模型应用于蒙古历史文献的研究,结合特定语言和领域,实现了自动关键词抽取,成果明显.

3.2.3 基于网络图的方法

复杂网络的研究起源于 20 世纪末两篇开创性的论文[39,40],近 20 年来取得了巨大的成就.复杂网络的建模采用图论的方法,表示成 $G=(V,E,W)$,其中,V 代表节点,E 代表边,W 代表权重.Cancho 和 Sole[41]第 1 次通过研究表明:人类语言也是一种复杂网络,具有复杂网络的小世界特性与无标度特性.定义单词为网络的节点,词与词之间的共现关系为边.本文将其称为语言网络图,有向语言网络图反映了元素之间的某种关系,无向语言网络图不考虑这种关系,加权语言网络图量化元素之间的联系强度,无权语言网络图不考虑联系强度,实际应用中的语言网络图可能是 4 种形式的组合.

基于复杂网络的中文文档关键词抽取是一种无监督方法,算法首先要构建文档的语言网络图,然后对语言网络图进行分析,在整个语言网络图上寻找起重要作用和中心作用的词或短语,将这些词或短语抽取出来作为关键词.根据特征词之间的连接方式,语言网络图的主要形式如图 2 所示.

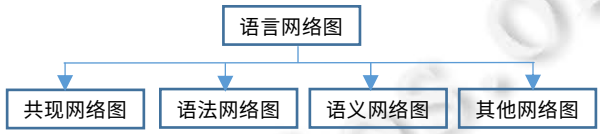

Fig.2 Main mode of language network graph

图 2 语言网络图的主要形式

共现网络图基于这样的分析:若两个词汇在同一窗口单元(比如邻接、同一个句子、同一个段落或同一章节等)中共现,则认为它们之间有一定联系.它们之间有一条边,边上的权重 W 表示两个词汇共现程度的大小.这种图可以是无向图也可以是有向图,无向图忽略了窗口单元中共现词汇的词序,有向图则考虑了词汇间的词序.

语法网络图建立在这样一种认识基础上:语法是一种稳定的连接关系,文档是由词汇按一定的语法规则组合起来的.比如,依存语法(dependency grammar)是一种充分利用句子中词汇信息的语法体系,句子中不同成分之间是不平等的,存在着支配与被支配、从属与被从属的关系.在句子中,如果词汇 A 依存于词汇 B,则在网络图中,词汇 A 指向词汇 B.这样,基于依存语法就可以构建一个有向图,节点之间的关系是它们语法上的依存关系.

Quillian[42]提出了语义网络的思想,语义网络图的理论基础是:词汇选择最根本原因在于,这些词汇本身具有的含义能够表达期望的内容.因此,文本中词汇最根本的联系在于语义层,可以通过词汇之间的语义联系将词汇进行连接构建文本的语义网络,节点是语义或概念,连接为节点间的语义关系,如上下位关系、部分-整体关系

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	$C_{EC_i}=\frac{1}{\lambda}\sum_j a_{ij}x_{ij}$	节点的分值是与它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	$CC_i=\frac{2\lvert e_{jk}\rvert}{k_i(k_i-1)}$	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,抽取出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):

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

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):

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

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,目的是识别出巨量数据中有效的、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%.Bougoui[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.好的关键词除了要与文档相关外,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):

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

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

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

$$F=\frac{2PR}{P+R} \tag{11}$$

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,∂为调节参数,当参数∂=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,特别是自动关键短语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) 关键短语的组合问题.关键短语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短语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,主流的自动关键短语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,有效组合关键短语,实现自动关键短语的抽取.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热潮,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: КовтунН 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 Reserarch 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, Pazienza 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 HCII. Swizerland: 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] 马宏炜,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提取算法研究.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.