

基于情感词典的引文文本情感识别研究*

左任銍^{1,2} 唐振华^{1,2} 黄晓^{1,2} 吴江^{1,2,3}

(1. 武汉大学信息管理学院, 武汉 430072; 2. 武汉大学电子商务研究与发展中心, 武汉 430072;

3. 武汉大学信息资源研究中心, 武汉 430072)

摘要: 基于情感分析的引证行为研究通过揭示科研论文引用情感的普遍规律, 进而探究引证行为的规律。本研究以OpinionFinder英文情感词典为基础, 借助SO-PMI算法计算引文文本中词语的点互信息值来判断词语之间的语义相似度, 从而构建了优化的引文领域情感词典。本研究从PLOS ONE期刊数据库收集了信息检索领域1 045篇文献的65 976处引用进行实证分析。结果表明, 这些引用中有84.02%为中性引用, 13.11%为正向引用, 2.87%为负向引用。本研究可以为其他引文领域情感词典的构建, 以及在更大规模引文数据中的应用提供借鉴。

关键词: 引文分析; 情感识别; 情感词典; SO-PMI; OpinionFinder

中图分类号: G254 **DOI:** 10.3772/j.issn.1673-2286.2022.02.002

引文格式: 左任銍, 唐振华, 黄晓, 等. 基于情感词典的引文文本情感识别研究[J]. 数字图书馆论坛, 2022(2): 10-17.

自引文分析提出到现在, 引文分析在理论和实践上都取得了广泛的发展, 在面向科技创新的战略情报研究和服务工作中发挥着重要作用^[1]。基于引文分析得到的论文文献被引次数、H指数和期刊影响因子等是目前学术影响力评价的主要指标^[2]。早期的引文分析主要集中在对引文著录的分析, 以引文的被引次数为基础。这种方法将不同引文简单地等同起来, 受限于全文数据的获取, 忽视了引文的引用位置、引用情感、引用功能、引用动机等信息, 从而忽略了引文在论文中的不同表现, 存在一定的局限性^[3]。

随着全文数据库和文本挖掘相关技术的发展, 引文内容分析得以实现。引文内容分析是对引用句和引文上下文的内容分析。对引文进行内容分析能够深入挖掘论文与引文之间的语义关联, 探究引文动机、引文功能、引文情感、引文重要性等影响因素, 从而进一步揭示引证规律^[4]。引文情感分析是引文内容分析的范畴, 也是当下引文分析的热点。其中, 许多学者对引文情感类别

进行梳理, 并将引文情感分为正面引用与负面引用^[5-6]或正面引用、负面引用、中性引用^[7], 也有将引用关系分为积极、消极、中立、无关^[8]的。

业界已经构建了不少权威的中英文情感词典, 并且许多学者在当前情感词典的基础上通过人工判断、自动化分类方法等方式构建了引文领域的情感词典^[7,9]。随着自动化分类方法的迭代, 构建引文领域情感词典的方法也需要不断更新。实际上, 这些新的自动化分类方法已经被学者在其他领域的情感词典中广泛应用, 如情感倾向点互信息算法 (Semantic Orientation-Point Mutual Information, SO-PMI)^[10-12]。因此, 引文领域的情感词典构建需要与时俱进。鉴于此, 本文借助当前在其他领域情感词典构建中常用的SO-PMI算法构建引文文本情感词典, 再利用该情感词典对这些引文文本进行情感识别并统计其结果。本文提出的引文文本情感词典构建方法与实证分析能够为其他学者在探究引证行为时提供借鉴, 从而揭示科研论文引用情感的普遍规律。

* 本研究得到国家教育部哲学社会科学研究重大课题攻关项目“网络环境下大数据新动能机制研究” (项目编号: 20JZD024) 资助。

1 相关研究

1.1 引文情感分析

引证行为是一种普遍存在于论文发表与科学研究过程中的行为现象。一些学者发现引证行为具有多种引用动机。Garfield^[13]指出有多种动机存在于论文引用中,并将这些动机分为15种,为之后的研究提供了指导。Willett^[14]研读了10篇论文并选择10个重要的引用动机对引文进行分类,然后比较了作者的引用动机和读者判断的引用动机。邱均平等^[15]将引用动机分为内在引用动机和外在引用动机两大类共5种引用动机,对引用动机的相互影响关系做了研究。后来,针对引文功能,尤其是引文情感分析,研究者开展了一系列重要实证研究。20世纪70年代起,不断有学者研究引文情感的分类框架。1975年,Chubin等^[16]采用分类器对引文情感进行分类时,将分类的最顶层设置为正面引用和负面引用。Teufel等^[17]将引文情感分为对引文的赞同、客观陈述引文、指出引文不足、与引文比较4类。Li等^[18]梳理区分了引文的12个功能以及3个引文情感极性。陆伟等^[4]对已有的引文内容标注体系进行了总结梳理,将引文内容分析划分为引文功能、引文重要性、引文情感倾向、引用动机4个维度,同时将引文的情感类别归为3类(正向引用、中性引用、负向引用)。引文情感分析逐渐从引文内容分析中脱颖而出,成为引证行为研究中重要组成部分。

1.2 基于情感词典的文本情感分析

基于情感词典的文本情感分析将文本看作众多词语的集合,即词袋^[19]。该方法首先根据现有的语料库和基础情感词典构建领域情感词典,然后对文本进行分词处理,最后使用构建的领域情感词典对文本中的词语进行情感得分计算,通过加权汇总得出文本的情感倾向。目前使用较多的情感词典主要有哈佛大学的GI (General Inquirer) 英文情感词典^[20]、匹兹堡大学提供的OpinionFinder主观情感词典^[21]、普林斯顿大学构建的英文情感词典WordNet^[22]、中国台湾大学的中文情感极性词典NTUSD^[11-12]、中国知网的情感词典HowNet^[11-12]、大连理工大学搭建的中文情感词汇本体库^[12]等。

根据文本所属领域的不同,学者构建了不同的领

域情感词典。Yang等^[10]利用拉普拉斯平滑法改进了SO-PMI算法,并构建了一个中文酒店评论情感词典。郭顺利等^[11]以HowNet、NTUSD等情感词典为基础,利用改进的SO-PMI算法和同义词词林扩展方法判别词语的情感类别,并以实际的图书评论作为语料构建了中文图书评论情感词典,并通过与人工标注的结果对比,计算准确率、召回率、F1值以证明该方法的可靠性。朱晓霞等^[23]在建立主题词表与情感词表的基础上,将点互信息(Point Mutual Information, PMI)和情感词典相结合计算情感词表中各元素的情感倾向值用于情感分类。叶霞等^[12]从HowNet、NTUSD、中文情感词汇本体库选取少量种子词,通过改进的SO-PMI算法对情感候选词进行情感极性判定,从而形成领域的正负情感词典。总之,PMI算法,尤其是改进后的SO-PMI算法在许多领域词典构建中有所应用,这证明了其在领域情感词典构建中的有效性。

早期的引文情感识别主要有问卷调查和人工判断两种方法^[16,24],这些方法存在效率低、规模小、主观性强等问题。随着自然语言处理技术和情感词典的发展,自动化分类方法广泛应用于领域情感词典的构建。有很多学者利用情感词典等语料库对文献的引文内容进行自动化引文情感识别研究。刘盛博等^[7]采用前人研究中人工选取的892个线索词作为情感词典搭建了一个基于引用内容的引文评价平台,通过句子主语识别的方式判断不同主语类型引用文献的情感表达差异。廖君华等^[9]结合中国知网HowNet情感词典和TF-IDF算法构建了抗衰老领域的情感词典并对PubMed数据库中的1135篇抗衰老领域论文进行了引文情感分析。徐琳宏等^[25]在分析引文情感表达方式的基础上提出了一套适用于引文情感表示的标注体系,并采用人机结合的标注策略与完善的引文标注系统,构建了规模较大的中文文献的引文情感语料库。Catalini等^[26]以1998—2007年免疫学杂志*Journal of Immunology*发表的15731篇论文为样本,研究了负向引用的特点和作用。

总的来说,对于引文情感这一重要的引文内容分析过程,已经有许多学者借助情感词典进行引文情感识别分析,但是仍然需要与时俱进,借鉴当前在领域情感词典构建应用广泛的SO-PMI算法相关研究以优化引文文本情感词典的构建方法。实际上,自动化情感分析方法还包括以机器学习和深度学习为基础的研究方法,此类方法依赖于标注好的训练集数据。由于引文情感自身的特点,即中性情感占绝大部分、正向情感次之、负向

情感极少,在标注训练数据集时容易造成数据集不平衡。对于机器学习或深度学习而言,如何处理不平衡的数据集是一个挑战^[27]。相比之下,基于情感词典的情感分析方法不需要大量标注,可操作性更强。鉴于此,本文采用基于情感词典的引文文本情感识别研究。

2 引文文本情感词典构建

2.1 SO-PMI算法与OpinionFinder情感词典

首先,构建情感词典需要计算词语之间的语义相似度,本研究采用的计算方法为SO-PMI算法;其次,大规模的引文情感分析需要借助全文数据。目前知名的全文数据库有生物医学领域的PubMed数据库、开放获取出版物PLOS ONE和Elsevier ConSyn数据平台,它们都是英文数据库,因此本文选择的是知名的英文情感词典OpinionFinder。

(1) SO-PMI算法。PMI是一种用于计算词语间的语义相似度的算法^[28]。PMI的基本思想是计算两个词语在文本中同时出现的概率,如果概率越大,其相关性就越紧密,关联度越高。SO-PMI算法引入PMI方法用于计算词语的语义相似度从而判断其情感倾向,从而达到捕获情感词的目的^[29]。首先分别选用一组褒义词和一组贬义词作为基准词,然后将待分类的词语分别与褒义词词组和贬义词词组中的每一个词计算PMI值并求和,最后用与褒义词词组计算得到的求和值减去与

贬义词词组计算得到的求和值得到一个差值,该差值即该词语的SO-PMI值。若SO-PMI值大于0,则判断这个词语的情感倾向为正向;若SO-PMI值小于0,则判断这个词语的情感倾向为负向;若SO-PMI值等于0,则判断这个词语的情感倾向为中性。

(2) OpinionFinder情感词典。OpinionFinder是由匹兹堡大学、康奈尔大学和犹他大学的研究人员联合开发的、用于处理文档并自动识别主观句子以及句子内主观性词语的系统。OpinionFinder情感词典是这套系统的基础,它包含8 219个英文词汇,并针对每个词汇的词性、是否为词干、情感极性以及词汇类别进行标记。其中情感极性分为正向、负向、既可正向又可负向、中性4类,词汇类别分为强主观词和弱主观词两种。强主观词是指该词汇在大多数语境下表现出明显的情感极性,弱主观词是指该词汇仅在一定的语境下表现出明显的情感极性。

2.2 基于情感词典的引文情感识别流程

引文情感识别与其他领域情感识别流程一致,重点是引文领域情感词典的构建。通过情感词典与引文文本中出现的词语获得种子词后,借助SO-PMI算法对种子词汇进行扩展,形成引文领域的情感词典,再通过该情感词典对实际的引文内容进行实证分析。基于情感词典的引文情感识别流程如图1所示。首先要获取引文内容数据。当前的全文数据库采用XML格式存储文献

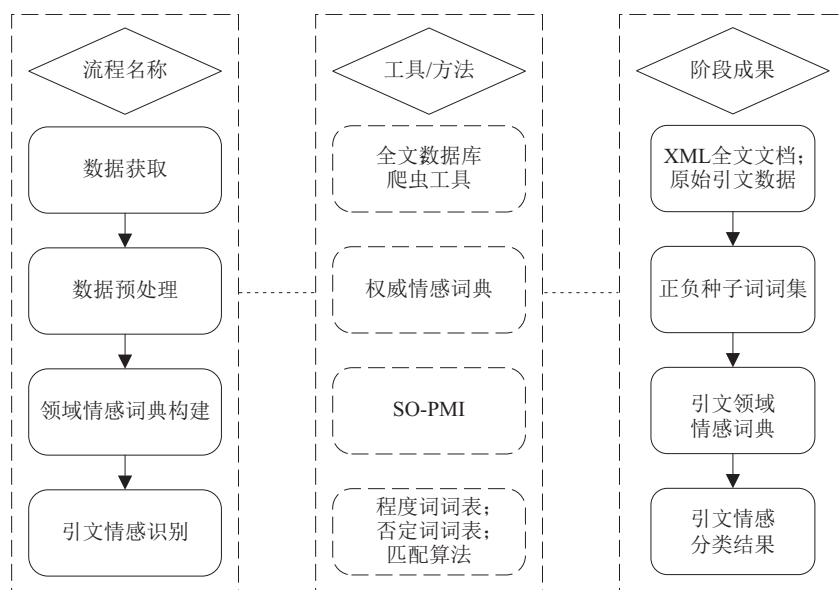


图1 引文情感识别一般流程

内容, 因此引文内容数据获取需要通过各种爬虫工具获取XML全文文档, 然后对XML文档进行解析, 获取每一处引文内容。其次是数据预处理。数据预处理主要包括分词、去除停用词等, 这些过程能够有效地提高计算效率, 保证结果的准确率。再次是构建引文领域情感词典。①要以一些已经标注好情感极性、通用的情感词汇作为种子词, 这些种子词来源于权威的情感词典; ②利用SO-PMI算法计算引文文本词语与种子词的差值, 判断引文文本词汇的情感倾向; ③以SO-PMI值作为这些词的情感权重, 构建以这些种子词为延伸的引文领域情感词典。最后是引文文本情感识别的实证分析。利用构建的引文领域情感词典和相应的程度词、否定词词典对引文数据进行加权计算, 可以得到每一处引文文本的情感分类, 实现引文文本的情感分析。

实际计算中, 引文文本数据量大且其中包括很多词频较低的专业词汇, 这些专业词汇不具有主观感情色彩。因此为了提高计算效率和准确性, 本文只计算词频大于10的引文文本词汇的SO-PMI值。同时, 由于浮点数计算精度存在的误差在多级累加后会放大, 因此需要为情感倾向判别设置一个阈值, 在观察部分典型中性词汇的SO-PMI值后, 本文将情感倾向判别的阈值设为10, 即当该词语的SO-PMI值大于10, 则该词的情感倾向为正向; 当SO-PMI值介于-10与10之间, 则该词的情感倾向为中性; 若SO-PMI值小于-10, 则该词的情感倾向为负向。

为了解决引文文本词汇与情感种子词汇在引文文本中共同出现次数为0这一问题, 本文在计算两个词汇共同出现的概率时引入拉普拉斯平滑计算改进SO-PMI算法^[30]。拉普拉斯平滑是解决零概率问题常用方法之一, 其假定样本足够大的时候, 对每个分量的计数加1, 其造成的估计概率变化可以忽略不计, 但可以方便有效地避免零概率问题。最终得到改进后的SO-PMI值计算方法如公式(1)所示。

$$SO-PMI(w1, w2) = \log_2 \left(\frac{q^2 [count(w1, w2) + 1]}{(q+2) \times count(w1) \times count(w2)} \right) \quad (1)$$

其中, q 为所有的引文数量, $count(w1)$ 为引文文本词汇 $w1$ 出现的次数, $count(w2)$ 是情感种子词汇 $w2$ 出现的次数, $count(w1, w2)$ 为词汇 $w1$ 和 $w2$ 共同出现的次数。

3 引文文本情感识别实证分析

3.1 引文文本数据获取与预处理

本研究的引文文本数据来自PLOS ONE期刊全文数据库。PLOS ONE是美国公共科学图书馆(PLOS)创办的PLOS系列期刊之一, 是全球知名的开放获取刊物。在获得该期刊中XML格式的文献数据后, 本研究通过数据解析获得引文文本, 紧接着对引文文本进行分词、去停用词、情感种子词集获取等数据预处理操作。

在PLOS ONE全文数据库网页中以“information retrieval”为检索词进行检索, 得到1 207个搜索结果。剔除没有XML格式内容的文献以及无关文献, 利用爬虫工具获取该领域XML格式的全文数据, 共获得1 045篇文档。在获得XML格式的全文数据后, 对其中的引用进行抽取。在多数情况下, 引用所在句子的边界就是引文上下文的边界, 即引用为引文句子^[35]。但在个别情况下, 相互关联的术语也可以出现在相邻的句子里, 即引文上下文的边界超出了引用所在的句子。因此, 也有研究将引用所在句子及其前后共3个句子或引用所在的段落一同作为引文上下文^[9,31]。综合考虑, 本文抽取引文标识前后两句话作为引文文本数据, 即引用节点前后的两个节点与引用标识节点本身形成一处引用, 共得到65 976处引用。

本研究的数据预处理主要包括两部分: 一是对获取的引文文本进行分词和去停用词处理; 二是将OpinionFinder情感词典分成正负情感种子词集。首先, 本文采用Python第三方库中的nltk库对引文文本进行分词处理。该库中的word_tokenize方法能够对文本进行分词处理。同时, nltk库中包含一个英文的停用词表, 可用于剔除停用词。其次, 针对正负情感种子词集的获取。将OpinionFinder情感词典中所有情感类型的词汇汇集到一个文本文件中, 按情感极性的不同分类, 得到2 718个正向情感词词集、4 910个负向情感词词集、570个中性情感词词集以及21个既可正向又可负向的情感词词集; 然后选择正向情感词词集和负向情感词词集中的“强主观”类型的词汇, 并将这些词汇与引文文本进行匹配, 引文文本出现频率大于10的“强主观”词汇被作为种子词, 得到正向种子词226个、负向种子词251个。为确保正负种子词的数量不对SO-PMI的计算产生较大影响, 本文人工删除了频率较低的种子词, 最终得到正向种子词200个、负向种子词200个。

3.2 引文文本情感识别分析

在得到情感种子词后,可以根据引文文本对情感词汇进行扩充,形成引文领域情感词典。结合选取的程度词和否定词词表,可以自动化计算引文情感得分,从而评估引用的情感倾向。

3.2.1 引文文本情感词、程度词与否定词选取

根据前文提到的构建方法,对正负种子词词集进行扩充得到引文文本情感词典。本研究选取引用中的所有词汇,首先剔除了词频低于10的引文词汇和数字形式的词汇,然后计算剩下词汇的SO-PMI值,并将SO-PMI值作为该情感词汇的情感倾向权重,如support的权重值为16.56、lack的权重值为-21.96、retrieval的权重值为6.75。因此,support为正向词汇,lack为负向词汇,retrieval为中性词汇。据此构建的引文文本情感词典具体信息如表1所示,其中正负向种子词个数分别为200,扩充后的正向词汇有1 969个,负向词汇2 116个,中性词汇9 422个。

表1 引文领域情感词典统计信息

情感类别	扩充前词汇数量/个	扩充后词汇数量/个	词汇示例
正向词汇	200	1 969	support, novel, respect, right, explicit……
负向词汇	200	2 116	lack, stress, argue, exploit, ambiguous……
中性词汇	0	9 422	retrieval, find, account, agency, mark……

在构建好引文文本情感词典后,还需选取程度词词表和否定词词表。本研究的程度词词表来源于中国知网HowNet情感词典中的英文程度词表。HowNet情感词典收录了170个程度词汇,分为most、very、more、-ish、insufficiently、over。本文参考这种分类方法但将over类中的词汇视情况归入了most或very类中,将most类中的程度词程度级数设为5、very类中的程度词程度级数设为4、more类中的程度词程度级数设为3、-ish类中的程度词程度级数设为2、insufficiently类中的程度词程度级数设为1.5,具体情况如表2所示。另外,本文将英文中常见的否定词如no、not、nothing、nowhere等构成否定词词表。

表2 程度词词表

程度类别	程度级数	词汇数量/个	词汇示例
most	5	44	absolute, absolutely, most, awfully, extremely……
very	4	19	very, unusually, deeply, especially, really……
more	3	7	more, further, so, such, comparatively……
-ish	2	8	fairly, passably, pretty, quite, rather, some……
insufficiently	1.5	6	just, light, merely, relatively, slightly……

3.2.2 引文文本情感识别结果分析

对引文情感类别的研究中,大多数研究将引文情感分为正向、中性、负向3类。究其原因,一方面是由于技术限制,自动化分类算法难以准确识别出过于细致的引文情感;另一方面,过于细致的引文情感分类会造成引文情感识别计算量增加,识别效率降低。因此,本研究也将引文情感类别划分为正向引用、中性引用、负向引用3类,如表3所示。

表3 引文情感分类表

引文情感类别	描述
正向引用	对所引内容的肯定或赞扬
	对所引文献中的方法、技术、数据等的应用
负向引用	明确指出引文的不足
中性引用	客观陈述前人研究工作,不含感情色彩

引文情感识别依赖于引文情感倾向计算,而引文情感的倾向计算是通过该引文所在引用的文本内容情感倾向综合计算得到,具体计算逻辑如图2所示。首先设置初始引文情感得分score为0。其次,从该引用文本内容的第一个词开始,判断该词与上一个词之间是否有程度词及否定词。如果有程度词,则将该情感词的情感权重为SO-PMI值乘以程度级数;如果有否定词,则该情感词的情感权重为SO-PMI值乘以 $(-1)^t$,t为否定词的个数。最后将该词的情感权重得分加到引用文本的情感得分中,重复该过程直到遍历该引用文本中的所有词,则可以得到该引文情感得分。若某引文的情感得分大于10,则该引文的情感倾向被归为正向类;若某引

文的情感得分小于10, 则该引文的情感倾向被归为负向类; 若某引文的情感得分介于-10与10之间, 则该引文的情感倾向得分被归为中性类。

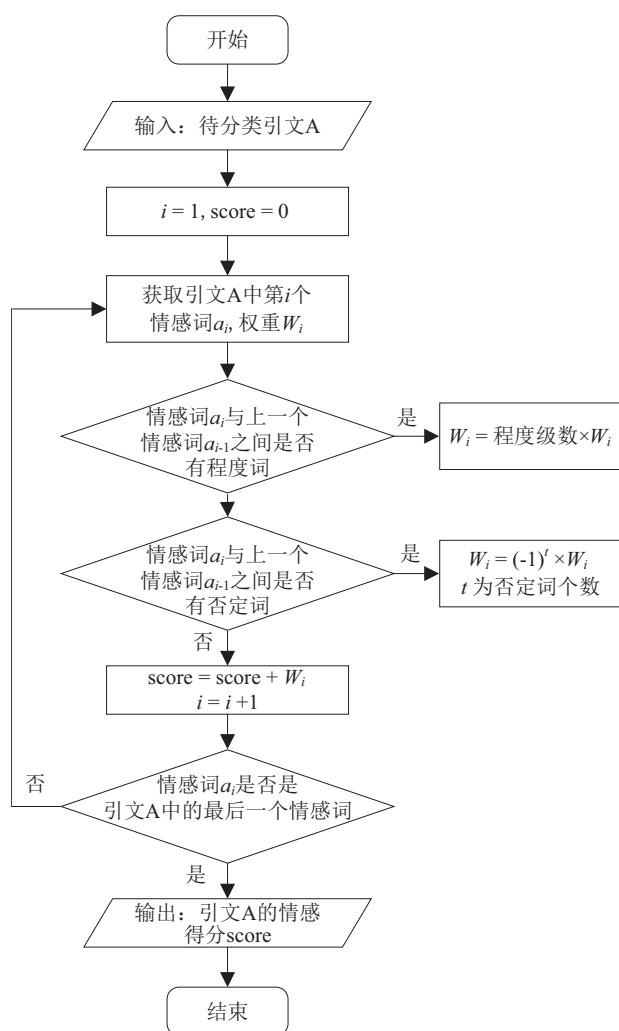


图2 引文情感识别计算

根据上述过程得到65 976处引用的情感得分, 并对结果进行统计, 如表4所示。结果显示, 中性引用占大多数, 为84.02%; 有13.11%的引用为正向引用; 只有2.87%的引用表现为负向引用。该结果与廖君华等^[9]采用TF-IDF算法结合情感词典对PubMed数据库中抗衰老领域文献进行引文情感分类的研究结果相近。在廖君华等的研究结果中, 有20.74%是正向引用, 1.43%是负向引用,

表4 引文情感倾向统计表

引用情感类别	引文数量/个	引文数量占比/%
正向引用	8 651	13.11
负向引用	1 893	2.87
中性引用	55 432	84.02

77.82%为中性引用。

由引文文本情感分类结果可以看出, 在科技文献创作中, 科研工作者更多是对前人研究工作的客观转述, 对引用内容明确表示肯定或对引用内容中的相关方法、数据延续应用的正向引用次之, 而只有少部分引用属于明确指出前文不足的负面引用。这一现象可以从多个方面进行解释。中性引用占据绝大部分的原因有两点: 第一, 科研文献的行文风格不同于日常表达, 它要求作者客观、严谨, 避免主观化、情绪化的表述, 力求专业、科学, 因此科研工作者在引用文献时更多是客观转述前人工作; 第二, 科研工作者的专业素养使其在对待研究成果上更加谨慎、冷静, 许多研究成果或多或少都会有些争议, 学者对待这些争议的态度是谨慎的, 因此在引用这些成果时不轻易表现自己的情感态度。对于正向引用的原因可能是学者在文献写作中引用了他人的研究方法和数据作为自身研究的根据, 这符合本文对正向引用的第二个描述。对于负向引用, 则是科研工作者对现有研究中缺陷的指出或对研究方法的改进等。一篇文献的发表需要经过严格的审核, 因此文献中很少有特别明显的错误, 再加上科学文献更多是立论而非驳论, 所以引文中负向引用所占比例较低。无论是正向引用、中性引用还是负向引用, 都是引证行为中的正常现象。如果说中性引用和正向引用说明科学研究的继承性, 那么负面引用的存在则说明科学研究不仅有继承性还有批判性, 科学研究的继承性和批判性共同推动着科学知识的螺旋式发展。

4 总结

随着全文数据库与自然语言处理等技术的发展, 引文分析已经进入到引文内容分析阶段, 对引用文本进行情感分析能够更好地探究学者的引证行为规律。基于情感分析的引证行为研究旨在通过对一定数量的论文进行引文情感分类来揭示科研论文引用情感的普遍规律, 进而探究引证行为的规律。然而, 目前基于情感词典的引文情感分析方法有待升级, 有必要将目前在其他领域情感词典构建应用广泛的SO-PMI算法运用在引文领域文本词典构建中, 并据此实现引文情感分析。

本文提出的基于情感词典的引文情感识别模型是新的领域情感词典构建方法在引文领域的一次尝试, 该方法具有可推广性, 后续的研究者可以将本文引文情感识别模型与方法应用到其他引文领域或者更大规

模的引文数据中加以验证。当然,本文也存在一定局限。该方法已经用于其他领域的情感词典构建,并且一些学者通过计算准确率、召回率、F1值等指标对比了该方法与人工标注情感分析的结果,证明了该方法的有效性。本文直接默认了该方法在引文领域情感词典构建中是有效的,并未进一步验证。未来的研究者可以通过人工标注的结果或者大规模的分析对比验证该方法在引文情感词典构建与情感识别分析的有效性。同时,未来研究将尝试使用不同方法构造领域情感词典,比较各方法之间的差异。

参考文献

- [1] 祝清松,冷伏海. 引文内容分析方法研究综述[J]. 情报资料工作, 2013(5): 39-43.
- [2] 李铮,邓三鸿,孔嘉,等. 学者学术影响力识别研究——基于引文全数据的视角[J]. 图书情报工作, 2020, 64(12): 87-94.
- [3] 刘盛博,丁莹,张春博. 引文分析的新阶段: 从引文著录分析到引用内容分析[J]. 图书情报知识, 2015(3): 25-34.
- [4] 陆伟,孟睿,刘兴帮. 面向引用关系的引文内容标注框架研究[J]. 中国图书馆学报, 2014, 40(6): 93-104.
- [5] MA Z, NAM J, WEIHE K. Improve sentiment analysis of citations with author modelling[C]//Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 2016: 122-127.
- [6] VEER MARTENS B V, GOODRUM A A. The diffusion of theories: a functional approach[J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 330-341.
- [7] 刘盛博,丁莹,张春博. 基于引用内容性质的引文评价研究[J]. 情报理论与实践, 2015, 38(3): 77-81.
- [8] ATHAR A. Sentiment analysis of citations using sentence structure-based features[C]//Proceedings of the ACL 2011 student session. 2011: 81-87.
- [9] 廖君华,刘自强,白如江,等. 基于引文内容分析的引用情感识别研究[J]. 图书情报工作, 2018, 62(15): 112-121.
- [10] YANG A M, LIN J H, ZHOU Y M, et al. Research on building a Chinese sentiment lexicon based on SO-PMI[C]//Applied Mechanics and Materials. Trans Tech Publications Ltd, 2013, 263: 1688-1693.
- [11] 郭顺利,张向先. 面向中文图书评论的情感词典构建方法研究[J]. 现代图书情报技术, 2016(2): 67-74.
- [12] 叶霞,曹军博,许飞翔,等. 中文领域情感词典自适应学习方法[J]. 计算机工程与设计, 2020, 41(8): 2231-2237.
- [13] GARFIELD E. Can citation indexing be automated?[J]. NBS Special Publication, 1965(269): 189-192.
- [14] WILLETT P. Readers' perceptions of authors' citation behavior[J]. Journal of Documentation, 2013, 69(1): 145-156.
- [15] 邱均平,陈晓宇,何文静. 科研人员论文引用动机及相互影响关系研究[J]. 图书情报工作, 2015, 59(9): 36-44.
- [16] CHUBIN D E, MOITRA S D. Content analysis of references: Adjunct or alternative to citation counting?[J]. Social Studies of Science, 1975, 5(4): 423-441.
- [17] LI X, HE Y, MEYERS A, et al. Towards fine-grained citation function classification[C]//Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013. 2013: 402-407.
- [18] YOUSIF A, NIU Z, NYAMAWA A S, et al. Improving citation sentiment and purpose classification using hybrid deep neural network model[C]//International Conference on Advanced Intelligent Systems and Informatics. Springer, Cham, 2018: 327-336.
- [19] 张辉,刘奕群,马少平. 文本情感分类中生成式情感模型的发展[J]. 计算机应用研究, 2014, 31(12): 3521-3526.
- [20] HARTMAN J, STONE P J, DUNPHY D C, et al. The general inquirer: a computer approach to content analysis[J]. American Sociological Review, 1966, 32(5): 859.
- [21] PIAO S, ANANIADOU S, TSURUOKA Y, et al. Mining opinion polarity relations of citations[C]//International Workshop on Computational Semantics (IWCS). 2007: 366-371.
- [22] ESULI A, SEBASTIANI F. Sentiwordnet: A publicly available lexical resource for opinion mining[C]//Proceedings of the fifth international conference on language resources and evaluation (LREC'06). 2006.
- [23] 朱晓霞,宋嘉欣,孟建芳. 基于主题—情感挖掘模型的微博评论情感分类研究[J]. 情报理论与实践, 2019, 42(5): 159-164.
- [24] MORAVCSIK M J, MURUGESAN P. Some results on the function and quality of citations[J]. Social Studies of Science, 1975, 5(1): 86-92.
- [25] 徐琳宏,丁莹,陈娜,等. 中文文献引文情感语料库构建[J]. 情报学报, 2020, 39(1): 25-37.
- [26] CATALINI C, LACETERA N, OETTL A. The incidence and role of negative citations in science[J]. Proceedings of the

- National Academy of Sciences, 2015, 112 (45): 13823-13826. 427-432.
- [27] 徐玲玲, 迟冬祥. 面向不平衡数据集的机器学习分类策略 [J]. 计算机工程与应用, 2020, 56 (24): 12-27.
- [28] KHAN F H, QAMAR U, BASHIR S. SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection [J]. Applied Soft Computing, 2016 (39): 140-153.
- [29] TABOADA M, ANTHONY C, VOLL K D. Methods for Creating Semantic Orientation Dictionaries [C] //LREC. 2006: 427-432.
- [30] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [EB/OL]. [2022-01-01]. <https://arxiv.org/ftp/cs/papers/0212/0212032.pdf>.
- [31] LIU S, CHEN C. The differences between latent topics in abstracts and citation contexts of citing papers [J]. Journal of the American Society for Information Science and Technology, 2013, 64 (3): 627-639.

作者简介

左任銓, 男, 1992年生, 博士研究生, 研究方向: 文本挖掘、社会网络计算、仿真模拟。

唐振華, 男, 1997年生, 硕士研究生, 研究方向: 科学计量与科学评价。

黃曉, 男, 1992年生, 博士研究生, 通信作者, 研究方向: 在线用户行为、科学评价, E-mail: xiaoh@whu.edu.cn。

吳江, 男, 1978年生, 博士, 教授, 博士生导师, 研究方向: 社会网络计算、社会仿真模拟。

Research on Emotion Recognition of Citation Text Based on Sentiment Dictionary

ZUO RenXian^{1,2} TANG ZhenHua^{1,2} HUANG Xiao^{1,2} WU Jiang^{1,2,3}

(1. School of Information Management, Wuhan University, Wuhan 430072, P. R. China; 2. Center for E-commerce Research and Development, Wuhan University, Wuhan 430072, P. R. China; 3. Center for Studies of Information Resources, Wuhan University, Wuhan 430072, P. R. China)

Abstract: Research on citation behavior based on emotion analysis explores the general law of citation emotion in scientific research papers, and further reveals the law of citation behavior. Based on the English Sentiment Dictionary-OpinionFinder and the SO-PMI, this study calculates the value of the point mutual information of words in the citation text to judge the semantic similarity between words, so as to construct an optimized emotion dictionary in the citation field. Our research collected 65 976 citations of 1 045 literatures in the field of information retrieval obtained from PLOS ONE journal database for empirical analysis. The results show that 84.02% are neutral citations, 13.11% are positive citations, and 2.87% are negative citations. This study can provide reference for the construction of emotional dictionary in other citation fields and the application of larger citation data.

Keywords: Citation Analysis; Sentiment Recognition; Sentiment Dictionary; SO-PMI; OpinionFinder

(收稿日期: 2022-01-17)