DOI:10.13530/j.cnki.jlis.140029

面向引用关系的引文内容标注框架研究*

陆 伟 孟 睿 刘兴帮

摘 要 引文内容分析能够帮助揭示文献引用关系的深层语义内涵。本文梳理了目前已有的引文内容标注体系,归纳出构建引文分类体系的三个主要维度,即引文功能,引文重要性,情感倾向。以支持文献引用关系分析为目标,针对引文内容分析设计出一个引文内容标注框架,其中包括揭示引文关系抽象性质的引文分类标注体系,描述被引文献具体内容的引用对象标注体系,以及记录引文客观特征的引文属性标注体系。具体的标注实验体现了该标注框架的可用性。图1。表6。参考文献56。

关键词 引文内容分析 引文功能 引文分类 引文关系 文献语义挖掘

分类号 G353.4

A Deep Scientific Literature Mining-Oriented Framework for Citation Content Annotation

Lu Wei, Meng Rui & Liu Xingbang

ABSTRACT Citation content analysis can reveal the connotative meaning of citations. This article organizes existing schemes of citation content analysis. Three main dimensions of designing citation classification scheme have been concluded. In order to support automatic scientific literature mining, this paper proposes a framework for citation content annotation, which contains a citation classification annotation scheme indicating literature's abstract attributes, a cited object annotation scheme describing literature's specific content, and a citation attribute annotation scheme recording citation's objective features. A pilot annotation shows the usability of proposed annotation framework. 1 fig. 6 tabs. 56 refs.

KEY WORDS Citation content analysis. Citation function. Citation classification. Citation relation. Scientific literature mining.

0 引言

引文在科研文献中十分普遍,扮演着重要的角色。作者在撰写的论文中引用他人的研究成果,一方面体现了作者对该成果的重视和兴趣,另一方面也在自己的研究成果中融入了他人的思想和方法[1]。鉴于引文的重要价值,相关的理论和应用得到广泛研究,其中一个重要方向是通过引文来度量学术成果的学术影响力,如文献被引频次及由其衍生的H指数[2]、期刊影响因子[3]等是目前主要的学术影响力评价工具。传统的引文分析将文献与文献之间的引用关系抽象为简单平等的线性关系,通过一篇文章被引用的频次来表示该文章学术影响力的高低。借助于这种对文献之间关系的简单表

* 本文系教育部人文社科基地重大项目"面向细粒度的网络信息检索模型及框架构建研究"(项目编号:10JJD630014)和国家自然科学基金面上项目"基于语言模型的通用实体检索建模及框架实现研究"(项目编号:71173164)的研究成果之一。

通讯作者:陆伟,Email:reedwhu@gmail.com

述,大规模的文献引用网络构建和文献评价成为可能。其实,作者在撰写文章过程中引用参考文献是一个复杂的过程。传统的引文分析方法只能告诉读者哪两篇文章之间具有引用关系,不能说明被引成果对于施引文献的具体贡献以及重要性,这种对引用关系的简化处理无法展示科研文献网络的真实情景。

事实上,引文不仅仅为相关文献建立了联系,通过深入分析引文的上下文内容可以从语义角度对文献间的引用关系进行理解。结合定性和定量方法对引文内容进行研究,描绘出施引文献和被引文献之间具体的情境关系,可以弥补传统引文分析中忽视引文语义细节的不足[4]。这些语义细节包括被引文献对于施引文献的作用、作者引用时的情感倾向等。大量基于引文内容的研究工作陆续开展。如 Small[5] 评估了使用文献被引数量评价其学术影响力的可靠性。Oppenheim 和 Renn[6]、McCain 和 Turner[7] 等结合引文内容研究了高被引文献所具有的被引特征。Hanney 等[8] 利用引文分类评估卫生领域研究成果的影响力,包括对这些成果在跨代引用过程中的影响力演变进行跟踪。利用引文内容还可以对引文索引技术进行改进。Garfield[9] 对作者的引用意图进行归类,探讨引文索引自动化构建的可行性。类似的,Lipetz[10] 和 Finney[11] 通过引用分类提高引文索引中文献之间的区分度。

在引文自动化处理方面,谷歌学术、中国知网等已有文献检索系统普遍只是对文章中的参考文献进行抽取,并在此基础上构建由单一引用关系组成的引文网络,缺乏语义层面的引文关系分析。随着自然语言处理和文本挖掘技术的成熟,从大量科研文本中自动化抽取语义信息成为可能,诸多学者在引文功能分类[12-14]、引文情感识别[15-16]、引文上下文抽取[17] 等方面取得了初步成果,为实现更深层次的科研文献语义信息抽取提供了良好基础。

为更好地支持文献语义关系挖掘,将自然语言处理、机器学习技术引入引文内容分析,需要一个系统的引文内容标注框架。本文通过对相关研究进行全面的调研和梳理,总结了目前引文内容分析研究中标注体系的优势和不足,认为已有分类体系缺乏对被引文献重要性及引用对象的重视。本文针对上述两点不足提出了一套引文内容标注框架:一方面结合参考文献对施引文献的重要性,对以往相对独立的功能类目进行组织;另一方面在框架中加入对引用对象及其类型的标注。同时本文利用提出的框架进行了初步的标注实验,验证了引文内容标注框架的可用性。

1 相关研究

引文在科研文献中十分普遍,体现了后来研究者对先前研究者成果的借鉴和认可,也为文献和文献之间建立了一条知识传递的纽带。传统的引文分析往往将文献与文献之间的引用关系简化为平等的线性关系,通过文章的被引数量来度量学术成果的影响力。然而随着研究的深入,传统的引文分析方法受到一些学者的质疑[18-19]:评价学术成果影响力不能单纯依靠文章被引数量这种简单的定量指标,简化的引文关系无法表现引用行为中的复杂意义。学者开始尝试结合语言学的研究方法,将研究深入到引文内容以解决上述问题[5,19]。引文内容分析,即基于引用句及其上下文内容对引文的性质进行分析,它通过引文内容的语法和语义特征,将被引文献对施引文献支持的作用和程度进行区分,深入探讨施引文献与被引文献之间的语义关联,进而揭示引文行为的本质。

由于引文内容的形态和特征各异,因此在实施引文内容分析之前,需要有针对性地制订一个引文分类体系,然后按照引文的对应特征对引文进行归类和分析。国外相关研究始于 20 世纪 60 年代。1965 年,Garfield[9] 提出了 15 种作者的引用动机,以探讨构建自动化引文索引的可行性。Lipetz[10] 定义了 4 组(施引文献的原创贡献、非原创贡献、一致性关系、施引文献对

被引文献的情感）共29项特征，以提高学术引文索引中不同文献的区分度。Finney[11]也以此为研究出发点，结合特征词和引文位置两个主要特征，设计了一个7类的分类体系。Herlach[20]创建了一个分类体系试图描述文献之间所有可能的关系，并发现如果一篇参考文献在文章中被多次提及则体现了其对原文具有较高的重要性。Frost[21]对参考文献的来源和情感倾向进行组合分类，研究作者的引用行为是否受到客观环境的影响。Oppenheim和Renn[6]为研究高被引文献被引用的原因，定义了一个包含7个类别的分类体系，包含背景、描述、对比、否定等类别，相比先前的体系更加清晰可用。Spiegel-Rösing[22]研究了文献中引文的功能分布，构建的13类分类体系具有良好的操作性。Moravcsik和Murugesan[23]尝试比较不同参考文献的质量，构建了一个基于二元选择的标注体系，很好地增强了标注者的判断力。Chubin和Moitra[24]将Moravcsik和Murugesan的方案合并为6个类目，将这6个类目按照从肯定到否定、从本质到补充、从基本到附属的角度进行了组织。总体而言，学者根据不同的研究目的提出了不同的分类体系，但也造成了一定的混乱。主要问题如Swales[25]指出，“大多数分类标注体系的跨领域适用性较差，并且需要标注者具备一定领域知识才能掌握”。Zhang等[4]构建了一个引文内容分析框架，试图解决目前引文分析研究中数值特征、语言特征以及社会文化特征研究之间分离的现状，为进一步引文分析研究提供了良好基础。

尽管许多学者试图通过引文内容对传统引文分析进行拓展，然而耗时的手工标注、格式化数据获取困难等问题阻碍其进一步发展和应用。如今，随着文献存储和检索技术的长足进步，学者开始尝试从自动化角度对引文内容进行研究。Garzone[12,26]较早尝试运用自动化技术对引文功能进行分类，列出35个类别并手工设定对应的规则，但该方法难以达到很好的召回率。Nanba和Okumura[27]设定了一个只有3类（基于、对比、其他）的简单体系以提高分类的准确性。Teufel[13,28]修改了Spiegel-Rösing[22]的分类体系，并使用机器学习方法改善引文功能分类的效果。Radoulov[14]修改了Garzone的分类方案，同时在体系中加入了引用对象的类型标注。此外，Iorio等[29]结合本体概念构建一个较为全面的引文分类体系。Xu等[30]创新性地将引文网络性质作为特征以期提高引文分类的准确度。除去引文功能分类之外，针对引文内容的自动化研究还包括自动摘要[31-33]，信息检索[34-38]，引文上下文识别[17,39]等方面。运用自动化技术对引文内容进行分析得到广泛认可，也将有更多相关成果出现。

国内也有一定数量的研究成果深入引文内容研究引文的相关性质。崔红[40]概括了11种引用动机，通过直接调查的方式获取数据，对科研学者的引文动机进行聚类分析。叶继元等[41]对负面引用现象进行了研究。陈晓丽[42]对引文的引用方式、内容类型以及引用力度进行了较为全面的分析，为后来研究者制定引文分类体系提供了良好的支持。赵青[43]从引用性质和引用深度两个角度对引文行为进行定性分析，引用性质体现作者的引用情感，引用深度体现参考文献与施引文献研究工作之间相关性的高低。文献评价方面，胡志刚等[44]以被引文献在文章中出现的引用次数作评价指标，显示出一定的应用价值，刘盛博等[45]利用引文分类改进了传统的引文评价机制。此外，祝清松等[46]对目前引文内容分析工作进行了综述。

2 引文内容标注框架设计

Zhang等[4]在研究中指出，实施引文内容分析的主要步骤是：首先对科研文献中的引文上下文内容进行识别和提取，其次需要制订一个支持进一步分析的标注体系。然而制订一个综合全面而不琐碎复杂的标注体系并非易事。一

个组织分类合理且能够全面表示引文特征的标注框架尤为重要,本文对目前影响力较大的引文分类体系研究成果进行整理,并按照其分类所依据的维度,将已有体系归为四类。

(1)引文功能

体现被引文献在施引文献中的作用、功能,是最为主要的分类维度。代表成果有 Oppenheim 和 Renn[6]、Spiegel-Rösing[22]、Moravcsik 和 Murugesan[23]等人的研究。

(2)引文重要性(引文质量)

体现被引文献对施引文献的重要性。代表体系有 Cano[19](定义 4 个重要性等级:本质、核心、有限、外围)、Moed[47](3 个期刊间引用影响力等级)、Wan[48](5 个参考文献重要性等级)等人的研究。

(3)情感倾向

体现施引作者对被引文献成果的情感倾向。代表成果有 Athar[15](4 类情感分类:积极、消极、中立、无关)的研究。

(4)引用动机

体现施引作者引用时的具体动机。代表成果有 Brooks[49](7 个引用目的)、Vinkler[50](分为专业动机和关系动机)等人的研究。

除了单一针对其中某一个维度进行分类设计之外,部分研究还将这四个维度进行一定程度上的融合,如 Teufel[13,28]在其体系中将情感倾向与引文功能进行融合。还有部分体系按照被引文献的文献类型、文献来源、引文出现位置等维度进行划分。本文认为这些维度更多体现的是引文的客观属性,故未加入讨论。更多关于引文标注体系的信息可以参见 Liu[51]、Bornmann 和 Daniel[52]的综述成果。

先前引文内容标注体系关注的四个维度中,引文动机倾向于从施引作者的主观视角进行研究,与本文所关心的研究方向并不一致。其余三个划分维度从抽象层面分析被引文献与施引文献的联系,均揭示了引文关系的重要性,也是本文进行框架设计的主要方向。本文很大程度上受到 Small[53]成果的启发,认为引用对象对于引文分析的研究具有重要价值,然而目前的成果中很少有针对引用对象的自动化研究。此外为支持进一步的自动化分析,引文内容的客观特征属性也需要进行标注。基于上述考虑,本文提出一个全面支持引文内容分析的引文内容标注框架,主要包括:①一个揭示引文关系抽象性质的引文分类体系;②一个描述被引文献具体内容的引用对象标注体系;③一个记录引文客观特征的引文属性标注体系。

2.1 引文分类标注体系

在对已有引文分类体系的整理中,本文总结出对引文进行分类的三个主要维度:引文功能、引文重要性以及引用情感倾向。在这三个划分维度中,引文功能直观体现了参考文献在施引文献中的作用,因而在大多数分类体系中处于核心位置。引用情感倾向也是学者较为关心的维度,直接体现了作者对于被引文献工作正面或负面的情感态度。由于科研文献中语言风格多为客观中立(Athar[15]在引文情感数据集构建过程中发现只有 14%的引文内容中表达了情感倾向),除去少数按照传统情感识别思路研究引文情感的成果[15-16,54]之外,也有作者[28,45]将对施引文献有支持作用的重要引用列为"正向"引用,这种融合不同维度的方法提供了很好的思路。

引文重要性衡量一篇参考文献对于其施引文献智力支持程度的大小,能够帮助读者了解哪些被引成果在作者的研究中贡献了重要作用。但是引文的重要性大小难以界定,往往依赖标注者的主观判断。值得注意的是,引文重要性的高低与引文功能的分布体现了较高的相关性[8,19],这启发本文尝试将这两个维度进行结合。引文功能本身也能够体现参考文献对于原文工作支持程度的大小,例如,属于"基于"功能的文献比"相关研究"功能的文献对原文具有更高的重要性。将引文功能与引文重要性相结合,一方面通过标注功能减少单纯对重要性进行标注的模糊性,另一方面根据重要性对不同

的引文功能进行排序，从而突出文章中最核心的若干参考文献。

按照上述设计思路，本文制订了一个结合引文重要性的功能分类体系（见表1），以及一个独立的作者引用情感倾向分类体系。本文参考先前工作中的类目设计并在试标注中进行调整，最终确定了15个功能分类类目。本文之所以设计一个较小粒度的功能体系，是希望展现一个深入全面的引文情景，而非为增强自动识别的准确率而在设计粒度上进行折中。

表1 结合引文重要性的引文功能分类体系

重要程度	功能类目	描述
非常重要	基于	施引文献工作以参考文献为起点
重要	启发	施引文献的研究受到被引文献的启发
	拓展	施引文献拓展或者修改了被引文献中的成果
	使用	施引文献使用了被引文献中的成果
	详细引用	施引文献详细引用了被引文献中的成果
一般	比较	施引文献工作与被引文献工作形成了对比
	相似	施引文献工作与被引文献工作内容近似
	肯定	施引文献肯定被引文献中的工作
不重要	相关研究	介绍与施引文献工作相关的其他研究
	简单引用	简单地引用了被引文献中的具体内容
	相关工作之间比较	对两个或多个被引文献的工作进行比较
	未来工作	被引文献对施引文献的进一步工作有所启示
	拓展阅读	通过查看参考文献以了解更多信息
非常不重要	历史背景	与施引文献工作有关的历史信息
	无关引用	与施引文献工作不相关的引用

Wan和Liu[48]在工作中设定了5个引文重要性等级，并规定了对应等级的特征。本文在对引文功能的重要性进行设定时参考了此划分标准，将15个功能类别按照相对重要性大小划分为5个等级。该划分方法虽然不能严格地反映作者对参考文献重要性的衡量，但是十分直观地区分出不同参考文献对于原文工作支持程度的大小。重要性划分的主要依据如下。

（1）非常重要

“基于”功能说明被引文献构成了施引文献工作开展的基础和前提，对于施引文献具有不可或缺的重要意义，因而“基于”功能对应于最高的重要性等级。

（2）重要

“启发”、“拓展”等引文功能表现出施引文献在观点、研究思路等方面参考了被引文献的成果。“使用”功能则表明施引工作使用了被引文献中的方法、工具。“详细引用”往往是对上述几种引用的具体阐述。这些工作促进了施引文献工作的开展和实施，是其有机组成部分。

（3）一般

“比较”和“相似”功能，通过对比来展现被

引工作与施引工作之间的异同,从侧面体现原文研究的特点。“肯定”则是通过引用被引文献的研究结论直接支持施引工作。在论证过程中,这三类引文在一定程度上支持了施引工作,但重要程度不及之前的四类引用。

(4)不重要

对于“相关研究”、“简单引用”、“相关工作之间比较”、“未来工作”、“拓展阅读”等五类功能的引文,虽然这些工作与施引文献具有一定的相关性,但施引文献往往只会对其工作进行简略描述,其在文中存在与否并不影响文章核心的论述,因而只具备较低的重要性。

(5)非常不重要

“历史背景”和“无关提及”两类功能的引文,一般是作者出于行文需要,通过引用对应的参考文献以引入要讨论的主题。这些被引工作与施引工作的相关性十分有限,删除它们并不会影响原文工作的完整性。

对于引用情感倾向分类的设计,本文没有改动,还是沿用传统的“正向”、“负向”及“中立”分类方法。情感倾向的标注相对次要,但具有情感倾向的引文内容仍具有重要研究价值。本文只对作者明确在内容中表现出赞扬或者否定情感的引文进行标注,其余均默认标注为“中立”。

2.2 引用对象标注体系

引用对象,是指作者在引用文献时提及成果中的具体内容对象。引文功能解释了作者“为什么”引用被引文献,而引文对象则表明作者具体引用了被引文献中的“什么”。Small[53]认为学者的引用行为将概念符号与对应的参考文献建立了联系,他研究了241篇化学论文,发现93%的文章使用相同的概念符号来表达参考文献中的工作。Small的研究揭示了这样一个事实:研究者倾向于用一致的“概念符号”来描述前人的工作。这启示我们可以将引用对象看作被引工作的“链接”,识别出这些具体的对象可以更好地改进学术本体构建和引文上下文范围识别的效果。

一些引文体系在功能中区分了不同的引用对象,如在Garzone[12]的体系中“使用”对应了5种对象类型(工具、公式、方法、条件、结果分析方法),这使体系变得复杂庞大。Radoulov[14]注意到类似不足,将引文对象抽取为单独的标注条目,设计了9种引用对象类型。但其目的仅仅是改进Garzone的分类体系,并没有考虑到对具体引用对象的研究。本文参考并改进了Radoulov的方案,设计了一个包含11个类型的引文对象分类体系(见表2)。在标注过程中不仅要求标注者标注出引用对象的类型,还需要标注出具体的被引用对象名称,以进一步研究这种“符号化”现象的性质。

表2 引用对象类型及描述

对象类型	描述
概念	引用被引文献中提出的概念的定义、介绍等
方法	使用被引文献中的具体方法
模型	引用被引文献中提出的模型方案
算法	引用被引文献中的具体算法
理论	引用被引文献中的理论
应用	引用被引文献中的应用成果
工具	使用被引文献中提供的工具
数据	使用被引文献中提供的数据
公式/推导	引用被引文献中的相关公式/推导
结果	引用被引文献中的实验结果
未提及	在文中没有提及明显的引用对象

2.3 引文属性标注体系

为支持进一步的自动化识别和分类,本文需要标注者在标注过程中记录引文的一些具体属性特征。具体内容见表3。

表 3 引文特征属性及描述

类别	名称	描述
文献特征	施引文献类型	施引文献的文献类型
	被引文献类型	被引文献的文献类型
引文特征	引文出现位置	引文标记在文章中出现的位置
	文章内被引频次	在施引文献中某一被引文献的被引次数
	同句内引文个数	同一个被引句中出现的引用个数
	是否是自引	施引文献与被引文献是否是自引关系
语法特征	引文上下文范围	描述被引文献的语句内容及范围
	特征词	体现引用功能、情感的特征词汇

3 标注实验

3.1 实验设置

“主题模型”是一种描绘文档潜在语义结构的概率模型，能够有效地表示蕴含于文档集中的潜在主题，因此产生了大量衍生模型及应用成果，其中最有影响力的成果是由 Blei 等提出的 LDA 模型[55]。鉴于“主题模型”研究涉及领域的广泛性，本文选取该领域文献作为实验的标注对象，从专家 Blei 推荐的“主题模型”相关文献列表[56]中随机选取 20 篇作为标注实验的样本，标注样本中共出现 365 篇参考文献和 673 条引文记录。这 20 篇文献均为“主题模型”研究的高影响力成果，涉及自然语言处理、数据挖掘、图形处理等多个主题的理论及应用研究，一定程度上保证了实验样本的广泛性。

目前，可获取的科研文献数据主要为 PDF 格式，并不便于标注，为此本文设计了一个基于 Web 的引文内容标注工具 WHU-CCAS 以辅助标注（见图 1）。该工具右侧显示当前标注文章的基本信息，左侧显示当前标注文章的参考文献列表。在选取某一篇参考文献之后，工具中部就会显示出对应的标注界面，内容包括已标注

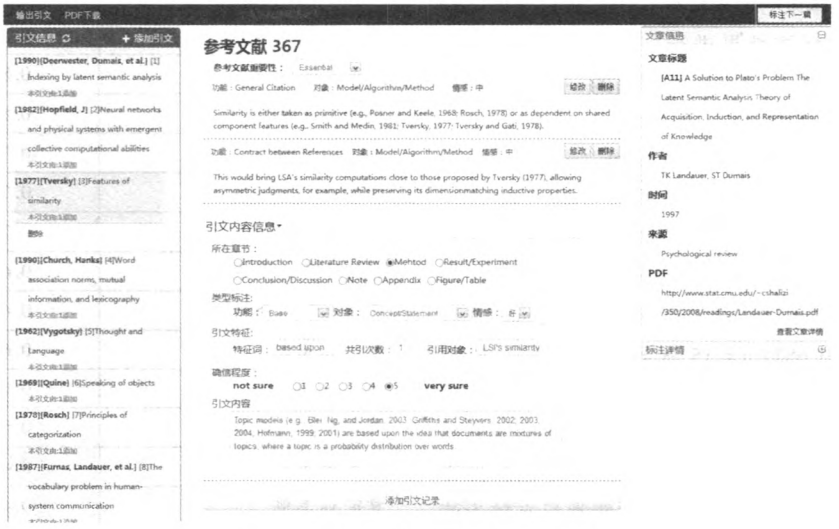

图 1 科研文献引文内容标注工具

的引文条目和添加新引文记录的表单。在标注引文记录时要求标注者对引文的分类信息、引用对象、所在章节位置以及上下文内容等信息进行标注。

本文的标注实验由第二作者和第三作者完成,在进行上文选定20篇文献的正式标注实验之前,两位标注者需要预先标注额外5篇文献,以统一基本的标注标准。对于每一个引文记录,标注流程描述如下:

(1)定位要标注的引文位置,判断引文所在章节的类型(引言、相关研究、方法、实验和结论);

(2)阅读引文所在语句,确认句子的词汇、包含引文个数以及是否自引等特征,大致判断该引文的功能和情感分类;

(3)确定该引文的上下文内容,从上下文中寻找相关特征;

(4)综合上述客观特征确定该引文的功能和情感类别,注意尽可能减少标注过程中的主观推理判断。

3.2 一致性分析

本文首先对两位标注者标注结果的一致性进行评估,结果一致性的高低一定程度上体现了该标注体系的可操作性。本文重点对引文功能分类的标注结果进行一致性评估和分析。Kappa系数是一个被广泛使用的一致性评价机制,其计算公式如下:

$$K=\frac{P(A)-P(E)}{1-P(E)}$$

其中$P(A)$表示标注结果一致性的实际观测值,$P(E)$表示标注结果一致性的期望值。本文两位标注者的引文功能标注结果Kappa一致性为$K=0.687$(n=15, N=673, k=2)。根据Teufel[28]给出的一致性参考指标($K\geq 0.8$表明十分可靠,$K\geq 0.69$表示可靠),该标注结果达到了一个相对可靠的一致性水平。在标注过程中,标注者表示相比于其他功能类目有较为清晰的判断特征,"简单引用"和"相关研究"这两类功能只能够通过所在章节和上下文内容来推断,且作者在文中的模糊表述也会增加判断的难度。本文将这两个功能进行合并以鉴定这两个类目上不一致的程度,合并之后Kappa值上升为0.806,达到了十分理想的一致性水平,这说明标注者对于"简单引用"和"相关研究"这两个类目的标注出现了一定分歧,有必要进一步调整两者的设置。

类似于Teufel[28]的分析,为了进一步判断标注过程中不同类目区分度的高低,本文进行了一对其他的二元一致性检验。即每次保留一个类目,同时将其他所有类目记为另一个类目,计算这种二元类目下的一致性系数,一致性系数越高表明标注者越能够将该类目与其他类目区分。区分度结果见表4,从表中数据可以看出,高于可信一致性标准$K=0.69$的有6个类目,表现出了较高的区分度。低于参考标准的8类中,"简单引用"和"相关研究"由于彼此之间区分度不足导致一致性较低。其余6类均只有较少的标注实例,由于Kappa系数对于标注频率低的类目更为敏感,因此少量标注结果不一致也会大幅度降低Kappa的一致性。

表4 不同功能类别的标注区分度

类别	一致性
历史背景 vs 其他	0.894
未来工作 vs 其他	0.856
比较 vs 其他	0.829
使用 vs 其他	0.797
详细引用 vs 其他	0.783
相似 vs 其他	0.746
相关研究 vs 其他	0.649
拓展阅读 vs 其他	0.613
基于 vs 其他	0.609
拓展 vs 其他	0.607
简单引用 vs 其他	0.569
启发 vs 其他	0.497
肯定 vs 其他	0.497
相关工作之间比较 vs 其他	0.284

此外在标注过程中还发现一些问题，一定程度上影响了标注者的判断。原则上要求标注者根据引文上下文中的客观线索进行标注，但是在实际标注中会出现一些客观线索无法支持分类的情况。如在 *Hierarchical Topic Models and the Nested Chinese Restaurant Process*（Blei，2004）中，作者基于中餐馆过程以及 LDA 模型提出了一个新的层次主题模型。然而文章中关于 LDA 这篇文献的引用只出现一次，且引文内容十分简单，没有体现较强的重要性。再者，部分引用内容容易引起标注者的理解歧义，如 *A markov clustering topic model for mining behaviour in video*（Hospedales et al，2009）中的一例。

Nevertheless, modeling the temporal order of visual events explicitly is risky, because noise in the event representation can easily propagate through the model, and be falsely detected as salient [9, 13].

文中提到了一个“错误识别”问题，但是通过具体内容很难判断出这个问题是由被引文献发现的还是指出了被引文献中的不足，这种情况只能借助领域知识或者通过阅读参考文献原文加以判断。

3.3 标注结果统计分析

综合两位标注者的标注结果得出引文的各项统计分布。表 5 是引文功能分类的标注结果分布，表格中还包括按照重要性等级分组后的引文频次统计。从表中数据可以看出，出现频次最高的引用功能是“相关研究”和“简单引用”，两者占所有引文数目的 67%。对于施引文献较为重要的 5 个功能的引文出现频次为 94，所占比例为 14%，这说明在学术文献中能够确定一定比例的重要引文，它们一定程度上体现了施引文献的主要思路来源。“比较”和“相似”的出现比例也有近 15%，比较的对象多为方法和结果。其余类目的引文出现频率较低，只占到整体的 5%左右，并未出现与文章研究主题无关的引用。从整体来看，本文标注的引文功能数量分布与其他成果[8,13,19]中的分布较为一致。

表 5 引用功能各类别标注信息统计

重要性	功能	引文个数	百分比(%)	按重要性分组个数	百分比(%)
非常重要	基于	11	1.63	11	1.63
重要	启发	3	0.45	83	12.33
	拓展	9	1.34		
	使用	55	8.17		
	详细引用	16	2.38		
一般	比较	69	10.25	99	14.71
	相似	28	4.16		
	肯定	2	0.30		
不重要	相关研究	287	42.64	465	69.09
	简单引用	163	24.22		
	相关工作之间比较	3	0.45		
	未来工作	5	0.74		
	拓展阅读	7	1.04		
非常不重要	历史背景	15	2.23	15	2.23
	无关引用	0	0.00		
合计		673		673	

引文的情感倾向数量分布上,标注为正面情感的引文数量为10个,占1.49%,负面情感的引文数量为16个,占2.38%,其余为中立情感,占96.14%,这说明学者进行施引主要的情感倾向为中立。在行文中表现出"正面"或者"负面"情感特征的引文只占不到5%,这一比例相比于Athar[15]统计的14%较低,这可能与本文标注文献的所属领域有关。

引用对象方面,标注结果显示引文中提及引用对象类型与提及具体引用对象的比例分别为34.7%和27.2%,这一比例足以说明引用对象出现的普遍性。同时,本文发现作者倾向于在文章中使用引用对象的名称来表示被引文献的工作,如在文章 *Finding Scientific Topics*(Griffiths,2004)中,对于重要参考文献 *Latent Dirichlet Allocation*(Blei,2003)的直接引用为7次,但在文章内容中"Latent Dirichlet Allocation"这一概念出现了8次,其中5次不是出现在引文内容中。类似的是,标注其余文章中也有使用"PLSA"、"LDA"等概念表示被引文献中的工作。这与Small[53]所研究的使用"概念符号"表示被引文献工作的现象一致,概念符号所在的语句往往与被引文献的内容相关,构成了对引文内容的良好补充。

表6展示了不同类型引用对象的数量分布,由于本文标注的文献是与"主题模型"相关的研究,因此"方法"、"模型"、"算法"等类型对象在引文中出现的频次最高。同时值得注意的是,由于一些概念所属的类型并不清晰,因此对于同一引用对象,在不同的文章中会出现不同的类型描述,如"Expectation Maximization"在不同的文章中被描述为"方法"(method)或者"算法"(algorithm)。

表6 不同引用对象类型出现的频次

引用对象类别	方法	模型	算法	应用	数据	结果	理论	工具	其他
个数	68	73	37	6	19	17	4	3	7

4 结论

本文回顾了引文内容分析的发展历程及现状,并对目前已有的引文内容标注体系进行了较为全面的梳理,归纳出引文分类依据的主要维度,并以支持文献语义关系挖掘为目标,设计出一个引文内容标注框架。该框架包括一个揭示引文关系抽象性质的引文分类体系,一个描述被引文献具体内容的引用对象标注体系,以及一个记录引文客观特征的引文属性标注体系,初步的标注实验验证了该框架的可用性。在下一步工作中,将对本文标注框架的不足加以改进,并对科研文献引文内容的自动化识别和处理进行研究。

参考文献

[1] Liu Y, Rousseau R. Interestingness and the essence of citation[J].Journal of Documentation, 2013, 69(4): 580-589.
[2] Hirsch J E. An index to quantify an individual's scientific research output[J].Proceedings of the National academy of Sciences of the United States of America, 2005, 102(46): 16569-16572.
[3] Garfield E. Citation analysis as a tool in journal evaluation[J].Science, 1972, 178(4060): 471-479.
[4] Zhang G, Ding Y, Milojević S. Citation content analysis (CCA): a framework for syntactic and semantic analysis of citation content[J].Journal of the American Society for Information Science and Technology, 2013, 64(7): 1490-1503.
[5] Small H. Citation context analysis[J].Progress in communication sciences, 1982, 3: 287-310.
[6] Oppenheim C, Renn S P. Highly cited old papers and the reasons why they continue to be cited[J].Journal of the

American Society for Information Science, 1978, 29(5): 225-231.
[7] McCain K W, Turner K. Citation context analysis and aging patterns of journal articles in molecular genetics[J]. Scientometrics, 1989, 17(1): 127-163.
[8] Hanney S, Frame I, Grant J, et al. Using categorisations of citations when assessing the outcomes from health research[J]. Scientometrics, 2005, 65(3): 357-379.
[9] Garfield E. Can citation indexing be automated[C]//Statistical association methods for mechanized documentation, symposium proceedings. 1965: 189-192.
[10] Lipetz B A. Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators[J]. American Documentation, 1965, 16(2): 81-90.
[11] Finney B. The reference characteristics of scientific texts[D]. City University (London, England), 1979.
[12] Garzone M A. Automated classification of citations using linguistic semantic grammars[D]. The University of Western Ontario, 1997.
[13] Teufel S, Siddharthan A, Tidhar D. Automatic classification of citation function[C]//Association for Computational Linguistics, 2006: 103-110.
[14] Radoulov R. Exploring automatic citation classification[D]. University of California, 1965.
[15] Athar A. Sentiment analysis of citations using sentence structure-based features[C]// Association for Computational Linguistics, 2011: 81-87.
[16] Athar A, Teufel S. Context-enhanced citation sentiment detection[C]//Association for Computational Linguistics, 2012: 597-601.
[17] Abu-Jbara A, Radev D. Reference scope identification in citing sentences[C]//Association for Computational Linguistics, 2012: 80-90.
[18] Collins H M. The TEA set: tacit knowledge and scientific networks[J]. Social Studies of Science, 1974, 4(2): 165-185.
[19] Cano V. Citation behavior: classification, utility, and location[J]. Journal of the American Society for Information Science, 1989, 40(4): 284-290.
[20] Herlach G. Citation patterns: mechanistically identifiable characteristics of citation links[D]. University of Chicago, 1973.
[21] Frost C O. The use of citations in literary research: a preliminary classification of citation functions[J]. The Library Quarterly, 1979, 49(4): 399-414.
[22] Spiegel-Rösing I. Science studies: bibliometric and content analysis[J]. Social Studies of Science, 1977, 7(1): 97-113.
[23] Moravcsik M J, Murugesan P. Some results on the function and quality of citations[J]. Social studies of science, 1975, 5(1): 86-92.
[24] 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.
[25] Swales J. Citation analysis and discourse analysis[J]. Applied Linguistics, 1986, 7(1): 39-56.
[26] Garzone M, Mercer R E. Towards an automated citation classifier[M]//Advances in artificial intelligence. Springer Berlin Heidelberg, 2000: 337-346.
[27] Nanba H, Okumura M. Towards multi-paper summarization using reference information[C]//IJCAI. 1999, 99: 926-931.
[28] Teufel S, Siddharthan A, Tidhar D. An annotation scheme for citation function[C]//Association for Computational Linguistics, 2009: 80-87.
[29] Di Iorio A, Nuzzolese A G, Peroni S. Identifying functions of citations with CiTalO[M]//The Semantic Web: ESWC 2013 Satellite Events. Springer Berlin Heidelberg, 2013: 231-235.
[30] Xu H, Martin E, Mahidadia A. Using heterogeneous features for scientific citation classification[C]//Proceedings of the 13th conference of the Pacific Association for Computational Linguistics. 2013.
[31] Qazvinian V, Radev D R. Scientific paper summarization using citation summary networks[C]//Association for Computational Linguistics, 2008: 689-696.
[32] Qazvinian V, Radev D R, Özgür A. Citation summarization through keyphrase extraction[C]//Association for Computational Linguistics, 2010: 895-903.
[33] Teufel S, Moens M. Summarizing scientific articles: experiments with relevance and rhetorical status[J]. Computational linguistics, 2002, 28(4): 409-445.
[34] Ritchie A, Teufel S, Robertson S. Creating a test collection for citation-based IR experiments[C]//Association for

Computational Linguistics,2006:391-398.
[35] Ritchie A, Teufel S, Robertson S. Using terms from citations for IR: some first results[M]//Advances in Information Retrieval. Springer Berlin Heidelberg, 2008: 211-221.
[36] Ritchie A, Teufel S, Robertson S. How to find better index terms through citations[C]//Association for Computational Linguistics,2006:25-32.
[37] Liu S, Chen C, Ding K,et al. Literature retrieval based on citation context[J].Scientometrics: 2014, 101(2): 1293-1307
[38] Ritchie A, Robertson S, Teufel S. Comparing citation contexts for information retrieval[C]//ACM,2008:213-222.
[39] Nakov P I, Schwartz A S, Hearst M. Citances: citation sentences for semantic analysis of bioscience text[C]// 2004:81-88.
[40] 崔红. 我国科技人员引文动机聚类分析[J].情报杂志, 1998, 17(2): 68-70. (Cui Hong. A cluster analysis of Chinese researchers' citation motivation[J].Journal of Intelligence, 1998, 17(2): 68-70.)
[41] 叶继元,袁培国,吴向东. 引文数据中的负面引用初探[J].新世纪图书馆, 2008(6): 22-23. (Ye Jiyuan, Yuan Peiguo, Wu Xiangdong. A primary study on negative citation in citation data[J].New Century Library, 2008 (6): 22-23.)
[42] 陈晓丽. 引文评价中的引文方式与力度因素[J].图书馆, 2000(6): 43-45. (Chen Xiaoli. The factor of citation type in citation appraisement[J].Library, 2000(6): 43-45.)
[43] 赵青. 文学学科引用性质与引用深度调查分析[J].情报杂志, 2010, 29(10): 46-50. (Zhao Qing. Research on citation character and citation depth in literature[J].Journal of Information, 2010, 29(10): 46-50.)
[44] 胡志刚,陈超美,刘则渊,等. 从基于引文到基于引用——一种统计引文总被引次数的新方法[J].图书情报工作,2013, 57(21): 5-10. (Hu Zhigang, Chen Chaomei, Liu Zeyuan, et al. From counting references to counting citations: a new way to calculate the total cited times references[J].Library and Information Service, 2013, 57(21): 5-10.)
[45] 刘盛博,丁堃. 基于引用内容的引文评价分析[C]//第九届中国科技政策与管理学术年会论文集, 2013. (Liu Shengbo, Ding Kun. Citation evaluation analysis based on citation context[C]//The Paper Collection of the 9th Conference of China Science and Technology Policy and Management, 2013.)
[46] 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, 34(5): 39-43. (Zhu Qingsong, Leng Fuhai. A review of research on citation content analysis method[J].Information and Documentation Services, 2013, 34(5): 39-43.)
[47] Moed H F. Citation analysis in research evaluation[M].Springer, 2006.
[48] Wan X J, Liu F. Are all literature citations equally important? Automatic citation strength estimation and its applications[J].Journal of the Association for Information Science and Technology, 2014,65(9):1929-1938.
[49] Brooks T A. Evidence of complex citer motivations[J].Journal of the American Society for Information Science, 1986, 37(1): 34-36.
[50] Vinkler P. A quasi-quantitative citation model[J].Scientometrics, 1987, 12(1): 47-72.
[51] Liu M X. Progress in documentation the complexities of citation practice: a review of citation studies[J].Journal of documentation, 1993, 49(4): 370-408.
[52] Bornmann L, Daniel H. What do citation counts measure? A review of studies on citing behavior[J].Journal of Documentation, 2008, 64(1): 45-80.
[53] Small H G. Cited documents as concept symbols[J].Social studies of science, 1978, 8(3): 327-340.
[54] Athar A, Teufel S. Detection of implicit citations for sentiment detection[C]//Association for Computational Linguistics,2012:28.
[55] Blei D M, Ng A Y,Jordan M I. Latent dirichlet allocation[J].The Journal of Machine Learning Research, 2003, 3 (1): 993-1022.
[56] Mimmo D. Topic modeling bibliography[EB/OL].[2014-07-16].http://mimno.infosci.cornell.edu/topics.html.

陆 伟 武汉大学信息管理学院教授,博士生导师。
通讯地址:湖北省武汉市武昌区八一路299号武汉大学信息管理学院。邮编:430072。
孟 睿 武汉大学信息管理学院硕士研究生。通讯地址同上。
刘兴帮 武汉大学信息管理学院硕士研究生。通讯地址同上。

(收稿日期:2014-06-19;修回日期:2014-07-31)