用于引文内容分析的标准化数据集构建*

张梦莹，卢　超，郑茹佳，章成志

摘　要　学术论文的全文数据越来越容易获取使大规模的引文内容分析成为可能。文章通过设计引文内容标注框架，开发引文内容标注系统，分别从引用对象、引文功能、引用情感、引文位置、引文重要性、标注自信度等方面进行标注。构建用于引文内容分析的标准化数据集并进行统计分析，可为引文内容的特征分析等基础性研究及学术预测等应用性研究提供数据支撑。

关键词　引文内容分析　引文标注系统　标准化数据集　引用功能　引用情感倾向

引用本文格式　张梦莹，卢超，郑茹佳，等. 用于引文内容分析的标准化数据集构建[J]. 图书馆论坛，2016 (8)：48-53.

Construction of Standardized Data Set for Citation ContentAnalysis

ZHANG Meng-ying，LU Chao，ZHENG Ru-jia，ZHANG Cheng-zhi

Abstract　As the structured data of academic literature becomes more and more accessible，it is likely toanalyze large-scale citation contentautomatically. In this paper，the framework of citation content annotation is constructed and a citation content annotation system is developed. Annotation is carried outon the objects，the functions，the sentiment，the location andthe importance of citations，and the degree of confidence. A standardized data set for citation analysis is then constructed and the statistical analysis is done，which provide data support for the basic research and applied research on citation content.

Keywords　citation content analysis；citation tagging system；standardized data set；citationfunction；citationsentiment

0　引言

自 Garfield 提出 SCI 以来，引文分析一直是图书情报学领域的研究重点和热点。最初由于技术不成熟且全文数据匮乏，学者们关注更易获取的题录和参考文献信息，但因忽略引文内容、位置、情感极性等被引情况，引文分析的结果缺乏内容层面的数据支持[1]。有关引文内容的少数研究一般是人工分析少量学术论文，结论缺乏普适性[2]。随着自然语言处理技术不断发展，学术文献全文数据获取难度下降，引文内容分析取得一定的成果。然而关于引文内容边界识别、引文功能及情感极性判定等基础性问题尚未出现公认的结论[3]，并且缺乏支持这些研究的公开的标注数

*本文系国家社会科学基金项目“在线社交网络中基于用户的知识组织模式研究”（项目编号：14BTQ033）、国家社会科学基金重点项目“大数据环境下社会舆情与决策支持方法体系研究”（项目编号：14AZD084）研究成果之一

据集。因此，构建可用于引文内容分析的较大规模的标注数据集具有重要意义。

近年国内外部分学者关注到上述问题并做了相关研究。Simone Teufel 等形成了自动识别引文功能的框架[4]；Athar 等利用机器学习方法对引文内容情感极性进行识别，精确度有待提高[5]；陆伟等对引文内容标注作了全面梳理，并开发出相应的标注平台，但庞杂的标注框架对标注者要求较高，且提供的标注数据规模较小[6]。为此，本文兼顾全面与易用，提出新的引文内容标注框架，并构建标注数据集进行统计分析。一方面，不断扩大标注数据的规模，为缺乏统一规范的全文数据和引文内容的自动化标引提供样本；另一方面，对其进行统计分析以了解和掌握数据集的基本特征，为后续大范围的引文内容分析、情感极性测度、动机识别等深入研究提供较为清晰和直观的数据。

1 相关研究概述

从引文分析理论的研究[7]到评估学术产出的衡量指标[8]，再到意识到缺乏引文内容信息的不足，以人工方式进行小样本的引文内容分析的相关研究逐渐出现。引文内容概念[9]提出后，大批学者投入引文内容特征与应用的研究之中。随后计算机技术的发展保证了大规模获取全文本信息及进行文本挖掘，引文网络研究、引文主题相似性研究、情感倾向分析等逐渐成为引文内容分析的新思路[10]。隐含引用、自引、转引等现象普遍存在，引文范围难以界定，而且引用动机的复杂使得判别的准确性难以保证，这都导致引文数据的充分性和准确性无法保证。如前所述，目前缺乏相对统一的引文内容标注框架和较大规模的引文内容标注数据集，而这些基础研究能在一定程度上解决或缓解上述难题。

1.1 引文内容标注框架研究概述

自 20 世纪 60 年代起，关于引文内容分类体系即引文内容标注框架的研究逐渐出现。Garfield[11]通过研究引文位置、上下文、重要性等因素，提出了 15 种引用动机，为后续研究提供了引导方向。Lipetz[12]定义了 29 种不同的引用原因，但没有很好地应用到具体的实证分析中。20 世纪 70 年代，一些研究人员根据其思路进行研究，但研究繁杂导致使用尤为困难。Oppenheim &Renn[2]整理出一个包括“ 历史背景”“ 相关工作的描述”“ 提供信息或数据”“ 比较”等 7 种类别的分类方法，使得分类体系逐渐清晰明了。此外，对引文分类体系的研究还存在一定程度的融合现象。Bilal Hayat Butt 等[13]将 Spiegel- Rosing 等[14]提出的 13 种引文动机类型分为 3 类概括性的情感类型。Simone Teufel[4]结合文章结构及引用情感提出一个 12 类引文功能的分类体系，但其工作仅倾向于施引文献与引文之间的对比。引文分类体系的研究为引文内容标注框架的开发提供了理论依据。Ying D 等[15]从语法和语义两方面分别对引文属性、被引属性以及两者之间的属性进行分析，构建了相对全面均衡的引文内容分析框架。该框架的提出虽有较大提升，但实际应用效果并不理想。陆伟等[6]的研究分为 15 个功能类目的引文分类体系，11 个类型的引用对象标注体系以及 8 个方面的引文属性标注体系，但因注重全面性却使其体系过于庞杂，缺乏易用性。

1.2 现有引文内容分析的数据集概述

专注于引文内容标注框架研究的相关文献，目前所使用的数据集仅有 3 个，见表 1。这表明目前用于引文内容研究的标注数据集的规模较小，一般以约 20 篇学术文献进行标注分析，且数据多数暂未公开。因此，构建一个较大规模且公开的标注数据集具有重要意义。

表 1 现有引文内容分析的标注数据集

作者	数据来源	标注文章数目	标注数据条数
陆伟等[6]	Blei 推荐的“ 主题模型” 文献[16]	20	673
Tefuel等[4]	CmpLg	26	548
Athar等[5]	ACL Anthology	20	1741

传统引文分析局限性的凸现，加之文本挖掘和自然语言处理等技术兴起，激发了学者对引文

内容研究的热情。为获得科学的研究数据，学者纷纷提出不同的引文内容标注体系，但尚未有较为完善的标注理论和方法。因此，本文通过对前人的研究进行梳理和总结，形成一套较为完整与易用的引文内容标注框架，并构建用于引文内容分析的标注数据集，进而帮助引文更好地应用到学术评价、提高检索性能、推荐系统以及学术预测等不同领域。

2 研究思路与方法

2.1 研究思路

本研究首先获取了一定规模的学术论文数据集；然后设计引文内容标注框架，并开发“引文内容标注系统”；接着随机选取 102 篇文献，分别从引用对象、引文功能、引文情感倾向、引文位置等方面进行标注，并对引文重要性、标注自信度进行评判，得到引文内容标注数据集，并进行了相应的统计分析。研究思路如图 1 所示。

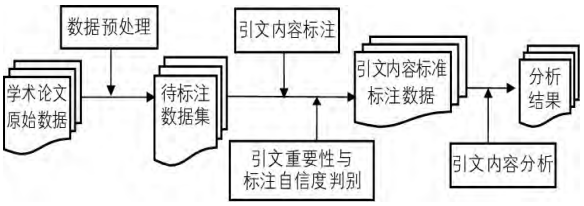

图 1 研究思路

2.2 数据

2.2.1 数据来源

Plos One(http://journals.plos.org/plosone/)是目前学术界非常有影响力的开放存取期刊，载文学科广泛，涉及从自然科学到社会科学等 10 多种学科。该期刊对所发表论文提供结构化全文下载，非常适合本研究。因此，本文抓取来自 Plos One 期刊 2006- 2015 年发表的 3414 篇文献，涉及 Cell Biology、Chemistry、Computer Science、Mathematics、Mental Health、Physics 等 6 个学科。

2.2.2 数据预处理

获得论文全文数据后，对所需信息进行抽取并存储至数据库，主要包括两部分：(1)题录信息。Plos One 中文献的题名、作者、通讯邮箱、发表时间、审查周期、论文类型及学科属性等内容。(2)引文内容信息。引文内容及其前后各两句话；引文内容所在的篇章结构及标题；引文内容中包含的引文的数目等信息。

2.3 方法

2.3.1 引文内容标注框架设计

正式标注实验前，预先制定并统一标注标准。通过对“引文内容标注框架”相关文献的调研和对已有引文分类体系的整理和分析，本标注框架分为 6 个部分，见表 2。

表 2 引文内容标注框架说明

标注对象	类别	描述
引用对象	背景	引用被引文献中的背景
	定义	引用被引文献中的定义
	数据	使用被引文献中提供的数据
	方法	使用被引文献中提供的方法
	理论	引用被引文献中的理论
	工具	使用被引文献中的工具
	结果	引用被引文献中的结果
	结论	引用被引文献中的结论
	观点	引用被引文献中作者的观点
	其它	其它
引文功能	背景	引文内容在施引文献中作为背景
	术语来源	引文内容在施引文献中作为术语来源
	研究基础	引文内容是施引文献的研究基础
	研究空白	引文内容提出了施引文献的研究空白
	相似性研究	引文内容作为施引文献中的相似性研究
	结果比较	引文内容在施引文献中以结果比较的形式出现
	评论	引文内容在引文献中以评论的形式出现
	比较	引文内容在引文献中以比较的形式出现
	相关研究	引文内容作为施引文献中的相关研究
	其它	其它
引用情感	消极	持消极情感，包括指出不足以及转折或否定性引用
	中立	持中立情感，即不含情感词的描述性引用
	积极	持积极情感，包括肯定或赞扬
引文位置	引言	引文在施引文献的引言部分
	文献综述	引文在施引文献的文献综述部分
	方法	引文在施引文献的方法部分
	结果	引文在施引文献的结果部分
	讨论	引文在施引文献的讨论部分
	结论	引文在施引文献的结论部分
引文重要性	1	非常不重要
	2	不重要
	3	一般
	4	重要
	5	非常重要
标注自信度	1	非常自信
	2	自信
	3	一般
	4	不自信
	5	非常不自信

2.3.2 引文内容标注平台实现

数据标注平台开发利用 Python 语言在 Django 1.8 框架下完成。在对系统需要的功能予以理解的基础上，设计系统的流程。引文内容标注系统包括两个部分：用户部分及管理员部分。用户部分包括用户注册、用户登录、用户标注；管理员部分包括管理员登录、管理员查看所有已标注的结果。每一个用户登录后按照分页内容选择文章，进入标注界面，对文献的每条引文内容进行标注，并提交。其中，标注界面包含文章基本信息，引文内容标注区域，自信度评价区域。

2.3.3 数据标注步骤

数据标注的步骤可分为三步：(1)根据数据标注策略依次对引文内容标注；(2)根据标注的引文内容相关信息和被引文献在施引文献中的角色评估其重要性；(3)根据标注者的标注情况对其当前引文内容条目的标注结果进行自信度的打分。

本实验分为两阶段：第一阶段在标注实验的 1/3 处据实际情况改进标注标准，以提高后续标注实验的质量。第二阶段则根据改进后的标准进行余下 2/3 的标注实验，对于全部标注数据中自信度为“ 3”及以下的标注结果进行二次标注，获得标注数据集。

2.3.4 标准化引文内分析数据集的分析

为保证标注结果的一致性，此次标注结束后两位标注者先各自对其标注的数据集进行统计分析，随后将分析结果进行比较，差异性较小。为进一步提高数据可信度，在进行数据获取与处理时，抽取标注自信度为“ 4”和“ 5”的全部数据进行分析。

在对标注数据集进行引文内容特征分析时，分别采用统计分析法和文献分析法。通过数据透视图进行数量和百分比的统计，并通过绘制饼图、柱状图等分析标注结果。随后对标注实验呈现出的结果进行解释分析，在此过程中主要采用文献分析法，利用其他相关研究中的权威结论验证本次标注实验结果的准确与普适性。

3 标准化数据集的分析

3.1 标注结果概述

本文随机抽取的学科、各学科文献数目及引文数目的分布情况如表 3 所示。引文数据来源广泛弥补了仅局限于类似“ 模型研究”等单一主题文献分析而导致由于学科属性使得分析结果过度偏离事实的缺陷，因此更具普适性。

表 3 引文内容分析标准数据集的数据分布情况

Discipline	#discipline	#citation content
Cell Biology	45	3046
Computer Science	37	2040
Physics	10	440
Mathematics	8	395
Chemistry	1	44
Mental Health	1	32
总计	102	5997

随后分析 5997 条引文数据的标注自信度，其中自信程度在 4 及以上的标注结果累计达 88.71%。此次标注实验基于本文设计的引文内容标注系统，从而建立引文内容分析的标准化训练集。由此看出，此标注系统的可用性较强。为进一步提高数据的可信度，对其进行数据筛选，过滤掉自信度较低的 11.29%的标注数据，挑选标注自信度为 4 和 5 的全部数据进行后期分析，共 5320 条。其中自信度为 4 的 4495 条，占 84.49%；自信度为 5 的结果共 825 条，占 15.51%。

3.2 引文内容标注指标结果分析

3.2.1 引用对象统计分析

标准化数据集的引用对象方面如图 2 所示。引用对象为“ 结论”“ 结果”的现象普遍存在，分别占 31.28%和 30.68%。科技文献是推动研究成果发展的手段，因此学者通常引用他人的方法、结果、结论来揭示其是否可以达到共同的认知。引用对象频率较高的其次对象是“ 方法”，出现次数多达 1034 条，占 19.44%。而“ 工具”“ 理论”“ 其它”最少，这是由于选取学科为自然科学，而成熟的理论与成形的软件等工具较多出现在社会科学文献中，这里出现的少部分是由于存在学科交叉现象所导致。

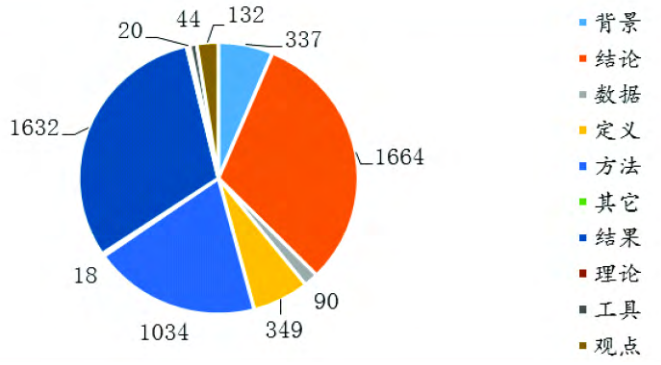

图 2　标准化数据集的引用对象统计

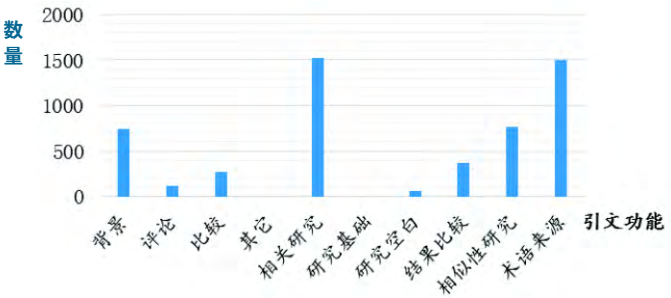

图 3　标准化数据集的引文功能统计

3.2.2　引文功能统计分析

标准化数据集的引文功能的标注结果如图 3 所示。“ 结果”“ 术语来源” 是出现频次最高的引文功能，二者占所有引文标注数据的 56.62%。刘宇等[17]提出多数引文是向读者提供研究来源信息，或罗列现有的相关研究成果。“ 比较”“ 结果比较” 也是常见的引用动机，一般用来与他人对比结果及方法。由于引用行为中真正起到实质性作用的引文甚少，因此“ 研究基础”“ 研究空白”“ 评论” 所占比例最低，三者仅占 3.34%。

3.2.3　引用情感统计分析

标注结果显示，5141 条引文数据表达中立的引用情感占96.64%，而明显带有情感色彩的引文数据仅占 3.36%，其中 113 条引文数据表达了积极的情感倾向，占 2.12%；66 条引文数据表达消极的情感倾向占 1.24%。本文上述研究结果与陆伟等[6]之前的研究结果(中立情感的引文占比 96.14%)极为相似。引用的情感是隐藏的，以避免学术上公开的批评，因此引用情感常是中立的，而带有明显情感色彩的引用中，学者们更倾向于积极引用。

3.2.4　引文位置统计分析

Hu[18]表示近半数的引文都高度集中于文章的引言部分，而本文标注实验与其结论相似，但研究结果略低。位于“ 引言” 位置的引文共 2318 条，占 43.57%。学者们通常在文章开始引用他人文献来引出自己的研究方法与思路，且多数相关综述被合并到引言中，因此“ 引言” 部分是引文出现的高频区。其次是位于“ 讨论” 的引文出现频次较高，共 1161 条，占 21.82%。Plos One 结构化全文数据使本文“ 结论” 处的引文仅占 0.45%，“ 文献综述” 处的仅占 0.23%。

3.2.5　引文重要性统计分析

最后本文给出引文重要性的统计分析结果，引文的重要程度为 2 和 3 的最多，分别为 2219 条以及 1901 条，二者占全部数据的 77.44%。对作为背景或相关性研究等引文的简单提及较为普遍而且重要性相对较小，而真正非常重要且具有影响力的引文应该是能够激发新的想法、方法的引用行为，例如作为本文的研究基础，仅占全部数据的 0.28%，可以看出，引用他人文献不一定代表其对自己文章的重要性高。

4　总结与展望

缺乏科学的引文数据是引文内容分析发展的瓶颈，而引文内容标注框架的研究提供了良好的解决方法。因此，本文对前人的研究进行梳理，开发了一个较为完整与易用的引文内容标注系统，通过对英文文献进行标注实验证实了该框架的可用性，并对引用对象、引文功能、引用情感、引文位置及引文重要性等方面的标注数据进行分析与讨论，构建了较为标准的数据集，具有重要的理论与应用价值。通过简单的数据统计，本文发现学者引用他人文献的行为与动机较为复杂，引用对象方面，结果和结论是最常见的；引文功能方面，一般仅作为相关研究简单提及或给出来源；引用情感方面，学者通常会避免学术上的批评而选择中立态度；引文位置方面，近半数

的引文出现在引言部分；引文重要性方面，多数引文没有实质性的作用，重要性偏低。今后还可优化引文分类标准，为引文内容标注体系的后续研究提供理论依据；也可开发新功能，收集更全面、准确的训练数据，为实现引文内容的自动标引与分析提供数据支持。在标注数据分析方面，今后可深入挖掘引用行为的特征及规律，进一步提升分析结果的应用水平，并为学术评价与推介、学科知识演化等应用研究提供有力支持。

参考文献

[1] 叶鹰. 高品质论文被引数据及其对学术评价的启示[J]. 中国图书馆学报，2010，36（1）：100-103.

[2] 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（3）：225-231.

[3] 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.

[4] Teufel S，Siddharthan A，Tidhar D. An annotation scheme for citation function[C]//Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue. Sydney，Australia，2009：80-87.

[5] Athar A，Teufel S. Context-enhanced citation sentiment detection[C]//Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Montreal，Canada，2012：597-601.

[6] 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（3）：93-104.

[7] Small H. Co-citation in the scientific literature：A new measure of the relationship between two documents[J]. Journal of the American Society for Information Science，1973，24（4）：265-269.

[8] 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.

[9] Small H. Citation context analysis[A]// Progress in communication sciences[M]. Norwood， NJ：Ablex Publishing，1982：287-310.

[10] 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，2013，38（6）：97-107.

[11] Garfield E. Can Citation Indexing Be Automated[C]// Proceedings of the Symposyum on Statistical Associca-tionMethods for Mechanized Documentation，Symposium. Washington，1963：189-192.

[12] 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.

[13] Butt B H，Rafi M，Jamal A，et al. Classification of Research Citations（CRC）[C]//Proceedings of the First Workshop on Mining Scientific Papers：Computational Linguistics and Bibliometrics. Istanbul，Turkey，2015：18-27.

[14] Spiegel-Rösing I. Science studies：bibliometric and content analysis[J]. Social Studies of Science，1977，7（1）：97-113.

[15] Ding Y，Liu X，Guo C，et al. The distribution of references across texts：Some implications for citation analysis[J]. Journal of Informetrics，2013，7（3）：583-592.

[16] David Mimmo. Topic Modeling Bibliograpyh[EB/OL]. [2014-07-16]. http://mimmo.infosci.cornell.edu/topics.html.

[17] 刘宇，李武. 引文评价合法性研究——基于引文功能和引用动机研究的综合考察[J]. 南京大学学报（哲学·人文科学·社会科学），2013，50（6）：137-148.

[18] Hu Z，Chen C，Liu Z. Where are citations located in the body of scientific articles? A study of the distributions of citation locations[J]. Journal of Informetrics，2013，7（4）：887-896.

作者简介 张梦莹，女，南京理工大学信息管理与信息系统专业本科生；卢超，男，南京理工大学管理科学与工程专业博士生；郑茹佳，女，南京理工大学信息管理与信息系统专业本科生；章成志，男，博士，博士生导师，南京理工大学信息管理系教授，通讯作者，E-mail：zhangcz@njust.edu.cn。

收稿日期 2016-06-12

责任编辑：何燕