图书馆杂志（2021年第4期 总第360期）
Library Journal(Vol.40 No.4)

基于本体和关联数据的全文引文分析方法研究*

石泽顺 肖 明（北京师范大学政府管理学院）

（图表扫码读取）

摘 要 针对传统引文分析法中存在的缺陷，提出了基于本体和关联数据的全文引文分析方法，以期利用本体和关联数据技术来揭示学术文献之间的全文引用关系，实现全新视角下的引文分析。首先选取《科学计量学》杂志引文分析领域的高被引文献组成实验数据集，并对全文引用信息进行抽取和标注。接着构建全文引文本体FCO对引文数据进行规范化描述，并发布为引文关联数据。最后，构建不同的SPARQL检索式对全文引用信息进行多维度抽取，对数据集的引文数量、引文功能、引文情感和引文位置的分布情况进行具体分析，实验验证基于本体和关联数据的全文引文分析方法的可行性。实验表明基于本体和关联数据的全文引文分析方法具有一定的可行性和实用性，对传统引文分析法的优化和全文引文分析方法的普及有着重要意义。

关键词 本体 关联数据 全文引文分析

Research on Full-text Citation Analysis Method Based on Ontology and Linked Data

Shi Zeshun, Xiao Ming (School of Government, Beijing Normal University)

Abstract This paper proposed a full-text citation analysis method based on ontology and linked data technology, aiming to reveal the full-text citation relationship between academic literature, and to achieve a new perspective of citation analysis. First of all, we selected high-cited papers in the "citation analysis" field of the *Scientometrics* journal to construct experimental data sets, and extracted the full-text citation data between citing papers and cited papers. Then, a Full-text Citation Ontology (FCO) was constructed to normalize the data and publish RDF triples as citation linked data. Finally, we constructed different SPARQL queries to extract the full-text citation information in multiple dimensions. We also analyzed the citation numbers, citation functions, citation sentiments, and citation locations in detail, hoping to verify the feasibility of this method. We found that our full-text citation analysis method has certain feasibility and practicability, which is of great significance to the optimization of the traditional citation analysis method and the popularity of the full-text citation analysis method.

Keywords Ontology, Linked data, Full-text citation analysis

0 引言

引文分析法（Citation Analysis）是一种揭示文献数量特征和内在规律的文献计量学方法。它主要利用数学、统计学和逻辑学中的归纳和综合等方法对学术文献、期刊、作者等对象的引用行为进行系统分析，进而得出一定的引证规律[1]。经过长期的发展，引文分析法在理论和实践方面都取得了一定成就，已经广泛应用在科学发展模式揭示、学科知识评价和学术前沿探测等领域，对科技创新和科研决策都有着重要意义。传统的引文分析方法和工具大多依赖于引文数据库，但随着研究的深入人们逐渐发现，基于引文数据库的引文分析方法存

* 本文系2016年度国家社会科学基金项目"基于语义识别的引文分析理论、方法与应用研究"（项目编号：16BTQ073）的研究成果之一。

在以下弊端[2]：①将所有引用行为视为是同等重要的；②各种统计指标均以文献作者本人标注的参考文献数量为依据；③只能揭示文献之间引用与否的简单关系，不能揭示更深层次的引用语义关系。在此背景下，研究人员从不同角度开展了优化和拓展传统引文分析法的研究。基于内容的引文分析方法[3]和全文引文分析法[4]也相继被提出，试图从施引文献与被引文献的全文视角分析深层次的引用语义关系。

自2006年Tim Berners-Lee[5]提出关联数据（Linked Data）的概念后，关联数据就一直是图书馆学、情报学的研究热点。关联数据通过RDF三元组将原本异构的数据互相关联起来，提供了一种全新的知识发现途径[6]。经过十余年的发展，关联数据技术在诸多领域都获得了广泛应用。在引文分析领域，关联数据的三元组结构与学术文献的引用行为之间有很多相通之处，因此可用于描述引文数据，并且发布为引文关联数据[7]。引文关联数据将各种引文知识单元进行重新组合和排序，从而形成了一个有序的、机器可读的、结构化的知识网络，这为大规模的引文数据处理和分析奠定了基础[8]。可以预见的是，在引文开放的大环境下，将会有越来越多的引文信息被发布成关联数据的格式，全文引文数据也将以机器可读的形式被更加规范化地存储起来。因此，对全文引文关联数据进行处理、分析和可视化，将会成为未来引文分析领域的研究热点。

1　相关研究进展

传统的引文分析方法和工具依赖于引文数据库。但国内外各大引文数据库都将所有引文看作是同等重要的，各种统计指标也均以被引用次数来衡量[2]。这种仅仅通过引用次数来衡量的方法只能表示引用与否，不能揭示出文献之间丰富的引用关系。自20世纪60年代起，一些学者开始进行引用行为与动机方面的研究，代表人物有Garfield[9]、Weinstock[10]、Lipetz[11]、Thorne[12]等。他们从理论角度进行分类，虽然没有进行实证研究，但为引用动机的研究提供了理论支持。另外Brooks[13]、Vinkler[14]、Bonzi[15]、Liu[16]、Case[17]、Tang[18]等人采取了调查访谈和实证研究的形式，通过对引文作者进行调查、访谈等来总结引用行为的原因和引用动机的类型。伴随着引用动机与引用行为研究的日臻成熟，全文引文分析方法被提上议案。与传统的基于文献题录信息分析不同，全文引文分析法将施引文献和被引文献的全文作为研究对象，因此能够揭示更深层次的引用关系。2013年，Ding等[3]提出基于内容的引文分析是下一代的引文分析，将其分为语法层面和语义层面。其中，语义层面是指利用文本挖掘和自然语言处理技术来分析引文的文本内容，进而揭示被引文献对施引文献的语义贡献。2014年，Zhang等[19]设计了基于引文内容的语义分析框架CCA。他们认为CCA可以对传统引文指标进行补充，并分析了其潜在价值及应用方向。2014年，赵蓉英等[4]提出，随着结构化全文数据的普及，全文本引文分析法将得到全面的发展。他们归纳了传统引文分析法的不足，认为全文本引文分析作为一种微观的、基于全文数据的引文分析方法，可以从根本上改变引文分析与科学计量的方法与结构；2014年，胡志刚[20]从全文引用的角度出发，选取《信息计量学杂志》（Journal of Informatics）的文献数据进行了引用位置、引用强度和引用语境的全文引文分析。同年，祝清松[2]在引文内容分析的基础上提出语义增强的引文分析方法。他将引文关键路径识别作为语义增强引文分析的应用方向，并选择量子失协和碳纳米管纤维领域进行了实证研究。陆伟等[21]梳理了现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：引文功能、引文重要性、情感倾向。他们针对引文内容分析设计出一个引文内容标注框架，并通过实验验证了该标注框架的可用性。

综上所述，全文引文分析方法作为一种全新的分析方法，具有自身独特的优势，已受到研究人员的广泛青睐。但需要注意的是，全文引文信息的标注与处理是一项非常复杂的工作，难以达到普及的程度。因此，当前国内外研究多为理论分析探讨，实证研究较少，全文引文分析方法的可行性和适用性都存在一定疑问。有鉴于此，本文首次提出了基于本体和

关联数据的全文引文分析方法。该方法旨在通过语义网中的本体和关联数据技术对引文数据进行规范化描述，利用SPARQL检索式来提取特定维度的引文数据，进而实现全新视角的引文分析，克服传统引文分析方法中存在的缺陷。

2 数据集构建与预处理

2.1 文献数据集构建

全文引文分析需要对作者引用参考文献的真实意图进行注释。但事实上，要想完全猜测和理解作者的引用意图是非常困难的。因此，标注全文引用信息要建立在对相关学科知识充分理解的基础上，尽可能地体会作者引用的实际意图。在文献数据选择上，本文选择引文分析作为实验研究领域，并且以《科学计量学》（Scientometrics）杂志2017年发表的引文分析领域的学术文献作为实验数据源。《科学计量学》创建于1979年，综合影响因子2.147，是目前文献计量学和引文分析领域最权威的期刊之一。该杂志主要关注科研活动的数量特征，重点放在通过数学统计方法研究科学发展和机制的趋势上。《科学计量学》杂志发表的文献类型主要有原创性研究、研究报告、综述、书评等。该杂志对研究人员和研究管理人员而言是不可或缺的，其全面的跨学科特性为研究人员、图书馆员和文献情报专家提供了极大的帮助[22]。本次检索的检索日期为2018年1月29日，最终检索结果为160条记录。

接着，按照Web of Science数据库的本地被引得分（Local Citation Score，LCS）指标进行降序排列，选取LCS不小于2的25篇论文（占15.6%）构成施引文献数据集CitingSet。这些文献是《科学计量学》杂志在引文分析领域最新、最重要的研究成果，具有一定的代表性。25篇施引文献共计包含1 271条参考文献（平均每篇有50.84条引文）。经过去重处理，最后得到1 199条（占94.3%）不重复的数据。同样地，对参考文献进行依次编号，得到被引文献数据集CitedSet。在此基础上，将施引文献集合CitingSet和被引文献集合CitedSet进行合并，构成本研究的最终实验数据集。

2.2 全文引用信息抽取和标注

从施引文献全文中可以抽取出带有引文标识的引用句，这些引用句与参考文献一一匹配对应，为全文引用信息的标注提供了可能。全文引用信息抽取和标注的主要流程如图1所示，包括引用句的抽取和引用信息标注两个重要步骤。首先，使用基于规则的方法（如正则表达式匹配等）和人工校对相结合的方式来提取出施引文献的引用句，这些引用句与参考文献列表一一匹配，构成了全文视角下引用行为的主体和客体。接着，由编码员从引用位置、引用功能和引用情感三个维度对引用句的引用信息进行标注。这三个维度之间彼此关联，互为补充，共同构成了全文引用分析的对象和研究内容。

2.2.1 引用句抽取

在全文引文分析视角中，引用信息都是基于文献中的具体引文内容（引文句）来产生的。在这种情况下，施引文献与被引文献之间的引用模式不再是简单的“引用”和“被引”的二元关系，而是变成了“施引文献”“引用句”和“参考文献”的三元关系。这与题录视角的引文分析有着本质的不同，因此分析结果也会产生较大差异。在构建好实验数据集以后，下一步就需要对施引文献的引文内容（引用句）进行抽取，从而与被引文献产生一一映射。笔者首先根据引用句的特征构建了统一的匹配规则。然后，在此基础上辅以人工校验和清洗，共计得到了1 010条引用句及其引用关系。获得的部分抽取结果如表1所示，包含施引文献编号、被引文献的编号、引用句内容以及参考文献的标题。最后，将1 010条引用句逐一编号，构成引用句集合。

需要注意的是，在不同的参考文献引用格式（如GB/T 7714、MLA、APA等）中，文献正文引用行为的标注方式是不同的。在前文构建的数据集中，所有的参考文献格式均是按严格 的APA（American Psychological Association）格式来进行标注的，所以其正文的引用句没有在上角方括号内进行标注，而是以圆括号的形式将被引作者和年代分别陈列出来。以引用 句 为“We may then have clustering solutions

that include many thousands or even many millions of publications (e.g., Boyack and Klavans 2014; Klavans and Boyack 2017; Waltman and Van Eck 2012).”为例，该句括号中的引用标志由作者姓名和发表年代组成，不同标识之间用分号隔开，表明该句话共引用了三篇不同的学术文献，分别是：Boyack 和 Klavans 在 2014 年发表的论文《Including cited non-source items in a large-scale map of science: What difference does it make》、Klavans 和 Boyack 在 2017 年发表的论文《Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge》和 Waltman 和 Van Eck 在 2012 年发表的论文《A new methodology for constructing a publication-level classification system of science》。通过与 CitedSet 中的被引文献匹配，得到该引用句与被引文献 #496、#1173 和 #43 产生三条引用关系。

2.2.2　全文引文信息标注

在全文引用行为分析过程中，笔者主要从引用功能、引用情感、引用位置三个角度展开讨论。根据相关文献，这三点是目前全文引文分析方法中最常用的三种分析维度。这三者之间彼此关联，互为补充，共同构成了全文引用分析的对象和研究内容。

（1）引用功能标注

引用功能表示被引文献对施引文献发挥的功能作用，如引用背景、引用信息、使用数据、使用方法、使用结论、观点扩展、观点驳斥等。引用功能的标注采用内容分析法的编码过程：①组成由三人构成的临时编码小组：两名经过提前培训的图书情报学专业的研究生（研究方向为引文分析）担任编码员，一名图书情报学科教授担任领域专家，负责编码结果检查与验收；②两名编码员分别独立地阅读施引文献，理解作者引用参考文献的实际含义。在此基础上对引用句中的引用功能进行标注；③对两位编码员的标注结果进行一致性检验，当编码一致性大于 90% 时认为编码准确率较高，编码结果通过；④由领域专家对编码结果进行统一校对，并产生最终的标注结果。需要注意的是，当单一的引用句不足以分析出特定的引用功能时，编码员必须返回引用句在文献中所处的位置，阅读引用句的上下文。这些上下文句子通常与引用句有着千丝万缕的关系，因此在引用功能标注上也会起到一定作用。

（2）引用情感标注

自引用情感的问题被提出以来，研究人员就从不同角度建立了引用情感的分类体系，用于对引用情感进行标注。这些分类体系不尽相同，但大致上可以分为积极（Positive）引用、中立（Neutral）引用和消极（Negative）引用三大类。本文在引用情感的标注方面采用了 David Shotton 和 Silvio Peroni[23] 的引用情感分类体系。这种方法把引用关系分为事实型关系和修辞型关系两大类，事实关系（如使用方法、结论、数据等）是对引用事实的具体描述，不具备情感上的态度倾向；而修辞关系（如表扬、综述、批评、讨论等）是具有明显情感倾向的引用关系，可以标注引用情感，并可进一步细分为积极、消极、中立三类。本文中引用功能与引用情感的分类情况如表 2 所示。

（3）引用位置标注

通常情况下，科学论文的章节标题各不相同，这为引用位置的统一分析带来了一定困难，所以必须进行标准化处理。大多数英语论文都有固定的章节格式（一级子标题），以四节式和五节式结构较为常见。在四节式论文著名的 IMRD[24] 结构体系中，第一节是论文的引言“Introduction”部分；第二节论述论文的数据和研究方法，常见的标题包括“Methods”“Methodology”“Data and methods”等；第三节是研究结果“Results”部分；第四节是结论“Conclusion”或讨论“Discussion”部分。此外，五节式和六节式的结构基本上都是对四节式的扩展，具有多样性特征 [20]。

图 2 是对施引文献数据集 CitingSet 中 25 篇文献的文献结构进行统计分析的结果。由该图可知，《科学计量学》杂志引文分析领域的高被引文献的段落结构主要由四节式和五节式构成（共 22 篇，占 88%）。其中，四节式论文的章节结构基本满足 IMRD 的分布规律；五节式论文大多是四节式的变形（如将“Discussion”和“Conclusion”分成两节，或

将“Methodology”和“Data”分成两节等）；另外还有部分五节式文献单独增加了一节，作为文献综述部分（如“Related work”等）。因此，基于以上分析和统一标准的考虑，笔者对施引文献的引用位置进行标准化处理，将引用位置统一定为“Introduction”“Related work”“Methods”“Results”和“Conclusion”的IRMRC五节式结构。另外，还有极少数引文出现在文献的摘要部分，标注为“Abstract”。编码员在进行引文位置标注时，需要先仔细阅读文献，理解文章的段落结构。然后，按照相应段落在文章中发挥的作用和功能进行标注。以施引文献#1为例，其第二节的标题虽然为“Clustering technique”，但经过阅读原文发现该节实际论述的就是数据和方法论，因此标注为“Methods”。同理，可以将#4和#24中的标题“A brief history of topic detection”和“Time frame and milestones”标注为“Related work”，将#11中的“Case study：3D printing”标注为“Results”等。经过标准化处理后的引用位置具有统一的标准，为后期进行引文位置分析奠定基础。此后，由编码员对引用行为发生时所处的章节结构进行位置标注。同样，两位编码员的编码结果需要进行一致性检验，最后将编码结果交由领域专家统一审核确认。

3 引文关联数据发布

数据集构建和处理完成之后，就可以考虑引文关联数据的发布了。通过关联数据技术，将全文引文数据转换成统一的、机器可读的RDF格式三元组，并发布在服务器上供用户进行SPARQL检索和查询，为后续的全文引文分析奠定基础。

3.1 全文引文本体构建

本文主要从引用功能、引用情感、引用位置三个角度出发，构建全文引文本体FCO（Full-text Citation Ontology，FCO），对全文引文信息进行全方位描述。在引文功能方面，本文在复用CiTO本体属性的基础上，共构建了包括“cito：reviews”“cito：citesForInformation”“cito：usesConclusionsFrom”“cito：obtainsBackgroundFrom”“cito：usesMethodIn”“fco：citesAsDefinition”等在内的38个引用功能属性，用于对全文引用功能信息的全方位标注；引用情感表示施引文献对被引文献的态度或立场，包括积极、中立和消极三大类，分别用“fco：positivelyCites”“fco：neutrallyCites”和“fco：negativelyCites”标注；引用位置标注施引文献对被引文献发生引用关系时所处的段落位置。在FCO本体中，引文位置属性用“fco：hasCitationLocation”进行统一标注，引文位置的具体值用字符型数据表示，包括“Abstract”“Related Work”“Introduction”“Methods”“Results”和“Conclusion”6个部分。构建好的FCO本体的类和部分属性如图3所示。

3.2 全文引文数据的RDF转换

按照前文标注好的引用信息，使用程序将全文引用数据转换成对应的RDF/XML格式的RDF代码。其中，引用功能1 538条，引用情感694条，引用位置2 023条RDF，共计4 255条三元组。以施引文献#1为例，其在全文信用信息方面有着如图4所示的网络节点关系（部分）。

转换后得到的RDF/XML代码如图5所示。

该段代码的主体部分从上到下依次分别展示了引用功能、引用情感、引用位置标注结果。引用功能部分的代码表示施引文献#1与被若干引文献之间产生了“cito：citesAsRelated”“cito：citesForInformation”“cito：discusses”“cito：extends”“cito：obtainsBackgroundFrom”“cito：usesMethodIn”“fco：citesAsDefinition”等功能属性；引用情感部分代码表示施引文献#1中立引用了被引文献#1098和#43，积极引用了被引文献#405；引用位置部分代码表示施引文献#1所包含的引用句#102处在文献段落结构的“Methods”部分。

3.3 SPARQL端口与语义查询

Apache公司的Jena Fuseki是一款语义仓储软件，它可以存储各种格式的RDF数据并通过HTTP协议来响应不同的SPARQL查询。Jena Fuseki内置了功能强大的SPARQL编辑器，可以支持不同功能字段的高亮显示，还可以选择检索结果的输出格式，包括XML、JSON、CSV、TSV、Turtle、JSON-LD、N-Triples等[25]。在Jena Fuseki中构建标题为“Web of Science”

的数据库，并上传所有的全文引用信息 RDF 三元组。最后在本地服务器“http://localhost:3030/”上成功发布为引文关联数据。

引文关联数据的语义查询需要通过 SPARQL 语句来实现。一个标准的 SPARQL 查询如图 6 所示。该 SELECT 子句一共定义了 4 个查询变量，分别为施引文献（?citing）、被引文献（?cited）、施引文献标题（?citingtitle）和被引文献标题（?citedtitle）。WHERE 子句中的三元组表示施引文献（?citing）引用了（cito:cites）被引文献（?cited）。该 SPARQL 语句可以查询出施引文献与被引文献之间所有的引用关系。

4　实验结果及分析

4.1　引文数量分析

在传统视角的引文分析方法中，引文的数量被默认为是参考文献的数量。由于每篇参考文献在文献末尾只列出一次，因此这种计算引文数量的方法实际上忽略了同一参考文献被多次引用的情况，由这样的引文数量所进行的引文分析，其结果也一定存在偏差。相比之下，全文引文分析从引文内容出发，标注每个引文句中的引用行为。因此，引文数量的计算更加科学。基于全文信息视角的引文数量分析，首先要检索出每篇施引文献所含的引文数量。

图 7 中展示的是查询每篇施引文献的引文数量的 SPARQL 代码。其中，前缀（PREFIX）部分复用的本体 / 词汇表包括 rdf、dct、cito、fco 和 xsd，相应的命名空间在尖括号中显示。SELECT 子句选定施引文献标题（?citingtitle）和引文数量（?citingcount）作为查询对象。WHERE 子句限定检索对象的查询范围，三元组表示的信息是引用句（?citingsentence）属于施引文献（?citingpaper），并且被引文献（?citedpaper）被施引句（?citingsentence）所引用。最后用 GROUP BY 和 ORDER BY 子句进行结果输出。

对两种视角下的引文数量分析进行对比。表 3 中显示的是 CitingSet 集合中每篇施引文献在传统视角和全文视角下的引文数量对比情况。

从表 3 可以看出，除文献 #15 在两种视角下的引用数量一致外（均为 84 次），其余 24 篇文献在全文信息视角下的引用次数都比题录视角下增加了 103%–236%。这种数量上的增加是符合常理的，并且出现多次引用同一篇文献的情况越多，其全文视角下的引用次数变化就会越大。经过仔细分析后发现，#15 出现两种视角下次数相同的原因是其所有的 84 次引用行为中均引用了不同的参考文献，但这样的标注引用方式在科学文献的实际分布中是非常少见的（样本中所占的比例仅为 4%）。这也从侧面说明：在题录视角下默认情况将所有参考文献同等对待，并且将参考文献数量定为引用次数的分析方法是不合理的，引文分析结果的可信度也会大打折扣。

4.2　引文功能分析

传统引文分析法只能判断文献之间引用与否，不能揭示更深层次的引用语义关系。引用功能分析的作用是对作者引用参考文献的目的进行深度剖析。对引用功能进行 SPARQL 检索的结果如图 8 所示。

在《科学计量学》杂志引文分析领域高被引论文的引用功能属性中，排名前两位的分别是 cito：reviews 和 cito：citesForInformation（共出现 1 393 次，占 75.7%）。cito：reviews 属性表示对被引文献进行回顾或文献综述，cito：citesForInformation 属性则代表引用了被引文献中的部分信息，这两种属性占引用功能的大多数。在科学论文的撰写过程中，通常需要对前人的研究进行总结回顾，因此使用最多的就是 cito：reviews 属性；同时，作者在撰写论文时经常需要使用参考文献中的一些信息来支持施引文献中引用句的撰写，因此也产生了大量的 cito：citesForInformation 属性。此外，这样的标注结果也与分析领域有着很大关系。引文分析法是对文献的引用和被引用现象进行分析的方法。与其他领域不同，引文分析领域常常将学术文献作为分析和研究的对象，这样特殊的分析模式也是产生大量 cito：citesForInformation 属性的原因。排名 3–6 位的引用功能 cito：usesConclusionsFrom、cito：obtainsBackgroundFrom、cito：usesMethodIn、fco：citesAsDefinition 分别代表施引文献引用

了被引文献中的结论、背景、方法和定义，这些引用功能的作用更加具体，相对比例维持在3%–5%之间。排名7–11位的引用功能出现次数较少，所占比例在1%–3%之间。另外由8个属性的出现次数极少，均不足20次，合计所占比例为2.88%。

通过以上分析可以看出，尽管全文引文本体FCO中包含了相当多的属性用于对引用功能进行全方位地标注，但在实际使用过程中，只有11个左右的引用功能属性比较常用（所占比例共计97.22%）。笔者将这11个常用的功能称为“F11”，在大多数情况下（尤其是在引文分析领域），用这11个属性来标注引文功能足够了。

接着，构建SPARQL检索式检索出每篇施引文献的引用功能属性及其次数，检索结果如图9所示。可以看出，不同引用功能属性在每篇施引文献中所占的比例不同。首先，cito：reviews属性在#15、#4和#22中的所占的比例非常高，分别达到100%、89.7%和69.2%。施引文献#15比较特殊，其引用功能属性只有cito：reviews一种。通过仔细对比发现其引用的84篇文献均作为文献回顾与综述，因此产生这样的结果。#4和#22则由于文献正文撰写了大篇幅的文献综述，因此cito：reviews引用功能属性所占比例较高。此外，有15篇文献的cito：reviews属性所占的比例在20%到55%之间，表明该属性在大部分文献的引用功能标注中均具有很强的作用。其次，cito：citesForInformation属性在各施引文献中的分布比较均匀，在25篇文献中所占的比例根据文献研究内容的以及作者的引用习惯在24%–73%之间波动。在#4和#16所占的比例较少，在#15中占比为0。最后，cito：usesConclusionsFrom属性占比较高的文献有#16、#5和#3，表明这些文献在撰写时更多地引用了参考文献的结论部分；cito：obtainsBackgroundFrom属性占比较高的文献由#1、#23和#21，表明这三篇施引文献引用参考文献作为背景信息的情况较多；cito：usesMethodIn属性所占比例较高的有#2、#1、#5和#12，表明这些文献更多地引用了参考文献中的研究方法。

4.3 引文情感分析

引用情感表明了作者对被引文献的态度倾向。不同的引文情感表示不同的引用目的。《科学计量学》杂志引文分析领域高被引论文的引用情感主要集中在中立引用（所占比例为95%）。在学术文献的撰写过程中，作者通常站在一个相对比较中立、客观的角度来进行分析。因此，中立态度地引用（包括cito：reviews、cito：discusses等）占据引用情感属性的绝大多数。同时，需要注意的是，依然有4%的积极引用和1%的消极引用存在。这些带有明显情感态度倾向的引用行为可以帮助我们快速定位相关，对了解知识背景和文献脉络有着重要作用。

同理，使用SPARQL语句检索出每篇施引文献的引用情感属性计数排列。最终得到的结果如图10所示。

从图11可以看出，在25篇施引文献样本中，只有#2、#3和#8三篇文献包含了消极引用，并且所占比例较少。有11篇文献包含了积极引用，其中#13较为特殊，6条引用情感态度均为积极。

4.4 引文位置分析

最后，引用位置的不同也表明了作者引用的不同目的。通过前文的统一标注，25篇施引文献中引用行为所处的位置已经全部被标注成“Abstract”“Introduction”“Related Work”“Methods”“Results”和“Conclusion”的六段式结构，因此可进行统一分析和处理。为了扩展研究角度，笔者将前文提及的引用功能和引用情感与引用位置相结合，进一步进行多维分析。为此，笔者编写了更加复杂的SPARQL代码，用于多维数据的查询和统计。使用如图11所示的SPARQL语句可以查询出发生在“Introduction”部分的引用情感属性，并且按次数进行降序排列。同理，可以检索出每一段落部分的引用情感属性。

引用位置在不同引用情感下的分布情况如图12所示。从中可以看出，消极引用（fco：negativelyCites）主要出现在结论（Conclusion）部分，这主要是由于学术文献的结论部分通常需要作者对研究研究进行个人总结，所以

会包含一些情感态度倾向比较明显的消极引用，如批评、不同意等；中立引用（fco：neutrallyCites）在各个段落位置的分布较均匀，并且出现在 Introduction、Related Work、Results 等部分的比例较大，在 Methods 和 Conclusion 部分所占的比例较少。通过对比原文后发现，这主要是由施引文献数据集中各段落内容的长短多少来决定的。通常情况下，论文的 Introduction、Related Work、Results 段落相对较长，描述的内容更多一些，所以更容易出现更多的中立引用；积极引用（fco：positivelyCites）出现在 Conclusion 和 Related Work 部分的比例较大，出现在 Conclusion 的原因与前面类似，即该部分通常会出现更多的态度倾向明显的观点总结。此外，作者在 Related Work 部分对相关文献进行回顾和总结时，有时也会添加一些个人情感的论述（如称赞、支持、扩展等）。

对每一段落部分的引用功能属性进行检索，结果如图 13 所示。从引用功能属性在不同引用位置中的分布情况来看，样本文献的 Abstract 部分由大量的 cito：reviews 和极少数的 cito：citesForInformation 属性构成，但由于样本点较少，所以结果不一定具有概括性。Introduction 位置的引用主要由 cito：citesForInformation、cito：reviews 和 cito：obtainsBackgroundFrom 三种功能属性构成（所占的比例共计为 84.4%）。这三种属性分别代表引用信息、文献回顾和引用背景知识，符合 Introduction 段落的撰写要求。值得注意的是，在样本文献的 Related Work 位置中，引用功能基本全部由 cito：reviews 构成（所占的比例为 83.8%），只有极少部分采用了其他引用功能属性。这表明引用功能 cito：reviews 与引用位置 Related Work 之间有着极强的相关性关系。最后，在 Methods、Results 和 Conclusion 部分，引用功能属性的分布情况比较一致，cito：reviews 和 cito：citesForInformation 属性占据绝大多数，其余属性出现次数较少。同时 cito：usesConclusionsFrom 属性在 Conclusion 部分分布较多，表明二者之间存在着一定的相关关系。

5　结语

本文从全文引文分析视角出发，对基于本体和关联数据的全文引文分析方法进行实证研究，详细步骤包括数据集的构建、全文引用信息抽取和标注、全文引文信息关联数据发布和引文分析试验等。在实验设计部分，我们选取了引文数量、引文功能、引文情感和引文位置四个角度分别构建对应的 SPARQL 检索式来进行查询分析，并且进一步扩展研究角度，进行了引文情感与引文位置、引文功能与引用位置的多角度综合分析。实验表明，与传统视角下的引文分析方法相比，全文信息分析方法的建模方式更加合理。本文所提出的基于本体和关联数据的全文引文分析方法具有一定的可操作性和实用性，这对传统引文分析法的优化和全文引文分析方法的普及都有着重要意义。

当然，本研究不可避免地也存在着一些不足之处。在全文引文数据的构建部分，笔者通过人工标注完成了实验数据集中文献的引用功能、引用情感和引用位置的属性标注，并在标注结果的基础上开展了全文引文分析实验。尽管全文引用信息的标注经过两名编码员的一致性检验和领域专家的人工校对，但标引结果的可靠性还是存在疑问的。事实上，要想从第三方角度完全理解作者在引用参考文献时的真实想法几乎是不可能的，这也是全文引文分析方法迄今为止发展受阻的最大原因。本文侧重于方法的提出和实验验证，因此对于数据编码方面出现的不可控因素，暂不作过多考虑。后续研究将从方法适用性的角度入手，继续考察基于本体和关联数据的全文引文分析方法在其他数据源（如 Scopus、CNKI 等）和其他学科领域（如物理、化学）的有效性。研究人员可以充分利用基于本体和关联数据的引文分析方法进行学术研究探索，辅助科研发现创新。

参考文献

[1] 邱均平. 信息计量学(九)第九讲 文献信息引证规律和引文分析法[J]. 情报理论与实践, 2001, 24(3): 236–240.

[2] 祝清松. 语义增强的引文分析方法与应用实验研究[D]. 北京: 中国科学院大学, 2014.

[3] Ding Y, Zhang G, Chambers T, et al. Content–based citation analysis: The next generation of citation

analysis[J]. Journal of the Association for Information Science and Technology, 2014, 65(9): 1820–1833.
[4] 赵蓉英, 曾宪琴, 陈必坤. 全文本引文分析——引文分析的新发展[J]. 图书情报工作, 2014, 58(9): 129–135.
[5] Berners–Lee T. Linked Data[EB/OL]. [2018–04–02]. https://www.w3.org/DesignIssues/LinkedData.html.
[6] 石泽顺, 肖明. 基于PoolParty的图情学科SKOS叙词表构建研究[J]. 图书馆学研究, 2017(23): 20–30.
[7] 郑巧. 基于引文关联数据服务的学术期刊资源建设[J]. 图书馆界, 2015(6): 9–11.
[8] 石泽顺, 肖明. 基于RelFinder的图情学科关联数据语义关系发现实践[J]. 图书情报工作, 2017, 61(17): 139–148.
[9] Garfield E. Can Citation Indexing Be Automated[C]. Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, Washington, 1964: 189–192.
[10] Weinstock M. Citation indexes (part 1). Encyclopedia of Library and Information Science[M]. New York: Marcel Dekker, 1971: 16–40.
[11] 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.
[12] Thorne F C. The citation index: another case of spurious validity[J]. Journal of Clinical Psychology, 1977, 33(4): 1157–1161.
[13] Brooks T A. Private acts and public objects: an investigation of citer motivations[J]. Journal of the American Society for Information Science, 1985, 36(4): 223–229.
[14] Vinkler P. A quasi–quantitative citation model[J]. Scientometrics, 1987, 12(1): 47–72.
[15] Bonzi S, Snyder H W. Motivations for Citation–A Comparison of Self Citation and Citation to Others[J]. Scientometrics, 1991, 21(2): 245–254.
[16] Liu M. A study of citing motivation of Chinese scientists[J]. Journal of Information Science, 1993, 19: 13–23.
[17] Case D O, Higgins G M. How Can We Investigate Citation Behavior? A Study of Reasons for Citing Literature in Communication[J]. Journal of the American Society for Information Science, 2000, 51(7): 635–645.
[18] Tang R, Safer M A. Author–Rated Importance of Cited References in Biology and Psychology Publications[J]. Journal of Documentation, 2008, 64(2): 246–272.
[19] Zhang G, Ding Y, Stasa Milojevic. Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content[J]. Journal of the Association for Information Science and Technology, 2013, 64(7): 1490–1503.
[20] 胡志刚. 全文引文分析方法与应用[D]. 大连: 大连理工大学, 2014.
[21] 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.
[22] Scientometrics: An International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science Policy[EB/OL]. [2018–04–02]. https://link.springer.com/journal/11192.
[23] Shotton D. CiTO, the Citation Typing Ontology[EB/OL]. [2018–04–20]. https://jbiomedsem.biomedcentral.com/articles/10.1186/2041–1480–1–S1–S6.
[24] Burrough–Boenisch J. International reading strategies for IMRD articles[J]. Written Communication, 1999, 16(3): 296–316.
[25] 石泽顺, 肖明. 基于网络叙词表的图情学科SKOS构建与可视化研究[J]. 情报学报, 2018, 37(3): 274–284.

石泽顺 北京师范大学政府管理学院，硕士研究生。研究方向：语义网、信息计量。作者贡献：设计论文整体研究思路并撰写论文。E–mail: shizs10@lzu.edu.cn 北京 100875
肖 明 北京师范大学政府管理学院，教授。研究方向：语义网、信息计量。作者贡献：修改论文，补充论据以及进行实验指导。 北京 100875

（收稿日期：2019–03–18 修回日期：2019–08–01）