DOI:10.13530/j.cnki.jlis.184005

科学论文功能单元本体设计与标引应用实验*

王晓光　李梦琳　宋宁远

摘　要　科学论文内容本体是科学论文内容结构和语义功能的形式化和规范化知识表示,对于科学论文的深度标引和知识挖掘具有重要意义。本文系统梳理了已有科学论文内容表示模型和内容本体,并以功能单元理论为基础,提出科学论文功能单元本体的设计思路,构建包含 28 个类和 5 种属性在内的科学论文功能单元本体 FUO。借助本体构建工具 Protégé,对科学论文功能单元本体 FUO 进行形式化表示。借助语义标注工具 GATE,利用功能单元本体 FUO 对论文进行初步的深度标引实验,检验了该本体的可用性。结果表明,功能单元本体 FUO 能够很好地表示科学论文内容组件的语义功能及其属性,揭示科学论文正文各部分的语义特征,可以用于面向知识发现的科学论文深度语义标引,为科学论文内容本体开发奠定了基础。图 4。表 9。参考文献 47。

关键词　科学论文　功能单元　内容本体　本体构建　深度标引

分类号　G254

Design and Application of Scientific Paper Functional Units Ontology

WANG Xiaoguang, LI Menglin & SONG Ningyuan

ABSTRACT

With the increasing of knowledge resources and the demand of knowledge mining, it is important to enrich the semantics of academic literature, which can not only help users quickly and accurately locate the knowledge units in scientific papers, but also can help readers to conduct comparative analysis and strategic reading. Therefore, it's essential to identify and describe the components and their semantic functions within scientific papers for promoting knowledge discovery and knowledge services.

Scientific paper content ontology is the standardized knowledge representation of scientific papers' content structure and semantic function. It is of great significance for the deep indexing, information extraction and knowledge mining of scientific papers. After a review on the existing researches of paper components and attributions as well as the published ontologies, the existing ontologies, limited to the fundamental theories, have some deficiencies in revealing the deep semantics of information embedded in scientific papers. In order to design and build a component ontology, which is more suitable for information extraction, the

* 本文系教育部人文社会科学重点研究基地重大项目"大数据资源语义表示与组织"(编号:16JJD870002)和国家社会科学基金重大项目"基于认知计算的学术论文评价理论与方法研究"(编号:17ZDA292)的研究成果之一。(This article is an outcome of the project "Semantic Representation and Organization of Big Data Resources" (No. 16JJD870002) supported by Foundation of Key Research Institute of Humanities and Social Science at Universities, Chinese Ministry of Education and the major project "Theories and Methods of Research Papers Evaluation Based on Cognitive Computing" (No.17ZDA292) supported by the National Social Science Foundation of China.)

通信作者:王晓光,Email:wxguang@whu.edu.cn,ORCID:0000-0003-1284-7164(Correspondence should be addressed to WANG Xiaoguang, Email:wxguang@whu.edu.cn, ORCID:0000-0003-1284-7164)

functional unit theory should be considered.

The functional unit theory is the fundamental theory that combines information tasks and genre analysis, which is more suitable for the development of scientific paper content ontology oriented to knowledge discovery. Based on the functional unit theory, a novel ontology named Scientific Paper Functional Units Ontology(FUO) is designed. After reviewing the 41 functional units, 28 components are redesigned, including background, goal, motivation, method-description, conclusion, contributions, etc. Based on the components, 12 classes and 28 subclasses are designed. The attributions of the classes are also designed by refering to Bio-Event ontology and News-Event ontology. The classes and attributions of FUO are formally represented with protégé 5.1. Then 10 research papers from JASIST are randomly selected to conduct a deep indexing experiment by using the GATE, a semantic annotation software. Finally, the distribution of different functional units within scientific papers is analyzed.

The originality of this research lies in the clear definition of the functional units with their attributes and the FUO which can reveal semantic features of scientific papers components in a more comprehensive and detailed manner. The results have also proved the potential availability of FUO for deep semantic indexing, semantic retrieval and knowledge discovery. This research deepens our understanding on scientific paper as a knowledge container from the perspective of information science. The limitation of this paper is the lack of considering the semantic relationships between content components of scientific paper. More detailed definition of the relationships and new components such as interactive tables, datasets, audios and videos should be studied in the future. 4 figs. 9 tabs. 47 refs.

KEY WORDS

Scientific papers. Functional units. Content ontology. Ontology construction. Deep indexing.

0 引言

随着科学研究事业迅速发展,科学知识增长开始加速,直接表现为科学论文数量的暴涨。科研工作者每年的阅读量在日益增长,而单篇文献的平均阅读时间却在逐渐下降,读者不得不在短时间内找到并阅读完自己感兴趣的文献[1]。从读者的理解对象来看,真正有价值的是这些科学文献内的特定部分而非全文。随着知识资源形态的多样化和用户知识需求的精准化,学术文献呈现出细粒度和语义化发展趋势[2]。在这样的环境下,帮助用户快速准确定位科学论文中的情报单元,并进行比较分析和战略阅读,就显得尤为重要。

科学论文的组成结构复杂,从形式结构来看,一般包括题目、摘要、表格、图片、参考文献、脚注、词组、句子等内容[3]。从逻辑结构来看,论文一般包括引言、文献综述、研究方法、结果和讨论等部分[4]。规范描述并准确表征论文不同内容部分的语义属性,是实现论文深度语义标引、情报抽取、知识挖掘和知识发现的基础[5-7]。科学论文内容本体设计在语义出版研究领域已经得到广泛关注。基于不同的理论和实践视角,不同形式的科学论文内容本体也陆续被提出[8-9]。但这些本体普遍关于论文的形式结构和修辞结构,没有对论文的语义功能结构进行理想的表达,限制了情报的自动抽取和知识发现。

本文参考已有的科学论文内容表示模型和本体模型,重点基于功能单元理论,设计了一种新型的科学论文功能单元本体(Functional Units Ontology,FUO),从内容组件的语义功能角度对科学论文的组织结构进行细粒度建模,并

采用标注实验对该本体进行可用性实验,检验该本体在科学论文内容语义功能表示上的能力。

1　相关研究综述

1.1　科学论文的内容组成部分及其属性

理解科学论文的内容组成部分及其结构具有重要意义。从语言学角度揭示作者的写作意图、文本功能、修辞结构等语义特征,对内容组件进行分类识别,有助于实现更高层次的知识组织与资源聚合[10]。基于此思想,研究者提出了诸多具有广泛影响力的科学论文内容结构模型,如引言—方法—结果—讨论(Introduction-Method-Result-Discussion,IMRD)模型。该模型将科学论文内容分为引言、方法、结果与讨论四大部分[11]。基于此模型,Teufel 在修辞理论指导下提出了科学论文论证块(Argumentative Zoning,AZ)模型,该模型表征了论文内部的论证结构[12]。随后,Teufel 又对这一模型进行扩展,提出了更细粒度的框架 AZ II,将作者情感倾向与文本修辞功能进行结合,强调对不同观点的比较[13]。Liakata 等[14-15]认为科学论文主要是为了阐明科学调查过程,总结实验结果,得出科学结论,因此提出了包括 11 个核心知识单元在内的核心科学概念(Core Scientific Concept,CoreSC)模型。de Warrd[16]也设计了一个包括 7 个类别的篇章块模型(Discourse Segment)。基于以上这些模型,研究者对科学论文进行了语义标注实验,构建了一些研究性语料库[17],分析了各种内容部分的语言学特征,探讨了科学论文的知识建构问题,并且为文本自动标注提供了一定的基础规则[18]。有研究表明,利用多种模型进行协同标注可以实现更丰富的语义揭示[19-20]。

除了对论文内容组成部分进行分类建模之外,对不同组成部分的状态和属性进行描述也必不可少。“元知识”(metaknowledge)是 Evans 在 2011 年提出的一种描述知识生产的时间、类型及作者观点态度的概念[21]。元知识是隐藏在论文表层符号之下的潜在知识情报,揭示这些知识有助于科学交流和情报分析。Thompson 等人对元知识进行了细致的定义,提出了 EventMine-MK 模型,并针对生物医学领域科学论文及新闻文本分别设计了 Bio-Event[22]、News-Event[23]等元知识模型,涉及知识类型、可信度、极性、程度、来源等多维属性。de Warrd 等[24]研究了论文内容的知识属性(Knowledge Attribution)和认知情态(Epistemic Modality),进而提出了科学论文中的命题属性模型,包含确定性等级、基础、来源三个方面。此外,在引文关系研究中,陆伟等也提出了引文的 8 种属性特征,如被引文献类型、被引频次、是否自引等[25],以支持更深层次的引文分析。由此可见,对科学论文内容组件的属性进行多维度的描述,不仅可以深入揭示内容组件的语义功能,还能更好地支撑基于文献的情报分析与知识发现。

1.2　科学论文内容本体

科学论文内容本体是借助本体技术对科学论文内容组成部分的规范化和形式化的表示,是科学论文组成结构的知识表示模型[26]。近年来,为了满足语义出版需要,科学论文内容本体逐渐产生。2007 年,Groza 等人提出的 SALT 本体定义了背景、讨论、结论、动机、情景等论文内容组件[27]。2011 年,W3C 提出了修辞块本体(ORB),将科学论文分为头部、主体、尾部三大部分,并参考 IMRD 模型对主体部分进行了细致定义,包括引言、方法、结果、讨论等部分[28]。2012 年,Peroni 等人提出了描述科学论文修辞结构的篇章元素本体 DEO[29],随后又提出了描述文献内容组件的 DoCO 本体[30],明确定义了诸如背景、作者贡献、材料、方法等多个内容组件。此外,Peroni 等人还结合引文本体 CiTO[31]、参考文献本体 BiRO、出版流程本体 PWO、出版角色本体 PRO,共同构建了用以支撑语义出版的本体集 SPAR[32]。在 SPAR 之外,Peroni 还提出了 AMO 本体,定义了科学论文的论证

结构[33]。

目前,科学论文内容本体在内容抽取、语义关联方面得到了应用,但主要集中在生物医学领域。Utopia Document[34]、Biotea[35]等项目借助DoCO本体对医学和生物学领域科学论文进行语义标引,同时结合领域本体,开发了知识库和语义出版系统。SLOR项目[36]也综合应用科学论文内容本体和引用本体对科学论文内容进行关联描述与语义发布。

总的来说,现有的论文内容本体多以修辞结构理论和语篇分析理论为基础,对论文内容组成部分进行表示和建模。这种本体模型因为理论视角的问题,无法深入揭示科学论文中的情报信息。事实上,科学论文中情报信息的揭示需要结合语言学的体裁分析、情报学的知识单元理论、信息搜寻理论和知识发现理论等来综合处理。为了构建更加适合情报分析和知识发现的科学论文内容本体模型,我们需要特别关注内容组成部分的语义功能。

2 功能单元理论

功能单元理论(Functional Units Theory)是由Zhang Lei[37]在2010年提出的关于科学论文语篇结构及内容组织方式的理论。Zhang Lei认为,功能单元是能够满足不同科学交流功能、实现知识传播任务的最小内容单元,这些内容组件分布在论文的引言、方法、结果和讨论四大部分[37]。功能单元理论借鉴了Swales的研究空间理论(Creating a Research Space,CARS)、语步分析[38]等体裁分析理论,识别出科学论文中的41个功能单元。同时,功能单元理论在信息使用模型的基础上,定义了读者阅读科学论文的五大常用信息使用任务,包括学习背景知识(Learn about background)、学习方法(Learn how to)、参考事实(Refer to facts)、参考论证(Refer to arguments)、跟进研究前沿(Keeping up),并将细粒度功能单元与具体的信息使用任务进行关联,揭示出不同类型情报的语义功能属性。

基于Zhang Lei提出的功能单元理论[39],表1详细表示了功能单元、论文IMRD结构模型和具体信息使用任务之间的关系。功能单元理论共区分了三类功能单元,参照IMRD结构模型定义的科学论文四大组成部分,第一类功能单元是与当前信息任务最直接相关的功能单元,第二类是在IMRD结构中对第一类功能单元的深入阐述,第三类则是与当前信息任务相关,但分布在其他IMRD结构中的功能单元。

从表1可以看出,不同的功能单元与特定信息使用任务之间存在一定的关联关系,特定部分的功能单元发挥着特定的作用,如在引言部分,功能单元"前人研究综述"和"前人研究贡献"在信息使用任务"学习背景知识"中发挥着最重要的作用。另外,同一功能单元在不同信息使用任务中也可以扮演不同的角色。各功能单元按照信息使用任务也形成了一定的层级结构。所以说,功能单元理论是面向信息任务、融合体裁分析的关于科学论文内容组织方式的基础理论。借助功能单元可以有效提升阅读过程中的内容导航、文献精度和深度理解,能够满足用户对于科研论文的不同需求,帮助用户实现战略阅读[39-41]。

相较于修辞、论证等语篇分析理论,功能单元理论具有以下三点特征,使其更加适合面向深度标引及情报发现的科学论文内容本体构建:①功能单元理论针对科学论文的语义功能和独特语境,规定了科学论文内容组件的类型及属性,定义更为全面、准确;②相较于一般的科学论文内容结构模型,功能单元理论对内容组件的定义更为细致,对情报功能的表达更为充分;③功能单元理论探讨了不同内容组件的功能及对用户信息使用任务的作用,将具体的内容组件与特定的信息任务进行了关联,可以用以支撑面向特定信息需求的检索与知识发现。因此,功能单元理论适合用于科学论文内容本体开发。

表 1　信息使用任务、IMRD 结构与功能单元之间的关系

信息使用任务	最相关的功能单元	同一组成部分的相关功能单元	其他组成部分的相关功能单元
学习背景知识	I:前人研究综述	I:前人研究贡献;前人研究空白;缩小主题范围;阐明定义	M:相关实验;论证方法 R:总结结果 D:提供主题既有知识;与前人研究的结果对比
参考事实	R:陈述结果	R:评估假设结果;总结结果	I:前人研究综述 M:概述实验步骤;描述任务 D:突出整体结果
参考论证	D:支持结果的解释	D:与前人研究的结果进行对比;突出整体结果;解释结果;提供主题既有知识;概括一般性结论;指出预期/意外结果;表明结果的意义	I:前人研究空白;强调主题重要性 M:论证方法 R:陈述补充结果;陈述未经验证的结果;陈述结果
学习方法	M:描述材料;描述任务;概述实验步骤	M:论证方法;陈述变量;概述数据分析过程;前人研究方法;描述实验参与者;陈述可靠性/有效性	I:总结方法 R:描述分析 D:评价方法
跟进研究前沿	I:前人研究空白	I:研究缘起;前人研究贡献;前人研究综述;强调主题重要性;介绍本研究	M:论证方法;相关实验;概述实验步骤 R:陈述结果 D:指出未来研究方向;突出整体结果

注:I 表示引言(Introduction),M 表示方法(Method),R 表示结果(Results),D 表示讨论(Discussion)。

3　科学论文功能单元本体设计

3.1　本体设计目标

科学论文功能单元本体(FUO)的设计目标是,从语义功能角度准确定义科学论文内容组件的类型及其属性,构建科学论文内容结构表示模型,并利用规范的本体表示技术,建立可共享和可重复使用的科学论文功能单元本体。

3.2　功能单元类型调整原则

功能单元理论借鉴了大量诸如体裁分析、语步理论等语言学基础理论,因而在类目设置上具有较强的语言学特征,存在功能重复、累赘等现象。例如,引言部分的“提出假设”(Present hypotheses)和结果(Results)部分的“重述假设”(Restate hypotheses)均表示具有“假设”语义功能的内容组件。同时,功能单元的具体类目除了对当前内容组件的语义功能进行概述之外,通常还包括情感倾向、来源等属性信息。例如,“前人研究空白”(Indicate a gap in previous research)既包含了“当下组件的语义功能”(Indicate a gap),又表示了组件的来源(Previous research)。

为了更加清晰地区分内容组件及其属性,我们首先对 Zhang Lei 提出的 41 个功能单元进行调整。一是将具有相似含义的类目进行合并,如将“强调主题重要性”(Claim importance of topic)和“陈述本研究价值”(State value of present

research)合并为“研究意义”(Significance)。二是排除掉含义较为模糊或适用性不强的类目,如“防止反诉”(Word off counterclaim)。三是新增类目。数据一直都是科学研究必不可少的一部分,因此本研究借鉴 DEO 本体,新增“数据”(Data)组件,用于对数据集本身及数据分析过程的描述。四是重新界定名称及含义。功能单元类目名称多为“动宾结构”,如“陈述结果”“阐明定义”等,本研究参照多数科学论文内容结构模型,将类目名称改为“名词”形式,如“结果”“主题相关定义”等,以准确界定其表示的含义。五是确定各个内容组件的属性信息,借鉴已有的元知识表示模型,从来源、组件类型、确定性程度等多个维度确定内容组件的属性。

3.3 本体中的类设计

基于以上调整原则,我们设计了包含 12 个一级类、28 个二级类的功能单元本体基本模型。其中,二级类具备一级类的语义特征。我们对方法、讨论、实验、数据等情报价值较高的部分进行了深入划分与界定。例如,将方法部分定义为方法描述(Method-Description)、方法选择(Method-Selection)、方法评估(Method-Evaluation),将实验部分定义为实验素材(Experiment-Material)、实验任务(Experiment-Task)、实验步骤(Experiment-Procedure)与实验变量(Experiment-Variable),将讨论部分定义为要点重述(Discussion-Recapitulation)、对比(Discussion-Comparison)、贡献(Discussion-Contribution)、局限性(Discussion-Limitation),如表 2 所示。表 2 中的第四列“共现框架”指的是与本文定义的组件类型具有相同类目的框架,如本文定义的“背景”类,同时也存在于 AZ、AZ-II、CoreSC、DiscourseSegment、DEO 这些模型中。

表 2 科学论文内容组件定义

一级类	二级类	具体解释	共现框架
背景(Background)	背景(Background)	被广泛接受的事实;主题既有的知识陈述	AZ、AZ-II、CoreSC、DiscourseSegment、DEO
主题(Theme)	研究范围(Scope)	缩小研究范围,明确研究主题	CoreSC、DiscourseSegment、DEO
	研究目标(Goal)	陈述本研究要解决的研究问题和旨在达到的研究目的	AZ、AZ-II、CoreSC、DiscourseSegment
	定义(Definition)	主题内容/关键词相关概念的界定	
缘起(Origin)	动机(Motivation)	描述进行本研究的理由	CoreSC、DEO
	意义(Significance)	陈述主题的必要性和重要性	
已有研究(Existing research)	他人研究(Others)	回顾该领域的他人研究成果	AZ、AZ-II、DEO
	本人研究(Own)	作者先前研究成果	AZ、AZ-II
	已有研究的价值(Value)	已有研究对后人研究的价值与意义	
	研究空白(Gap)	指出现有研究的不足	
假设(Hypothesis)	假设(Hypothesis)	对未经证实的现象和事实提出的假定	CoreSC、DiscourseSegment

续表

一级类	二级类	具体解释	共现框架
方法(Method)	方法选择(Method-Selection)	实验方法、实验技术等的选择理由	AZ-II、CoreSC、
	方法描述(Method-Description)	简要描述本研究或实验中使用到的方法	AZ-II、CoreSC、DiscourseSegment、DEO
	方法评估(Method-Evaluation)	事后对方法的有效性进行评估	
实验(Experiment)	实验素材(Experiment-Material)	描述实验参与者、实验材料等	DEO
	实验任务(Experiment-Task)	描述实验任务	
	实验步骤(Experiment-Procedure)	描述进行实验的详细步骤	
	实验变量(Experiment-Variable)	介绍实验的研究变量	
数据(Data)	数据描述(Data-Description)	指数据本身(实验数据集、公式、表格、代码等)	DEO
	数据分析(Data-Analysis)	对实验过程中涉及的数据进行陈述、说明和分析	DEO
结果(Result)	结果描述(Result-Description)	陈述从研究或实验中得到的各种直接相关结果	AZ-II、CoreSC、DiscourseSegment、DEO
	结果评估(Result-Evaluation)	评估、分析关于假设的结果	DiscourseSegment、DEO
结论(Conclusion)	结论(Conclusion)	对结果的总结、归纳及推论,得出本研究的结论和见解	AZ-II、CoreSC、DEO
讨论(Discussion)	要点概括与重述(Discussion-Recapitulation)	通过重述研究问题、研究目的、理论或方法等来加强讨论	DEO
讨论(Discussion)	对比(Discussion-Comparison)	与前人研究成果进行对比	AZ、AZ-II
	贡献(Discussion-Contribution)	讨论本研究(结果)带来的贡献和启示	DEO
	局限性(Discussion-Limitation)	指出本研究的不足	AZ-II
未来工作(Future)	未来工作(Future)	陈述研究的下一步工作	AZ-II、DEO

相较于已有的修辞块本体、篇章元素本体等，本研究提出的本体具备多层次、多粒度的特征，能够更全面、细致地揭示科学论文内容组成部分的语义功能特征，能够满足论文深度标引的需求。

3.4 本体中的属性设计

科学论文中的功能单元通常都带有语境信息，如作者态度、观点确定性程度等。这些语境信息的缺失会影响计算机对文本的正确解读[42]。因此，科学论文功能单元本体必须表示这些属性。我们在参考 Bio-Event、News-Event 两个模型以及 de Warrd 的研究成果基础上，设计了 5 类属性，如表 3 所示。

表 3 科学论文功能单元本体中的属性

属性	属性值示例	具体解释	参考框架
确定性程度（Certain Level）	① low	缺乏事实依据，确定性程度低	Bio-Event de Warrd
	② high	有客观依据，确定性程度高	
情感倾向（Tendency）	① positive	对事实或观点为正向情感态度，包括支持、肯定或点明优势	News-Event
	② negative	对事实或观点为负向情感态度，包括反对、否定或指出不足	
	③ neutral	中立态度，未表明态度	
知识类型（Type）	① investigation	通过研究调查得到的知识	Bio-Event
	② observation	通过实验或现象直接观察得到的知识	
	③ explanation	通过总结、归纳和推理来进行解释和阐述	
	④ general	一般类型，通常指科学事实	
来源（Source）	① own	来源于作者本人的研究	Bio-Event News-Event deWarrd
	② others	来源于他人研究	
时态（Tense）	① past	描述状态为过去	News-Event
	② present	描述状态为一般/现在	
	③ future	描述状态为将来	

在描述科学论文内容组件时，确定性程度用以表示内容的语义确信度，根据是否有客观事实为依据，可以将属性值定为高和低。如：基于客观数据得到的结论，其确定性程度为“高”，而由推理或猜测得到的假设，确定性程度为“低”。情感倾向揭示的是内容所传达的正负向情绪，包括评价、判断、推测等方面。如支持某一观点时，情感倾向为正向；表达某一结果的不足时，态度是负向的；客观陈述事实则为中立态度。知识类型包括调查、观察、解释及一般类型，与研究方法密切相关，如通过实验法得到的内容通常是由观察得到的，而通过问卷法得到的结论通常归为由调查得到的。来源包括作者本人的研究、引用他人的研究。时态包括过去、一般/现在、将来。如在陈述已有研究或实验过程时，通常使用过去时态；在展望未来研究时，使用将来时态。以上属性信息均可通过科学论文组件中特定的线索词反映出来，这同时也为机器自动标注、知识挖掘与知识发现提供了文本语义层面的基础。

特定的内容组件会呈现出相应的属性特征,具体如表4所示。先从属性角度来看,一般涉及科学事实、数据和引用的,都带有确定性程度和来源属性。情感倾向属性一般不存在于“主题”“背景”“实验”“未来工作”组件中,因为这些组件很少涉及评价、判断等观点性内容。知识类型属性不用于“缘起”“假设”“未来工作”组件。时态属性是普适的,在英文论文中所有内容组件都会有对应的时态特征。再从内容组件角度来看,“已有研究”“数据”“结果”“结论”和“讨论”组件的知识形态都能涉及表3定义的5种属性类型,因为它们都以科学事实和数据为基础,需要注明引用来源,同时也会有判断、评价性内容产生,故带有情感倾向。“假设”组件的属性特征比较清晰,通常确定性程度值为“低”,情感倾向为正向或负向。“缘起”和“方法”组件会带有情感倾向属性,因为大部分研究在研究动机和意义上都会表现出正向的态度,同时方法部分会涉及有效性评估。“实验”组件一般是客观陈述实验的素材、步骤、任务、变量,不存在事实、数据和结论、评判,所以不带有确定性程度和情感倾向属性;但是在选择实验变量时,通常需要借鉴和引用相关研究已经论证过的变量,因此带有来源属性。而“研究缘起”“假设”“未来工作”都不涉及引用,故没有“来源”属性,其中“未来工作”组件只是一般性陈述未来的研究方向,所以也不会涉及确定性程度、情感倾向和知识类型属性。

表4　内容组件与属性信息的匹配

内容组件	属性				
	确定性程度	情感倾向	知识类型	来源	时态
背景	√		√	√	√
主题	√		√	√	√
缘起		√			√
已有研究	√	√	√	√	√
假设	√	√			√
方法		√	√	√	√
实验			√	√	√
数据	√	√	√	√	√
结果	√	√	√	√	√
结论	√	√	√	√	√
讨论	√	√	√	√	√
未来工作					√

3.5　基于 Protégé 的科学论文功能单元本体表示

在开发了本体模型后,我们使用 Protégé5.1 对本体进行了表示。本体中的类如图1所示,属性如图2所示。

4　基于 FUO 的科学论文深度语义标引实验

4.1　深度语义标引方法

深度标引(Deep indexing)是近年来被广泛

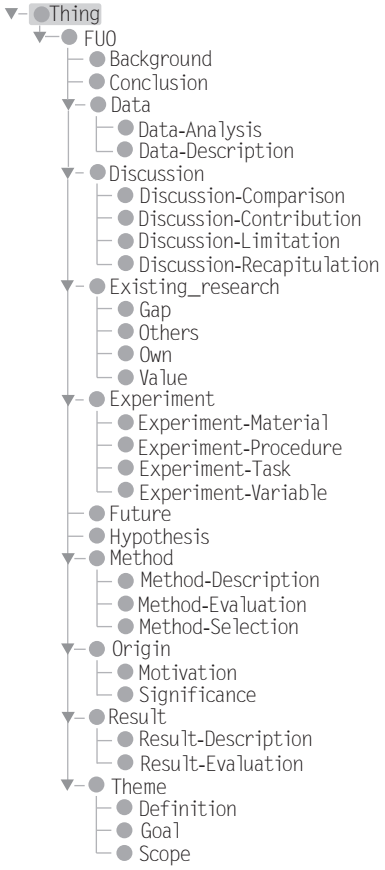

图 1 本体大类及其子类树形图

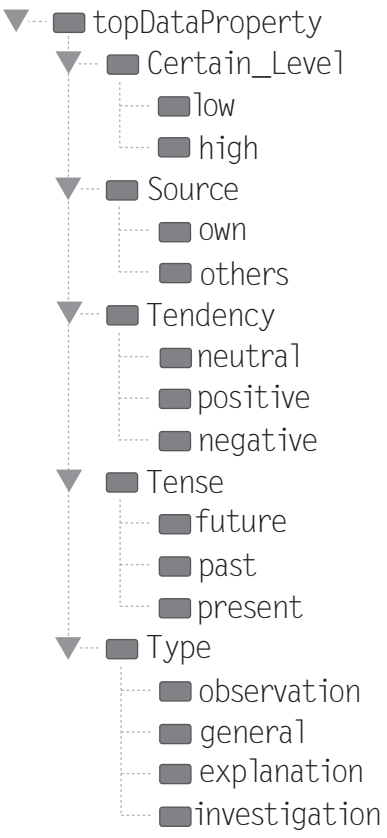

图 2 本体属性树形图

接受的新型文献组织方式,意在通过对文献内部不同粒度数字资源进行标引与组织,表征并揭示图表、数据及相关内容中蕴含的潜在知识,以满足用户高精度、细粒度的检索需求,实现知识挖掘、共享与重组目标[43],已在信息检索领域达成了广泛共识[44]。科学论文内容的深度标引需要准确定义科学论文内容组件及其属性信息[45]。标引流程如图 3 所示。

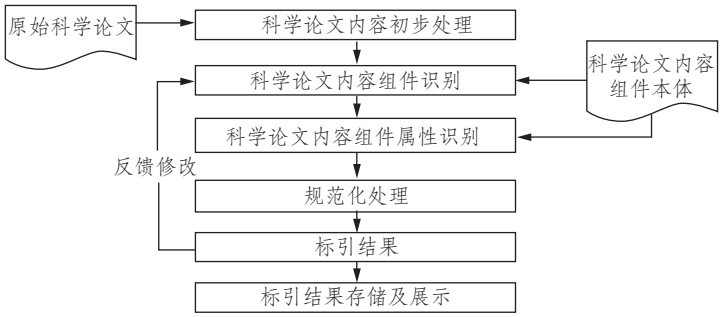

图 3 科学论文内容深度语义标引流程

在比较分析了标注工具 GATE[46] 和 Annotea[47] 后,我们选择了 GATE 作为标引工具。GATE 由谢菲尔德大学开发,是一款集本体标注、自然语言处理等功能模块为一体的文本分析工具,最新版本为 8.4.1。借助 GATE 平台可以使用本体直接对科学论文内容部分进行标引,并将标引结果以 XML 的形式进行存储。GATE 可以对不同语义单元进行可视化呈现,其具体操作流程如图 4 所示。

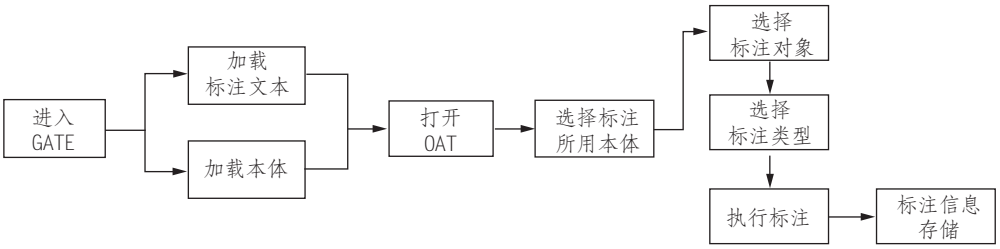

图 4　GATE 平台的标引流程

4.2　标引结果存储示例

标注实验结果可以被保存为 XML 文档,示例如表 5 和表 6 所示。表 5 展示了学习方法部分的标注结果代码。这一部分主要包括方法和实验两大类及其子类,主要属性包括来源、知识类型、时态等。

表 6 所示为研究背景部分的标引结果,这部分侧重于对已有研究的梳理及对比,属性主要包括来源、知识类型、情感倾向、时态等。

4.3　深度标引实验结果

为了验证功能单元本体 FUO 对科学论文语义结构和功能的表示能力,本研究从学术期刊 *Journal of the Association for Information Science and Technology* 中随机抽取了 2015—2018 年间发表的 10 篇研究型论文作为实验对象,借助 FUO 进行深度语义标引。该期刊是美国科学技术情报学会的会刊,主要发表情报科学与技术方面的论文,论文格式十分规范。随机抽取的 10 篇论文如表 7 所示。

表 5　学习方法标引结果代码示例

信息使用任务及相关内容组件	XML 代码
学习方法(Learn how to do)	<method-descriptionsource = “own” tense = “present”> Our method is based on the following steps…</method-description> <experiment-materialsourse = “others” tense = “past” type = “general” > We have used a corpus of five scientific journals: PLoS Biology, PLoS Computational Biology,…. in the XML format. </experiment-material> <experiment-procedure source = “own”>section categorization and part-of-speech-tagging… </experiment-procedure> <experiment-procedure source = “own” tense = “past”> our first task was to categorize the sections according to the…To do this, we analyzed the titles… <experiment-procedure source = “own” tense = “past”> the extracted citation contexts were processed using TreeTagger…</experiment-procedure> ……
• 方法(Method) 方法选择(Method-Selection) 方法描述(Method-Description) 方法评估(Method-Evaluation) • 实验(Experiment) 实验素材(Experiment-Material) 实验任务(Experiment-Task) 实验步骤(Experiment-Procedure) 实验变量(Experiment-Variable)	

表 6　研究背景标引结果代码示例

信息使用任务及相关内容组件	XML 代码
研究背景(Learn about background)	<backgroundtype="general" tendency="neutral" tense="past"> Citation analysis has been the subject of numerous studies during the last decades and there has been a constant interest in producing a theory of citations…</background> <value tendency="positive" source="others">The works of Cronin [5-7], Small [21] and Leydesdor[11] are among the most important in this domain and showed the importance of this research….</value> <gap tendency="negative" source="others" tense="present">Their large-scale study examines only the number of citations according to the positions in the text and in the sections but does not rely on further linguistic analyses of the citation contexts…</gap> <discussion-comparison type="explanation" tendency="positive" source="own" type="present">These corpora can serve as a reference data for other works…our study tries to extend this type of approach…</discussion-comparison> ……
• 已有研究(Existing Research) 他人研究(Others) 本人研究(Own) 已有研究价值(Value) 现有研究空白(Gap) • 背景(Background) • 主题(Theme) 研究范围(Scope) 主题定义(Definition) • 讨论(Discussion) 对比(Discussion-Comparison)	

表 7　十篇样本论文列表

论文编号	论文标题
1	Reader Characteristics, Behavior, and Success in Fiction Book Search
2	Cognitive Modeling of Age-Related Differences in Information Search Behavior
3	Matched Control Groups for Modeling Events in Citation Data: An illustration of Nobel Prize Effects in Citation Networks
4	Citations, Mandates, and Money: Author Motivations to Publish in Chemistry Hybrid Open Access Journals
5	Assessing Perceived Organizational Leadership Styles Through Twitter Text Mining
6	Cultural Diversity of Quality of Information on Wikipedias
7	Scientists' Data Reuse Behaviors: A Multilevel Analysis
8	How Online Social Interactions Influence Customer Information Contribution Behavior in Online Social Shopping Communities: A Social Learning Theory Perspective
9	Imperfect Referees: Reducing the Impact of Multiple Biases in Peer Review
10	Online Consumer Reviews and Sales: Examining the Chicken-Egg Relationships

我们遵照前文所述流程和方法对这 10 篇论文进行了标引,并对标引结果进行了统计分析,以观察不同类型的功能单元在科学论文中的分布情况,结果如表 8 和表 9 所示。

表 8　各样本论文中含有的功能单元种类数量

论文编号	论文 1	论文 2	论文 3	论文 4	论文 5	论文 6	论文 7	论文 8	论文 9	论文 10
类目数量	30	31	28	25	28	26	31	32	28	30
类目占比(%)	90.9	93.9	84.8	75.8	84.8	78.8	93.9	97.0	84.8	90.9

表 9　不同功能单元出现的频率

论文数量/篇	10	9	8	7	6	5	4
对应的功能单元种类	背景、研究范围、研究目标、动机、他人研究、方法描述、实验素材、实验任务、实验步骤、实验变量、数据描述、数据分析、结果描述、结论、贡献、局限性、确定性程度、情感倾向、来源、时态、知识类型	方法选择	意义、现有研究价值、研究空白、结果评估	要点概括与重述	假设、对比	定义、方法评估、未来工作	本人研究

表 8 揭示了每篇论文含有的功能单元种类数量。从表 8 可以看出,不同的论文含有的功能单元种类数量不同,有 5 篇论文含有 30 种以上功能单元,2 篇论文分别含有 25 和 26 种单元类型。再结合表 9 来看,不同类型的功能单元在文章中出现的频率也不同,有些功能单元如“背景”“研究目标”十分常见,有些功能单元,如“本人研究”和“方法评估”比较少见。综合这些统计数据可知,使用功能单元本体 FUO 可以对科学论文内容进行深度语义标引,具有较好的可用性。

5　讨论

5.1　多本体协同下的论文深度语义标引

科学论文是知识的容器,蕴含着不同类型的情报与知识。科学论文功能单元本体不受领域限制,但仅仅是一种视角。研究者还可以从不同的理论和观察视角提出更多的论文内容本体,揭示论文的情报结构和功能特征。事实上,为了全面揭示一篇科学论文内涵的科学知识和情报,还必须协同使用不同本体对科学论文内容进行多角度的深度语义标引。需要特别指出的是,科学论文内容本体与领域本体(医学、生物学等)不同,两者是不同类型的本体。在论文深度语义标引中需要使用内容本体进行语义功能标引,使用领域本体进行主题标引,两者缺一不可。

5.2　科学论文功能单元本体中的关系定义

科学论文内容组件间的关系在科学论文语篇有机构成中扮演着重要角色,但关系的定义和识别较为复杂。虽然修辞结构理论、论证理论都对内容组件间的关系进行了定义,但不同理论的语义关系定义存在冲突。功能单元理论对此考虑也不够细致,所以在我们的本体模型中暂时没有考虑这种关系定义。未来,可以借鉴体裁分析、语篇分析、情报分析等理论,对内容组件间的语义关系进行尝试性定义。

5.3　面向科学情报分析的内容组件重组

科学论文的不同部分隐藏着不同功能和价值的科学情报,揭示这些细粒度的潜在的情报具有重要意义。本研究利用功能单元本体 FUO 对论文进行初步的深度语义标引实验,统计分析了不同功能单元的分布情况。事实上,如果

对所有论文进行深度语义标引,就可以对细粒度的内容片段进行重新组织,从而辅助情报分析和科研工作者的战略阅读,大大提高情报工作的效率和科研工作者理解特定科研领域宏观知识的效率。

6 总结

本文以功能单元理论为基础,面向科学论文情报表示和揭示需求,提出了一种新的科学论文功能单元本体 FUO。借助该本体和语义标注工具 GATE,对 10 篇样本论文进行深度语义标引实验,结果表明该本体适用于科学论文的深度语义标引,能够在一定程度上表示和揭示科学论文中各部件的语义功能及其属性。

该研究借助本体技术,首次清晰地定义了科学论文的功能单元及其属性,提出的功能单元本体 FUO 在信息组织、语义检索、知识发现、情报分析等领域拥有广阔的应用空间。但是,随着科学论文的数字化转型,越来越多的内容组件形式出现,如互动表格、数据集、音视频等,如何对这些内容组件进行建模表示,需要进一步考虑。不管如何,本文的研究深化了我们对论文这种知识容器的理解,完善了情报学视角下的科学论文语义结构与功能理论。未来,我们将继续细化功能单元的分类,并尝试从其他理论视角提出更丰富的科学论文内容本体。

参考文献

[1] Renear A H, Palmer C L. Strategic reading, ontologies, and the future of scientific publishing[J]. Science, 2009, 325(5942):828.

[2] 黄永, 陆伟, 程齐凯,等. 学术文本的结构功能识别——在学术搜索中的应用[J]. 情报学报, 2016, 35(4):425-431.(Huang Yong,Lu Wei,Cheng Qikai,et al. The structure function recognition of academic text: application in academic search[J]. Journal of the China Society for Scientific and Technical Information,2016, 35(4):425-431.)

[3] Bishop A P. Document structure and digital libraries: how researchers mobilize information in journal articles[J]. Information Processing & Management, 1999, 35(3):255-279.

[4] Sandusky R J, Tenopir C. Finding and using journal-article components: impacts of disaggregation on teaching and research practice[J]. Journal of the American Society for Information Science & Technology, 2008, 59(6): 970-982.

[5] 温有奎, 徐国华. 知识元链接理论[J]. 情报学报, 2003, 22(6):665-670.(Wen Youkui, Xu Guohua. Knowledge element linking theory[J]. Journal of the China Society for Scientific and Technical Information,2003, 22(6):665-670.)

[6] 赵火军, 温有奎. 基于引文链的知识元挖掘研究[J]. 情报杂志, 2009, 28(3):148-150.(Zhao Huojun, Wen Youkui. Research on mining knowledge elements based on citation chain[J]. Journal of Information, 2009, 28(3):148-150.)

[7] Sateli B, Witte R. Semantic representation of scientific literature: bringing claims, contributions and named entities onto the Linked Open Data cloud[J]. Peer J Computer Science, 2015, 1: e37.

[8] Elhadad M K, Badran K, Salama G I. A novel approach for ontology-based dimensionality reduction for Web text document classification[C]// Ieee/acis, International Conference on Computer and Information Science. IEEE, 2017:373-378.

[9] Fang W, Guo Y, Liao W. Ontology-based indexing method for engineering documents retrieval[C]// IEEE International Conference on Knowledge Engineering and Applications. IEEE, 2017:172-176.

[10] 曹树金,李洁娜,王志红. 面向网络信息资源聚合搜索的细粒度聚合单元元数据研究[J]. 中国图书馆学报, 2017, 43(4):74-92.(Cao Shujin, Li Jiena, Wang Zhihong. Research on the meta-data schema for fine-grained aggregation units of internet resources[J].Journal of Library Science in China,2017, 43(4): 74-92.)

[11] Burroughboenisch J. International reading strategies for IMRD articles[J]. Written Communication, 1999, 16(3):296-316.

[12] Teufel S, Carletta J, Moens M. Anannotation scheme for discourse-level argumentation in research articles[C]// Meeting of the European Chapter of the Association for Computational Linguistics, 2001:110-117.

[13] Teufel S. The structure of scientific articles: applications to citation indexing and summarization[J]. Studies in Computational Linguistics, 2010, 38(2):443-445.

[14] King R D, Liakata M, Lu C, et al. On the formalization and reuse of scientific research[J]. Journal of the Royal Society Interface, 2011, 8(63):1440-1448.

[15] Liakata M, Thompson P, Waard A D, et al. A three-way perspective on scientific discourse annotation for knowledge extraction[C]//Proceedings of the Acl Workshop on Detecting Structure in Scholarly Discourse, 2012: 37-46.

[16] de Waard A, Buitelaar P, Eigner T. Identifying the epistemic value of discourse segments in biology texts[C]// Proceedings of the Eighth International Conference on Computational Semantics. Association for Computational Linguistics, 2009: 351-354.

[17] Liakata M, Saha S, Dobnik S, et al. Automatic recognition of conceptualization zones in scientific articles and two life science applications[J]. Bioinformatics, 2012, 28(7):991-1000.

[18] de Waard A. A pragmatic structure for research articles[C]// International Conference on Pragmatic Web, Icpw 2007, Tilburg, the Netherlands, October. DBLP, 2007:83-89.

[19] Mihăilă C, Kontonatsios G, Batista-Navarro R T B, et al. Towards a better understanding of discourse: integrating multiple discourse annotation perspectives using UIMA[C]// Linguistic Annotation Workshop and Interoperability with Discourse, 2013:79-88.

[20] Guo Y F, Korhonen A, Liakata M, et al. Identifying the information structure of scientific abstracts: an investigation of three different schemes[C]//Proceedings of the Workshop on Biomedical Natural Language Processing, 2010:99-107.

[21] Evans J A, Foster J G. Metaknowledge[J]. Science, 2011, 331(6018):721-725.

[22] Thompson P. Enriching a biomedical event corpus with meta-knowledge annotation[J]. Bioinformatics, 2011, 12(1):393.

[23] Thompson P, Nawaz R, Mcnaught J, et al. Enriching news events with meta-knowledge information[J]. Language Resources & Evaluation, 2016:1-30.

[24] de Waard A, Maat H P. Epistemic modality and knowledge attribution in scientific discourse: a taxonomy of types and overview of features[C]// The Workshop on Detecting Structure in Scholarly Discourse. 2012:47-55.

[25] 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.(Lu Wei, Meng Rui, Liu Xingbang. A deep scientific literature mining-oriented framework for citation content annotation[J].Journal of Library Science in China,2014, 40(6):93-104.)

[26] 王晓光,宋宁远. 科学论文内容本体比较研究[J]. 数字图书馆论坛, 2017(8):2-7.(Wang Xiaoguang, Song Ningyuan. Review on the scientific paper component ontologies[J].Digital Library Forum,2017(8):2-7.)

[27] Groza T, Moller K, Handschu S, et al. SALT: weaving the claim Web[M]. Springer Berlin Heidelberg, 2007.

[28] 马雨萌,祝忠明. 科学篇章修辞块本体标准及其应用分析[J]. 情报杂志, 2012, 31(10): 112-116.(Ma Yumeng, Zhu Zhongming. The ontology of rhetorical blocks standard and its application analysis[J]. Journal of

Information,2012, 31(10): 112-116.)

[29] The discourse element ontology[EB/OL]. [2017-09-15]. http://www.sparontologies.net/ontologies/deo/source.html

[30] Constantin A, Peroni S, Pettifer S, et al. The document components ontology (DoCO)[J]. Semantic Web, 2016, 7(2): 167-181.

[31] Shotton D. CiTO, the citation typing ontology[J]. Journal of Biomedical Semantics, 2010, 1(S1): S6.

[32] Peroni S. Semantic Publishing: issues, solutions and new trends in scholarly publishing within the Semantic Web era[D]. Universita' di Bologna, 2012.

[33] The argument model ontology[EB/OL]. [2017-09-15].http://www.essepuntato.it/2011/02/argumentmodel.

[34] Attwood T K, Kell D B, Mcdermott P, et al. Utopia documents: linking scholarly literature with research data[J]. Bioinformatics, 2010, 26(18): i568-i574.

[35] Castro L J G, Mclaughlin C, Garcia A. Biotea: RDFizing PubMed central in support for the paper as an interface to the Web of data[J]. Journal of Biomedical Semantics, 2013, 4(1): 1-22.

[36] Parinov S, Kogalovsky M. Semantic linkages in research information systems as a new data source for scientometric studies[J]. Scientometrics, 2014, 98(2): 927-943.

[37] Zhang L, Kopak R, Freund L, et al. A taxonomy of functional units for information use of scholarly journal articles [J]. Proceedings of the American Society for Information Science and Technology, 2010, 47(1): 1-10.

[38] Swales J. Genre analysis: English in academic and research settings[M]. Cambridge: Cambridge University Press, 1990.

[39] Zhang L, Kopak R, Freund L, et al. Making functional units functional: the role of rhetorical structure in use of scholarly journal articles[J]. International Journal of Information Management, 2011, 31(1): 21-29.

[40] Zhang L. Grasping the structure of journal articles: utilizing the functions of information units[J]. Journal of the American Society for Information Science and Technology, 2012, 63(3): 469-480.

[41] Zhang L. Linking information through function[J]. Journal of the Association for Information Science and Technology, 2014, 65(11): 2293-2305.

[42] Nawaz R, Thompson P, Mcnaught J, et al. Meta-knowledge annotation of bio-events[J]. Lrec, 2010.

[43] Hyer M. Deep indexing: harnessing the power of data discovery [EB/OL]. [2017-09-15]. https://www.osti.gov/home/system/files/07-08-08_CENDI_ProQuest_Hyer.pdf.

[44] Liu Y, Guo Y, Wu S, et al. Deep index for accurate and efficient image retrieval[C]// ACM on International Conference on Multimedia Retrieval. ACM, 2015: 43-50.

[45] Liakata M, Teufel S, Siddharthan A, et al. Corpora for the conceptualisation and zoning of scientific papers[C]// International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta. DBLP, 2012: 105-108.

[46] Cunningham H, Wilks Y, Gaizauskas R J. GATE: ageneral architecture for text engineering[C]// Conference on Computational Linguistics. Association for Computational Linguistics, 1996: 1057-1060.

[47] Kahan J, Koivunen M R, Prud' Hommeaux E, et al. Annotea: an open RDF infrastructure for shared Web annotations[J]. Computer Networks, 2002, 39(5): 589-608.

王晓光 武汉大学信息管理学院教授。湖北 武汉 430072。

李梦琳 武汉大学信息管理学院硕士研究生。湖北 武汉 430072。

宋宁远 武汉大学信息管理学院博士研究生。湖北 武汉 430072。

(收稿日期:2018-01-19;修回日期:2018-03-12)