

# 基于纳米出版物的中文学位论文语义组织研究\*

李春秋 徐曾旭林 宋宁远 王晓光

**摘 要** 纳米出版物在细粒度揭示科学论文内容、规范描述科学概念等方面具有一定优势,已被运用于语义出版与语义组织等领域。但囿于纳米出版物在表征论断语义、实现论断间语义关联等方面的不足,现有纳米出版物未能多维度、多粒度揭示科学论文的语义特征与结构特点,由此限制了纳米出版物的应用与服务。本研究复用领域本体,通过改进纳米出版物通用模型,提出了适用于特定领域、特定体裁科学论文论断表征的方法,开展应用实践探讨。针对信息检索领域中文学位论文的语义特征与语义关联,本文扩展了纳米出版物通用模型,细化了学位论文纳米出版物的论断类别,构建了中文学位论文纳米出版物模型;并选取信息检索领域的若干中文学位论文为实验对象,生成纳米出版物的 RDF 命名图及 Turtle 数据,在此基础上分别基于案例分析和数据集应用开展实证研究,以验证本研究所构建模型的适用性。本研究提出的纳米出版物改进方法与扩展模型,可为纳米出版物在具体领域的应用研究和中文学位论文的语义组织提供借鉴。图 12。表 17。参考文献 17。

**关键词** 纳米出版物 语义组织 中文学位论文 语义出版 信息检索

**分类号** G254

## Nanopublication-based Semantic Organization of Chinese Dissertation

LI Chunqiu, XU Zengxulin, SONG Ningyuan & WANG Xiaoguang

### ABSTRACT

With advantages of revealing contents of scientific paper and formally describing scientific concepts at fine granularity, nanopublication has been widely used in the fields of semantic publishing and semantic organization. However, due to nanopublication's weakness in specific fields including the representation of assertion semantics and the realization of semantic linkage between assertions, the existing nanopublication failed to reveal semantic features and structure characteristics of scientific papers from multi-dimension and multi-granularity, thus limiting its application and service. In view of this, the research reuses domain ontology, improves nanopublication common model, proposes representation approach to specific domain and type of scientific paper's assertions, and conducts application practices. With a focus on the semantic features and linkage of Chinese dissertations in information retrieval domain, the research expands the common structure of nanopublication model, classifies the specific assertion types, and designs description models of nanopublication for Chinese dissertations on

\* 本文系中央高校基本科研业务费专项资金资助课题项目“基于纳米出版物模式的中文学术论文的语义组织研究”(编号:310422112)的研究成果之一。(This article is an outcome of the project “Nanopublication Based Semantic Organization of Chinese Academic Paper”(No. 310422112) supported by the Fundamental Research Funds for the Central Universities.)

通信作者:宋宁远,Email:songny\_w hu@126.com,ORCID:0000-0001-5601-1487( Correspondence should be addressed to SONG Ningyuan, Email:songny\_w hu@126.com,ORCID:0000-0001-5601-1487)

information retrieval. The research selects certain numbers of Chinese dissertations on information retrieval as experiment samples, and creates RDF named graphs and Turtle data for nanopublication. On this basis, empirical research is carried out through case analysis and data set application in order to further verify the usability of the proposed models.

The proposed approach to improve nanopublication and extend description models in the research could provide reference to nanopublication's application in specific domain and semantic organization of Chinese dissertation. The proposed model excels in information retrieval by revealing semantic characteristics of specific statements about experiment data such as experiment parameter, experiment model and test collection. The model covers the core classes of information retrieval and formalizes their relationships, which provide description model for semantic data to automatically extract assertions and semantic relationships. By using term recognition, entity extraction, machine learning and data cleaning, the model proposed in this study helps the assertion extraction and automatic annotation of the Chinese dissertation, and also provides models and methods for automatic construction of nanopublications. There are limitations for describing specific semantics in other specific domains when applying the model to creating nanopublications of Chinese dissertations with various structural and semantic features.

Scientific papers in nature language are complex on content semantics. It is difficult to identify experiment tasks and procedures, and necessary experiment assessments are also required. Therefore, in future, it is necessary to further establish a large-scale, high-quality and inter-linked scientific paper corpus based on innovated description model of scientific contents to provide a data foundation for extracting and revealing assertions in scientific papers. Scientific paper is composed of knowledge units with semantic features and logic relationships. The future application of nanopublication in scientific papers shall focus on formal description and semantic relationships at the fine granularity of knowledge units, with a purpose to construct multi-level, multi-granularity and multi-dimension content datasets of scientific papers. 12 figs. 17 tabs. 17 refs.

#### KEY WORDS

Nanopublication. Semantic organization. Chinese dissertation. Semantic publishing. Information retrieval.

## 0 引言

学位论文是重要的文献信息源之一,内容详实且具有较高的原创性、学术价值与信息价值,是阶段性研究成果的集成。但由于结构版式固定、机器可读性差、篇幅较长等原因,传统出版模式下的学位论文不能充分满足用户快速定位知识、准确检索知识和高效挖掘知识关联的需求,从而影响学位论文的有效利用及其在科学交流体系中的价值体现<sup>[1]</sup>。

随着语义技术的发展,情报学、出版学及计

算机科学等领域不断探索运用 RDF、关联数据、本体等知识组织工具,创新科学论文内容的表现形式,提升知识的机器可读性<sup>[2]</sup>。2009年,肖顿<sup>[3]</sup>提出了语义出版概念,试图将语义网和文本分析等技术与出版实践相结合,以增强出版物的语义内涵并提高知识单元的关联程度<sup>[4]</sup>。同年,概念网络联盟<sup>[5]</sup>设计了一种全新的科学知识表示与组织模型,即纳米出版物。该模型重点使用机器语言,基于陈述表达科学论断并将科学论断相互关联,实现对论文细粒度知识单元的语义建模、表征与关联。纳米出版物有助于提升知识的机器可读性,打破知识固有边

界,实现知识的可链接、可溯源、可复用以及学位论文知识价值的再利用<sup>[6]</sup>。

纳米出版物在表征知识结构、实现知识对象独立出版、生成与发布机器可读的论文内容数据集等方面具有优势,但在表征科学论断的语义特征、揭示论断间语义关系、实现论断语义关联等方面尚存不足<sup>[7]</sup>。为了更好地发挥和利用纳米出版物的优势,实现对中文学位论文的准确表征,还需要扩展纳米出版物通用模型,结合领域需求与描述目的,构建适用于中文学位论文的纳米出版物表达模型。针对这一问题,本研究选取知识更新快、结构严谨、方法明确、实验规范的信息检索领域中中文学位论文为研究对象,借助领域本体表征科学论断的语义特征与语义关系,基于纳米出版物通用模型提出面向信息检索领域中中文学位论文的纳米出版物表达模型,并构建纳米出版物网络模型,实现细粒度知识单元的封装与发布,优化中文学位论文知识组织模式,以促进中文学位论文知识资源价值最大化,辅助研究者开展溯源或推理等知识发现活动。

## 1 研究综述

### 1.1 纳米出版物结构

2009年,概念网络联盟<sup>[5]</sup>基于开放标准首次提出纳米出版物概念,将其阐释为以“论断”为单位的“具有科学意义、机器可读、最小的可出版信息单元”,其最小元素为“概念”。2010年,保罗·格鲁斯<sup>[6]</sup>等首次提出纳米出版物模型的结构及命名图形式的RDF序列化方法,并提供了模型示例。2015年,Open PHACTS进一步确立了纳米出版物的核心模型与通用结构<sup>[8]</sup>,包含三类基本信息(头部、论断和出处)和两个辅助要素(出版信息和纳米出版物ID)。其中,论断是纳米出版物的基础,表现为主语—谓语—宾语三元组形式;出处信息是论断的元数据;出版信息是纳米出版物的作者、时间、版权及版本等信息;头部界定了纳米出版物与论断、

出处、出版信息间的关系;纳米出版物ID由URI标识符组成,可视作纳米出版物唯一标识符,而完整性密钥由一系列URI标识符组成,可用于标识作者身份和出版物版本<sup>[8]</sup>。

纳米出版物的核心思想是将科学论断及其语境信息关联,形式化表示科学信息,有利于科学信息的整合、查询、推理等处理。但纳米出版物通用模型仅提供了表征独立论断的方式,缺乏对论断语义特征及论断间语义关系的定义。因此,在实际应用中需根据具体需求,复用领域本体实现模型扩展,以确保富语义论断无歧义、可被唯一识别,并表征纳米出版物创建者和创建时间等背景信息。

### 1.2 纳米出版物项目实践

纳米出版物改进了科学论文和数据集的出版方式,目前被运用于生命科学、计算机科学、人文与社会科学等领域,一定程度上解决了文献数量多、语义关联不足、知识利用效率低等问题。表1列举了代表性纳米出版物项目。

科学论文的纳米出版物构建多围绕篇幅较短的英文期刊论文展开。如Lipani<sup>[9]</sup>和廖建军<sup>[10]</sup>等采用结构化语义标引技术,识别并抽取分散于文献各处的知识内容,结合领域特征将其发布为符合纳米出版物模型的知识元,以实现知识结构的语义化表达。此外,关于纳米出版物的适用性也被广泛探讨。如Golden<sup>[11]</sup>等提出纳米出版物的论断应根据领域特性和实际需求构建,在形式化表达不同学科领域学术论文科学论断时,应考虑学科领域特点及应用目的;Kuhn<sup>[12]</sup>等认为纳米出版物仅适用于部分领域的科学论断。相较于以理论探索为主的人文社科领域学术论文,自然科学领域实证性研究论文的内容主要有方法、实验及数据等,论证结构较固定且实验结论较客观,科研成果可重复进行实验验证,更适用于纳米出版物表征。因此,为了提升纳米出版物的适用性,有必要结合特定领域、特定体裁的科学论文语义特征扩展纳米出版物通用模型。

表 1 纳米出版物主要项目

领域	项目名称	年份	简介
生命科学	Open PHACTS <sup>①</sup>	2011	建立开放药理学空间,整合药理数据至可交互的基础构架,发布纳米出版物指南
	BEL2nanopub <sup>②</sup>	2017	将 BEL 文档转化为医学数据的纳米出版物,生成论断、出处和出版信息
计算机科学	Feedback Looping NPs <sup>[13]</sup>	2011	使用博客与 MediaWiki 等关联数据,集成外部资源提取科学论断,构建纳米出版物
	ADmIRE <sup>[9]</sup>	2014	基于信息检索领域本体与科学文献,生成信息检索实验的纳米出版物,设计自动化提取流程
	Nanopublication 知识服务模型 <sup>[10]</sup>	2017	设计分布式 Nanopublication 知识服务模型及医学文献知识检索实验系统
艺术	EMA <sup>③</sup>	2014	构建适用于音乐领域语义检索与知识发现的纳米出版物知识库,提供音乐寻址能力 API
历史	Early Modern European Peace Treaties Online <sup>④</sup>	2015	以 Turtle 形式将条约元数据发布为关联开放数据
哲学	EMTO Nanopub <sup>⑤</sup>	2017	关注哲学事实,开展事实表达,提供早期现代哲学史的数字资源,以网页形式可视化呈现

综上,纳米出版物通用模型在一定程度上建构了知识实体间的关联,但在表征论断的类型、语义特征及论断间语义关系等方面尚有欠缺,需结合领域特征优化与创新纳米出版物模型。为准确表征信息检索领域中学位论文中的论断,建构以纳米出版物模型为基础的论断知识网络,需结合领域特征准确确定论断间语义关系、论断的语义特征,并开展语义建模。

## 2 中文学位论文的纳米出版物模型

### 2.1 结构分析

学位论文具有一定的结构特征,包括题名、署名、导师、摘要、关键词等题录信息以及引言、背景、意义、文献综述、研究方法、实验过程、数

据分析和结论等内容信息。题录信息表现为结构化数据,提供学位论文出处及出版信息。内容信息以论断为基本结构,且由特定的逻辑关系关联。针对学位论文的形式与内容特点,本研究设计了适用于中文学位论文的纳米出版物模型,如图 1 所示。

论断是纳米出版物模型的核心,源自非结构化数据集或科学论文主要内容中的论断或实验数据,表述实验研究过程或科学观点和结论,在文献中通常表现为具有特定语义特征(背景、实验方法和数据集描述等)的论断。通过表征与关联论断间语义关系,各独立的纳米出版物可相互充分关联,有助于构建高度关联的纳米出版物网络。纳米出版物的出处信息揭示了论断来源,如论断作者、论断发布时间和 URL

① <http://www.nanopub.org/guidelines/1.8>

② <https://github.com/tkuhn/bel2nanopub>

③ <https://mith.umd.edu/research/enhancing-music-notation-addressability>

④ <http://dm2e.eu/open-humanities-awards-early-modern-european-peace-treaties-online-final-update/index.html>

⑤ [http://emto-nanopub.referata.com/wiki/EMTO\\_Nanopub](http://emto-nanopub.referata.com/wiki/EMTO_Nanopub)

等<sup>[10]</sup>。针对学位论文题录信息特点,本研究补充了 ORCID、学位、导师、研究方向、所属学科、分类号、语种和作者单位等信息。在中文学位

论文纳米出版物模型的基础上,本研究复用了 PAV、SIO、BIBO、IR 等本体与 FOAF、DC 等词表来表征语义关系,见表 2。

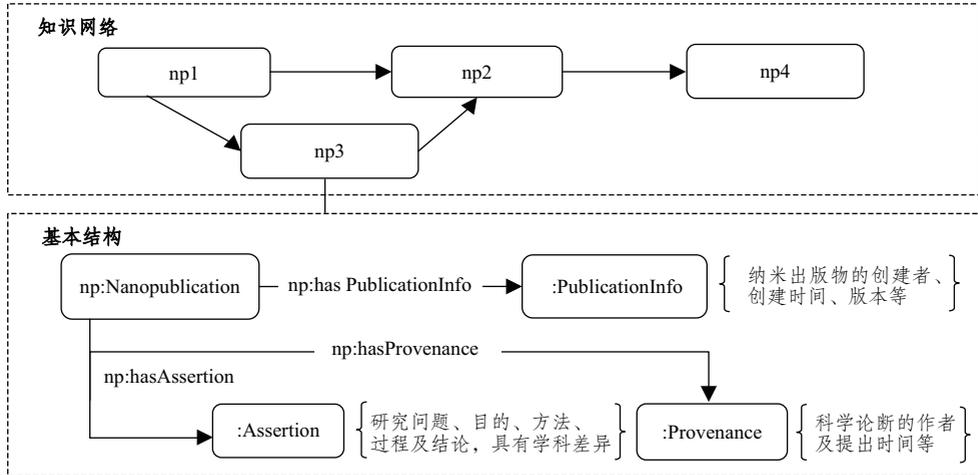


图 1 中文学位论文纳米出版物模型结构

表 2 中文学位论文纳米出版物模型复用的本体与词表及其命名空间

本体/词表	前缀	命名空间
Nanopublication Ontology (纳米出版物本体)	np	http://www.nanopub.org/nschema#
Provenance, Authoring and Versioning Ontology (出处、创作及版本 PAV 本体)	pav	http://purl.org/pav/
Friend of A Friend (FOAF 词表)	foaf	http://xmlns.com/foaf/0.1/
RDA Element Sets (RDA 元素集)	rdau	http://rdaregistry.info/Elements/u/
Nature Publishing Group Core Ontology (Nature 核心本体)	npg	http://ns.nature.com/terms/
DCMI Metadata Terms (都柏林核心元数据术语集)	dcterms	http://purl.org/dc/terms/
Semanticscience Integrated Ontology (语义科学集成本体)	sio	http://semanticscience.org/ontology/sio.owl
Bibliographic Ontology (书目本体)	bibo	http://purl.org/ontology/bibo/
Information Retrieval Ontology (信息检索本体)	ir	http://ifs.tuwien.ac.at/~admire/ir_ontology/ir#

## 2.2 内容分析

本研究复用 11 个核心类与 21 个关键属性,以有效描述学位论文所含的实验模型、科学结论、实验结果及论断来源等论断信息,并根据论

断间的逻辑关系将纳米出版物关联形成纳米出版物网络。

(1) 论断分析:论断实例采用 np: Assertion 类表征。为准确表征信息检索领域中学位论文

文的论断特点,本研究在 Lipani<sup>[14]</sup>等研究的基础上,根据信息检索本体与该类学位论文涉及的主要任务、实验等要素,共复用 11 个核心类及 7 个对应属性以揭示论断类型、语义特征及语义

关系,见表 3 和表 4。其中,M 表示必选,R 表示推荐,O 表示可选。

(2) 出处信息分析:本研究复用 PAV 本体以具体描述论断作者、论断提出时间等,采用

表 3 中文学位论文纳米出版物模型中 `np:Assertion` 的类别列表

类	注释	取值类型	约束
<code>ir:EvaluationThing</code>	一系列评估的事件合集,子类包括信息检索任务、实验和评价措施	URI	M
<code>ir:Task</code>	信息检索任务	URI	M
<code>ir:Experiment</code>	信息检索实验,子类包括信息检索系统、测试集合、运行程序和评价方法	URI	M
<code>ir:EvaluationMeasure</code>	评价措施,子类包括评价参数、正排序结果和反排序结果	URI	R
<code>ir:Run</code>	运行程序,在测试集合上运行特定信息检索系统的执行结果	URI	M
<code>ir:TestCollection</code>	测试集合,子类包括集合和主题	URI	M
<code>ir:Collection</code>	实验运行中应用的集合	URI	M
<code>ir:Topics</code>	实验运行中对应的主题	URI	O
<code>ir:IRSystem</code>	信息检索系统,子类包括事件组合、系统核心组件和模型等	URI	M
<code>ir:IRSystemCore</code>	信息检索系统核心组件	URI	R
<code>ir:IRModel</code>	信息检索模型,子类包括语言模型、概率模型和潜在语义索引等	URI	O

表 4 中文学位论文纳米出版物模型中 `np:Assertion` 类的属性

属性名称	定义域	值域	取值类型	约束	注释
<code>ir:innerEvaluation</code>	<code>ir:EvaluationThing</code>	<code>ir:Task</code> , <code>ir:Experiment</code> , <code>ir:EvaluationMeasure</code>	URI	M	评估的事件合集
<code>ir:hasExperimentComponent</code>	<code>ir:Experiment</code>	<code>ir:Run</code> , <code>ir:TestCollection</code>	URI	M	实验的主要组成部分
<code>ir:hasCollection</code>	<code>ir:TestCollection</code>	<code>ir:Collection</code>	URI	R	实验运行中的测试集合
<code>ir:hasTopics</code>	<code>ir:TestCollection</code>	<code>ir:Topics</code>	URI	R	实验测试集合的主题信息
<code>ir:belongsToIRSystem</code>	<code>ir:Run</code>	<code>ir:IRSystem</code>	URI	M	运行实验的信息检索系统
<code>ir:hasIRSystemCore</code>	<code>ir:IRSystem</code>	<code>ir:IRSystemCore</code>	URI	R	信息检索系统核心组件
<code>ir:fromIRSystemToIRModel</code>	<code>ir:IRSystem</code>	<code>ir:IRModel</code>	URI	O	信息检索系统生成模型

RDA 元素集的属性描述论断作者的相关信息。核心本体和 DC 元数据术语集所定义的属性,见论断来源论文相关信息的描述主要复用 Nature 表 5。

表 5 中文学位论文纳米出版物模型中 np:Provenance 类的属性

属性名称	定义域	值域	取值类型	约束	注释
pav: authoredBy	: Assertion	foaf: Person	URI	M	论断作者
pav: createdOn	: Assertion	/	literal	M	论断提出时间
pav: derivedFrom	: Assertion	npg: Publication	URI	R	论断来源的学位论文
foaf: name	foaf: Person	/	literal	M	论断作者的姓名
rdau: P60649	foaf: Person	/	literal	R	论断作者的导师信息
npg: hasSubject	foaf: Person	npg: Subject	URI	R	论断作者的学科专业
rdau: P60514	foaf: Person	/	literal	O	论断作者学位授予年度
rdau: P60001	foaf: Person	/	literal	R	论断作者学位授予单位
dcterms: title	npg: Publication	/	literal	O	论断来源学位论文名称
bibo: url	npg: Publication	/	literal	R	论断来源学位论文 URL
dcterms: language	npg: Publication	/	literal	O	论断来源学位论文语种

(3) 出版信息分析:表 6 通过复用 PAV 本体术语 dcterms: created 等来描述纳米出版物的出版属性 pav: createdBy、pav: version 和 DC 元数据出版信息。

表 6 中文学位论文纳米出版物模型中 np:PublicationInfo 类的属性

属性	定义域	值域	取值类型	约束	注释
pav: createdBy	np: Nanopublication	foaf: Person	URI	M	纳米出版物的作者
dcterms: created	np: Nanopublication	/	literal	M	纳米出版物的生成时间
pav: version	np: Nanopublication	/	literal	R	纳米出版物的版本号

### 2.3 RDF 命名图分析

纳米出版物的 RDF 命名图<sup>[15]</sup>可直观反映纳米出版物结构及其语义关系,包括头部图、论断图、出处图和出版信息图等。图 2 表示纳米出版物头部信息,借助属性 np: hasAssertion、np: hasProvenance 及 np: hasPublicationInfo 分别描述纳米出版物与论断、出处和出版信息间的关系。

论断图通过复用领域本体或专业词表,规范表征论断的语义特征及语义关系。图 3 是信息检索领域中中文学位论文纳米出版物的论断图示例,包含事件评估和实验组成等。该示例使用 ir: innerEvaluation 描述实验任务、实验过程及实验结果,使用 ir: hasExperimentComponent 描述实验数据集和运行细节,使用 ir: belongsToIRSystem 表示信息检索系统。

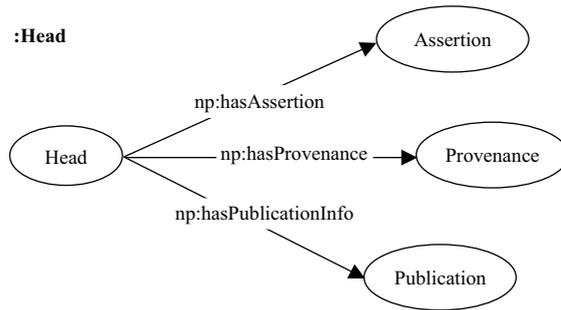


图 2 中文学位论文纳米出版物的头部图示例

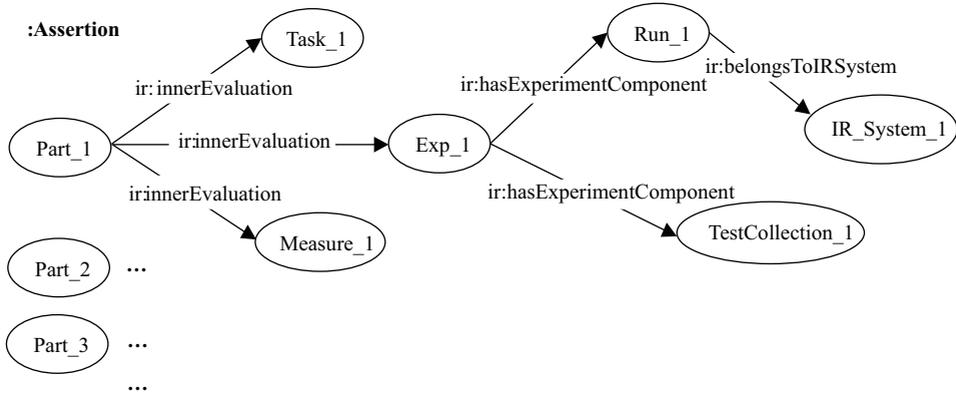


图 3 中文学位论文纳米出版物的论断图示例

图 4 复用 PAV 本体属性 pav: authoredBy、pav: createdOn 和 pav: derivedFrom 等分别描述论断作者、论断提出时间及论断来源论文等背景信息,用于追踪论断出处。图 5 描述了中文学位

论文纳米出版物的出版信息,通过属性 pav: createdBy、dcterms: created 和 pav: version 分别描述生成者、生成时间及相应版本等。

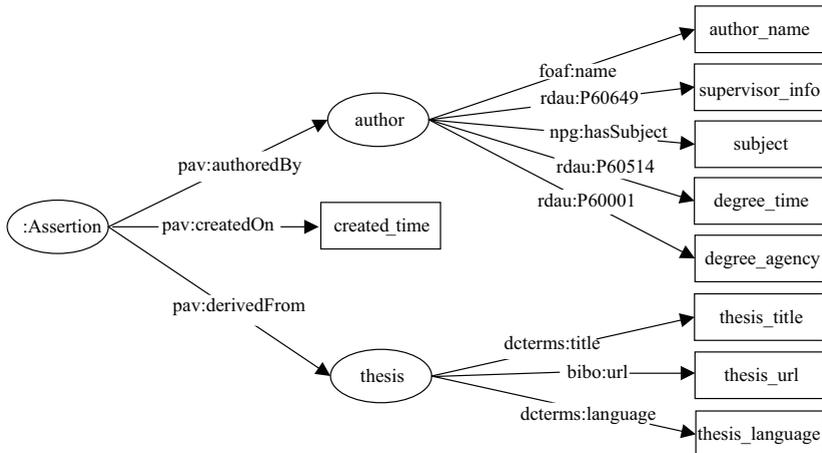


图 4 中文学位论文纳米出版物的出处图示例

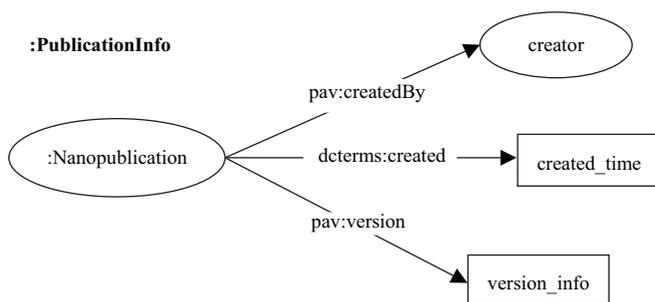


图5 中文学位论文纳米出版物的出版信息图示例

### 3 信息检索领域中文学位论文纳米出版物构建实验

本研究以信息检索领域中文学位论文为研究对象,设计并开展纳米出版物的构建实验,以验证研究设计的纳米出版物扩展模型的适用性。

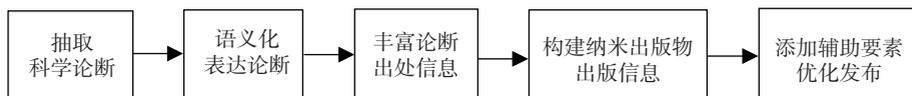


图6 中文学位论文纳米出版物生成步骤

(1) 抽取科学论断:信息检索领域的相关研究基于科学实验开展,所含论断通常表现为事实性的,可以通过篇章结构解析定位并抽取论断<sup>[16]</sup>。

(2) 语义化表达论断:借助自然语言处理技术对抽取的论断文本进行断句或分词处理,经语义分析后,将每个论断转化为三元组形式,使用 URI 标识符标识,并参考领域本体或专业词表描述论断的语义特征。

(3) 丰富论断出处信息:定义论断来源论文的作者、组织机构、论断提出时间和地点等,包含论文 URL、标题、ORCID 等信息,可通过 CNKI 和万方数据库等获取。

(4) 构建纳米出版物出版信息:定义纳米出版物的生成者、生成时间、版权和版本等信息。

### 3.1 实验步骤

本研究在 Open PHACTS 提出的生命医学领域数据纳米出版物发布步骤<sup>[5]</sup>的基础上,结合中文学位论文的结构特点,制定中文学位论文纳米出版物的生成步骤,见图 6。

(5) 添加辅助要素优化发布:定义纳米出版物 ID、完整性密钥等,遵循 Kuhn<sup>[12]</sup> 推荐的原则,采用可信任的 URIs 控制作者身份和版本等。

### 3.2 实验对象选取

考虑学位论文数据的全面性和权威性等因素,本研究选取万方中国学位论文全文数据库<sup>①</sup>为数据源。以主题词“信息检索”进行检索,限定检索条件为“中文”和“学位论文”,论文出版时间限定为 2014 至 2019 年,抽取主题词相关性排序前十的论文页面信息用于验证模型的适用性,见图 7。

图 8 是检索结果按相关度排序得到的与检索词相关性最高的中文学位论文示例。其中,

① <http://c.wanfangdata.com.cn/thesis>

上图是学位论文的摘要及所含主要论断信息， 息等论断出处信息， 纳米出版物的出版信息可  
下图是论文作者、在线出版时间及学位相关信息 在最终发布时定义。

url	题名	摘要内容	关键词	作者	学位授予单位	授予学位	学科专业	导师姓名	学位年度	语种	分类号	在线出版日期
<a href="http://w">http://w</a>	基于LightLDA的信息检索方法研究	如今, 互联网	信息检索	韩潇	华中师范大学	硕士	计算机软	何婷婷	2017	中文	TP391	2018年1月26日
<a href="http://w">http://w</a>	信息检索查询性能预测	查询性能预测	信息检索	陶永全	江苏大学	硕士	计算机应	吴胜利	2015	中文	TP391	2015年10月12日
<a href="http://w">http://w</a>	基于改进向量空间模型的信息检索	互联网的发展	信息检索	吴莹	华南师范大学	硕士	软件工程	蒋运承	2016	中文	TP391	
<a href="http://w">http://w</a>	基于关键词的关系数据库时态信息检索	随着大数据时	时态数据	张晓民	大连海事大学	硕士	计算机科	张俊	2017	中文	TP311.131	2017年07月31日
<a href="http://w">http://w</a>	LDA在信息检索中的应用研究	随着互联网的	语义主题	何锦群	天津理工大学	硕士	计算机技	赵德新	2014	中文	TP391.3	2014年7月15日
<a href="http://w">http://w</a>	多源教育资源信息检索与融合关键技	近年随着互联	教育资源	陈肖崑	东南大学	硕士	计算机科	徐立臻	2017	中文	TP311.13	2018年05月28日
<a href="http://w">http://w</a>	基于PageRank和贝叶斯网络的信息检索	随着互联网技	信息检索	吴龙龙	华南师范大学	硕士	计算机科	蒋运承	2016	中文	TP391	
<a href="http://w">http://w</a>	基于语义的对等网络信息检索技术研究	P2P网络以其	对等网络	吕俊	江苏大学	硕士	计算机技	吴胜利	2016	中文	TP393.09	2016年11月3日
<a href="http://w">http://w</a>	基于深度神经网络的音乐信息检索	音乐分类从本	音乐情绪	赵天坤	北京邮电大学	硕士	计算机科	张雷	2015	中文	TN912.34	
<a href="http://w">http://w</a>	面向信息检索的微博权威性计算方法	随着无线网技	微博 文本	危艳华	华中师范大学	硕士	计算机技	涂新辉	2017	中文	TP391.1	2018年1月26日

图7 信息检索领域相关度排序前十的中文学位论文

**目录** 基于LightLDA的信息检索方法研究

[↓ 下载](#)
[□ 在线阅读](#)
[☆ 收藏](#)
[↗ 导出](#)
[↻ 分享](#)

**摘要:** 如今, 互联网上的数据量呈指数级增长趋势, 各种各样的数据与日俱增。面对如此浩瀚的数据海洋, 如何从中快速、准确地获取用户需要的信息越发成为一个亟待解决的问题。这也是如今的信息检索技术所要面对的巨大挑战。

目前, 信息检索中引入语义信息的主流做法是使用机器学习方法LDA训练主题模型。尽管融入LDA主题信息后, 检索性能有所提升, 但是LDA模型的算法复杂度太高, 导致其训练主题信息时容易受限于语料的规模和主题的数量, 因此不能很好地解决如今大数据时代面临的检索问题。2015年微软分布式、高性能工具LightLDA的开源, 便看到了这一问题解决的希望。本文正是着眼于大数据时代面临的检索问题, 探讨了LightLDA在信息检索中应用的可行性和有效性, 主要工作包括以下两个方面:

第一, 将LightLDA应用到信息检索模型中。利用LightLDA对几个规模较大的TREC数据集进行了主题信息的训练, 并将训练后的主题信息融入到语言模型框架中, 构建了基于主题信息的检索模型(简称LLBDM); 然后在此基础上, 利用信息熵的概念尝试构建了新的检索模型(简称LMIE)。最后, 把这两种模型的效果和信息检索中的Baseline做了比较, 并分析了相关参数对模型的影响。通过实验, 验证了LightLDA在信息检索模型中的可行性和有效性。

第二, 将LightLDA应用到伪相关反馈中。利用LightLDA对伪相关反馈文档进行了主题信息的训练, 然后基于Rocchio伪相关反馈框架和上述主题信息, 构造了伪相关反馈模型Rocchio-LightLDA。最后, 把Rocchio-LightLDA模型和伪相关反馈中的Baseline模型做了比较, 并分析了相关参数对模型的影响。通过实验, 验证了LightLDA在伪相关反馈中的可行性和有效性。

通过以上两方面的研究, 成功地把LightLDA应用到了信息检索领域, 为如今大数据时代面临的检索任务提供了一种可行的解决方案, 同时对基于海量数据检索的相关问题也有一定的借鉴意义。

**1 论断信息**

**2 论断出处信息**

关键词: [信息检索](#) [检索模型](#) [主题信息](#) [伪相关反馈](#)

作者: 韩潇

学位授予单位: [华中师范大学](#)

授予学位: 硕士

学科专业: 计算机软件与理论

导师姓名: 何婷婷

学位年度: 2017

语种: 中文

分类号: TP391

在线出版日期: 2018年01月26日

图8 中文学位论文示例的信息界面

### 3.3 内容信息获取

(1) 论断抽取: 首先基于 CISP (Core Information about Scientific Papers) ① 篇章结构模型, 人工识别摘要文本中的研究目的、动机、研究对象、研究方法、实验过程、观察、实验结果和结论等组件, 见表 7。再参照信息检索本体核心类与

属性, 结合论断命名图, 对不同组件进行描述与抽取, 得到相应的论断, 包括实验任务、实验组成、信息检索系统及评估结果等, 见表 8。获得关键论断信息后, 借助信息检索本体表证论断间语义关系, 生成 RDF 三元组。

① <https://www.aber.ac.uk/en/media/departmental/impacs/computerscience/pdfs/ReportCISPshort.pdf>

表 7 中文学位论文示例中论断语句抽取

论断类型	语句
实验步骤	LightLDA 模型训练 TREC 数据集,生成主题信息
	主题信息融入 LM 语言模型框架,构建 LLBDM 模型
	LLBDM 模型结合信息熵概念,构建 LMLIE 模型
	LightLDA 模型训练伪相关反馈文档 PRF,生成主题信息
	主题信息结合 Rocchio 伪相关反馈框架,构造 Rocchio-LightLDA 模型
评估实验	LLBDM、LMLIE 模型对比信息检索中 Baseline 模型,性能表现更优
	Rocchio-LightLDA 模型对比伪相关反馈中的 Baseline 模型,性能表现更优
实验任务	LightLDA 在信息检索模型中应用的可行性和有效性
	LightLDA 在伪相关反馈中应用的可行性和有效性

表 8 中文学位论文示例中论断一和论断二的描述框架

描述对象	类	描述内容	取值类型
实验任务	ir:Task	(一)LightLDA 在信息检索模型中应用的可行性和有效性 (二)LightLDA 在伪相关反馈中应用的可行性和有效性	literal
实验组成	ir:Experiment	由信息检索系统和测试集合构成	URI
测试集合	ir:TestCollection	(一)TREC 数据集和主题信息 (二)伪相关文档 PRF 和主题信息	URI
信息检索系统	ir:IRSystem	由信息检索系统核心组件和生成的信息检索模型构成	URI
信息检索系统 核心组件	ir:IRSystemCore	(一)LightLDA,输入为 TREC 数据集,输出为主题信息 (二)LightLDA,输入为 PRF 文档,输出为主题信息	URI
信息检索模型	ir:IRModel	(一)LLBDM,基于 LM 模型,输入为主题信息 (一)LMLIE,基于 LLBDM 模型,输入为信息熵 (二)Rocchio_LightLDA,基于 Rocchio 模型,输入为主题信息	URI
评估结果	ir:EvaluationMeasure RankedResult	(一)LLBDM 与 LMLIE 性能均优于 Baseline 模型 (二)Rocchio_LightLDA 性能优于 Baseline 模型	URI

(2) 出处信息获取:表 9 使用了表 5 定义的 学位论文作者、论断来源论文信息外,其他取值  
 属性描述上述示例摘要中的论断出处信息。除 类型均为 literal。

表9 中文学位论文示例对应的论断出处信息描述记录实例

描述对象	属性	描述内容	取值类型
学位论文作者	pav: authoredBy	由姓名、导师、专业、学位时间和学位授予单位组成	URI
学位论文作者姓名	foaf: name	韩潇	literal
指导老师	rdau: P60649	何婷婷	literal
学科专业	npg: hasSubject	计算机软件与理论	literal
学位授予年度	rdau: P60514	2017	literal
学位授予单位	rdau: P60001	华中师范大学	literal
论断提出时间	pav: createdOn	2018年01月26日	literal
论断来源论文信息	pav: derivedFrom	由标题、链接和语种组成	URI
论断来源论文标题	dcterms: title	基于 LightLDA 的信息检索方法研究	literal
论断来源论文链接	bibo: url	Y3298156#	literal
论断来源论文语种	dcterms: language	中文	literal

### 3.4 实验结果

纳米出版物模型的形式化表达可采用 W3C RDF<sup>①</sup> 数据模型实现,本研究基于示例论文生成信息检索领域中文学位论文纳米出版物及其 Turtle 形式数据,并采用 RDF 命名图表达。

(1) 前缀与头部信息: 纳米出版物推荐使用

URI 命名事物,可通过名称空间前缀简化表达 URI,如前缀 ir 对应信息检索专用词表。头部信息由 :Head 声明,通过 :Nanopublication 声明纳米出版物实例,采用纳米出版物本体定义的属性描述类间关系。本研究所使用的前缀及头部信息见表 10。图 9 展示该示例与论断、出处及出版信息间的关联。

表10 中文学位论文示例的前缀与头部信息

头部信息的 RDF 数据	注释
<pre>@ prefix : &lt;http://www.example.org/nanopub/this-ir-example&gt;. @ prefix np: &lt;http://www.nanopub.org/nschema#&gt;. @ prefix ir: &lt;http://ifs.tuwien.ac.at/~admire/ir_ontology/ir#&gt;. @ prefix pav: &lt;http://purl.org/pav/&gt;. @ prefix foaf: &lt;http://xmlns.com/foaf/0.1/&gt;. @ prefix rdau: &lt;http://rdaregistry.info/Elements/u/&gt;. @ prefix npg: &lt;http://ns.nature.com/terms/&gt;. @ prefix uby: &lt;http://purl.org/olia/ubyCat.owl#&gt;. @ prefix rdfs: &lt;http://www.w3.org/2000/01/rdf-schema#&gt;. @ prefix dcterms: &lt;http://purl.org/dc/terms/&gt;. @ prefix sio: &lt;http://semanticscience.org/resource/&gt;. @ prefix bibo: &lt;http://purl.org/ontology/bibo/&gt;. @ prefix xsd: &lt;http://www.w3.org/2001/XMLSchema#&gt;. @ prefix wf: &lt;http://www.wanfangdata.com.cn/details/detail.do?_type=degree&amp;id=#&gt;.  :Head { : a np: Nanopublication; np: hasAssertion: Assertion; np: hasProvenance: Provenance; np: hasPublicationInfo: PublicationInfo. }</pre>	<p>纳米出版物和复用的词表及本体的前缀信息</p> <p>纳米出版物头部信息表征了纳米出版物实例和类间关系</p>

① <https://www.w3.org/RDF>

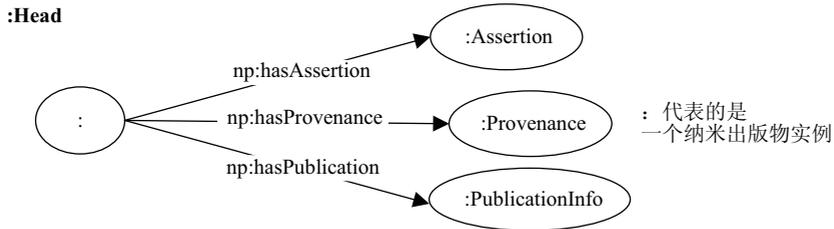


图9 中文学位论文示例的头部位

(2) 论断信息: 该示例论文由两个信息检索实验组成。表 11 展示论断一的实验主体信息, 其中 :Part\_1 描述实验一的组成。:Task\_1 代表信息检索任务, 指出该实验任务需要检验 LightLDA 在信息检索模型中应用的可行性和有效性, 其中 @ zh 表示该信息以中文字符串表达。:Exp\_1 旨在描述实验一的主要组成部分,

由 :IR\_Collection\_1 和 :Run\_1 组成, 分别描述该实验中信息检索的数据集和实验运行过程。其中 :IR\_Collection\_1 作为信息检索的测试集类, 对应论断中提及的 :TREC 和 :Topic\_info 两类数据集。Topic\_info 是整个实验的中间产物, 由 LightLDA 模型应用于 TREC 数据集生成。

表 11 中文学位论文示例中论断一主体信息

论断一主体信息的 RDF 数据	注释
<pre>:Part_1 a ir:EvaluationThing;   ir:innerEvaluation :Task_1;   ir:innerEvaluation :Exp_1;   ir:innerEvaluation :Measure_1.</pre>	:Part_1 描述论断一对应的完整实验
<pre>:Task_1 a ir:Task;   rdfs:comment "LightLDA 在信息检索模型中应用的可行性和有效性"@ zh.</pre>	:Task_1 描述实验一的主要任务目标
<pre>:Exp_1 a ir:Experiment;   ir:hasExperimentComponent :IR_Collection_1;   ir:hasExperimentComponent :Run_1.</pre>	:Exp_1 描述实验一的主要组成部分
<pre>:IR_Collection_1 a ir:TestCollection;   ir:hasCollection :TREC;   ir:hasTopics :Topic_info.</pre>	:IR_Collection_1 描述实验一中 IR 的数据集
<pre>:TREC a ir:Collection.</pre>	:Run_1 描述实验一中实验运行过程
<pre>:Topic_info a ir:Topics;   rdfs:comment "主题信息"@ zh.</pre>	

如表 12 所示, :Run\_1 是论断一的核心部分, 定义实验运行过程, 描述 :IR\_System\_1 中系统运行状况。该信息检索实验系统核心是

LightLDA 模型, 输入 TREC 数据集后可生成主题信息, 运行后生成两个模型: LLBDM 和 :LMLIE。如表 13 所示, :Measure\_1 描述实验一生成的两

个模型分别与:Baseline 进行对比评估的结果, 图表征论断一的基本信息。分析相关参数对模型的影响。图 10 使用 RDF

表 12 中文学位论文示例中论断一的实验运行信息

论断一的实验运行信息的 RDF 数据	注释
<pre> :Run_1 a ir:Run;       ir:belongsToIRSystem ;IR_System_1.  :IR_System_1 a ir:IRSystem;       ir:hasIRSystemCore ;LightLDA;       ir:fromIRSystemToIRModel ;LLBDM;       ir:fromIRSystemToIRModel ;LMLIE.  :LightLDA a ir:IRSystemCore;       sio:SIO_000230 ;TREC;       sio:SIO_000229 ;Topic_info;       rdfs:comment "Light Latent Dirichlet Allocation"@en.  :LLBDM a ir:IRModel;       sio:SIO_000641 ir:LM;       sio:SIO_000230 ;Topic_info;       rdfs:comment "LightLDA Based Document Model"@en.  :LMLIE a ir:IRModel;       sio:SIO_000641 ;LLBDM;       sio:SIO_000230 ;Information_entropy;       rdfs:comment "Language Model-LightLDA-Information Entropy"@en.  :Information_entropy a npg:Concept.</pre>	<p>:LightLDA 模型是该信息检索实验系统的核心 sio:SIO_000230 has input sio:SIO_000229 has output</p> <p>:LLBDM 是基于 ir:LM 和 :Topic_info 生成的模型 sio:SIO_000641 has basisir :LM 对应语言模型</p> <p>:LMLIE 是基于:LLBDM 和 :Information_entropy 生成的模型</p> <p>:Information_entropy 对应信息熵</p>

表 13 中文学位论文示例中论断一的实验评估信息

论断一的实验评估信息的 RDF 数据	注释
<pre> :Measure_1 a ir:EvaluationMeasureRankedResult.  :Measure_1 {   :LLBDM sio:SIO_000699 ;Baseline;   :LMLIE sio:SIO_000699 ;Baseline. }  :Baseline a ir:IRModel.</pre>	<p>:Measure_1 描述实验一生成的模型与:Baseline 对比评估的结果</p> <p>sio:SIO_000699 is greater than or equal to</p>

同理,:Part\_2 描述论断二中实验二的组成。:IR\_Collection\_2 描述实验二的测试集数据,其中:PRF 对应伪相关反馈文档,:Topic\_info 为 LightLDA 应用于:PRF 上生成的主题信息。:Run

\_2 描述以:LightLDA 为核心的 IR 系统:IR\_System\_2 及新模型:Rocchio\_LightLDA。通过将模型:LightLDA 应用于:PRF 数据集,输出:Topic\_info 主题信息,结合:Rocchio 伪相关反馈框架构

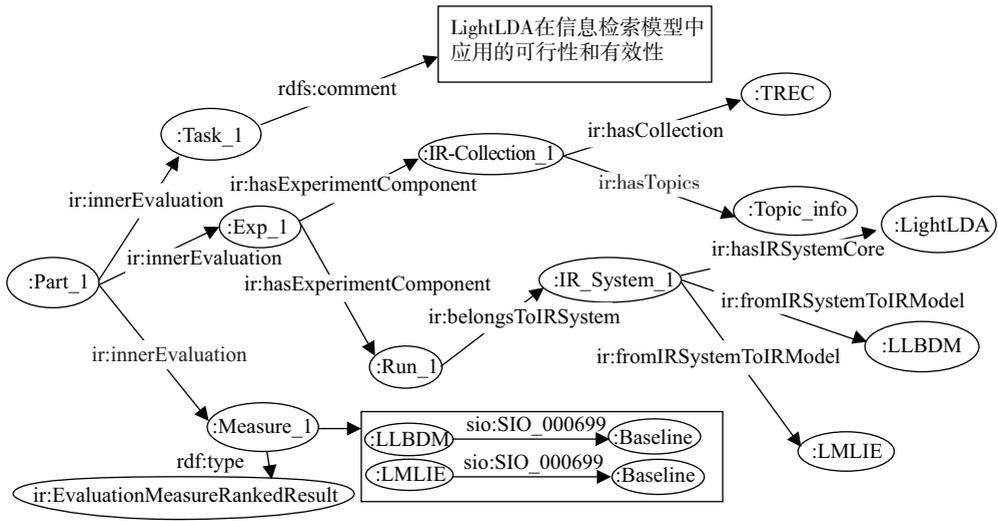


图 10 中文学位论文示例论断一的 RDF 图

造模型; Rocchio\_LightLDA。: Measure\_2 描述实验 2 生成的模型; Rocchio\_LightLDA 与 :Baseline 对比评估的结果。同理也可使用 RDF 图表征论断二的基本信息。

本研究基于表 14 和表 15 中的信息分别生成了出处图(见图 11)和出版信息图(见图 12)(RDF 命名图)。其中,纳米出版物的创建者信息采用 ORCID 标识。

(3) 出处及出版信息:遵循模型结构要求,

表 14 中文学位论文示例的出处信息

出处信息的 RDF 数据	注释
<pre> :Provenance { :Assertion pav; authoredBy :author. :author a foaf:Person; foaf:name "韩潇"@zh; rdau:P60649 "何婷婷"@zh; npg:hasSubject "计算机软件与理论"@zh; rdau:P60514 "2017"~xsd:date; rdau:P60001 "华中师范大学"@zh.  :Assertion pav; createdOn "2018-01-26"~xsd:date.  :Assertion pav; derivedFrom :thesis. :thesis a npg:Publication; dcterms:title "基于 LightLDA 的信息检索方法研究"@zh; bibo:url&lt;http://www.wanfangdata.com.cn/details/detail.do?type=degree&amp;id=Y3298156#&gt;; dcterms:language "中文"@zh.                     </pre>	<p>:Provenance 描述该论断的出处信息</p> <p>:author 描述论断作者相关信息</p> <p>pav:createdOn 描述论断生成时间</p> <p>:thesis 描述论断来源论文相关信息</p>

表 15 中文学位论文示例的出版信息

出版信息的 RDF 数据	注释
<pre> :PublicationInfo { : pav:createdBy &lt;https://orcid.org/0000-0002-8181-1168&gt;. : dcterms:created "2019-04-15T20:24:11+01:00"^^xsd:dateTime. : pav:version "1.0". } </pre>	<pre> :PublicationInfo 描述该纳 米出版物出版信息 : 指纳米出版物实例 </pre>

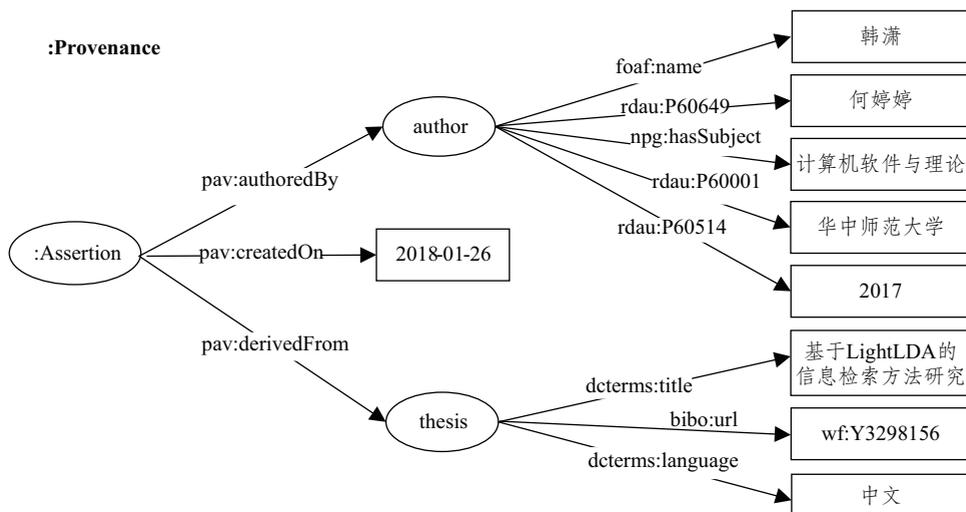


图 11 中文学位论文示例的出处图

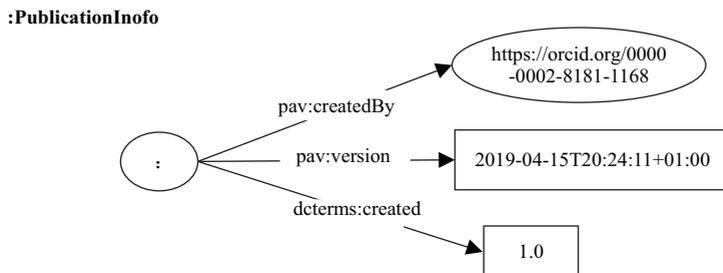


图 12 中文学位论文示例的出版信息图

#### 4 模型验证分析

本研究选取图 7 所示信息检索领域的十篇中文学位论文分别构建纳米出版物,以验证模型及方法的有效性。首先分析摘要文本特征,发现其所描述的科学论断主要是关于“实验过

程”,多含“论文主要做了如下工作”“主要研究工作如下”及“本文的主要任务有”等语句,通过表达先后顺序的连接词描述研究工作。因此,通过识别与定位重点语句和连接词对摘要内容进行预处理,内容拆分结果见表 16。随后,分别从实验任务、实验步骤及评估实验维度抽取论断,见表 17。

表 16 中文学位论文数据集摘要内容拆分结果片段

URL	题名	工作内容 1	工作内容 2
Y3298156	基于 LightLDA 的信息检索方法研究	第一,将 LightLDA 应用到信息检索模型中。利用 LightLDA 对几个规模较大的 TREC 数据集进行了主题信息的训练,并将训练后的主题信息融入到语言模型框架中,构建了基于主题信息的检索模型(简称 LLBDM);然后在此基础上,利用信息熵的概念尝试构建了新的检索模型(简称 LMLIE)。最后,把这两种模型的效果和信息检索中的 Baseline 做了比较,并分析了相关参数对模型的影响。通过实验,验证了 LightLDA 在信息检索模型中的可行性和有效性。	第二,将 LightLDA 应用到伪相关反馈中。利用 LightLDA 对伪相关反馈文档进行了主题信息的训练,然后基于 Rocchio 伪相关反馈框架和上述主题信息,构造了伪相关反馈模型 Rocchio-LightLDA。最后,把 Rocchio-LightLDA 模型和伪相关反馈中的 Baseline 模型做了比较,并分析了相关参数对模型的影响。通过实验,验证了 LightLDA 在伪相关反馈中的可行性和有效性。

表 17 中文学位论文数据集论断抽取结果片段

URL	题名	实验步骤 1	实验任务 1	评估实验 1	实验步骤 2	实验任务 2	评估实验 2
Y3298156	基于 LightLDA 的信息检索方法研究	①利用 LightLDA 对 TREC 数据集进行了主题信息的训练;②将训练后的主题信息融入到语言模型框架中,构建了基于主题信息的检索模型(简称 LLBDM);③利用信息熵的概念尝试构建了新的检索模型(简称 LMLIE)。	验证了 LightLDA 在信息检索模型中的可行性和有效性。	把这两种模型的效果和信息检索中的 Baseline 做了比较,并分析了相关参数对模型的影响。	①利用 LightLDA 对伪相关反馈文档进行了主题信息的训练;②然后基于 Rocchio 伪相关反馈框架和上述主题信息,构造了伪相关反馈模型 Rocchio-LightLDA。	LightLDA 在伪相关反馈中的可行性和有效性。	最后,把 Rocchio-LightLDA 模型和伪相关反馈中的 Baseline 模型做了比较,并分析了相关参数对模型的影响。

在摘要拆分与论断抽取基础上,本研究应用所构建的模型生成相应的纳米出版物 RDF 数据。实验结果表明,该模型较清晰地表征了信息检索领域中学位论文中科学论断的语义特征,支持形式化发布实验方法和实验数据,保证了科学实验数据的完备性和可重复性。在传统出版物基础上加工生成的纳米出版物,可以弥补传统出版物在知识关联上的缺陷<sup>[17]</sup>,提高知识内容的机器可读性,辅助用户高效获取知识。对于结构多变、领域多样的中文学位论文纳米

出版物的生成而言,本研究构建的模型可能存在特定领域或特定语义的表达局限,但提出的模型扩展方法具有实际应用价值和参考价值。

本研究在纳米出版物生成实验中利用了传统出版物信息,纳米出版物与传统出版物对于学术交流与知识传播都具有重要的作用。相较于传统出版物,纳米出版物可集成不同知识资源,将知识单元结构化、关联化,提高知识单元的机器可读性。基于此,知识挖掘和语义组织可有效帮助研究者开展知识发现。

## 5 结语与展望

本研究针对纳米出版物通用模型在论断语义表达方面的不足,通过复用领域本体和语义出版本体构建了中文学位论文纳米出版物模型。本研究重点关注信息检索领域中文学位论文的语义组织研究,在纳米出版物通用模型的基础上,通过复用信息检索领域本体对科学论断语义特征及论断间语义关系的形式化表征,提出了更适用于中文学位论文的纳米出版物模型。该模型具有较强的科学论断表达能力,可描述与揭示实验参数、实验模型和测试集合等特定语义特征。该模型也较全面地涵盖了信息检索核心类,并规范了类间关系,为自动抽取论断和语义关系提供了语义数据表达模型。本研究还开展了一系列实证应用探索,生成了纳米出版物数据集,验证了所构建模型的适用性,能为纳米出版物在具体领域的语义组织应用研究提供参考。

现阶段,大规模批量处理纳米出版物 RDF 数据有技术难度,大多数研究团队采用既有 RDF 数据集来发布纳米出版物,而针对科学论

文原始文本的纳米出版物数据集建设相对较少。本研究构建的模型可利用术语识别和实体抽取等手段,借助机器学习与深度洗涤等技术,实现中文学位论文语句层的论断抽取与自动标注,为自动化构建纳米出版物提供模型与方法支撑。自然语言表达的科学论文内容语义较复杂,存在识别实验任务难、实验步骤详细程度不统一、实验评估环节缺乏等问题。未来还需进一步创新科学论文内容表示模型,建构大规模、高质量、相互关联的科学论文语料库,为实现科学论文论断信息的大规模抽取与表示提供数据基础。科学论文由具有一定语义特征与逻辑关联的知识单元构成。未来纳米出版物在科学论文方面的应用研究仍需关注细粒度知识单元的形式化表示与语义关联,探索建设多层次、多粒度、多维度的科学论文内容数据集。大数据时代,图情档及出版机构正从信息服务向知识服务转型,通过采用纳米出版物等语义出版模型,可加强知识标引深度和可信赖引用。未来可先由论文作者提交纳米出版物以构建所需基础数据,再由出版社或研究发布者通过词表自动对应和数据生成技术自动转换并发布纳米出版物,实现知识资源的自动整合、链接和互操作。

## 参考文献

- [1] 索传军,盖双双. 知识元的内涵、结构与描述模型研究[J]. 中国图书馆学报,2018,44(4):54-72. (Suo Chuanjun,Gai Shuangshuang. The connotation,structure and description model of knowledge unit[J]. Journal of Library Science in China,2018,44(4):54-72.)
- [2] 王晓光,李梦琳,宋宁远. 科学论文功能单元本体设计与标引应用实验[J]. 中国图书馆学报,2018,44(4):73-88. (Wang Xiaoguang,Li Menglin,Song Ningyuan. Design and application of scientific paper functional units ontology[J]. Journal of Library Science in China,2018,44(4):73-88.)
- [3] Shotton D. Semantic publishing:the coming revolution in scientific journal publishing[J]. Learned Publishing,2009,22(2):85-94.
- [4] 王晓光,陈孝禹. 语义出版:数字时代科学交流系统新模型[J]. 出版科学,2012(4):81-86. (Wang Xiaoguang,Chen Xiaoyu. Semantic publishing:a new model for scientific communication[J]. Publishing Journal,2012(4):81-86.)
- [5] Mons B,Velterop J. Nano-publication in the e-science era[C]//Workshop on Semantic Web Applications in Scientific Discourse,2009.

- [ 6 ] Groth P, Gibson A, Velterop J. The anatomy of a nanopublication[J]. *Information Services and Use*, 2010, 30(1-2): 51-56.
- [ 7 ] 王晓光, 宋宁远. 语义出版物的内容组织架构研究——基于纳米出版物和微型出版物的比较分析[J]. *出版科学*, 2017, 25(4): 20-27. ( Wang Xiaoguang, Song Ningyuan. Content schema of semantic publication: a comparative analysis based on nanopublication and micropublication[J]. *Publishing Journal*, 2017, 25(4): 20-27. )
- [ 8 ] 牛丽慧, 欧石燕. 纳米出版及其应用研究进展[J]. *图书情报工作*, 2018, 62(7): 125-133. ( Niu Lihui, Ou Shiyuan. Research advances of nanopublication and its applications[J]. *Library and Information*, 2018, 62(7): 125-133. )
- [ 9 ] Lipani A, Piroi F, Andersson L, et al. Extracting nanopublications from IR papers[C]//*Information Retrieval Facility Conference: Multidisciplinary Information Retrieval*, 2014: 53-62.
- [ 10 ] 廖建军. 基于 Nanopublication 的知识服务架构解析[J]. *图书情报工作*, 2017, 61(17): 131-138. ( Liao Jianjun. Knowledge service architecture analysis based on nanopublication[J]. *Library and Information Science*, 2017, 61(17): 131-138. )
- [ 11 ] Golden P, Shaw R. Period assertion as nanopublication: the PeriodO period gazetteer[C]//*The 24th International Conference on World Wide Web*. 2015: 1013-1018.
- [ 12 ] Kuhn T, Barbo P E, Nagy M L, et al. Broadening the scope of nanopublications[C]//*The 10th Extended Semantic Web Conference*. LNCS 7882, 2013: 487-501.
- [ 13 ] Clare A, Croset S, Grabmuller C, et al. Exploring the generation and integration of publishable scientific facts using the concept of nano-publications[C]//*In CEUR Workshop Proceedings*, 2011: 13-17.
- [ 14 ] Lipani A, Piroi F, Andersson L, et al. An information retrieval ontology for information retrieval nanopublications[C]//*International Conference of the Cross-Language Evaluation Forum for European Languages: Information Access Evaluation*, Springer, Cham, 2014: 44-49.
- [ 15 ] Carroll J J, Bizer C, Hayes P, et al. Named graphs, provenance and trust[C]//*International World Wide Web Conference*. ACM Press, 2005: 613-622.
- [ 16 ] 苏云梅, 武建光. 纳米出版物语义模式及其与知识发现的关系[J]. *中华医学图书情报杂志*, 2015, 24(12): 15-18. ( Su Yunmei, Wu Jianguang. Nanopublication semantic pattern and its relation with knowledge discovery[J]. *Chinese Journal of Medical Library and Information Science*, 2015, 24(12): 15-18. )
- [ 17 ] 吴思竹, 李峰, 张智雄. 知识资源的语义表示和出版模式研究——以 Nanopublication 为例[J]. *中国图书馆学报*, 2013, 39(4): 102-109. ( Wu Sizhu, Li Feng, Zhang Zhixiong. Research on semantic representation and publishing schema of knowledge resource: take nanopublication as an example[J]. *Journal of Library Science in China*, 2013, 39(4): 102-109. )

李春秋 北京师范大学政府管理学院讲师。北京 100875。  
徐曾旭林 中国科学院文献情报中心硕士研究生。北京 100190。  
宋宁远 南京大学信息管理学院博士后。江苏 南京 210023。  
王晓光 武汉大学信息管理学院教授。湖北 武汉 430072。

(收稿日期: 2021-03-11)