情报理论与实践
Information Studies:Theory & Application
ISSN 1000-7490,CN 11-1762/G3

《情报理论与实践》网络首发论文

题目：　科学事件知识图谱构建研究
作者：　白如江，周彦廷，王效岳，王志民
网络首发日期：　2020-03-18
引用格式：　白如江，周彦廷，王效岳，王志民．科学事件知识图谱构建研究．情报理论与实践. http://kns.cnki.net/kcms/detail/11.1762.G3.20200317.1708.008.html

●白如江[1]，周彦廷[2]，王效岳[1]，王志民[3]

（1.山东理工大学科技信息研究所， 山东 淄博 255049；2.中国人民大学信息资源管理学院， 北京 100872；3.山东理工大学齐文化研究院， 山东 淄博 255049）

科学事件知识图谱构建研究*

摘 要：[目的/意义]在新的信息背景下，以文献为单元的知识组织方式已经无法满足用户的需求，对科学文献的语义化与结构化的知识组织研究成为目前图情领域研究的热点之一。[方法/过程]文章通过提出科学事件的概念，在现有科研元数据以及知识元框架的理论基础上，构建了科学事件元数据模型，将科学元事件划分为科学事件主语、科学事件谓语、科学事件宾语三个部分。利用 LTP 语言云根据本文所构建的科学事件元数据模型，对图情领域的科学文献进行科学事件的语义化与结构化知识组织，将所得数据存入图数据库 Neo4j 中并构建了以图情领域为例的科学事件知识图谱。[结果/结论]实验结果证实了文章所提方法可行有效。

关键词：知识元；元数据；知识元描述模型；科学文献；知识图谱

Research on the Construction of Knowledge Graph of Scientific Events

Abstract: [Purpose / significance] Under the new information background, the document-based knowledge organization has been unable to meet the needs of users. The research on semantic and structured knowledge organization of scientific documents has become one of the hot topics in the field of graphic information. [Method/ process] By putting forward the concept of scientific events, this paper constructs a metadata model of scientific events on the basis of existing metadata of scientific research and knowledge meta-framework, and divides scientific meta-events into three parts: subject of scientific events, predicate of scientific events and object of scientific events. Based on the metadata model of scientific events constructed in this paper, the LTP language cloud is used to organize the semantic and structured knowledge of scientific events in the field of graphics and information. The data are stored in the graph database Neo4j and the knowledge graph of scientific events is constructed, taking the field of library and information science as an example. [Method/ process] The experimental results show that the method proposed in this paper is feasible.

Keywords: knowledge element; metadata; description model of knowledge element; scientific literature; knowledge graph

目前科学文献的知识组织方式主要是以单篇文献为单位，科研人员对能够快速的、准确的从数量巨大且日益快速增长的科学文献中获取所需知识的需求持续增长。因此，现有的科学文献的知识组织方式已无法满足目前知识获取的需求，探索基于科学文献的科学事件知识图谱在知识组织方面势在必行。

*本文为教育部哲学社会科学研究重大课题攻关项目“稷下学派文献整理与数据库建设研究”（项目编号：19JZD011）和山东省高等学校人才引育计划“科技大数据研究创新团队”的研究成果。

1 相关研究

进入 21 世纪以来，新一轮科技革命正在孕育兴起，全球科技创新呈现出新的发展态势和特征。习主席指出“面对科技创新发展新趋势，世界主要国家都在寻找科技创新的突破口，抢占未来经济科技发展的先机。我们不能在这场科技创新的大赛上落伍，必须迎头赶上、奋起直追、力争超越”[1] 。

科学研究水平成为综合国力的重要组成部分，各国对科学研究的投入大幅度的增加，科学研究成果的产出量也随之爆发式的增长[2-4]。科学论文作为科学研究活动的主要的知识形态展现成果，同时也是科学研究成果的主要表现形式。据有关统计，在近几十年里，科学研究成果的产出量在全球范围里大约以九年为一个周期进行实现翻倍式的数量增长[5]。根据“Science”杂志的报道，平均每二十秒钟，便有一篇科学文献被发表[6]。

当前科学文献的知识组织方式大多还是以文献为单位，对科学文献内在知识内容的描述与揭示较少，虽然科学文献本身存在一些基础的元数据，如标题、作者、发表时间、所属期刊等。从计算机处理和读者用户两个角度出发，读者可以从科学文献的外部元数据获得关于该科学文献的特征信息，却无法获取有关该科学文献内容的语义化信息，使得读者无法快速的获得所需的知识。计算机擅长处理结构化的数据，然而信息主要以非结构化的形式存储和传播，为了让计算机能够处理这些信息，就需要理解这些非结构化形式数据蕴含的语义，分析其中的语义单元之间的表达形式，从而将其转成为结构化的形式。然而，计算机在面对缺失了结构化、语义化信息的科学文献资源时，难以对其进行适当的处理，无法提供更加智能、快速、准确的科学文献资源检索服务，从而无法进一步提高读者获取所需知识的效率[7]。科学文献的主要信息是非结构化的形式，如何将科学文献中非结构化的信息换成结构化、语义化的信息，并组织成为知识是目前图书馆学所面临的主要问题之一。

事件的概念在各个研究领域具有不同的解释，目前人们对事件还未有统一的认识。科学研究事件（以下简称“科学事件”）在科学研究活动中广泛的存在。科学文献作为科学研究活动的成果性产物，同样也是对科学事件的记录与表述，我们可以将一篇篇科学文献看做一个个科学事件，探索不同科学文献间的联系，从本质上讲便是探寻不同科学事件间的联系。通过引文来发现科学文献间的关联关系存在着一些弊端，目前存在大量的自引、虚引、魅引现象，而且对一篇科学文献的引用还存在很多可能的原因，可能是对某个概念的解释、对引文的评判、对某一方法的改进等，很难准确地说明两篇科学文献间存在的语义关系[2]。从科学事件的角度对科学文献进行知识的组织是对目前存在的科学文献元数据以及科学文献间存在的关系的补充，科学事件不单含有所要表征的科学文献的元数据，还具有科学文献的元数据所不具有的语义信息。

知识的语义化、结构化和有序化，是知识共享和应用的基础。知识的编码化和数字化形成了知识库。图示结构是一种能有效表示数据之间结构的表达形式，因此，人们考虑将数据中蕴含的知识用图的结构进行形式化表示，基于图论的知识图谱是计算机技术的发展与知识工程的研究相结合的产物。知识图谱作为知识库的继承者，具有可推理性、准确性、智能化等优点[8]。

在此背景下，本课题为以科学文献为研究对象，定义了对应波普尔 “第三世界”理论的科学事件，对科学事件进行识别和抽取并构建知识图谱，有助于满足科研工作者的知识需求，并通过可视化展示，有利于帮助科研工作者把握科技创新方向。

2 科学事件

科学文献是一系列科学事件的载体，科学文献的文本中记录了这些科学事件，要想获得这些科学事件，首先需要明确科学事件的定义。

科学事件是蕴含在科学文献中的事件，是关于科学研究的事件，其与客观世界中的事件具有很大的差异性，这种差异性主要取决于思维方式。科学研究活动是客观世界与脑部思维活动相结合的产物，科学研究的推动因素很大一部分是根据客观世界所获取的信息通过大脑的思维活动所产生。在波普尔的三个世界理论中阐述了存在的三个世界：第一个世界是物理世界，对应着我们所谓的客观世界，简称为世界 1，发生在世界 1 的事件便是我们所常知的具有事件、地点、人物等因素的物理事件；第二个世界是精神世界，是一切脑部思维活动、主观精神活动的世界，对应着我们所谓的精神世界，简称世界 2，世界 2 对世界 1 具有反馈作用，并会通过人们的躯体反馈出来，发生在世界 2 的事件便是我们的精神活动，本文称为精神事件；第三个世界是人类精神产物的世界，从广义上来说也是一切主观精神的活动的产物世界，如语言、艺术、文字、科学研究、理论猜测等一切抽象的以及具体的精神产物[9-13]。如图 1 所示。

科学事件是对科学研究活动的过程的陈述，所以科学事件是属于世界 3 范畴的事件，即是波普尔的“三个世界”理念中的客观的精神世界中的事件，区别于物理事件于精神事件，他具有以下特点：

1）非物理性。区别于物理事件，物理事件具有明确的时间与空间的属性，而科学事件的时间与空间并不明确，难以确定科学事件发生的具体时间与地点。

2）可描述性。区别于精神事件，精神事件是我们在大脑中的思维活动，具有难以表述的特点，而科学事件是明确的，是可以表述出的知识。

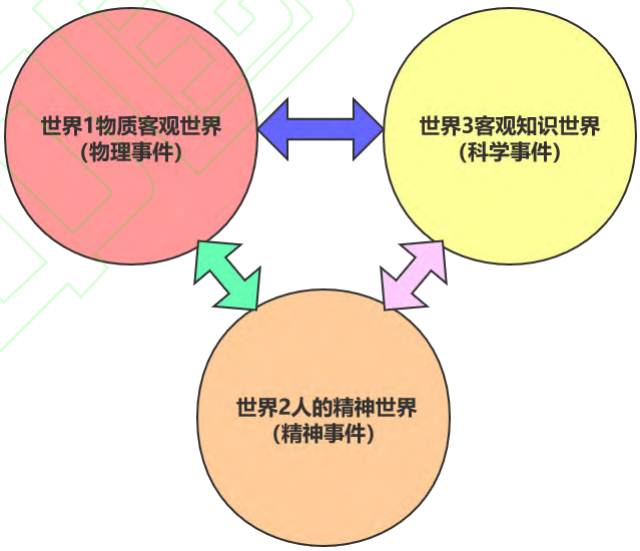

图 1 波多尔的“三个世界”

通过对科学文献的大量前期调研，根据科学文献的特点，本文认为每一篇科学文献只包含一件科学事件，科学文献的语言论述方式是属于世界 3 客观知识世界的语言论述方式，与世界 1 物质客观世界的语言论述方式是具有极大的差异性的，科学文献是通过“提出问题→研究问题→解决问题”这一流程的学术成果，科学文献是这一学术流程的描述总结性成果，其内容的语义性总结便是科学文献的题目，题目便是对该科

学文献所含科学事件的语义表述，如：“多维度视角下学科主题演化可视化分析方法研究——以我国图书情报领域大数据研究为例”“语义角色标注及其在科技情报分析中的应用研究”等。

2.1 科学元事件

科学事件是多层次的，多维在于其实体与关系的多维度性，多层次性在于其结构是多层次的。一篇科学文献是对一个问题的提出以及解决，所以一篇科学文献只包含了一个科学事件，但是一个科学事件是由多个更细粒度的科学事件单元组成，本文将这种更细粒度的科学事件单元称之为科学元事件（Scientific Meta-events），科学事件是对单篇科学文献的宏观事件性的描述，而科学事件通过科学元事件对科学文献进行微观的事件性描述，其结构如图 2 所示。

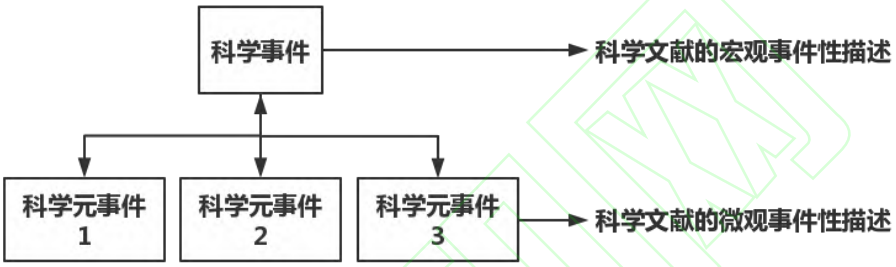

图 2　科学事件结构图

“元事件”（Meta-events）是一个很抽象的名词，根据中国人民大学索传军教授对知识组织的相关研究[14]，本文定义一个“元事件”是由三部分组成：Subject（加菲尔德）——Action（提出）—— Object（引文索引概念）。在科学文献中存在的“元事件”，这里的“元事件”便是“科学元事件”。

科学元事件的叙述中会包含一些科学文献的语言特点，在语言的论述上可能会出现两种情况：①存在完整的句法结构，主语、谓语、宾语；②句法结构不完整，只含有谓语、宾语，缺少主语。

以“首先对数据进行预处理，然后进行基于 PLDA 的主题识别。”为例，可以从中提取出两个科学元事件，第一个科学元事件为“数据（主语）——进行（谓语）——预处理（宾语）”，第二个科学元事件为“（缺少主语）——进行（谓语）——基于 PLDA 的主题识别（宾语）”，识别出的第一个科学元事件为句法结构完整的科学元事件，识别出的第二个科学元事件为缺少主语的科学元事件，具体如图 3 所示。

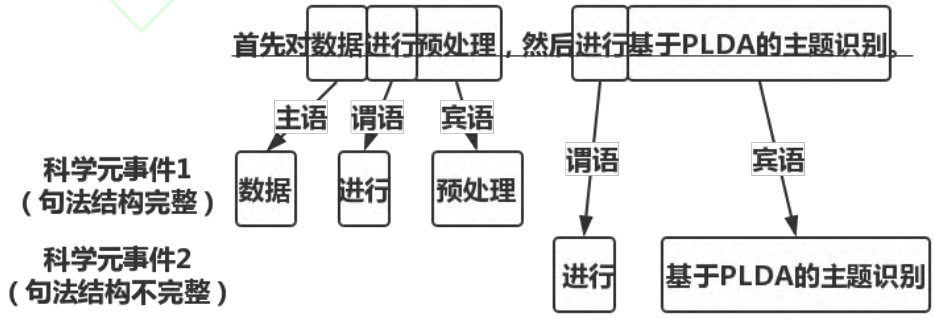

图 3　科学元事件举例

2.2 科学事件定义

基于以上研究，本文对**科学事件的定义为**：科学事件为某一同源关系下所有科学元事件所构成的集合，科学事件 SE={SME1,SME2,SME3,.....,SME*n*}，其中 SMEn 为科学元事件，n 为科学元事件的数量。

2.3 科学事件元数据模型

知识图谱的元数据模型是知识图谱的数据模型，包含了构建该知识图谱所需的数据类型以及这些数据类型之间的关联关系，知识图谱中每个节点都为一个实体，实体与实体间的连线为实体间的关系，知识图谱的元数据模型是对实体以及实体间的关系的模型化描述。在构建知识图谱之前，首先要需要进行知识建模，即根据数据构建知识图谱的元数据模型。

为了准确、有效的识别出蕴含在科学文献中的科学事件，在借鉴现有学术元数据以及知识元框架[14-17]的基础上，本文根据科学文献的外部特征以及科学事件的定义来构建知识图谱所需的元数据模型，基于科学事件的知识图谱元数据模型，本文简称为科学事件元数据模型，科学事件元数据模型的框架设计图如图 4 所示。

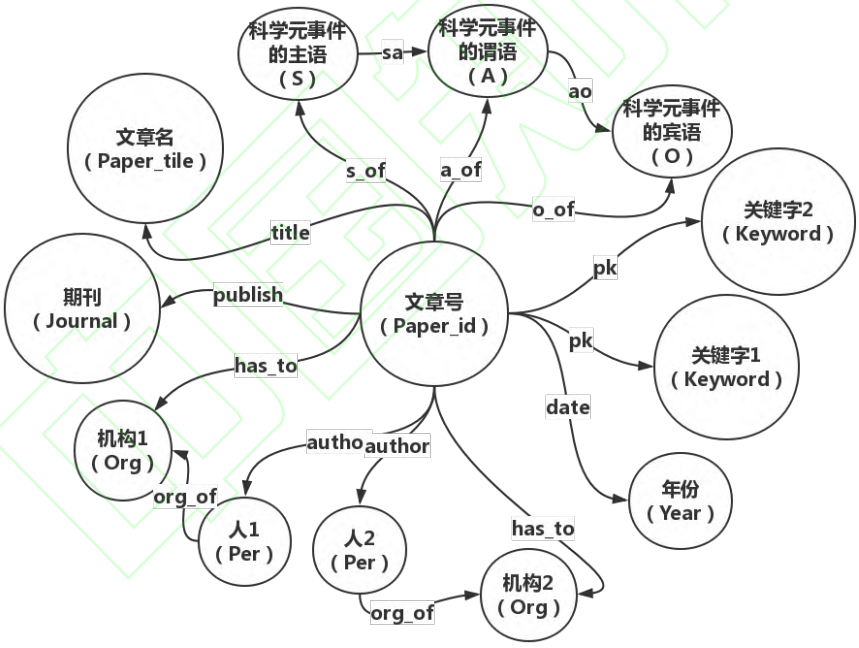

图 4　科学事件元数据模型的框架设计图

2.3.1 科学事件元数据模型的实体　**实体（Entity）是知识图谱的基本单元，也是文本中承载信息的重要语义单元，本文根据科学文献的外部特征以及科学事件的定义来确定构建基于科学事件的知识图谱所需要的实体，科学事件元数据模型实体的具体定义如表 1 所示。**

表 1　科学事件的实体

实体	中文名	描述
Paper_id	文章号	来源于科学文献的文章编号。是可精准查询该科学文献的唯一标识，具有唯一性
Paper_title	文章名	科学文献的文章名，科学事件的来源科学文献。可能会出现文章名重名的现象
Keyword	关键字	科学文献的关键字。可以通过关键字简单的了解到科学文献的研究内容、研究对象、研究方法的信息，可以帮助读者更好的了解科学事件的全貌
Year	年份	科学文献的公开发表年份，也是科学事件的发生年份。通过科学文献的公开发表，可以给科学事件提供时序信息
Per	人	科学文献的作者，也是科学事件的主要参与者。科学文献的作者通常会对应多个人
Org	机构	科学文献作者的所属机构，即该科学文献的所属机构，科学事件的主要发生地点。科学文献可能因其不同作者的机构不同，抑或是同一作者对应多个机构，同一科学文献会出现对应多个机构的现象
Journal	期刊	科学文献的期刊，也是科学文献的来源。一篇科学文献只对应一个期刊
S	科学元事件的主语	科学元事件的主语。是科学元事件的主要组成部分之一，可能会出现该科学元事件的主语是其他科学元事件的宾语的现象
A	科学元事件的谓语	科学元事件的谓语，即科学元事件的触发词。是科学元事件的主要组成部分之一，是识别科学元事件的主要根据
O	科学元事件的宾语	科学元事件的宾语。是科学元事件的主要组成部分之一，可能会出现该科学元事件的宾语是其他科学元事件主语的现象

2.3.2 科学事件元数据的关系 **关系（Relation）是知识图谱的重要组成部分，是知识图谱中节点与节点间的连线，通过关系我们可以将实体与实体相连，并根据关系进行推理查询，表 2 是科学事件元数据模型的实体间关系的具体定义。**

表 2　科学事件的关系

关系	中文名	描述
title	名称关系	存在于文章号（Paper_id）与文章名（Paper_title）之间，链接箭头方向为由文章号（Paper_id）指向文章名（Paper_title）
author	作者关系	存在于文章号（Paper_id）与人（Per）之间，链接箭头由文章号（Paper_id）指向人（Per）
org_of	隶属机构关系	存在于人（Per）与机构（Org）之间，链接箭头由人（Per）指向机构（Org）
has_to	属于关系	存在于文章号（Paper_id）与机构（Org）之间，链接箭头由文章号（Paper_id）指向机构（Org）
date	时间关系	存在于文章号（Paper_id）与年份（Year）之间，链接箭头由文章号（Paper_id）指向年份（Year）
publish	出版关系	存在与文章号（Paper_id）与期刊（Journal）之间，链接箭头由文章号（Paper_id）指向期刊（Journal）
pk	研究领域关系	存在于文章号（Paper_id）与关键字（Keyword）之间，链接箭头由文章号（Paper_id）指向关键字（Keyword）
s_of	主语关系	存在于文章号（Paper_id）与科学元事件的主语（S）之间，链接箭头由文章号（Paper_id）指向科学元事件的主语（S）
a_of	谓语关系	存在于文章号（Paper_id）与科学元事件的谓语（A）之间，链接箭头由文章号（Paper_id）指向科学元事件的谓语（A）
o_of	宾语关系	存在于文章号（Paper_id）与科学元事件的宾语（O）之间，链接箭头由文章号（Paper_id）指向科学元事件的宾语（O）
sa	主谓关系	存在于科学元事件的主语（S）与科学元事件的谓语（A）之间，链接箭头由科学元事件的主语（S）指向科学元事件的谓语（A），表示了科学元事件的动态流动方向
ao	谓宾关系	存在于科学元事件的谓语（A）与科学元事件的宾语（O）之间，链接箭头由科学元事件的谓语（A）指向科学元事件的宾语（O），表示了科学元事件的动态流动方向

3 实证研究

为了有效的从科学文献中抽取出科学事件，本文将以科学文献的摘要为实验文本，进行科学事件的抽取研究。摘要又称概要、内容提要。摘要是以提供文献内容梗概为目的，不加评论和补充解释，简明、确

切地记述文献重要内容的短文。其基本要素包括研究目的、方法、结果和结论。具体地讲就是研究工作的主要对象和范围，采用的手段和方法，得出的结果和重要的结论，即摘要蕴含了整篇科学文献的科学元事件的信息[18-20]。在科学事件的抽取方法中，因科学事件是属于世界 3 客观知识世界的事件，跟世界 1 物质客观世界是有区别的，所以在事件的研究思路上，本文选择利用外部图书情报领域词典和语义分析来识别科学事件的触发词和科学事件的事件论元，所要抽取出的目标科学元事件如图 5 所示。

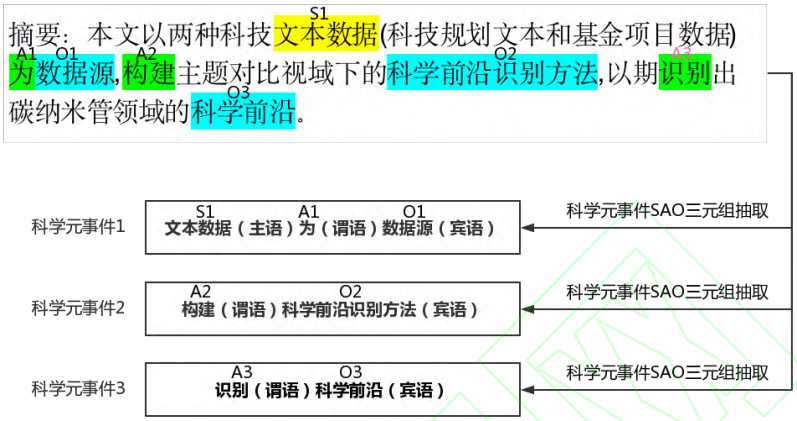

图 5 科学事件的自动抽取示例

3.1 研究思路

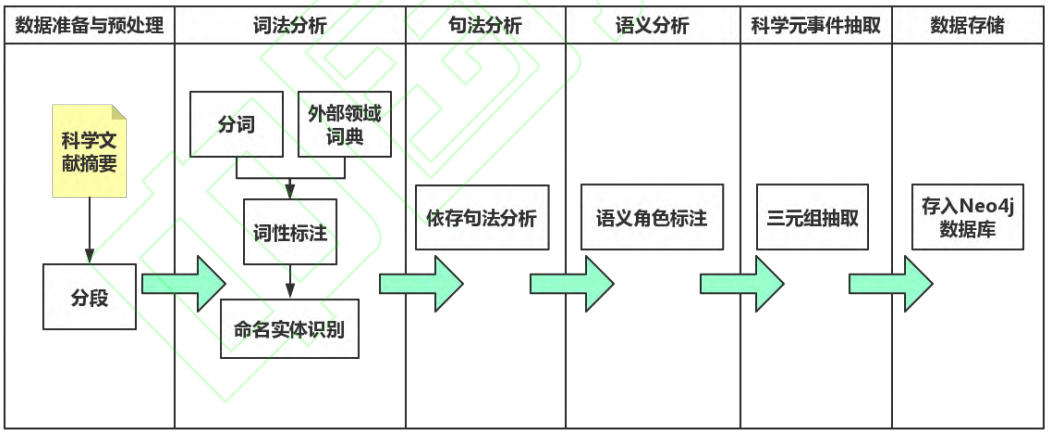

图 6 科学事件知识图谱构建研究路线图

具体步骤如下（见图 6）：

第一步：数据准备与预处理。从知网检索图情领域的 CSSCI 核心期刊论文数据，并将其外部特征与摘要下载并存储，对所获得的论文数据做一下简单的统计，并对摘要进行分段处理，为以下的实验做好数据基础。

第二步：词法分析。外部的领域词典对分好段的摘要文本数据进行分词处理，利用外部领域词典可以防止领域词汇的丢失，调高分词效果，并对分好词的文本数据进行词性标注，利用对词性标注完成的文本数据进行命名实体识别，以期高准确度的识别出文本数据所含的领域实体。

第三步：句法分析。对上一步处理好的数据进行依存句法树的构建，并对其进行依存句法的分析，以

期得出科学文献文本中的识别出的命名实体在句子中的句法成分，为下面对实体的语义标注做准备。

第四步：语义分析。对依存句法分析好的文本数据进行语义标注，进行句子级的语义分析，识别出科学文献中句子里的科学事件论元。

第五步：科学事件抽取。对进行完成语义分析的文本数据结构，对科学文献文本中识别出的主语、谓语、宾语进行抽取，并按照 SAO 三元组的格式进行存储表示，构建科学元事件。

第六步：数据存储。将抽取出的 SAO 三元组进行数据对齐，转换其格式，存入图数据库 Neo4j。

3.2 实验

3.2.1 实验环境与数据集

1）硬件。ubuntu 16.04 系统（64 位），Intel（R）Core i5-3360M CPU @ 2.8 GHz，4.0G，500G Solid State Drive。

2）配置环境。Python 3.6.8，Neo4j，java 1.8.0_181，Apache POI，pyltp，jieba。

3）数据集。本文的研究数据源为图书情报领域的 20 种中文社会科学引文索引（Chinese Social Science Citation Index，CSSCI）的核心期刊的科学文献数据，CSSCI 在我国人文社会科学评价领域具有权威地位，选用 CSSCI 所收录的全部图书情报领域的核心期刊的科学文献作为数据源，并对其进行科学事件的识别，能更好的、更全面、更具权威性的揭示我国图书情报领域的科学发展脉络，以期为相关科研工作者提供助力。

检索数据库：维普中文期刊全文数据库；数据获得方式：通过与维普公司合作的方式获得所需数据源；时间跨度：2008 年 1 月 1 日至 2017 年 11 月 23 日；检索结果：46476 篇；检索日期：2017 年 11 月 23 日。

3.2.2 数据预处理 **通过对实验数据的摘要文本内容进行提取，并进行文本格式的转换，使用 python 引用 pyltp 的 Segmentor 同时引用 ltp 语言云发布的 cws.model 模型实现对科学文献的摘要进行句子分段处理，将一段文本中的多个句子分隔。**

以科学文献《数字图书馆员教育：LIS 课程面向国际化》的摘要原文为例，所得到的句子文本分段结果，如图 7 所示。

1. 政治、经济、社会、教育与科技环境的变化对LIS专业人员工作的众多领域产生冲击。
2. 在世界的许多角落，数字化信息的创造、储存与扩散无日不在进行，因而，数字图书馆在这个虚拟信息环境中扮演着举足轻重的角色
3. 许多国家都已开展对数字图书馆员的教育，尽管方法不同，但面临的问题类似。
4. 在数字图书馆员教育的众多问题中，已经显现并需要考虑和调查的有：多学科尺度在数字图书馆员教育中的重要性；图书馆、档案
5. 由IFLA教育与培训部，以及其他国际项目所做的工作，也将在文中述及。

图 7 句子文本分段结果

3.2.3 词法分析 首先利用 python 的 jieba 分词的 HMM 隐形马尔科夫模型模式，并引入提前构建好的外部图情领域专有词词典对所得的科学文献的句子进行分词处理。

通过引用外部领域词典的 jieba 分词的 HMM 隐形马尔科夫模型的分词，并将分词所得结果通过使用 python 引用 pyltp 的 Postagger 和 LTP 平台已经训练好的 pos.model 模型对其实现词性的标注，词性标注的结果如图 8 所示。

政治/n 、/w 经济/n 、/w 社会/n 、/w 教育/v 与/c 科技/n 环境/n 的/u 变化/v 对/a LIS/ws 专业/n 人员/n 工作/n 的/
多/a 领域/n 产生/v 冲击/v 。/w
在/p 世界/n 的/u 许多/a 角落/n ,/w 数字化/v 信息/n 的/u 创造/v 、/w 储存/v 与/c 扩散/v 无日/nt 不/d 在/p 进行/
因而/c ,/w 数字/n 图书馆/n 在/p 这个/r 虚拟/v 信息/n 环境/n 中/nd 扮演/v 着/u 举足轻重/i 的/u 角色/n 。/w
许多/a 国家/n 都/d 已/d 开展/v 对/a 数字/n 图书馆/n 员/n 的/u 教育/v ,/w 尽管/d 方法/n 不同/a ,/w 但/c 面临/v
问题/n 类似/v 。/w 在/p 数字/n 图书馆/n 员/n 教育/v 的/u 众多/a 问题/n 中/nd ,/w 已经/d 显现/v 并/c 需要/v 考
和/c 调查/v 的/u 有/v ：/w 多/a 学科/n 尺度/n 在/p 数字/n 图书馆/n 员/n 教育/v 中/nd 的/u 重要性/n ;/w 图书馆/
档案馆/n 、/w 博物馆/n （/w LAM/ws ）/w 等/v 文化/n 机构/n 的/u 日渐/d 融合/v ;/w 新/a 模式/n 的/u 教/v 与/c
,/w 尤其/d 是/vl 新/a 教学法/n 和/c 电子/n 学习/v ;/w 甚至/d 需要/v 再行/v 考虑/v 信息/n 专业/n 人员/n 的/u 新

图 8 词性标注结果

对词性标注的结果通过使用 python 引用 pyltp 资源包的 NamedEntityRecognizer 和 LTP 平台已经训练好的 ner.model 模型，根据词性的标注实现命名实体识别，命名实体识别结果如图 9 所示。

角色。 许多国家都已开展对 数字图书馆 员的教育,尽管方法不同,但面临的问题类似。在

数字图书馆 员教育的众多问题中,已经显现并需要考虑和调查的有：多学科尺度在 数字图书馆 员教

图 9 命名实体识别结果

3.2.4 句法分析 对实体名识别的结果通过使用 python 引用 pyltp 资料包的 Parser 和 LIP 平台已经训练好的 parser.model 模型，根据命名实体识别的结果和词性分析的结果实现依存句法树的构建，并对结果进行句法分析，其所得结果如图 10 所示。

1 政治 政治 n n _ 12 定中关系 _ _
2 、 、 wp w _ 3 标点符号 _ _
3 经济 经济 n n _ 1 并列关系 _ _
4 、 、 wp w _ 5 标点符号 _ _
5 社会 社会 n n _ 1 并列关系 _ _
6 、 、 wp w _ 7 标点符号 _ _
7 教育 教育 v vn _ 1 并列关系 _ _
8 与 与 c c _ 9 左附加关系 _ _
9 科技 科技 n n _ 10 定中关系 _ _
10 环境 环境 n n _ 1 并列关系 _ _
11 的 的 u u _ 1 右附加关系 _ _
12 变化 变化 v vn _ 21 主谓关系 _ _
13 对 对 p p _ 17 状中结构 _ _
14 LIS LIS ws nx _ 16 定中关系 _ _
15 专业 专业 n n _ 16 定中关系 _ _

图 10 句法分析结果

3.2.5 语义分析与三元组抽取 **通过使用 python 引用 pyltp 资料包的 SementicRoleLabeller 和 LIP 平台已经训练好的 pisrl_win.model 模型，根据句法分析的结果实现语义角色的标注，构建语义依存关系，并对结果进行句法分析。**

将语义分析中的主语、谓语、宾语，通过正则表达式的方法，按照 SAO 三元组的数据结构（Subject-Action-Object）从语义角色标注的结果中抽取出来并进行存储，每个 SAO 三元组即是一个科学元事件。

根据举例的科学文献摘要，得到 5 个科学元事件：

1）科学元事件 SAO1=（政治变化工作众多领域，产生，冲击）；

2）科学元事件 SAO2=（数字化信息创造，在，进行）；

3）科学元事件 SAO3=（数字图书馆，扮演，举足轻重角色）；

4）科学元事件 SAO4=（新模式教育，是，新教学法）

5）科学元事件 SAO5=（以及，供，其实践）。

经过抽取所得 SAO 三元组 98726 组，部分结果如图 11 所示。

S	A	O
政治变化工	产生	冲击
数字化信息	在	进行
数字图书馆	扮演	举足轻重角色
许多国家	开展	对教育
新模式教	是	新教学法
信息学院	提供	全面教育
未来图书馆	具备	统计多方面知识
档案工作	得到	承认
没有	相应	影响力
档案工作者	来源于	工作价值内涵
其	进行	分析
档案工作具	界定	我们职业
	增加	档案工作
中国	出版	一批
广东通史	是	其中重要一部
一部	具备	权威特征
该文	结合	北大图书馆近期相关实践
读者	是	高校图书馆宝贵资源之一
教师读者	有着	挖掘巨大潜力
高校图书馆	突破	现有学科服务模式
各种资源	提高	学科服务水平
中国高等教	遵循	统一标准

图 11 科学元事件的抽取结果

3.2.6 科学事件构建 **科学事件是同源关系的科学元事件的集合，所以通过将科学元事件的 SAO 三元组可组织成科学事件。**

通过利用 Apache POI 的框架，结合正则表达式，将同源关系科学元事件进行合并，如将例子中的识别出的同源关系科学元事件 SAO1 至科学元事件 SAO5 进行合并，得到的结果为：$政治变化工作众多领域#产生#冲击$数字化信息创造#在#进行$数字图书馆#扮演#举足轻重角色$新模式教育#是#新教学法$以及#供#其实践$。其中$与$之间代表为一个科学元事件，$与#之间是科学元事件的主语，#与#之间为科学元事件的谓语，#与$之间是科学元事件的宾语。

通过组织科学元事件的主语、谓语、宾语得到的科学元事件识别结果共 98726 个，如图 12 所示。

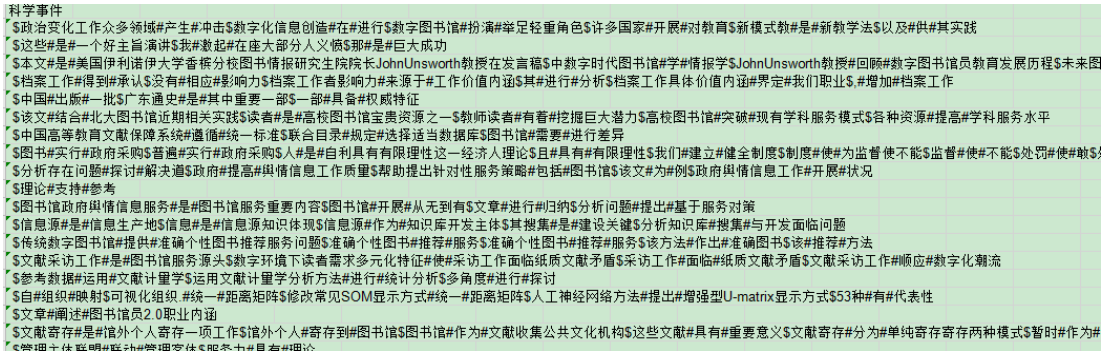

图 12 科学事件结果

3.2.7 数据存储 将科学文献的外部特征处理结果与科学元事件关系的对应结果利用 Neo4j 的 Cyber 语言的 load 功能导入到 Neo4j 图数据库中。

共得到科学元事件主语实体 S 共 153252 个节点，科学元事件谓语实体 A 共 153253 个节点，科学元事件宾语实体 O 共 154437 个节点。

根据本文所构建的科学事件元数据模型将节点与关系信息一一导入进 Neo4j 图形数据库中，得到了图情领域科学事件知识图谱，结果如图 13 所示。

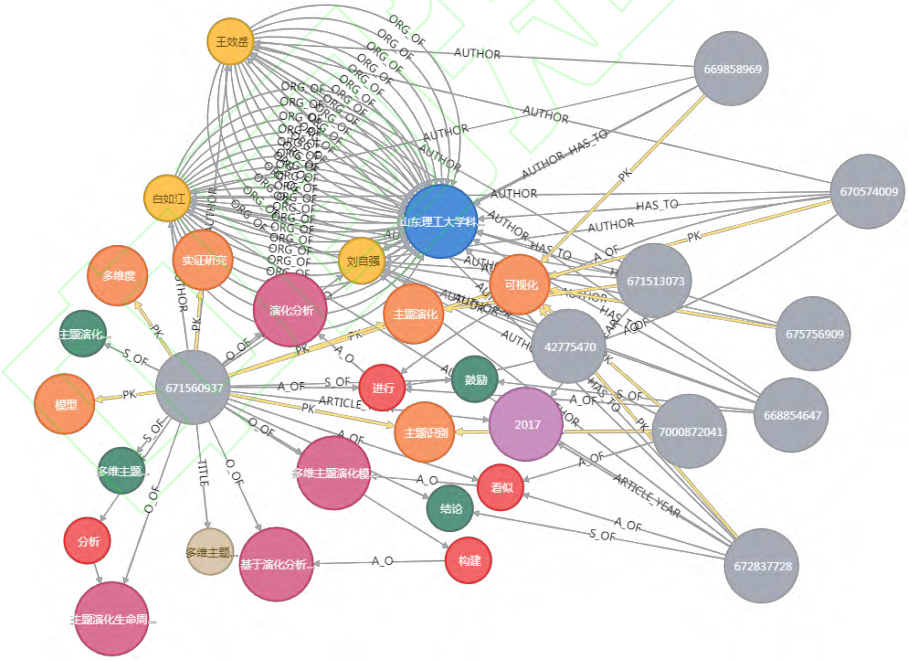

图 13 知识图谱导入数据结果（部分）

3.3 应用实例

图情领域科学事件知识图谱已经构建完成，作为一个知识图谱，可以对其进行查询，Neo4j 图数据库的查询语言为 Cypher，如需对节点与关系进行查询需用到 MATCH、WHERE 以及 RETURN 语句，在查询时还需要使用 LIMIT 语句限制显示的数据范围，因为数据库较大，限于笔者的硬件条件，不经限制数据范

围会使数据库瘫痪。

例 1：对作者为“刘自强”的部分文章进行查询其科学元事件，并限制数据大小为 10，得到结果如图 14 所示：

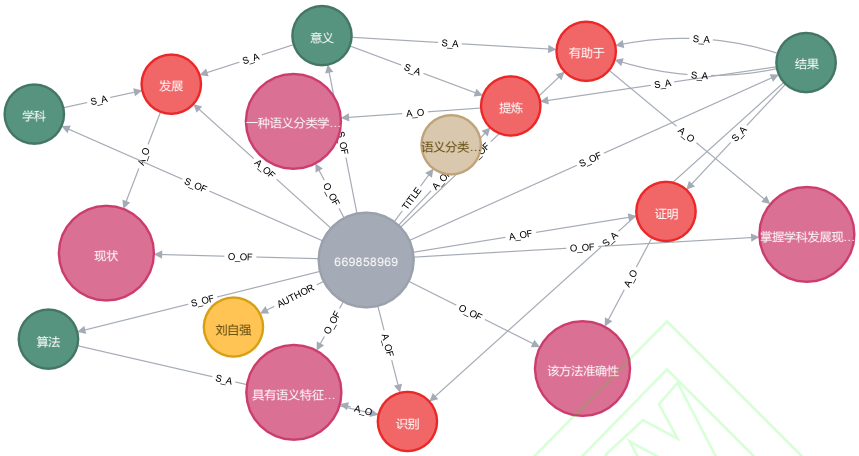

图 14 知识图谱科学元事件查询实例 1（部分）

可以得到文章号为 669858969，文章名为《语义分类的学科主题演化分析方法研究——以我国图书情报领域大数据研究为例》的 9 个同源科学元事件：

1）意义（S），提炼（A），一种语义分类学科主题演化分析方法（O）。

2）意义（S），有助于（A），掌握学科发展现状情况（O）。

3）意义（S），发展（A），现状（O）。

4）学科（S），发展（A），现状（O）。

5）算法（S），识别（A），具有语义特征社区（O）。

6）结果证明（S），证明（A），该方法准确性（O）。

7）结果（S），识别（A），具有语义特征社区（O）。

8）结果（S），提炼（A），一种语义分类学科主题演化分析方法（O）。

9）结果（S），有助于（A），掌握学科发展现状情况（O）。

例 2：查询如何能得到“LDA”，即查询科学元事件宾语为“LDA”的科学元事件，限制数据大小为 10，得到的结果如图 15 所示：

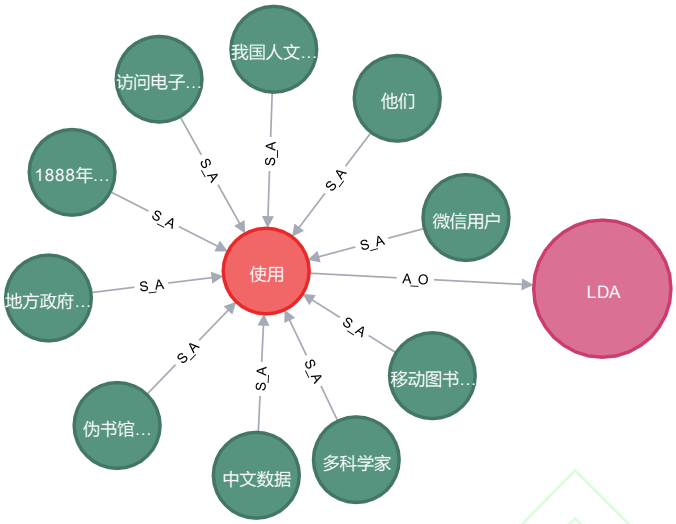

图 15 知识图谱科学元事件查询实例 2（部分）

LDA 为一种聚类模型，在图情领域科学事件知识图谱中检索科学元事件宾语为“LDA”的科学元事件，限制了数据量为 10，可以得出 10 个科学元事件：

1）中文数据（S），使用（A），LDA（O）。

2）多科学家（S），使用（A），LDA（O）。

3）1888 年傅云龙（S），使用（A），LDA（O）。

4）微信用户（S），使用（A），LDA（O）。

5）移动图书馆用户（S），使用（A），LDA（O）。

6）他们（S），使用（A），LDA（O）。

7）访问电子资源（S），使用（A），LDA（O）。

8）地方政府评估结果（S），使用（A），LDA（O）。

9）伪书馆藏（S），使用（A），LDA（0）。

10）我国人文社会科学研究者（S），使用（A），LDA（O）。

例 3：对例 2 进行扩展，查询可以使用“LDA”做些什么的语义信息，即查询科学元事件主语为“LDA”的科学元事件，限制数据大小为 10，得到的结果如图 16 所示：

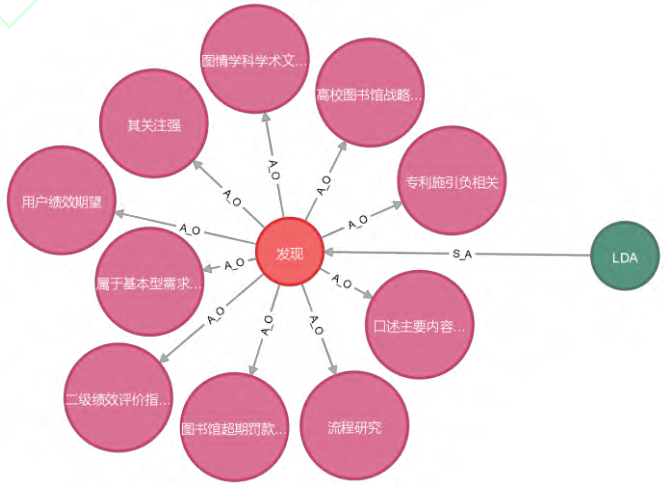

图 16 知识图谱科学元事件实例 3（部分）

通过检索，得到可以使用 LDA 发现什么的语义信息，从而与实例 2 组成事件链，以“LDA”为科学元事件的主语，限制了数据量为 10，可以得到以下 10 个科学元事件：

1）LDA（S），发现（A），图书馆超期罚款行为履行行为（O）。

2）LDA（S），发现（A），二级绩效评价指标中开放时间度（O）。

3）LDA（S），发现（A），属于基本型需求项目包括（O）。

4）LDA（S），发现（A），用户绩效期望（O）。

5）LDA（S），发现（A），口述主要内容规律（O）。

6）LDA（S），发现（A），流程研究（O）。

7）LDA（S），发现（A），专利施引负相关（O）。

8）LDA（S），发现（A），高校图书馆战略发展规划制定考量国家发展战略外部环境（O）。

9）LDA（S），发现（A），其关注强（O）。

10）LDA（S），发现（A），图书馆超期罚款行为履行行为（O）。

实例 2 与实例 3，LDA 同时可作为科学元事件的主语和科学元事件的宾语，科学元事件间首尾相接，便可以组成完整的科学事件链。

4 结束语

本文通过提出科学事件的概念，在现有科研元数据以及知识元框架的理论基础上，构建了科学事件元数据模型，将科学元事件划分为科学事件主语、科学事件谓语、科学事件宾语三个部分。本文从理论和实证两个方面对科学事件进行了解析，将所得数据存入图数据库 Neo4j 中并构建了以图情领域为例的科学事件知识图谱，证实了本文所提方法可行有效。

由于篇幅有限，本文仅选取图情领域中的部分论文进行实证分析。后续将选取大样本实例对科学事件元数据模型的特征和结构进行摸索，从而进一步验证本文所提的研究设想。□

参考文献

[1] 习近平. 在中国科学院第十七次院士大会、中国工程院第十二次院士大会上的讲话[J]. 当代劳模, 2014(6):14-17.

[2] 索传军, 盖双双. 单篇学术论文的评价本质、问题及新视角分析[J]. 情报杂志, 2018, 37(6):106-111

[3] 白如江, 冷伏海, 廖君华. 一种基于科技规划文本的研究前沿主题地图构建方法[J]. 图书情报工作, 2017(61):121.

[4] 白如江, 冷伏海, 廖君华. 科学研究前沿探测主要方法比较与发展趋势研究[J]. 情报理论与实践, 2017(5):37-42.

[5] BORNMANN L, RÜDIGER M. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references[J]. Journal of the Association for Information Science and Technology, 2015, 66(11):2215-2222.

[6] LAL D , LAL R . The Rise of Open Access[J]. Indian Journal of Microbiology, 2011, 51(4):416-417.

[7] 李鲲, 姚长青, 张均胜. 一种基于文献的科研事件库构建方法[J]. 情报理论与实践,

2017(9):133-138,143.
[8] 姚艳玲, 开滨. 基于知识图谱的智能信息处理领域可视化分析[J]. 科技通报, 2017(6).
[9] 师宏睿. 关于波普尔三个世界理论的信息学阐释[J]. 图书馆理论与实践, 2002(2):35-36.
[10] 刘迅. 论图书馆学情报学理论的共同基础——关于波普尔世界 3 理论的思考[J]. 情报科学, 1982(1):13-20.
[11] 杜汝楫. “三个世界”的学说--波普哲学介绍之一[J]. 哲学动态, 1982(6):33-37.
[12] 袁曦临. 图书馆学、情报学理论创新中的知识论研究[J]. 情报资料工作, 2006(3):8-11.
[13] 闻凤兰. 波普尔的第三世界理论及其建构原因[J]. 内蒙古民族大学学报：社会科学版, 2008, 34(4):61-64.
[14] 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.
[15] TAN Z, LIU C, MAO Y, et al. AceMap: A Novel Approach towards Displaying Relationship among Academic Literatures[C]// International Conference Companion on World Wide Web. 2016.
[16] SINHA A, SHEN Z, SONG Y, et al. An Overview of Microsoft Academic Service (MAS) and Applications[C]// the 24th International Conference. ACM, 2015.
[17] TANG J, ZHANG J, YAO L, et al. ArnetMiner: extraction and mining of academic social networks[C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008.
[18] 黎新伍. 学士学位论文英文摘要的写作体会[J]. 科技信息：学术版, 2006(11):350.
[19] 张清. 解析科技论文的撰写范式要求[J]. 价值工程, 2012, 31(14):319-321.
[20] 关于论著文章的中英文摘要的书写要求[J]. 临床小儿外科杂志, 2006(2):391-391.

作者简介：白如江（ORCID:0000-0003-3822-8484），男，1979 年生，博士，研究馆员。研究方向：知识组织。**周彦廷**（ORCID:0000-0001-7624-2637，通讯作者），男，1994 年生，博士生。研究方向：知识组织。**王效岳**（ORCID:0000-0002-7100-7758），男，1961 年生，博士，教授。研究方向：数据挖掘与信息处理技术。**王志民**，男，1949 年生，博士，教授。研究方向：数字人文。
作者贡献声明：白如江，拟定研究命题和思路设计。**周彦廷**，负责数据收集分析和论文撰写。**王效岳**，论文框架确定。**王志民**，论文细节修改。
录用日期：2020-03-10