doi:10.3772/j.issn.2095-915x.2021.06.005

标准文献知识图谱构建的模型设计与集成方法

开放科学（资源服务）标识码（OSID）

赵伟 张览 望俊成

中国科学技术信息研究所 北京 100038

摘要：[目的 / 意义] 随着大数据的迅速发展，面向标准领域的知识服务已经成为当前大数据背景下标准信息化发展的前沿热点问题，基于知识图谱技术开展标准文献资源服务，对于揭示标准知识的整体关联性，让标准数据发挥更大的效用，具有重要的研究意义。[方法 / 过程] 本文在解析标准文献特征及内容知识结构的基础上，提出了用于描述标准知识的标准知识单元五元组，建立了面向标准文献的资源层—描述层—映射层的三层知识图谱构建模型。进一步提出了基于规则和基于 LDA 主题模型的标准文献知识图谱构建的集成式方法。[结果 / 结论] 通过建立起标准文献的知识图谱，有助于扩展实体关系类型，并以期为后续实证研究提供理论和方法支撑。

关键词：标准文献；标准知识；知识图谱；知识抽取

中图分类号：G35

Model Design and Integrated Method for the Construction of Standard Literature Knowledge Graph

ZHAO Wei ZHANG Lan WANG Juncheng

Institute of Scientific and Technical Information of China, Beijing 100038, China

Abstract: [Purpose / Significance] With the rapid development of big data, knowledge service in the field of standards has become a frontier hot issue in the development of standard informatization. Carrying out standard literature resource service based on knowledge atlas technology has important research significance for revealing the overall relevance of standard knowledge and

基金项目 中国科学技术信息研究所重点工作项目“金融大数据建设与知识服务（二期）——金融科技知识图谱构建”（ZD2020-03）。

作者简介 赵伟（1975-），博士，研究员，研究方向为科技资源管理，Email：zhaowei@istic.ac.cn；张览（1996-），硕士研究生，研究方向为产业竞争情报；望俊成（1984-），博士，副研究员，研究方向为科技政策与科技管理、文本数据可视化、大数据治理。

引用格式 赵伟 , 张览 , 望俊成 . 标准文献知识图谱构建的模型设计与集成方法 [J]. 情报工程 , 2021, 7(6): 58-66.

making standard data more effective. [Method / Process] On the basis of analyzing the characteristics and knowledge structure of standard documents, this paper puts forward the five tuples of standard knowledge units for describing standard knowledge. A three-layer knowledge map construction model of resource layer description-layer mapping-layer for standard documents is proposed. Further, an integrated method for the construction of standard literature knowledge map based on rules and LDA topic model is proposed。[Results /Conclusions]Through the construction knowledge map of standard literature, can help to extend entity relationship types,andthat will provide theoretical and methodological support for subsequent empirical research.
Keywords: Standard literature; standard knowledge; knowledge graph; knowledge extraction

引言

目前，随着大数据的迅速发展，知识图谱技术已成为科技文献资源服务的重要创新手段。近年来不少机构和学者投入到知识图谱研究中，借助其强大的语义处理能力将领域知识有序地组织起来，揭示知识的整体关联性，科技文献知识图谱得到了越来越多的重视[1]。然而，由于数据覆盖不全、精确度不高、受概念范围的影响等原因[2]，很难规范科技文献的实体及其关系，而且科技文献还包括了前瞻性预测和猜想，甚至包括一些由于同行评议把关不严或认识不到位而产生的错误理解和阐释，导致可视化结果与客观事实不符，这也是知识图谱技术在这些年无法在科技文献领域获得令人满意的应用的重要原因之一。

标准文献作为十大科技文献资源之一，蕴含着丰富的科技知识，既是标准的重要载体和表现形式，也是科研人员收集标准情报的主要来源[3]。其具有科技文献的共性特点，还与其他科技文献存在显著不同，即标准文献是基于具有可操作性的最低限制性要求而形成的，是真实可靠的。标准文献知识图谱属于典型的领域知识图谱，在规范且丰富的数据基础上，建立起一套通用型规则抽取体系和可视化模型，实现全覆盖和高质量的标准领域知识图谱，对于揭示标准知识的整体关联性，为标准研究人员提供有组织的标准文献知识集合，让标准数据发挥更大的效用，具有重要的研究意义。

总体上，国外学者们围绕标准文献知识图谱开展的研究并不多见，我国对标准文献的相关研究陆续有了试验探索。在早期的科学知识图谱研究中，以标准文献网络的结构特征为研究目标，多以引用关系为核心，以文献计量学、社会网络分析方法以及聚类等为研究方法进行分析[4-8]。这一阶段的研究以建立标准间的链接网络为主，尚未达到标准内容知识元的粒度。随着数据挖掘技术的运用和数字化的标准信息获取手段的加强，学者们逐渐向标准知识关联的领域深入研究。郭德华[9]指出应根据标准文献的知识关联关系开展支持知识关联检索、动态跟踪等功能的标准文献知识链接服务。甘克勤等[10]基于K均值聚类算法、模糊C均值聚类算法等文本聚类方法应用于标准文献题录数据并进行聚类试验，结果表明在核心词汇抽取的准确性上效果较好，但在分词和聚类的准确

doi:10.3772/j.issn.2095-915x.2021.06.005

性上需进一步提升。语义网技术的推广促进了知识服务的发展[11-12]，在此浪潮下，潘薇[13]、甘克勤[14]、梁薇[15]、李抵非等[16]进一步围绕语义网环境下的标准知识关联的理论和方法开展了积极探索。然而总体上，标准文献分析和知识关联等的相关研究主要体现在基于传统情报学、科学计量学的应用，在内容挖掘和知识图谱的构建应用方面少有涉足。目前标准文献的研究还存在一些制约因素，如标准文献的非结构化数据格式处理困难、标准文献的信息组织模式不清晰和微观分析方法与宏观分析方法应用不协调等。已有标准间关系的抽取深度和广度还不够，并在很大程度上影响着知识图谱构建最终的效果。因此，有待对标准文献的要素及其知识抽取方面开展更多探索。

1 标准文献的基本概念与特点

在当今知识经济时代，标准反映了该国的经济、技术和生产水平，其重要性日益凸显。标准文献作为标准的重要信息载体和表现形式，概念分为狭义和广义两种。狭义的标准文献是指由技术标准、管理标准、工作标准及其他具有标准性质的规范性文件所组成的一种特定形式的科技文献体系，简称标准；广义的标准文献指与标准化活动有关的所有文献，除了狭义概念下的各类标准外，还包括标准分类资料、标准检索工具、标准化期刊、标准化专著、标准化手册、定制图册等其他出版物。

标准文献作为一种特殊的文献，除具有一般科技文献的属性和作用外，其自身在结构、形式、内容、制定及适用范围等方面均具有独特而明显的特点[17]，具体表现在：（1）具有法律约束力。标准是参与生产工作、管理、设计制造的准入门槛和遵守依据，标准化法明确规定必须执行强制性标准，鼓励自愿采用推荐性标准。（2）具有统一的产生过程和专门的编写格式。国家设立了标准制修订的流程规范，专门规定了标准文献的编排格式，并设有固定的代号。（3）具有时效性。标准文献通常情况下代表了底线和门槛，起到准入作用，其目的是确保规格或安全。随着经济发展、标准化对象的变化和科学技术水平的提高，标准文献也要不断更新换代，因而产生了废止无效的标准文献。（4）具有明确的适用范围和用途。标准文献的“范围”结构概括了该篇标准的适用范围和不适用范围，简明扼要地说明了标准化对象和要解决的问题。（5）不同种类和级别的标准在不同范围内贯彻执行。

2 标准文献的结构解析

标准知识元和知识关联模式是识别、研究和应用标准知识的基本出发点。构建标准文献知识图谱，需要对标准文献的组成要素、层次和知识关联逻辑进行分析，进而确定标准文献文本特征的抽取任务和模型选择。因此，标准文献的结构解析是采用知识图谱对其进行表达的基础。

2.1 标准文献的结构

标准要素是组成标准文献的基本单元，标准文献的内容都是由各种要素构成的。根据GB/T1.1—2020《标准化工作导则第1部分：标

准化文件的结构和起草规则》，标准要素的划分有 3 种方式。依据要素的性质，可将标准中的要素划分为“规范性要素”和“资料性要素”；依据要素在标准中所处的位置，标准要素可划分四类：“规范性一般要素”“规范性技术要素”和“资料性概述要素”“资料性补充要素”[18] 如表 1 所示。

表 1　标准的要素

按要素性质划分	按要素在标准中所处的位置划分	要素的编排	要素的状态
规范性要素	规范性一般要素	标准名称	必备要素
		范围	必备要素
		规范性引用文件	可选要素
	规范性技术要素	术语和定义	可选要素
		要求	可选要素
		需考虑的因素	可选要素
		规范性附录	可选要素
资料性要素	资料性概述要素	封面	必备要素
		目次	可选要素
		前言	必备要素
		引言	可选要素
	资料性补充要素	资料性附录	可选要素
		参考文献	可选要素
		索引	可选要素

规范性要素不一定是必备要素，资料性要素也可能是可选要素，这几个概念间具有交叉关系。要实现标准文献知识图谱，数据内容必须覆盖所有必备要素、规范性技术要素、部分规范性一般要素和部分资料性规范性要素。因此，标准文献实体应在标准封面、前言、范围、规范性引用文件等部分进行抽取。

标准的层次划分和设置采用部分、章、条、段、列项和附录的形式[18]，如表 2 所示，对任何一份标准来说，其编排都至少要有章、条、段三个层次，其编排方式为层层嵌套。

表 2　标准的层次

层次名	编号（以GB/T1.1—2020为例）	标题
部分	第1部分	标准化文件的结构和起草规则
章	6	文件名称和结构
条	6.1	文件名称
条	6.1.1	通则
段	无编号	无
列项	a) b) c)	无
条	6.1.2	可选元素的选择
条	6.1.2.1	引导元素
……	……	……
附录		附录A
……	……	……

2.2　标准文献的知识关联关系

知识具有关联属性，标准文献之间的知识关联是各标准文献知识元之间存在的各种关系的总和。研究标准文献的关联，可以使各知识元形成系统的知识关联网络，发现其潜在的逻辑关系[18]，有助于加强对标准文献的利用，便于使用者快速准确地获取技术标准中的技术知识。

知识按可被直接获取和理解的程度可划分为显性知识及隐性知识。标准文献的显性知识又可分为直接关系和间接关系，直接关系指两

doi:10.3772/j.issn.2095-915x.2021.06.005

份标准之间通过一条通路即可连接的关系，通常包含引用关系、采用关系、修改关系和代替关系。其中引用关系类似于学术论文中的引用关系，可以解释标准技术发展的脉络；采用关系是标准文献特有的关系，它反映了一国标准在其他国家和地区标准化领域的影响力。间接关系则指两份标准之间通过一个或多个连接点建立的关系，这个连接点通常表现为归口单位、提出单位、起草单位和起草人等。

除了诸如相互引用和采用之类的显性关联外，标准文献还可以通过主题内容关系链接在一起，以形成不易直接发现的隐性关联。标准中的“范围”部分可视为专利和论文的“摘要”部分，规定了该标准的适用范围和标准化对象，这一篇章结构蕴藏着丰富的标准信息，可视为多个关键词的集合，同一领域内的标准文献可能共同对同一主题下的某些关键词进行规范约束，跨领域之间的标准文献也可能从该领域所属的维度分别对某一项标准化对象设立标准规范。通过分析和发现标准文献之间的隐性关联，可以获得大量潜在的隐藏知识，从而使标准文献可以创造更大的价值。

3 标准文献知识图谱 RDM 模型

标准文献知识效用的最大化取决于从整体上对相互关联的各类标准知识进行系统、灵活的应用。因此首要解决的是获取多个标准文献知识之间的关系，其次将其分解为知识单元并进行知识链接，最后用知识图谱进行表达。

本文从细粒度知识单元视角出发，通过对标准文献逻辑结构的分析，提出了用于描述标准文献知识的知识单元五元组（E, A, R, T, H），并结合科技文献的资源语义空间的描述 [19] 提出了标准文献知识图谱 RDM 模型（Resource-Description-Mapping，资源—描述—映射），模型如图 1 所示。该模型由文献资源、知识单元描述和知识单元映射三个层面支撑并连接起整个标准文献的知识图谱。正如前文所述，知识图谱通过三元组（实体—属性—属性值）和实体—关系—实体的形式表达知识。这一方式同样适用于标准文献的知识表示，标准文献的实体可从显性特征和隐性特征两方面表示，显性标准实体即为常规的易于判别的实体，这些实体的确定依赖于标准文献严格划一、有规律的描述风格；隐性标准实体指的是通过自然语言处理，将隐性知识显性化得到的那些实体，而显性标准实体和隐性标准实体的分布又依赖于标准文献严格的编排体系。

因此，本文提出的标准知识单元五元组与知识图谱三元组存在重合的要素，但是通过知识单元五元组进行描述是为了更准确、更全面地表示实体关系，最终的知识表示方式仍然符合知识图谱的三元组的逻辑。首先，标准文献的内容由多个知识单元组成，多个知识单元之间的不同组配方式又能反过来表示标准文献要素；其次，利用人工或半人工、自动的抽取方法提取标准知识内容中所包含的主题知识；最后，将这些具有实体概念意义的标准知识单元通过知识链接的方式映射到知识图谱的实体关系表达上，这样就建立了多层次关联关系，形成标准文献知识图谱。

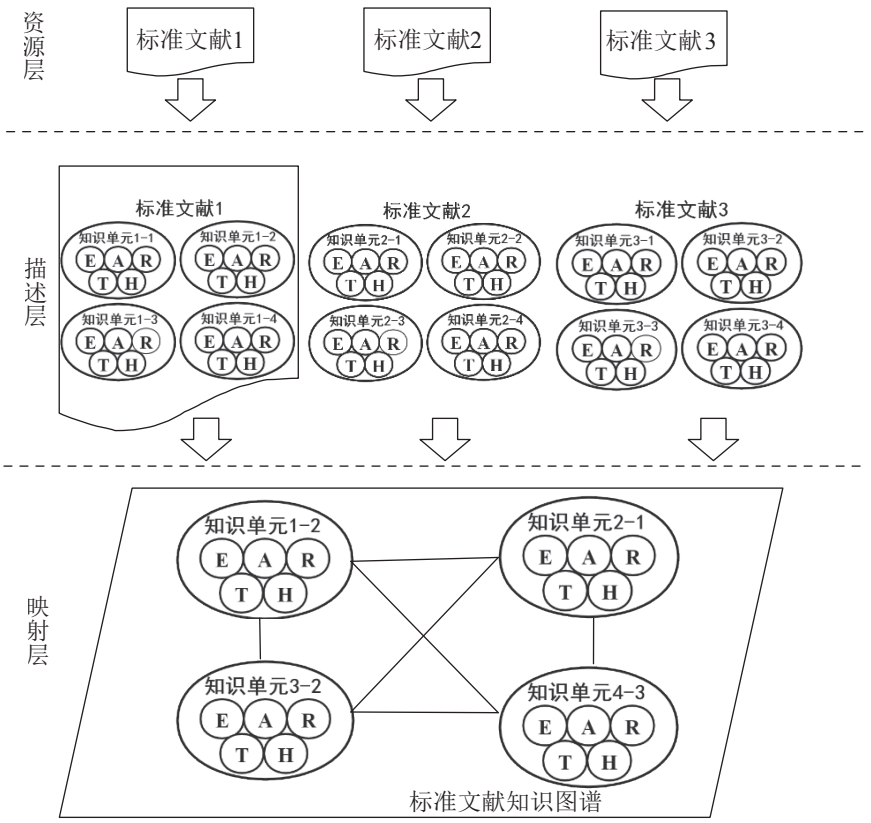

图 **1**　标准文献知识图谱 **RDM** 模型

标准文献知识图谱 RDM 模型由 *SLR*、*KuD* 和 *KuM* 三要素构成，模型可表示为：

$$RDM=(SLR, KuD, KuM) \quad (1)$$

第一层为标准文献资源层 *SLR*（Standard Literature Resource），表示该模型所组织的标准文献资源集合。通常按照标准文献主题内容所属学科和标准化专业领域进行组织，本研究依据国际标准分类法（ICS）和中国标准文献分类法（CCS）划分标准文献资源层级。

第二层为标准文献知识单元描述层 *KuD*（Description of Knowledge unit），表示从标准文献资源中抽取出的知识单元 *Ku*（Knowledge unit）经过描述与表示后所形成的集合。本研究使用标准文献知识单元五元组来描述标准文献的基本属性：

$$Ku=(E, A, R, T, H) \quad (2)$$

其中，*Ku* 代表标准文献知识单元；*E* 为标准知识单元的实体（Entity）；*A*（Attribute）为标准知识单元的属性集合，包括：标准编号、标准名称、发布时间等；*R*（Relation）为标准知识单元之间的关系，包括代替关系、引用关系以及主题关联关系等显性关系和隐性关系；*T*（Topic）为描述标准知识单元主题的主题概念集合；*H*（Hierarchy）为标准文献的要素层次结构，是不同知识单元在标准文献内的分布位置。

第三层为标准文献知识单元映射层 *KuM*（Mapping of Knowledge unit），表示若干个知

doi:10.3772/j.issn.2095-915x.2021.06.005

识单元所构建的知识图谱，表示为：

$$KuM=(Tc, Ec, Rtc, Rec)=$$
$$(Tc, Ec, (tci, tcj, rtc), (eci, tci, rec) \quad (3)$$

Tc 表示标准知识的主题概念集合，每一个节点代表着一个标准知识单元的主题概念，由多个关键词组成；*Ec* 表示标准文献实体集合，每一个节点代表一个标准知识单元的实体概念；*Rtc* 表示标准知识单元主题概念之间的关系集合，每一个语义关系可被描述为一个标准三元组 (*tci*, *tcj*, *rtc*)，*tci* 和 *tcj* 分别为两个标准主题概念，*rtc* 表示两个主题之间的关系；*Rec* 表示标准知识单元主题概念与实体之间的关系集合，同标准主题概念间关系一样，每一个关系可描述为一个标准三元组 (*eci*, *tci*, *rec*)，*eci* 为第 *i* 个标准实体，*tci* 为第 *i* 个标准知识单元的主题概念，*rec* 表示标准实体与主题概念之间的关系。

本研究设计的标准文献知识图谱 RDM 模型核心在于知识单元五元组的解构，将实体 *E* 之间的关联关系进行分解，以层次结构 *H* 为骨架，分为显性关联和隐性关联。其中显性关联主要，指通过实体属性 *A* 和 *R* 建立的知识单元之间的链接，隐性关联主要指通过主题 *T* 建立的知识单元之间的链接。如在“封面”部分，通常包含有属性，如标准代码、推荐等级、分类号等；在“前言”部分，通常包含实体，如采用标准、代替标准、提出单位、归口单位、起草单位以及采用方式属性；在“范围”部分，包含由关键词组成的主题要素；在“规范性引用”部分，包含引用文件实体及其代码属性，因此通过知识单元五元组可以比较充分的揭示标准文献知识。

4 标准文献知识图谱构建的集成方法

在前文所建立的 RDM 模型中，资源层和描述层可通过知识抽取来实现，描述层和映射层可通过知识链接的方式来实现，即 RDM 模型的建立过程主要包括标准文献的知识抽取和知识链接两个步骤，分别作用于不同的章节结构，从而建立起标准文献的知识图谱如图 2 所示。

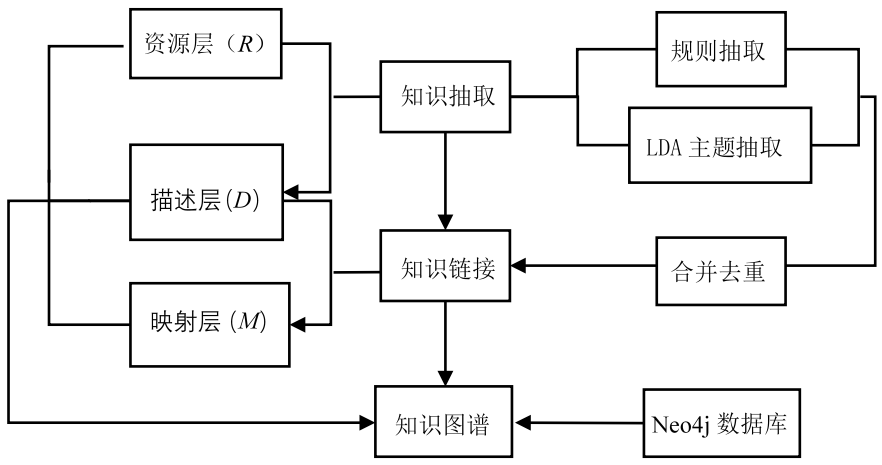

图 2 基于 **RDM** 模型的知识图谱构建的总体流程图

（1）标准文献知识抽取。根据不同知识单元在标准文献内所处的位置，分别利用规则抽取和 LDA 主题抽取方法，对标准的显性知识和隐性知识单元进行匹配和抽取，形成标准文献实体。

（2）标准文献知识链接。对标准文献知识单元之间相互链接所依靠的语义关系进行人工合并去重，再利用 Neo4j 数据库将标准知识单元的各知识单元转换到标准文献知识图谱中。

其中，如前文所述，在知识抽取的过程中，本文基于规则方法进行标准文献显性知识抽取，基于 LDA 主题模型进行隐性知识抽取。对于显性知识而言，由于其存在于标准封面和标准前言中，组织方式相对规范，因此适合用基于规则的方法进行信息抽取。对于隐性知识而言，其存在于标准文献正文中，相较于其他知识更加复杂，无法用规则抽取的方式得到合适的描述字段。而 LDA 是一种文档主题生成模型，可以用来分析一篇文档的若干主题分布，近几年在情报学领域中得到了广泛应用，包括用于科技文献文本分类、科学主题演化与科技文献相似度计算等研究领域。标准文献的初始文本是 PDF 格式的，属于非结构化数据，需要先对初始文本做格式转换，XML 格式的数据是半结构化，具有清晰的逻辑结构，便于后续操作，但要实现标准全文的三元组抽取还很困难。因此，可考虑将研究范围限定在前言部分、范围部分和规范性引用文件部分。从非结构化表示的标准文献中抽取出结构化的实体属性关系，并以三元组的形式存放到文件中，其研究成果可有助于标准知识库构建、标准搜索引擎和标准信息检索的实现。因此，本文提出采用 LDA 主题模型进行隐性知识的抽取，数据范围限定在“范围”部分。

通过上述集成方法可实现与 RDM 模型的良好映射，它适用于标准这一特定领域。规则抽取技术成熟，通过人工定义模板可以保证准确性，在垂直领域中表现良好；而 LDA 技术不同于专门用于知识抽取的技术，它是关键词、主题词抽取的主流技术，LDA 抽取的效果对标准主题的揭示更有说服力。因此，本文认为通过规则和 LDA 模型的知识抽取集成方法使基于标准文献挖掘得到的关系是有用且有效的。

5 结论

本文在解析标准文献特征及内容特点的基础上，分解了标准知识结构，构造了标准知识单元五元组 (E, A, R, T, H)，并基于五元组设计了标准文献知识图谱 RDM 模型，从资源层、描述层和映射层解构了标准知识图谱构建的理论模型。针对标准关联关系的识别与挖掘，补充和拓展标准文献知识关联的相关研究。除了一般性的题录关系，本研究进一步拓展的关系体现在两点，即四种不同程度的相互采用的知识关联关系和基于主题关联维度的多元关系。标准文献中“范围”部分可视为论文文献中的“摘要”，其中存在揭示主题内容的信息，通过提取主题词，挖掘标准实体和主题词汇之间的语义关联可以建立标准主题间的关系。

在此基础上，进一步提出构建 RDM 模型的集成方法，即 RDM 模型的建立过程主要包括标准文献的知识抽取和知识链接两个步骤，分别作用于不同的章节结构，从而建立起标准

doi:10.3772/j.issn.2095-915x.2021.06.005

文献的知识图谱，有助于扩展实体关系类型。

需要指出的是，上述标准文献知识图谱的构建研究仍处于初级阶段，缺少高质量语料库，未实现基于机器学习、深度学习等技术实现知识抽取。下一步可在现有研究的基础上结合机器学习方法，选择更加适合的实体关系抽取方法。

参 考 文 献

[1] 漆桂林 , 高桓 , 吴天星 . 知识图谱研究进展 [J]. 情报工程 , 2017, 3(1):4-25.

[2] 王颖 , 钱力 , 谢靖 , 等 . 科技大数据知识图谱构建模型与方法研究 [J]. 数据分析与知识发现 , 2019, 3(1):15-26.

[3] 张国华 . 标准文献是一种重要情报源 [J]. 中国标准化 , 2006(7):66-69.

[4] 李景 , 汪滨 , 周洁 , 等 . 我国标准制定的领域动态趋势分析——基于国家标准馆 2006—2007 年度国内外标准文献新到馆藏 [J]. 图书情报工作 , 2009, 53(1):56-60.

[5] 刘华 . 基于文献计量的国内外信息与文献国家标准对比研究 [J]. 图书情报工作 , 2011, 55(12):51-55.

[6] 王季云 , 王宇 . 基于社会网络分析法的标准网络测度指标研究——以五个主管部门国家标准网络比较为例 [J]. 中南财经政法大学学报 , 2016(5):21-29.

[7] 张君维 , 马志远 . 聚类方法在标准引用分析中的应用 [J]. 机械工业标准化与质量 , 2016, 459(6):40-42.

[8] 张正敏 , 郑东林 , 孙冬梅 . 基于引力模型的标准文献联系测度研究与应用 [J]. 标准科学 , 2019(547):93-96.

[9] 郭德华 . 标准文献知识链接服务模式研究 [J]. 图书情报工作 , 2011, 55(9):76-79.

[10] 甘克勤 , 丛超 , 张宝林 , 等 . 基于划分的文本聚类算法在标准文献中的试验与对比研究 [J]. 标准科学 , 2013(475):48-51.

[11] Berners-lee T, Hendler J, Lassila O. The semantic web[J]. Scientific American, 2001, 284(5):34-43.

[12] Noura M, Gyrard A, Heil S, et al. Automatic knowledge extraction to build semantic web of things applications[J]. IEEE Internet of Things Journal, 2019, 6(5):8447-8454.

[13] 潘薇 , 李景 , 马志远 . 语义网环境下的标准知识关联与服务探析 [J]. 标准科学 , 2014(9):34-36+44.

[14] 甘克勤 , 马志远 , 张明 . 标准文献关联可视化研究与实践 [J]. 标准科学 , 2015(1):34-38.

[15] 梁薇 , 马万钟 , 王文君 . 基于知识地图的标准知识可视化模型研究 [J]. 标准科学 , 2015(7):62-64.

[16] 李抵非 , 田地 , 胡雄伟 . 基于深度学习的中文标准文献语言模型 [J]. 吉林大学学报 , 2015, 45(2):596-599.

[17] 葛郁葱 . 标准文献的特点及其检索方法 [J]. 情报杂志 , 2009, 28(2):166-167.

[18] GB/T 1.1—2020, 标准化工作导则第 1 部分：标准化文件的结构和起草规则 [S]. 北京 : 中国标准出版社 , 2020.

[19] 李祯静 , 秦春秀 , 赵捧未 , 等 . 科技文献的资源语义空间：一种细粒度知识组织方法 [J]. 情报杂志 , 2019, 38(2):158-165+180.