标准知识数字化表达通用模型与自动抽取技术研究

马小雯[1] 孙红军[2*] 刘彦林[1] 甘克勤[2]

（1.之江实验室；2.中国标准化研究院）

摘 要：标准数字化是国内外标准化发展的重要领域和方向。研究以食品与农产品领域标准为研究对象开展标准知识数字化表达的通用模型与自动抽取技术研究，明确了国内外标准化数字化研究现状与问题，提出了标准知识数字化表达的通用模型，开展了标准知识数字化自动提取技术研究，实现对表达模型的知识要素的自动标注和抽取，并据此形成具有语义关联的标准知识。最后，以2000项食品与农产品领域标准为例进行标准知识数字化表达模型与自动提取技术的实证研究。

关键词：标准知识，数字化，通用模型，自动抽取，语义关联

DOI编码：10.3969/j.issn.1674-5698.2024.01.012

Research on Universal Model of Digital Representation of Standards Knowledge and Automatic Extraction Technology

MA Xiao-wen[1] SUN Hong-jun[2*] LIU Yan-lin[1] GAN Ke-qin[2]

（1. Zhejiang Lab; 2. China National Institute of Standardization）

Abstract: Standards digitization is an important field and direction of development at home and abroad. This research takes the standards in the field of food and agricultural products as the research object to carry out the research on the universal model of digital expression of standards knowledge and the automatic extraction technology of standards knowledge, defines the current situation and problems of standardization digitization research at home and abroad, puts forward the universal model of digital expression of standards knowledge, and carries out the research on digital automatic extraction technology of standards knowledge. The knowledge elements of the representation model are automatically labeled and extracted, and the standards knowledge with semantic association is formed accordingly. Finally, the digital representation model and automatic extraction technology of standards knowledge are studied by taking 2,000 food and agricultural product standards as examples.

Keywords: standards knowledge, digitization, universal model, automatic extraction, semantic association

---

基金项目：本文受之江实验室开放课题“面向图表识别的标准数字化知识提取标准研究”（项目编号：K2022NH0AB02）、中国标准化研究院院长基金项目“数字标准馆标准体系构建及关键标准研制与应用”（252023Y-10411）资助。

作者简介：马小雯，硕士，之江实验室智能科技标准化研究中心工程师，研究方向为标准数字化。

孙红军，通信作者，博士，副研究员，研究方向为标准数字化。

刘彦林，学士，高级工程师，研究方向为标准数字化、数据要素标准化。

甘克勤，硕士，中国标准化研究院国家标准馆副馆长，研究方向为标准数字化。

0 引言

以新一代信息技术为代表的新一轮科技革命和产业变革加速演进，经济、产业数字化转型成为时代趋势。标准作为经济活动和产业发展的技术支撑，以及国家基础性制度的重要方面，无论在深度还是在广度上都即将受到这一趋势的影响。《国家标准化发展纲要》指出，“发展机器可读标准、开源标准，推动标准化工作向数字化、网络化、智能化转型”。标准数字化转型已成为新时代我国重点产业发展的战略任务，对增强我国产业发展安全、参与全球市场竞争具有重要意义。

随着我国食品与农产品行业的迅速发展，企业规模不断增长，食品与农产品行业的安全形势比较严重，面临的挑战和竞争前所未有，同时暴露出的安全、健康、环境问题也愈来愈多，在新产品研制面临的对象、要求的技术条件、新工艺、新技术应用等方面的安全与环保问题日益突出。为进一步加快标准数字化转型步伐和有效解决食品与农产品领域安全与环保问题，本研究将以食品与农产品领域标准为研究对象开展标准知识数字化表达模型与自动提取技术研究，首先，明确国内外标准化数字化当前研究现状与问题；其次，通过文献和实地调研，提出标准知识数字化表达模型；再次，开展标准知识数字化自动提取技术研究，实现对表达模型的知识要素的自动标注和抽取，并据此形成具有语义关联的标准知识库；最后，以2000项食品与农产品领域标准为例进行标准知识数字化表达模型与自动提取技术的实证研究，以验证理论或技术的可行性。

1 国内外研究现状

有关标准知识数字化表达模型主要集中于以下3个方面。

（1）在图书文献领域，越来越多的信息研究机构正在推进语义解析，支持各种细粒度的知识单元关联与计算，不仅包括段落、表格、人物、机构，还包括知识点、概念等复杂本体关系的构建。并通过XML系列置标语言的描述和标记，与特定领域的各种知识相关联，支持可计算、可推理的智能检索与语义知识发现。国外已推出文献知识表达服务，将传统以文献为中心的搜索平台，转换为以事件为中心和RDF为基础的复合语义架构。许多国际信息研究机构已经在语义解析方面进行诸多实践，卓有成效。数字技术和数字环境在颠覆传统资源形态的同时，也在全面改造信息资源建设与服务模式。国家科技图书文献中心（NSTL）构建科技知识组织体系共享服务系统（STKOS），收录615,384个概念，2,321,681个术语，应用于NSTL数以亿计的外文期刊内容的本体揭示，形成NSTL更具语义特征的知识搜索和关联体验。

（2）在商业应用方面，知名医学数据库PubMed通过医学主题词（MeSH），对自然语言表达的医学文献进行规范化处理和标引，表明文章核心内容，实现基于语义树的引导式搜索。PubMed凭借其语义级别的标引，在医学领域得到广泛应用，在知网以PubMed为关键词搜索，可以查到2000余篇论文是基于PubMed产出的科研成果。目前，国内也有一些数字化公司开发产业数字大脑平台，即按照产业链的思路，对某一企业发展的上下游企业、所需人才、技术、资源进行语义化关联，实现对企业或产业的动态跟踪和管理。

（3）在标准知识层面，2019年，ISO/IEC正式提出了一种名为SMART（Standards Machine Applicable, Readable and Transferable）标准数字化的新型标准概念[1-3]。将标准数字化发展划分为

注：① 阶段0，即形成纸质版标准。
阶段1，传统数字化格式，采取简单PDF格式。
阶段2，机器可识别。包含标准文本结构化的内容，可利用软件识别文件结构并进行基本处理。
阶段3，机器可执行。可根据应用场景选择性地访问赋有语义的标准内容，可利用应用程序界面对标准内容执行较复杂的操作。
阶段4，机器可决策，也称为智能标准。机器能够以更为复杂的方式执行或解析标准内容。

5个阶段，包括："纸质文本（阶段0）""开放数据格式（阶段1）""机器可读文档（阶段2）""机器可读内容（阶段3）""机器可交互内容（阶段4）[①]"。ISO/IEC在工业领域已经提出并积极实践了面向机器可读的工业通用语义知识库。目前，各国际标准组织及部分先进国家部分标准数字化已达到阶段2，并率先在食品和农产品、信息技术、智能装备、航空航天等领域开展了面向阶段3~4标准数字化的应用和探索。

在标准知识领域，我国尚缺少统一标准知识数字化表达模型，即如何明确标准文献关键知识的组织要素是本研究的重点。同时，在我国，由于我国食品和农产品安全领域不同标准文本内容及结构的差异，我国食品和农产品环保安全知识数字化技术推进缓慢，整体还处于纸质标准电子化、结构化的标准数字化初级阶段（阶段1）针对特定标准知识尚未实现自动化标注与抽取，尚未有对食品和农产品领域标准数字化转型过程中建立类似于ISO/IEC面向机器可读的标准知识抽取与知识库，存在检索标准资源不全，检索手段落后、查全率和查准率低、检索质量不高等问题，与国外存在较大差距。

2 标准知识数字化表达通用模型与自动抽取技术研究

2.1 基于知识本体理论的标准知识数字化表达的通用模型

为更好对标准文献结构进行结构化、知识化、可视化分析，本文基于语义网理论，基于知识本体理论，采用叙词表等组织方式（示例见表1），开展了标准知识三元数据模型研究，深化标准文献的多粒度内容描述和知识关系的表达揭示，对标准化对象、指标项等实体概念进行语义关联。通过对国家标准、行业标准的内容主题分析与标引，涵盖工作场景、业务流程、应用设备等多种组织维度，对同专业的各个类型的标准按照相同或相似的要素结构进行分析分解，在分析归纳的基础上提炼出了既适合于结构化分解标准文献的技术指标，又能适应不同类型标准揭示标引的统一数据分解模型，构建了较为通用标准的知识模型和人工加工方法，形成了标准数字化的通用模型和方法的相关标准，率先创新性地提出了本体（标准化对象）–体例（标准段落结构）–标准指标的三元数据结构。

其中，本体和体例均需要建立同义词和上下位的关系，标准指标则还包括指标项、指标值、计量单位、限定类等，从而实现文献碎片化分析，实现对标准知识的数字化表示，这样就通过三元组数据模型，将标准内容转化为具有语义关联关系的数据。值得强调的是，由于标准文献结构和形式各异，即使同一标准文献也可能由文字、数值、图表以及引用等不同内容结构组成。因此，为更好理解上述三元数据模型，本文后续将通过具体例子实证检验不同内容结构下的本体（标准化对象）–体例（标准段落结构）–标准指标的确定问题。

2.2 基于自然语言处理和机器学习的标准知识数字化抽取技术

为大幅度降低标准知识标准化和抽取的人工成本，开展基于自然语言处理和机器学习的半自动化标准知识组织技术研究，通过对半结构化数据及非结构化数据做半自动化处理[4-6]：以人工处理的结构化数据为训练集，应用机器学习框架，针对半结构化数据，实现自动的实体与关系标注；以人工构建的词表和语法规则范式为基础，针对非结构化数据，实现实体识别与消歧、关系标注，并构建标准知识库。再由专家对关键信息进行总结，通过迭代的方式优化标注结果，供专家筛选判断，以此加快标准知识的构建过程。具体如下。

（1）针对自然语言文字为主的失信信息，采用基于规则的方法，如：使用正则表达式或者巴克斯范式等规则框架的模式，配合词表进行范式匹配，基于规则的模板匹配，基于语义规则的解析等，实现描述性内容的实体识别和关系抽取。

（2）针对表格为主的失信信息，采用基于机器学习的方法，如：基于朴素贝叶斯的文本分类，基于深度学习的段落分类，基于神经网络的句子分类等，实现关键要素的实体识别和关系分类。

（3）针对需重点分析的失信信息，采用基于统

计的方法，如：基于词袋模型的文本分类，基于统计特征的段落分类，基于统计模型的句子分类等，实现细粒度的知识图谱的构建。

（4）针对其他类型的失信信息，采用基于搜索的方法，如：基于搜索引擎专业的关键词表的段落和句子抽取，实现失信内容的细粒度命中。

3　食品和农产品标准知识的实证研究

本文基于“标准化对象—体例—指标项—取值范围—指标值—计量单位—限定条件”等知识组织模型（如图1所示），通过人工或已有标注的食品和农产品的训练数据集（见表1），利用自然语言处理和机器学习等技术实现了对2000项食品和农产品标准知识的自动高精度标注和抽取。

表1　食品和农产品标准知识训练数据集（示例）

	标准化对象（T1）	指标项（T2）	指标值（T3）	限定条件（L）
Oracle	1181	1672	2550	365
食典通	13495	9591	11284	0
实验数据	13788	10545	13542	14634

限于篇幅，本研究仅展示了鲜苹果和乳制品标准知识的抽取结果，见表2和表3。根据表2所示，在鲜苹果中优等品的大型果的质量要求中，对质量等级要求是果径（最大横切面直径）≥70mm，通过上述标准知识数字化表达模型，将标准内容转化为具有语义关联关系的数据。根据表3所示，乳制品中乳粉的色泽应呈均匀一致的乳黄色或具有应有的色泽。

4　研究结论

本研究将以食品与农产品领域标准为研究对象开展标准知识数字化表达模型与自动提取技术研究，首先，明确国内外标准化数字化当前研究现状与问题；其次，通过文献和实地调研，创新性地提出标准知识数字化表达模型；再次，开展标准知识数字化自动提取技术研究，实现对数字化表达模型知识要素的自动标注和抽取，据此形成具有语义关联的标准知识；最后，以2000项食品与农产品领域标准为例进行标准知识数字化表达模型与自动提取技术的实证研究，验证理论或技术的可行性。研究发现如下：（1）构建了适用于标准知识的数字化表达模型，即本体（标准化对象）–体例（标准段落结构）–标准指标的三元数据结构模型，通过上述标准知识数字化表达模型，能够将标准技术内容转化为具有语义关联关系的数据。（2）提出

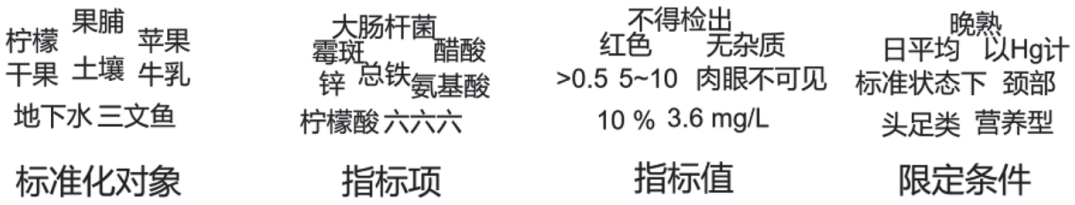

图1　食品和农产品标准知识分类（示例）

表2　GB/T 10651规定的鲜苹果产品标准知识数字化表达和抽取结果示例

产品标准元数据中文名称	标准化对象	一级体例	二级体例	指标项	取值范围	指标值	计量单位	一级限定类	二级限定类
产品标准对象内容	鲜苹果	质量要求	质量等级要求	果形		具有本品种应有的特征		优等品	
	鲜苹果	质量要求	质量等级要求	果形		允许果形有轻微缺点		一等品	
	鲜苹果	质量要求	质量等级要求	果径（最大横切面直径）	≥	70	mm	优等品	大型果

表3 NY/T 657-2021规定的绿色食品 乳与乳制品产品标准知识数字化表达和抽取结果示例

标准化对象	一级/二级体例	指标项	指标值	一级限定条件	二级限定条件
乳制品	产品标准/技术要求	色泽	色泽均匀一致，呈乳白色或微黄色	发酵乳	发酵乳
			具有与添加剂成分相符的色浮		风味发酵乳
			呈均匀一致的乳白色或乳黄色，有光泽	炼乳	淡炼乳、加糖炼乳
			具有辅料应有的色泽		调制炼乳
			呈均匀一致的乳黄色	乳粉	乳粉
			具有应有的色泽		调制乳粉
		滋味、气味	具有发酵乳特有的滋味、气味	发酵乳	发酵乳
			具有与添加成分相符的滋味和气味		风味发酵乳
			具有乳的滋味和气味	炼乳	淡炼乳
			具有乳的香味，甜味纯正		加糖炼乳
			具有乳和辅料应有的滋味和气味		调制炼乳
			具有纯正的乳香味	乳粉	乳粉
			具有应有的滋味、气味		调制乳粉
		组织状态（外观）	组织细腻、均匀，允许少量乳清析出；风味发酵乳具有添加成分特有的组织状态	发酵乳	发酵乳、风味发酵乳
			组织细腻，质地均匀，黏度适中	炼乳	淡炼乳、加糖炼乳、调制炼乳
			干燥均匀的粉末	乳粉	乳粉、调制乳粉

了基于自然语言处理和机器学习的标准知识数字化提取技术，利用自然语言处理和机器学习等技术实现了对2000项食品和农产品标准知识的自动高精度标注和抽取，为我国标准化工作迈向ISO/IEC提出的阶段3“机器可读文档”提供技术参考。

参考文献

[1] 刘曦泽,牛娜娜,王益谊. SMART标准用例分析与启示[J]. 标准科学, 2022(12):63-67.

[2] 马超. 面向机器可读标准的电力标准数字化述评与展望[J]. 中国电力, 2023,56(8):216-229.

[3] 崔静,王立玺. 标准数字化工作路线图探究[J]. 信息技术与标准化, 2023(06):43-46.

[4] 姬发家,朱莹,阴皓,等. 基于混合神经网络的自然语言处理技术研究[J]. 电子设计工程, 2023,31(10): 92-96. DOI:10.14022/j.issn1674-6236.2023.10.020.

[5] 王江鹏. 基于深度学习的自然语言处理技术发展分析[J]. 中国安防, 2022(12):40-43.

[6] 翟旭京,白宇,刘艳茹,等. 基于机器学习的配电网监控信息批处理方法[J]. 微型电脑应用, 2023,39(05):39-42+50.