情报杂志
Journal of Intelligence
ISSN 1002-1965,CN 61-1167/G3

《情报杂志》网络首发论文

题目：面向数字知识管理的智能内容研究进展
作者：王思丽，祝忠明
网络首发日期：2018-12-14
引用格式：王思丽，祝忠明．面向数字知识管理的智能内容研究进展[J/OL]．情报杂志.
http://kns.cnki.net/kcms/detail/61.1167.G3.20181213.1506.006.html

网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

网络首发时间：2018-12-14 13:40:50
网络首发地址：http://kns.cnki.net/kcms/detail/61.1167.G3.20181213.1506.006.html

面向数字知识管理的智能内容研究进展*

王思丽[1,2,3]　祝忠明[1,3]

（1. 中国科学院西北生态环境资源研究院文献情报中心　兰州　730000）
2. 中国科学院大学　北京　100049；3. 中国科学院兰州文献情报中心　兰州　730000）

摘　要　[**目的/意义**]随着人工智能和机器学习技术的不断发展推动，智能内容作为一种情报内容组织与挖掘的新策略模式，已成为相关科技行业领域的研究热点与制胜点之一。本研究将有助于系统理解与全面掌握智能内容研究与应用的方法和技术路线，为内容驱动的数字知识管理与服务提供理论方法支撑。[**方法/过程**]采用文献研究法、演化分析法、跨学科研究法、比较分析法、案例分析法与归纳总结的方法，对智能内容的研究背景意义、概念定义、典型应用案例和研究发展现状等进行了分析探讨。[**结果/结论**]面向数字知识管理的智能内容研究是发展趋势也是行业挑战，数字图书馆领域应考虑借鉴相关的 XML/JSON 内容结构化实现方法标准和智能内容开发模式策略，致力于模块化、结构化、语义丰富、无格式、可重用和可配置化的数字知识管理系统的研发与应用。

关键词　数字知识管理　智能内容　内容智能　内容管理　内容策略
中图分类号　G250.7　　**文献标识码**　A

Research Progress on Intelligent Content Oriented to Digital Knowledge Management

Wang Sili[1,2,3]　Zhu Zhongming[1,3]

（1. Literature and Information Center of Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou　730000;
2. University of Chinese Academy of Sciences, Beijing　100049;
3. LanzhouLiterature and Information Center of Chinese Academy of Sciences, Lanzhou　730000）

Abstract　[**Purpose/Significance**]With the continuous development of artificial intelligence and machine learning technology, intelligent content as a new strategy and model for the organization and mining of information content has become one of the research hotspots and winning points in the related field. This research will help to systematically understand and fully grasp the methods and technical routes of intelligent content research and application, and provide theoretical methods for content-driven digital knowledge management and services. [Method/Process] Using literature research method, evolution analysis method, interdisciplinary research method, comparative analysis method, case analysis method and summary method, the research meaning, concept definition, typical application cases and research application status of intelligent content are discussed in detail. [**Result/Conclusion**]Intelligent content research for digital knowledge management is a development trend as well as an industry challenge. The digital library field should consider the relevant XML/JSON content structured implementation method standards and intelligent content development model strategies, and strive to study, development and application of intelligent digital knowledge management systems that are designed to be modular, structured, semantically rich, format free and, as a consequence, reusable, discoverable, configurable and adaptable.

Key words　digital knowledge management　intelligent content　content intelligence　content management　content strategy

---

基金项目：中国科学院兰州文献情报中心 2018 年文献情报创新能力建设项目“基于深度学习的领域本体自动构建方法研究”（编号：Y8AJ012005）的研究成果之一。
作者简介：王思丽（ORCID：0000-0002-2126-3462），女，1985 年生，博士研究生，馆员，研究方向：知识发现与知识组织，知识管理系统建设等；祝忠明（ORCID：0000-0002-2365-3050），男，1968 年生，博士生导师，研究馆员，中国科学院兰州文献情报中心资源系统建设部主任，研究方向：知识发现与知识组织，知识管理系统建设等。

0 引 言

数字知识管理系统是重要的知识内容组织、管理和服务工具,能够有效地促进知识交流与共享。目前,随着人工智能和计算机网络技术的不断进步,内容变革的步伐和范围进一步加剧,传统的支持单一内容类型结构和功能集中固定的单块架构模式的数字知识管理系统正面临严重生存危机,智能内容作为一种新兴的情报内容组织和挖掘的策略模式,是发展趋势也是行业挑战,将为数字知识管理系统建设带来契机。

本研究将首先系统化梳理智能内容产生的背景意义、概念定义及演化特征以揭示什么是智能内容,为进一步研究奠定理论背景基础;其次通过对智能内容的典型应用案例分析,着重归纳总结了面向数字知识管理的智能内容研究进展与应用现状。最后进行小结,概述了智能内容驱动的数字知识管理系统构建的必要性。

1 什么是智能内容

1.1 智能内容产生的背景及意义 智能内容,英文文献资料中多表述为“Intelligent Content”或“Smart Content”或“Content Intelligence”,中文常将其概述为利用人工智能进行内容挖掘(生产/消费)的过程。相关调研表明,从1956年AI被提出,至今内容挖掘共经历了大约四个里程碑性质的阶段:

(1)第一阶段:大概时间间隔为1956-1968,该阶段的主要特点是依靠计算机简单辅助人工制作编辑内容,内容挖掘是以人工为主导,但受计算机等媒体自身的支配。常见的如期刊/报纸/杂志的编辑,电视/电影/舞蹈/话剧的编导等,是基本完全依赖专业人员围绕某一系列特定主题进行精心编辑/编排/编导优质内容进行内容生产与呈现。

(2)第二阶段:大概时间间隔为1969-2005,该阶段的主要特点是利用互联网搜索引擎驱动查找内容,内容挖掘是以搜索为主导,受计算机网络技术的支配。代表性标志分别是1969年互联网的诞生,到1989年万维网技术的突破,再到1998年谷歌搜索引擎的出现,2000年中文搜索引擎百度的上线等,基本上用户所有的内容查找和获取都可以依赖搜索服务进行实现。但前提是用户必须明确自己内容挖掘的目标,即预先知道想要获取什么,因为该方式并不能覆盖未知的领域。目前第二阶段发展依然迅猛。

(3)第三阶段:大概时间间隔为2006-至今,该阶段的主要特点是以用户生成内容为主导的社交网络时代。标志性事件是2006年谷歌收购YouTuBe,随后越来越多的互联网科技公司开始争先恐后地搭建各种社交网络平台并通过吸引用户去完成平台上的内容生产及发布,其他用户可通过平台推送去实时获取所需内容。国外知名的如Wikipedia、Facebook、Twitter等,国内的如微博、微信、知乎、豆瓣等。这种模式常基于用户主动关注或订阅去自动帮助用户获取优质的兴趣相关的内容,内容获取变得更方便,内容范围变得更广。

(4)第四阶段:大概时间间隔为2010-至今,以深度学习[1]为代表的机器学习技术热度骤然高涨,是被称为以机器学习驱动变革的智能内容时代。在该阶段中,用户进行内容挖掘不再完全依赖编辑或编导,也不必非得预先知道自己想要什么然后去搜索,也无需必须去关注很多网络平台的微博或公众号,机器学习技术驱动的用户在线行为分析(点赞/支持/分享/评论/收藏/浏览/检索等)将和内容生产者策划整合内容的理念相结合,自动为内容找到对应的受众群体进行量身定制性地主题内容聚合、推送与呈现,以最大化达到“内容即服务”,“所见即所得”,“所想即所得”的效果。近年来,国内外已经有很多社交/媒体/电商等平台率先涉足智能内容领域,主要研究应用包括用户画像、智能推荐、智能问答、机器翻译/创作、图像/语音/人脸识别等。如Twitter、Facebook等过去都是以时间优先顺序为机制进行内容推送,如今通过定位和分析用户兴趣,基于内容对用户的重要度优先算法进行智能推送。自2014年起美联社、今日头条、新华社等先后开始基于自然语言处理技术和智能推荐技术等研发机器人进行自动数据采集、加工、写稿和智能分发等。正如Mode Media公司的首席执行官Samir Arora所述,“智能内容比我们自己更清楚我们是谁,在什么位置,喜欢看什么内容”[2]。

总的来说,这四个阶段虽然是由于相关技术不断进步而推动的由低阶到高阶的内容挖掘过程,但目前都并未被完全取代,未来也将会长期共存,不同的是以什么为主导,是否最大化地解放了人工及运用了智能化技术。在2016年的世界经济论坛年会中,埃森哲首席执行官Pierre Nanterme认为“未来内容运营将无法抵挡数字化带来的前所未有的业务中断”,并将这一变化称为“第四次工业革命”[3]。第四次工业革命浪潮下包含了一系列智能化技术:S(Social Network,社交网络),M(Mobile Media,移动传媒),A(Artifical Intelligence,人工智能),B(Big Data,大数据),C(Cloud Computing,云计算),常被简称为SMABC;但也带来了快速化规模化的扰乱性危机:内容消费者可以轻易借助SMABC技术享受更多的互动和个性化体验,不再单纯只是被动等待或仅忠于某一家互联网科技公司。内容生产者可以通过快速融合SMABC技术来接近用户、优化成本和革新业务,大幅提高内容质量和生

产力,竞争无边际成本。因而趋势所逼,更多数字化行业不得不采取内容转型,像上文提到的 Facebook、Twitter 都曾一度面临生存困境,最终依靠智能内容破解了盈利难题。我国 2017 年的百度云智峰会,2018 年的搜狐 WORLD 大会等,主题都包含大数据与智能内容营销,并提出智能内容技术将成为新一轮产业革命的核心驱动力,而基于用户画像的智能内容营销将是制胜点。同时,美国内容科学公司的创始人 Colleen Jones 发布的“2018 内容预测及综述”[4]一文中指出,内容的可及时查找性将成为当下最大的挑战,并根据 IBM 报告“What is Watson”所调研,当前互联网上非结构化数据约占全部数据的 80%,而计算机无法直接从非结构化数据中获得有意义的内容信息。智能内容的核心理念是通过借助机器进行自主学习并且很大程度上不需要依赖人工重新编程而实现将内容数据结构化、语义化、可视化等将可能有效解决这一难题。综上所述,未来人工智能和机器学习技术可能会历经多次转型,智能内容正处在这种转型的初始阶段,是起始点也是转折点,具有重要的研究意义和良好的应用前景。

1.2 **智能内容的概念定义及演化** 自 2007 年以来,国内外学术界逐渐开始关注智能内容,从不同的专业领域视角来描述标准结构化数据分析与非结构化数据挖掘之间的协作关系,认为智能内容的首要特征是将普通非结构化内容与结构化内容区分开来的一种属性或特质,并由此涌现了一些具有代表性的概念定义:

2007 年,英国布洛尔研究公司(Bloor research)的 Gerry Brown 率先提出了“内容智能[5]”这一术语,认为内容智能是商业智能和内容管理的结合,旨在探索转化和利用其公司日益庞大的非结构化数据的策略。随后,Gerry Brown 进一步将“内容智能”解释为:“内容智能提供了散乱的企业数据和文本孤岛的完整的 360 度全景视图。它提升了将文本和数据作为一个整体进行切片,分块,深入挖掘和报道的能力。”但由于初期对内容智能的研究常依赖于大量的数据点和工具分析,很容易使数据的影响性变弱,再加上由于其在公司的领导力不足及缺乏专门的资源配置等,“内容智能”并未被广泛研究采用。

2008 年,美国 Rockley 集团总裁 Ann Rockley 发表了“什么是智能内容”[6]一文,对智能内容进行了明确定义:“智能内容被设计为模块化,结构化,可重用,无格式和语义丰富,因此可自动发现,可重新配置和适应性强”。Ann Rockley 认为智能内容是对数字文本,图像,视频,音频或多媒体数据进行改编后添加的编码,允许自动处理用于各种用途并适用于不同设备和接口的访问。创建智能内容的过程涉及删除格式和添加语义元数据,语义元数据标签可有助于实现内容模块化并允许自动组装、格式转换和交付。以 Ann Rockley 为代表的 Rockley 集团不仅创造了智能内容这一术语,近十年来也一直致力于开发智能内容管理策略和基础信息架构的研究。Rockley 集团于 2010 年开始创立了智能内容会议(Intelligent Content Conference,ICC)[7],迄今为止 ICC 会议已举办了多届,最近的一届于 2018 年 3 月 20-22 在美国拉斯维加斯举办,会议主题是讨论无格式、模块化、结构化的内容创建和分发方法,旨在为智能内容定义可行方法与最佳实践。同时,Ann Rockley 在统一内容策略,企业内容管理和智能内容研究方面发挥了重要作用,目前已出版了多部有关内容策略和智能内容研究的著作,其代表性的著作是 2012 年出版的“管理企业内容:统一内容战略”[8],随后又于 2015 年出版了“智能内容:入门”[9]一书,重申了智能内容的定义、五大特点及提供一些典型应用案例以供分析学习。

2010 年,美国基于神经科学的预测分析技术公司 Evolve24 的研发副总裁 David Geddes 基于自己长达 25 年的信息资源整合、商业情报分析和内容数据管理经验,将智能内容阐述为:“智能内容是准备就绪的全球市场情报,以企业的速度在企业中共享,从而推动战略性和运营性业务决策。”[10] David Geddes 认为智能内容应建立在一套完整的相关文件的基础上,通过电子数据库、在线新闻、博客、论坛、讨论组、视频、微博等的智能搜索引擎实时汇总、分析、实体提取、情感评分、主题识别、数据挖掘等,以近实时方式提取情报信息,并将信息以用户友好的方式排列在相关门户网站中。他认为智能内容是内容管理转向内容智能的结果,既是望远镜,可以看到新兴的趋势,机会和风险;同时又是显微镜,可以检查受众的态度,心理状况或相关统计信息,因此可用于对未来市场的动态预测和建模,以便革新科技行业进行商业决策的模式。

2014 年起,美国内容科学公司开始致力于内容智能及内容策略研究,近年来已发布多个智能内容系统及影响力研究的白皮书,涵盖了内容智能相关的概念、实践、标准和工具等。代表性论著是其创始人 Colleen Jones 在 2015 年发布的“一个具有影响力的智能内容系统的 4E”[11]。随后受 Ann Rockley 的影响,在 2016 年发表了文章“什么是内容智能”[12],重新定义了内容智能的概念,提出了智能内容系统的基本框架,并预测内容智能将成为未来一种重要的战略实践:“内容智能是代表将内容数据和业务数据转化为具有影响力的内容策略和战略的可操作理解的系统和软件。”该概念主要包含三个关键点:①一个系统化的方法:内容智能并不是一个已完成的命题。它需要一个框架、一系列流程及人员参与。②集成化的软件:需要将正确的

软件工具集成到框架和流程中,否则开发内容智能是不可能的。③内容影响力:Colleen Jones 认为通过内容来操作、理解、评估数据是关键,内容智能最大的成功不是在于获取已创建和发布的内容,而是必须具备能够收集多个数据源并针对有关内容问题执行分析和解释的一种影响力。基于上述概念,Colleen Jones 指出内容智能不单是指人工智能 AI、商业智能 BI 或智能内容其中之一,应是三者领域互补的重叠区域,是内容评估的演变,是一种更全面和更复杂的内容自动化处理模式。

目前国内对智能内容的概念定义等基础理论体系研究较少,一般以国外的基础理论方法为基础,相关研究主要集中在智能内容技术及应用探索方面。如腾讯发布的智能推荐 API[13],以腾讯海量的用户行为数据和在游戏、电商、金融、资讯等多领域积累的产品大数据为依托,以数据、算法、系统为核心,为用户提供基于海量用户画像和实时大数据机器学习的内容个性化推荐服务。如亿欧智库在 2017 年发布了人工智能+内容生产的研究报告[14],对当前国内外人工智能在内容生产领域的应用现状进行了深入分析与探讨。

综上研究,本文发现:

(1)智能内容和内容智能虽是两种文字表述方式,但实质都是研究内容管理、转化和利用的一种自动化处理模式和策略,旨在强调如何最大化自动化实现从散乱的文本和数据中提取结构化信息并转化为可用知识进行情报决策服务,内容越智能化,服务将越精准化。从 Bloor research 初步的将文本和数据分块集成挖掘理念到 Rockley 集团的致力于内容模块化和结构化的创建和转换方法研究,到 Evolve24 的基于相关智能化技术实时提取和推送情报信息的认知,再到内容科学公司的从框架流程、软件系统和内容影响力分析等多角度多方位的概念阐释,是一个贯穿基本理论、方法实践、技术路线再到综合分析实现的一个逐步发展深入的演变过程。

(2)智能内容和内容智能两种研究模式也存在一些差异。从语法和概念本义来看,智能内容相当于一个名词短语,是一种狭义的概念,侧重于从内容数据自身的组织结构出发来处理和挖掘内容,它应是实现内容智能的基础步骤。而内容智能相当于一个动词短语,是一种更广义的概念,侧重于提倡从框架流程、软件工具、人员组成等多方面因素综合考虑内容的结构化处理与应用研究,它是智能内容向系统化应用发展的必然经历和理想结果。

(3)总体来看,内容智能与智能内容的研究彼此影响与受益,应是相辅相成异曲同工的,但由于内容智能考虑的因素更多,难以系统化实现与工程化复用,因此目前国内外研究较多的是智能内容,且已形成了基本的理论方法体系。智能内容作为一个跨多领域的转型期的新概念,被誉为当下知识产品和知识服务的发展需求和趋势,是可快速应对用户消费需求和内容特征变化的个性化内容技术,其基础理论方法已率先在国内外其他行业领域如社交、传媒、电商、搜索引擎等领域得到了广泛应用实践且效果显著。但同时,智能内容作为一种内容组织和实施的策略和技术导向,应是伴随着智能技术的发展和突破而不断更新迭代的,因此其技术工具和实践应用研究并未成熟,尚有很大的研究空间。

2 面向数字知识管理的智能内容研究

2.1 典型案例分析 智能内容作为一种新兴的内容组织策略和技术导向,目前已在面向数字知识管理的相关研究如行业内容结构化组织标准、企业内容结构化组织模式、知识库内容结构化设计及服务丰富策略、基于内容即服务 CaaS 架构的基础内容管理平台建设等中已得到了诸多共识与实践应用。以下是 4 个典型案例分析:

2.1.1 DITA DITA[15-17] 是指达尔文信息分类体系架构(Darwin Information Typing Architecture),最早是由美国 IBM 公司为解决内容生产和发布中的标准化描述、规范化组织、统一化存储等问题而提出,目前是由国际结构化信息标准促进组织 OASIS 所支持。DITA OASIS 标准主要用于设计、编写、管理和发布规范化的基于 XML 体系结构的、面向主题的内容信息,目前的最新版本是 OASIS 于 2015 年发布的 DITA 1.3 规范,较之以前版本有了很大改进:将对主题和地图的新关注作为核心文档内容类型;针对不同的受众群体提供了不同的版本规范,包括核心 DITA 版本,技术内容版本,专用于培训和教学的版本,并扩展了在教学、医疗、出版、营销、制药等行业领域的应用;此外还提出了开箱即用的轻量级半组件化 DITA 架构规范和即将构造设计全新的模块化 DITA2.0 的策略思想。

DITA 的核心内容架构模型如图 1 所示:

图 1 DITA 核心内容架构模型[16-17]

(1)DITA 交付情境:用于定义 DITA 适用的应用

场景。

(2)DIAT 类型化主题结构:主题是 DITA 中最基本也是最顶层的结构单元,不允许被嵌套,其他内容结构都将围绕主题进行分类描述与组织。

(3)跨多信息类型的专业词表或主题领域:以 XML 为载体,通过 DTD 和 Schema 定义基于主题的结构化内容。针对特定领域,允许通过扩展主题标签,加入新的专业词表或领域词表,实现对新的主题类型和领域信息的支持和共享。

(4)DITA 公共结构:由于 DTD 和 Schema 主要是 XML 的技术规范,为了支持其他序列化信息类型的重用和共享,DITA 公共结构用于定义不同于 DTD 和 Schema 的一些元数据结构和基于 OASIS 标准的语义化表示的表格结构等。

当前 DITA 主要用于技术手册、交互培训、教材、标准、报告、商业文档、旅游书籍的编写等,其最显著的特性就是利用 DITA 内容架构模型生产的单源内容人和机器都可理解,可通过不同方法进行自动重用,支持多渠道内容交付与输出,即所谓的"人机可读,一次编写,多重引用,多元发布"。同时,DITA Wiki 社区已提供了 DITA OASIS 标准的独立开源实现工具包 DITA-OT[18],支持基于领域主题的结构化内容创作与发布,并可将结果输出为多种文档类型格式,如 XHTML/HTML、Java/Eclipse 帮助、PDF、ODF、Word RTF 等。DITA-OT 已在多种创作工具如 Adobe FrameMaker、OXygen XML Editor 等,多个 CMS 如 Astoria CMS、easyDITA 等,多类交付系统如 DITAweb、ePublisher 等,多类打印制作工具如 DITA InPrint 8.13、Mita3.4 PDF 等中得到了大量应用。

2.1.2 Quark Quark 是美国一家具有近 40 年发展历史的软件公司,旨在为创建、管理、发布和交付结构化内容提供端到端的自动化解决方案,在企业动态出版和语义出版研究中颇具影响。Quark 认为当前的大部分基于 XML 的内容结构化创建工具过于技术化,表面上对用户隐藏标签结构但实质上在内部又基于标签至上的元素名称严格验证文档结构,存在严重的跨域编辑[19]问题,且其内容发布不能够自动适用于多渠道,设计丰富的布局样式和专业输出需要程序员级别的技术来定义所需的样式表。因此,Quark 致力于使得非专业技术人员也能够创建结构化的、语义丰富的、基于 XML 的、自适应多渠道的智能内容,并提出了 Quark 智能内容定义:开放的、用户可配置的基于 XML 的结构化内容组织模式[19]。

Quark 借鉴了 DITA 及许多 XML 结构化的实现思想,提出了基于内容类型的根类和类层次的概念:一组包含基本类型的内容是最顶层的结构单元(根类),其他所有内容都需要被语义化描述为这些根类别之一。Quark 智能内容对内容类型的专门化是基于语义标签的结构化规则,与 HTML、DITA 有所不同,如表 1 所示:

表 1 Quark 与 HTML、DITA 的内容类型描述对比

内容类型	Quark	HTML	DITA
Sections	section	div	topic
Blocks	p	p	p
In-lines	tag	em, strong, etc.	phrase
Lists	ul, ol	ul, ol	list type="type"
Tables	tables	table	table
Images	image	img	image
Media	Media	video, object	object
Metadata	XML meta fragment	tag attribute = "value"	tag attribute = "value"

(1)HTML 常驱动 CSS 样式或触发标签特定的 JS 来实现对内容类型的专门化描述与组织,如常见的基于 class 属性的编码表示:

…

该表示通常是非结构化的甚至是研发者随意书写的可以被随时替代的符号,并不含语义,且并不限制使用及验证 class 规则的值。

(2)DITA 使用面向主题的分层结构表示,严格的基于元素名称验证文档结构,如:

<concept class ="-topic/topic concept/concept">…</ concept>

该描述方法人和机器都容易理解:元素"concept"属于类"topic"。

(3)Quark 预定义了常见的几种内容类型,针对每一种内容类型,都制定了基于语义的专门化规则,支持基于标准操作过程创建文档并将每个文档限制为一个且仅限一个"purpose"以区分验证。如:

<section type="purpose">…</section>

同时,Quark 也预定义了多种规则供选择与配置。此外,目前 Quark 已发布了多个智能化产品系统[20],如 Quark 语义出版平台、Quark XML 创作工具、Quark 智能销售平台 Docurated、Quark 智能打印平台 App Studio 等。这些系统基本都既有商业版又有开源版,可供相关研究与应用。

2.1.3 Elsevier - Article of Future"Article of Future"(未来文章)[21]是 Elsevier 自 2009 年以来一直在进行的项目,旨在研究探索更好地呈现知识库中在线期刊文章并丰富其内容的方式。Elsevier 通过大量的用户需求调研和市场统计分析,基于内容的结构化、可读性、可发现性和可扩展性等原则制定了"未来文章"的内容设计策略:①文章是根本基础。未来文章应支持当前主流的"PDF"格式外观,便于在线阅读。②嵌

入式内容丰富服务。新的内容元素应无缝整合到文章中的自然位置。③上下文情境信息展示。将在原始文章的文本旁提供补充性的内容和功能以及来自外部数据库的相关信息。④清楚的导航。文章的概述和目录应链接到文章中相应的章节或数字。⑤干净的设计。为鼓励和吸引用户浏览文章视觉负载应达到最小化。⑥自定义视图。用户界面可根据屏幕尺寸进行自动调整,自适应特定主题。

基于上述内容设计策略,2012 年起,Elsevier 对 ScienceDirect 知识库中的所有文章都采用了这种更加动态和用户友好的格式-交互式 HTML 格式:主要是对文章内容进行结构化分块,对介绍、结果、数值、讨论等进行标签式导航。在内容视图页面支持三窗格式布局展示,每个窗格都可以独立滚动,同时显示交互式文本和图像,如图 2 所示。

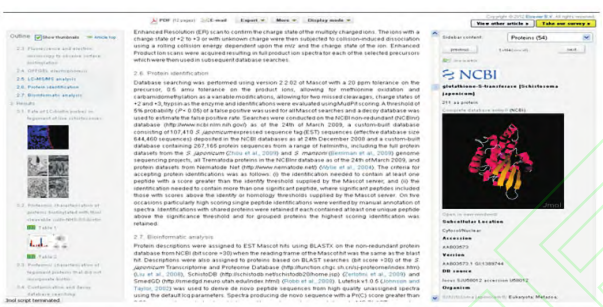

图 2　Elsevier 未来文章的内容视图示例[22]

(1)中间内容区:主要显示原始文章和基础数据、图表等。

(2)左侧导航区:提供带有可点击的章节标题、图像、表格缩略图的目录。

(3)右侧附加区:提供访问补充信息和附加功能,也可直接从文章内容或通过下拉菜单访问。并根据每个学科领域的特殊性和文章内容的不同,提供不同的功能。如在电化学类文章中展示化合物的分子结构;在古生物学类文章中展示 3D 化石模型,并提供它们被发现地点的空间分布信息;在地球科学、生命科学类文章中嵌入交互式谷歌地图等。

此外,Elsevier 还将智能内容研究引入临床医学领域,发布了 ClinicalKey 医学信息平台[23],基于爱思唯尔合并医学分类法对大量医学内容信息进行了深度语义标引,以提供更为丰富的主题分类服务。

2.1.4　Contentful　Contentful 是基于 CaaS 架构的基础内容管理平台[24],旨在将内容作为服务运行,提供 API 集合帮助开发人员管理,集成和交付各种平台类型上的数字知识内容。通过分析可以看到,Contentful 的基本数据模型是:将内容组织到空间中,允许将一个项目平台中所有相关资源自由组合在一起,包括内容实体,媒体资产及将内容本地化为不同语言的设置及自适应国际化标准语言环境的翻译配置等。每个空间都支持用户创建不同的内容类型,并包含一些基本元数据信息等。

Contentful 主要内容架构模式是:

(1)RESTful API:基于 RESTful API 将内容数据封装为标准化的结构化的 JSON 格式实现对内容、资产、翻译的全面程序化控制。

(2)微服务架构:基于容器的管理模式,完全分离的写入和读取 API 确保在不中断应用程序的情况下实现容错服务。

(3)缓存技术 + API 负载均衡技术:采用高级缓存技术,可与外部内容传输网络紧密集成,提供 API 的有效负载。

(4)可移动数据同步技术:基于 API 实现对系统中图像自动压缩、格式转换以及离线下载。提供 Sync API 允许通过增量更新或已更改的内容,将空间中的所有内容的本地副本保持为最新。

Contentful 主要提供了四类内容服务 API:

(1)内容传输 API:用于将内容从 Contentful 发送到应用程序,网站和其他媒体的只读 API。内容是作为 JSON 结构化数据,图像,视频和其他媒体是作为二进制文件来传递的。该 API 可通过分布式内容传输网络获得,离用户最近的服务器将智能提供所有内容。这将最大限度地减少延迟,尤其有利于移动应用在多个数据中心托管内容,也大大提高了内容的可用性。

(2)内容管理 API:一个用于管理内容读写的 API,并支持多种管理方式。如:①从 WordPress,Drupal 等知识库自动导入。②与其他后端系统如电子商务系统集成。③用户自定义内容编辑体验等,该方式是指在原 API 的基础上构建其他 Web 应用程序。

(3)内容预览 API:内容传输 API 的变体,将内容提供给用户之前智能预览。主要使用一个预览访问令牌使得内容管理员和作者可以在上下文情境中查看他们的工作,就像真正发布一样。

(4)图像 API:该 API 允许用户自定义调整和裁剪图像,更改背景颜色并将其转换为不同的格式。通过使用图像 API 进行上述转换,用户可以上传高质量的资源,准确传递应用所需的信息,并且仍然可以获得缓存的所有优势。

同时,Contentful 也提供了上述四类内容服务 API 的开源软件工具包,供扩展研究和应用。此外,Contentful 还支持当前内容模块化开发和工程化复用的相关标准编程语言和技术,如基于 JavaScript 脚本语言的 Node. js、Angular、Vue. js 等的应用开发等;基于 API 查询标准 GraphQL 检索数据并将其作为服务器连接到 React 前端 UI 的应用开发等;基于 Java、Rubby、PHP、

Python 等主流编程技术的应用开发等。

2.2 **研究进展总结** 从上述案例分析可以看出,目前面向数字知识管理的智能内容研究重点是内容的结构化组织方法和策略模式,与传统的支持单一内容类型结构和功能服务集中固定的架构模式有很大不同,并取得了一定的实践应用效果。但除此之外,还有更多其他智能内容相关研究和应用,总体来说,目前主要研究有:

(1)面向数字知识管理的智能检索研究:主要针对内容的相关性检索、关联度排序、情感评分等,或与外部资源整合时的主题识别及情景语义计算等,或基于图像/语音识别将其转换为文本或指令进行检索等研究。涉及的关键技术包含主题识别/挖掘、主题聚类/分类、机器学习、机器翻译、图像/语音识别、文本与语音转换等,常用的相关性开源工具包有 Elastic-Search、Apache Solr、Apache Mahout 等,相关性计算模型有布尔模型,向量空间模型 VSM,N-gram 统计语言模型,概率模型,机器学习模型等。如 Asma ElAdel[25] 于 2011 年提出了基于快速小波变换和深度卷积神经网络的图像智能检索工具。如 Joao P. Carvalho[26] 于 2017 年提出了基于智能检索和用户情感分析的智能专家系统 MISNIS,对 twitter 进行主题挖掘。如朱佳晖[27] 于 2017 年对基于深度学习的主题建模方法进行了实验研究。

(2)面向数字知识管理的智能推荐研究:主要针对内容及服务的个性化展示、智能化推荐等进行研究,涉及的关键技术包括关联计算、知识推理、知识图谱、用户画像、协同过滤、机器学习等。如 Nicola Capuano[28] 于 2009 年提出了基于 SAPI 规则的智能内容框架,可以基于用户偏好和情景分析自动调整内容和服务。如 Do-Eun Cho[29] 于 2013 年提出了基于协同过滤和内容相关性识别 的自适应智能推荐方案。如 Jose Aguilar[30] 于 2017 年提出了基于知识表示范式和推理机制的智能推荐系统框架。如黄立威等[31] 于 2017 年对基于深度学习的推荐系统进行了对比分析。

(3)面向数字知识管理的智能 CMS 研究:主要是从内容管理系统构建的底层框架和基本架构出发,对内容的结构化组织和管理、无限内容类型的动态定义和定制、复杂数字内容对象的关联和存储、自适应多渠道多租户的内容发布和利用模式等进行研究。目前主流的架构模式包括 API First 模式、模块化与组件化(插件化)模式、微服务架构模式、纳米出版模式、语义出版模式等。代表性的技术工具以 Java(包含 JavaScript、Jquery、Ajax 等)、Python、C#、Ruby、PHP 等主流的 Web 软件编程语言为主,辅以 Portlet、Nodejs、Angular、Vue、RESTful 等新兴的 Web 组件化技术框架。如 Jonathan P. Leidig 等[32] 于 2014 年提出数字图书馆领域应基于自然语言处理、启发式解析和挖掘、社交网络构建等技术将元数据结构化和规范化,建立能够智能分析与摄取内容的智能 CMS。如郭栋等[33] 于 2015 年提出了一个基于微服务架构和 Docker 轻量级容器技术的 PaaS 平台。如美国的日立数据系统[34] 于 2016 年推出了内容智能管理工具集 HCI, 支持跨异构数据孤岛和不同位置连接并汇聚多类型结构化数据,分布式动态集群和个性化定制连接和访问数据服务的 API 等,目前已在荷兰银行 Rabobank、美国国家档案局、中国电信等内容平台建设中得到了应用。如我国的中科汇联软件公司[35],一直致力于为用户提供 3C(Content/内容管理、Communion/交流协作、Commerce/电子商务)软件服务,于 2013 年与清华大学开展战略合作,转向人工智能和机器学习的研发,并于 2014 年推出了基于智能化、组件化、自适应 PC 端和移动端等技术模式的 easySite 智能内容管理平台,及基于 Portlet 技术、WebOS 整合功能和自定义工作流引擎机制的 easyPortal 智能协同系统,目前已在中国人民银行、中国海关总署、中国地震局、中国南航等的内容管理建设中得到了应用。除了商业化的 CMS 转向智能 CMS 研究,许多主流的开源 CMS 也逐渐向智能化发展,如 DSpace、Drupal、Fedora、Dataverse、Hydra 等开源系统,本身并不是模块化和组件化的架构模式,但开始逐渐支持语义化、结构化的关联数据,并提供了可扩展和互操作的 API。同时也涌现出了一批新兴的真正基于模块化、组件化、API first 或微服务架构模式的开源智能 CMS,如 Contentful[24]、Pubsweet[36]、Nuxeo[37]、Gentics Mesh[38] 等。

(4)面向数字知识管理的智能问答、智能创作研究等:主要是研发并应用机器人实现完全自动化的信息搜集与抽取、主题识别与挖掘、内容分析与整合的整个工作流程,实时动态的提供机器自动问答服务、机器自动完成写稿/编剧/编辑/图像设计/音频合成/视频剪辑/三维建模/动画生成任务等。但目前主要应用在体育、财经、地震、民生等内容结构比较固定的新闻媒体领域以及电商、游戏、娱乐等财力雄厚足以支撑耗资巨大的智能研究与应用的大型商业知识管理系统中,科技领域的研究还相对较少,因此不再细述。

3 结 语

研究表明,智能内容的产生是人工智能和机器学习技术发展的必然导向,已成为当前的研究热点和制胜点之一。智能内容为数字知识内容管理与服务等行业领域应对当前日益加剧的内容变革趋势及用户不断增长的需求变化提供了方向、策略与技术路线。作为

数字知识管理系统建设相关的研究人员,笔者认识到当下智能内容驱动的数字知识管理系统构建势在必行,应考虑借鉴相关的 XML/JSON 内容结构化实现方法标准和智能内容开发模式策略,致力于模块化、结构化、语义丰富、无格式(格式外观等演示信息与内容数据分离)、可重用、可发现与可配置化的数字知识管理系统的研发与应用,将基于 Restful API、模块化、组件化、大数据分布式索引与 NoSQL 存储等作为基本的内容架构开发模式,将实现人机友好的数字内容交互式格式和自适应多渠道内容的展示与服务作为基本的内容服务目标,不断促进智能化数字知识管理系统的研究进展与突破,为内容驱动的知识服务提供理论方法与技术工具支撑。

参 考 文 献

[1] Onal K D, Zhang Y, Altingovde I S, et al. Neural information retrieval: at the end of the early years [J]. Information Retrieval Journal, 2017: 1-72.

[2] ARORA S. Smart cars, smart thermostats — now here comes smart content [EB/OL]. [2016-02-22]. https://venturebeat.com/2016/02/22/smart-cars-smart-thermostats-now-here-comes-smart-content/.

[3] Nanterme P. Digital disruption has only just begun [EB/OL]. [2016-01-17]. https://www.weforum.org/agenda/2016/01/digital-disruption-has-only-just-begun/.

[4] Jones C. 5 Content Predictions for 2018 (and a Roundup) [EB/OL]. [2018-01-05]. https://review.content-science.com/2018/01/5-content-predictions-for-2018-and-a-roundup/.

[5] Brown G. Content intelligence: Content management meets business intelligence [EB/OL]. [2018-03-10]. https://searchbusinessanalytics.techtarget.com/podcast/Content-intelligence-Content-management-meets-business-intelligence.

[6] Rockley A. What is Intelligent Content? [EB/OL]. [2018-03-18]. https://www.eiseverywhere.com/ehome/69264/137386/.

[7] 2018 Intelligent Content Conference Las Vegas, Nevada [EB/OL]. [2018-03-18]. https://www.intelligentcontentconference.com/.

[8] Rockley A, Cooper C. Managing Enterprise Content: A Unified Content Strategy [M]. USA: XML Press, 2012.

[9] Rockley A, Cooper C, Abel S. Intelligent Content - A Primer [M]. USA: XML Press, 2015.

[10] Grimes S. This is Content Intelligence, According to 4 Experts [EB/OL]. [2010-10-07]. https://www.cmswire.com/cms/information-management/this-is-content-intelligence-according-to-4-experts-008811.php.

[11] Jones C. The 4 Es of an Influential and Intelligent Content System [EB/OL]. [2015-01-27]. https://review.content-science.com/2015/01/4-es-influential-intelligent-content-system/.

[12] Jones C. What Is Content Intelligence? [EB/OL]. [2016-02-11]. https://review.content-science.com/2016/02/what-is-content-intelligence/.

[13] 智能推荐 TIR-腾讯云[EB/OL]. [2018-03-20]. https://cloud.tencent.com/product/ir.

[14] 崔 粲. 2017 人工智能+内容生产研究报告[R]. 亿欧智库, 2017: 1-40.

[15] 范 炜. 达尔文信息类型架构 DITA 研究[J]. 情报杂志, 2009, 28(11): 172-175.

[16] 叶宇姗, 解 凯, 曾庆涛, 等. 数字出版中的 DITA 技术[J]. 北京印刷学院学报, 2016, 24(4): 29-32.

[17] Introduction to the Darwin Information Typing Architecture[EB/OL]. [2018-07-06]. https://www.ibm.com/developerworks/xml/library/x-dita1/index.html.

[18] DITA-OT [EB/OL]. [2018-05-06]. http://www.dita-ot.org/.

[19] The Beginner´s Guide to Smart Content [R]. Quark Software Inc, 2014: 1-24.

[20] Smart Content: XML Authoring for Non-Technical Users, Schema, Document [EB/OL]. [2018-05-08]. http://www.quark.com/Solutions/Content-Automation/What-is-Smart-Content.aspx.

[21] Aalbersberg I J. The Article of the Future [EB/OL]. [2012-09-21]. https://www.elsevier.com/connect/the-article-of-the-future.

[22] Designing the Article of the Future [EB/OL]. [2018-07-12]. https://www.elsevier.com/connect/designing-the-article-of-the-future.

[23] Elsevier Launches ClinicalKey for Individual Clinicians, Providing Comprehensive Content for 41 Specialties [EB/OL]. [2012-9-19]. https://www.elsevier.com/about/press-releases/clinical-solutions/elsevier-launches-clinicalkey-for-individual-clinicians,-providing-comprehensive-content-for-41-specialties.

[24] Contentful: Content Infrastructure for Digital Teams [EB/OL]. [2018-05-16]. https://www.contentful.com/.

[25] Eladel A, Zaied M, Amar C B. Fast DCNN based on FWT, intelligent dropout and layer skipping for image retrieval [J]. Neural Netw, 2017, 95: 10-18.

[26] Carvalho J P, Rosa H, Brogueira G, et al. MISNIS: An Intelligent Platform for Twitter Topic Mining [J]. Expert Systems with Applications, 2017, 89: 374-388.

[27] 朱佳晖. 基于深度学习的主题建模方法研究[D]. 武汉: 武汉大学, 2017.

[28] Capuano N, Maio G R D, Ritrovato P, et al. Improving Access to Services through Intelligent Contents Adaptation: The SAPI Framework [J]. Lecture Notes in Computer Science, 2009, 5736: 248-258.

[29] Cho D E, Yeo S S, Kim S J. An Adaptive Intelligent Recommendation Scheme for Smart Learning Contents Management Systems [C]//Park J, Ng J Y, Jeong H Y, et al. Multimedia and Ubiquitous Engineering, Lecture Notes in Electrical Engineering. Dordrecht, Netherlands: Springer, 2013, 240: 57-66.

[30] Aguilar J, Diaz P V, Riofrio G. A general framework for intelli-

gent recommender systems [J]. Applied Computing and Informatics, 2017 13(2): 147-160.

[31] 黄立威,江碧涛,吕守业,等. 基于深度学习的推荐系统研究综述[J]. 计算机学报,2017,40:1-30.

[32] Leidig J P, Fox E A. Intelligent digital libraries and tailored services[J]. Journal of Intelligent Information Systems, 2014, 43(3): 463-480.

[33] 郭 栋,王 伟,曾国荪. 一种基于微服务架构的新型云件PaaS 平台[J]. 信息网络安全,2015(11): 15-20.

[34] Hitachi Content Platform Anywhere [EB/OL]. [2018-06-05]. https://www. hitachivantara. com/en-us/products/cloud-object-platform/content-platform-anywhere. html.

[35] 中科汇联-HuiLan Technology [EB/OL]. [2018-06-08]. http://www. huilan. com/web/cp/index. html.

[36] Pubsweet [EB/OL]. [2018-07-10]. https://pubsweet. org/.

[37] Nuxeo [EB/OL]. [2018-07-10]. https://www. nuxeo. com/.

[38] Gentics Mesh - The open source headless CMS for developers [EB/OL]. [2018-07-10]. https://getmesh. io/.