张成昱

数字化文献的知识解构研究*

摘　要　数字化文献解构模式及其理论是对于把固化在数字化文献中的知识对象化地解析成基于范畴和相关关联的本体概念体系的方法和理论的总和。知识解构理论还包括解构的逆过程,即知识的重构。参考文献 4。

关键词　知识管理　知识解构　数字化文献　知识对象　广义 Ontology 方法
分类号　G253

ABSTRACT　The deconstruction patterns of digitized documents and their theory is the sum of methods and theories for the deconstruction of knowledge objects in digitized documents into category-based and correlated ontological notion systems. The theory of knowledge deconstruction also includes the reverse process, i. e. the reconstruction of knowledge. 4 refs.

KEY WORDS　Knowledge management. Knowledge deconstruction. Digitized document. Knowledge object. Generalized ontology method.
CLASS NUMBER　G253

资源和服务是图书馆,乃至数字图书馆的两大核心要素,它们也是本文的核心研究对象。包含电子图书、电子期刊和各种文献型数据库在内的数字化文献构成了数字图书馆信息资源的主体;而数字化文献基于网络的传播则是数字图书馆的主要信息服务方式。

数字化文献是以计算机为基本信息载体,以网络为主要传播介质的文献类型,是数字化的知识对象(基本范畴)的集合。数字化文献解构模式及其理论是对于把固化在数字化文献中的知识对象化地解析成基本范畴和相关关联的本体概念体系的方法和理论的总和,这也是数字化文献有别于传统文献的重要特征之一。由于知识的解构是一个可逆的过程,所以知识解构理论还包括解构的逆过程,即知识的重构,知识传播的受众按自身个性化的需要把已经离析的知识要素或者知识对象重新组合成新的知识载体单位的过程,这也是一个数字化文献的再创造过程。

目前的数字化文献大多仍来源于传统文献,是通过对印刷型文献进行内容复制式地数字化制作而产生的。对于电子图书之类的数字化文献进行基于页面、段落和字符的物理标识性的解构,如基于目录和关键词的线索化功能以及基于字的全文检索功能已经比较成熟,成为各种文献服务系统必备的功能。而基于内容和概念的知识化解构还主要停留在基于分词的全文检索和基于简单语法和语义分析的自动摘要和索引方面。笔者在具体实现和应用数字化文献的知识解构和重构时选择了在人工智能领域被广泛用于知识表示的 Ontology 技术作为基本方法,并为了适应数字化文献相关研究的特殊需要而提出广义 Ontology 方法来试图使两者有机地结合起来。

1　面向内容层次结构的电子图书格式标准

电子图书是数字化文献中最受关注的类型之一,对它的内容组织结构的研究也随着各种应用系统和资源的出现而不断深入。这种研究的成果主要体现在各种电子图书格式和标准上。目前还没有统一的电子图书文件格式,不同的公司因版权或商业因素等原因,往往各行其事,采用不同的文件格式。国外常用的就有:PDF、OEBPS、微软的 LIT、韩国的 EBKS、日本的 JepaX。国内的方正、超星、书生、国图等公司也均有自己不同的电子图书格式。

美国的开放电子图书论坛在 1999 年底提出了开放电子图书出版物结构 Open eBook Publication Structure,简称 OEBPS,以规范电子图书的制作格式,实现电子图书阅读系统的互操作,它提供了电子图书的一种内容规范。OEBPS 是一种相对宽泛的、轻量级的规范。它定义了两个 XML DTD,即包括 DTD 和

* 本文得到北京市教育科学十五规划青年专项课题"基于分布式教育资源组织的网上大学生学术社区建设的应用研究"(CIA02354)资助。

基本 OEBPS 文档 DTD。其中，OEBPS 包文件（格式良好而且有效的 XML 文件）构成某个完整出版物的框架，阅读器系统利用它来查找和组织出版物的各个组成部分。基本 OEBPS 文档 DTD 则从形式上定义了本规范中所描述的 XHTML 子集。

OEBPS 出版物是若干 OEB 文档、一个包文件以及其他文件的集合。这里的其他文件通常表现为各种媒体类型，包括结构化文本和图形，它们共同构成该出版物中不可分割的组成部分。符合本规范的出版物中必须包含且仅包含一个 OEB 包文件，该文件可被用来指定 OEB 文档、图形以及组成 OEB 出版物的其他对象，还可用来指定它们彼此之间是如何相互关联的。

清华大学图书馆从 1998 年开始和超星公司合作，研究和开发的多层检索技术是以图像为信息展示方式，同时能够支持全文检索的技术。该技术是通过在整个电子图书系统各模块提供相应功能来实现的。多层检索技术是本系统克服纸介质出版物数字化过程效率和质量矛盾的核心技术之一。

多层检索技术对电子图书的信息空间使用了多种信息的表示方法，提供多种层次的信息服务。电子图书的逻辑结构被分为 3 个层次。

第一层：基于超星电子图书的图像格式的信息表示层。这一层是直接面向读者的，是用户获得信息的最直接方式。该层次的主要信息源，就是图像形式的电子图书。

第二层：基于 OCR 的全文信息应用层，利用全文可以对电子图书进行检索，实现对信息的二次利用。在本系统中，第二层信息主要包括电子图书 OCR 全文，以及相应的全文位置信息。

第三层：电子图书的描述层，即元数据层，主要为数据库内的电子图书的条目信息。

2 数字化文献知识解构的研究框架

2.1 研究对象和内容

对于数字化文献知识解构的研究要在总结现有数字图书馆系统应用和开发经验的基础上，建立用以阐释数字化文献的知识化过程之本质的知识解构和重构理论框架，使之体系化、完备化和实用化。与上述研究同步，建立一个基于现有 Ontology 技术和理论，并加以扩展的广义 Ontology 方法模型，并加以应用研究。

数字化文献知识解构的研究需要解决以下两个方面的问题。

首先是如何构建一个完备、全面并且具有实用性的数字化文献知识解构和重构理论体系，这包括：(1)作为数字化文献组成部分的知识对象的捕捉和规范问题。(2)知识对象集合和广义 Ontology 在应用中的对应关系，即如何利用广义 Ontology 的运算和操作，实现知识对象集合的应用。(3)知识对象重构和用户环境的关系问题。

其次是要解决实现数字化文献知识解构的方法论研究，就是对于所谓广义 Ontology 方法的研究：(1)附加 Ontology 的构造方法和内容与概念的变迁机制。(2)广义 Ontology 结构中接口的定义和实现算法。(3)广义 Ontology 方法研究中 ontology 内部的操作：更新、伸缩、重排的精确定义和系统实现算法的实现问题。(4)广义 Ontology 方法研究中 Ontologies 之间的运算：合并，融并，比对的精确定义和系统实现算法的实现问题。

2.2 研究途径和思路

数字化文献的研究要遵循这样的途径：确定并创建一个具有一定学科专指性的文献集合作为研究对象信息源；根据该信息源制订相关核心 Ontology 规范并创建一个实例；开发并运行一个针对上述信息源，符合基于上述 Ontology 的数字化文献远程教育系统；分析和研究系统服务性能，评估和完善数字化文献解构和重构理论和广义 Ontology 方法。

我们可以从以下几个角度开展对于数字化文献知识解构的研究。

(1)数字化文献应用系统的实证分析研究。电子图书和电子期刊是目前应用较为广泛，也比较成熟的数字化文献类型，在知识采集、组织和传播机制方面各具特点。它们都是以物理的册或卷为基本知识承载单位，以平面的字符文本信息为主要符号系统。前者侧重于知识传播的系统性，后者更侧重于知识传播的时效性。

(2)数字化文献的网络传播特性研究。数字化文献的存在状态和它的传播状态是有区别的，知识解构过程其实是在数字化文献进行传播时才开始进行，并与传播过程基本同步。这意味着尽管人们可以以最利于知识解构的方式在数字化文献中描述和保存知识，但附着在数字化文献上的知识的标记永远不可能是知识本身，而它们成为知识本身的过程就是知识被传播的过程。

(3)基于数字化文献的虚拟教学环境建设的应

用研究。以网上课堂和远程教育为实例的虚似教学环境中对于数字化文献的需要具有特殊性和迫切性。传统教学的"教"是以教师的传授为主,而虚拟教学的"教"则是由教师的远程传授和教学文献的主动推送相结合而实现,后者起到越来越重要的作用。我们侧重于研究数字化文献中知识要素和教师授课内容的实时关联机制、网上课堂中基于数字化文献的交互功能的实现以及如何通过数字化文献的知识重构和主动推送实施个性化教育。

2.3 研究的可行性

数字化文献具有知识化的载体特征,为知识管理提供了基础。在数字化文献中,知识标识的线索化方式不被局限在线性空间里。知识要素之间可以随意建立起系统可以操作的链接,这就使数字化文献可以按知识本来的概念-关联模型来加以组织和处理。

知识解构和重构理论具有一定的研究基础和广泛的应用背景。对于数字化文献的组织、访问和发布已经有了成熟的研究基础,基于字和词的文本全文索引模式已经具有对数字化文献进行知识解构的雏形。网络的发展为数字化文献的应用提供了便捷的条件。对于某些特殊对象,如图像、表格、多媒体等的描述也得到相应元数据的支持,这些使得以知识对象为基本元素的数字化文献的解构成为可能。

广义 Ontology 方法是对 Ontology 技术的合乎逻辑的拓宽和发展。Ontology 技术在被用来作为数字图书馆领域的研究手段时,和原有图书情报学常见的主题词表等具有很多相通之处,很容易把相关的理论和方法移植过来。而作为本文提出的广义 Ontology 方法研究则牢牢建筑在原有 Ontology 技术的基础上,只是更适合于本研究的应用环境。

3 知识解构:内容对象向知识对象的嬗变

知识解构理论是一个论述如何把附着在数字化文献上的静止的知识信息解析成具有足够颗粒度的知识要素及其相互关系的集合的方法和理论。数字化文献是一种知识的间接载体,它直接承载的其实是以某种格式记录知识的符号序列,这个符号序列是由某个特定的符号语言体系严格加以定义和规范的。因此对于数字化文献的知识解构要分为两个步骤加以研究:一是对特定数字化文献集合,进行基于相应符号语言的文本解析(这个"文本"不是狭义的字符文本概念);二是对由符号进行标记和描述的知识本体进行提取、确认和组织,并分析和揭示这些知识本体之间的关系。

这两个步骤同时代表了数字化文献研究的两个层次:前者是针对标记的研究,以符号作为处理对象,比较适于人工智能的知识解构机制的实现;后者则是针对标记所指的研究,是对更接近于知识本质的对象进行的。

在针对数字化文献的知识解构模式的研究中,必须定义、规范和分析在知识解构过程中起到重要作用的几个过程:

(1)在数字化文献中,不仅需要对一个知识对象的物理位置进行确定,还需要对它在文献内容中的逻辑位置进行确定。比如对于一个数学公式对象,不仅由符号、函数和字母组成的数学式本身属于这个对象,该公式附近说明、限制和解释该公式的文本信息也应该属于这个对象,如对于自变量定义域的说明,变量物理意义的解释,公式使用条件等。

(2)在数字化文献中,知识对象的定义、范畴、语义表达方式和属性结构都必须满足结构化和规范化要求。这一方面体现在知识对象的各个要素都由相应 Ontology 中的规范词表所控制,还体现在它具有充分的机器可理解性。这将为软件系统的开发提供可操作性。

(3)在数字化文献中,如何界定一个知识对象的概念范畴对对象的解析极为重要。在静态 Ontology 的框架中,知识对象的范畴刚在系统中生成就被固化了。而在广义 Ontology 中,它受在内容和语义上相邻的对象,或者说语境所影响,在不断伸缩和变化着。这同时也决定了对它的限定应该是一个基于模糊集的概率模型。

(4)在数字化文献中,各个知识对象最终是以一个相对独立的形态存在着的,或者说它们可以脱离原来所属的物理载体,单独被用户或系统所访问、编辑和描述。然而这个离散的过程必须保留对象之间的各种相关关联关系,并依据这些关联关系确定在新的对象集合中,相互之间的位置和因果关系。离散是知识解构的主要过程。

在针对数字化文献的知识重构模式的研究中,则必须定义、规范和分析在知识重构过程中起到重要作用的几个过程:

(1)在应用系统中,知识对象的相互关系往往通过比较来获得。两个知识对象的比较可以归结为两个知识对象的属性集合之间的相似性计算。相似度可以是一个数值,也可以是一个数值向量,这取决于

系统如何定义两个知识对象的同一关系和如何描述两个知识对象的不同程度。

(2)在应用系统中,知识对象的关联不仅来自知识对象被解构的过程,还可能来自知识对象在系统中被使用的过程。关联的继承、更新、转移和发散,和基于新的关联网络而对于整个知识对象集合互动机制的调整都通过广义 Ontology 的关联运算而实现。

(3)在异构的应用系统之间,知识对象可以通过基于元数据的互操作机制实现知识的传播。一个知识对象在传播的目标系统中不仅要被辨识和理解,还要有机地融合到目标系统的知识体系中,与目标系统中原有的知识对象建立有效关联。

(4)来自不同系统的相似对象可以在传播的目标语境中合为一个新的对象。相同的属性在新的系统中重新得到统一的定义,相似的属性被修正成同一属性,不同的属性根据系统需要而增加或删除。知识对象的归并是知识对象集合的归并基础。

4 知识解构的工具:广义 Ontology 方法

Ontology 技术为实现对于数字化文献的知识解构提供了有效方法。Ontology 是人工智能研究领域常用的知识表示方法,在信息资源建设中也具有重要作用[1~2]。我们在把 Ontology 技术应用到数字化文献的解构和重构中的同时,还对 Ontology 的应用方式进行方法论研究,提出广义 Ontology 方法作为 Ontology 技术在该领域推广应用的理论基础。

Ontology 最精练的定义是由 Gruber 所给的:对某种概念化体系的规范说明。它包含了两层含义:一个是对某个领域进行抽象和归纳,也就是实现一个概念化的过程;另一个则是如何对这个概念化结果的表达。Ontology 本身就具有强烈的方法论色彩。

我们在这里提出的广义 Ontology 方法就是在 Ontology 的方法论基础上进行了适应其应用环境的扩展和演绎。对于广义 Ontology 方法的研究主要包括以下几个部分:

(1)广义 Ontology 除了包含概念、关系、函数和公理之外,还增加用于 ontologies 之间进行各种运算的接口描述信息。接口是一个包含异构 ontologies 中概念限制条件和关联关系的结构化的数据集合,包含用于互操作的元数据和概念间的对应关系和相似度信息等。广义 Ontology 在结构上的扩展是为了适应源自数字化文献的知识对象集合之间相互作用的需要。

(2)广义 Ontology 不是严格面向领域的,而是在很大程度上面向某些特定的信息载体的,因此在本课题中为叙述方便可以分为两个大类:一个是面向领域和内容的核心 Ontology,另一个是面向载体和表达的附加 Ontology。两者在结构上相似,但在概念(类)的含义和范畴上有很大差异。后者是广义 Ontology 动态性的实现基础。

(3)核心 Ontology 是领域内通过对概念和关系的收集、整理和规范,加以精确地定义,并加以编码后生成的。同时也可以利用已有同类 Ontology 加以集成和复用。附加 Ontology 则是针对具体文献载体或者文献载体的集合,在相应文献的范围内,收集、整理和规范概念和关系,并在应用中加以修正和精确化。附加 Ontology 还可以是其他附加 Ontology 运算的产物。

(4)在单个 Ontology 内部对更新、伸缩、重排等操作进行定义,这些操作被用于广义 Ontology 在整个系统应用过程中的进化。

(5)在两个或多个 Ontology 之间定义合并、融并、比对等运算,这些运算被用于广义 Ontology 之间各种导致它们发生变化的相互作用的实现。

5 知识解构理论的应用前景

5.1 促进基于互联网的虚拟教育环境建设

以数字化文献为主体的数字化教育信息资源是虚拟教学环境的重要组成部分,也是最能体现数字化环境在教育过程中作用的因素[3]。在传统教学模式下,文献的使用是和课堂教学在很大程度上相分离甚至相矛盾的;而在虚拟教学环境中,由于数学化文献的内容获取的便利性,以及包含其知识要素的各个数字对象可以有机地和整个教育信息发布系统融合在一起,它可以大大促进虚拟教学的质量,使得可以最大限度体现分布式、交互性和开放化的网络教育成为可能。

5.2 对数字图书馆、e-learning 和电子商务等应用系统开发的影响

知识解构理论可以成为数字图书馆等与信息资源服务和传播有关的应用系统开发的重要理论基础。

目前上述各类系统在对数字化文献进行组织、使用和交易时往往只能基于一个相对比较粗糙的物理载体,比如一本图书或一篇论文等,而无法针对承载知识的实际数字对象进行操作,这不利于提高数字化文献在承载和传播知识方面的效率和质量[4]。造成

这个问题的根本原因就在于数字化文献本身的知识化程度有很大差别，只有当数字化文献的格式可以把它所承载的知识信息充分线索化，使所有知识要素及其相互关系得以揭示，数字化文献以及与之相关的应用系统才能够真正发挥数字化和网络化的威力。

而知识解构理论就是这样一个论述如何把附着在数字化文献上的静止的知识信息解析成具有足够颗粒度的知识要素及其相互关系的集合的方法和理论。

5.3 对网络传播媒介的知识发布效果的影响

目前的网络传播中，许多信息源的主体是由数字化文献构成的，尤其是那些由传统媒体如报纸、图书和期刊等通过介质数字化而获得的资源。这些数字化文献在传播过程中如何发挥其传播效果，或者说如何最大限度地扩大相关信息在特定媒体领域中的影响范围和力度，很大程度上取决于数字化文献本身是否符合网络这个传播媒体的特征。

利用数字化文献的知识解构理论，一方面可以从传播学的层面对数字化文献的组织方式加以优化，并在数字图书馆相关应用研究中体现传播学的特色，另一方面把数字化文献的知识解构方法应用到远远超出图书馆范围的所有传播领域，对于目前一些传统媒体的“上网”提供重要的理论指导。这有助于目前应用很广的电子期刊、电子报纸和电子图书等摆脱所面临的无法突破相应的传统媒介的模式，而难以独立发展的困境。

6 结论

对于数字化文献知识解构的研究体现出以下几个方面的创新。

首先是理论抽象上的创新。知识解构和重构理论可以揭示数字化文献的传播本质，它不仅可以把传统图书馆学的分类、编目和主题等理论和方法概括到自己的理论框架中，更重要的，它可以成为解析数字文献所承载的物理内容（即符号）和逻辑内容（即符号所指代的知识）之间的外在和内在联系的方法论基础。

其次是理论应用上的创新。通过数字化文献的知识解构，以往文献与教学的相互作用上升到数字化文献和虚拟教学环境之间融合无间的有机统一体。它改变了传统教学模式中，文献只能起到辅助和边缘作用的状况，为各种网络教育系统开发中实现对数字化文献的充分利用提供了可靠的理论基础。

最后是对于Ontology方法的扩展和创新。广义Ontology在原有基础上进行了创造性扩展，提出了广义Ontology方法的概念和基本内容，并把它应用到以互联网为主要传播媒介的数字化信息资源的组织结构和获取方式的研究中，这有助于解决非结构化信息资源在互联网环境中信息定位问题。

参考文献

1 耿方萍，朱祥华. 基于本体的网络资源表示研究. 计算机应用，2003，23(4)
2 楼向英. Ontology：概念及其在数字图书馆中的应用. 图书馆杂志，2002，21(11)
3 Rusch-Feja D. The Open Archives Initiative and the OAI Protocol for Metadata Harvesting: Rapidly forming a new tier in the scholarly communication infrastructure. Learned Publishing, 2002,15(3)
4 Harnad S. Electronic preprints and postprints. Encyclopedia of Library and Information Science, Marcel Dekker, Inc, http://www. ecs. soton. ac. uk/ ~ harnad/Temp/eprints. htm. 2003.

张成昱 北京大学信息管理系博士研究生，清华大学图书馆系统部副研究馆员。通信地址：北京清华大学。邮编100084。
（来稿时间：2004－09－30）

关于来稿中“参考文献”著录的两点说明

请向本刊投稿的作者，除要严格按照国家标准《文后参考文献著录规则》对参考文献予以著录外，还应注意以下事项：

1. 凡文后有参考文献的论文，应在论文正文的相应位置，用角注形式标出参考文献的序号。

2. 参考文献如果是网上文献，著录项目不应省略责任者和文献题名；并且不能只著录网站主页，而应尽可能著录文献所在页面，并加注作者从网上查得该文献的日期。

例如：

包冉. 国家网格在路上——2003网格技术与应用研讨会侧记. http://www2. ccw. com. cn/04/0402/b/0402b17-1. asp（查询于2003－03－11）