基于细粒度聚合单元元数据的书目资源聚合研究

卫宇辉

摘　要　随着书目资源中多源异构数据的增多,信息片段呈现出扩散分布的特征,这增加了用户获取目标资源的难度、降低了信息检索效率,而聚合细粒度信息资源、构建信息单元之间的关联关系对于知识服务效率的提升具有重要作用。本文通过对聚合单元进行细粒度划分,根据语篇单元、句群单元、节段单元的属性特征及DC、LOM元数据,构建了细粒度聚合单元元数据框架;从聚合单元知识组织体系出发,结合细粒度聚合环境下聚合单元之间的关联关系,构建了基于聚合单元元数据框架的细粒度信息语义组织模型,并基于该模型提出书目关系扩展及映射方案;通过分析聚合单元元数据框架下的书目数据聚合层次及其他书目框架实体之间的关联,实现了书目元数据细粒度与揭示内容的细化、扩展,为细粒度书目资源聚合及检索提供了理论基础。图7。表4。参考文献23。

关键词　书目资源　细粒度　聚合单元　元数据

Research on Bibliographic Resource Aggregation Based on Fine-Grained Aggregation Unit Metadata

Wei Yuhui

Abstract: With the increase of multi-source heterogeneous data in bibliographic resources, information fragments show the characteristics of diffusion distribution, which increases the difficulty for users to obtain target resources and reduces the efficiency of information retrieval. Aggregation of fine-grained information resources and construction of association relationships between information units play an important role in improving knowledge service efficiency. Based on the fine-grained division of aggregation units and the attribute characteristics of discourse units, sentence group units and segment units, as well as DC and LOM metadata, this paper constructs a fine-grained aggregation unit metadata framework. Starting from the knowledge organization system of aggregation units and combining the association relationships among aggregation units in fine-grained aggregation environment, a fine-grained information semantic organization model based on metadata framework of aggregation units is constructed, and a bibliographic relationship expansion and mapping scheme is proposed based on the model. By analyzing the aggregation level of bibliographic data under the aggregation unit metadata framework and the relationships among other bibliographic framework entities, the refinement and expansion of fine-grained bibliographic metadata and revealed content are realized, which provides a theoretical basis for aggregation and retrieval of fine-grained bibliographic resources. 7 figs. 4 tabs. 23 refs.

Keywords: Bibliographic Resources; Fine Granularity; Aggregation Unit; Metadata

网络文献作为各类文章的载体,聚集了许多关联度较低的学术信息资源,利用传统的数字化手段无法形成规范化的知识系统,导致学术信息检索困难、精确度不高[1]。随着网络信息资源的快速更新、文献实体内容的不断变更以及文献数据库的逐渐丰富,如果缺乏对文献编目数据的及时更新,则会直接影响文献检索的精准度。书目关系是一种能够描述信息资源形态特征和内容特征的书目记录间关系,是促进信息资源内容深化、结构序化的主要途径[2]。因此,建立规范化的书目数据、挖掘书目之间的关系是实现文献书目自动更新的重要基础,对于文献资源检索、读者服务、文献资源建设具有重大意义。

目前,信息资源聚合作为知识服务领域中的重要基础,已成为国内外信息组织与检索领域探讨的热门话题,国内外学者开展了大量相关研究

DOI:10.13666/j.cnki.jnlc.2020.0609

并取得了显著成果,例如信息资源聚合的概念及相关理论[3,4]、信息资源聚合的效果评估[5,6]、信息资源聚合手段和技术[7,8]、信息资源聚合的应用[9,10]等。而实现网络信息资源聚合的关键问题在于从细粒度层面深入挖掘信息资源之间的关联及特征,现有的细粒度网络学术资源研究主要集中在关联数据[11-13]、知识元[14,15]、粒度划分[16,17]等层面,这些研究为细粒度网络学术资源的抽取、识别与关联分析奠定了理论基础。但关于细粒度网络学术资源的划分研究侧重于从形式结构的角度出发来构建元数据框架[18],基于逻辑结构划分细粒度网络学术资源的研究较少,同时也缺乏相应的元数据描述标准。在专题数据库开发中,书目的著录会以资源类别(比如图书以种类区分、期刊以刊名区分)为最小单元,无法充分揭示书目的内容特征、学术价值和发挥专题数据库的文献整合利用功能,易于造成数据冗余。而基于逻辑结构划分细粒度的网络学术资源,则会遵循"有利于检索发现,有利于读者使用,有利于提高开发效率"原则,根据实际情况选择"章节"或者"篇"为最小著录单元,然后再以逻辑关系进行组配,实现细粒度书目数据的自动化整合。因此,本文对以下问题进行了探索:(1)如何基于逻辑结构和形式结构划分细粒度聚合单元;(2)如何构建反映多类型网络文献资源信息单元层级的信息组织框架;(3)如何定义信息组织框架中的知识概念、关系,揭示聚合单元与实体之间的关联;(4)细粒度网络资源聚合模式下的检索效果如何。针对这些问题,本文通过设计细粒度聚合单元元数据框架,建立了聚合单元元数据框架下的细粒度信息组织模型并基于元数据框架构建了书目扩展关系的映射方案,实现细粒度书目数据的聚合,通过设计检索任务检验效果,为细粒度聚合单元环境下的书目分析提供了一定的理论基础。

1 聚合单元划分的依据与方法

目前,大多数元数据方案主要基于书目关系来实现对文献资源的聚合与检索,集中于对可检索书目资源的揭示与关联,较少关注文献资源实体及其内容组织结构。结合文献资源的内容组织与知识单元,本文以文献资源的逻辑结构、形式结构及不同知识单元之间的关系为依据,划分出不同层级的聚合单元,拓展元数据设计方案,丰富书目资源聚合与检索方式,从而实现基于书目关系与文献知识单元的书目资源聚合功能。

1.1 逻辑结构分析

(1)节段单元

节段单元是指根据文献框架与逻辑思路对文章内容进行划分得到的结果[19]。划分文献节段单元的价值体现在三个方面:(1)让读者能够根据文献各级标题了解全文的组织结构,从而判断该文献与自身实际需求是否相符;(2)帮助读者快速检索所需的段落内容并进行精确定位,节省信息资源查找时间;(3)用文献各级标题的关键词来描述节段单元主题,有利于文献主题聚合功能的实现。

(2)句群单元

句群单元则是指具备修辞目的的语篇结构。信息资源句群单元的划分以体裁和体裁分析为主要依据,体裁是指社会交际活动的分类,例如学术论文、新闻报道、法律文件等;体裁分析是指从体裁角度出发,通过深层解析特定语篇的微观结构和宏观结构来掌握语篇的特定认知结构。由于不同类型文献体裁的分析结果各不相同,必须综合语篇的交际功能与话语意图进行体裁分析。本文选取开源期刊论文进行体裁分析,以 CARS 模型为基础[20],进行句群单元划分。Swales 在 1990 年提出引言结构分析模型,即 CARS(Create a Research Space)模型,包括确定研究领域、确定研究定位、把握研究契机 3 个语步(move),以及相应语步的步骤(step)。语步是作者写作目的的总体概况,步骤是为实现语步目的的详细描述。以《结合地理信息的引文分析研究现状》(鲁超、刘清,《情报科学》2011 年第 2 期)为例,该文属于非实证型,论文组件包括介绍、理论分析、论证及结

论,介绍部分的语轮/语步划分结果如表 1 所示。

表 1　语轮/语步划分结果

语轮	语步	句群单元
语轮 1:提出研究领域的主题	语步 1:阐述定义	引文分析就是……对各种关联和相关关系进行分析研究
	语步 2:归纳问题的客观知识	以便揭示其数量特征和对象间内在的规律,……预测科学发展趋势等。
	语步 3:介绍研究对象的产生及发展过程	普赖斯于 1956 年发表的著作……期刊共引、主题共引和类的共引等。
	语步 4:收窄论题	期刊共引可以运用到期刊……利用共引理论来探讨科学范式。
语轮 2:提出已有研究的不足	语步 1:论述开展研究的理由	在引文思想产生和发展的过程中……使得引文分析具备了实用工具。
	语步 2:阐述以往研究的贡献	1998 年美国情报科学研究所……推动了基于引文的文献计量方法的应用。

1.2　形式结构分析

文献的形式结构包括摘要、图标、正文及参考文献等组成要素。本研究通过分析文献的形式结构对不同组成要素进行拆分,并将反映文献外部特征的要素(摘要、标题、作者、关键词、机构、参考文献等)作为文献元数据信息,通过解析论文正文部分,对图表、句群进行抽取,经过逻辑结构分析后得到由句群单元、节段单元组成的细粒度聚合单元。文献中的图表通常概括了全文的重点研究内容,是对文献主要观点的形象描述,图表的提取对于文献资源聚合及检索具有重要作用。但对图表单元必须给予相应的描述以便于用户理解,可以将主题明显的、能够解释图表的句群单元与图表单元标题进行关联匹配,从而为图表提供相应的情境信息。

1.3　不同层级聚合单元之间的关系

综合文献的逻辑分析及形式分析结果可知,语篇单元、句群单元、图表单元及节段单元共同构成了细粒度网络文献资源的聚合单元,这些分布于不同层级的聚合单元之间均存在一定关系[21],篇章单元包含图表单元、节段单元和功能单元,且都是一对多的关系:图表单元需要篇章单元和具有相对完整意义的相关句群单元进行解释。因此,图表单元需要与提及该图或表的句群单元相关联,由于可能存在不止一个句群单元提及图或表的情况,句群单元也可能不只提到一个图或表,所以图表单元与功能单元是多对多的关系:从形式结构上看,句群单元包含于节段单元之中。节段单元与句群单元是一对多的关系,节段单元可以指示句群单元所在的物理和逻辑结构位置。如图 1 所示。

2　聚合单元元数据框架设计

2.1　聚合单元属性特征

本研究中聚合单元属性特征及其包含的元素分别有复用 DC 元数据元素、LOM 元数据元素以及新增元素:(1)复用 LOM 元数据。LOM 元数据中的粗粒度聚合单元能够重新组合、复用,符合本文的元数据研究目的。(2)复用 DC 元数据。本研究包括细粒度聚合单元和粗粒度聚合单元,

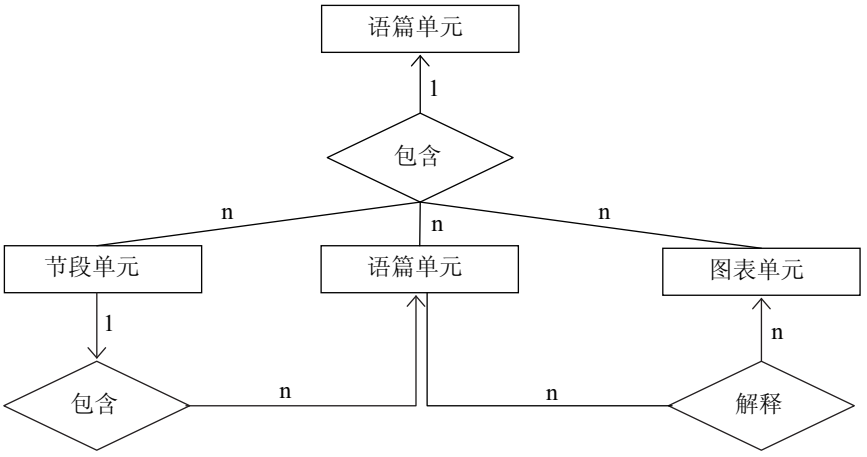

图 1　不同层级聚合单元的关系

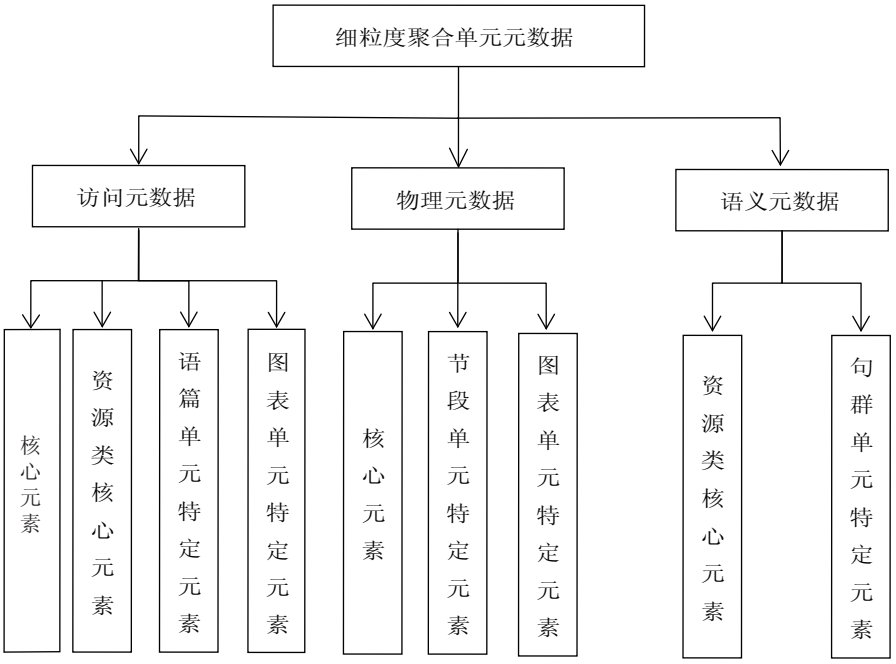

图 2　细粒度聚合单元元数据框架

复用 DC 元数据具备较好的可移植性和访问属性。(3)新增元素。为充分描述聚合单元的属性特征还应对特征独特的新增元素进行分析。语篇单元的新增元素包括“体裁类型”“相关信息”，期刊论文按照理论、实证、综述进行分类能够帮助用户查找其所需的文献资料，因此需要增加“体裁类型”元素。另外，新增“聚合层级”元素来描述句群单元与节段单元所处的层级位置，以揭示不同细粒度聚合单元之间的关联，从而实现细粒度元数据聚合。

2.2　聚合单元元数据框架

语义元数据、访问元数据及物理元数据共同

组成元数据框架(见图2),分别描述聚合单元的内容特征、外部特征及物理形态。访问元数据核心元素包括来源、标识、关键词,语篇单元特定元素包括作者、时间、分类、语言类型、资源类型、相关信息、体裁类型,图表单元特定元素为描述,资源类核心元素为标题;语义元数据资源类核心元素为话语意图,句群单元特定元素为语义功能;物理元数据核心元素包括存储位置、聚合层次,图表单元特定元素为图表类型,节段单元特定元素为节段单元层级。

下面以各类元数据的具体某一元素为例介绍其对应的著录方式,具体内容如表2所示。

表2 元数据著录方式

元数据名称	来源	聚合层级	语义功能
标签	Source	Aggregated Level	Semantic Function
定义	聚合单元的来源	聚合单元物理粒度层级	对句群单元内容的语义功能描述
注释	著录来源信息	著录聚合单元的粒度层级	著录反映句群单元内容语义功能的受控词汇
元素修饰词	资源来源;篇章单元来源;节段单元来源	无	无
著录细则	著录网络信息资源的来源信息、指定语篇单元唯一标识符、节段单元标识符	备选项(语篇、句群、节段、图表),分别对应语篇单元、句群单元、节段单元及图表单元	利用体裁微观分析结果作为受控词汇进行著录

3 基于聚合单元元数据框架的细粒度信息聚合设计

3.1 基于聚合单元元数据框架的知识组织模型设计

本文基于支持知识发现的聚合单元元数据框架设计了细粒度信息聚合的知识组织框架,如图3所示。该模型主要包括五个步骤:(1)资源采集与预处理。采集信息资源的主题及非主题特征并对其进行规范性描述。(2)识别主题与聚合单元。识别不同粒度聚合单元的主题,根据体裁分析结果划分聚合单元。(3)构建聚合单元本体。构建用于聚合处理和语义描述的知识体系。(4)资源描述。根据聚合单元本体识别聚合单元语义并进行标注,形成多维复合的语义概念。(5)聚合与呈现。将与用户需求语义相匹配的聚合单元进行重组,进行可视化呈现,实现交互功能。

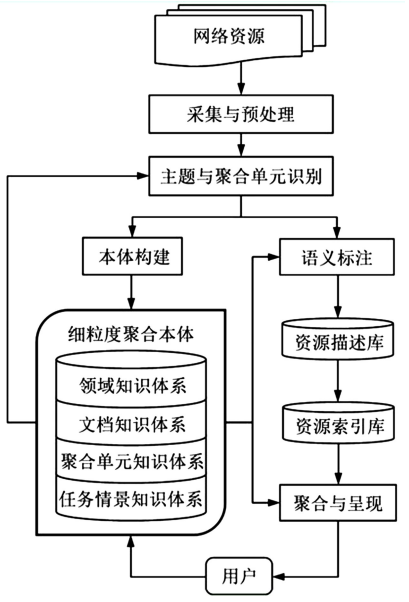

图3 基于聚合单元元数据框架的知识组织模型

3.2 基于聚合单元元数据的标注

在细粒度聚合单元元数据框架下,聚合单元元数据是信息组织的基本单元。在细粒度聚合过程中,元数据是描述各层级聚合单元、关联聚合单元、揭示文档粒度属性的重要工具,对聚合单元元数据进行标注是实现细粒度信息聚合的重要基

础[22]。在对不同层级聚合单元进行标注时,根据细粒度聚合单元本体所属层级构建数据库表,并结合语义元数据、物理元数据以及访问元数据的属性特征设置相应的字段。同时结合语篇单元、句群单元、节段单元对应数据库表之间的关联,确定表与表之间的关系,如图 4 所示。

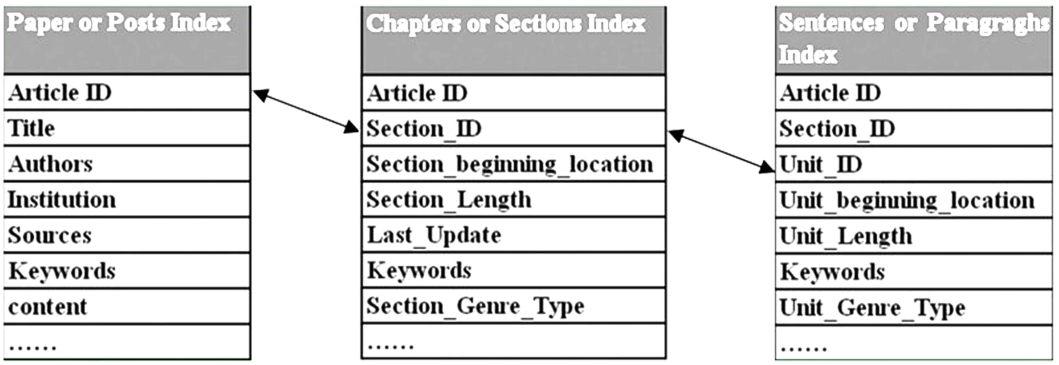

图 4　不同层级聚合单元标注与索引数据表

聚合单元元数据通常采用 XML 技术进行标注,利用由陈述、资源、属性组成的 RDF 数据模型来描述元数据信息,RDF 数据模型除了具备语义互操作功能外,还能在与元数据交换过程中保持其语义不变[23]。在划分不同层级的细粒度聚合单元后,本文根据聚合单元本体实现语义标注,按照聚合单元层级组织相互关联的细粒度元数据来形成知识体系,为文档检索奠定基础。

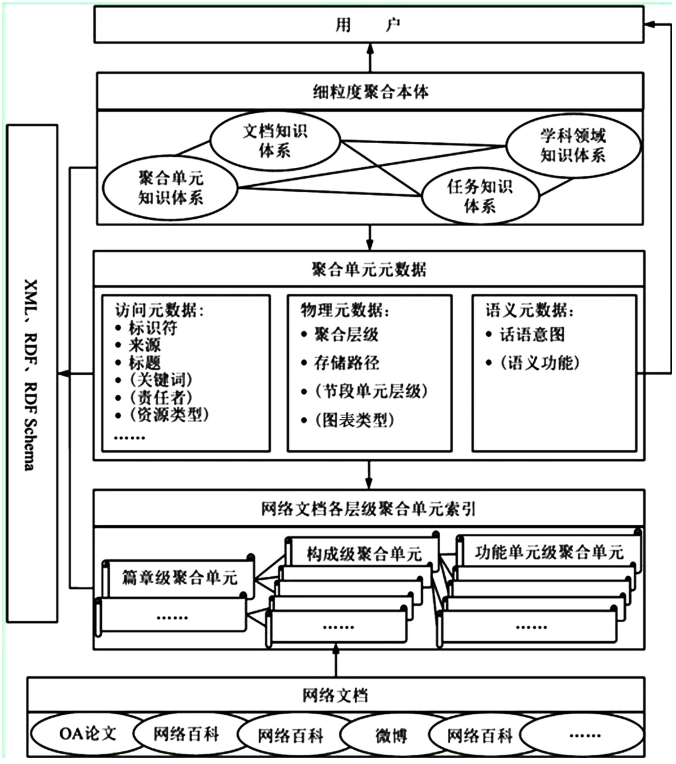

图 5　基于聚合单元元数据框架的细粒度信息语义组织模型

3.3 基于聚合单元元数据框架的细粒度信息语义组织模型设计

基于聚合单元元数据框架及其知识组织模型,以及细粒度聚合环境下聚合本体的语义聚合原理,本文构建了细粒度信息语义组织模型,如图5所示。在细粒度信息语义组织模型中,将网络信息资源划分为不同层级的聚合单元后,通过聚合单元属性提取、元数据标注及索引来构建细粒度聚合本体,将具有语义关联的单元聚合在一起,形成丰富的复合本体,为用户提供实现语义关联的网络信息资源。

3.4 基于聚合单元元数据框架的细粒度信息语义组织模型的书目关系扩展

(1)基于元数据的书目关系扩展

为解决现有元数据方案在书目关系扩展上的局限性,本文将实体资源划分为资源、主题、人、机构以及地点等类别,各个实体均具有自身属性和属性值,例如人的属性包括性别、姓名、出生日期、国籍、所属机构、作品等;地点的属性包括城镇、地区、国家等;机构的属性包括地区、作品、员工等;书目资源的属性包括标题、名称、出版社、类型、责任者、语言、版权、标识号、来源等;主题的属性包括责任者、概念外链、上/下位主题等。对实体及其属性的划分有利于区分各类资源的表现形式,实现具有相同属性的实体之间的关联。

根据书目关系中的实体及其属性,可建立基于聚合单元元数据的书目关系扩展框架。通过对实体关系进行扩展,发现各个实体之间、实体属性之间、实体与实体属性之间均存在一定的关联,即不同层次的书目关系。基于元数据的书目关系分类体系(如表3所示)反映了各实体之间的关联,该关联是扩展书目关系的重要基础,有利于实现对实体与属性之间、实体属性之间的关系扩展,从而在书目检索过程中关联更多与检索条目相关的实体和属性。

(2)基于元数据的书目关系扩展映射

根据元数据框架扩展元数据书目关系后,需要设计相应的映射方案,以直观地呈现以关系为主线的资源描述结果。鉴于目前图书馆采用的书目编目方案以MARC格式为主,本文以CNMARC为例分析其扩展书目关系与字段的映射情况,表4介绍了部分实体之间的关系,反映了书目与机构、人、书目、地点、主题等实体之间的关系和字段对应情况。

表3 基于元数据的书目关系分类体系

类型	书目资源	人	机构	主题
书目资源	具备继承/描述/连续/等同/整体-局部等关系及共有特性	贡献/创立	贡献/出版/收藏	用于描述资源/是资源的属性
人	贡献/创立	等同/血缘关系/共有特性	人从属于机构关系/共有特性	人创建了主题/主题是人
机构	贡献/出版/收藏	人从属于机构关系/共有特性	继承关系/共有特性	机构创建了主题/主题是机构
主题	用于描述资源/是资源的属性	人创建了主题/主题是人	机构创建了主题/主题是机构	类同/整体-局部关系

表4　基于CNMARC的扩展书目关系及字段映射方案

类型	书目关系	字段表现
书目-书目	继承关系	A 200 $ a=B 200 $ a,A 205 $ a≠B 200 $ aOR A 200 $ a=B 200 $ d=> B 为继承关系,205 $ a、205 $ b 为版本说明,若为翻译关系,200 205 $ z 为语种说明 A 440 非空 OR A 441 非空 OR A 442 非空 OR A 443 非空 OR A 444 非空 OR A 445 非空=>A 与 A 440、A 441、A 442、A 443、A 444、A 445 位继承关系
	描述关系	300 $ a 作注释说明
	连续关系	A 304 非空=>参照 304 字段 OR A 440 非空=>A 是 A 440 的后刊 OR A 430 非空=>A 是 A 430 的前刊 记录头标区第 7 个字符="s"=>
	等同关系	A 200 $ a=B 200 $ a A 200 $ f=B 200 $ f A 200 $ b≠B 200 $ b=>OR B 为等同关系,200 $ b 为载体类型说明 452 非空=>A 与 452 $ a 为等同关系 记录头标区第 6 个字符说明资源类型
	整体-局部关系	A 410 非空 OR A 411 非空 OR A 436 非空 OR A 461 非空 OR A 463 非空 OR=> A 200 $ a 与 A 410 $ a、A 411、A 436、A 461、A 463 是整体局部关系
	共有特性	具有一个或多个等同字段 Eg:A 210 $ c= B 210 $ c=>A 与 B 为同一出版社出版发行
书目-人	主要责任关系	If A 701 $ a $ b 非空=> A 701 $ a $ b 是 A 的主要责任者, $ 4 是责任方式 else A 200 $ f 非空=>A 200 $ f 第一个逗号前为 A 的主要责任者
	主题关系	A 105 第 12 位字符≠y=> A 是传记
	次要责任关系	If A 702 $ a $ b 非空=> A 702 $ a $ b 是 A 的次要责任者, $ 4 是责任方式 else A 200 $ f 非空=>A 200 $ f 第一个逗号前为 A 的次要责任者
书目-机构	贡献关系	A 711 OR A 712 非空=> A 711,A 712 著录的机构对该书目实体做出贡献
	出版关系	A 210 $ c 非空=> A 210 $ c 是 A 的出版机构,出版地为 A 210 $ a
	收藏关系	A 905 $ a 非空=> A 馆藏单位是 905 $ a

续表

类型	书目关系	字段表现
书目-主题	主题关系	606 $a、606 $x、690 $a、330 $a
	其他概念关系	605 $a、605 $x
书目-地点	创建地	A 102 $a 是 A 的出版国(出版国用代码版国) A 210 $a 是 A 的出版城市
	主题关系	607 $a、607 $x
	其他地点	A 701 $e 非空=> A 701 $e 为 A 的发生地点

(3)实例分析

本文以 MARC 记录为例(如图 6 所示),根据 CNMARC 的扩展书目关系及字段映射方案分析其关系的层次与构成。通过解析 MARC 记录发现存在两个层级关系,分别为实体与属性之间的关系和实体与实体之间的关系。其中,“信息组织”的出版日期为“20040928”,该书目的标识符为“413 页”“ 7-04-015340-8”,分别体现了 200 $a 与 010 $a 、200 $d 之间、200 $a 与 210 $d 之间的关系,反映了实体与实体属性间的关系;“高等教育出版社”与“信息组织”、“戴维民”与“信息组织”分别存在出版关系和创作关系,“高等学校”作为“信息管理”的下位主题,分别体现了 200 $a 与 200 $f 、200 $a 与 210 $c 、690 $a 与 690 $x 之间的关系。

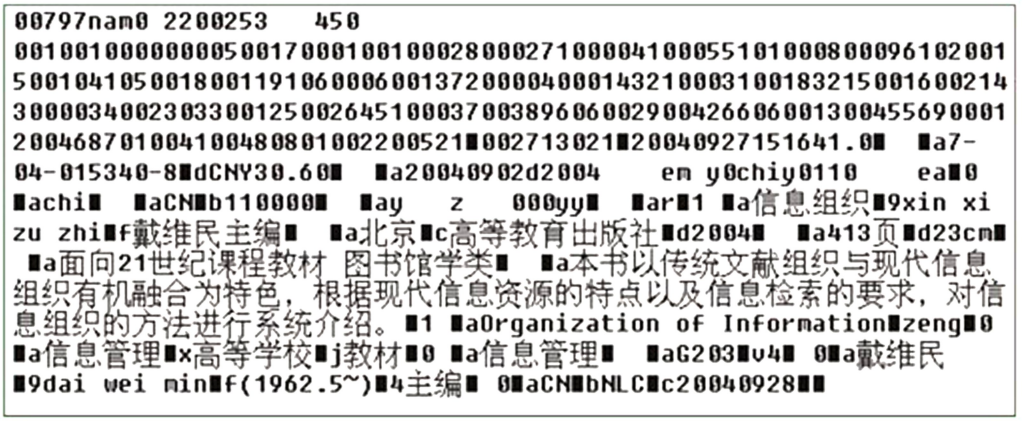

图 6 MARC 记录

4 聚合单元元数据框架下细粒度模型的书目数据聚合

4.1 聚合单元元数据框架下细粒度模型的书目数据聚合层次设计

书目数据聚合涉及信息层、数据层和知识层,书目数据的聚合机制反映了文献的内部特征及外部联系。从内部特征来看,不仅包括文献的题名及基本内容信息,还包括文献的转载信息、出版信息、收录情况等;从外部组织关系来看,反映了文献资源之间的从属关系、引用关系以及作者、机构、标题、内部主题和发行卷期等概念性内在关系。

细粒度聚合环境下,数据层实现对多源异构书目数据的整合,参考相关标准对文献元数据进

行设置、著录、标注和审校，从内部信息及外部关系来揭示文献特征，搜集期刊的影响因子、刊物信息描述、期刊收录情况、投稿指南、期刊分类信息等数据作为聚合的数据基础；信息层实现对书目数据的序化，通过建立元数据方案对信息资源进行规范化描述，揭示数据的外在关联及内部特征，该过程需要利用聚合本体或 RDA、MARC、DC、LOM 等元数据进行语义标注，实现书目数据的语义关联；知识层利用基于聚合单元元数据的细粒度知识组织体系来揭示实体之间的联系及本质特征。

在书目数据聚合过程中，首先利用分类法与叙词表划分期刊文献的细粒度聚合单元，建立基于聚合单元的元数据框架；然后，采用语义网技术对逻辑关系进行定义，根据不同层级聚合单元之间的逻辑关系深入描述元数据属性并进行语义规范；最后，结合关联数据建立文献关联，实现细粒度信息聚合。该过程实现了数据层、信息层与知识层的有效聚合，形成了发现知识的聚合本体，从而为文献资源的关联发现、语义检索及导航检索奠定了基础。

4.2 书目聚合层次与书目框架实体的关联

聚合单元元数据框架下细粒度信息语义组织模型通过以下结构层次来实现书目数据聚合，如图 7 所示。该聚合机制的原理在于：利用书目数据对期刊资源核心元素进行附注，比较分析期刊资源间的关联度与内部特征，从而实现期刊资源在不同信息层面的聚合。在现实应用中，可借助元数据关联技术实现多层级知识检索系统的设计与开发，这有利于资源获取与知识的自动发现，能够为信息检索与利用提供更便捷、快速的途径。

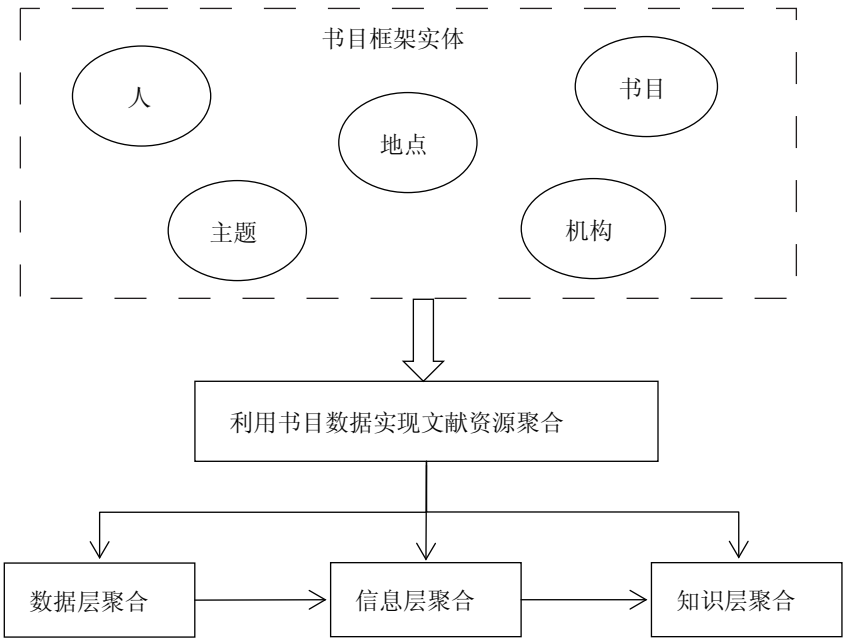

图 7　书目聚合层次与书目框架实体的关联

该聚合机制中包含两个层级结构：(1) 基于细粒度信息单元属性及语义关联建立文献资源数据。在实体-属性书目关系体系下，根据实体之间的关联及其自身属性来组织文献资源编目数据，对不同层级聚合单元的核心元素进行语义描述，建立规范化书目数据。期刊书目编目过程

中,对编目工作涉及的主题标引、版本项、发行项、附注项、文献题名、标准编号、载体形态项以及获得方式项进行描述与著录,重新编排书目数据,再根据这些内容分析期刊资源的再版、改名、流传等情况。(2)利用书目数据聚合文献资源。一是运用资源表征描述实现资源的数据层聚合,将书目数据置于更加广泛的网络环境中,使不同语言形式和包装形式的书目元数据都能在国际范围内展现,从而形成具有较强关联的元数据系统,为增强资源学科分类属性、促进信息层关联聚合奠定基础。二是借助元数据本体实现资源的信息层聚合,结合用户任务建立核心元素集合,从而提供目次表跳转、关键词检索以及跟随链接等服务。三是利用标准词表实现资源的知识层聚合,运用学科内容定制、学科知识索引、资源推荐与导航等元素,促进期刊资源在知识层面上的聚合,为用户提供语义检索、关联发现、文献索引等知识服务。

4.3 书目资源聚合模式下的检索示例

在传统的粗粒度文献检索模式下,由于书目资源之间的关联度较低,且内在内容逻辑联系分散,导致书目资源无序排列在文献中,检索主要通过选择对应类型的数据库来获取部分数据;而在细粒度聚合环境下,通过书目资源的关联聚合就能实现语义检索,精确获取目标数据。因此,在书目资源细粒度聚合模式下信息检索流程可细化为以下五步:

第一步,确定检索词汇。用户根据其所需资源选取适当的检索词汇,表达检索对象的主题、作者、标题、类型等信息。本文以“爱迪生”为检索词,分析细粒度信息聚合机制下的信息检索途径。第二步,识别检索点。该过程通过构建基于细粒度聚合单元的元数据标准,提高识别书目资源的可视化程度,对检索实体进行识别,筛选出具有相似特征的实体。由“爱迪生”确定检索实体为“Thomas Alva Edison”,中文名为托马斯·阿尔瓦·爱迪生,发明家、企业家,拥有四大发明。根据该描述实体,识别出“爱迪生的发明”“爱迪生人物故事介绍”“爱迪生发明与专利介绍”等类型的文章。第三步,选择目标源。根据用户需求选取载体、内容、来源均能符合其需求的资源。在细粒度信息聚合模式下,书目检索资源包含各种载体形式、出版形式的信息资源,因此对于“爱迪生”这一检索词,目标源可以设置为报纸、增刊、正刊、会议集等形式。第四步,获取资源。通过细粒度信息聚合模式实现书目数据的初步筛选,呈现相同或相似内容的所有资源,利用载体类型、出版社、发表时间、语言类型等检索条件缩小检索范围,实现书目资源的精确检索。第五步,浏览检索结果集。序化检索选定的书目资源,建立检索结果集中各实体之间的关联。例如,根据介绍“爱迪生”人物事迹的文章,关联出其他相关的书目及文章,由初始检索目标关联查找到书目 1,由书目 1 关联查找到人物 2,再由人物 2 关联查找到其他书目或文章,经过多次关联积累更多书目资源,扩展检索结果集的范围,为用户提供更全面、详实的信息资源。

5 结语

针对网络文献资源的细粒度聚合问题,本文根据逻辑结构和形式结构对聚合单元进行了细粒度划分,根据不同层级聚合单元的属性及关系特征建立了细粒度元数据方案,在此基础上对元数据进行语义标注和规范化著录,构建了聚合单元元数据框架下的细粒度信息语义组织模型。该模型通过对书目资源各类实体与数据层、信息层、知识层的聚合来实现对信息资源的分解、重组,从而实现检索系统的知识发现、语义检索等功能;通过深入挖掘并扩展书目资源之间的关联关系,为用户提供更全面、高效、便捷的知识服务。

参考文献

1 Trace C B, Dillon A. The evolution of the finding aid in the United States-from physical to digital document genre[J]. Archival Science, 2012, 12(4):501-519.

2 邱均平,王菲菲.基于共现与耦合的馆藏文献

资源深度聚合研究探析[J].中国图书馆学报,2013(3):25-33.

3 Santos R L T,et al.Aggregated search result diversification [C]//International Conference on Advances in Information Retrieval Theory.Berlin:Springer,2011:250-261.

4 张玉峰,何超.馆藏资源聚合结果的层次可视化方法研究[J].情报理论与实践,2013(8):41-44.

5 王学东,等.多模态网络主题资源聚合与实证研究[J].情报科学,2014(7):9-13.

6 Chuklin A,et al.A comparative analysis of interleaving methods for aggregated search[J].Acm Transactions on Information Systems,2015,33(5):5-46.

7 胡昌平,等.基于社会化群体作用的信息聚合服务[J].中国图书馆学报,2010(3):51-56.

8 马费成,等.基于关联数据的网络信息资源集成[J].情报杂志,2001(2):167-170,175.

9 Harting O,Langegger A.A database perspective on consuming linked data on the Web[J].Datenbank Spektrum,2010,10(2):57-66.

10 张小峰.基于关联数据的图书馆学术资源推荐研究[J].图书馆学研究,2012(5):87-89.

11 温有奎,焦玉英.基于范畴论的知识单元组织与检索研究[J].情报学报,2010(3):387-392.

12 姜永常,等.基于知识元的知识组织及其系统服务功能研究[J].情报理论与实践,2007(1):39-42.

13 郭少友,等.基于细粒度语义化描述的医学文本检索[J].情报理论与实践,2015(8):130-134.

14 丁培.科学文献与科学数据细粒度语义关联研究[J].图书馆论坛,2016(7):24-33.

15 刘平峰,等.基于模糊等价关系的文本多粒度划分方法[J].情报学报,2012(6):589-594.

16 陈勇跃,等.知识检索中的知识抽取与可视化研究[J].情报科学,2010(11):1719-1723.

17 李波.专题数据库开发中的文献粒度问题研究——以《桂西北少数民族历史、文化资源数据库》为例[J].新世纪图书馆,2014(6):57-60.

18 马翠嫦,等.网络学术文档细粒度关联与聚合的信息组织机制研究[J].现代情报,2019(12):37-45,54.

19 于晖.词汇衔接模式与语篇体裁分析[J].北京科技大学学报:社会科学版,2008(2):109-113.

20 肖珑,等.中文元数据标准框架及其应用[J].大学图书馆学报,2019(5):29-35.

21 曹树金,等.面向网络信息资源聚合搜索的细粒度聚合单元元数据研究[J].中国图书馆学报,2017(4):74-92.

22 欧石燕.面向关联数据的语义数字图书馆资源描述与组织框架设计与实现[J].中国图书馆学报,2012(6):58-71.

23 金华.基于书目框架的期刊元数据语义聚合探究[J].图书馆工作与研究,2019(9):55-60.

(卫宇辉　馆员　中国版本图书馆)

收稿日期:2020-06-15