目录学思想在数据结构化过程的传承与应用

Inheritance and Application of Bibliographic Mechanism in the Structuring Process of Unstructured Data

彭贤哲 郑建明 李佳新 石进
PENG Xianzhe ZHENG Jianming LI Jiaxin SHI Jin

（南京大学信息管理学院，南京，210023 / School of Information Management, Nanjing University, Nanjing, 210023）

摘要：[目的/意义] 信息资源时代下，数据类型多元化特征显著，透析数据结构化过程中蕴含的目录学思想，有助于解决非结构化数据管理与利用的难题。**[研究设计/方法]** 首先辨析数据结构化的本质过程，并揭示其中蕴含的目录学机理和标引分类思想，说明用目录学思想指导数据结构化过程的可行性，并借由目录工作运用的文献揭示、书目索引编纂、文献标引分类、文献组织等传统方法，解析不同类型非结构化数据的特点，指导其关联整合、索引指示、标引分类、组织重构等主要结构化过程，最终实现非结构化数据的“辨章学术、考镜源流”。**[结论/发现]** 数据结构化基本承袭了以分类标引等为核心的书目思想，在本质上是作为致用之学的目录学在当下环境的延续和发扬。**[创新/价值]** 有助于制定数据结构化过程的范式流程，增强非结构化数据结构化解析过程的复用性。

关键词： 目录学；非结构化数据；数据分类标引；数据组织

中图分类号：G250.7 **DOI：**10.13366/j.dik.2024.01.080

引用本文： 彭贤哲，郑建明，李佳新，等．目录学思想在数据结构化过程的传承与应用 [J]. 图书情报知识，2024，41(1)：80-91.(Peng Xianzhe, Zheng Jianming, Li Jiaxin, et al.Inheritance and Application of Bibliographic Mechanism in the Structuring Process of Unstructured Data[J]. Documentation, Information & Knowledge, 2024,41(1)：80-91.)

Abstract: [Purpose/Significance] With the increasingly emergence of multi-type data in the era of information resources, analyzing bibliographic ideas in the process of data structuring is helpful to the management and application of unstructured data. **[Design/Methodology]** This study analyzes the essential process of data structuring, reveals the bibliographic mechanism, indexing and classification thought contained in this process, and explains the feasibility of using bibliographic ideas to guide data structuring. Based on the literature description, indication, classification,and organization used in catalog work, the characteristics of unstructured data are identified, and the main structural processes are guided to complete, so as to realize "Distinguishing to Show the Academy, Researching to Define the Origins". **[Findings/Conclusion]** Data structurization basically inherits the bibliographic ideas that mainly include classification and indexing, and reflects the continuation and development of bibliography as a practical science in the current environment. **[Originality/Value]** Above mentioned approach displays the standard routine of data structurization, and further strengthen reusability of this process.

Keywords: Bibliography; Unstructured data; Classification for data; Organization for data

1 引言

现代信息技术的高速发展，推动了世界的数字化进程，人们的各项网络活动时刻生产着数字信息，同时又依赖各项数字信息改变着自身的生活。正因如此，人们兼具了数字信息生产者与使用者的双重身份，由此产生了一系列数字信息与现实世界耦合相生的新兴技术，如元宇宙、VR虚拟现实、数字孪生等。在此背景下，现实世界与数据世界的交织愈发紧密，极大促进了现存数据容量的增长、丰富了数据类型，确保了数据对客观世界描述的全面性，但亦导致数据管理过程出现数据结构混乱、质量不一等现实性问题。当下，根据结构类型不同可将数据分为结构化数据、半结构化数据和非结构化数据（如表1）。将无固定结构的半结构化、非结构化数据转化为具备特定结构数据的有序化过程，即数据结构化。通过统一数据的类型和结构，可有效解决如今数据管理混乱的难题。

与半结构化、结构化数据不同，非结构化数据的组织方式多样，不符合数据模型，一般无法通过简单直接的方式分析处理，将其存储在二维逻辑清晰的关系型数据库中。因此，数据结构化过程的研究对象以非结构化数据为主。将非结构化数据进行结构化不仅便

[基金项目] 本文系国家社会科学基金项目“面向国家安全的科技情报态势感知研究”（21BTQ012）的研究成果之一。（This is an outcome of the project “Research on Scientific Intelligence Situation Awareness for National Security”(21BTQ012) supported by National Social Science Foundation of China.）

[通讯作者] 石进（ORCID：0000-0002-1621-6944），博士，教授，研究方向：智能目录学、大数据分析，Email:shijin@nju.edu.cn。（Correspondence should be addressed to SHI Jin, Email:shijin@nju.edu.cn, ORCID：0000-0002-1621-6944）

[作者简介] 彭贤哲（ORCID：0000-0002-0131-8227），博士研究生，研究方向：目录学、大数据分析与技术，Email:pengxz_tm@163.com；郑建明（ORCID：0000-0002-7989-4435），博士，教授，研究方向：数字信息资源管理、目录学基础理论，Email:zhengjm@nju.edu.cn；李佳新（ORCID：0000-0001-7780-3723），硕士研究生，研究方向：信息组织、情报学，Email:mf21140060@smail.nju.edu.cn。

表1 不同类型数据的特点[1]
Table 1 Features of Different Types of Data

数据类型	举例	特点
结构化数据	二维表	先有结构后有数据，二维结构
半结构化数据	HTML 文档、XML 文档	先有数据后有模式，无规则性结构，数据内容与结构混杂在一起
非结构化数据	图像、文本、音频、视频	模式和结构具有多样性

于数据管理，更有利于挖掘其中潜在的高价值信息。就当前数据分布和利用特点而论，现今广泛存在的数据多为非结构化数据，但数据领域价值和利用率较高的却为结构化、半结构化数据。作为主体存在的非结构化数据，虽然具有信息丰富、使用价值高的特点，但其多样的形式增大了数据分析处理的难度，降低了数据蕴藏价值的显性化程度，限制了非结构化数据的利用管理和价值挖掘。为此，规范非结构化数据的结构化解析技术，增强该技术的复用性，具备相当可观的潜在价值[2]。

当前，关于数据结构化的研究多以实践应用为主，鲜有深入该过程本质属性的理论研究。在现有数据结构化的实践研究中，以方法研究最能体现其机理，如邹波[3]参照文件目录树、索引等数据组织管理方法，设计并构建了海量非结构化数据组织管理系统MUDOMS；冀中[4]通过视频内容分析技术，提取语义内容以获取视频的摘要、索引信息等结构化数据，促进了视频信息的传播利用；刘娜[5]结合非结构化视频数据的颜色和运动特征创建目录索引信息，提高了视频的查询检索效率。此外，亦存在少量针对不同类型非结构化数据进行统一管理利用的研究，如杨伟[6]以包含文本和地图的时空轨迹数据为研究对象，提出了系列轨迹数据结构化处理的模型与算法；曹磊[7]以文本和图像数据为研究对象，构建不同类型非结构化数据的混合索引，统一其表示和访问方法。

综上而言，现有研究中的结构化处理对象多以文本、视频数据为主，并借助自然语言处理、数据挖掘、深度学习等技术以过滤、加工、组织非结构化数据。不同研究者或为促进数据的高效统一化管理，或为开发挖掘数据潜在价值，完成了特定类型的非结构化数据的结构化过程。由此看来，现有关于数据结构化过程的研究指向性突出，多关注于该过程的实践效果及特定领域的应用，鲜有深入开展对数据结构化过程的理论探究，数据结构化的具体实现过程在不同的研究中存在较大的差异性。即便如此，数据结构化过程在不同研究中亦存在部分共性特征，即绝大部分研究利用了词表、索引、标引、摘要等目录学工具，保证了结构化解析过程的实现。

目录学作为文献管理组织的致用之学，其强调的“辨章学术、考镜源流”思想大意为：“辨别学术使其彰显，稽考源流使其明晰”[8]，由此衍生的工具已被广泛应用于结构化过程，说明其中蕴含的文献资源“有序化”的核心思想与数据结构化过程的本质属性别无二致，该过程是目录学思想在当代大数据环境下的具体体现和实例写照。为此，本文立足目录学角度，落实“理论指导实践，实践验证理论”的指导方针，剖析数据结构化的具体过程，挖掘其中蕴含的“辨章学术、考镜源流”的目录学思想，从本质属性视角透析数据结构化的步骤、实现方式以及目录学机理，辨明该过程的方法机理和内核本质，建立标准化范式与流程，提高该过程的效率和性能，实现非结构化数据的规范化管理、高效能利用，丰富结构化过程机理方向的研究。

2 数据结构化过程机理解析

将非结构化数据转化为结构化数据，即在非结构化数据中挖掘隐藏结构[9]，消除噪音数据以去繁从简，关联孤立数据以分类重构，组织散乱数据以序化存储，将海量不确定的非结构化数据过滤为明确的结构化信息的过程[10]。数据结构化过程虽具有大量的实践基础，但从理论层面了解该过程的研究却较为匮乏，为此，本文将从蕴含机理、本质过程、实现方式三个方面解析数据结构化，加强对该过程的理论认识。

2.1 目录学机理

目录学要领在于“考镜源流、辨章学术”，分类在这一过程中的地位举足轻重。如若缺乏分类工作，书

籍的组织管理则无统一章法可循，这将大大降低读者查询利用书籍的效率。为此，目录学通过构建书籍的知识门类，为书籍的合理使用及高效管理提供了现实可行的途径。在此过程中，基于“同则同之，异则异之”[11]、“合乎义例者曰类，不合者谓之不类”[12]等体现分类基本原理的观点，目录学衍生出了众多系统分类体系，如六分法、四分法、八分法等[13]，用于书籍的辨同别异、知识管理、查询利用。

基于分类原则，目录工作的最终导向在于将文献资源有序结构化以便高效利用，在此过程中，标引作为分析、提炼、总结书籍信息的重要步骤，重在以浓缩、显性化的标记字段反映书籍的整体内容，方便读者快速发现、获取所需书籍资源。故此，目录学作为一门致用之学，其中蕴含的以书籍资源知识显性化、信息浓缩化、查询便捷化为中心的标引思想，在目录工作中具有十分突出的地位。

目录作为将非结构化文本凝练为结构化信息的典型工作实践，其中强调的分类和标引思想历久弥新，对于当前数据结构化的过程亦具有借鉴和参考价值。如图1，以分类和标引为核心的目录学思想不仅可用于指导文献资源的结构化，亦适用于实现图像、视频等非结构化数据的结构化。

识别非结构化数据中隐含的结构属性，挖掘非结构化数据中可用于分析处理的“原子结构”，根据具体需求确定非结构化数据中结构单元的分割层次，进而识别抽取实体及其相互之间的关系，这些实践工作是对目录学标引思想的延续与继承。在信息技术的加持下，结构化过程从以往针对专业文献的篇、章、段、句进行著录、摘要、索引和综述，切换为分析、提炼、总结泛化的非结构化数据中不同粒度的记录单元。结构化过程中分析单元粒度的确定应参考目录工作中分类法确定分割单元时“即类求书，因书究学”[12]、“类例既分，学术自明”[14]的准则，权衡数据语义的完整性、管理的精细化、查询的便捷性。究其本质，该过程为精简非结构化数据的内部组织，实现“文献知识”或“数据知识”单元化，用于分析、组织、存储、检索之需。

非结构化数据的标注完成了对数据的内部解析，通过拆解非结构化数据的多样化结构，实现数据中隐含结构属性的提取和注释。与目录工作流程类似，标注之后的步骤在于对提取的注释数据分门别类，根据某一特定分类体系和逻辑结构组织并揭示数据，据此重塑数据结构并将其系统化。该步骤是关联、存储数据的基础，由此将非结构化数据的隐性结构显性化，重塑非结构化数据的混沌结构。

2.2 数据结构化的本质过程

数据结构化的落脚点，在于将非结构化数据凝练为可序化表达的二维逻辑数据结构，以节点和节点连边的“属性—属性值”二维结构，采用键值对格式进行存储[15]。该过程主要在于提取非结构化数据中结构化单元的属性特征，实现关联分类、序化组织，构建二维逻辑以表示非结构化数据，降低非结构化数据的信息熵值，为此，可将其本质过程分为以下三步。

（1）根据需求确定研究对象、关键特征，抽取对应粒度的结构化单元，实现精简标准化存储。离散、无序的非结构化数据蕴含潜在的利用价值，不加过滤的存储方式虽然可令数据高度保真，但却为后续的管理利用带来了阻碍。因此，数据结构化的首步工作即是

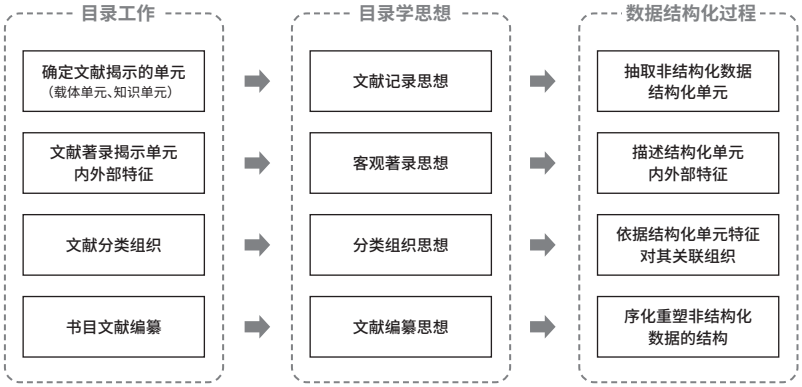

图1 目录工作、目录学思想、数据结构化过程对应关系

Fig.1 The Relationship Among Bibliographic Work, Bibliographic Theory and Structure Process

在具体场景中明确的需求导向下，确定数据的目标用途、关键属性、分析粒度、质量要求、结构字段等内容，继而标注关键属性特征以获取相应粒度单元的结构化字段，衡量描述数据的实体、关系属性特征的价值，根据需求赋予属性特定的权重，根据属性权重大小判断是否足以保存其对应的特征值，过滤冗余无关信息，以获取精简后可代表非结构化数据并满足具体需求的多维特征字段，构建节点属性表。该过程虽不可避免带来一定的信息损失，但在确保满足具体需求的前提下，有助于实现格式的高度统一规范、语义特征的显性化，实现数据的精简标准化存储，进而扩展数据后续调用管理的操作空间。

（2）描述单元特征，构建高度关联组织。非结构化数据的精简存储初步实现了数据的单元化过程，着力于完成“数据单元节点”的提取构建过程，但节点之间仍处于相对离散的孤立状态。下一步可利用数据的内外特征，抽取非结构化数据之间的显性或隐性关联路径，搭建“数据单元节点”的“网络关系”，拟定节点之间连边的属性表，这有利于数据的灵活管理，提高数据组织的系统严密性。

（3）重塑结构，序化存储，促进数据的高效传递利用。存储组织非结构化数据的目的在于利用，但原始的非结构化数据格式多样不一、语义信息隐匿，不利于数据的高效传递，而精简存储的数据将内外部特征以字段形式表达，可提高非结构化数据传递的专指性，即根据用途提取相关特征以便于数据的直接利用传输，以高度精简的传递内容优化传递效率、改善数据使用成效。此外，基于结构化字段关联组织的非结构化数据，可确保数据传递的完备性和精准性，展现非结构化数据的潜在利用价值。

2.3 结构化过程实现方式

数据结构化的本质过程是一定的，但在具体实践中的实现方式是多样的。该过程的具体实现方式包括识别数据中隐藏的结构属性值、筛选关联离散数据及组织分类重建系统结构。其中涉及的技术手段包括基于命名实体识别、关系抽取、知识图谱、语义分析的文本结构化，基于主体识别、语义分割、内容解析的图像结构化，以及基于关键帧提取、语义特征标记、目标跟踪的视频结构化。

在已有非结构化数据的结构化研究中，非结构化数据的结构化过程实现方式虽然多样，但一般可归纳为三步。如图2所示，视频数据的结构化过程归纳为镜头分割、关键帧提取、场景重构[16]，音频数据的结构化过程归纳为音频分割、分类及平滑后处理[17]，文本数据的结构化过程归纳为分词处理、内容定位和结构化字

图2 非结构化数据的结构化过程

Fig.2 Structuring Process of Unstructured Data

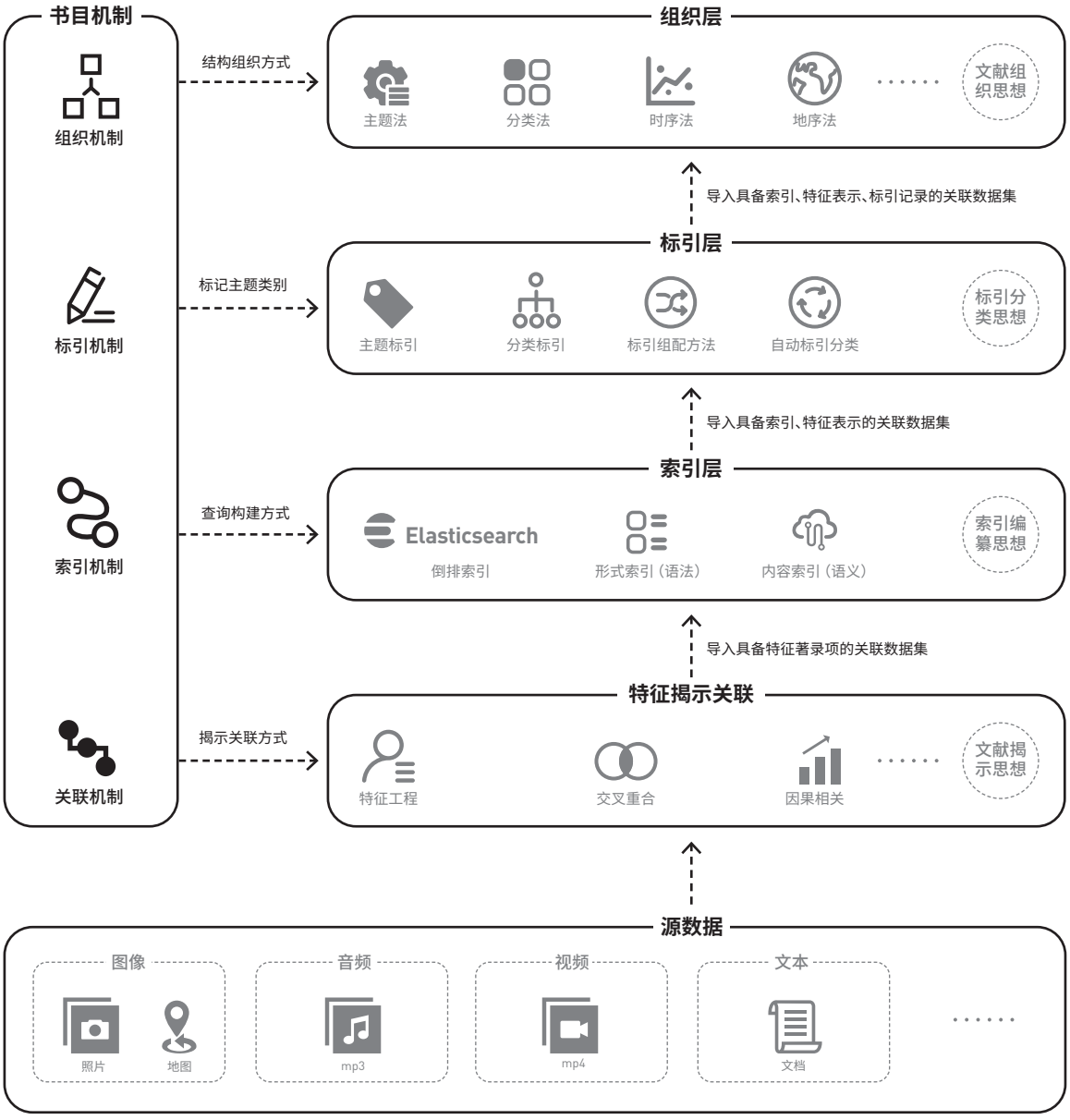

图3 数据结构化过程中的书目机制

Fig.3 Bibliographic Mechanism in the Structuring Process of Unstructured Data

段提取[18]，图像数据的结构化过程归纳为特征提取、结构化表达、场景表达[19]。

不同类型的非结构化数据的结构化实现方式虽多样不一，但本质过程符合上文所述的三步，具备显著的趋同性，与传统目录工作互通有无。围绕数据“有序结构化”而展开的三个标准化过程，流露出数据结构化过程中蕴含的目录学思想，说明数据结构化的本质在于实现目录学倡导的资源“有序化”，是作为“致用之学”的目录学在大数据时代背景下的具体呈现。

3 数据结构化过程中书目机制的体现

常见的非结构化数据主要为文本、图像、音频、视频等多媒体数据，为实现古典目录学强调的“辨章学术、考镜源流”的目标，具体结构化实现过程注重辨别分类用以彰显特征，倚重稽考源流借以关联组织，依照先后顺序分为关联整合数据、索引指示数据、标引分类数据、重构组织数据四步。具体过程如图3所示，先通过揭示关联步骤建构数据之间的有机联系，将孤立

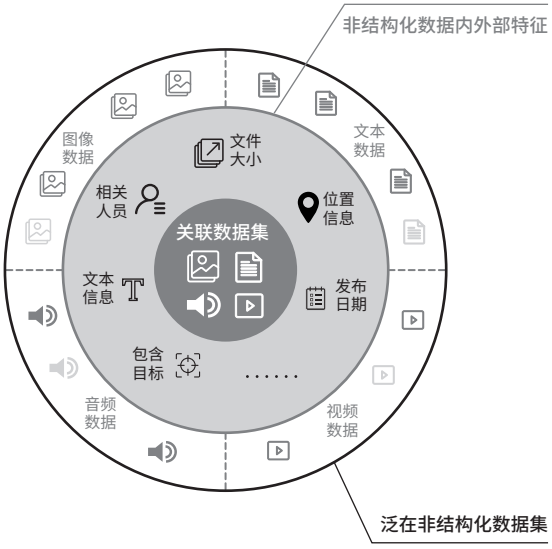

图4 网络环境下非结构化数据的关联
Fig.4 Association of Unstructured Data in Internet Era

数据“汇集成流”；进而借助索引指示环节透析数据来源及存储位置，使无序数据“有源可循”；再者根据标引分类过程辨析并赋予数据类别标签，为隐晦数据“注解标类”；最后利用多样组织方法解析并重建数据结构框架，令繁杂数据“系统规范”。

3.1 关联整合机制

大数据时代，各种类型的非结构化数据由于没有固定的结构和模式，在未结构化之前通常较为分散，在未加以著录之前多为孤岛数据，相互之间关联性较差，很难构成完整严密的系统。依据数据特征之间的重复、交叉、包含等集合关系构建关联机制，可有效解决“数据孤岛”的问题。如图4所示，将泛在的非结构化数据转换为结构化数据，首先即是在确定分割单元内外部特征的基础上，建立非结构化数据的关联机制，进行整合分析。根据目录学思想，文献可借助内外部特征的揭示产生关联，非结构化的数据亦可以此为鉴，通过揭示数据的内容与外部形式特征完成非结构化数据的关联整合。

3.1.1 内容关联

非结构化数据的内部特征，主要表现为文本、图像、音频、视频的内容信息。依据数据内部特征的交叉重合、指引借代及语义相近关系，产生了共现、链接、语义三种关联方式[20]，其中蕴含的关联机制，在于衡量非结构化数据内部特征之间的相似性、相关性、一致性，实现该过程的方法主要包括相关性分析、主成分分析、因子分析、聚类分析等。

就文本而言，通过规则和词典的方法、基于机器学习的方法及基于深度学习的方法[21]，可得到文本的内容特征，实现文本在语法、语义、语用层次的关联，将分散无序的孤立文本整合为统一整体。与文本数据类似，音频数据中同样蕴含结构化字段信息，如音色、音调、旋律等[22]，根据音频信息固有的内容特征相似性亦可产生关联。关于图像数据的关联方法主要分为两种[23]，一种方法是基于图像的描述内容特征产生关联，如由标题关键词组构成的主题图像集系统，可加强图像数据与文本数据的关联程度；另一种方法是基于图像内容挖掘潜在关联模式，可细分为像素类、特征类和语义类三个层次的关联模式；将两种方法结合，综合图像的描述内容信息和固有内容信息构建关联模式，可实现图像中两类特征的关联。视频数据作为音频数据与多帧图像数据的集合体，针对音频和图像数据的关联方法同样适用于视频数据，但视频数据某些专有的内容特征有待使用特定方法手段实现关联，特别是构成视频数据的图像之间本身即存在时序和逻辑上的内容关联[24]，可据此整合分散、孤立的视频数据。此外，不同类型的非结构化数据在提取出结构化的特征之后，可根据特征的相似性产生对应关联[25]。

3.1.2 形式关联

目录工作在关联文献时使用的关联结点包含文献的内容特征和外在形式特征，以图书的关联为例（如图5），内容特征为关键词，外在形式特征包括作者、出版社、丛书、版本，基于特征的重合交叉关系，构建了不同图书单元的关联网络。数据结构化过程可以此为参考，将非结构化数据的外在形式特征作为关联依据，这些特征具体可包括发布者信息、文件名、发布时间、语言类型、来源出处等。

根据多媒体数据发布的时间产生先后顺序的关联，在结合数据固有内容基础上分析发布者信息的异同点，有助于后续过程中非结构化数据在时序上的分类与组织。非结构化数据的来源出处构成的网络结构，在一定程度上亦可展现数据之间的相互引用摘录情况，避免同源异构数据的冗余性。外在形式特征相对内容特征更为清晰，可借助简单的逻辑判断实现不同数据之间的关联，继而结合内容特征关联，构成非结构化数据的完整关联体系。

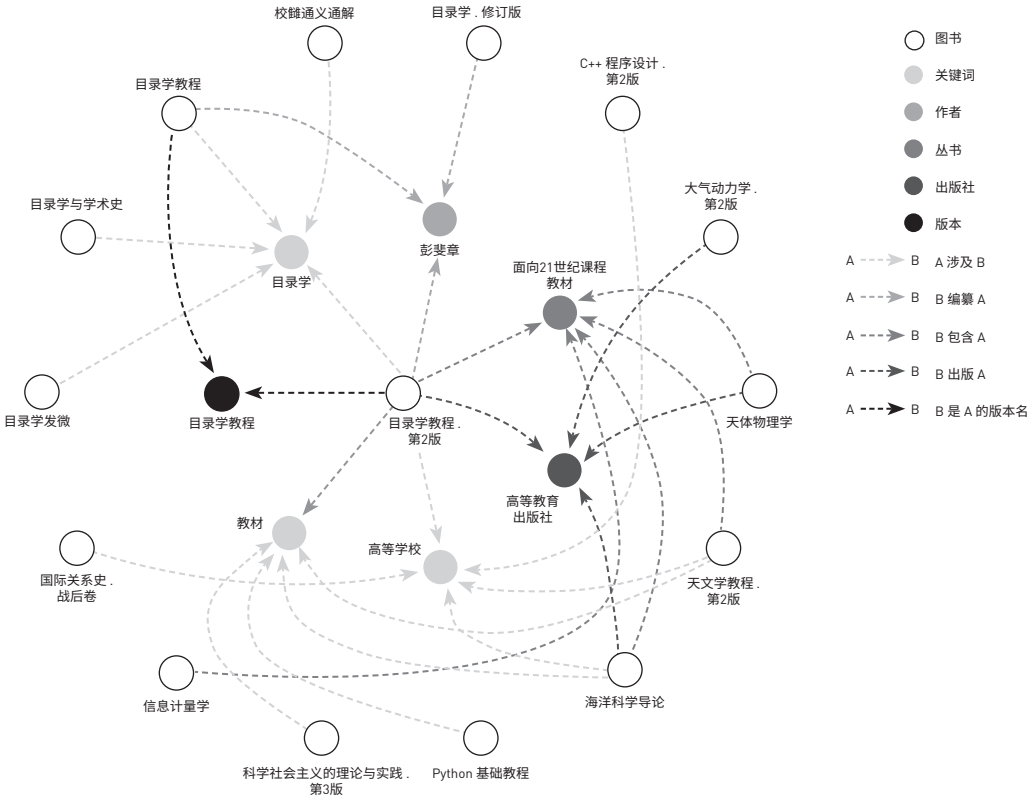

图5 基于内容特征和形式特征的图书关联示意图

Fig.5 Association Diagram Based on Content and Formal Features from Books

目录学通过揭示文献的内外部特征，可掌握不同文献内容的广度和深度的关系，展现文献内容之间的联系[26]。非结构化数据的结构化过程在此指导下，可通过提取不同类型数据的内容和外在形式特征，融合特征并实现关联匹配，进而将孤立的非结构化数据整合为一个紧密关联的统一整体。通过形式关联机制，可实现特定主题非结构化数据的信息聚合功能，这类似构建特定专题书目，依据特征脉络梳理信息资源，将具备共同特征的非结构化数据“汇集成流”。

3.2 索引指示机制

索引是提高数据查询效率最常用和有效的方法，重在指示数据位置，提高信息查询和检索的速度[27]。目录学中的索引构建编纂方法可直接应用于非结构化文本数据，其特点在于记录文献中个别事项和内容作为检索单元，指引文献位置。在书目工作中，完备的索引一般具备四个条件[26]：（1）必须由众多索引款目组合而成；（2）索引款目由著录标目、修饰语和参照项三者组成；（3）索引款目按一定次序组成；（4）索引款目能够通过一定方式联系标目。满足以上四个条件的索引可以指示某一类特定字段信息的文献群出处，且索引之间存在一定的关联。在借鉴书目索引的基础上，数据索引亦应满足与之对应的要求。

首先，数据的索引需建立在关联非结构化数据的基础上，由此产生的单个索引能指示与之相似或相关的某一类数据群的出处。针对文本类型数据，这些索引包括标目信息（标题、关键词、主题词等）、说明注释信息、存放位置信息；针对图像类型数据（如图6），索引主要由标识信息（文件名、主题内容等）、图像补充说明信息（图像大小、像素、客观描述信息等）、图像存放位置信息构成。

再者，非结构化数据的索引编排有一定的顺序，非结构化文本数据的索引编排方法包括字顺法、主题法等，音频、视频、图像等非结构化数据可在其文本描述信息基础上借鉴文本索引的编排方法，同时亦可根据自身的专有特点构建特定的索引编排方法，如音频的索引可依据音高、音色等专属特点排序。索引的序化便于非结构化数据的查询检索，是数据结构化过程不可或缺的一环。由此生成的索引既不能等同于目次内容的字顺排列，亦不能简单地把大小章节标题全部做

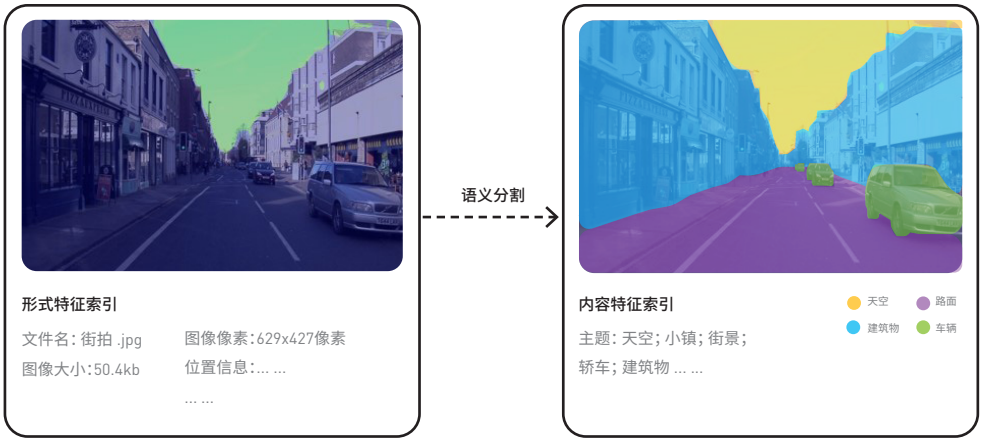

图6 图像数据的索引构建示意图

Fig. 6 Construction Diagram for Index of Image

成索引款目[28]，索引的编排方式必须多样灵活，方能揭示原始非结构化数据的内容和位置信息。

根据构建目录索引的要求提取生成非结构化数据的索引，重在指示数据的位置信息，便于非结构化数据查询检索的多样化、精准化和高效化，是数据实现结构化的关键一步。索引机制的构建，主要依据非结构化数据“时空逻辑特征”具备的顺序性特点，为每一条非结构化数据附加位置信息，增强数据“可溯性”，便于在“数据流”中精准提取信息。

3.3 标引分类机制

原始的非结构化数据由于逻辑结构相对分散、类型多样，不利于分析与管理，为此，需用其他手段提取原始数据中蕴含的价值信息，以新生的结构化数据揭示、描述并替代非结构化数据。文献标引作为目录工作中揭示文献内容的一种手段[29]，亦可以数据标引的形式表现在数据的揭示工作中，以此自动分类并生成标签[30]。

数据标引有利于非结构化数据的结构化管理，是整个大数据应用的基础，也是实现知识发现和数据创新的关键所在[31]，目前关于非结构化数据的标引主要包括网络信息自动标引[32]、数字图像标引[33]以及音视频信息标引[34]。

目录学中，按照标引目的的不同可将标引分为主题标引和分类标引，将该思想同样应用于数据结构化标引过程时，前者在于构建主题词表，挖掘数据价值信息便于利用；后者注重提取辨识信息用于分类。非结构化数据一般涉及到多主题，主题标引运用到了概念限定组配和概念相交组配两种方法[35]。概念限定组

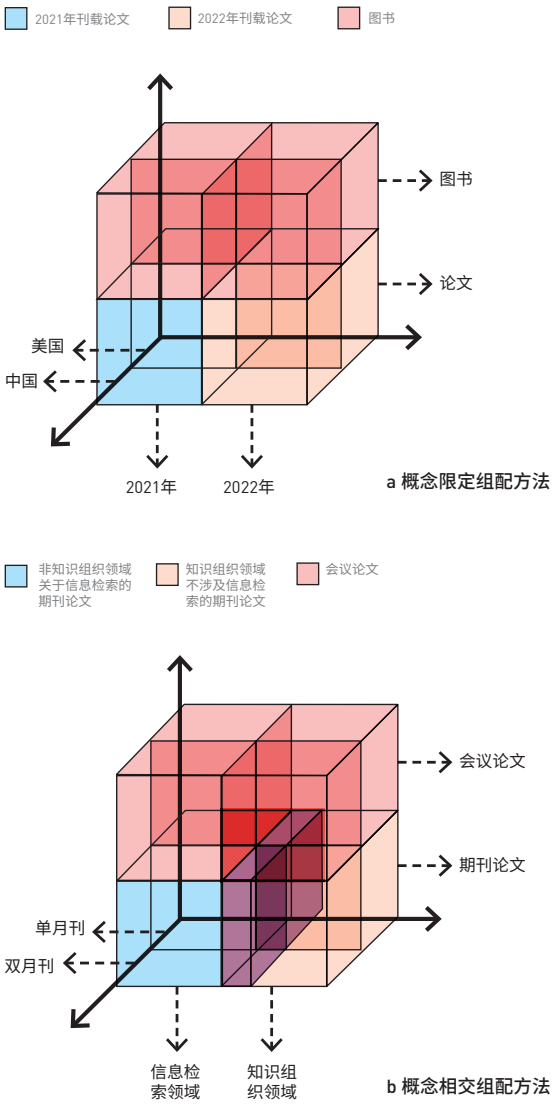

图7 主题标引构成分类标引示意图

Fig.7 The Diagram Illustrating that Subjects Represent Classification

配依据主题词间存在的语义和语法关系，用一个或者多个不存在相交关系的主题词表示限定概念与被限定概念之间的并列关系。以“2022年刊载论文”为例（图7a），组配“2022年”和“论文”两个主题概念，提升概念专指度。概念相交组配，指多个主题词之间具备概念交叉关系，如将“信息检索”与“知识组织”这两个概念进行组配（图7b），即可得到专指度更高的概念—“基于知识组织的信息检索”。通过分析非结构化数据的特征信息构建基础主题词表，每个主题代表一个特定维度，根据主题的交叉组合方法即可揭示非结构化的数据内容，由此建立的主题空间系统将在保证结构化的条件下最大程度代替并表示非结构化数据。

在获取主题标引信息之后，可据此开展分类标引工作，按照分散和集中的要求录入相应的类目。如果非结构化数据仅包含一个主题，则可将此条数据归入特定的类目之中。当数据中出现多主题时，如果主题之间相似度较高，则可能构成多个主题对应一个类目的结果；如果主题之间相似度较低，一条数据则可能有多个分类标引。根据主题标引组合获取分类标引，形成严格的、结构化的、分门别类的集合，可用于表达多个事件、对象之间的关系[36]。

针对非结构化数据的标引机制，赋予非结构化数据主题标签，在此基础上增添类别属性，划分数据类群，达到“辨章”的目的。其中主题标签对数据辨析程度的深浅，影响分类体系的系统性与完备性，决定了标引步骤能否区分辨别数据。

3.4 重构组织机制

文献经过书目加工后形成的各种款目，被收入到各种书目索引中，需按照科学的方法使之有序化，形成一个有组织的严密整体，揭示出相关的文献特征，以便读者选择[26]。文献的这种整合组织机制，规定了文献在特定载体上存储和组织的方法，以便快速地访问、处理和管理文献。类似地，非结构化数据在经过关联整合、索引构建、标引分类之后，已经具备了关联、索引及标引信息，亦可参考文献组织机制，综合利用上述方法以排序、组织、重构非结构化数据，实现完整的结构化过程。目录学中，文献组织方法包括分类法、主题法、时序法、地序法等[26]，由于文献资源多为文本类型，因此这些方法多可以直接应用于非结构化文本数据的组织。

其中，时序法、地序法作为以外部特征为依据的数据组织方法，以图像、音频、视频等非结构化数据中关于时间、位置的标引、索引、关联信息作为组织依据，可从时间和空间维度排序重组数据。如上海图书馆的数字人文项目[37]，即为利用历史文献、历史地图等非结构化数据中蕴含的时间、地点信息组织而成。根据外部特征方法进行重构的结构化数据较为系统地揭示了原始数据信息，但该类方法的语义揭示程度较浅，在组织过程中仍需纳入揭示数据内容特征的方法进行优化重构。

而分类法、主题法这类以内容特征为依据的文献组织方法，需结合该类型数据的专有特点方可适用组织非结构化数据。如使用文献的分类法组织音乐这类音频数据的过程中，可对其进一步改良和扩展，将音乐的风格韵律等专有特点纳入分类依据。由主题法指导的数据组织主要依赖于非结构化数据的内容特征和主题标引信息，将数据中论述事物对象的主题，用规范化的术语表示，将同一主题的数据加以集中，以适应查询、检索、阅览之需。根据主题编排组织的结构化数据应具备以下特点[26]：（1）能够灵活准确地以自然语言或受控语言直接表达概念；（2）将与研究对象有关的各种类型的非结构化数据集中在一起，强化数据的综合全面性；（3）主题之间是互相独立的，可按字顺法等计算不同主题数据组织顺序的权重，便于非结构化数据的高速检索查询。

基于外在形式特征的时序法、地序法构建了非结构化数据体系的宏观组织框架，而基于内容特征的主题法丰富了宏观组织框架下辖的微观组织单元，该类单元可根据内容特征产生关联，跨越时间和空间宏观框架的限制，实现非结构化数据不同层级的网状连接，提高非结构化数据的查全率。如上海图书馆人名规范库[38]，依据人物的内外部特征构建人物本体，实现了图像、文献等非结构化数据的组织关联、结构化呈现与管理。

非结构化数据的重构组织是数据分析、挖掘、管理的基础，通过借鉴目录学文献组织方法，可在著录关联基础上揭示文献资源潜在的学术规范时空逻辑秩序，由此制定的数据组织机制通过解析数据单元进而梳理其内在时空逻辑顺序，建立“学术化”规范性的结构化数据系统，完成非结构化数据结构化过程的最后一步。

4 数据结构化过程中的书目思想应用

当下数据结构化过程的实践应用，在大数据分析与处理、人工智能系统构建、元宇宙的探究等新兴技术中均有体现。这些新兴技术虽然具备灵活且多变的外在表现形式，但蕴含其中的数据结构化过程依旧围绕着信息知识结构化展开，经过提炼与整合碎片化知识，建立信息之间的“有机联系”，完成无序信息的组织、揭示与利用，从而提供多样化的服务。数据结构化在新兴技术的实践应用，一定程度上丰富了该过程的最终导向，但其本质却大同小异，尤其在内容、方法论、系统三个层面，呈现出了现代信息化技术对书目思想的传承与应用。

4.1 内容层面

大数据分析与处理相关技术主要在于数据的采集与筛选、数据的表示与描述、数据的关联与聚类、数据的组织与存储等等，人工智能系统构建的核心技术包括计算机视觉、机器学习、知识表示等等，元宇宙的实现也必须建立在对现有世界的描述表示数据的提炼、关联、组织基础上。从上述技术处理内容来看，一者在于提炼描述原始数据，通过特征工程、文本表示、表示学习等技术，以有价值的代表性特征数据，精炼表示原始信息，与书目思想中的文献揭示内容异曲同工；二者注重关联整合孤立数据，通过文本分类、特征聚类、关联规则等机器学习算法，连接孤立的代表性特征数据，系统串联信息单元，与书目思想中文献组织内容殊途同归。

书目思想在数据结构化过程中内容层面的体现，在于其提供了每一个结构化步骤的预期需求。如参照文献揭示思想在于确定、揭示描述文献单元的内外部特征。数据结构化第一步的目标即是通过特征抽取、矢量表示等计算机可理解、编码、分析的方式，抽取、存储结构化单元的关键特征。有鉴于此，书目思想作为数据结构化过程内容层面的指导思想，规范了结构化步骤的顺序、要求与目的。

4.2 方法论层面

书目思想所提供的方法论在于剖析书目结构原理以科学有效地揭示文献资源信息，具体实施过程涉及文献的选择、揭示、著录、排列等，实施方法包括类序、叙录、索引、文摘等。其通过构建著录、索引、标签等书目数据，描述、关联、组织多样且丰富的文献资源信息。

书目思想所提供的方法论，亦可见于现有的数据结构化的实践应用过程中。如依据文献组织的主题法、分类法提出的数据关联聚类算法，可系统整合原始数据、生成类别标签并产生关联。启发于书目思想提供的方法，数据结构化处理的实践过程构建了类似书目数据的二次数据，如特征数据、索引数据、标引类别、导航数据等等，以此为干、以干带面、纲举目张，是数据处理实现自动化关联、智能化组织的基础。

4.3 系统层面

大数据分析与处理技术目的在于发掘原始数据隐藏的规律与模式，人工智能系统构建核心在于从训练数据集中获取用于决策计算的数据特征，元宇宙实现的关键在于组织整合用于描述原始世界或数据的二次数据，这些技术所涉及的数据结构化处理结果，均是在描述、关联和组织数据基础上，构建一个能反映原始数据变化规律、预测未来数据走势信息的数据系统，便于数据的治理与挖掘。这与书目思想建立严密规范的图书系统以便于图书的管理与利用的意图是一致的。

以文献资源结构化为目的的书目思想，在于提取二次书目数据构建文献资源管理框架，增强原始文献数据组织的系统性，这同样对数据结构化过程提出了系统层面的要求。为此，实践过程最后获取的结构化数据，整体结构应当尽可能规范严密，这也是数据结构化的目标。该目标的实现可依托表示学习技术识别提取数据内外部特征、计算特征权重以衡量重要性程度、根据特征聚类关联算法构建网络，在分割数据单元基础上进行关联整合，赋予结构化数据系统性的特点。

5 结语

大数据时代下，非结构化数据广泛而普遍，蕴含丰富且宝贵的信息，具备十分可观的挖掘价值。通过实现非结构化数据的结构化分析、存储、管理、利用，开启非结构化数据的“矿藏大门”，探索发现其中的潜在价值。目录学作为文献资源结构化过程的理论精髓，其中蕴含的分类和标引思想历久弥新，亦充分体现在非结构化数据的结构化实践过程中，可将其作为指导

数据结构化过程的核心思想。

为此，本文通过解析数据结构化过程的多样呈现形式，剖析其本质属性，挖掘不同实践过程中共同蕴含的目录学机理，提出了数据结构化过程需建立的关联、索引、标引、组织四种机制，以达到“辨章学术、考镜源流”的目的，将其作为一个标准化过程，增强数据结构化过程的复用性，在不断迭代优化的基础上提高该过程的效率及性能。

总结来看，本文在前人研究基础上，从理论层面探析了数据结构化的本质过程、目录学机理、书目机制以及书目思想在其中的应用，较为系统全面地论述了数据结构化过程对目录学思想的传承与应用。但是，本文的研究视角整体上仍然过于宏观，对于数据结构化过程的细节把控多有不足，对该过程中应用的技术所依托的目录学方法原理尚未明晰，后续工作有待于通过案例分析方法，聚焦数据结构化过程的具体步骤，从微观实践层面剖析该过程中蕴含的目录学思想，并进一步在实际场景中加以验证与应用。

作者贡献说明

彭贤哲：文献调研，框架设计，论文撰写与修改；

郑建明：提供思路，框架设计，论文撰写与修改；

李佳新，石进：论文修改。

参考文献

[1] 彭宇，庞景月，刘大同，等 . 大数据 : 内涵、技术体系与展望 [J]. 电子测量与仪器学报 , 2015,29(4)：469-482.(Peng Yu, Pang Jingyue, Liu Datong, et al. Big Data: Connotation, Technical Framework and its Development[J]. Journal of Electronic Measurement and Instrumentation, 2015,29(4)：469-482.)

[2] 李战怀 , 王国仁 , 周傲英 . 从数据库视角解读大数据的研究进展与趋势 [J]. 计算机工程与科学 ,2013,35(10)：1-11.(Li Zhanhuai,Wang Guoren,Zhou Aoying. Research Progress and Trends of Big Data from a Database Perspective[J]. Computer Engineering and Science,2013,35(10)：1-11)

[3] 邹波 . 海量非结构化数据的组织研究与实现 [D]. 武汉 : 华中科技大学 ,2008：13.(Zou Bo. Research and Implementation of Massive Unstructured Data Organization[D].Wuhan:Huazhong University of Science and Technology, 2008：13.)

[4] 冀中 . 基于多模态信息的新闻视频内容分析技术研究 [D]. 天津 : 天津大学 ,2007：9.(Ji Zhong. Research on News Video Content Analysis Technology Based on Multimodal Information[D]. Tianjin: Tianjin University,2007：9.)

[5] 刘娜 . 视频结构化索引和检索研究与实现 [D]. 武汉 : 华中科技大学 , 2009：45.(Liu Na. Research and Implementation of Video Structure Index and Retrieval System[D]. Wuhan:Huazhong University of Scicence and Techonology, 2009：45.)

[6] 杨伟 . 时空轨迹数据的结构化处理与行为语义感知 [D]. 武汉 : 武汉大学 , 2019：141.(Yang Wei. Structured Processing and Behavioral Semantic Perception of Spatio-Temporal Trajectory Data[D]. Wuhan: Wuhan University, 2019：141.)

[7] 曹磊 . 大规模非结构数据索引和可视化的研究 [D]. 天津 : 天津大学 , 2012: 22.(Cao Lei. Research on Large-scale Unstructured Data Processing of Index and Visualization[D]. Tianjin:Tianjin University, 2012: 22.)

[8] 傅荣贤 . 论章学诚“辨章学术 , 考镜源流”理念的本质 [J]. 大学图书馆学报 , 2016,34(2)：111-117.(Fu Rongxian. Dredging on the Essence of Zhang Xuecheng's Concept of Distinguishing to Show the Academy Researching to Define the Origins[J]. Journal of Academic Libraries, 2016,34(2)：111-117.)

[9] 郑建明 . 当代目录学 [M]. 北京 : 科学出版社 , 2020：243.(Zheng Jianming. Contemporary Bibliography[M]. Beijing: Science Press. 2020: 243.)

[10] 王青兰 , 王喆 , 曲强 . 新型国家公共卫生信息系统建设 : 提高系统韧性的思考 [J]. 改革 ,2020(4)：17-27.(Wang Qinglan, Wang Zhe, Qu Qiang. The Construction of New National Public Health Information System:Reflections on Improving the System Resilience[J]. Reform,2020(4)：17-27.)

[11] 王先谦 . 荀子集解 [M]. 北京 : 中华书局 , 1988：419.(（Qing Dynasty) Wang Xianqian. Collected Annotations of Xunzi[M]. Beijing: Zhonghua Book Company, 1988：419.)

[12] 姚名达 . 中国目录学史 [M]. 上海 : 上海古籍出版社 , 2002：7, 49.(Yao Mingda. History of Chinese Bibliography[M]. Shanghai: Shanghai Classics Publishing House, 2002：7, 49.)

[13] 鲁雯舒 . 中国古代私家藏书研究 [D]. 哈尔滨 : 黑龙江大学 , 2019：41.(Lu Wenshu. Research on Private Collections in Ancient China [D]. Harbin: Heilongjiang University, 2019：41.)

[14] 郑樵 . 通志二十略 [M]. 王树民点校 . 北京 : 中华书局 , 1995：1806.(（Song Dynasty) Zheng Qiao. General Zhi Twenty Little [M]. Proofread by Wang Shumin. Beijing: Zhonghua Book Company, 1995：1806.)

[15] 叶春蕾 . 基于 Hadoop 的高校图书馆大数据关键技术研究 [J]. 数字图书馆论坛 , 2017(5)：33-38.(Ye Chunlei. Study on the Key Technology of University Library's Big Data Based on Hadoop [J]. Digital Library Forum, 2017(5)：33-38.)

[16] 王筱斐 . 基于视频内容的智能视频摘要系统 [D]. 北京 : 北京邮电大学 , 2017：7-8.(Wang Xiaofei. Intelligent Video Abstract System Based on Content[D]. Beijing: Beijing University of Posts and Telecommunications, 2017：7-8.)

[17] 李莉，李胜广，谭林 . 多维警务感知指挥系统设计 [J]. 中国安防 , 2016(12)：88-91.(Li Li ,Li Shengguang, Tan Lin. Design of Multidimensional

Police Perception Command System[J]. China Security & Protection,2016(12):88-91.)

[18] 尚小溥 , 许吴环 , 赵红梅 , 等 . 中文超声文本结构化与知识网络构建方法研究 [J]. 图书情报工作 ,2019,63(16):112-120.(Shang Xiaopu,Xu Wuhuan,Zhao Hongmei,et al. Research on Structure and Knowledge Network Construction of Chinese Ultrasonic Text[J]. Library and Information Service,2019,63(16):112-120.)

[19] 许鹏飞 . 图像结构化特征表达方法研究 [D]. 哈尔滨 : 哈尔滨工业大学 , 2013: 23.(Xu Pengfei. Research on Structured Feature Representation in Images[D]. Harbin: Harbin Institute of Technology, 2013: 23.)

[20] 常娥 , 孟祥保 . 图书馆资源组织中的数据关联特征研究 [J]. 图书馆论坛 ,2016,36(2):49-56.(Chang E,Meng Xiangbao. The Characteristics of Data Association in Library Resources Organization[J]. Library Tribune,2016,36(2):49-56.)

[21] 黄胜 , 王博博 , 朱菁 . 基于文档结构与深度学习的金融公告信息抽取 [J]. 计算机工程与设计 ,2020,41(1):115-121.(Huang Sheng, Wang Bobo, Zhu Jing. Information Extraction of Financial Announcement Based on Document Structure and Deep Learning[J]. Computer Engineering and Design,2020,41(1):115-121.)

[22] 熊回香,薛姣,李青维,等 . 视频信息的对象关联交叉检索研究 [J]. 图书情报工作 , 2011,55(S2):295-299.(Xiong Huixiang, Xue Jiao, Li Qingwei, et al. Research on Object Association Cross-retrieval of Video Information [J]. Library and Information Service, 2011,55(S2):295-299.)

[23] 杜琳,陈云亮,朱静 . 图像数据挖掘研究综述 [J]. 计算机应用与软件 , 2011,28(2):125-128,223(Du Lin, Chen Yunliang, Zhu Jing. Overview of Research on Image Data Mining[J]. Computer Applications and Software, 2011,28(2):125-128,223.)

[24] 彭宇新 , 綦金玮 , 黄鑫 . 多媒体内容理解的研究现状与展望 [J]. 计算机研究与发展 ,2019,56(1):183-208.(Peng Yuxin,QiJinwei,Huang Xin. Current Research Status and Prospects on Multimedia Content Understanding[J]. Journal of Computer Research and Development,2019,56(1):183-208.)

[25] 刘忠宝,赵文娟 . 融合语义特征和分布特征的跨媒体关联分析方法研究 [J]. 情报学报 , 2021,40(5):471-478.(Liu Zhongbao, Zhao Wenjuan. Research on Cross-Media Correlation Analysis by Fusing Semantic Features and Distribution Features[J]. Journal of the China Society for Scientific and Technical Information,2021,40(5):471-478.)

[26] 彭斐章 . 目录学教程 [M]. 北京: 高等教育出版社 ,2004:1,143-146.(Peng Feizhang. A Course of Bibliography[M]. Beijing:Higher Education Press, 2004:1,143-146.)

[27] 葛微 . 大数据索引和查询优化技术与系统研究 [D]. 南京大学 , 2019:4.(Ge Wei. Research on Big Data Indexing and Query Optimization Technology and System[D]. Nanjing University, 2019:4.)

[28] 于倩倩 . 专著内容索引的编制方法探析——以《情报检索语言与智能信息处理丛书》为例 [J]. 图书馆学研究 ,2012(19):36-42.(Yu Qianqian. Discussion on the Compilation Method of Monograph Content Index—Taking Information Retrieval Language and Intelligent Information Processing Series as an Example[J]. Research on Library Science,2012(19):36-42.)

[29] 聂鑫 . 我国书目信息标准化工作发展述略 [J]. 情报科学 ,2004,22(8):982-984.(Nie Xin. The Development of Bibliography Standardization in China[J]. Information Science,2004,22(8):982-984.)

[30] 江华丽 , 曹祺 , 陈刚 . 政府公开信息自动标引的设计与实现 [J]. 数字图书馆论坛 ,2019(1):43-49.(Jiang Huali, Cao Qi, Chen Gang. Design and Implementation of Automatic Indexing of Government Public Information[J]. Digital Library Forum,2019(1):43-49.)

[31] 唐京春 , 张新新 . 专业出版大数据建设路径分析 [J]. 科技与出版 ,2019(1):59-63.(Tang Jingchun, Zhang Xinxin. Analysis on the Construction Path of Professional Publishing Big Data[J]. Science and Technology & Publication, 2019(1):59-63.)

[32] 沈静 , 周金治 , 马建国 . 基于 UCL 的网页信息自动标引技术研究 [J]. 现代图书情报技术 ,2008(8):58-62.(Shen Jing, Zhou Jinzhi, Ma Jianguo. Research on the Auto Indexing Technology about Web Page Based on UCL[J]. New Technology of Library and Information Service,2008(8):58-62.)

[33] 包翔 , 刘桂锋 . 基于特征包的数字图书馆图像检索系统框架设计及实证研究 [J]. 图书馆杂志 ,2020,39(8):57-65.(Bao Xiang,Liu Guifeng. Framework Design and Empirical Study of Digital Library Image Retrieval System Based on Bag of Features[J]. Library Journal,2020,39(8):57-65.)

[34] 徐志华 , 周金治 , 李莲春 . 基于 UCL 的网络音视频自动标引 [J]. 电视技术 ,2011,35(12):87-89.(Xu Zhihua, Zhou Jinzhi, Li Lianchun. Automatic Indexing for Audio and Video of Network Based on UCL[J]. Video Engineering,2011,35(12):87-89.)

[35] 毕贞 , 向禹 . 档案主题标引与分类标引的比较分析 [J]. 资源信息与工程 ,2019,34(2):199-201.(Bi Zhen,Xiang Yu. Comparative Analysis of Archival Subject Indexing and Classified Indexing[J]. Nonferrous Metals Abstract,2019,34(2):199-201.)

[36] 刘润达 , 彭洁 , 涂勇 . 一种多维关键词与分类关联的科学数据资源分类导航平台构建方案 [J]. 现代图书情报技术 ,2010(9):74-78.(Liu Runda, Peng Jie, Tu Yong. A Scientific Data Resource Navigation Platform Scheme Based on Multi-Facet Keyword and Classification Connection[J]. New Technology of Library and Information Service,2010(9):74-78.)

[37] 夏翠娟 . 中国历史地理数据在图书馆数字人文项目中的开放应用研究 [J]. 中国图书馆学报 ,2017,43(2):40-53.(Xia Cuijuan. The Opening and Application of Chinese Historical Geography Data in Digital Humanities Projects of Libraries[J]. Journal of Library Science in China, 2017, 43(2) 40-53.)

[38] 夏翠娟 . 面向人文研究的“数据基础设施”建设——试论图书馆学对数字人文的方法论贡献 [J]. 中国图书馆学报 ,2020,46(3):24-37.(Xia Cuijuan. The Construction of "Data Infrastructure" for Humanities Research:The Methodological Contribution of Library Science to Digital Humanities[J]. Journal of Library Science in China,2020,46(3):24-37.)

(收稿日期:2023-03-12)