图书馆建设
Library Development
ISSN 1004-325X,CN 23-1331/G2

《图书馆建设》网络首发论文

题目：　基于层次聚类的图书元数据语义聚合研究
作者：　彭贤哲，石进
网络首发日期：　2024-06-12
引用格式：　彭贤哲，石进．基于层次聚类的图书元数据语义聚合研究[J/OL]．图书馆建设．https://link.cnki.net/urlid/23.1331.G2.20240611.1704.002

网络首发时间：2024-06-12 09:11:18
网络首发地址：https://link.cnki.net/urlid/23.1331.G2.20240611.1704.002

基于层次聚类的图书元数据语义聚合研究*

彭贤哲，石 进

南京大学信息管理学院 南京市 210023

[摘要] 实现多源异构的图书资源的深度融合，有利于拓展图书服务的广度和全面性，促进智慧图书馆体系的建设，其中，多样异构、名称不一的图书元数据的语义聚合是深度融合多来源图书信息的关键所在。为此，本文通过对比分析图书在不同平台分布的元数据的差异性，以 BERT-Base-Chinese 作为词嵌入模型，从元数据字段本身、属性值两个角度，分析元数据之间的语义相似度和距离，并基于距离矩阵实现层次聚类，自动构建元数据之间的映射对应关系，实现名称或属性相似的图书元数据之间的语义聚合。实验结果表明，该模型识别的映射关系精准度达到了 93.33%，大大降低了元数据聚集与融合过程中需付出的人力成本。此外，图书元数据语义聚合方式获得的结果具备横向扩展的空间，聚合过程亦可迭代复用，在其他信息聚合场景也具有一定的兼容性和通用性。

[关键词] 图书元数据；层次聚类；BERT 模型；语义相似度；语义距离

[分类号] G250.7

Research on Semantic Aggregation of Book Metadata based on Hierarchical Clustering

Peng Xianzhe, Shi Jin
School of Information Management, Nanjing University Nanjing 210023

[Abstract] Achieving deep fusion of heterogeneous book resources from multiple sources is crucial for expanding the breadth and comprehensiveness of library services and promoting the development of intelligent library systems. Among these challenges, the semantic aggregation of diverse and differently named book metadata plays a pivotal role in facilitating

·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”（项目编号：21BTQ012）的研究成果之一。

the deep integration of book information from various sources. This study aims to address these challenges by conducting a comparative analysis of the differences in metadata distribution across various platforms. Leveraging BERT-Base-Chinese as the word embedding model, we analyze the semantic similarity and distance between metadata using both metadata fields and attribute values. By employing a hierarchical clustering approach based on the distance matrix, we automatically establish mapping relationships among metadata instances, enabling the semantic aggregation of book metadata with similar names or attributes. Experimental results demonstrate an impressive mapping relationship precision rate of 93.33%, significantly reducing the human effort required during the metadata aggregation and fusion process. Furthermore, the proposed semantic aggregation approach for book metadata exhibits extensive applicability, allowing its iterative reuse in other information aggregation scenarios while maintaining compatibility and generality.

[Keywords] Book Metadata; Hierarchical Clustering; BERT Model; Semantic Similarity; Semantic Distance

1 引言

图书作为知识的实际载体与传播介质，不同平台持有的图书信息之间虽然通常预留了访问获取链接，但多源异构的图书信息由于元数据格式名称并不统一，使其在内容层面的关联整合层次较浅，呈现出了“联而不合”的局面。针对于此，通过构建不同平台的图书元数据字段的映射关系，可拓展资源聚集融合的范围，实现多源图书信息的自动聚合以构建更为完整多样的著录信息，形成高度准确、广泛全面、开放共享的融合服务资源，适应“十四五”发展规划中全国智慧图书馆体系的发展与完善[1]，响应“以人为本、绿色发展、方便读者”的号召[2]，提高图书服务的广度和深度，解决图书信息之间“联而不合”的问题。

当前，图书资源检索发现系统主要是通过在多个开放获取平台中搜集海量图书数据，以预索引的形式向用户提供一站式检索服务[3]，它集成了多来源的图书元数据资源，以元

数据仓储为基础[4, 5]，将元数据作为图书资源发现系统的核心。而如何构建异构元数据的映射与规范，使之形成统一元数据标准，是多来源信息实现自动化语义聚合的关键所在。为此，本项研究以图书作为研究对象，通过分析计算图书元数据文本或属性值之间的语义相似度和距离，实现元数据的层次聚类，促进多来源的图书资源在语义层面的聚集融合，提升不同平台的图书著录信息之间的关联整合深度，力求集成更加多元的图书著录元数据用于加工揭示馆藏资源，为读者提供较为全面完整的信息服务。

2 相关研究

资源聚合，作为将来源不同、观点不同的信息自动或半自动地、及时地转换成能为人机决策提供有效支持的表示的有效方法[6]。其中，图书资源聚合的本质是对不同类型、不同来源、不同语种的资源进行聚集和依据资源的内容特征在语义层次上进行融合[7]，经历了从资源整合到资源聚合的发展过程[8]，早期的数字资源整合研究主要集中在联机公共访问目录（Online Public Access Catalogue，OPAC）[9-11]、跨库检索[12, 13]、文献引用[14]以及元数据[15]等方面。随着关联数据[16, 17]、本体[18-20]和语义网[21, 22]等技术和方法的快速兴起与发展，以资源集成为目的的资源整合研究，逐渐过渡为以资源深度融合为目的的聚合研究，并对资源服务和知识发现相关研究产生了重要影响。

然而，大部分资源聚合技术本身需要大量人工干预，自动化程度较低，往往不具备扩展性和复用性。相比之下，基于元数据著录的资源聚合方法充分利用元数据的结构语义和元数据项本身语义的相似程度，通过预先构建不同资源之间的异构元数据映射关联规则，可以大大降低该方法后续使用过程中人工干预的程度[23]，操作简便可行，提升了资源聚合过程的复用性和扩展性。为此，已陆续有学者开展了基于元数据映射关系的馆藏资源分面语义关联[23]、多平台来源的学位论文元数据的映射转换机制构建[3]、基于ＢＩＢＦＲＡＭＥ２．０的期刊元数据语义聚合[24]、书目资源的细粒度聚合单元元数据框架的构建[25]、基于元数据名称相似度构建决策树实现的文献资源融合[26]，但现有基于元数据的资源聚合研究多关注框架模式、机制验证等理论层面的探讨，实践层面也仅是从案例分析角度出发，且多是通过人工手段预先构建元数据映射规则用于资源聚合，少有基于元数据映射关系自动化构建进而可扩展复用的实验应用层面的研究。

为弥补现有研究的不足，本项研究从实证分析视角出发，以多个平台的图书元数据字段及其属性值为样本，分析字段文本或属性值之间的语义相似度和语义距离,引入层次聚类算法实现图书元数据的语义聚合，促进多来源信息的关联、重组与融合，为图书资源元数据的聚合提供一个可扩展复用和迭代优化的流程框架，加深多源异构信息之间的关联整合程度，集成更为广泛有效地揭示图书资源的元数据项，提升图书服务的广度与深度。

3 研究数据与研究方法

3.1 研究数据

图书检索发现系统中的图书资源来自不同的数据库，每个数据库拥有各自的元数据标准，如何对这些资源进行关联整合，首要工作就是分析各种图书数据库元数据格式的特征，继而实现不同元数据格式的统一转换，建立一种“多对一”的元数据聚合方案。为此，本项研究首先选择图书样本范围，然后筛选图书信息的来源平台，用于图书元数据语义聚合的实证分析过程。

3.1.1 图书样本

为保证图书在多个平台均有记录，本文初步选择了《中文学术图书引文索引》(Chinese Book Citation Index，CBKCI)统计报告 2017 年底公布的三大类“被引排名前十位古籍图书”作为研究样本，其作为“对学术图书的学术质量和影响力进行综合分析”的中文学术图书评价工具[27]，由此筛选获得的图书在多个平台和场景下已有一定的记录基础[28]，因此适用于作为多来源图书的元数据字段聚合的研究样本。

3.1.2 图书服务平台

当下中文图书服务应用平台主要包括图书馆平台（国家图书馆、上海图书馆）、数字资源集成商平台（读秀、超星）、图书电商平台（京东、当当网）、传统馆配商电子书平台（畅想之星）、出版商平台（科学出版社）、网络文学阅读平台（豆瓣阅读、微信读书）等[29]，由于传统馆配商电子书平台专注于为出版方、图书馆搭建信息对接桥梁[30]，出版商平台图书收录范围相对局限，本文主要选择了图书元数据著录范围相对广泛的其他 4 类平台，考虑到不同平台的普及流传程度和开放性、图书数据的可获取性和覆盖范围，这 4 类

平台具体包括国家图书馆、上海图书馆、南京大学图书馆、读秀、京东、当当网、豆瓣这7个图书服务平台，作为上述10本图书的元数据的获取来源渠道。

综上所述，本文以10本图书在7个图书服务平台均有记录作为过滤筛选条件，获得了6本图书在7个平台的元数据字段和对应属性值，共计获得119个图书元数据信息，其中图书元数据分布在当当网平台有14个、京东平台有18个、豆瓣平台有18个、读秀平台有16个、国家图书馆平台有17个、南京大学图书馆平台有19个、上海图书馆平台有17个。

表1 图书的元数据记录字段

平台	元数据字段	多个平台共有的图书记录
上海图书馆	语言；分类；主要作者；文献类型；丛书名；中图法；书名；载体形态；版；出版地；索书号；出版社；团体作者；附注；主题；出版时间；ISBN	西方古典哲学原著选辑 士与中国文化 中国文化要义 现代西方法理学 西方美学史 殷虚卜辞综述
京东	内页插图；丛书名；页数；用纸；正文语种；店铺；编辑推荐；版次；开本；内容简介；包装；目录；商品编码；品牌；出版社；出版时间；产品特色；ISBN	
南京大学图书馆	中图法分类号；ISBN及定价；载体形态附注；一般附注；出版发行项；提要文摘附注；团体责任者；简体题名；其他责任者；版本说明；丛编项；使用对象附注；载体形态项；题名.责任者；marc记录；个人责任者；科图法分类号；书目附注；学科主题	
国家图书馆	语种；分类；摘要；文献类型；来源数据库；关键词；出版.发行地；制作时间；丛编题名；出版发行时间；出版.发行者；载体形态；目次；责任者；版本说明；标识号；所有责任者	
当当网	开本；书摘插画；丛书名；是否套装；编辑推荐；书名；内容简介；目录；包装；纸张；商品详情；国际标准书号ISBN；所属分类；作者简介	
读秀	中图法分类号；出版发行；丛书名；参考文献格式；图书概览；主题词；开本；作者；作者简介；原书定价；引证文献；免责声明；ISBN号；内容提要；试读图书馆文献传递；页数	
豆瓣	页数；装帧；评分；出版年；副标题；我要写书评；论坛；目录；原文摘录；丛书；内容简介；短评；ISBN；出版社；作者；作者简介；丛书信息；定价	

3.2 研究方法

3.2.1 元数据的语义相似度与距离

（1）利用BERT（Bidirectional Encoder Representation from Transformers）进行文本向量表示。

BERT作为Google研究所2018年提出的一种预训练语言模型，将Transformer编码器的双向训练机制应用于语言建模，采用注意力机制动态捕捉输入与输出间的关系[31]。该模型输入层中的每一个字符都由字符向量(Token Embeddings)、段向量(Segment Embedding)和位置向量(Position Embedding)加和产生，可根据输入文本得到每个输入字符在上下文中的向量表示。相比Word2Vec、Doc2Vec等神经网络语言模型，BERT模型可根据不同语境获取字符或词语的动态向量表示，利用丰富的上下文语义特征有效解决多义性问题，并在

多项 NLP 任务中已取得了较好的应用[32]。为此，本文利用 BERT-Base-Chinese 作为词嵌入模型，对中文图书元数据字段和属性值进行向量表示，生成对应的文本向量。

（2）利用余弦相似度计算文本相似度。

在获得两段文本对应的文本向量后，采用余弦相似度计算向量相似度，借以衡量文本之间的相似度：

$$\cos(A,B)=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}=\frac{A\cdot B}{|A|\times|B|}$$

式中，A,B 分别为两段文本的表示向量，A_i表示向量 A 的第 i 维度值，B_i表示向量 B 的第 i 维度值。由此计算得出的相似度分布范围是[-1,1],该值越接近 1,表示两个向量相似度越大，两段文本内容越相似。

（3）图书元数据的语义相似度和距离计算

计算图书元数据的语义相似度，可从元数据的字段名称和对应属性值两个角度展开。基于字段名称的元数据语义相似度，可用两个元数据的字段文本的余弦相似度表示。而根据元数据对应属性值计算的语义相似度，主要可以采用多本图书在两个元数据对应的属性值的余弦相似度集合的最小值、平均值、中值和最大值表示。对于两个文本 A 和 B，基于语义距离 dis(A,B)与语义相似度 sim(A,B)的负相关关系[33]，元数据的语义距离可定义为：

$$dis(A,B)=1-sim(A,B)$$

基于该公式，元数据之间的语义相似度越大，其语义距离越短，相同元数据的语义相似度为 1，语义距离为 0，变化范围为[0,2]。根据上述公式，图书元数据的语义距离，可根据不同的语义相似度计算方式获得。总结来看，本文计算的图书元数据之间的语义距离包括 5 种，包括基于字段内容的 1 种、基于对应属性值的 4 种。

3.2.2 基于语义距离的层次聚类

层次聚类作为一种无监督的聚类算法，其优势是没有加入过多的人工主观判断依据，通过最大轮廓系数获取最优聚类个数[34]，计算出元数据间潜藏的客观规律并加以识别、聚类，使聚类结果更加客观可靠[35]，适用于本项研究文本量较小、计算复杂度不高的语义聚合情况。其基本思路是通过计算不同元数据间的语义距离,在不同层次对元数据样本进行划

分，从而形成一棵有层次的嵌套聚类树[36]，其中，不同类别的原始元数据是树的最低层，树的顶层是一个聚类的根节点[37]。

本文主要以元数据作为聚类对象，通过已构建的图书元数据的 5 种语义距离矩阵，利用层次聚类方法处理距离矩阵以实现元数据类别的划分，自动化生成不同平台的图书元数据的映射对应关系，有助于对比分析与重组融合不同来源的图书信息在不同元数据字段下属性值。对于最终获得的聚类结果的精准性，采用召回率(R)、精确率(P)和调和平均值($F_{measure}$)来计算[38]，公式如下：

$$R=\frac{TP}{TP+FN} \qquad P=\frac{TP}{TP+FP}$$

$$F_{measure}=(1+\beta^2)\times\frac{P\times R}{(\beta^2\times P)+R}$$

式中，TP 代表在真实类别中处于相同类别且在聚类中也被分到相同簇的数据对数量，FP 表示在真实类别中处于不同类别但在聚类中被分到相同簇的数据对数量，FN 为在真实类别中处于相同类别但在聚类中被分到不同簇的数据对数量，由于此项研究中 R 与 P 重要性相当，故取β为 1，$F_{measure}$可用 F1 表示。

4 图书元数据语义聚合

由于待识别的元数据来源广泛，所采用元数据标准、编写习惯等不同，其元素和属性的名称不统一，融合名称各异的元数据是组织与聚合多源图书资源的基础[39]。为实现元数据语义化映射的自动化过程，必须通过模型化方式使得机器可以自动识别不同来源的元数据之间的对应关系[40]。为此，图书元数据的语义聚合过程，可从描述分析多来源图书的元数据、基于元数据字段内容和属性值的层次聚类、融合层次聚类结果三个步骤展开。

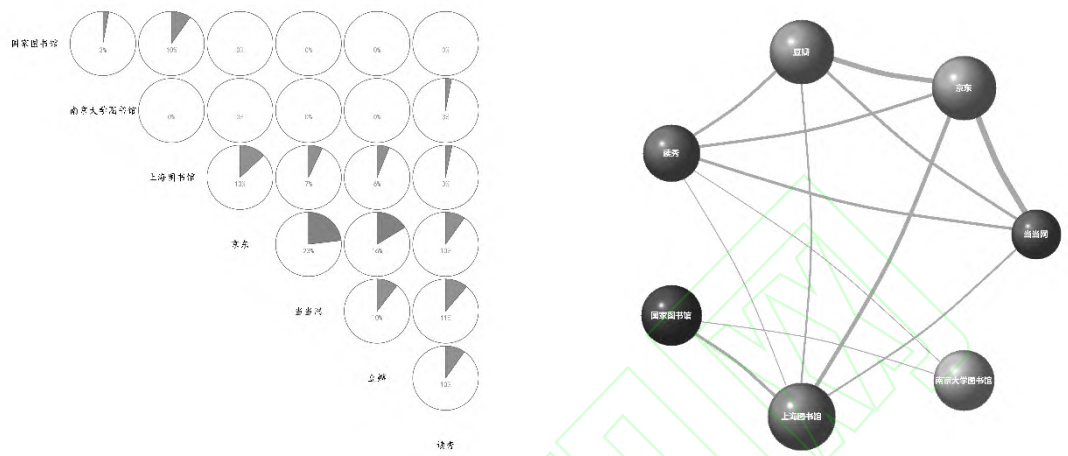
4.1 元数据字段及其属性值

虽然不同平台构建图书元数据的标准和规划不一致，但平台之间的元数据字段仍存在一定的重合，可以直接依据这些重合的元数据字段用以融合不同平台的图书资源。为此，本项研究采用 Jaccard 系数衡量不同平台的图书元数据的重合情况[41]，公式如下：

$$Jaccard(A,B)=\frac{|A\cap B|}{|A\cup B|}$$

式中，A 是平台 A 拥有的图书元数据集合，B 是平台 B 拥有的图书元数据集合。分析来自 7 个图书服务平台的 119 个图书元数据字段，共有 92 项不同名称的图书元数据字段，结

果表明（见图 1a），图书馆类型平台与其他类型平台的元数据重合度较低，尤以国家图书馆和南京大学图书馆最为显著，而电商和网络阅读类型的图书平台之间的元数据重合度相对较高，以京东平台最为显著，但平台之间的图书元数据 Jaccard 系数不超过 23%，且大多在 10%以内。



a 不同平台元数据的 Jaccard 系数　　b 不同平台元数据共现网络

图 1 不同图书服务平台之间的元数据重合情况

此外，在不同平台图书元数据的共现网络中（见图 1b），节点大小代表平台拥有的图书元数据个数，两个平台的连线粗细代表它们共有的图书元数据个数，结果表明，豆瓣、读秀、京东、当当网之间共有的图书元数据虽然较多，但不超过 6 个，大部分平台之间重合的图书元数据较少，基本保持在 3 个以下。整体看来，7 个平台的元数据字段共现网络仍较为稀疏，无法基于重合的元数据实现多个平台图书资源的关联整合。

总结不同图书服务平台之间的元数据重合情况，直接根据不同平台共有的图书元数据实现图书资源的关联整合程度较浅，仅能完成少部分元数据对应属性值的融合和重组。造成这种结果的原因很大程度上在于不同平台的服务场景不同，或如豆瓣平台关注图书的点评交流，或如京东平台强调图书的售卖交易，导致不同平台对图书资源的揭示角度存在差异，表现为著录元数据的名称和种类存在较大的区别。为此，有必要在图书元数据字面匹配的关联方法的基础上，依据图书元数据字段内容或对应属性值的语义距离构建它们之间的映射对应关系，加深图书资源的关联整合程度。

4.2 基于字段内容的层次聚类

对于正式出版的图书，供应商一般可提供 MARC 格式的书目记录，但对于线上图书的元数据，往往采用的是供应商建库时自定义的格式，或称非 MARC 格式，因此，即使是同类图书资源，不同供应商提供的元数据记录在格式和命名上仍可能存在较大差异[42]，如描述图书所属丛书的元数据的字段名称就包括丛书信息、丛书名、丛书三种。虽然不同平台描述图书同一内容的元数据的格式不同、名称不统一，但它们对应的字段内容之间的语义距离理应是相近的。为此，可通过图书元数据字段内容的语义距离实现层次聚类，构建图书元数据的映射对应关系。

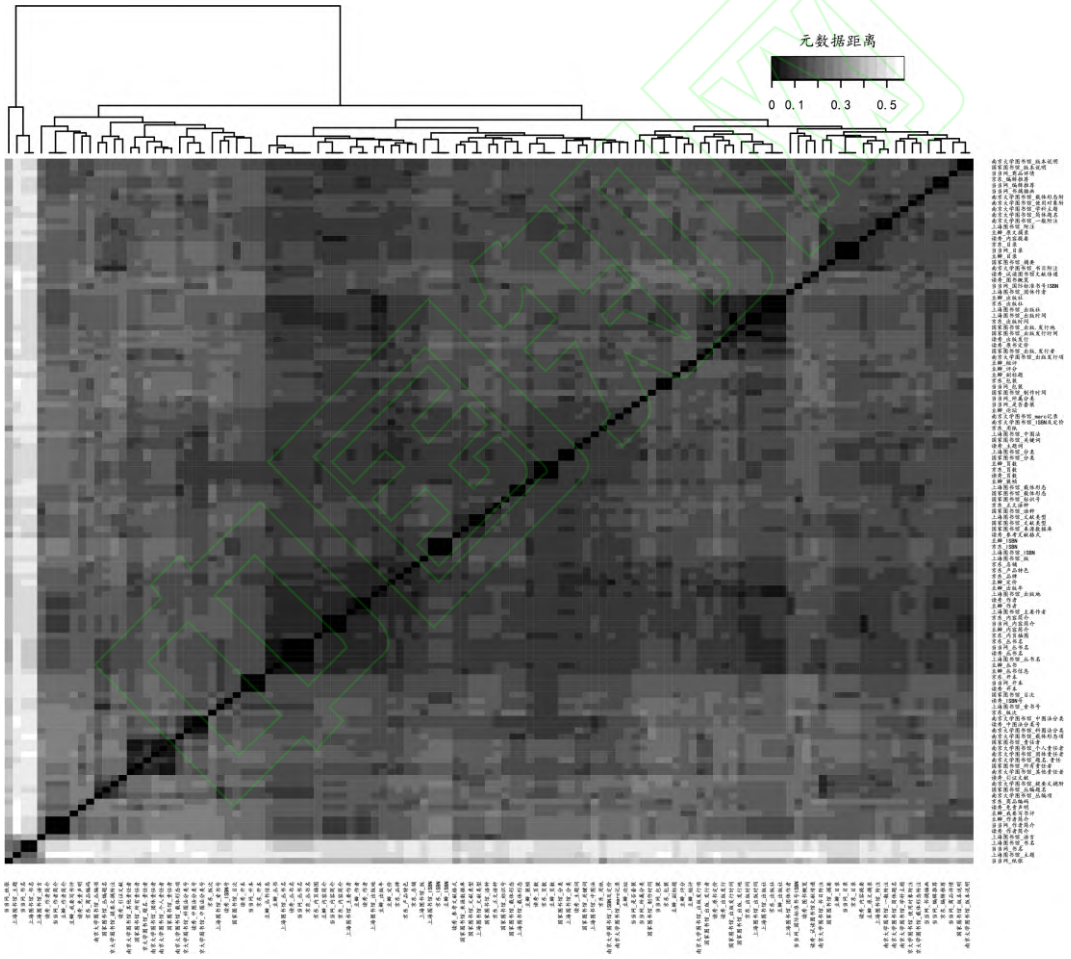


图 2 图书元数据字段内容之间的语义距离热力图

根据图书元数据的语义距离构建的热力图（见图 2）发现，大量图书元数据之间的语义相似度保持在 0.7 以上，语义距离一般在 0.3 以下，图书元数据有进一步融合的空间，如描述图书所属丛书信息的元数据字段包括丛书、丛书名、丛书信息，虽然名称不一致，

但元数据之间的语义相似度超过了 0.9，语义距离在 0.1 以下，可据此通过层次聚类方法构建元数据的分类对应关系。

如图 3 所示，基于文本的语义距离进行层次聚类的具体过程主要分为三步：（a）计算选定截断点语义距离获得的聚类结果的轮廓系数曲线，获得保证轮廓系数最大的最优截断距离；（b）将获得的最优截断语义距离作为判断元数据是否属于同一类别的标准，短于或等于此距离归为同类，否则归为不同类别，以此生成以不同颜色区分类别的层次聚类树；（c）依据层次聚类树，生成聚类结果列表。基于图书元数据字段内容的层次聚类结果表明，将元数据分为了 79 种，产生映射对应关系的元数据总计 21 个，部分名称不同但描述内容一致的图书元数据被划分成同一类，包括描述图书所属丛书信息、出版发行信息、作者或责任者信息、附注信息等元数据字段。

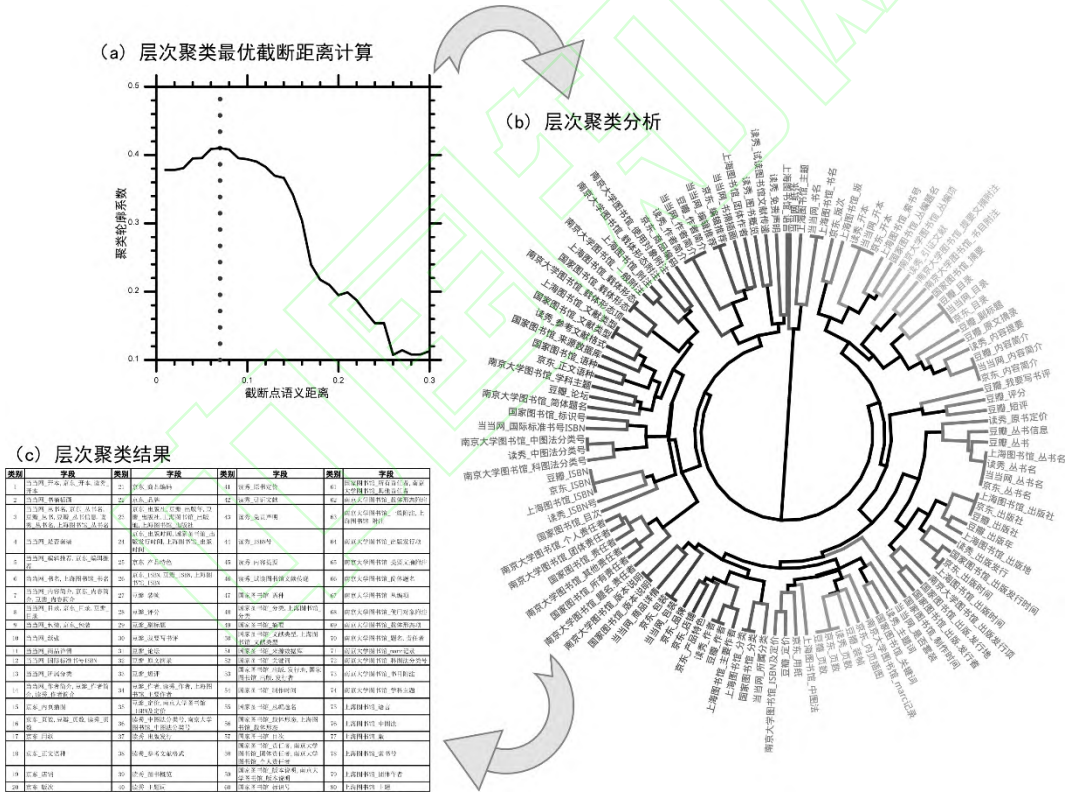


图 3 基于图书元数据字段内容的层次聚类分析过程

虽然基于图书元数据字段内容的层次聚类分析促成了部分相同类别的元数据字段的合并，但由于元数据字段文本短、包含信息量少的特点，因此仍存在大量的同类元数据仍被划分为不同类别，包括读秀平台的 ISBN 号、国家图书馆平台的丛编题名以及大量描述图书主题的元数据等等。与元数据字段内容不同，元数据对应的属性值可由多条记录构成，且

部分属性值文本相对较长，将元数据对应的属性值之间的语义距离作为衡量元数据类别划分的依据，有助于减小同类元数据被确定为不同类别的错误率。

4.3 基于属性值的层次聚类

基于图书元数据对应属性值计算的语义距离共计有四种，分别为根据元数据多条属性值记录计算得到的语义距离值集合中的最小值、平均值、中值和最大值。根据四种语义距离获得的热力图（见图 4），元数据对应属性值的语义距离的变化范围为[0,1]，相比字段内容变化范围更大。不同平台描述图书同一内容的元数据的格式不同、名称不统一，其中某些元数据对应的字段内容之间的语义距离虽然较大，但基于属性值的语义距离却较小，尤以字段文本内容较短的元数据最为显著，如目录和目次、丛编项和丛书名、内容提要和内容简介等。为此，通过图书元数据对应属性值的语义距离实现层次聚类，可弥补基于字段内容实现图书元数据层次聚类结果的不足。

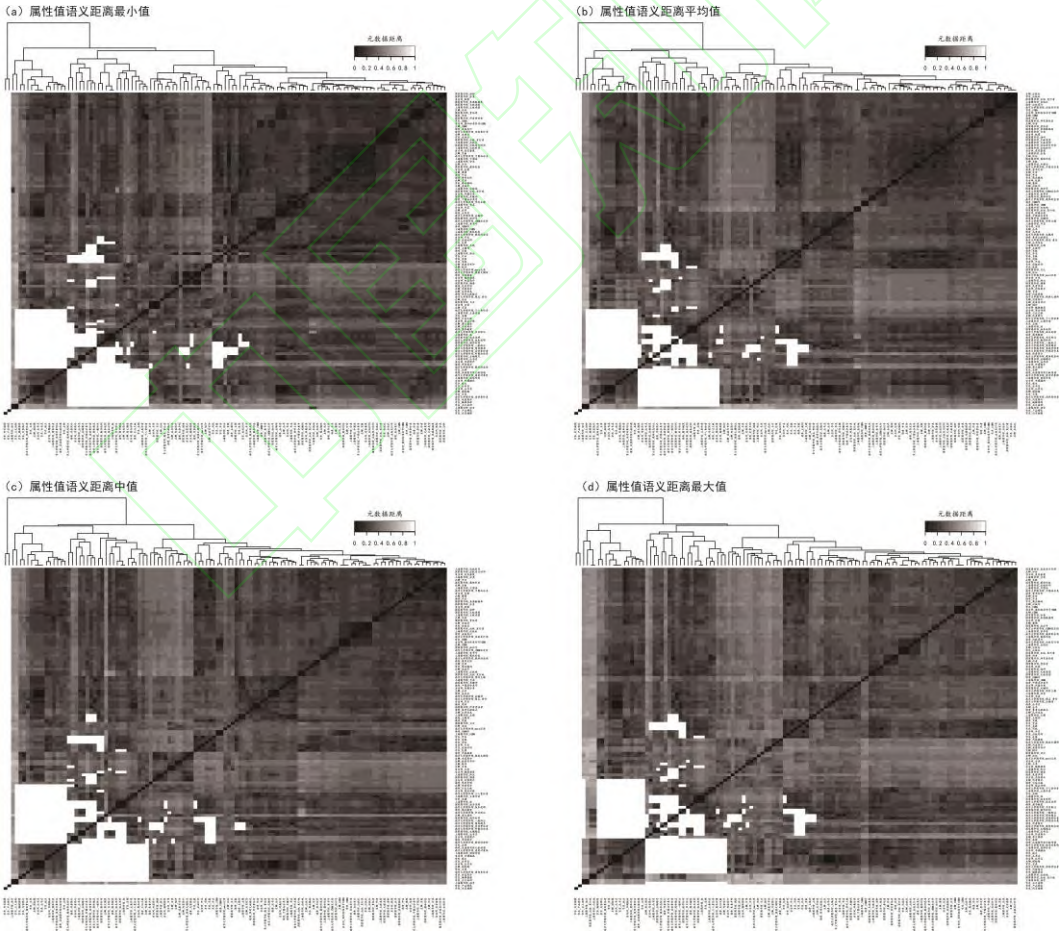


图 4 图书元数据对应属性值之间的不同种类语义距离的热力图

参照基于文本的语义距离进行层次聚类的具体流程，分别获得了基于图书元数据对应属性值的四种语义距离的层次聚类树。分析图 5 发现，之前通过图书元数据字段语义距离的聚类结果中被误分为不同类别的元数据，在层次聚类树中被正确地分为了同一类别，如中图分类号、分类、中图法字段均被划分为了同一类别。此外，根据四种类型的图书元数据对应属性值的语义距离构建映射关系的阈值条件不同，即实现最优层次聚类的截断语义距离不统一，层次聚类的结果也存在较大的差异性。

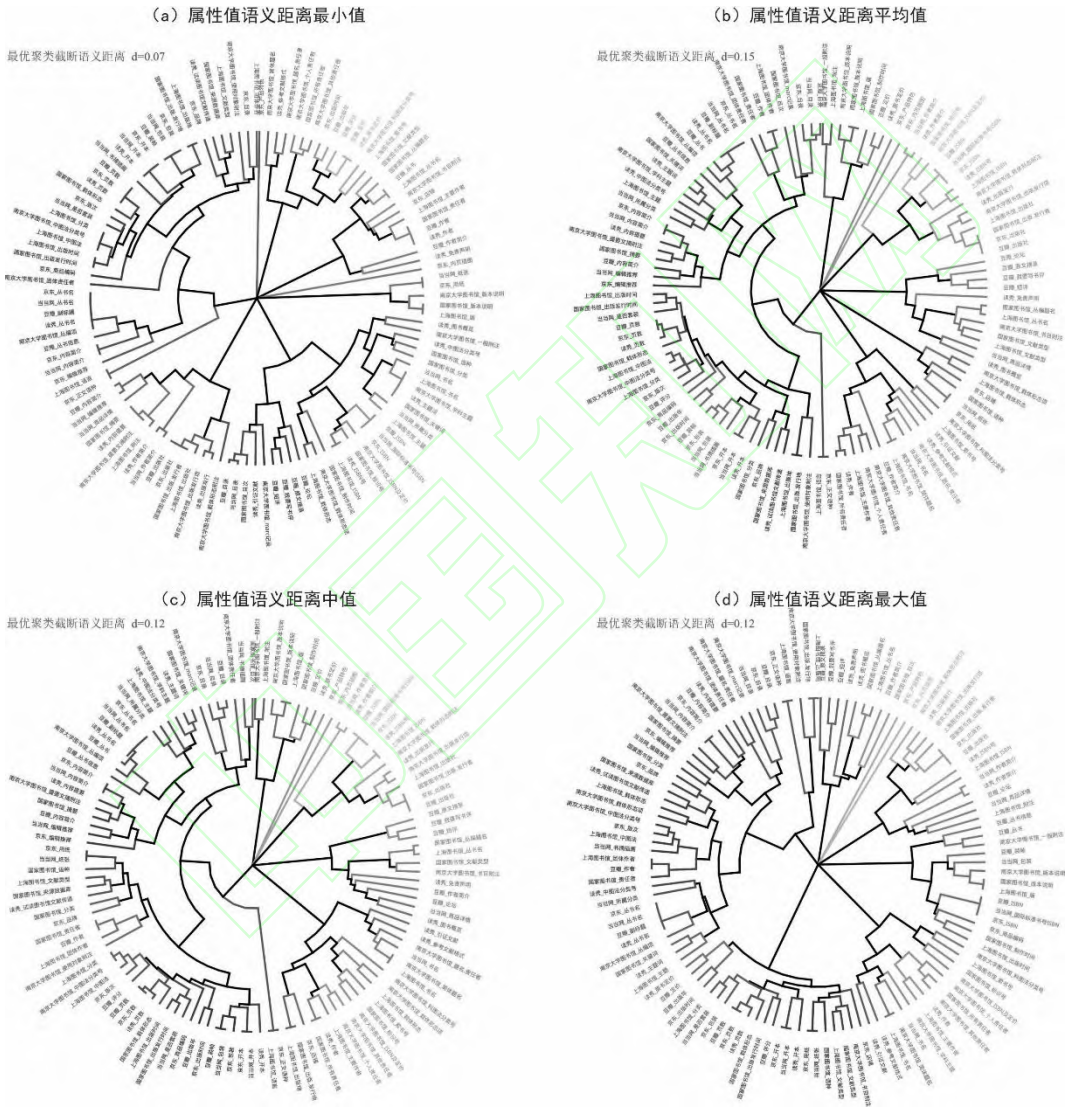


图 5 基于图书元数据对应属性值的语义距离获得的层次聚类树

由于聚类截断的语义距离不同，基于语义距离的最小值、平均值、中值和最大值得到的聚类种数依次为 65、53、56、58，产生映射对应关系的元数据总数分别为 45、63、58、58，不同的聚类结果有各自的优势，拥有其独有准确的图书元数据对应关系，如 5（a）中

目录与目次字段、图 5（b）中试读图书馆文献传递与来源数据库字段、图 5（c）中来源数据库与文献类型字段、图 5（d）中内容提要与内容简介字段等。为此，有必要进一步融合不同类型的聚类结果，提高建立元数据映射对应关系的全面性。

4.4 聚类结果融合

表 2 基于图书元数据不同类型语义距离的层次聚类结果

聚类方式	元数据字段	合并字段个数	聚类方式	元数据字段	合并字段个数
基于元数据属性值语义距离最小值的层次聚类实现的元数据映射关系	丛书名,副标题,丛编项,丛书	4	基于元数据属性值语义距离中值的层次聚类实现的元数据映射关系	页数,载体形态,载体形态项	3
	编辑推荐,内容简介	2		正文语种,语言	2
	目录,目次	2		版次,中图法分类号,分类,中图法	4
	包装,装帧	2		出版社,出版.发行者	2
	纸张,用纸	2		我要写书评,短评	2
	国际标准书号 ISBN,ISBN,ISBN 号,标识号	4		作者,责任者,团体作者,所有责任者	4
	正文语种,语言	2		定价,原书定价	2
	出版社,出版.发行者	2		出版发行,出版发行项	2
	作者,责任者	2		主题词,关键词,学科主题	3
	定价,原书定价	2		摘要,提要文摘附注	2
	出版发行,出版发行项	2		来源数据库,文献类型	2
	主题词,关键词,学科主题,书名	4		出版.发行地,出版地	2
	内容提要,摘要,提要文摘附注,附注	4		版本说明,版	2
	出版.发行地,出版地	2		标识号,ISBN 及定价	2
	出版发行时间,出版时间	2		其他责任者,个人责任者	2
	版本说明,版	2	基于元数据属性值语义距离最大值的层次聚类实现的元数据映射关系	书摘插画,版次,中图法分类号,中图法	4
	所有责任者,其他责任者	2		丛书名,副标题,丛编项,丛编题名	4
	中图法分类号,分类,中图法	3		是否套装,包装,分类,品牌	4
基于元数据属性值语义距离平均值的层次聚类实现的元数据映射关系	丛书名,副标题,丛书,丛编项,丛编题名	5		纸张,用纸	2
	是否套装,出版发行时间,出版时间,出版年	4		国际标准书号 ISBN,ISBN	2
	书名,简体题名	2		页数,载体形态	2
	内容简介,内容提要	2		正文语种,语言	2
	包装,装帧	2		出版社,出版.发行者	2
	纸张,用纸	2		出版时间,出版年,制作时间	3
	国际标准书号 ISBN,ISBN,ISBN 号	3		评分,出版发行时间	2
	所属分类,主题	2		我要写书评,短评	2
	页数,载体形态,载体形态项	3		丛书,一般附注	2
	正文语种,语言	2		作者,责任者,团体作者,主要作者	4
	版次,中图法分类号,分类,中图法	4		定价,原书定价	2
	出版社,出版发行,出版.发行者,出版发行项	4		出版发行,出版发行项	2
	我要写书评,短评	2		主题词,关键词	2
	作者,责任者,团体作者,所有责任者,主要作者	5		试读图书馆文献传递,来源数据库	2
	定价,原书定价	2		摘要,提要文摘附注	2
	主题词,关键词,学科主题	3		文献类型,书目附注	2
	试读图书馆文献传递,来源数据库	2		出版.发行地,出版地	2
	摘要,提要文摘附注	2		版本说明,版	2
	文献类型,书目附注	2		所有责任者,其他责任者,个人责任者	3
	出版.发行地,出版地	2		团体责任者,题名.责任者	2
	版本说明,版	2		简体题名,书名	2
	标识号,ISBN 及定价	2	基于元数据字段内容语义距离的层次聚类实现的元数据映射关系	丛书名,丛书,丛书信息	3
	一般附注,附注	2		出版社,出版年,出版地	3
	其他责任者,个人责任者	2		出版时间,出版发行时间	2
基于元数据属性值语义距离中值的层次聚类实现的元数据映射关系	丛书名,副标题,丛书,丛编项,丛编题名	5		作者,主要作者	2
	是否套装,出版发行时间,出版时间,出版年	4		定价,ISBN 及定价	2
	书名,简体题名	2		出版.发行地,出版.发行者	2
	包装,装帧	2		责任者,团体责任者,个人责任者	3
	纸张,用纸	2		所有责任者,其他责任者	2
	国际标准书号 ISBN,ISBN,ISBN 号	3		一般附注,附注	2
	所属分类,主题	2			

对比基于图书元数据之间的 5 种语义距离实现的层次聚类详细结果（见表 2），图书元数据的映射对应关系差别较大，仅选择其中一类作为最终聚类结果会存在较大的疏漏。针对这个问题，通过将 5 种聚类结果中存在交集的元数据映射关系结果求并集，可有效融合扩展元数据之间的对应关系，如融合基于图书元数据对应属性值语义距离的最小值、平均值的层次聚类结果中[作者，责任者]、[所有责任者，其他责任者]、[作者，责任者，团体作者，所有责任者，主要作者]和[其他责任者，个人责任者]四组元数据映射关系结果，生成作者、主要作者、责任者、团体作者、所有责任者、其他责任者、个人责任者 7 个图书元数据的映射对应关系。通过扩展融合不同类型的层次聚类结果，有助于提高元数据映射关系的全面性。

但是，5 种层次聚类详细结果中构建的元数据映射关系并不一定全部准确，如[主题词，关键词，学科主题，书名]、[是否套装，出版发行时间，出版时间，出版年]、[是否套装，包装，分类，品牌]等。在融合不同类型的聚类结果时，基于错误的元数据映射关系将会扰乱其他聚类结果，导致原本被正确分为两类的图书元数据被化分为同一类，如在以[主题词，关键词，学科主题，书名]这个错误分类结果去融合其他聚类结果时，将会使[主题词，关键词，学科主题，主题，所属分类]和[书名，简体题名]融合为同一类，使得最终错误构建的元数据映射关系数增加 10 个。

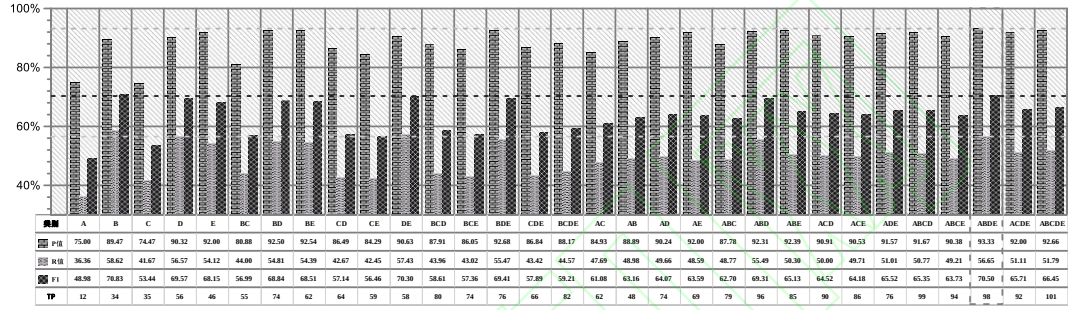
融合不同类型的聚类结果并不能简单地直接将所有结果组合，需考虑不同类型聚合结果的 31 种组合方式，兼顾图书元数据映射关系的精度和准确度，通过召回率和精确率分析与评估多样化的聚合结果。

5 元数据聚合的评估与应用

在根据多种方式获得图书元数据的层次聚类结果后，由于现有研究尚未有专用于自动构建元数据映射关系的同类模型，致使无法对比同类模型展现本模型的性能和质量。为此，有必要根据人工拟定的正确分类方案评估图书元数据聚合结果的精准性，并在此基础上探讨聚合结果的扩展应用空间，彰显聚合图书元数据的意义和价值。

5.1 聚合结果质量评估

为便于后续探讨多样化聚类结果的组合方式，使用不同的字母指示不同的聚类结果。在图 6 中，A 代表了基于元数据字段内容实现聚类的元数据映射关系，B、C、D、E 分别代表了基于元数据属性值的最小值、最大值、平均值、中值实现聚类的元数据映射关系，多个字母的组合代表了各个字母所代表的元数据映射关系的融合结果，P 值、R 值和 F1 分别代表了不同聚合结果的精确率、召回率及其二者的平均值，TP 代表在真实类别中处于相同类别且在聚类中也被分到相同簇的数据对数量。



类别	A	B	C	D	E	BC	BD	BE	CD	CE	DE	BCD	BCE	BDE	CDE	BCDE	AC	AB	AD	AE	ABC	ABD	ABE	ACD	ACE	ADE	ABCD	ABCE	ABDE	ACDE	ABCDE
P值	75.00	89.47	74.47	90.32	92.00	80.88	92.50	92.54	86.49	84.29	90.63	87.91	86.05	92.68	86.84	88.17	84.93	88.89	90.24	92.00	87.78	92.31	92.39	90.91	90.53	91.57	91.67	90.38	93.33	92.00	92.66
R值	36.36	58.62	41.67	56.57	54.12	44.00	54.81	54.39	42.67	42.45	57.43	43.96	43.02	55.47	43.42	44.57	47.69	48.98	49.66	48.59	48.77	55.49	50.30	50.00	49.71	51.01	50.77	49.21	56.65	51.11	51.79
F1	48.98	70.83	53.44	69.57	68.15	56.99	68.84	68.51	57.14	56.46	70.30	58.61	57.36	69.41	57.89	59.21	61.08	63.16	64.07	63.59	62.70	69.31	65.13	64.52	64.18	65.52	65.35	63.73	70.50	65.71	66.45
TP	12	34	35	56	46	55	74	62	64	59	58	80	74	76	66	82	62	48	74	69	79	96	85	90	86	76	99	94	98	92	101

图 6 不同层次聚类方法获得的图书元数据映射关系的精准性

根据图 6，P 值、R 值、F1 的变化范围分别为[74.47%，93.33%]、[36.36%，58.62%]、[48.98%，70.83%]，通过单一方式计算获得的聚合结果的 TP 值普遍较小，构建的元数据映射关系较少，其中 A、C 两类聚合结果的 P 值、R 值、F1 较小，B、D、E 三类聚合结果的 P 值、R 值、F1 相对较高，差异性并不很显著。对于通过组合方式获得的聚合结果，组合的个数越多，TP 值越大，但 P 值、R 值和 F1 并没有出现增大的现象，说明聚类结果的融合提高了元数据映射关系的全面性，却不能保证结果的准确性，其中尤以融合了 C 的聚合结果精准性较低。综合分析，融合了 A、B、D、E 的聚合结果正确构建的元数据映射关系总计 98 个，精确率高至 93.33%，召回率和 F1 值也较高，适合作为最终扩展融合的聚类结果（见表 3）。

表 3 最终获得的层次聚类结果及正确分类方案

方式	元数据字段	合并字段个数	方式	元数据字段	合并字段个数
组合 ABDE 获得的聚类结果	出版社,出版发行项,出版发行时间,出版.发行者,出版发行,出版时间,出版.发行地,是否套装,出版地,出版年	**10**	正确分类	丛书名,副标题,从编项,丛书,从编题名,丛书信息	**6**
	个人责任者,团体责任者,责任者,所有责任者,主要作者,其他责任者,团体作者,作者	**8**		编辑推荐,内容简介,内容提要,摘要,提要文摘附注	**5**
	附注,提要文摘附注,一般附注,摘要,内容简介,编辑推荐,内容提要	**7**		主题词,关键词,学科主题,主题,所属分类	**5**
	国际标准书号 ISBN,ISBN 号,ISBN,定价,ISBN 及定价,原书定价,标识号	**7**		出版社,出版发行,出版.发行者,出版发行项	**4**
	从编项,丛书名,丛书信息,副标题,丛书,从编题	**6**		出版发行时间,出版时间,出版年,制作时间	**4**

	名				
	学科主题,简体题名,关键词,书名,主题词	**5**		中图法分类号,分类,中图法	**3**
	分类,中图法分类号,版次,中图法	**4**		正文语种,语言,语种	**3**
	文献类型,试读图书馆文献传递,来源数据库,书目附注	**4**		版本说明,版,版次	**3**
	页数,载体形态项,载体形态	**3**		页数,载体形态,载体形态项	**3**
	目录,目次	**2**		试读图书馆文献传递,来源数据库,文献类型	**3**
	装帧,包装	**2**		出版.发行地,出版地,出版社	**3**
	用纸,纸张	**2**		附注,一般附注	**2**
	语言,正文语种	**2**		目录,目次	**2**
	版本说明,版	**2**		包装,装帧	**2**
	所属分类,主题	**2**		书名,简体题名	**2**
	我要写书评,短评	**2**		是否套装,包装	**2**
正确分类	作者,责任者,团体作者,所有责任者,主要作者,其他责任者,个人责任者,题名.责任者,团体责任者	**9**		我要写书评,短评	**2**
	国际标准书号 ISBN,ISBN,ISBN 号,标识号,ISBN及定价,定价,原书定价	**7**		纸张,用纸	**2**

对比人工拟定的正确分类方案，最终聚合结果构建的元数据映射关系准确度高达93.33%，基本囊括了不同来源、不同名称的图书元数据之间的确切对应关系，同时也不免将少数不同类别的语义距离过于接近、信息量有限的元数据集合过度融合，如出版信息包含的出版者和出版时间两个类别。因此，通过层次聚类方法自动化构建的图书元数据映射关系，基本展现了描述同一内容的异源元数据之间的对应状况，后续仅需辅以简单的人工判断即可完成元数据的聚集与融合，大大降低了此过程中需付出的人力成本。

5.2 元数据聚合的应用

基于层次聚类方法的图书元数据语义聚合方式，最大的优势在于聚合结果的可扩展性、聚合过程的可复用性。这两点具体表现在该种聚合方式的开源性和动态性，当需要在原有基础上聚合其他平台的图书信息时，亦可以复用之前执行的聚合过程，根据这些平台图书元数据的字段内容和对应属性值，计算其与原有平台不同类别的元数据集合的语义距离，参考之前计算获得的不同层次聚类方式的截断距离，构建待划分类别的图书元数据与已划分类别的图书元数据之间的映射对应关系。这种动态可扩展的元数据映射关系，有助于扩大资源聚合的范围，提升图书资源的广度和全面性。

此外，这种以多源图书资源为例、通过层次聚类方法构建的元数据语义聚合方式，除具备开放兼容和迭代优化的特点外，还有一定横向扩展的空间，仅需更换元数据的描述对象，选择性地保留或优化图书元数据语义聚合的流程框架，即可应用在其他信息聚合场

景，如学者信息、新闻信息等。由此说明，基于元数据语义距离的聚合流程框架并不局限于特定研究，具备一定的兼容性和通用性，具有潜在的应用价值。

6 结语

图书元数据作为用于揭示图书资源的描述著录字段，聚合分布在不同平台的图书元数据属性及属性值，有助于图书馆提升馆内资源的加工揭示程度，为读者提供深入全面的信息服务。为此，针对当前大部分资源聚合技术自动化程度较低、扩展性和复用性不足的问题，通过引入 BERT 模型计算描述相同内容、名称各异的元数据的字段内容、对应属性值之间的语义相似度，本文设计了基于层次聚类方法的多源图书元数据语义聚合模型，构建了元数据之间的映射对应关系，提高了资源聚合的自动化程度，拓展了模型的适用范围，并以国家图书馆、上海图书馆、南京大学图书馆、读秀、京东、当当网、豆瓣这 7 个图书服务平台的元数据为例进行了效果验证。

实验结果表明，通过层次聚类方法实现语义聚合的模型突破了传统元数据融合多依赖名称相似度的瓶颈，大大提升了元数据映射关系的识别能力和整个过程的自动化水平，并能够按照此流程扩展和融合其他来源的图书元数据信息以及应用在其他信息聚合场景，具备一定的开源性和通用性。但本研究还存在一些问题有待后续改进，主要包括：（1）图书样本选择范围有限，文本较短的图书元数据字段内容和属性值的信息量有限，对于根据 7 本图书获得的元数据聚合结果影响较大，样本范围有待于进一步扩大，方可提高模型的精准性；（2）仅构建了中文图书资源元数据的映射对应关系，后续需探讨同一类元数据的名称和对应属性值的融合方式，增强模型的完整性。

参考文献

[1] 曹海霞，侯新宇，杨洋等. 展望“十四五”，促进智慧图书馆大发展——第二届中国高校智慧图书馆(馆长)论坛会议综述[J]. 新世纪图书馆，2021(10)：93-96.

[2] 王世伟. 未来图书馆的新模式——智慧图书馆[J]. 图书馆建设，2011(12)：1-5.

[3] 葛梦蕊，杨思洛，李超. 学位论文资源发现系统多源元数据映射研究[J]. 图书情报知识，2018,183(03)：45-54.

[4] 窦天芳，姜爱蓉. 资源发现系统功能分析及应用前景[J]. 图书情报工作，2012,56(07)：38-43.

[5] 秦鸿，钱国富，钟远薪. 三种发现服务系统的比较研究[J]. 大学图书馆学报，2012,30(05)：5-11+17.

[6] Boström, H.，Andler, S. F.，Brohede, M.， et al. On the Definition of Information Fusion as a Field of Research[J]. Neoplasia, 2007,12(2)：98-107, IN101.

[7]赵蓉英，谭洁. 基于共词分析的馆藏资源语义聚合研究[J]. 情报资料工作，2014 (04)：34-38.

[8] 张云中. 从整合到聚合:国内数字资源再组织模式的变革[J]. 数字图书馆论坛，2014 (06)：16-20.

[9] 范亚芳，边佳平. 对高校图书馆虚拟馆藏资源整合的分析研究[J]. 图书情报工作，2002 (09)：60-63.

[10] 庞跃霞，曹丽娟，丁申桃. 高校图书馆馆藏目录整合方法探讨[J]. 图书馆杂志，2006 (04)：40-42.

[11] 章成志，苏新宁. 信息资源整合的建模与实现方法研究[J]. 现代图书情报技术，2005 (10)：60-63.

[12] NSTL.跨库检索[EB/OL].[2014-07-16]. http://cds.nstl.gov.cn/.

[13] Chang, C.，Lu, W. AGRICULTURAL CROSS LANGUAGES INFORMATIONRETRIEVAL SCHEMA BASED ON MUTI-THESAURUSMAPPING[C]. Computer and Computing Technologies in Agriculture II, Volume 1, 2009, Boston, MA.

[14] ISI web of knowledge[EB/OL].[2014-07-16]. http://wokinfo.com/.

[15] Collections in the age of e-research;realizing potential through curationand aggregation[EB/OL].[2014-07-10]. http://www.clir.org/dlf/forums/fall2010/22PalmerDLF.pdf.

[16] 董坤，谢守美. 基于关联数据的 MOOC 资源语义化组织与聚合研究[J]. 情报杂志，2016,35(06)：177-182.

[17] 赵芳. 基于关联数据的网络社区学术资源聚合模式研究[J]. 图书馆学研究，2016,381(10)：49-52+101.

[18] Carlsson, C.，Brunelli, M.，Mezei, J. Decision making with a

fuzzy ontology[J]. Soft Computing, 2012,16(7): 1143-1152.

[19] Carlsson, C., Mezei, J., Brunelli, M. Fuzzy Ontology Used for Knowledge Mobilization[J]. International Journal of Intelligent Systems, 2013,28(1): 52-71.

[20] Carlsson, C., Brunelli, M., Mezei, J. Fuzzy ontologies and knowledge mobilisation: Turning amateurs into wine connoisseurs[J]. International Conference on Fuzzy Systems, 2010: 1-7.

[21] Nebot, V., Berlanga, R. Building data warehouses with semantic web data[J]. Decision Support Systems, 2012,52(4): 853-868.

[22] Oualhi, O. L., Mohamed T. Dynamic Generation of Adaptative Teaching Material for Semantic Web Approach[C]. IADIS Multi Conference on Computer Science and Information Systems, 2012.

[23] 黄文碧. 基于元数据关联的馆藏资源聚合研究[J]. 情报理论与实践, 2015,38(04): 74-79.

[24] 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 (09): 55-60.

[25] 卫宇辉. 基于细粒度聚合单元元数据的书目资源聚合研究[J]. 国家图书馆学刊, 2020,29(06): 90-101.

[26] 李静, 胡潜, 李想等. 基于决策树的多源文献元数据融合研究[J]. 图书情报工作, 2022,66(06): 118-125.

[27] 叶继元. 《中文图书引文索引·人文社会科学》示范数据库研制过程、意义及其启示[J]. 大学图书馆学报, 2013,31(01): 48-53.

[28] 李明, 李江, 陈铭等. 中文学术图书引文量与 Altmetrics 指标探索性分析及其启示[J]. 情报学报, 2019,38(06): 557-567.

[29] 张会田. 纸电融合模式下的中文电子书馆配应用平台建设[J]. 图书馆学研究, 2021 (07): 51-58.

[30] 杜玉玲, 赵旭鹏. 国内馆配电子书平台 PDA 方案对比分析[J]. 图书馆学研究, 2018,423(04): 40-47.

[31] Devlin, J., Chang, M.-W., Lee, K., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J].

ArXiv, 2019,abs/1810.04805.
[32] 赵展一，李贞贞，钟永恒等. 融合专利类别与语义信息的企业潜在技术关系测算方法研究[J]. 情报理论与实践，2023,46(03)：200-208.
[33] 葛斌，李芳芳，郭丝路等. 基于知网的词汇语义相似度计算方法研究[J]. 计算机应用研究，2010,27(09)：3329-3333.
[34] 张靖，段富. 优化初始聚类中心的改进 k-means 算法[J]. 计算机工程与设计，2013,34(05)：1691-1694+1699.
[35] 陈娟，吴卓青，邓胜利. 基于层次聚类法的“知乎”用户细分与行为分析[J]. 情报理论与实践，2018,41(07)：111-116.
[36] Arslan, O.，Guralnik, D. P.，Koditschek, D. E. Coordinated Robot Navigation via Hierarchical Clustering[J]. IEEE Transactions on Robotics，2016,32(2)：352-371.
[37] Loukides, G.，Shao, J.-H. An Efficient Clustering Algorithm for k-Anonymisation[J]. Journal of Computer Science and Technology，2008,23(2)：188-202.
[38] 张卫，王昊，邓三鸿等. 面向数字人文的古诗文本情感术语抽取与应用研究[J]. 中国图书馆学报，2021,47(04)：113-131.
[39] 葛红梅，徐晶晶，刘靓靓等. 印本与数字期刊元数据差异与融合实践[J]. 图书馆杂志，2022,41(10)：35-41.
[40] 李慧佳，马建玲，张秀秀等. 元数据语义化映射过程研究——以中科院机构名称规范控制库为例[J]. 图书馆论坛，2017,37(12)：72-79.
[41] 张吉玉，张均胜. 考虑时序的单篇科技文献新颖性评估方法[J]. 图书情报工作，2022,66(17)：93-105.
[42] 华苏永. 基于 FOLIO 平台的图书馆编目工作思考[J]. 图书馆杂志，2023,42(03)：52-58+82.

[作者简介]

彭贤哲 1995 年生，南京大学信息管理学院博士研究生，研究方向为目录学、大数据分析与技术，E-mail:pengxz_tm@163.com；
石 进 1976 年生，南京大学信息管理学院教授，博士生导师，研究方向为情报学、大数据分析与技术、智能目录，E-mail: shijin@nju.edu.cn。