引文格式:康力成,姜莉莉.中国省域综合地图集通用内容结构知识构建方法研究[J].武汉大学学报(信息科学版),2026,51(3):475-489.DOI:10.13203/j.whugis20250038

Citation: KANG Licheng, JIANG Lili. Method of Constructing General Content Knowledge System for Provincial Comprehensive Atlas of China[J]. Geomatics and Information Science of Wuhan University, 2026, 51(3): 475-489. DOI: 10.13203/j.whugis20250038

中国省域综合地图集通用内容结构知识构建方法研究

康力成[1,2] 姜莉莉[1,2]

1 中国科学院地理科学与资源研究所地理信息科学与技术全国重点实验室,北京,100101

2 中国科学院大学资源与环境学院,北京,100049

摘 要:地图集作为系统化的科学作品,其内容结构需遵循严谨的逻辑框架与核心主题。内容设计作为地图集编制的核心环节,奠定了整本图集科学性的基础。针对图集内容设计缺乏标准化参考的问题,提出一种融合自然语言处理与知识图谱技术的省域综合地图集通用内容结构体系构建方法,旨在推动地图集内容设计的自动化与智能化。通过收集整理国内外98部区域综合地图集,构建结构化资料库,并基于预训练语言模型(pre-trained language models,PLM)实现子图组文本聚类,利用地图集数据集与知识图谱嵌入增强PLM的语义理解能力。地图集知识与知识图谱的融合训练有效提升了PLM在专题子图组聚类任务下的准确性与鲁棒性,聚类准确率最高提升11.46%。构建了图组—子图组—图幅三级层次化内容结构体系,可为省域地图集内容设计提供科学依据,促进内容标准化与结构规范化,提升地图集的完备性与科学性。

关键词:地图集;内容结构;文本聚类;知识图谱;PLM

中图分类号:P208 **文献标识码**:A **收稿日期**:2025-04-25

DOI:10.13203/j.whugis20250038 **文章编号**:1671-8860(2026)03-0475-15

Method of Constructing General Content Knowledge System for Provincial Comprehensive Atlas of China

KANG Licheng[1,2] *JIANG Lili*[1,2]

1 State Key Laboratory of Geographic Information Science and Technology, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

2 College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: **Objectives**: The content structure of an atlas, as a systematic scientific work, must adhere to a rigorous logical framework and coherent thematic organization. Content design constitutes the foundation of atlas scientific quality. However, the absence of standardized references for atlas content structuring limits consistency and automation. We propose a general content structure system for provincial comprehensive atlases by integrating natural language processing and knowledge graph technologies, aiming to advance intelligent and automated atlas content design. **Methods**: A structured database was constructed based on the collection and organization of 98 regional comprehensive atlases from domestic and international sources. A general content structure framework for provincial atlases was then derived through subgroup-based analysis. Hierarchical text clustering of atlas subgroup texts was conducted using pre-trained language models (PLM). To enhance semantic representation, atlas domain knowledge was incorporated through knowledge graph embedding techniques, thereby improving the semantic understanding capacity of PLM. **Results**: A three-tier hierarchical and standardized content structure system, organized as group-ssubgroups-maps

基金项目:科技部第三次新疆综合科学考察项目(2021xjkk0303)。

第一作者:康力成,硕士,主要从事专题地图与地学信息图谱研究。kanglicheng22@mails.ucas.ac.cn

通信作者:姜莉莉,博士,副研究员。jiangll@igsnrr.ac.cn

was successfully established, comprising 4 groups, 55 subgroups, and 289 map categories. Hierarchical clustering of 2 319 atlas subgroup texts using PLM revealed distinct inflection points in clustering evaluation metrics, corresponding to the typical range of subgroup numbers within thematic groups. Furthermore, the integration of atlas domain knowledge through knowledge graph embedding significantly enhanced clustering robustness and precision, improving overall clustering accuracy by up to 11.46%. **Conclusions:** The proposed general content structure system provides a scientific foundation for provincial atlas content design, enhances structural standardization, and improves the completeness and scientific rigor of atlas compilation. Furthermore, the integration of knowledge graph embedding offers a novel framework for enhancing PLM performance in domain-specific text clustering tasks, contributing to improved clustering accuracy and intelligent atlas design methodologies.

Key words: atlas; content structure; text clustering; knowledge graph; PLM

地图集是根据专题内容、空间范围和时间视角系统地组织并以统一的地图语言组合的一套有针对性的汇编地图[1],其内容设计取决于地图集的性质和用途[2]。按照内容特征分类,可将地图集分为普通地图集、专题地图集和综合性地图集[3],其中综合性地图集是兼有普通地图和专题地图的地图集,内容完备、类型多样、图幅众多,全面系统地展现一个区域的自然环境、资源、经济、社会发展、历史人文特征。综合性地图集的内容特征决定了其内容设计是一个复杂的系统工程[4],要求图集设计过程中深度融合地图学、统计学、自然地理学、经济学等多学科知识。综合性地图集的主题和内容结构设计直接决定了图集的内在特性和系统程度,反映了图集编制者的认知观[5]。尽管19世纪以来国内外地图工作者编制了大量高水平的地图集,积累了丰富的理论与实践经验[6-9],但图集内容设计工作仍然依赖主观经验,标准化范式的缺失使得图集质量评估缺乏统一的标准。多学科知识整合效率的低下,导致地图集编制的经济与时间成本居高不下,图集设计阶段时长往往以月为单位,通常长达1~6个月。本文探索构建综合性地图集标准化内容结构,旨在为综合性地图集内容设计提供系统性的科学参考,提高综合性地图集内容设计的整体质量和效率。

在地图科学技术数字化与智能化的大背景下,地图集"设计—编绘—出版"全过程的智能化成为一种新的发展趋势。王家耀院士指出,通过构建地图集总体设计知识库和推理机制,可实现地图集总体设计的智能化[10]。在省域综合地图集的内容设计方面,黄丽娜等[11]利用自然语言处理技术挖掘省域地图集内容体系中的相似组成部分,提取出省情综合地图集统一的内容体系,制作了省情综合地图集内容图谱。这项研究揭示了中国省情综合地图集的层次化组织特点,提取了图集内容的指纹图谱,为省级尺度的综合地图集内容设计提供了框架指导。需要指出的是,该研究的样本局限于中国省域地图集,尚未涵盖其他区域尺度的综合地图集内容模式,且缺乏基于地理本体论的内容图谱量化评估体系。尽管如此,该研究为区域地图集通用内容结构体系的设计提供了有益的启示和参考。

综合性地图集按区域尺度可分为国家地图集、省域地图集、市地图集等,其内容设计受区域尺度影响较大,不同区域尺度的综合性地图集在内容体系设计、资料选择和处理、地理信息表示方法等方面存在一定差异[12],可供参考的资料数量也存在差异。中国省级行政区数量较多,省域地图集出版量大,编制经验丰富,其中不乏优秀作品,并且内容体系较为规范化。考虑到上述问题,本文着眼于省域地图集内容结构设计开展研究。

1 研究基础与技术框架

1.1 区域综合地图集内容结构分析

现代地图集的内容结构通常遵循图组—子图组—图幅的层次组织。图组定义为一组设计风格、比例尺、投影和主题内容统一的地图,用于展示某一区域的连续或系统性地理信息。子图组是按领域细分的图组,由相同或相近主题的多幅地图、图表与文字组成。图幅指的是单幅地图,以地图形式呈现单个或多个制图指标。

中国内地省域地图集通常由专题地图与普通地图组成。早期的省域地图集内容结构与地图册类似,地图集内容主体为市、区县普通地图,专题地图仅占一小部分,大多未设置图组层。进入21世纪,中国省、市地图集内容设计逐渐成熟,

形成了序图组、人口资源环境图组、经济社会图组、区域详图(普通地图)的基本图组划分。其中,序图组通常包含行政区划、区位、地势、卫星影像、历史沿革、对外交流等内容。人口资源环境与经济社会图组为专题图组,将省域自然地理数据与经济社会统计数据以地图形式可视化展示。部分省域地图集添加了区域发展规划(又称可持续发展、科学发展)图组,介绍当地的发展战略、政策实施、环境保护等内容,具体内容通常依据国家战略和各地区特殊政策而定。少部分省域地图集为了反映区域文化特色,还添加了额外的专题图组,例如《广东省地图集》(2020年版)新增了历史文化图组,《内蒙古自治区地图集》(2023年版)新增了人文撷萃图组。中国的市、区县地图集在内容结构上也往往借鉴省域地图集的框架。

1.2　知识图谱

知识图谱是一种结构化的语义知识库,以“实体—关系—实体”三元组为基本单位,构成网状的知识结构,是现实世界知识的图形表示[13]。知识图谱通过将某一领域的知识体系转化为计算机可识别和存储的结构化语义知识库,有助于促进该领域与计算机科学、统计学、信息科学等学科的深度融合,从而为该学科研究提供重要帮助[14]。在地学领域,多模态地球科学学术知识图谱为地学的数据挖掘和信息检索提供数据和服务支持[15]。知识图谱的跨学科知识关联与整合能力为地图集的设计与制作提供了便利。李连营等[16]利用结构化图集数据与非结构化社交媒体数据,构建了海洋地图可视化方法知识图谱。彭楚骄等[17]从地图集中提取了统计数据的可视化方法,构建了“统计数据—可视化方法”知识图谱,并搭建了专题地图可视化的推荐机制。韩德志等[18]从现有地图集中抽取空间数据、数据特征、可视化维度、应用领域、可视化形式五类知识,构建泛地图可视化知识图谱,建立输入数据与知识图谱内数据类型、空间数据特征、可视化维度的相似度计算模型,实现了泛地图可视化形式的精准推荐。MAI等[19]与钟琳颖等[20]的研究表明,利用领域知识和资料搭建知识图谱,使用文本挖掘技术提取制图要素为叙事地图的设计与制作提供了很大的帮助。

1.3　知识增强预训练模型

预训练语言模型(pre-trained language models,PLM)是预先使用大规模无标注文本数据进行自监督训练的语言模型,生成了泛在领域知识的嵌入表示,能够学习到词语的上下文相关性[21]。PLM在自然语言处理(natural language processing,NLP)领域取得了巨大的成功,在许多NLP任务下表现出最先进的性能。在PLM成为NLP领域研究主流工具的背景下,融合知识图谱与PLM成为了提升PLM性能的一种主要途径,知识增强预训练模型便是二者结合的一种常见方法[22]。知识增强型PLM(knowledge-enhanced PLM,KE-PLM)指的是将外部知识整合到PLM中,以克服预训练模型存在的可解释性差、鲁棒性弱和缺乏推理能力等问题,提升模型在知识密集型下游任务(如实体识别、关系抽取、问答系统等)中的表现[23]。知识图谱是知识增强的一种常见形式,K-BERT[24]、KG-BERT[25]和KEPLER[26]等模型是基于知识图谱的KEPLM,它们有效提高了语言模型在文本分类、聚类等任务下的性能。

在上述研究的基础上,本文将探索文本聚类与知识图谱技术如何应用于省域地图集通用内容结构体系的构建。

2　研究方法

2.1　地图集数据收集与结构化处理

考虑到本文主要研究目标是构建综合性地图集内容结构体系,因此需要了解地图集内容的发展过程、国内外差异,以及尺度差异。因此,共收集国内外综合性地图集98部。其中,国内地图集66部,包括省域、市、区县、流域、经济带地图集,以及《“一带一路”南亚资源环境地图集》。鉴于国内外区域综合性地图集的编制目的均在于从多维度系统展示区域发展状况,其内容体系结构具有较强的共性特征。因此,将国外地图集纳入分析有助于完善中国省域综合性地图集的通用内容结构体系。本文选取32部国外地图集,主要为国家级综合地图集。

不同区域综合地图集的层次结构存在差异,按层次结构划分,存在图组—子图组—图幅、图组—图幅以及子图组—图幅3种内容结构。将98部地图集的内容结构整理为统一的资料库。资料库中图集层次结构需要对齐,按照地图集内容结构,将地图集统一整理为图组—子图组—图幅层次结构。部分地图集存在多层子图组,统一将一级子图组视为子图组。一级以下的子图组视为三级图组和四级图组等,暂时不纳入研究范围。普通地图图组(一般命名为区县详图)的图

幅为区域下辖一、二级行政区划的详细地图,如省域地图集的普通地图是省下辖市、区县地图。由于普通地图图组内容的固定性,本文未将普通地图图组纳入研究范围,仅针对专题地图图组进行研究。

构建统一的资料库不仅要完成结构对齐,还要完成语义对齐与消歧,适当清洗数据,以便后续对资料库文本进行分析研究。首先,将国外地图集内容统一翻译为简体中文。其次,中国的《香港发展地图集》《台湾省地图集》则按照普通话的用词习惯修改文本。在处理地图集目录文本时,若地名标签出现较频繁(单独作为定语出现,且不是"粤港澳大湾区""长江经济带"等专有名词),则会去除地名标签。时间标签也会尽可能删除,表示不同时间相同制图指标变化的地图,如历史沿革地图,按制图指标合并为一个图幅。若目录中专题图幅末尾出现"地图""图"和"示意图"等地图定义标签,也会删去。此外,尽可能去除标点符号。缺主语或指代过于模糊的文本,参考上级内容补充上下文。删除只有文字介绍,没有地图或图表的子图组尽可能删除区域、事件、数量标签。

资料库中收集98部区域综合地图集的内容结构,包括485个图组、3 063个子图组,选择性收集了其中40部地图集的5 689个图幅。

2.2 基于文本聚类的地图集通用内容体系构建

聚类分析是一种无监督学习技术,旨在将数据集中的样本或特征变量根据相似性进行分组,使得同一类中的元素之间相似度较高,而不同类之间的相似度较低。聚类分析的常用算法有*k*-means、层次聚类和基于密度的带噪声应用空间聚类等。文本聚类是聚类分析在文本数据上的应用,根据文本的相似性将文本划分为若干个簇,使得高相似度文本属于同一簇,低相似度文本分为不同簇。通过词嵌入模型将文本转化为词向量,即可采用聚类分析的方法挖掘文本的潜在结构与模式。

在文本聚类中,文本的相似度通常从字面相似度、语义相似度与风格相似度角度衡量,相似度数值通常在0到1之间,值越大,表示文本之间的相似度越高。基于距离的文本聚类利用相似度函数来度量文本对象之间的接近程度,通过分词与特征提取将文本转化为数值向量,计算向量之间的距离或相似度得到文本相似度的度量。余弦相似度是NLP领域最常用的相似度函数[27]。余弦相似度可作为向量距离直接代入聚类计算过程。

PLM通常都具有分词器,可完成中文文本的分词处理与向量化,得到文本的嵌入表示。经过大规模百科文本训练的PLM通过利用大规模文本数据进行自监督学习,捕捉了词语和句子在不同上下文中的语义关系,相似领域的文本更有可能相互出现在各自的上下文中。因此,PLM可以识别文本的领域相关性。

2.3 基于知识图谱嵌入的PLM增强

地图集的内容结构是以地图语言为形式表达的结构化领域知识。而训练PLM使用的是大量非结构化文本数据,这使得它们缺乏结构化的领域知识。因此,需要利用领域知识图谱来引导PLM分析地图集文本,以帮助模型更有效地识别地图集内容的知识关联性。地图集图组、图幅文本通常为短文本(长度大多小于20字),短文本所表达的信息是有限的,缺少上下文信息,这影响了短文本的相似性度量与文本聚类的准确性[28]。百科知识图谱可以补充短文本的背景知识[29],以解决短文本的这种稀疏性问题。本文选用了CN-DBpedia中文通用百科知识图谱[30]作为输入模型的知识图谱,CN-DBpedia是由复旦大学知识工场实验室研发并维护的大规模通用领域百科知识图谱,致力于为机器语义理解补充背景知识。根据地图集的图组、子图组数据从知识图谱中提取了与地图集内容设计有关知识的子图,共获得459 593个三元组,占原知识图谱数据的8.89%。

参考K-BERT、KEPLER、CokeBERT等KE-PLM的技术方法[31],本文设计了一种融合知识图谱嵌入的增强PLM。模型包括嵌入模块、融合模块与训练输出模块,如图1所示。

1)嵌入模块。嵌入模块以TransH算法[32]为核心,将知识图谱中的实体和关系表示为低维向量,以便在向量空间中进行推理和计算。TransE是传统的知识图谱嵌入算法,只能处理实体的一对一关系。TransH是TransE的改进,允许每个关系拥有一个超平面,实体的向量表示可以在该超平面上进行投影,当同一个实体参与不同关系时,可以在不同的超平面上得到不同的表示。与TransE相比,TransH可以处理实体之间多对多的关系,能更准确地表达较为复杂的实体关系。

2)融合模块。融合模块通过将知识图谱嵌入与PLM融合,将结构化知识注入到PLM中。由于知识图谱嵌入的维度与PLM的底层嵌入维

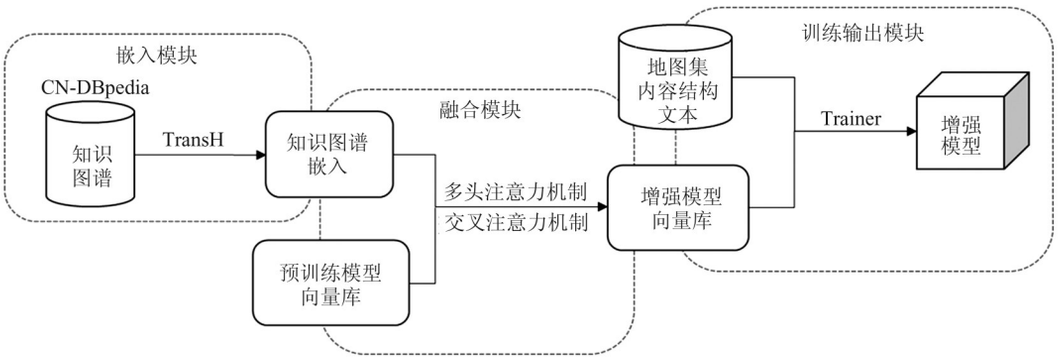


图 1　融合知识图谱嵌入的增强预训练模型

Fig. 1　Enhanced PLM Incorporating Knowledge Graph Embedding

度往往不一致,本文采用非线性变换层将知识图谱嵌入与 PLM 的嵌入维度对齐。非线性变换通过多层感知机将知识图谱嵌入映射至 PLM 的维度空间,并借助联合训练优化语义一致性,实现跨模态特征融合。

为了有效整合知识图谱嵌入和 PLM 的语义表示,设计了双重注意力机制。由多头注意力机制[33]和交叉注意力机制两个注意力机制模块组成。多头注意力机制是一种重要的神经网络组件,它可以通过并行的多个注意力头来学习不同的子空间表示。在融合层的上下文中,多头注意力机制用于结合文本表示和知识图谱嵌入,计算其交互信息。给定查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$,注意力的计算过程[33]可以表示为:

$$A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

式中,$A(\cdot)$ 为单头缩放点积注意力函数;d_k 是键向量的维度,用于缩放点积以防止值过大。在多头注意力机制中,上述过程被扩展为多个头,每个头都有其独立的参数集:

$$M(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=C(h_1,h_2\cdots,h_i)\boldsymbol{W}^0 \tag{2}$$

$$h_i=A(\boldsymbol{Q}\boldsymbol{W}_Q^i,\boldsymbol{K}\boldsymbol{W}_K^i,\boldsymbol{V}\boldsymbol{W}_V^i) \tag{3}$$

式中,$C(\cdot)$ 为拼接函数;$M(\cdot)$ 为多头注意力映射函数;$\boldsymbol{W}_Q^i$、$\boldsymbol{W}_K^i$ 和 $\boldsymbol{W}_V^i$ 是投影矩阵;$\boldsymbol{W}^0$ 是输出变换矩阵。通过多个头的并行计算,多头注意力机制在不同的表示子空间中同时执行注意力操作。

交叉注意力机制是多头注意力机制的一个变种,专门用于融合不同来源的嵌入信息。在本文中,交叉注意力机制用于将 PLM 的语义表示与知识图谱嵌入进行双向交互,两种交叉注意力操作合并生成一个融合的表示。这种融合方法可以捕捉文本和知识之间的复杂交互关系,从而提高模型的语义表示能力和分类性能。

3)训练输出模块。训练输出模块基于 Transformers 库[34]的 Trainer 构建。Transformers 库是 Hugging Face 开发的开源 Python 库。Trainer 是 Transformers 库的一个高级接口,支持 Transformer 架构 PLM 的训练和微调。根据下游任务需要建立训练数据集,通过 Trainer 训练与知识图谱嵌入融合的 PLM,输出为新的模型。

3　省域地图集通用内容结构设计

3.1　技术路线

为构建省域地图集普适性内容结构体系,以合理分配专题地图内容,以地图学理论为指导,整合国内外国家、省域、市、县地图集内容设计经验,结合 NLP 技术对地图集结构化资料库展开深度挖掘。研究方案与技术路线如图 2 所示。首先,将地图集资料库分为序图组、专题图组、发展规划图组 3 个部分,利用 PLM 生成子图组语义向量,通过层次聚类将 3 部分各自的子图组归并为子图组类,得到省域地图集通用子图组。随后从图组、子图组中提取出专题子图组训练数据集,融合百科知识图谱微调 PLM,用于专题子图组类的自动识别与语义关联建模,并进一步刻画人口资源环境等要素在不同专题图组中的组织关系。最后,通过计算语义相似度矩阵对图幅资料进行语义归并,得到图幅类,基于语义相似度将图幅类分配至子图组类。最终形成由图组—子图组—图幅的标准化层次结构组成的省域地图集通用内容结构体系。

3.2　基于层次聚类的子图组归并

图组层是对地图集专题内容的概括性区分,将地图集专题内容分为 3～6 个大类。图幅层往往表达的是具体的制图指标,一个领域可以包含多个图幅。现代地图集中出现的子图组层表示地图集内容的具体领域划分,将制图指标按研究

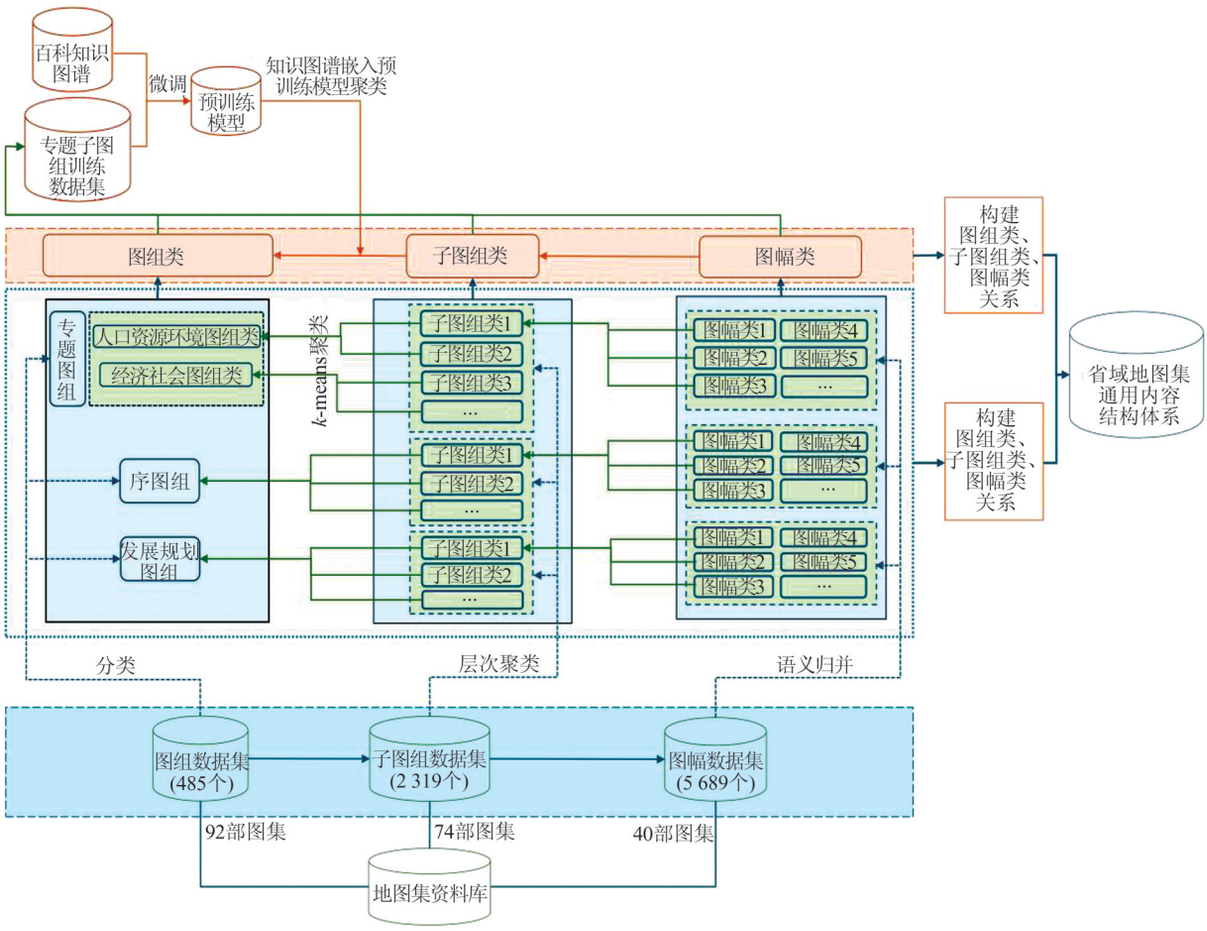


图2 省域地图集通用内容结构设计研究方案

Fig. 2 Research Programon the Design of a General Content Structure for Provincial Atlases

领域归入特定类别。因此,本文从子图组入手,分析地图集内容的多样性及其共性。

从原理上看,*k*-means聚类通过迭代将数据划分为预设的*k*个互斥的簇,划分结果为非层次化的独立簇,相对适用于大规模数据的简单划分。而层次聚类通过逐步合并(自底向上)或分裂(自顶向下)构建树状结构,允许簇之间存在嵌套关系,适合表达数据的多粒度特性,相对适用挖掘数据的多层级主题。在实验中也发现,基于划分的*k*-means聚类方法在将子图组聚类为图组(聚类个数10以下)的任务下取得了较好的结果,而基于层次的层次聚类方法在子图组、图幅语义归并任务上(聚类个数10以上)表现出较好的性能。结合常见聚类方法的特点与实际聚类效果,本文采用层次聚类构建子图组类,采用*k*-means聚类将子图组聚合为图组。首先,采用层次聚类将子图组归并为若干子图组类,再采用层次聚类方法对全部子图组进行归并,形成若干子图组类;在此基础上,进一步采用*k*-means聚类对所得子图组类进行划分,将其聚合为若干大类,即实现子图组类向图组层级的归属分配。实验过程中发现,序图组与发展规划图组中的子图组及图幅文本对聚类结果产生了较强干扰。序图组往往混合地形、历史等多类型专题图幅,主题边界模糊,易导致语义空间分布离散。而发展规划图组则多围绕具体政策文本展开,具有明显的行政表述特征,在对子图组名称进行地名删除等预处理操作后,容易产生语义歧义,从而影响文本向量表示的稳定性。结合§2.1对区域综合地图集内容结构的分析结论,将序图组、专题图组与发展规划图组分别构建聚类模型并独立开展分析,以减少跨主题语义干扰,提高文本聚类结果的准确性与稳定性。

3.2.1 专题图组

专题图组是描述区域自然环境、经济社会、历史人文等专题内容的图组,反映制图区域在上述领域的发展现状。专题图组子图组来自74部图集,共2 198个子图组。每部图集平均有3.30个专题图组,中位数为2个,四分位范围为2~4个。每部图集平均约有29.70个子图组,中位数

为 27 个,四分位范围为 19.25～35.75 个,95% 的范围是 10～64 个。子图组的划分反映了地图集作者对地图集制作涉及的领域知识在整体知识上的划分。基于这个理念,可认为地图集专题图组的一般子图组个数是地图集专题领域知识的细分分类数。因此,首先采用层次聚类方法对子图组进行聚类,生成子图组类,聚类个数范围取 10～64,使用 Hugging Face 平台上常用的 5 种中文 PLM 进行聚类:bert-base-chinese[35]、m3e-base[36]、xiaobu-embedding-v2[37](简称为 xiaobu)、zpoint_large_embedding_zh[38](简称为 zpoint)和 Yinka[39]。在子图组聚类实验中,还添加了 6 部没有图组但具有代表性子图组的地图集,包含 121 个子图组,子图组数据总量为 2 319 个。参考文本聚类相关文献[40-41],采用轮廓系数与邓恩指数两个聚类内部评价指标来评价不同模型、不同聚类数的聚类效果。选用统计学常用的逆方差加权法计算两个指标的方差,取方差倒数为权重,计算聚类个数的加权得分,通过比较加权得分得到最佳模型与最佳聚类个数。加权得分计算公式为:

$$S_i = \frac{as_i + bd_i}{a + b} \tag{4}$$

式中,s_i与d_i分别为某种模型某一聚类数下聚类结果的轮廓系数与邓恩指数;a和b分别为轮廓系数与邓恩指数的权重。在专题图组子图组层次聚类中,a和b分别为 205.36 与 245.94。加权得分变化如图 3 与表 1 所示。

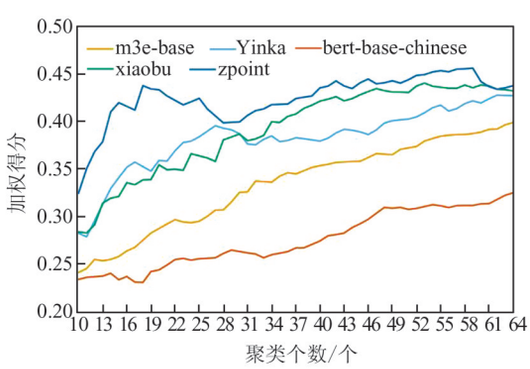


图 3　专题图组子图组层次聚类加权得分

Fig. 3　Weighted Scores for Hierarchical Clustering of Subgroups in the Thematic Group

图 3 显示,使用 zpoint 模型在聚类个数为 59 时,聚类效果最佳,该模型的层次聚类效果在聚类个数 59 处出现拐点。在整体上,zpoint 模型的聚类效果也是最好的,该模型对地图集文本聚类任务的适应性较好,xiaobu 和 Yinka 次之,bert-base-chinese 和 m3e-base 的聚类结果整体上相对较差。实际上,在轮廓系数与邓恩指数得分更高时,簇的大小也更均匀,更不容易出现多个不同内容子图组聚为一类的情况。加权得分对应的聚类数普遍较高的原因是轮廓系数与邓恩指数一般随聚类个数增加而上升,但邓恩指数的波动性较强。两个指标的方差大致都接近 0.004,加权得分权重也较为接近。确定聚类个数为 59 对应的距离阈值 0.69 对专题图组的子图组进行层次聚类,将子图组数据归并为子图组类,每个簇为一个子图组类。完成聚类后计算聚类中心词,对每个类内向量取算术平均,得到几何意义上的类中心作为类的均值向量,通过余弦距离衡量类内每个向量与均值向量的相似性,选择距离最小的样本作为聚类中心词,将聚类中心词作为子图组类名称。观察类名称及其内容,去除其中与省尺度不相关的名称(如“外交”)、包含地名(如“韩国”)的类,将可合并的类进行合并(如“通信”并入“邮政通信”),共删除 10 个类,7 个类被合并。根据类的内容适当修改类名称,如将“水利设施”改为“水资源与水利工程”。筛选出的 42 个子图组类包含 2 140 个子图组,占原专题图组子图组数据的 92.28%,绝大多数的子图组数据得到了

表 1　专题图组子图组层次聚类加权得分前 20 名

Table 1　Top 20 Weighted Scores for Hierarchical Clustering of Subgroups in the Thematic Group

模型	聚类个数/个	轮廓系数	邓恩指数	加权得分	排名
zpoint	59	0.319 1	0.571 2	0.456 0	1
zpoint	58	0.318 1	0.570 8	0.455 3	2
zpoint	57	0.317 8	0.570 5	0.455 0	3
zpoint	55	0.315 6	0.569 3	0.453 4	4
zpoint	56	0.316 2	0.566 6	0.452 2	5
zpoint	54	0.314 2	0.568 0	0.452 0	6
zpoint	53	0.313 6	0.563 4	0.449 2	7
zpoint	52	0.312 0	0.563 2	0.448 4	8
zpoint	46	0.296 8	0.569 0	0.444 6	9
zpoint	51	0.304 4	0.561 2	0.443 9	10
zpoint	49	0.305 9	0.557 9	0.442 7	11
zpoint	42	0.293 4	0.568 1	0.442 5	12
zpoint	60	0.314 6	0.549 2	0.442 0	13
zpoint	45	0.295 5	0.562 9	0.440 7	14
zpoint	48	0.304 4	0.555 3	0.440 6	15
xiaobu	53	0.299 4	0.559 0	0.440 3	16
zpoint	50	0.301 5	0.557 1	0.440 3	17
zpoint	47	0.304 0	0.553 8	0.439 6	18
xiaobu	60	0.316 1	0.541 1	0.438 3	19
xiaobu	58	0.313 5	0.543 3	0.438 3	20

保留。由于zpoint模型的聚类效果较好,序图组与发展规划图组的子图组也将使用zpoint模型进行层次聚类。

3.2.2 序图组

序图组是展示区域地理概况的图组,通过行政区划、地势形态、遥感影像等要素呈现区域地理本底特征。国内地图集常设置序图组作为图集第一个图组,国外的国家地图集通常也会在开头设置国家、领土概况图组。序图组中的图幅往往占据一个展开页,两侧以文字介绍为主,一个子图组往往只有一个图幅。序图组来自77部地图集,每部图集子图组数量平均值为7.53个,中位数与众数为7个,四分位范围为5~9个,中间95%数据范围是3~17个。子图组层次聚类的聚类个数取3~17个。计算聚类结果的轮廓系数与邓恩指数,作为聚类结果评价指标。通过逆方差加权法计算聚类个数的加权得分,a与b分别为662.91与236.56,得到聚类个数为17时加权得分最高,如表2所示。确定聚类个数为17对应的距离阈值对序图组子图组进行层次聚类。计算聚类中心词作为类标签,观察类内容,去除其中与省尺度不相关、包含地名、纯文本页以及与专题图组重复的部分,保留5个子图组类,按类大小排序为:政区、地势、卫星影像、历史沿革、对外交往。

表2 序图组子图组层次聚类加权得分

Table 2 Hierarchical Clustering Weighted Score of Subgroups in the Preface Group

聚类个数/个	轮廓系数	邓恩指数	加权得分
3	0.108 0	0.436 9	0.194 6
4	0.112 9	0.471 8	0.207 3
5	0.117 8	0.424 4	0.198 5
6	0.136 1	0.462 2	0.221 9
7	0.150 4	0.600 0	0.268 6
8	0.175 9	0.488 3	0.258 1
9	0.175 9	0.528 4	0.268 6
10	0.175 9	0.573 0	0.280 4
11	0.175 8	0.587 6	0.284 1
12	0.184 6	0.578 2	0.288 1
13	0.186 5	0.580 9	0.290 3
14	0.204 4	0.567 7	0.299 9
15	0.218 0	0.578 0	0.312 6
16	0.221 4	0.598 9	0.320 6
17	0.225 2	0.616 8	0.328 2

在中国省域地图集中,通常在省域行政区划地图后展示一幅中国地图,命名为“中国政区”或“区位”“地理位置”“某省在中国的区位”,这部分子图组在层次聚类中未能与行政区划内容区分开来,但十分常见,共有54个子图组符合类似名称要求,因此在序图组通用子图组中添加了“中国政区”,最终序图组共有6个通用子图组。

3.2.3 发展规划图组

发展规划图组是具有中国特色的图组,又称可持续发展、科学发展图组,一般展示区域经济政策、发展建设规划、产业布局、生态保护等表示未来发展趋势的内容,裴秀奖金奖作品大多具有发展规划图组。数据来自25个地图集,共164个子图组,每部图集发展规划图组平均子图组个数为6.56,中位数为6,四分位范围5~8,中间95%范围是4~12。子图组层次聚类的聚类个数取中间95%数据范围4~12。计算轮廓系数与邓恩指数评价聚类结果,通过逆方差加权法计算聚类个数的加权得分,a与b分别为2 022.52与241.81,得到聚类个数为10时加权得分最高,如表3所示。确定聚类个数为10对应的距离阈值对发展规划图组的子图组进行层次聚类,计算聚类中心词作为类标签,观察类内容,去除其中与省尺度不相关、包含地名的类以及子图组数量少于6的类,最终保留7个子图组类,根据类的内容适当修改类标签,按类大小排序为:区域经济和产业发展布局、总体规划、乡村振兴、国土空间规划、生态建设、历史文化名城保护、综合交通发展规划。

表3 发展规划图组子图组层次聚类加权得分

Table 3 Hierarchical Clustering Weighted Score for Subgroups in the Development Planning Group

聚类个数/个	轮廓系数	邓恩指数	加权得分
4	0.103 9	0.488 9	0.145 1
5	0.100 8	0.474 2	0.140 7
6	0.118 2	0.532 9	0.162 5
7	0.117 0	0.587 1	0.167 2
8	0.125 1	0.601 2	0.176 0
9	0.139 4	0.629 0	0.191 7
10	0.157 8	0.644 6	0.209 8
11	0.150 7	0.608 7	0.199 7
12	0.158 4	0.639 0	0.209 7

3.3 基于KEPLM的专题图组设计

中国现代省域地图集往往将专题地图分为自然资源环境、经济社会两个部分,因此形成了人口资源环境、经济社会两个反映区域发展现状的固定专题图组。中国现代省域地图集专题图

组内容的分配基本相似,子图组对应的图组基本可作为子图组的真实标签。少部分内容分配存在一定分歧,尤其是人口、劳动力、民族这部分内容,有时被分在资源环境图组,有时被分在经济社会图组。自然资源环境、经济社会两类图组名称都各自相似、类间易于区分,且具有这两类明显标签图组的国内外地图集数量达50部,超过地图集总数的一半,占包含图组—子图组结构的74部地图集中的67.57%。研究发现,PLM在k-means聚类k=2时可以天然地将子图组、图幅区分为自然资源环境与经济社会两类,PLM经过了千万级中文文本数据的训练,其内部向量空间已足以反映领域知识的分布与领域知识之间的邻近程度。参考文本聚类相关文献[42-43],采用聚类准确率(accuracy, ACC)、调整兰德指数(adjusted rand index, ARI)与归一化互信息(normalized mutual information, NMI)3个指标与原地图集中的子图组分类比较,评价聚类结果如表4所示。模型m3e-base、xiaobu、zpoint、Yinka在3个指标中都取得了较高的得分,4个模型中在该任务下总体得分最高的是m3e-base,其聚类ACC达到了0.924 2,达到较高水平。这说明中国地图集专题图组的自然资源环境、社会经济二元分类是相当科学的,符合领域知识的自然分类与常识。但小部分子图组的分类仍存在分歧,如第一产业(农林牧渔)、旅游、水利工程等方面的子图组,采用不同模型得到的分类结果不同。将资源环境与经济社会二元划分的现有地图集专题内容作为训练数据集,对模型进行微调,同时利用百科知识图谱补充语义信息,可以引导模型提升聚类精度。

表4　不同PLM k=2时聚类结果评价

Table 4　Evaluation of k=2 Clustering Results for Different PLM

模型名称	ACC	ARI	NMI
bert-base-chinese	0.500 0	−0.000 8	0.000 1
m3e-base	0.924 2	0.719 4	0.609 2
xiaobu	0.867 9	0.541 0	0.453 8
zpoint	0.868 6	0.543 2	0.456 5
Yinka	0.869 4	0.545 4	0.456 3

本文将所有同时包含自然资源环境、经济社会图组并且以上述内容为主要专题的地图集图组与子图组整理为资源环境经济社会子图组数据集。数据集共1 332条数据,由子图组和标签组成,来自50部地图集,其中图组名作为标签,所有自然资源环境、经济社会两类图组名称统一记为资源环境或经济社会两种标签。数据集内1 332个子图组中749个属于经济社会图组,583个属于资源环境图组。标签统一设置为0和1,分别代表资源环境与经济社会。将数据集分为训练集、验证集、测试集,比例为816∶180∶336。训练集与验证集用于模型训练,测试集则用于评价训练效果。

通过TransH训练得到提取子图后的CN-DBpedia的知识图谱嵌入,训练次数为40次,嵌入维度分别设置为64、128、256和512[44]。将知识图谱嵌入分别与xiaobu、zpoint、Yinka模型融合,这3个模型是Hugging Face平台预训练模型榜单MTEB Leaderboard汉语区在2024年7月综合排名前10的模型。使用知识增强模型、没有知识增强的Trainer微调模型与原模型对测试数据集进行k=2的k-means聚类,计算聚类评价指标ACC、ARI、NMI,结果如图4与表5所示。

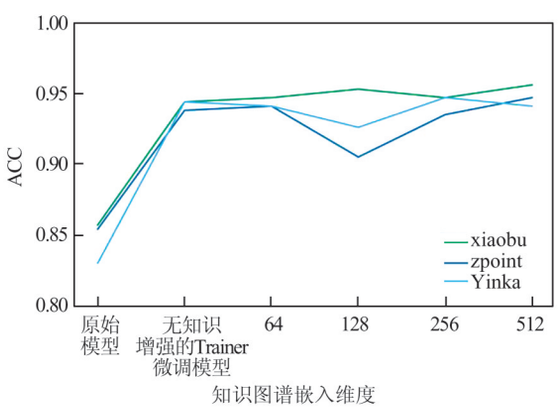


图4　测试数据集在聚类数k=2条件下的聚类准确率

Fig. 4　Accuracy of Clustering for the Test Dataset with the Number of Clusters k=2

表5　不同训练条件下xiaobu模型k=2聚类结果评价

Table 5　Evaluation of k=2 Clustering Results of xiaobu Under Different Training Conditions

模型	ACC	ARI	NMI
xiaobu	0.857 1	0.508 7	0.425 6
xiaobu+Trainer	0.943 5	0.785 9	0.686 6
xiaobu+CNDB(64)+Trainer	0.946 4	0.796 6	0.705 0
xiaobu+CNDB(128)+Trainer	0.952 4	0.818 1	0.743 7
xiaobu+CNDB(256)+Trainer	0.946 4	0.796 6	0.705 0
xiaobu+CNDB(512)+Trainer	0.955 3	0.828 9	0.754 5

图4与表5显示,融合了嵌入维度为512的百科知识图谱的xiaobu微调模型聚类精度最高,ACC达到了0.955 3,与原始模型相比提升了11.46%。取该模型对筛选后的42个子图组类名

称进行$k=2$的k-means聚类,计算聚类中心词并比较两个中心词与“资源环境”“经济社会”两个词的语义相似度,取相似度高者作为输出的聚类标签。最终利用该模型将42个子图组类分配至资源环境与经济社会两个专题图组,其中人口资源环境图组包含14个子图组类,经济社会图组包含28个子图组类。综合序图组、专题图组、发展规划图组内容,参考现有省域地图集进行适当排序,形成省域地图集通用内容结构体系如图5所示。图5显示了图组—子图组层次结构的省域地图集通用内容结构体系,其中专题图组分为人口资源环境与经济社会两个图组。

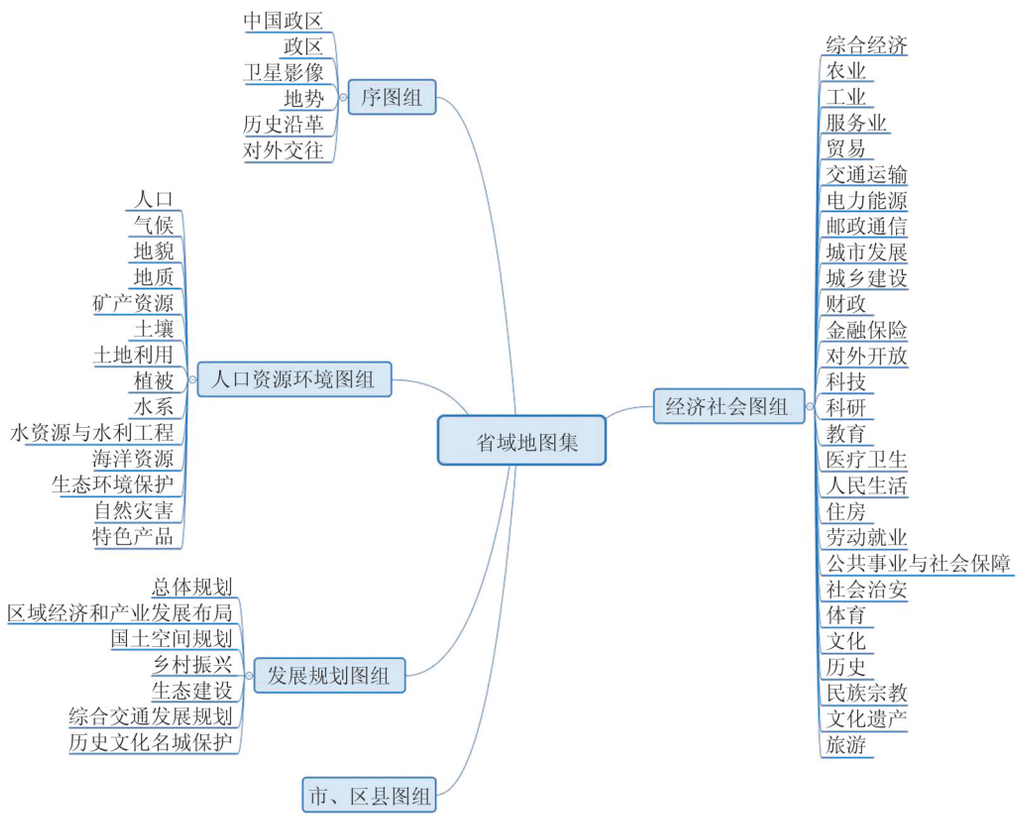


图5 省域地图集通用内容结构图谱

Fig. 5 Common Content Structure Graph of Provincial Atlases

3.4 图组-子图组-图幅三级内容体系

在建立图组—子图组层次结构体系后,接下来需要为各个子图组设计图幅,才能形成包含图组—子图组—图幅完整层次结构的地图集内容结构体系。

图幅以地图形式呈现单个或多个制图指标,反映了子图组所描述领域可制图的指标。本文收集了来自40部地图集的共5 689个图幅名称,将图幅对应的子图组与通用子图组类匹配。xiaobu模型计算地图集图组与子图组文本相似度数值的标准差较大,易于区分不同文本内容。因此,选用xiaobu模型计算所有图幅的语义相似度矩阵。根据图幅文本向量特征,选用0.85作为阈值,语义相似度大于等于该阈值即认为图幅名称相同,将相似度互相大于0.85的图幅名称合并为一个图幅类。一般来说,每个子图组类筛选出不少于3个图幅。若子图组中图幅数量较少,则优先选取出现频数相对较高或来源地图集权威性较高的图幅作为代表,以保证所构建内容结构体系的代表性与科学性。通过对样本地图集图幅内容进行统计与筛选,最终构建形成由4个图组、55个子图组、289个图幅组成的图组—子图组—图幅三级层次结构的省域综合性地图集通用内容结构体系。由于总体图幅数量较多,若全部展示不便于直观表达,因此选取具有代表性的子图组进行示例说明。在各子图组中,气候子图组通常包含的图幅数量较多,其图幅总数达到490幅,在样本地图集中平均图幅数为13.61幅,具有较强的典型性和代表性。因此,以气候子图组为例,对子图组内部推荐设计的图幅类型进行展示。在所构建的通用内容结构体系中,气候子图组隶属于人口资源环境图组。根据图幅筛选原则,在气候子图组中选取图幅数量不少于3个,共得到29个具有代表性的通用图幅。省域地图集

内容结构体系中气候子图组的通用图幅如图6所示,这29个图幅从内容主题上大致可归纳为气候区划、气温、太阳辐射、风和降水5个方面。

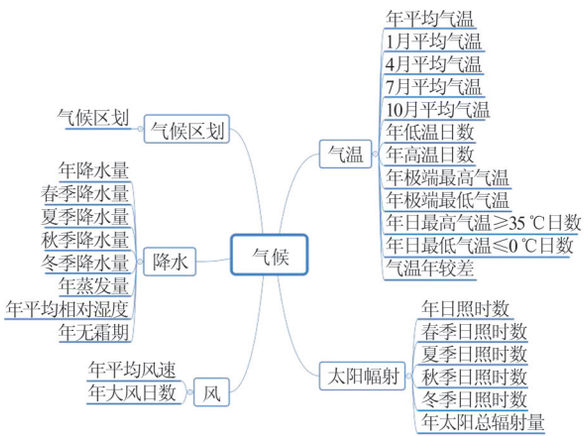


图6　气候子图组通用图幅

Fig. 6　General Maps for the Subgroup of Climate

4　讨　论

本文在专题图组二元聚类实验中,利用知识图谱嵌入微调PLM,通过文本聚类成功将子图组科学精确地分配到人口资源环境与经济社会两个图组中。劳动力、民族等内容在各省域地图集中图组归属的分歧较大,增强后的模型将这些内容分配到了经济社会图组中,表明这些内容与经济社会的关联度更高,关系更紧密。在各省域地图集中图组归属分歧很大的“人口”则被分配至资源环境图组,参考现有省域地图集专题图组命名,由资源环境内容组成的图组命名为人口资源环境图组。在一些地图集中分配到经济社会图组的水利工程、自然灾害、特色产品等内容被分配到了人口资源环境图组,在部分地图集被分配到人口资源环境图组的旅游子图组则被模型归入经济社会图组。

目前的专题图组内容分配有些不平衡,人口资源环境图组只有14个子图组,而经济社会图组则有28个子图组。专题图组的子图组类未根据类大小筛选,这使得一些不常见的内容出现在了省域地图集通用内容结构体系中。在实际的地图集编制中,不仅要全面考虑地图集制图指标,还要根据数据提供情况、区域特色等条件限制对地图集内容进行筛选。文献[11]在进行中国省情综合地图集内容图谱构建的研究时引入了支持度的概念,表示指标主题的频繁度,文中采用50% 支持度从地图集资料库中挖掘到相对稳定的最大频繁项集数。参考这项研究,本文设计了一种指标,称为单部地图集子图组内容量,表示某类子图组在一部地图集中具有的子图组数量。该指标的计算方法为:

$$v=\frac{q}{n} \tag{5}$$

式中,v为单部地图集子图组类内容量;q为子图组类包含的子图组总数;n为地图集总数。已知在专题图组数据库中,包含子图组—图幅结构的地图集总数为80。例如生态环境保护子图组类包含129个子图组,则生态环境保护子图组类的单部地图集子图组类内容量为1.612 5,即平均每部地图集具有1.612 5个子图组属于生态环境保护类。v值越高,代表着某类子图组在地图集中出现的可能性越高。50% 支持度可转换为v=0.5,那么也可以设置阈值v=0.5作为区分子图组类在地图集内容结构中的频繁度标准,v=0.5对应类大小为40,这代表着类大小不低于40的子图组类在多数地图集中出现。因此,将v值不低于0.5的子图组类提取出来,作为省域地图集内容结构设计的基础子图组。其余子图组则作为备选子图组,根据省域特征选择性的纳入设计。序图组与发展规划图组的子图组也按照相同规则筛选,得到省域地图集基本内容结构图谱,如图7所示。图7展示了多数区域综合地图集都会纳入设计的内容,在省域地图集内容设计中它们将成为固定内容。在省域地图集基本内容结构体系中,人口资源环境与经济社会两个专题图组的核心内容是较为平衡的,分别有11和13个子图组。研究数据来自国内外区域综合地图集,综合了国内省、市地图集以及国外国家地图集的内容结构特点,因此在输出的省域地图集通用内容结构体系中,子图组的内容很可能比原先中国省域地图集相同子图组的内容要多,如生态环境保护子图组,单个子图组更可能占据多个展开页。

省域地图集通用内容图谱中显示的子图组与实际地图集子图组类一一对应。图7显示的专题图组中24类基础子图组是具有子图组—图幅结构的80部地图集超过半数拥有的内容。然而这80部地图集专题图组部分没有任何一部具有全部24类基础子图组,仅有《吉林省地图集》(2009年版)与《长江经济带地图集》(2019年版)包含23类专题基础子图组。平均每部地图集仅包含19.49类专题子图组、14.93类专题基础子图组。在后续的省域地图集编制中,若能参考省域地图集通用内容结构体系与基本内容图谱进行地图集内容设计,可提高地图集内容的完备性与

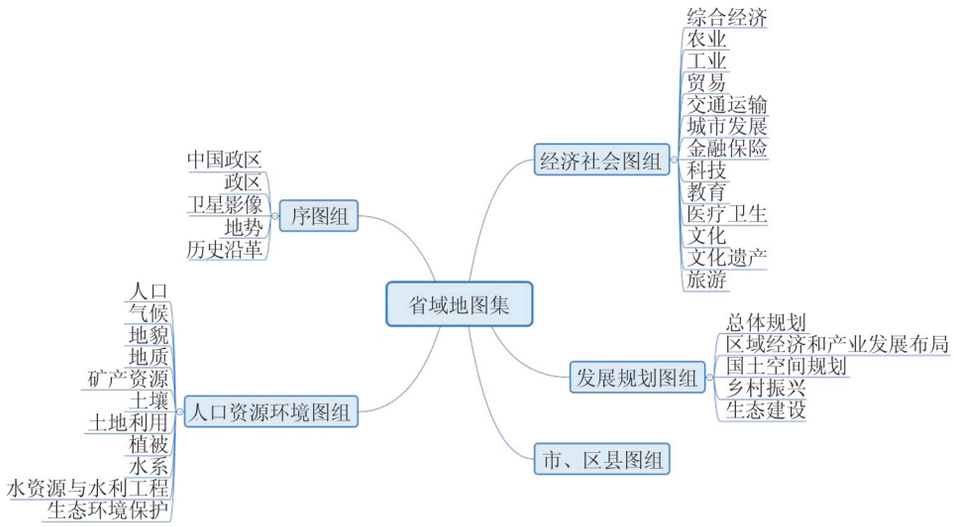


图7 省域地图集基本内容结构图谱

Fig. 7 Basic Content Structure Graph of Provincial Atlases

结构的科学性,为读者呈现更丰富的专题内容。

从专题子图组测试数据集聚类结果图4与表5中可以直观地看出,经Trainer训练微调后的模型聚类精度明显高于原始模型。以xiaobu模型为例,训练微调后ACC平均提升了10.69%,ARI和NMI的提升更加明显,分别达到了58.27%和68.95%。然而,聚类ACC提升的主要贡献来自Trainer训练,未加入知识图谱嵌入的模型经Trainer训练微调后ACC提升了10.07%。而知识图谱对模型文本聚类准确度的提升并不明显,融合知识图谱与不融合知识图谱微调的xiaobu模型相比,ACC平均上升了0.71%。嵌入维度为512的知识图谱对xiaobu模型ACC的提升为1.26%,ARI和NMI的提升分别为5.47%与9.89%。zpoint与Yinka的微调结果也表现出类似的规律,但聚类评价指标随知识图谱嵌入维度不同的变化特征与xiaobu不同。xiaobu模型融合知识图谱后,ACC均比不融合知识图谱的模型高,但数值上下波动,知识图谱嵌入维度为512的模型ACC最高,嵌入维度为256的模型ACC则是最低的。zpoint与Yinka则出现了在某些知识图谱嵌入维度情况下ACC比无知识图谱嵌入的结果低,尤其是嵌入维度为128时。嵌入维度由128提升至512的过程中,ACC则呈上升趋势。在实验结果中,知识图谱嵌入维度与聚类评价指标的关系并不明确,这可能是因为PLM本身经过了千万级文本的训练,而知识图谱的数据量仅为十万数量级,与预训练数据相比显得比较微小。WU等[44]与陈加兴等[45]在实验中发现,随着知识图谱嵌入维度的增加,模型性能可能会先提高后下降。一般来说,高维嵌入能够对语义信息进行更细粒度的编码,从而提高模型性能[45-46]。但是,过高的嵌入维度可能稀释知识图谱信号,导致模型出现过拟合,从而使得模型性能下降[47]。由于算力条件限制,本文的知识图谱嵌入维度最高设置为512。

本文也存在一些局限性,以下内容可作为后续的研究方向:(1)在知识图谱嵌入对PLM聚类ACC提升并不明显的情况下,暂未针对知识图谱嵌入对PLM聚类ACC提升效果展开深入研究,若后续算力条件允许,可设置更高的知识图谱嵌入维度数,根据聚类评价指标的变化比较知识图谱与PLM嵌入维度之间的关系对聚类ACC的影响;(2)百科知识图谱子图提取方法为简单的文本匹配,可能破坏了知识图谱结构,在未来的工作中可研究更优的子图提取方法,保证不破坏知识图谱结构的同时提取出最适用于地图集内容分类的图谱;(3)本文使用的PLM是截至2024-07的常用中文PLM(bert-base-chinese、m3e-base)以及在Hugging Face平台PLM榜单MTEB Leaderboard综合排名前10的中文开源PLM(xiaobu、zpoint、Yinka),可能缺少时效性,后续可利用最新、更先进的中文PLM开展研究。

5 结 语

本文引入知识图谱与NLP分析国内外区域综合地图集的内容结构内在特征,构建了由图组—子图组—图幅层次结构组成的省域地图集通用内容结构体系,可为区域综合地图集内容设计

提供具有科学性的可靠参考,也为提高地图集内容设计的自动化和智能化程度探索了实现路径。本文设计了一种基于知识图谱嵌入的预训练模型增强方法微调预训练模型,利用百科知识图谱将结构化领域知识融入模型向量库,在专题子图组分类任务下聚类 ACC 提升幅度最高达 11.46%,为基于预训练模型的文本聚类提供了一种能有效提高聚类精度的预训练模型微调方案。在未来的研究工作中,将在省域地图集内容结构设计智能化方面作进一步的探索,探讨省域特色的量化提取方法,根据区域特色对地图集内容进行取舍,从而为省域地图集内容结构如何体现区域特色提供可参考的设计方案。由通用内容结构体系与特色内容设计方案构成省域地图集内容设计知识库,利用检索增强生成技术实现省域地图集内容结构智能化设计。

参 考 文 献

[1] VOZENILEK V. Aspects of the Thematic Atlas Compilation[C]//CARTOCON 2014 Conference, Olomouc, Czech Republic, 2014.

[2] 赵曦. 地图集的设计与编制研究:以《陕西省地图集》为例[D]. 西安:长安大学, 2011.
ZHAO Xi. The Study of the Atlas Design and Compilation: By the Example of *the Atlas of Shaanxi Province*[D]. Xi'an: Chang'an University, 2011.

[3] 祝国瑞. 地图学[M]. 武汉:武汉大学出版社, 2004.
ZHU Guorui. Cartography[M]. Wuhan: Wuhan University Press, 2004.

[4] 何宗宜, 宋鹰, 李连营. 地图学[M]. 武汉:武汉大学出版社, 2016.
HE Zongyi, SONG Ying, LI Lianying. Cartography[M]. Wuhan: Wuhan University Press, 2016.

[5] 崔文宏. 区域综合地图集内容结构设计研究[J]. 地理空间信息, 2019, 17(9): 53-55.
CUI Wenhong. Research on the Content and Structure Design of Regional Comprehensive Atlas[J]. *Geospatial Information*, 2019, 17(9): 53-55.

[6] MEACHAM J E, ALLAN S. State Atlas Design[J]. *Cartographic Perspectives*, 2003(44): 42-50.

[7] 周振发, 赵大龙.《黑龙江省地图集》的设计与创新[J]. 测绘与空间地理信息, 2017, 40(12): 171-173.
ZHOU Zhenfa, ZHAO Dalong. Design and Innovation of *Heilongjiang Province Atlas*[J]. *Geomatics & Spatial Information Technology*, 2017, 40(12): 171-173.

[8] 王秀斌. 区域综合地图集总体设计思路探讨:《福建省情地图集》总体方案设计研究[J]. 测绘通报, 2008(1): 59-61.
WANG Xiubin. A Study of Comprehensive Design of *the Situation Atlas of Fujian Province*[J]. *Bulletin of Surveying and Mapping*, 2008(1): 59-61.

[9] 徐虹艳, 李伟锋, 吴俊华.《宁波市地图集》的设计与研制[J]. 测绘与空间地理信息, 2016, 39(12): 200-202.
XU Hongyan, LI Weifeng, WU Junhua. Design and Research of *Ningbo Atlas*[J]. *Geomatics & Spatial Information Technology*, 2016, 39(12): 200-202.

[10] 王家耀. 地图科学技术:由数字化到智能化[J]. 武汉大学学报(信息科学版), 2022, 47(12): 1963-1977.
WANG Jiayao. Cartography: From Digital to Intelligent[J]. *Geomatics and Information Science of Wuhan University*, 2022, 47(12): 1963-1977.

[11] 黄丽娜, 许国光. 中国省情综合地图集的内容图谱构建与主题表达可视分析[J]. 武汉大学学报(信息科学版), 2023, 48(5): 665-677.
HUANG Li'na, XU Guoguang. Construction of Content Tupu and Visual Analysis Subject Expression for Provincial Comprehensive Atlases in China[J]. *Geomatics and Information Science of Wuhan University*, 2023, 48(5): 665-677.

[12] 刘聪, 洪钧. 县市区级综合性地图集内容结构与信息表达方式初探[J]. 地理空间信息, 2010, 8(6): 126-128.
LIU Cong, HONG Jun. Study on the Content Structure and Information Expression of Comprehensive Atlas of Counties and Districts[J]. *Geospatial Information*, 2010, 8(6): 126-128.

[13] SHETH A, PADHEE S, GYRARD A. Knowledge Graphs and Knowledge Networks: The Story in Brief[J]. *IEEE Internet Computing*, 2019, 23(4): 67-75.

[14] MA X G. Knowledge Graph Construction and Application in Geosciences: A Review[J]. *Computers & Geosciences*, 2022, 161: 105082.

[15] DENG C, JIA Y T, XU H, et al. GAKG: A Multimodal Geoscience Academic Knowledge Graph[C]// The 30th ACM International Conference on Information and Knowledge Management, Queensland, Australia, 2021.

[16] 李连营, 彭楚骄, 郭邦祁, 等. 海洋地图可视化方法的知识图谱构建[J]. 测绘地理信息, 2022, 47(1): 77-80.
LI Lianying, PENG Chujiao, GUO Bangqi, et al.

Construction of Knowledge Map of Marine Map Visualization Method [J]. *Journal of Geomatics*, 2022, 47(1): 77-80.

[17] 彭楚骄, 李连营, 郭邦祁, 等. 使用知识图谱推荐专题地图可视化方法[J]. 测绘科学, 2022, 47(6): 188-194.
PENG Chujiao, LI Lianying, GUO Bangqi, et al. Recommendation of Thematic Map Visualization Method Using Knowledge Graph[J]. *Science of Surveying and Mapping*, 2022, 47(6): 188-194.

[18] 韩德志, 郭仁忠, 陈业滨, 等. 基于可视化维度理论的泛地图知识推荐方法[J]. 地球信息科学学报, 2024, 26(1): 110-120.
HAN Dezhi, GUO Renzhong, CHEN Yebin, et al. Knowledge Recommendation Method of Pan-Map Based on Visualization Dimension System[J]. *Journal of Geo-Information Science*, 2024, 26(1): 110-120.

[19] MAI G C, HUANG W M, CAI L, et al. Narrative Cartography with Knowledge Graphs[J]. *Journal of Geovisualization and Spatial Analysis*, 2022, 6(1): 4.

[20] 钟琳颖, 沈婕, 毛威, 等. 基于时空知识图谱的古城叙事地图设计与实现[J]. 遥感学报, 2021, 25(12): 2421-2430.
ZHONG Linying, SHEN Jie, MAO Wei, et al. The Design and Realization of the Ancient City Narrative Map Based on Spatiotemporal Knowledge Graph [J]. *National Remote Sensing Bulletin*, 2021, 25(12): 2421-2430.

[21] WANG H F, LI J W, WU H, et al. Pre-Trained Language Models and Their Applications[J]. *Engineering*, 2023, 25: 51-65.

[22] MIN B N, ROSS H, SULEM E, et al. Recent Advances in Natural Language Processing *via* Large Pre-Trained Language Models: A Survey [J]. *ACM Computing Surveys*, 2023, 56(2): 1-40.

[23] HU L M, LIU Z Y, ZHAO Z W, et al. A Survey of Knowledge Enhanced Pre-Trained Language Models[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2024, 36(4): 1413-1430.

[24] LIU W J, ZHOU P, ZHAO Z, et al. K-BERT: Enabling Language Representation with Knowledge Graph[PP/OL]. V1.(2019-09-17)[2024-06-15]. https://arxiv.org/abs/1909.07606.

[25] YAO L, MAO C, LUO Y. KG-BERT: BERT for Knowledge Graph Completion[PP/OL]. V2.(2019-09-07)[2024-06-21]. https://arxiv.org/abs/1909.03193.

[26] WANG X Z, GAO T Y, ZHU Z C, et al. KEPLER: A Unified Model for Knowledge Embedding and Pre-Trained Language Representation [PP/OL]. V3.(2019-11-13)[2024-05-07]. https://arxiv.org/abs/1911.06136.

[27] AGGARWAL C C, ZHAI C X. A Survey of Text Clustering Algorithms[M]. Boston: Springer US, 2012.

[28] SONG G, YE Y M, DU X L, et al. Short Text Classification: A Survey [J]. *Journal of Multimedia*, 2014, 9(5): 635-643.

[29] 荆琪, 段利国, 李爱萍, 等. 基于维基百科的短文本相关度计算[J]. 计算机工程, 2018, 44(2): 197-202.
JING Qi, DUAN Liguo, LI Aiping, et al. Short Text Correlation Calculation Based on Wikipedia [J]. *Computer Engineering*, 2018, 44(2): 197-202.

[30] XU B, XU Y, LIANG J Q, et al. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System [C]//The 30th International Conference on Industrial, Engineering, Other Applications of Applied Intelligent Systems, 2017.

[31] SU Y S, HAN X, ZHANG Z Y, et al. CokeBERT: Contextual Knowledge Selection and Embedding Towards Enhanced Pre-Trained Language Models[J]. *AI Open*, 2021, 2: 127-134.

[32] WANG Z, ZHANG J W, FENG J L, et al. Knowledge Graph Embedding by Translating on Hyperplanes [C]//AAAI Conference on Artificial Intelligence, Québec, Canada, 2014.

[33] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All you Need[C]//The 31st Conference on Neural Information Processing Systems, LongBeach, CA, USA, 2017.

[34] WOLF T, DEBUT L, SANH V, et al. Transformers: State-of-the-Art Natural Language Processing [C]//Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020.

[35] bert-base-chinese[EB/OL]. [2024-05-01] https://huggingface.co/google-bert/bert-base-chinese.

[36] m3e-base [EB/OL]. [2024-07-14] https://huggingface.co/moka-ai/m3e-base.

[37] xiaobu-embedding-v2 [EB/OL]. [2024-06-30]. https://huggingface.co/lier007/xiaobu-embedding-v2.

[38] zpoint_large_embedding_zh [EB/OL]. [2024-06-05]. https://huggingface.co/iampanda/zpoint_large_embedding_zh.

[39] Yinka[EB/OL]. [2024-05-30]. https://huggingface.co/Classical/Yinka.

[40] 陈丹, 褚宏伟, 吴雅琴, 等. 基于熵加权子空间算法的旅游数据聚类分析[J]. 旅游研究, 2021, 13(5): 18-31.
CHEN Dan, CHU Hongwei, WU Yaqin, et al. Clustering Analysis of Tourism Data Based on Entropy Weighting *k*-Means Subspace Algorithm [J]. *Tourism Research*, 2021, 13(5): 18-31.

[41] 李梦杰, 刘建国, 郭强, 等. 基于文本挖掘的互联网教育课程主题发现与聚类研究[J]. 上海理工大学学报, 2018, 40(3): 259-266.
LI Mengjie, LIU Jianguo, GUO Qiang, et al. Topic Discovery and Clustering Research for Online Courses Based on Text Mining[J]. *Journal of University of Shanghai for Science and Technology*, 2018, 40(3): 259-266.

[42] 谢士尧, 王小梅. 基于深度文本聚类的论文与专利数据融合方法研究[J]. 数据分析与知识发现, 2024, 8(4): 112-124.
XIE Shiyao, WANG Xiaomei. Paper and Patent Data Fusion Based on Deep Text Clustering [J]. *Data Analysis and Knowledge Discovery*, 2024, 8(4): 112-124.

[43] 吴建华, 魏宁, 张勇生, 等. 语言空间智能: 学科交叉创新的机遇与挑战[J/OL]. 武汉大学学报(信息科学版), 2025. https://ch.whu.edu.cn/article/doi/10.13203/j.whugis20240351.
WU Jianhua, WEI Ning, ZHANG Yongsheng, et al. Linguistic-Spatial Intelligence: Opportunities and Challenges of Interdisciplinary Innovation [J/OL]. *Geomatics and Information Science of Wuhan University*, 2025. https://ch.whu.edu.cn/article/doi/10.13203/j.whugis20240351.

[44] WU Y H, HUANG Y A, LI J Q, et al. Knowledge Graph Embedding for Profiling the Interaction Between Transcription Factors and Their Target Genes [J]. *PLoS Computational Biology*, 2023, 19(6): e1011207.

[45] 陈加兴, 胡志伟, 李茹, 等. 融合描述信息和结构特征的知识图谱链接预测[J]. 计算机科学与探索, 2024, 18(2): 486-495.
CHEN Jiaxing, HU Zhiwei, LI Ru, et al. Knowledge Graph Link Prediction Fusing Description and Structural Features[J]. *Journal of Frontiers of Computer Science and Technology*, 2024, 18(2): 486-495.

[46] ABDELAZIM H, THARWAT M, MOHAMED A. Semantic Embeddings for Arabic Retrieval Augmented Generation (ARAG)[J]. *International Journal of Advanced Computer Science and Applications*, 2023, 14(11): 01411135.

[47] 齐如煜, 尹章才, 顾江岩, 等. 高精地图的知识图谱表达[J]. 武汉大学学报(信息科学版), 2024, 49(4): 651-661.
QI Ruyü, YIN Zhangcai, GU Jiangyan, et al. Knowledge Graph Expression of High Definition Map[J]. *Geomatics and Information Science of Wuhan University*, 2024, 49(4): 651-661.