引文格式: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.DOI:10.13203/j.whugis20200547

Citation: 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. DOI: 10.13203/j.whugis20200547

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

黄丽娜[1,2] 许国光[1,3]

1 武汉大学资源与环境科学学院,湖北 武汉,430079

2 数字制图与国土信息应用工程自然资源部重点实验室,湖北 武汉,430079

3 重庆市勘测院,重庆,401120

摘 要:省情综合地图集使用专题地图对省域的资源禀赋和社会发展水平进行全面展示,是一项复杂的知识系统。通过对地图集中大量、复杂、非结构化的内容进行模式化重构,建立地图集的内容图谱,挖掘中国省情综合地图集的内容组织规律。首先,构建图集词汇向量和计算语义相似度,提取图集内容表达的标准主题词;然后,结合图集的"图组→图幅→指标"编排结构,对主题词进行语义层次聚类,构建图集内容表达的树型图谱;在此基础上,挖掘各省综合地图集内容图谱的频繁子图,形成图谱指纹,以此标识出中国省情综合地图集主题表达的共性特征。研究结果表明,中国省情综合地图集的主题内容具有层次化组织特点,可将指纹图谱作为框架,指导省情综合地图集的主题选择和内容组织。借助内容图谱和指纹图谱分析,还可进一步揭示出中国各省综合地图集在内容表达上具有明显的聚类特征和多样性特征。研究结果可为新编省情综合地图集设计提供依据。未来还可进一步探究不同类型地图集的内容图谱构成,丰富地图集设计与编制理论。

关键词:省情综合地图集;内容图谱;指纹特征;可视分析;地图集设计

中图分类号:P283 **文献标识码**:A **收稿日期**:2021-01-19

DOI:10.13203/j.whugis20200547 **文章编号**:1671-8860(2023)05-0665-13

Construction of Content Tupu and Visual Analysis Subject Expression for Provincial Comprehensive Atlases in China

HUANG Li'na[1,2] *XU Guoguang*[1,3]

1 School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China

2 Key Laboratory of Digital Mapping and Land Information Application Engineering, Ministry of Natural Resources, Wuhan 430079, China

3 Chongqing Survey Institute, Chongqing 401120, China

Abstract: **Objectives**: Provincial comprehensive atlas is a complex knowledge system. It uses thematic maps to comprehensively display the resource endowment and social development level of a province. We reconstruct the large, complex and unstructured contents of the atlas to establish the atlas content tupu, explore the basic content organization rules of provincial comprehensive atlas in China. **Methods**: First, we construct the vocabulary vectors of thematic subjects presented in atlas, calculate their semantic similarities, and extract the standard expressions of subject words. Second, following the compile order of "map group, map sheet, and cartographic index", the semantic hierarchical clusters of subject words are figured out, and the content map of atlas are constructed in a tree-wised graph. Finally, the frequent subgraphs of content tupu among provincial atlases are examined to form an atlas fingerprint to identify the common fea-

基金项目:国家重点研发计划(2017YFB0503502)。

第一作者:黄丽娜,博士,副教授,主要研究领域为地理信息可视化与认知、专题地图设计理论与方法。linahuang@whu.edu.cn

通讯作者:许国光,硕士,工程师。xgg19950302@163.com

tures of subject expressions in the comprehensive atlas of China's provinces. **Results**: It is shown that the thematic contents of provincial comprehensive atlas are organized hierarchically, and the fingerprint tupu can be used as a framework to guide the thematic selection and content organization during the comprehensive atlas compilation. With the benefit of content tupu and fingerprint tupu, it can be further revealed that the provincial comprehensive atlas in China has obvious clustering characteristics and diversity characteristics in content expression. **Conclusions**: The research results provide a basis for the design of provincial comprehensive atlas. In the future, we will explore the content tupus of different types of atlas to enrich the theory of atlas design and compilation.

Key words: provincial comprehensive atlas; content tupu; fingerprint feature; visual analysis; atlas design

省情综合地图集采用统一协调的地图语言全面反映省域的资源禀赋和社会经济水平。相较于长文本的地方志和普查报告,省情地图集基于空间图形思维呈现地学知识,为人们认知人居环境时空特征提供了直观、高效的手段[1-2]。目前,中国各省均编制和再版了省情综合地图集,这些省情综合地图集是展示当地自然人文环境的重要资料,对国民经济建设、教育科技、地理信息服务等也具有科学参考价值[3]。

与其他类型的图集相比,省情综合地图集的内容表达更注重完备性。为展现区域特色,省情综合地图集通常包含概览、历史变迁、资源生态、交通旅游、发展规划、区域地理等专题图组;内容往往涉及自然、人口、经济、文化、历史各领域数十个专业方向和部门,图幅数量巨大,上图指标更是数不胜数。然而,完备不等同于"大而全",应选择必需内容上图,立体表达出区域内各地理要素间的有机结合规律[4],因此,编制省情综合地图集也是一项复杂的知识系统工程[3]。如何系统完备地选择上图内容,准确有序地表达专题指标,是编制地图集需要解决的首要问题[5-6]。纵观现有省情综合地图集成果,其内容表达的顶层设计主要依赖编纂者对制图区域的主观理解[7-10],选题视角见仁见智,指标设计百花齐放,知识体系尚未形成统一认识。

许多学科领域通过构建图谱助力知识表达和分析。例如,生物学领域利用DNA指纹图谱标识同类物种中稳定的基因性状[11];化学领域使用酸根、盐基等来表示无机物的固有化学属性;地理学领域也尝试借助遥感技术探索地学现象与过程的谱系特征[12-15]。受图谱思维的启发,本文将图谱技术与地图集知识工程结合起来,以构建图谱,对地图集中大量、复杂和非结构化的制图内容进行模式化表达,挖掘中国省情综合地图集的内容组织规律,进而分析中国省情综合地图集内容表达的多样性。本文研究工作不仅是对地图制图理论的扩展,还可为省情综合地图集的设计与编制提供参考依据。

1 内容图谱构建思路

1.1 中国省情地图集内容表达特点

中国省情综合地图集的内容表达普遍遵循"图组→图幅→指标"的层次关系[4,7-9]。使用树结构表示省情综合地图集的内容体系,不难发现,各省地图集的内容体系的构成指标和组织形态虽有不同,但也存在多个相似分支。如图1所示,序图组中都有行政区划图,表示政区分布,人口资源环境图组中都有人口图,都表达人口数量和人口密度指标;序图组的历史沿革图(《浙江省地图集》)与安徽概览图组的历史名人图(《安徽省地图集》),图幅名相似;"全省多年进出口总额"(《浙江省地图集》)与对外贸易图中"进出口总额"(《安徽省地图集》)指标近义。这些相似结构在不同图集中频繁出现,印证了综合图集在内容选题和指标表达方面的规律性。提取各图集中频繁出现的主题及其关联结构,标识省情综合图集按某种规律表达特定主题的图谱特征,可为制定省情综合地图集内容体系提供科学参考。

1.2 内容图谱的构建框架

地图集内容图谱应包含两部分信息:图集所表达的主题内容以及主题内容之间的结构特征。一方面,主题内容来源于地图集中图组名、图幅名以及通过视觉变量呈现的制图指标,需提炼它们的语义信息,以获取相应的主题实体或概念。考虑到省情综合地图集制图指标种类繁多、专业性强、数据非结构化、选择灵活(相同主题可根据制图需要和数据可获得性选择近似指标),需要提取专业贴切、有代表性的主题词构建内容图谱。另一方面,省情综合图集的编排逻辑分明,由若干个制图指标联合表示某一专题图幅,多个专题图幅组成某一专题图组,进而构成地图集,

属于典型的part-of层次结构。但是,图集内容的编排组织又不等同于图谱结构。实际编图时可能用多个图幅来表示一个专题内容,如《江苏省地图集》的“水资源”专题由“水系”“水资源”“水利工程”3个图幅组成;或把多个专题内容合并为一个图幅/图组,如《安徽省地图集》将“科技”与“教育”两个专题合并为“科技教育”图幅,因而还需要结合指标表达的语义特征和上下文来解析主题词以及主题词之间的关联,最终形成地图集的内容图谱。

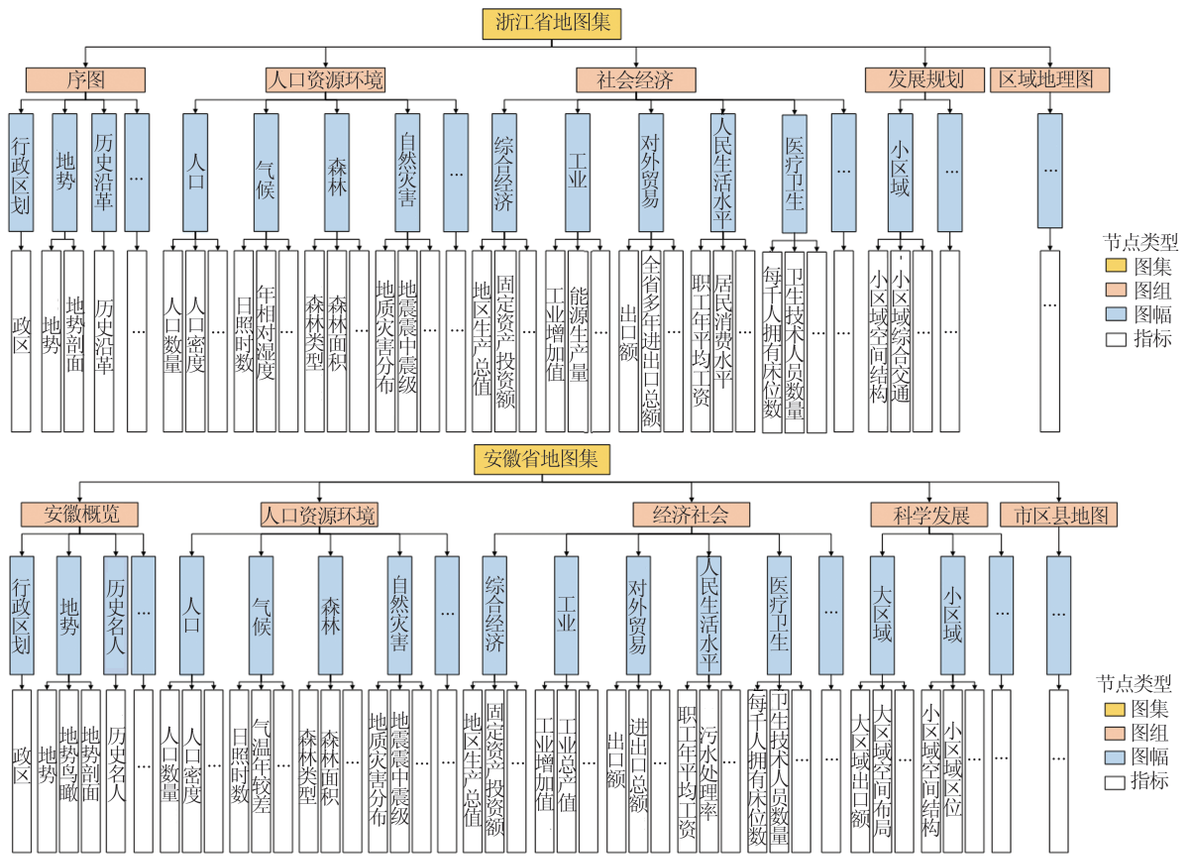


图 1　省情综合地图集内容体系示例

Fig.1　Content Systems Deriving from the Provincial Comprehensive Atlases

为降低过多人工参与而导致的主观性,本文引入自然语言处理技术进行省情综合地图集的主题词提取和关系构建,技术框架如图 2 所示。首先提取地图集内容并构建主题词的向量模型;然后基于词向量计算主题词的语义相似度,选择代表性词语作为主题词的标准化表达;最后通过语义聚类重构主题词对应图组、图幅与制图指标的层次逻辑关系,形成地图集内容图谱。提取各图集内容图谱中主题组织的频繁模式,则得到图集内容的指纹图谱。

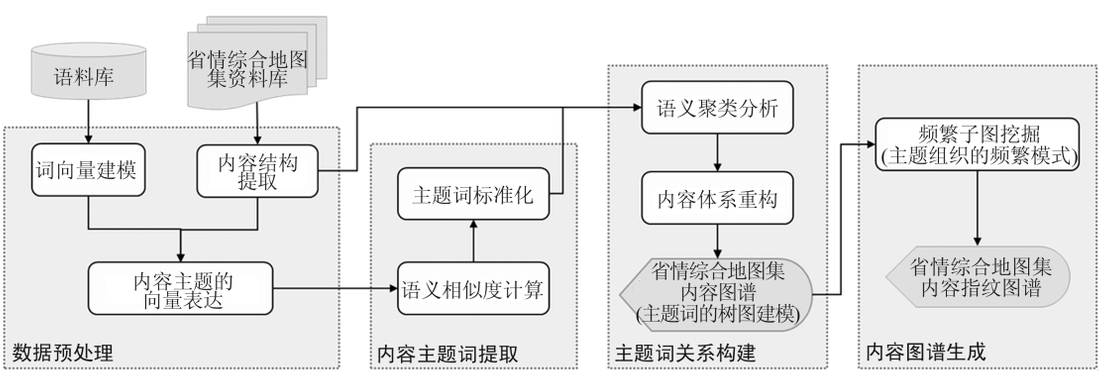


图 2　省情综合地图集内容图谱构建技术框架

Fig.2　Framework of Content Tupu Construction for the Provincial Comprehensive Atlas

2 地图集内容图谱构建方法

2.1 图谱主题词提取

理想条件是利用领域本体词汇抽取主题词,但地图学领域目前尚未建立地图集的本体词汇库。由于地图集中词语均是高度凝练的且与专题密切相关,在一定程度上可将图组名、图幅名、制图指标看作图集内容表达的主题词。为保证图谱的专业性和普适性,将所有图集的图组名、图幅名和制图指标均作为潜在主题词,图谱主题词则是其中语义最具代表性的词语或词组。

借鉴文献[12]的研究,本文基于语义相似度和上下文关联提取主题词。语义相似度采用词嵌入(WordEmbedding)模型计算[16]。基本思想是将词语映射到N维实数向量中,每一个维度代表词语的某个潜在特征,反映词语的句法和语义信息。现有研究提出了众多词向量训练框架,如词向量模型(Word2Vec)、神经网络语言模型(neural network language model,NNLM)、对数双线性模型(Log-bilinear model,LBL)、克劳伯-威斯通(Collobert & Weston,C&W)模型、循环神经网络语言模型(rerrent neural network language model, RNNLM)、全局向量(global vector, GloVe)模型、词序模型(Order)等[17-20]。根据文献[21],在词义相关性、同义词检测、文本分类等方面,各训练框架性能差异不大。其中,Word2Vec由于利用了上下文信息来训练词向量,可以较好地表达不同词语之间的相似和类比关系[19-20]。注意到地图集中许多制图指标是由多个词语组合而成的复合词,例如"高速公路里程""年最高气温日数"等,可进一步使用句向量,如词频-逆文档频率加权(term frequency-inverse document frequency,TF-IDF)模型[22],来构建图集的主题词向量。

计算两个主题词向量的夹角余弦值。若两主题词的语义相似度小于指定阈值,则认为它们分别表示不同的主题,均作为标准主题词保留下来;反之,若相似度大于阈值,则取其中一个主题词作为标准主题词,另一个主题词用标准主题词替换。当多个主题词的相似度均大于阈值时,选取共有最大相似词作为标准主题词。由于图集内容按图组、图幅、制图指标进行层次组织,应分别进行标准主题词提取。

2.2 主题词组织重构

针对地图集中一幅图表示多个专题或多幅图表达一个专题的情况,还需要梳理图幅主题词之间的关联。本文采用人工与自动结合的方式进行主题词组织重构。首先分析各图集的图幅安排,结合多数图集的图幅编排策略对交叉主题词进行分解,再根据各主题词的语义相似度,将复合主题词拆解为若干子主题、把语义紧密的子主题归并为一个主题。以"科技""教育""文化"为例,多数图集采用3幅图表示,而安徽省图集合并表示为"科技教育",黑龙江省图集合并表示为"科技文化"。结合语义相似度分析,如图3所示,"科技""教育""文化"在语义空间中相互远离,故将"科技教育"和"科技文化"分解为3个简单主题词。

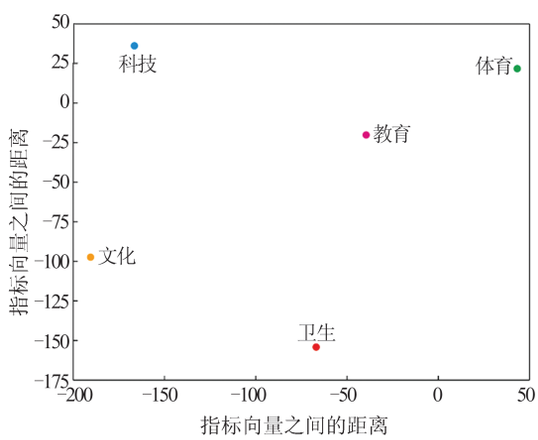


图3 几个主题词的语义邻近分析

Fig.3 Semantic Proximity Analysis of Several Subject Headings

在此基础上,按"制图指标→图幅→图组"的顺序对各主题词进行逐层归并。常用的分类算法有朴素贝叶斯分类算法、K近邻算法、支持向量机(support vector machine, SVM)、决策树等[23-26]。研究表明,SVM算法对于文本分类任务的效果较好[26],故本文选择SVM算法进行主题词的重分类。随机选择两部地图集训练主题词分类器,利用分类器对其余地图集的主题词进行聚类,得到各图集的主题词层次树结构,即地图集的内容图谱。

2.3 内容图谱的频繁模式挖掘

地图集内容图谱的频繁模式指多部图集中频繁出现的内容主题及其关联结构。在图集的"图组→图幅→指标"主题词层次树结构中,频繁模式具体指大部分地图集的主题词树图中均包含的子图。地图集内容主题的最大频繁子图即地图集内容的指纹图谱。

地图集内容主题的频繁子图定义如下:对于

图 $G=\{V,E\}$ 和 $G'=\{V',E'\}$，V 和 V' 是主题词，E 和 E' 是主题词之间的关联，如果 $V\subseteq V'$ 且 $E\subseteq E'$，则称 G 是 G' 的子图，G' 是 G 的超图。在图数据库 GDB 中，图 G' 的数量称为图 G 的支持度，若支持度大于某一阈值，则称图 G 为 GDB 的频繁子图，如果所有图 G' 均不是频繁的，则称图 G 为 GDB 的一个最大频繁子图。对于深度至指标主题的最大频繁子图，指标主题是频繁的，且图幅主题也是频繁的；若深度至图幅主题，则图幅主题是频繁的，但指标主题不一定频繁。

注意到地图集数量较少而每部图集的主题词繁多，可将挖掘地图集内容主题的最大频繁模式转换为提取事务数据的最大频繁项集。将地图集内容图谱按照“图组”“图组-图幅”“图组-图幅-指标”的形式转化为事务集合，如图 4 所示。若某事务集是频繁的，则应是满足最小支持度的共同子集，这些事务集的交集就是最大频繁项集。把最大频繁项集的事务数据还原为树图结构，便可得到地图集主题内容的最大频繁子图。

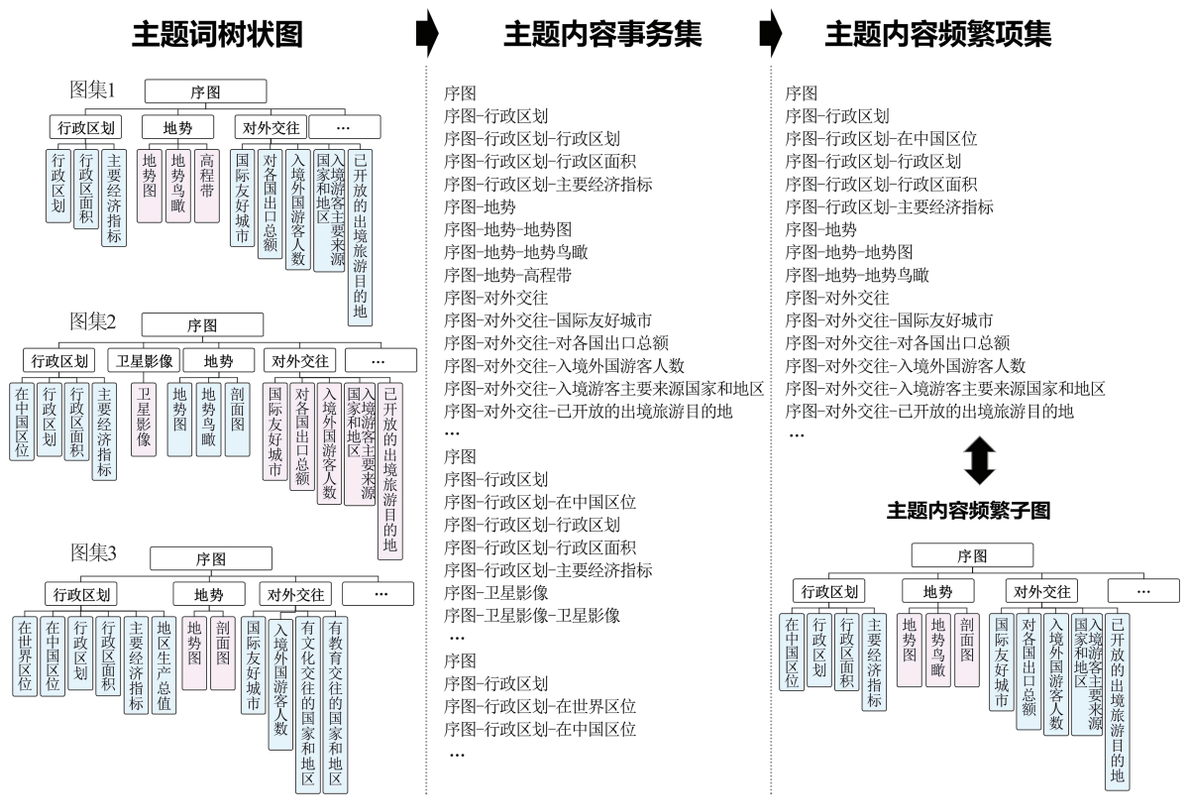


图 4　内容图谱最大频繁子图的挖掘过程示例

Fig.4　Workflow of Mining the Maximum Frequent Subgraph from a Content Tupu

3　中国省情综合地图集的内容图谱建立

3.1　数据获取与预处理

本文数据集为中国 2004—2015 年间公开出版的省情综合地图集，包括《江苏省地图集》《江西省地图集》《浙江省地图集》《山东省地图集》《吉林省地图集》《陕西省地图集》《安徽省地图集》《河北省地图集》《湖北省地图集》和《黑龙江省地图集》，涉及 2 000 余幅专题地图。这些地图集覆盖中国多个省区，较全面地代表了中国各地省情综合地图集的编制情况；平均发行量在 4 000 本左右，且大部分获得了优秀地图裴秀奖金奖的荣誉[27]，具有较强的社会影响力。

从地图集中采集内容主题词并进行如下处理：

1）指标清洗：采集制图指标时，过滤其中的空间分辨率和时间分辨率信息。例如，不同图集中“各市公路客运量”和“全省多年公路客运量”均表示“公路客运量”，仅制图单元或年份不同，则将其统一归为“公路客运量”指标。又如，“每千人拥有医疗机构床位数量”与“每万人拥有医疗机构床位数量”可统一为“每 N 人拥有医疗机构床位数量”。类似地，如“××率”与“××比重”属于同义词，统一记录为“××比重”。

2）简/别称对照：对于地图集中的习惯性简称、别称，如“旅客运输量”与“客运量”和“货物运

输量”与“货运量”等,在《财经大辞典》中认为后者是前者的简称,故根据相关工具书建立简称别称对照表。本文使用的工具书有《财经大辞典》《现代汉语大辞典》《新华汉语词典》《中国冶金百科全书》《安全环保》《化合物》《中国土木建筑百科辞典》等。

3)语料库构建:以中文维基百科语料库和全国及各省的社会经济公报为主体,补充分词表和停用词表[28]。为统一表达形式,使用Python的OpenCC(Open Chinese Convert)工具将词汇库中的繁体字转化为简体字[29],然后利用LTP分词工具将成段的文本分解为词语集合[30]。去除其中的停用词和非中文字符,建立语料库,为后续构建主题词向量提供基础。

3.2 内容主题获取

Python的Gensim库提供了Word2Vec模块,可用于建立主题词向量以及进行语义相似度计算[31],但要设置合适的训练参数。本文利用Chen等[32]建立的wordsim240和wordsim296数据集作为辅助,首先通过Word2Vec训练生成数据集的词向量;然后计算由向量表示的成对词语相似性与人工标注的成对词语相似性的相关系数,将其作为语义紧密度指数,计算基于向量模型推算的类比词(如根据“小学”之于“基础教育”推算出“大学”对应的是“高等教育”)与人工标注的类比词之间的语义相似度,将其作为词类比指数。Word2Vec框架有连续词袋(continuous bag of words,CBoW)和Skip-gram两种训练模式[19-20]。对比两种模式训练结果的语义紧密度和词类比指数,选择指数较高的模型和概率函数,最终确定采用Skip-gram模型和Hierarchicalsoftmax概率函数训练生成主题词向量。经多次实验,设词向量维度为120,上下文范围(窗口尺寸)为5,降采样值为1×10^{-5},迭代次数为5。

使用TF-IDF方法计算主题词向量[19,22],计算公式如下:

$$tf_{i,d}\,f_{i,d}=tf_{i,d}\cdot\lg\frac{|D|}{\left|\{d\in D|i\in d\}\right|} \tag{1}$$

$$I_d=\sum_{i=1}^{n}w_i\cdot tf_{i,d}\,f_{i,d} \tag{2}$$

式中,$tf_{i,d}\,f_{i,d}$表示主题词d中基本词语i的权重,对主题词中每个词语的词向量加权求和即可得到该主题词的向量表达I_d;$tf_{i,d}$为词语i在主题词d中出现的频率;D为语料库;w_i表示词语i的词向量。

计算主题词向量的夹角余弦值时,本文参考了文献[33],选择0.85作为阈值来判断主题词的语义相似度。根据前述算法,基于主题词的语义相似判断最终确定的标准主题词如表1所示。各图集的图组名、图幅名和制图指标将统一表示成标准主题词构建内容图谱。若同一图幅中出现多个相同的标准主题词,则删除重复的主题词。

表1 标准主题词提取结果示例

Tab.1 Examples of Extracting Standard Subject Headings

主题词1	主题词2	语义相似度	标准主题词
总人口数量	人口数量	0.922 570 29	人口数量
少数民族聚集地分布	少数民族分布	0.926 992 44	少数民族聚集地分布
高速公路里程	高速公路长度	0.866 226 84	高速公路里程
年最高气温日数	年高温日数	0.950 536 51	年高温日数
平均高温日数	年高温日数	0.972 380 63	年高温日数
年最高气温日数	平均高温日数	0.945 518 99	年高温日数
活立木蓄积量	森林蓄积量	0.954 365 14	森林蓄积量
活立木蓄积量	林木蓄积量	0.955 687 64	森林蓄积量
有林地蓄积量	森林蓄积量	0.981 543 70	森林蓄积量
林木蓄积量	森林蓄积量	0.978 464 05	森林蓄积量
活立木蓄积量	有林地蓄积量	0.954 480 33	森林蓄积量
有林地蓄积量	林木蓄积量	0.980 648 91	森林蓄积量

3.3 内容图谱生成

应用Python机器学习库Scikit-Learn提供的SVM模块进行地图集标准主题词的重分类[34]。本文采用《湖北省地图集》作为学习集、《江苏省地图集》作为测试集训练主题词分类器。当分类准确率在90%以上时,得到最终分类器。以“科技”“教育”“文化”“卫生”“体育”专题图为例,基于SVM语义分类的指标主题分组结果见图5。

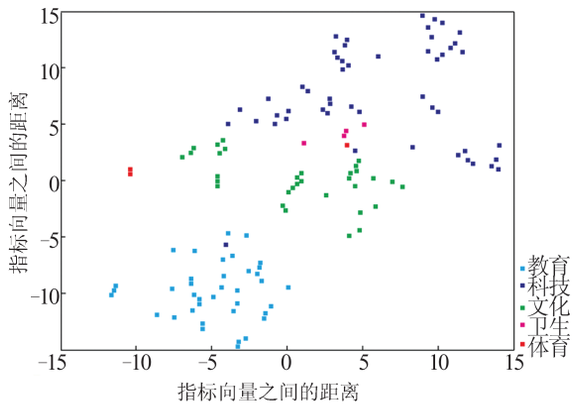


图 5　“科学”“教育”“文化”“卫生”“体育”专题图的主题词重分类结果

Fig.5　Re-classification of Subject Headings on Thematic Maps of Science, Education, Culture, Health and Physical Education

以主题词为节点,按主题词重分类结果建立各图集专题组织的层次树,得到地图集的内容图谱。图 6 展示了《浙江省地图集》内容图谱的专题信息以及组织结构。对比可知,序图组中政区、地势和历史沿革等图幅的主题词与其上图指标主题词几乎相同,上图指标主题极少。相对而言,经济社会图组表达的主题最为广泛,涉及 22 个图幅主题,其次是人口资源环境图组,有 18 个主题。图集中交通运输、对外贸易、民营经济和专业市场、医疗卫生、文化等主题的内容尤为丰富,上图指标多达数十个。

3.4　指纹图谱提取

从各省地图集的内容图谱中挖掘主题表达的最大频繁模式,提取图集指纹图谱。分析支持度阈值变化与最大频繁项集数量变化的关联(见图 7)发现,随着支持度阈值增加,最大频繁子图数渐趋稳定,阈值拐点出现在 50% 附近,采用 50% 支持度可从地图集资料库中挖掘到相对稳定的最大频繁项集数。故本文采用 50% 作为最小支持度阈值。

中国省情综合地图集的指纹图谱如图 8 所示。指纹图谱标识了中国省情综合地图集的内容选题和组织特点:一般包含序图、资源环境、社会经济和县市区地理图 4 个图组主题。在 50% 的支持度下,序图组包含政区、地势图幅主题;人口资源环境图组包含人口、劳动力、地质、矿产、气候、水系、水资源、土壤、自然灾害等图幅主题;社会经济图组包含综合经济、工业、农业、国内贸易、对外贸易、金融等图幅主题。

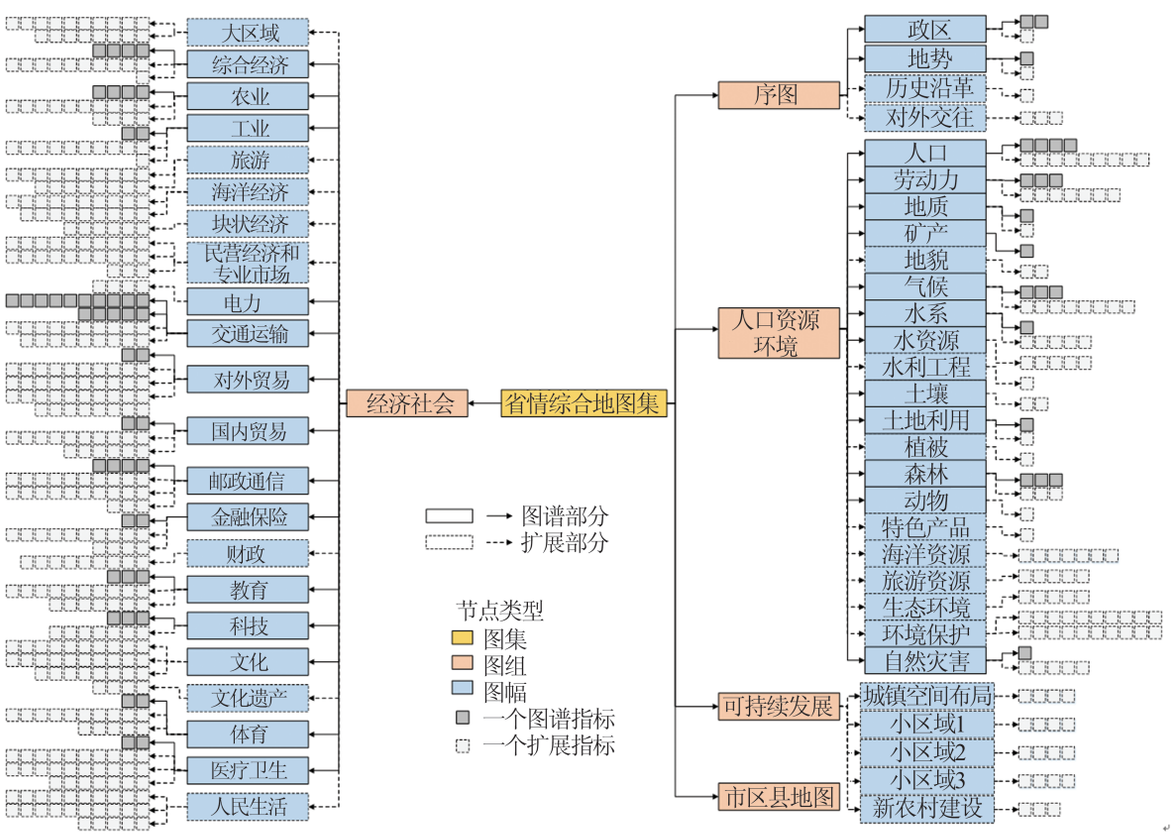
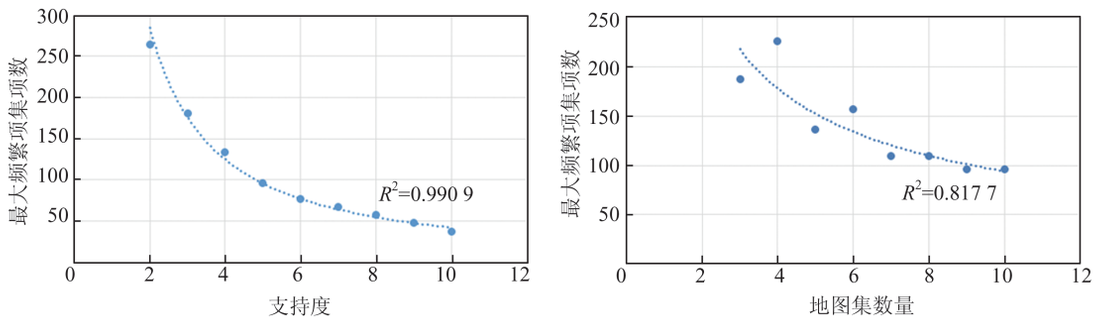


图 6　《浙江省地图集》的内容图谱

Fig.6　Content Tupu of *Zhejiang Province Atlas*



(a) 最大频繁子图数变化　　(b) 50%支持度下的最大频繁子图数变化

图7　支持度阈值与地图集内容频繁模式挖掘结果的回归分析

Fig.7　Regression Analysis Between Support Thresholds and Frequent Patterns of Atlas Content

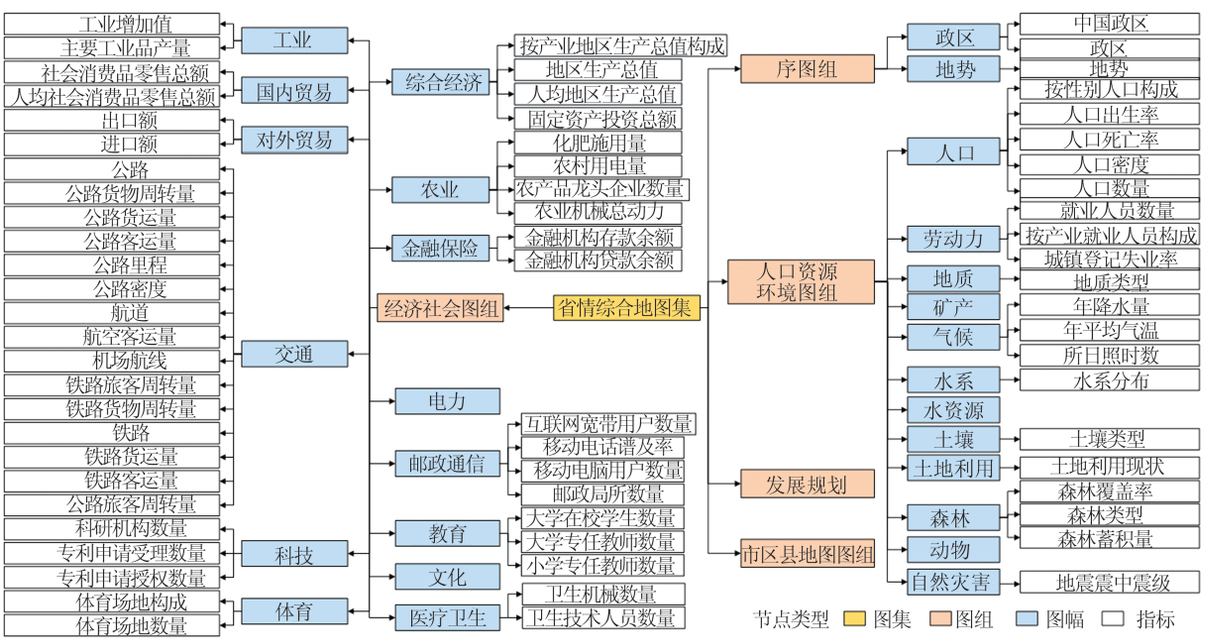


图8　50%支持度条件下的省情综合地图集内容的指纹图谱

Fig.8　Content Fingerprint of Provincial Comprehensive Atlas with the 50% Support Threshold

根据指纹图谱,序图组的图幅主题词与指标主题词几乎相同,这表明序图组内容较简单。相比之下,人口资源环境图组和经济社会图组的图幅主题更丰富,说明这两个图组内容复杂,需要通过多个专题图幅和大量制图指标来表达。其中,人口资源环境图组中矿产、水资源、动物等和社会经济图组中电力、文化等图幅的指标主题稀少,这是由于不同的省情综合图集选择了语义差别较大的制图指标来表示,体现了各省地图集内容表达的多样性;气候、人口、综合经济、交通等图幅的指标主题较多,表明中国各省地图集在表达这些图幅主题时对指标的选择具有一定共识。

指纹图谱为省情综合地图集的内容设计提供了结构化基础,可用于指导实际省情综合地图集内容选题和编制。要满足地图集内容的系统性和完备性需要,地图集中应至少包含指纹图谱中所列的图组主题、图幅主题和指标主题。结合实际制图时的数据资料情况、表达方法设计、图面配置效果等条件,各省区在制作地图集时因地制宜地选择相同或近似指标,既可通过指标含义及其数值差异体现区域特色,又有利于各省省情的横向对比。

4　基于内容图谱的地图集表达差异分析

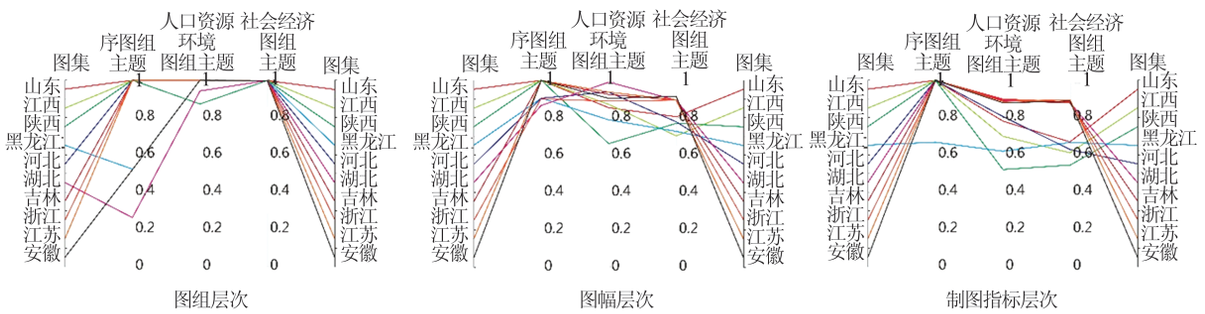
4.1　主题差异的聚类特征分析

将各省图集的内容图谱看作是指纹图谱的实例化,从内容图谱中标记图谱指纹。计算其实际主题词与标准主题词之间的语义相似度,求和并进行标准化处理:

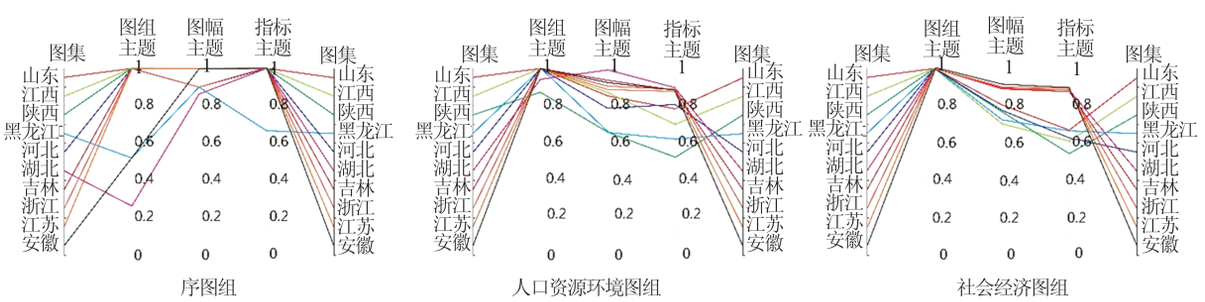
$$S=\frac{1}{n}\sum_{i=1}^{m}s_i \tag{3}$$

式中,s_i为实例i与其图谱指纹之间的语义相似度;n为指纹图谱中的主题词数量;m为实例中与指纹图谱对应的主题词数量。

利用平行坐标系对计算结果进行可视化分析(见图9),发现各省情综合地图集在主题表达上具有明显的聚类特征。图组层次上,除黑龙江省、安徽省、湖北省和陕西省的地图集在序图组有分离,各图集的图组主题与图谱的平均相似程度均接近1,说明各图集的图组主题几乎相同。在图幅层次上,虽然人口自然资源图组的语义相似度有一定波动,分布约在0.7~1之间,但序图组和社会经济图组的多数图幅都相对收敛;序图组的语义相似程度收敛于0.9和1两处,社会经济图组的语义相似程度多数收敛于0.9处。在指标层次上,湖北省、吉林省、浙江省、江苏省和安徽省地图集的制图指标在序图组、人口资源环境图组、社会经济图组与指纹图谱的平均语义相似度分别达到1、0.9和0.9,呈现出明显的聚集特点。



(a) 各组织层次语义相似度的平行坐标轴表达



(b) 各专题图组语义相似度平行坐标轴

图 9　基于平行坐标轴的各省综合地图集与指纹图谱主题词语义差异聚类分析

Fig.9　Cluster Analysis on Thematic Difference of Subject Headings in Compreshensive Provincial Atlas and Fingerprint Tupu Based on Parallel Axises

按照图组专题分析各图集主题与图谱主题的语义相似度。注意到各省图集的人口资源环境图组的语义相似度呈离散分布,特别是在图幅层次分散于0.6~1、制图指标层次上分散于0.5~0.9,表明各省图集选择了略有差异的主题来表达。但在社会经济图组方面,各图集大体可分成两组。湖北省、吉林省、浙江省、江苏省和安徽省的地图集为一组,图幅和指标主题与图谱的语义相似度一致,均为0.9;其他地图集可近似看作一组,图幅主题与图谱的语义相似度介于0.7~0.8,指标主题与图谱的语义相似度在0.6上下。这些均印证了中国省情综合图集的内容表达既有一致性,也有多样性。

4.2　内容选题的多样性分析

固定内容图谱中指纹组合的标记顺序,比对各省图集的主题规模和构成,如图10所示。各图集均表示了大量的其他主题,呈现多态特征。结合图幅的具体内容,如《浙江省地图集》增加了“海洋资源”“民营经济”“长三角经济圈”等图幅,体现了明显的区域特色。浙江省是沿海省份,地属长三角经济圈,经济发达,商业贸易繁荣;省内大量中小型企业通过产业聚集形成了许多产业集群,如海宁皮革、义乌商品等,是中国民营经济发展的典型。这些图幅强化了区域特色。又如,《黑龙江省地图集》增加了“草地”“自然与农业区划”“垦区经济”等图幅。黑龙江省作为农业大省,农业水平发达,物产丰饶,拥有丰富的森林植

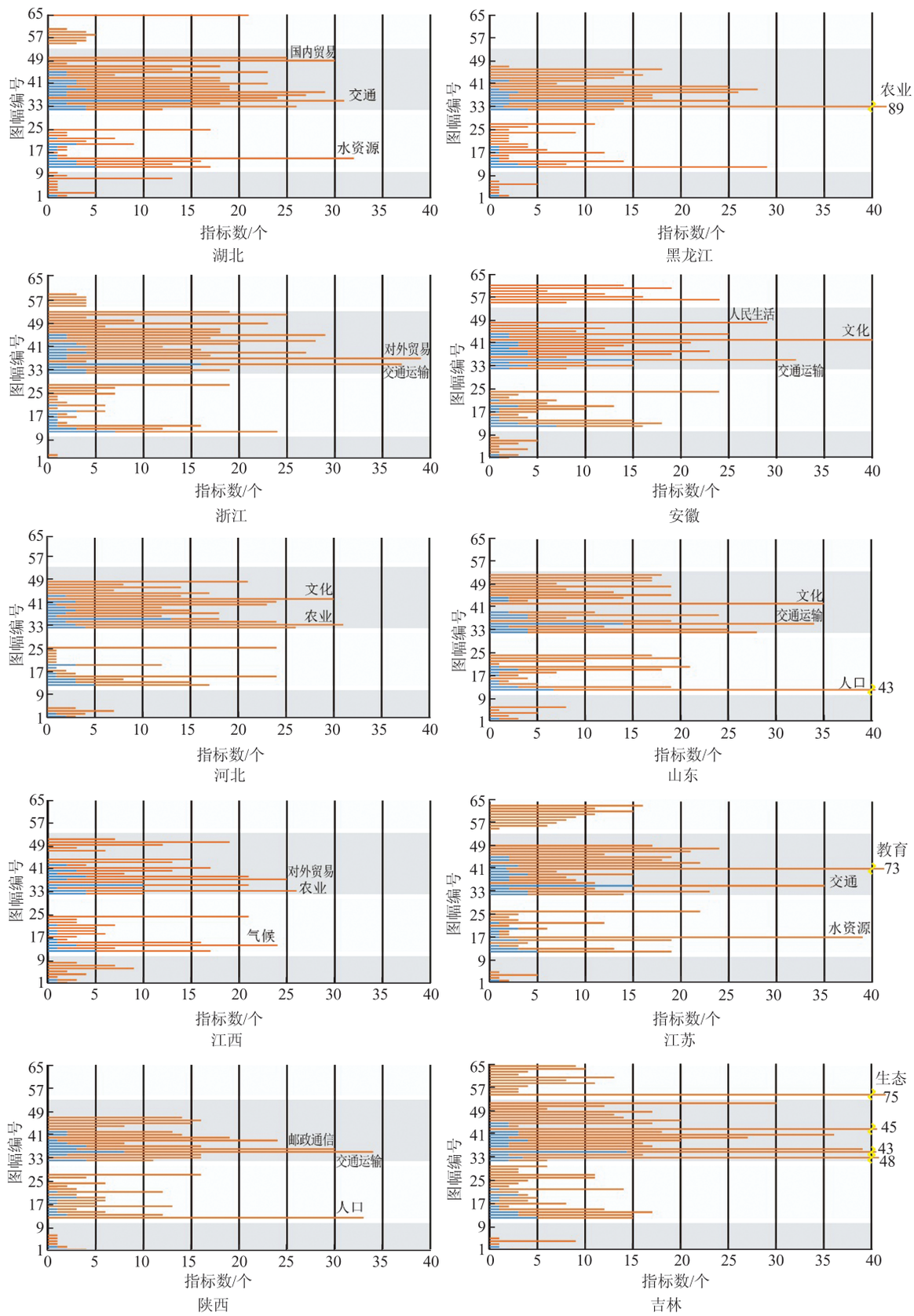


图10 各省综合地图集内容构成与特征图幅

Fig.10 Content Composition and Characteristic Maps of the Explored Provicial Atlases

被资源和广袤的草场资源,并且形成了特有的垦区生产形式。在指纹图谱的框架上补充这些图幅可以更立体地突出制图区域的个性信息。

进一步采用Jenks自然断裂法提取各图集指标数最多的图幅[35],可知,《湖北省地图集》指标最多的专题图是水资源、交通运输和国内贸易,这与湖北省的“千湖之省”“九省通衢”区位形象一致。湖北省内除长江、汉江干流外,省内各级河流绵亘蜿蜒,湖泊星罗密布,水资源丰富。并且湖北省地处中国中部,是国内公路、铁路和航空运输的重要枢纽,交通运输部也将湖北省定为交通强国建设试点地区之一[36]。又如《江苏省地图集》指标最多的专题图是教育、水资源、交通。作为中国的教育大省,江苏省建设了多所科研院所、高等院校和职业教育学校,教育水平一直位于全国前列。由于地处长江下游地区,河渠纵横、水网稠密,中国五大淡水湖就有两个位于江苏省。江苏省内路网发达,不仅农村公路网密度和高等级公路比重全国领先、各区市全部开通了动车,而且内河等级航道与四级以上航道里程均居中国第一。

利用内容图谱还有助于发现地图集内容表达中可能存在的问题。如《江西省地图集》“对外贸易”图幅的指标数位居前三,却缺少指纹图幅“国内贸易”。结合江西省的区域发展情况[37],其对外贸易并没有特别突出的优势,强调对外贸易而忽略国内贸易,可能缺乏合理解释。

5　结　语

本文借助自然语言处理技术提取地图集的主题词及其组织结构,实现了地图集的内容图谱构建,并通过频繁模式挖掘技术建立了中国省情综合地图集的内容指纹图谱。内容图谱和指纹图谱为定性定量分析图集内容表达提供了有效途径。一方面,利用指纹图谱可以揭示中国省情综合地图集内容的共性特征。中国近十年编制的省情综合地图集通常按“图组→图幅→指标”三层结构组织内容。类似于化学领域利用化学分子式指导生成化合物,可使用本文提取的指纹图谱指导新编省情综合图集编制的内容设计和指标选择。以指纹图谱为框架,再结合地方特色调整图集内容,有助于实现图集内容表达的系统性、完备性和区域适应性。也可提取已编图集的内容图谱与指纹图谱进行比对,若发现该图集内容图谱与指纹图谱不一致时,则需进一步分析是否为根据区域特色进行的适应性调整,抑或存在内容缺失或错漏。另一方面,分析各图集内容图谱与指纹图谱之间的主题语义差异,发现中国各省综合地图集的主题表达具有聚类特征和多样性特征。指纹图谱是从已编图集的内容图谱中挖掘得到的,会随着已编图集的变化而相应发展,因此,若未来中国省情综合地图集的指纹图谱持续发生变化,借鉴使用基因图谱研究物种谱系特征的思路,可利用指纹图谱的变化探究中国图集内容表达的时空演变特征。

本文研究仍存在一些不足之处。由于目前尚未建立面向省情综合地图集的本体词汇库,本文将图集的所有词语均作为潜在主题词,采用自下而上的方法提取图谱主题词,主题词存在冗杂信息。若能利用图集相关领域的本体词汇库,例如省情综合图集常用的统计年鉴构建和扩充本体词汇,可以提升图谱主题词的普适性和专业性。此外,本文通过SVM聚类算法进行主题词重分类,仅考虑了“图组→图幅→指标”的上下层次结构,存在一定局限性。图集主题词之间具有丰富的语义关联[38],例如,is-a、part-of,kind-of等,仍待进一步挖掘主题词之间的逻辑关系,实现各主题词的相互联结。根据文献[39],可基于地图集本体词汇库建立语义关联规则,通过匹配主题词与本体词汇,推理出主题词之间的关联逻辑,进而将内容图谱从树结构拓展至网络结构,更全面、细致地描述图集内容组织的谱特征。

构建内容图谱为研究地图集选题特征和组织模式提供了新思路。现有的综合地图集种类众多,按制图范围分为世界、国家、城市等尺度,按区域分布分为中国和外国。在未来的研究工作中,也可借助地图集内容图谱探讨不同类型地图集的编排特色,丰富地图集设计与编制理论。

参　考　文　献

[1] Chen Shupeng, Yue Tianxiang, Li Huiguo. Studies on Geo-Informatic Tupu and Its Application[J]. *Geographical Research*, 2000, 19(4): 337-343. (陈述彭, 岳天祥, 励惠国. 地学信息图谱研究及其应用[J]. 地理研究, 2000, 19(4): 337-343.)

[2] Ye Yanjun, Zhang An, Qi Qingwen. The Characteristic and Value of Provincial and Regional Maps in near Modern China[J]. *Journal of Geo-Information Science*, 2016, 18(1): 57-67. (叶妍君, 张岸, 齐清文. 中国近代省区地图特色与价值的探讨[J].

地球信息科学学报, 2016, 18(1): 57-67.)

[3] He Zongyi, Song Ying, Li Lianying. Cartography [M]. Wuhan: Wuhan University Press, 2016. (何宗宜, 宋鹰, 李连营. 地图学[M]. 武汉: 武汉大学出版社, 2016.)

[4] Zou Yujun. Review and Current Development of Provincial Atlas in China[J]. *Geomatics and Information Science of Wuhan University*, 1960, 4(1): 39-47. (鄒毓俊. 中国编制省地图集的回顾和近况[J]. 武汉大学学报(信息科学版), 1960, 4(1): 39-47.)

[5] Gómez S L S, Sancho C J, Bosque S J. Atlas Design: A Usability Approach for the Development and Evaluation of Cartographic Products[J]. *The Cartographic Journal*, 2017, 54(4): 343-357.

[6] Buckley A. Atlas Mapping in the 21st Century[J]. *Cartography and Geographic Information Science*, 2003, 30(2): 149-158.

[7] Wang Bing. Discussion on Design Features of New Atlas of Hubei Province[J]. *Geospatial Information*, 2015, 13(3): 177-178. (汪冰. 浅谈新编《湖北省地图集》的设计特点[J]. 地理空间信息, 2015, 13(3): 177-178.)

[8] Wang Yueping. Key Points for Updating and Compiling Anhui Province Atlas[J]. *Standardization of Surveying and Mapping*, 2019, 35(3): 55-57. (汪跃平.《安徽省地图集》更新编制实践[J]. 测绘标准化, 2019, 35(3): 55-57.)

[9] Zhou Zhenfa, Zhao Dalong. Design and Innovation of Heilongjiang Province Atlas[J]. *Geomatics & Spatial Information Technology*, 2017, 40(12): 171-173. (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.)

[10] Robinson A C, Demšar U, Moore A B, et al. Geospatial Big Data and Cartography: Research Challenges and Opportunities for Making Maps that Matter [J]. *International Journal of Cartography*, 2017, 3(sup1): 32-60.

[11] Shao Xuehua, Liu Niu, Lai Duo, et al. Genetic Diversity Analysis and DNA Fingerprint Mapping of 28 Varieties of Phyllanthus Emblica L. Based on ISSR Molecular Marker[J]. *Journal of Northwest A & F University* (*Natural Science Edition*), 2020, 48(8): 129-136. (邵雪花, 刘牛, 赖多, 等. 28份余甘子品种遗传多样性的ISSR分析及指纹图谱构建[J]. 西北农林科技大学学报(自然科学版), 2020, 48(8): 129-136.)

[12] Cai D, Chen Y M, Gao C. Primary Research on Geo-Informatic Tupu for Crime Spatio-Temporal Analysis[C]//Geo-Informatics in Resource Management and Sustainable Ecosystem, Wuhan, China, 2015.

[13] Hu Xiaodong, Luo Jiancheng, Xia Liegang, et al. Adaptive Water Body Information Extraction Using RS Tupu Computing Model[J]. *Acta Geodaetica et Cartographica Sinica*, 2011, 40(5): 544-550. (胡晓东, 骆剑承, 夏列钢, 等. 图谱迭代反馈的自适应水体信息提取方法[J]. 测绘学报, 2011, 40(5): 544-550.)

[14] Yang Cunjian. The Idea of Geo-Information Tupu and Its Practices[J]. *Journal of Geo-Information Science*, 2020, 22(4): 697-704. (杨存建. 地学信息图谱思想与实践探索[J]. 地球信息科学学报, 2020, 22(4): 697-704.)

[15] Liu Hai, Wang Xingling, Chen Xiaoling, et al. Study on Structure of Ecosystem Services Value Assisted with Geo-Information Tupu in Jiangxi Province[J]. *Geomatics and Information Science of Wuhan University*, 2012, 37(1): 118-121. (刘海, 王兴玲, 陈晓玲, 等. 利用地学信息图谱的江西省生态服务价值结构研究[J]. 武汉大学学报(信息科学版), 2012, 37(1): 118-121.)

[16] Yang Yang, Liu Longfei, Wei Xianhui, et al. New Methods for Extracting Emotional Words Based on Distributed Representations of Words[J]. *Journal of Shandong University* (*Natural Science*), 2014, 49(11): 51-58. (杨阳, 刘龙飞, 魏现辉, 等. 基于词向量的情感新词发现方法[J]. 山东大学学报(理学版), 2014, 49(11): 51-58.)

[17] Cho K, van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation [C]// Empirical Methods in Natural Language Processing, Doha, Qatar, 2014.

[18] Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation[C]//Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014.

[19] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality [J]. *Advances in Neural Information Processing Systems*, 2013, 26: 3111-3119.

[20] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [EB/OL]. [2020-06-01]. https://arxiv.org/abs/1301.3781.

[21] Lai Siwei, Xu Liheng, Chen Yubo, et al. Chinese Word Segment Based on Character Representation Learning[J]. *Journal of Chinese Information Pro-*

cessing, 2013, 27(5): 8-14. (来斯惟, 徐立恒, 陈玉博, 等. 基于表示学习的中文分词算法探索[J]. 中文信息学报, 2013, 27(5): 8-14.)

[22] Arroyo-Fernández I, Méndez-Cruz C F, Sierra G, et al. Unsupervised Sentence Representations as Word Information Series: Revisiting TF-IDF [J]. *Computer Speech & Language*, 2019, 56: 107-129.

[23] Minaee S, Kalchbrenner N, Cambria E, et al. Deep Learning Based Text Classification: A Comprehensive Review [EB/OL]. [2020-07-02]. https://arxiv.org/abs/2004.03705.

[24] Aly R, Remus S, Biemann C. Hierarchical Multi-label Classification of Text with Capsule Networks [C]// The 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Florence, Italy, 2019.

[25] Kim J, Jang S, Park E, et al. Text Classification Using Capsules [J]. *Neurocomputing*, 2020, 376: 214-221.

[26] Zhu Fangpeng, Wang Xiaofeng. Text Classification for Ship Industry News [J]. *Journal of Electronic Measurement and Instrumentation*, 2020, 34(1): 149-155. (朱芳鹏, 王晓峰. 面向船舶工业新闻的文本分类[J]. 电子测量与仪器学报, 2020, 34(1): 149-155.)

[27] CSGPC. Evaluation Rules of Peixiu Award for Excellent Maps [EB/OL]. [2020-10-08]. http://www.csgpc.org/bencandy.php?fid=157&id=5054.

[28] Guan Qin, Deng Sanhong, Wang Hao. Chinese Stopwords for Text Clustering: A Comparative Study [J]. *Data Analysis and Knowledge Discovery*, 2017, 1(3): 72-80. (官琴, 邓三鸿, 王昊. 中文文本聚类常用停用词表对比研究[J]. 数据分析与知识发现, 2017, 1(3): 72-80.)

[29] OPENCC. Open Chinese Convert [EB/OL]. [2020-10-08]. https://github.com/BYVoid/OpenCC.

[30] LTP. Language Technology Platform [EB/OL]. [2020-10-09]. https://github.com/HIT-SCIR/ltp.

[31] Řehůřek R. Models. Word2Vec-Word2Vec Embeddings [EB/OL]. [2020-10-11]. https://radimrehurek.com/gensim/models/word2vec.html.

[32] Chen X, Xu L, Liu Z, et al. Joint Learning of Character and Word Embeddings [C]//The 24th International Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015.

[33] Rui Weikang, Liu Kai. Semantic Similarity Based Opinion Extraction [C]// The 5th Conference on Natural Language Processing and Chinese Computing, Kunming, China, 2016. (芮伟康, 刘凯. 基于语义相似度的评论观点抽取[C]//第五届自然语言处理与中文计算会议, 中国, 昆明, 2016.)

[34] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-Learn: Machine Learning in Python [J]. *Journal of Machine Learning Research*, 2011, 12(85): 2825-2830.

[35] MS. Clustering via Jenks Natural Breaks [EB/OL]. [2020-10-08]. https://github.com/MSH19/Clustering-via-Jenks-Natural-Breaks-Matlab.

[36] Comprehensive Planning Department. Notice of the Ministry of Transport on Promulgating the First Batch of Pilot Units for the Construction of Transport Power [EB/OL]. [2020-10-08]. http://xxgk.mot.gov.cn/jigou/zhghs/201910/t20191025_3288812.html. (综合规划司. 交通运输部关于公布第一批交通强国建设试点单位的通知[EB/OL]. [2020-10-08]. http://xxgk.mot.gov.cn/jigou/zhghs/201910/t20191025_3288812.html.)

[37] Jiangxi Academy of Social Sciences. Jiangxi Economic and Scocial Development Report 2020 [EB/OL]. [2020-10-08]. http://jx.people.com.cn/n2/2020/0801/c190260-34199246.html. (江西社会科学院. 江西经济社会发展报告(2020)[EB/OL]. [2020-10-08]. http://jx.people.com.cn/n2/2020/0801/c190260-34199246.html.)

[38] Chen Jun, Liu Wanzeng, Wu Hao, et al. Basic Issues and Research Agenda of Geospatial Knowledge Service [J]. *Geomatics and Information Science of Wuhan University*, 2019, 44(1): 38-47. (陈军, 刘万增, 武昊, 等. 基础地理知识服务的基本问题与研究方向[J]. 武汉大学学报(信息科学版), 2019, 44(1): 38-47.)

[39] Jiang Bingchuan, Wan Gang, Xu Jian, et al. Geographic Knowledge Graph Building Extracted from Multi-sourced Heterogeneous Data [J]. *Acta Geodaetica et Cartographica Sinica*, 2018, 47(8): 1051-1061. (蒋秉川, 万刚, 许剑, 等. 多源异构数据的大规模地理知识图谱构建[J]. 测绘学报, 2018, 47(8): 1051-1061.)