

面向产业区域布局评价需求的大模型应用

刘昌斯 马艺晏 冉迅 李欣悦

中国航天系统科学与工程研究院 北京 100076

DOI: 10.12238/ems.v7i4.12737

[摘要] 本文为解决企业选择战略新兴产业发展区域评价得需求,通过构建产业布局评价指标体系,利用“结构化爬虫数据+agent+prompt+大模型”的技术框架,实现对战略性新兴产业布局评价的自动化采集、处理和深度理解分析,探讨区域战略性新兴产业评价的新方法、新工具,同时将大模型获取数据与熵权法确定指标权重相结合,可实现对战略性新兴产业进行综合客观地评价的效果。经研究论证,其商业航天项目可考虑在京津冀、长三角、粤港澳区域加大投资;低空经济项目可考虑在粤港澳、京津冀区域进行开展;氢能项目适合在长三角区域发展;新材料项目可考虑在长三角、成渝、粤港澳区域进行扩大生产;电子信息与智慧产业可考虑在成渝、粤港澳区域发展。大模型技术在本文中的应用,使评价结果更精准、更高效、更科学,为政策制定者和企业提供跟有效的决策依据。

[关键词] 产业区域布局评价;大模型技术应用;熵权法;战略性新兴产业;企业

引言

近年来,企业传统经济增长方式主要依赖于高投入和资源高消耗的带动,部分项目产业层次较低,发展空间不足,缺乏新动力,难以实现企业的高速发展,需要强有力产业的支持。因此某企业积极响应国家战略部署,加快培育和发展战略性新兴产业,打造具备特色化、集群化战略新兴产业发展区域。在如何选择适合本企业战略新兴产业发展区域方面,本文以自然语言处理技术和深度学习算法为基础大模型为基础对战略性新兴产业区域布局进行实证定量评价,对本企业五大战略性新兴产业在京津冀、长三角、粤港澳、成渝产业区域发展进行评估,为决策者提供数据支持和决策参考。

1 现状研究及指标体系设计

1.1 现状研究

战略性新兴产业布局评估是指对特定区域内战略性新兴产业的空间分布、规模结构、发展水平及相互关系进行系统分析和评价,以揭示战略性新兴产业发展中存在的问题和潜力,为产业规划和政策制定提供依据。传统的战略性新兴产业布局评估方法主要依赖于专家经验和统计数据,存在主观性强、效率低等问题。随着数据获取和处理能力的提升,基于数据的产业布局评估方法逐渐兴起,以大模型为代表的新一代信息处理技术以强大的自然语言处理和深度学习算法为基础,能够以惊人的速度浏览并理解海量的知识库内容,够迅速识别并提取关键信息,还能通过复杂的语义分析,理解这些信息之间的内在联系与逻辑关系,迅速构建对产业区域状况的全方位认知,准确捕捉到影响区域发展的关键要素与潜在趋势,更客观、更准确地评估战略性新兴产业布局状况,为优化产业结构、推动产业升级提供有力支持。

1.2 评价指标体系设计

战略性新兴产业代表着科技创新和未来产业发展的方向,在发展初期较为依赖产业政策环境,在发展过程中其全

局性、长远性、导向性和动态性是战略性新兴产业所具有的四大特征。本文在遵循目的性、技术进步性、弹性、规模性、效益性、可持续性六大原则下,遴选出能够体现战略性新兴产业特色和优势的评价指标体系,具体包括产业政策环境、需求收入弹性系数、技术进步水平、区位商、影响力系数、战略性新兴产业动态评价指标。

2 大模型应用

2.1 产业知识库构建

2.1.1 目标网站选择与分析

面向产业区域研究需求,考虑到数据质量因素,选取北京、河北、天津、上海、浙江、安徽、江苏、广东、广西、香港、澳门、四川、重庆、成都、陕西、新疆等地区的税务局、财政部、国资委等官方网站作为目标网站。由于网站存在政策或技术层面的限制,在进行数据爬取前,分析各网站的数据结构、更新频率及访问限制,为后续爬虫设计提供依据。

2.1.2 爬虫设计与实现

对目标网站进行分析之后,采用爬取策略如下:优先使用如“八爪鱼”等第三方爬虫工具爬取数据,对于第三方爬虫工具无法爬取的网站,设计爬虫程序单独处理。

大部分政府网站政策发布网页采用分页系统组织展示政策新闻列表,通过分页导航模式(如“上一页”、“下一页”,以及直接跳转至特定页码)提供用户浏览服务,基于该机制设置爬取策略,可完成大部分政府网站数据的爬取。

对于特定无法用爬虫工具爬取的网站,设计递归爬虫程序,采用Python等编程语言,结合Scrapy、BeautifulSoup等库,实现对目标网站的深度爬取,爬虫设计考虑反爬虫机制、异常处理等功能,采取温和的爬取策略,如降低请求频率、模拟真实的用户行为等爬取数据。

2.1.3 数据处理

由于各个政府网站数据格式不统一，爬取的格式也存在差异，因此需要进行数据清洗处理。爬取数据主要包括政策新闻标题和内容文本，利用正则表达式、自然语言处理等技术，对文本数据进行分词、去停用词、词性标注等处理，提取关键信息如时间、地点、政策名等。将清洗后的数据以文本方式，按照政策时间、地点、政策名、政策内容分类存储，

并根据地域分成不同的文本文件，支撑大模型使用。

2.2 大模型选型

在国内人工智能领域，百度文心一言、阿里通义千问及月之暗面 kimi 凭借其先进的技术优势与广泛的应用潜力，占据了举足轻重的地位。三个大模型功能对比如下表 1：

表 1 主流大模型功能对比

大模型名称	同时上传文件数	单个文件大小	文件格式
文心一言	10 个	50MB	Word、PDF、TXT、PPT
通义千问	100 个	150MB	PDF、Word、Excel、Markdown、EPUB、Mobi、txt
kimi	50 个	100 MB	pdf、doc、xlsx、ppt、txt、图片

将知识库中的数据提交到文心一言、通义千问和 kimi。经过测试，kimi 由于对上传字数的限制无法实现一次分析。文心一言和通义千问可深入分析上传文档，提炼出关键结论。针对缺乏上传文档支持的问题，通义千问会利用其庞大的语料库进行检索和分析，以获取额外的支持材料。文心一言在本次分析中展现出了强大的数据理解分析能力和准确性，因此，将其作为本研究的技术基础。

2.3 大模型验证与优化

2.3.1 理解能力验证

将知识库中的数据导出后，上传到大模型并提出问题，大模型接收到请求后，利用其内置的算法和预训练模型，对上传的文档进行深度阅读和理解。通过分析文档中的文本内容，识别出关于战略新兴产业和未来产业的关键词、短语及其关联关系。随后根据这些信息，生成思维导图。见图 1：

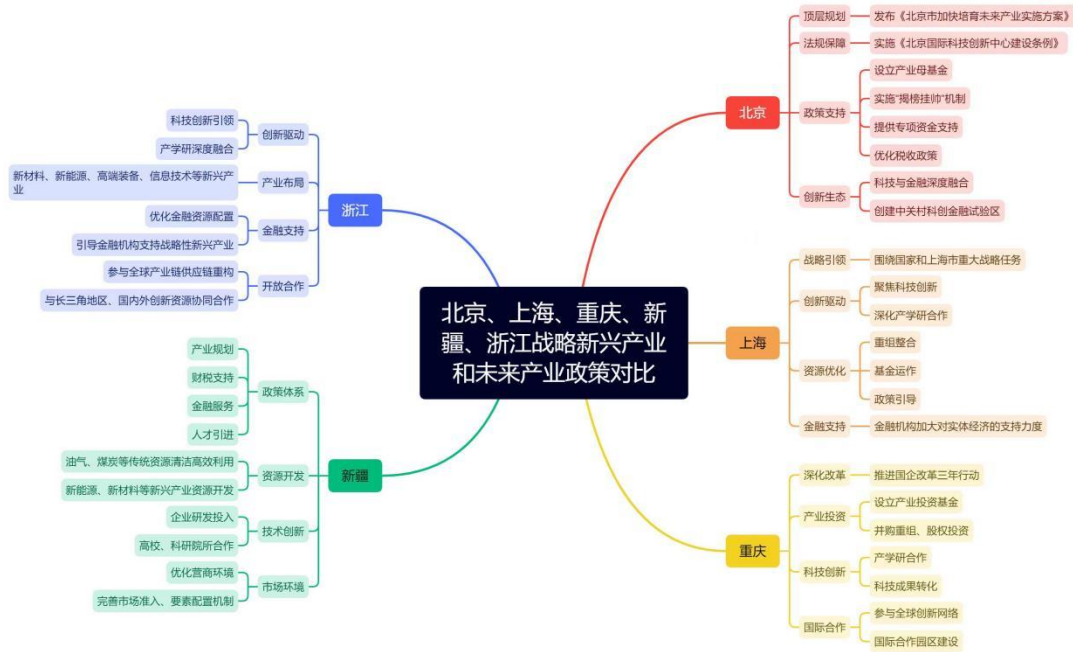


图 1 战略性新兴产业与未来产业思维导图

对比知识库中各产业方向的政策内容与多次生成的思维导图中各产业方向的政策内容，正确率达到 100%，由此可初步验证大模型在文档理解方面的准确性。

2.3.2 问答结果验证

在以上操作的基础上，向大模型提交多个产业相关问题，将涉及到的地域、产业和政策的答案在知识库中进行人工比对和解读，验证了大模型回答的正确性。由于模型涉及到的问题较多，对超出爬虫获取到的数据范围外的问题，通过对问题加入搜索范围进行提问，并对大模型返回结果的参考网

站内容进行整理解读，验证了通过搜索引擎完成问题解答的方法。

2.3.3 提示调优

结合产业问题结果分析验证情况，从以下三方面优化提示的结构和内容，提升产业大模型理解分析能力和应用价值。

明确大模型数据搜索范围。在提出问题的过程中将范围设置为国资委官网、政府官网、新华社等官方媒体网站，优化回答速度和正确性。

明确答案输出方式。在提出的问题中加入答案输出方式，

如“请使用 markdown 格式输出……思维导图”，markdown 可快速通过软件转换形成思维导图，提升答案的准确性和提问效率。

明确答案格式。将所有问题加入“请回答是或者否，如果是请列出具体规划方案名称”，增强模型对复杂文本的理解能力和文本识别的准确性、适应性。

2.4 产业大模型应用

2.4.1 问题输入设计

结合产业问题结果分析验证情况，运用明确问题、提供上下文、逐步引导、限定范围等方式优化提示的结构和内容，探索提升产业大模型理解分析能力和任务适配性的路径，并进行分析应用输出产业布局评价模型指标体系。

产业区域布局评价模型构成较为复杂，因此采用思维链的方式逐步引导，将复杂任务分解为一系列简单的步骤，明确每一步的提示和预期输出，以获取高质量的指标评分。

2.4.2 结果输出与优化

将模型输出结果与实际数据进行对比，验证模型的准确性和有效性，保证产业区域评价模型输入的准确性。

3 实例应用

3.1 权重模型构建

3.1.1 熵权法的计算

熵权法是在各指标的变异程度基础上，利用信息熵计算出各指标的熵权，再修正各指标的权重，从而得出较为客观的指标权重。

3.2 实例研究

基于“结构化爬虫数据+agent+prompt+大模型”的技术框架，实现对产业数据的自动化采集、处理和深度理解分析，并获得上述指标的计算值，依据熵权法经处理后得到标准化矩阵 R，利用公式（1）、（2）、（3）计算得到 6 个评价指标 5 大产业的熵权值，利用公式（4）计算每个区域 5 大产业加权综合得分，如表 2 所示。

标准化后每个区域得分如表 3 所示。

表 2 4 区域的综合加权得分

战略性新兴产业方向	京津冀区域综合得分	长三角区域综合得分	粤港澳区域综合得分	成渝区域综合得分
商业航天	0.1737	0.1663	0.1567	0.1456
低空经济	0.1996	0.1446	0.2002	0.1268
氢能	0.1172	0.1389	0.1232	0.1001
新材料	0.0965	0.1274	0.1163	0.1223
电子信息与智慧产业	0.0995	0.0956	0.1125	0.1211

表 3 5 大产业在 4 区域的标准化后得分

战略性新兴产业方向	京津冀区域综合得分	长三角区域综合得分	粤港澳区域综合得分	成渝区域综合得分
商业航天	100.00	95.74	90.21	83.82
低空经济	99.70	72.23	100.00	63.34
氢能	84.38	100.00	88.70	72.07
新材料	75.75	100.00	91.29	96.00
电子信息与智慧产业	82.16	78.94	92.90	100.00

3.3 结果分析

某企业商业航天项目可考虑在京津冀、长三角、粤港澳区域加大投资；低空经济项目可考虑在粤港澳、京津冀区域进行开展；氢能项目适合在长三角区域发展；新材料项目可考虑在长三角、成渝、粤港澳区域进行扩大生产；电子信息与智慧产业可考虑在成渝、粤港澳区域发展。

4 结语

本文通过构建评价指标体系，利用“结构化爬虫数据+agent+prompt+大模型”的技术框架，实现对战略性新兴产业布局评价的自动化采集、处理和深度理解分析，探讨区域战略性新兴产业评价的新方法、新工具。将大模型获取数据与熵权法确定指标权重相结合，可实现对战略性新兴产业进行综合客观地评价的效果。通过对某企业战略性新兴产业进行实证评价，表明模型分析的结果符合企业在四个区域的

发展现状，证明本文的评价达到了满意的决策结果，可为政策制定者和企业提供科学的决策依据，通过模型的应用，可以促进企业产业结构的优化升级和区域经济的协调发展。

[参考文献]

[1]张继国, Vijay P Singh. 信息熵—理论与应用[M]. 北京: 水利水电出版社, 2012

[2]刘洪昌. 中国战略性新兴产业的选择原则及培育政策取向研究[J]. 科学学与科技管理, 2011 (3): 87-92

[3]胡振华, 黎春秋, 熊勇清. 基于“AHP — IE — PCA”组合赋权法的战略性新兴产业选择模型研[J]. 科学学与科学技术管理, 2011, 32 (7): 104-110

[4]贺正楚, 吴艳. 战略性新兴产业的评价与选择[J]. 科学学研究, 2011, 29 (5): 678-683