引文格式：花叶. 地名数据语义消歧及文化属性挖掘方法[J]. 地理空间信息, 2026, 24(6):73-77. doi:10.3969/j.issn.1672-4623.2026.06.013

地名数据语义消歧及文化属性挖掘方法
——以上海市为例

花　叶
（上海市测绘院，上海 200063）

摘　要：随着大数据技术的迅猛发展，地名数据的采集规模与应用范围持续扩大，然而这些不同来源的地名数据会存在重复、冗余、不一致性等问题，严重影响了数据分析结果的可靠性及其在实际应用中的效率。传统数据融合方法通常依赖大量人工干预，处理成本较高且难以适应大规模数据环境。本研究立足于实际应用需求，系统探讨基于语义分析的地名数据处理路径，并整合本体推理、上下文感知与深度学习技术，以实现多源异构地名数据的消歧、融合、挖掘与信息提取。

关键词：语义分析；多源时空数据；文化属性挖掘；地名消歧；多源数据融合
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2026）06-0073-05

Methods for Semantic Disambiguation and Cultural Attribute Extraction in the Geographical Name Data: A Case Study of Shanghai City, China

HUA Ye
(Shanghai Surveying and Mapping Institute, Shanghai 200063, China)

Abstract: With the rapid advancement of big data technology, the collection scale of geographical name data and its application scope continue to expand. However, geographical name data from diverse sources often suffer from issues such as duplication, redundancy, and inconsistency, which significantly influence the reliability of analytical results and their practical efficiency. Traditional multi-source data fusion methods typically require extensive manual intervention, incur high processing costs, and struggle to adapt to large-scale data environments. Addressing practical application needs, we systematically explored a geographical name data processing approach based on semantic analysis, and integrated ontology reasoning, context-aware technique, and deep learning technology to achieve disambiguation, integration, mining, and information extraction of multi-source heterogeneous geographical name data.

Key words: semantic analysis, multi-source spatial-temporal data, cultural attribute mining, geographical name disambiguation, multi-source data fusion

地名数据除具备地理信息数据常见的空间（几何）属性和时间属性外，还蕴含丰富的语义信息。由于其来源广泛且具有较强的随机性，这类数据在正确性、精度和完整性方面往往难以满足传统测绘地理信息资料的应用要求，且常以非结构化形式存在。沿用传统方法进行清洗与融合，不仅需投入大量人工干预，还易导致信息丢失或冗余度增高，进而影响数据效能、质量及后续应用与分析的效果[1]。

针对多源地名数据时空维度不一致、噪声显著、格式异构及语义信息处理困难等问题，本文以基于规则库与深度学习的语义分析方法为核心，结合多源数据融合技术框架，构建地名数据清洗、属性融合与文化内涵挖掘的智能化处理流程，以提升地名数据的质量、完整性及文化属性挖掘效能，为智慧城市、数字文旅等场景提供高质量的地名数据支撑。

1　地名消歧

在多源地理信息融合场景下，同一地理实体往往呈现多源异构的地名标识、属性信息及空间表述特征（如位置描述的模糊性、坐标精度差异或历史沿革中的曾用名/别名记录），亟需构建自动化的空间语义判别机制以实现实体一致性归一化——即确保该地理实体在融合数据中仅对应唯一的坐标与地名，同时按时间序列完整保留其历史地名演变信息以消歧[2]。多源数据融合实践中面临显著挑战：其一，地名表述的语义异构性导致同物异名现象频发（如“豫园”与“上海老城隍庙”的指代重叠、“中山路”在不同行政区划中的空间指向差异）；其二，空间位置描述的主观

作者简介：花叶（1977—），高级工程师，研究方向为地图与地名文化时空数据库以及可视化表达，E-mail：1804243238@qq.com。

性（如文本型位置描述的模糊边界）与坐标数据的精度差异（如不同测绘标准下的坐标偏移）引发空间定位冲突；其三，属性描述的维度差异（如规划资源数据中的用地性质属性与智慧交通数据中的通行属性存在语义断层）加剧了数据整合的逻辑不一致性。当前研究可借助地名属性中的语义特征提取技术（如基于规则库的语义模式识别与深度学习驱动的上下文语义分析[2]），结合时空属性维度进行多模态综合判别，从而在降低人工干预的同时提升数据处理的质量与效率。

1.1 基于语义要素关联的地名去重

在不同来源的地名数据处理过程中，首要步骤是实施重复性处理，即赋予每个地名唯一的地理位置坐标。

以上海市天地图地名地址库建设为例，数据来源包括测绘部门提供的基础地形数据、导航公司提供的兴趣点（POI）数据，以及民政与规划管理部门提供的行政地名数据[3]。首先，针对多源文本地址（如“上海市黄浦区南京东路100号”），构建基于规则的分词器与正则表达式模板，实现对行政区划层级（省、市、区、街道/镇）、道路名称、门牌号、楼栋号、房间号等基本要素的初步切分与类别标注；其次，针对非标准地址表述（如“浦东新区张江高科技园区碧波路690号A座”），引入条件随机场（CRF）与双向长短期记忆网络（BiLSTM）相结合的序列标注模型，利用预训练词向量（如Word2Vec或BERT）对地址字符串进行字符级或词语级特征嵌入，输出“B-省/市/区/街道/路/号/楼/室”等标签序列，从而提取结构化地址成分[4]。在此基础上，对门址字段（如“门牌号”）进行专门处理：通过规则校验（如“数字+号”模式）结合模糊匹配（如“甲/乙/丙”后缀识别）消除格式差异；针对缺失门牌号的地址（如“漕宝路”），则利用邻近POI空间插值或基于路网的地址插值算法进行补充。最终，将上述结构化字段存入统一地址字典，并关联至地名实体，为后续相似度计算中“地址”权重的精确匹配提供标准化输入，如表1所示。

表1 地名要素结构化提取表

要素类型	作用	示例
类别	实体语义分类（层级化）	餐饮>中餐>川菜
	空间位置确定（标准化）	“淮海中路1001号”—环贸IPAM店附近
门址	唯一性标识符（强验证）	021-1235678（上海***单位）
电话	实体核心名称（同义词归一）含简称、曾用名、别名等	“交大”=“上海交通大学” “环球港”=“上海月星环球港”
专名	实体语义分类（层级化）	餐饮>中餐>川菜

如图1所示，将地名信息通过以上方式提取关键要素后，进行复合权重设计，如结合编辑距离、Jaccard系数、词向量（如BERT）的名称相似度，调整空间自适应半径距离（根据地名的密集程度），电话号码的匹配（先要进行一定规则清洗，如统一格式、去分机号等），依据“分类+地址（或电话）”、“分类+名称”、“名称+空间距离”等不同权重组合，并设置复合权重进行相似度的计算（图2）。

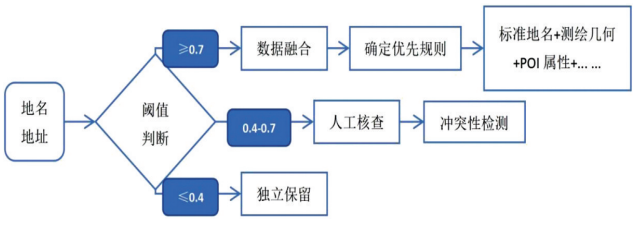


图1 相似度计算判断流程

1.2 基于语义规则库的地名标准化处理

地名标准化处理是指以官方正式命名文件或地方志为依据，明确辨识地名的标准名称、别名、简称、曾用名等属性，该过程亦为地名消歧的重要环节，对地名管理、规划建设与实际应用具有重要意义。在实际操作中，常采用分级与分类相结合的处理策略。由于该类资料多呈非结构化文本格式，需从中抽取描述地名位置信息的文本段，并与现有地名数据库中的地理实体进行快速匹配，非结构化文本中重构地理实体的空间语义以确定其标准地名，如图3所示。

通常情况下，通过叠加行政区划属性与详细地址信息可实现高效匹配。例如，某地名命名文件中记述：“合生时代锦苑，位于上海市浦东新区三林街道西营南路”，通过提取“浦东新区三林街道+西营南路”这一地址组合，可将其与地名数据库中的“合生前滩一号”实体相对应，进而确定其标准地名为“合生时代锦苑”，而“合生前滩一号”则可作为别名予以保留。该类地名位置信息的识别常可借助上下文特征实现，如出现“位于”“坐落于”等关键词。

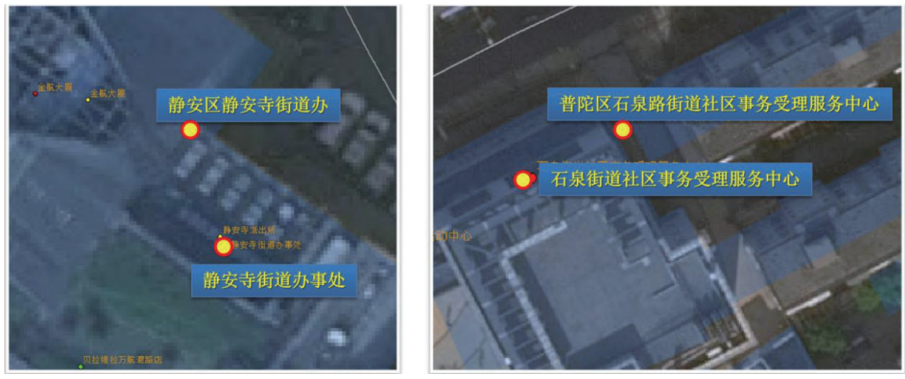


图2　天地图–上海地名重复数据自动提取（相似程度70%以上结果）

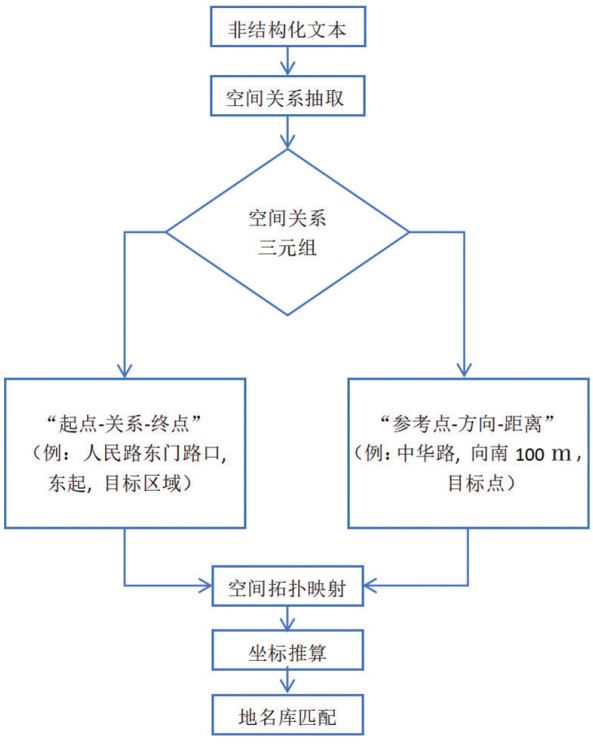


图3　空间关系解析技术路线

然而，部分地名在文献中的描述较为模糊，或主要以空间关系进行表述，例如：“东起人民路东门路口，南至中华路东刘家弄口，西到方浜中路光启路口，北达人民路龙潭路口”。针对此类空间描述，可构建描述空间关系的词汇库与语法规则库，并基于规则匹配算法实现空间关系抽取[5]。进一步可应用支持向量机、Winow 等算法构建分类器，以有效识别文本中的空间关系表达。

在构建空间关系语义计算模型时，可将地名位置词库与空间拓扑关系表进行映射匹配，如“包括”对应空间关系“IN”（包含）、“相邻”对应“EC”（外部连接）、“相隔”对应“DC”（离散）、“在...内”对应“PO”（部分重叠）等。此外，描述相对距离的语句，如“往东南100 m左右”“向西约50 m”等，作为空间约束条件，进而推算具体位置信息。最终，借助此类位置信息与现有地名数据库进行匹配，可实现地理实体的快速定位与属性关联[6]。

2　地名文化属性挖掘与融合

在地名标准化处理中，已初步应用了基于语义规则库的自动化处理方式，但该方法对规则库的完备性要求较高，应用范围存在一定局限性，尤其面对段落较长、结构复杂的文本时，需进一步引入更深入的语义分析技术。语义分析的核心目标在于理解自然语言文本（包括词汇、短语、句子与段落）的实际含义，使机器能够近似于人类理解语言所表达的内容。因此，针对书籍与文献中地名文化属性的挖掘与融合，采用机器学习方法具有显著适用性。针对不同来源的文献资料，可通过分析其文本体例与格式特征，制定相应的信息抽取策略[7–8]。

2.1　基于SAO结构的语义分析方法

目前，基于“主语–动作–宾语”（SAO）结构的语义提取方法（图4）广泛应用于科技文献的知识挖掘中，例如专业技术领域的创新点识别与发明专利的技术分析。其核心原理是从非结构化文本中识别并抽取出SAO结构，依据特定规则进行标注，从而识别出句子中与数据属性相关的关系短语及其所关联的对象[9–10]。

以上海红色地名文化挖掘课题为例，通过对一定规模的标注语料库进行训练，可依据语法规则库对各类文献中与红色地名相关的时间、地点、事件、人物等信息实现自动分类与标识，进而完成匹配与提取，从而实现将地名从文本中的“孤立标签”提升为“关系网络中的节点”，更能支撑“地名叙事”的构建，形成基于地名的知识图谱。需要特别指出的是，该方法对地名专名词典与语法规则库的完备性要求较高。若某些地名未被收录于专名词典中——例如“上海民众反日救国联合会总部”中的“民众反日救国联合会”——则相应信息可能难以被识别，从而导致遗漏[11]。

表2 部分红色地名样本数据提取结果

名称	分类	时间	地点	相关人物
中国共产党第一次全国代表大会会址	会议旧址	1921年	黄浦区兴业路76号（原望志路106号）	毛泽东、李达、李汉俊、张国焘、刘仁静、何叔衡、董必武、陈潭秋、王尽美、邓恩铭、包惠僧、陈公博、周佛海等
中国共产党第二次全国代表大会会址	会议旧址	1922年	静安区老成都北路7弄30号（原南成都路辅德里625号）	陈独秀、张国焘、李达、杨明斋、罗章龙、王尽美、许白昊、蔡和森、谭平山、李震瀛、施存统等
中国社会主义青年团中央局机关遗址	机构旧址	1920–1925年	静安区大沽路398弄中凯城市之光居住区（原址大沽400—402号）	俞秀松、袁振英等
上海学生联合会会所旧址	机构旧址	1919年	黄浦区淮海中路567弄14号	有李果、邹韬奋、钱杏邨等
鲁迅故居	人物纪念地	1933–1936年	虹口区山阴路132弄9号（大陆新村9号）	鲁迅、许广平
五卅烈士墓遗址	历史运动纪念地	1925年	虹口区新同心路1号（原方家木桥）	顾正红、何秉彝等
复旦公学旧址	教育、宣传机构旧址	1905年	长宁区华山路1626号（复旦中学）	马相伯、于右任
《新青年》编辑部旧址	教育、宣传机构旧址	1920年	黄浦区南昌路100弄	陈独秀、陈望道

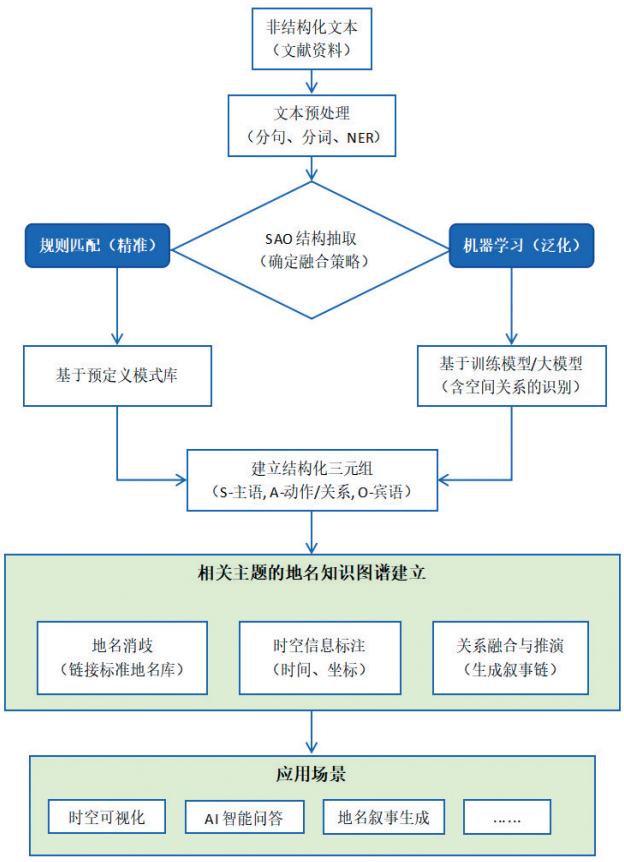


图4 基于SAO的地名文化挖掘技术流程

2.2 基于预训练语言模型的方法

与依赖预先文本标注或构建语法规则库的SAO（subject–action–object）方法相比，预训练语言模型（如BERT、GPT等）通过在大规模无标注语料上进行自监督预训练，从而能够自动提取并整合目标信息，实现对有效内容的自动化整理与输出[12]。本文以上海市红色地名文化调研课题为应用案例（表2），语料来源包括《上海党史知识读本》《上海市重要革命遗址通览》等公开出版物；在此基础上，依据领域专家意见，精选若干具有代表性的红色地名作为研究样本，并针对不同分析任务的需求，选用相应的预训练语言模型进行文本处理与分析。

以Google提出的BERT模型为例，该模型属于判别式模型，能够对文本进行自动分类，并有效识别文本中地名、人名、机构名等实体信息，实现高效的信息抽取与归类，通过调用BERT–base–uncased等模型[13]，可对非结构化文本实现自动抽取与地名实体识别，提取内容聚焦于历史人物、历史事件，并明确地理实体的主题、时间及地址等关键信息。

通过语义分析技术识别出地名和地址后，进行地名地址库匹配，筛选出准确信息，并将坐标信息和地名库中的属性信息、多媒体信息以及地名编码等挂接到地理实体信息上[14]。如图5所示。

ChatGPT作为生成式人工智能模型，能够基于已有资料进行上下文关联分析，生成逻辑连贯的文本，从而实现新文本的自动创作。在本研究的红色地名文化调研中，首先，将《上海党史知识读本》《上海市重要革命遗址通览》等文献的节选段落输入该模型，利用其上下文理解能力，自动生成符合特定红色地名的结构化解说文本，涵盖地理坐标描述、历史沿革概述、重大事件关联叙事及现状特征说明等要素。其次，在已有地名数据存在缺失或表述不一致的情况下，通过向ChatGPT提供多源文本片段与提示模板，模型可依据提示生成统一格式的地名释义条目，如“名称由来—历史事件—现状用途”三段式结构。此



图5　基于自动提取地名语义信息的应用平台

外，借助模型的生成多样性，也可对同一红色地名生成多条不同表述风格（如学术简报型、公众科普型）的文本，以支持不同应用场景下的语料需求。尽管当前在生成准确性与效率方面仍与BERT等判别式模型存在一定差距，但由于生成式模型具备通过自监督学习不断产生新颖样本数据的能力，其在构建大规模、多变体的红色地名语料库方面展现出独特优势，从而为专业地名学研究提供丰富的辅助素材与研究资料[15]。

3　结　语

基于语义分析的多源时空数据清洗与文化属性挖掘，正从规则驱动迈向大模型智能认知阶段。未来地名语义分析将呈现四大趋势：一是多模态语义理解，融合遥感影像、街景图片、语音及社交媒体信号，实现地名时空穿越式认知；二是生成式知识发现，借助大语言模型自动生成历史沿革、红色事件讲解及文旅导览文本，推动地名从静态标识向动态文化载体转变；三是轻量化与边缘部署，使智能服务下沉至手机、车载导航与AR眼镜等终端；四是跨学科协同生态，打破地理信息、历史、语言、社会学与计算机科学壁垒，共建“数字地名学”交叉研究平台与开放文化计算基础设施。最终目标是构建融合空间、时间与人文维度的数字化语义关联网络，实现从“地理信息”到“人文智能”的跨越，为智慧城市注入文化底蕴，为传统文化数字化传承开辟新路径。

参考文献

[1]　陆明华.“天地图·上海”与主节点数据融合方案研究[J].测绘与空间地理信息,2016,14(7):169-172

[2]　杜超,李玲,张玲.“天地图·四川”与主节点数据融合研究[J].测绘,2019,62(3):33-36

[3]　王春光,尹志永,吕从,等.地理信息公共服务平台数据融合方法初探[J].科技创新导报,2015,12(22):106-108

[4]　杨珩.GIS在上海市地名信息化中的综合应用[J].地理空间信息,2019,17(1):29-33

[5]　张雪英,闾国年,杜咪,等.大数据驱动的地名信息获取与应用[J].现代测绘,2017,40(3):1-5

[6]　葛咏,满旺,任周鹏,等.地名制图应用研究:以地名普查成果为例[J].测绘地理信息,2021,46(2):21-27

[7]　王建.关于红色资源挖掘与利用研究:以北京市延庆区为例[J].理论建设,2020,40(6):100-106

[8]　郭仁忠,应申.论ICT时代的地图学复兴[J].测绘学报,2017,61(10):1274-1282

[9]　李晓曼,宋红燕.面向专利情报研究的SAO语义结构分析方法述评[J].情报科学,2020,41(5):619-627

[10]　刘瑜,张飞舟,王劲峰.地理空间人工智能:概念、框架与挑战[J].测绘学报,2022,51(12):2405-2417

[11]　李亚云.多元历史空间信息挖掘与可视化表达技术研究:以上海红色地名文化为例[J].工程勘察,2024,52(1):25-28

[12]　李德仁,王密,肖晶,等.论无所不在的时空智能[J].遥感学报,2025,29(6):95-105

[13]　花叶,姜朝芳.数字信息技术赋能城市地名管理的探索与实践[J].遥感学报,2024,33(5):102-107

[14]　Han X, Zhu Z, Yu P, et al. Open Prompt: An Open-source Framework for Prompt-learning[C]//Proceedings of ACL 2022 System Demonstrations,2022:121-130

[15]　赵鹏军,孙胤社,王涵.基于深度学习的城市地名空间分布演化模拟[J].地理研究,2023,42(5):1123-1136