引文格式：徐磊,孙伟,白洁,等. 基于互联网数据的地形图变化快速发现技术研究[J]. 地理空间信息, 2025, 23(6):117-120. doi:10. 3969/j.issn. 1672-4623.2025.06.026

基于互联网数据的地形图变化快速发现技术研究

徐　磊[1]，孙　伟[1]，白　洁[1]，杨　霞[1]，娄　坤[1]
（1. 武汉市测绘研究院，湖北 武汉 430022）

摘　要：传统的基本地形图更新主要采取人工巡检和定期复测的方法，导致更新周期长、数据现势性无法保证的问题，这极大地限制了使用价值和应用范围。为了更有效地捕捉城市变化信息，利用爬虫技术对互联网上丰富的地理信息变化数据进行获取，解析为地理信息变更要素。这些数据依据形式的不同，被细分为文本数据、结构化数据和瓦片数据。文本数据利用大语言模型对项目名称、建设状态、地址等关键信息进行提取。结构化数据提取字段中的项目名称、状态、地址等信息后进行存储。瓦片数据采用互联网纯道路数据，经过二值化、两期瓦片正反向相减、TMS规则配准以及栅格图像矢量化等处理，获得道路变化区域。最后，将这些变化信息汇总，利用项目名称和建设状态对冗余数据去重后，形成城市变化信息，提交给专业工作人员进行审核，并流转分发给外业人员，以进行实地的数据更新。以武汉市2023年12月的变化信息为例，此方法不仅快速准确，而且显著缩短了基本地形图的更新时间，从而确保了数据的时效性。

关键词：地理信息；快速发现；爬虫；大语言模型；图像处理
中图分类号：P283　　**文献标志码：**B　　**文章编号：**1672-4623（2025）06-0117-04

Research on Rapid Detection Technology of Topographic Map Changes Based on Internet data

XU Lei[1], SUN Wei[1], BAI Jie[1], YANG Xia[1], LOU Kun[1]
(1. Wuhan Geomatics Institute, Wuhan 430022, China)

Abstract: The traditional update methods of basic topographic maps are mainly manual inspection and regular resurvey. The long update cycle and unguaranteed data current status greatly limit its use value and application scope. To more effectively detect urban change information, we employed crawler technology to acquire and analyze rich geo-information change data on Internet, and parsed it into geo-information change elements. We categorized these data into text data, structured data and tile data according to different forms. Text data utilizes large language models to extract key information such as project name, construction status and address. Structured data extracts information such as item name, status, and address from fields and stores them. Tile data adopts pure road data from Internet, which undergoes processes such as binarization, forward and reverse subtraction of two-phase tiles, TMS rule registration, and raster image vectorization to obtain road change areas. Finally, we summarized these change information, de-duplicated redundant data by project name and construction status, and submitted the urban change information to professional staff for review and distributed to field personnel for on-site data updates. Taking the change information in Wuhan in December 2023 for example, we found that this method was not only fast and accurate but also significantly shortened the update time of basic topographic map, ensuring the timeliness and freshness of data.

Key words: geographical information, rapid discovery, web crawler, large language model, image processing

当前，我国互联网地理信息主要分布在多个类别的网站上。一方面，互联网媒体平台如新闻网站和社交网络等，承载着大量的地理信息内容；另一方面，各级政府部门和附属机构的官方网站，包括国家、省、市、县各级的建设、水利水务、交管、自然资源、发改委等部门，以及招投标系统、土地市场网和政府数据开放中心等，都是地理信息的重要来源。此外，互联网地图也提供了丰富的地理信息。特别值得一提的是，政府官网及其附属网站、知名新闻媒体以及互联网地图所提供的数据，因其极高的真实性、精确的地理信息描述、完备的属性以及及时的更新频率，成为了人们获取地理信息的重要渠道[1]。因此，本文采用互联网信息挖掘技术实现基本地形图变化信息的快速发现[2-9]。

1　技术路线

构建一个目标网站库，涵盖政府公开数据（如武汉市城乡建设局、武汉市建设工程招投标监督管理系统等）、新闻媒体（如武汉晚报、荆楚网等）以及互联网地图服务（如高德地图、百度地图等）。构建的

第一作者简介：徐磊（1991—），硕士，工程师，主要从事GIS开发工作，E-mail：820771543@qq.com。

网站库如表1所示。

表1 网站库

网站名	数据内容	数据类型
武汉市城乡建设局	建设要闻	新闻文本
	重大建设项目	新闻文本
	惠民惠农惠企	新闻文本
	施工许可查询	结构化文本
	竣工验收备案查询	结构化文本
武汉市建设工程招投标监督管理系统	招投标信息	结构化文本
武汉市公共资源交易平台招标投标交易系统	中标结果公告	结构化文本
湖北省电子招投标交易平台	中标结果公告	结构化文本
湖北省公共资源交易电子服务系统	中标结果公告	结构化文本
武汉土地市场网	用地项目公示	结构化文本
武汉市公共数据开放平台	建筑工程施工许可信息	结构化文本
	建设工程竣工档案验收	结构化文本
武汉市公安局交通管理局	占道施工交管信息	pdf文本
武汉市发展和改革委员会	重大项目清单	pdf文本
新闻媒体	武汉晚报、长江日报、掌上武汉、长江网、长江云、汉网、湖北新闻网、荆楚网、网易新闻	新闻文本
微博用户	中国建筑、观察者网、新浪湖北、国网湖北电力、中建三局、武汉交通、中国基建报	新闻文本
瓦片数据	高德地图、百度地图	瓦片数据

目标网站中涉及到基本地形图变化的信息，通过数据对比分析，发现导致地形图变化的信息中往往存在某些特定的词汇，以此构成数据筛选的词汇库，如开工、竣工、通车、开通、落成、封顶、贯通、建成、交房、交付。利用Scrapy爬虫框架，每天凌晨自动抓取目标网站库中前一天的信息，对于瓦片数据可以一个季度进行统一获取。通过定义的词汇库对新闻标题、微博标题、招投标公告标题进行初步筛选，选出标题中存在词汇库中词汇的数据。对于结构化的数据主要包含施工许可、竣工验收、招投标结果，注定伴随着地形图变化，可直接进行存储。

根据获取的数据类型，将数据划分为文本数据、结构化数据和瓦片数据三类。文本数据利用qwen大语言模型的推理能力，将需要提取的信息和文本一同构成prompt。例如：某某新闻内容 提取文本中的项目名称、地址和建设状态。利用LLM（大语言模型）来提取工程名称、地址和建设状态关键字。

结构化数据具有明确的项目名称、地址、建设状态等信息，可直接提取后进行数据存储。

对收集到的文本信息和结构化数据进行对比分析，若出现项目名称和建设状态都一致的情况则进行去冗余处理。去重后的数据利用地理编码技术将提取的项目地址空间化，从而构建每天的地理信息变化信息库。

对于瓦片数据，使用高德的纯道路数据，通过二值化、影像正反向相减、GDAL栅格数据矢量化、TMS规则坐标配准等技术手段，确定变化区域。

最终，利用管理平台展示变化信息，由专业人员进行审核，形成最终的变化信息，并分发给外业人员进行数据更新。

2 关键技术

2.1 变化信息采集

针对政府公开数据平台、新闻媒体和互联网地图等网站的信息爬虫时，选择使用Scrapy框架。

Scrapy框架采用了异步IO和Twisted网络库，使得爬虫在抓取网页时具有极高的效率和稳定性。相较于传统的同步IO方式，异步IO能够同时处理多个请求，大大提高了爬虫的并发能力，从而在短时间内抓取到更多的数据。并提供了丰富的扩展和中间件支持，可根据实际需求定制爬虫的功能和行为。同时提供了强大的调度器和持久化机制，可管理大量的爬取任务和数据存储。

Scrapy框架凭借其高效、灵活和可扩展的特性，在地形图变化信息采集任务中发挥了重要作用。由于不同的网站拥有不同的网页结构，利用Scrapy框架，为每个网站定制各自的爬虫程序。爬虫程序每天凌晨自动抓取目标网站库中前一天的信息，对于瓦片数据一个季度进行统一获取。

在数据采集过程中，我们建立了一套词汇库，对抓取的数据进行初步筛选，筛选出可能变化的信息。爬取的信息依据类型被细分为文本数据、结构化数据和瓦片数据三类。

2.2 数据处理

2.2.1 关键字提取

在处理文本数据时，传统的NPL技术在提取关键词时存在一定的局限性。其需要依赖于大量的相关文本数据来构建关键词词汇库。其准确性和效果往往受到训练数据和模型的影响，如果训练数据不足或者模型不够先进，那么提取的关键词可能不够准确或者全面。因此，需要寻找一种更加高效、准确且适应性强的方法来提取关键词，以满足地理变化信息采集和处理的需求[10-13]。

借助qwen大语言模型的优势，使用RAG（retrieval-augmented generation)技术，将获取的文本同“提取文本中项目名称、地址和建设状态（开工、建设中、竣工）”这段话共同组成一个prompt，并传递给LLM进行文本分析，实现了对地理变化信息的深度挖掘和精准提取。该模型具备强大的自然语言处理能力，能够准确理解文本语义，提取出工程名称、地址、建设状态（开工、建设中、竣工）等关键信息。相较于传统的信息提取方法，此模型在准确性和效率上都表现出了显著的优势。这一技术的应用，极大地提升了我们在地理变化信息采集和处理方面的能力。

对于结构化信息，数据中通常包含工程名称、地址、建设状态等信息，可直接使用工程名称作为关键字。例如招投标、施工许可、竣工备案等信息。

2.2.2　冗余去重

在处理众源数据时，不可避免地存在冗余和重复信息。城市建设中的工程项目都存在其专有名词，无论在新闻、招标和备案信息中都相对统一，例如武汉协和医院质子医学中心，在微博、新闻、武汉市城乡建设局建设信息中都是采用同一命名。利用此特点，提取的项目名关键字结合建设状态可作为查重的标签进行对比，从而去除这些重复信息。数据对比中，将项目名称和建设状态都相同的数据认为是同一项目，保留一个录入到变化信息库中。对于新获取的数据在提取工程名称、地址、建设状态这3个关键词后，都会同变化信息库中的数据进行对比，从而去除冗余数据。

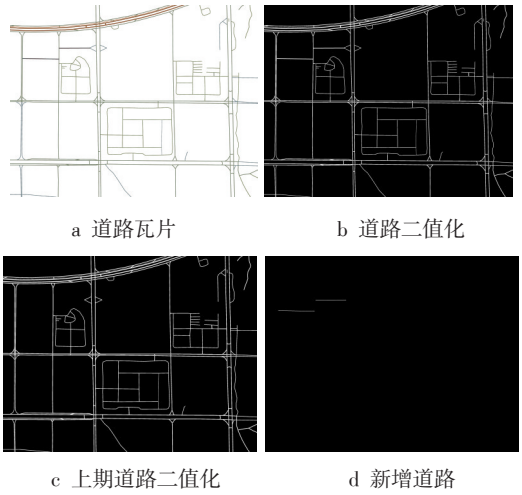
2.2.3　地理编码空间化

地理编码可将工程地址转换为具体的坐标点。通过地理编码，可将地址信息转化为精确的经纬度坐标，从而将地理变化信息与实际的地理位置相对应。通过与高德的专业地理编码服务相结合，系统可将获取的项目地址转换为相应的坐标信息。

2.2.4　瓦片数据变化信息提取

为了实现道路地理变化信息的精确提取，采用了高德API来获取纯道路数据。通过高德API，我们能够获取到高精度的道路矢量数据，这些数据为后续的图像处理提供了准确的基准[14-15]。

在获取到道路数据后，采用了二值化的方法对图像进行处理。二值化依据特定阈值将图像转换为只有黑白2种颜色，突出道路信息，去除背景噪声，得到清晰的道路图像。然后采用两期影像正反向相减的方法，提取出道路的在不同时间段内的变化信息，如新增的道路、消失的道路等。为将提取出的变化信息与实际的地理位置相对应，采用TMS规则进行配准。TMS规则是一种常用的地图切片规则，它能够将地理坐标转换为具体的像素坐标。通过TMS规则配准，能够将提取出的变化信息准确地映射到实际的地理位置上。为了将栅格数据转换为矢量数据，采用GDAL库进行栅格转矢量的操作。通过GDAL栅格转矢量，能够将提取出的变化信息转换为矢量格式，方便后续的数据存储和分析。瓦片数据变化信息提取如图1所示。



a 道路瓦片　　b 道路二值化

c 上期道路二值化　　d 新增道路

图1　瓦片数据变化信息提取

3　实施成果

依据本文的方法对武汉市2023年12月份的数据汇总情况，共发现566处存在地理信息变化的工程项目。其中开工93项、施工中4项、竣工469项。

由于在竣工备案中存在一个项目中不同建设单元单独竣工备案的情况，会导致竣工信息较多。通过同媒体新闻的相关消息比对，本文的方法准确地发现了12月份工程的建设变化情况，例如：武汉地铁19号线即将开通、武汉武车一路即将完工、华润项目中小学竣工等民生教育工程。数据来源于主流报社、官方网站、政府公开数据、官方微博，数据具有极高的可靠性，可作为基本地形图的快速更新的依据。

4　结　语

本研究基于互联网中关于城市的建设新闻、备案工程、招投标工程和道路变化等信息，为基本地形图变化发现带来了新的解决方案。该方法充分收集丰富的互联网信息，并利用LLM、地理编码、图像处理技术提取其中的变化信息，有效识别城市变化区域，为基本地形图的快速更新提供了数据支撑，显著缩短了更新周期，节约了成本。相较于以

往利用NPL技术提取文本关键信息，需要大量训练新闻文本和名词库等信息，提取的信息也会根据文本中词汇出现的频率而发生偏差。本文利用LLM的强大推理能力，能更准确地提取关键信息，为数据准确去重提供了保障。

参考文献

[1] 张大骞,刘善磊,潘九宝,等. 基于互联网的省级重要基础地理信息变化快速发现技术研究[J]. 测绘与空间地理信息, 2019,42(8):27-29

[2] 关茜,龙云涛,司连法,等. 一种基于众源地名数据构建全球中外文地名数据资源建设方法[J]. 测绘通报,2023(2):167-171

[3] 龚婧媛,孙海燕,钱志坚. 基于众源地理数据的居住环境空间分异研究:以武汉市为例[J]. 测绘地理信息,2019,44(6):11-15

[4] 李德仁. 迎接地理信息产业与IT产业的大融合[J]. 测绘地理信息,2015,40(5):1-3

[5] 单杰, 秦昆, 黄长青, 等. 众源地理数据处理与分析方法探讨[J]. 武汉大学学报(信息科学版),2014,39(4):390-396

[6] 李思谕,龚竞,杨军. 省级地理信息动态更新技术体系研究[J]. 测绘通报, 2021(1):130-137

[7] 周全斌. 省级基础地理信息更新机制[J]. 甘肃科技,2010,26(12):49-50

[8] 温秀萍. 省级基础地理数据更新技术研究与实践[M]. 测绘通报,2009(1):68-71

[9] 范攀峰. 县域基础地理信息变化发现机制及方法研究[J]. 测绘地理信息,2017, 42(6):73-75

[10] 张晓丽. 面向新闻领域的关键词提取方法研究及系统实现[D]. 太原:山西大学,2021

[11] 余鑫. 基于AI的新闻关键内容及热点分析系统的技术与实现[D]. 北京:北京邮电大学,2020

[12] 刘德荣. 基于语义聚类和加权TextRank的新闻关键词提取方法的研究与实现[D]. 北京:北京邮电大学,2021

[13] 郭俊枫,赵仁亮,郑娇龙. 面向网页文本的地理要素变化发现[J]. 地理信息世界,2015,22(1):52-56

[14] 王艳东,魏广泽,刘波,等. 众源矢量路网融合系统的设计与实现 [J]. 地理空间信息,2021,19(6):1-5

[15] 田文文. 基于自发地理信息的空间数据变化发现与更新方法研究[D]. 武汉:武汉大学, 2013

（上接第112页）

参考文献

[1] Kalman R E. A New Approach to Linear Filtering and Prediction Problems[J]. Journal of Fluids Engineering, 1960, 82(1): 35-45

[2] 周先林,张慧君,和涛,等. GPS/INS松耦合组合导航的自适应卡尔曼滤波算法研究[J]. 时间频率学报,2020,43(3):222-230

[3] 张友民,戴冠中,张洪才. 卡尔曼滤波计算方法研究进展[J]. 控制理论与应用,1995,11(5):529-538

[4] Huang Y, Zhang Y, Wu Z, et al. A Novel Adaptive Kalman Filter with Inaccurate Process and Measurement Noise Covariance Matrices[J]. IEEE Transactions on Automatic Control,2018,63(2):594-601

[5] 李松,唐小妹,孙鹏跃,等. GNSS/INS紧组合最大熵卡尔曼滤波算法[J]. 全球定位系统,2020,45(4):1-8

[6] 葛志敏,江金光,张超,等. 改进抗差自适应EKF算法在GNSS/INS组合导航中的应用[J]. 大地测量与地球动力学,2023,43(7):740-744

[7] 胡佳. 基于GNSS/SINS的高精度组合导航定位技术研究[D]. 桂林:桂林电子科技大学,2022

[8] 黄友灿,陈玉林,薛阳,等. 一种GNSS定位结果质量评价指标[J]. 大地测量与地球动力学,2023,43(11): 1 167-1 171

[9] Duchoň F, Hanzel J, Babinec A, et al. Improved GNSS Localization with the Use of DOP Parameter[J]. Applied Mechanics and Materials,2014,611: 450-466

[10] Su C, Guo S, Liu X, et al. Signal Quality Assessment of BDS-3 Preliminary System[J]. Journal of Electronics & Information Technology,2020,42(11): 2 689-2 697

[11] Lu Y, Zhong L D, Di Z, et al. Quality Assessment Method of GNSS Signals Based on Multivariate Dilution of Precision[C]//European Navigation Conference (ENC), Helsinki Finland,2016:1-6

[12] 陈起金. 基于A-INS组合导航的铁路轨道几何状态精密测量技术研究[D]. 武汉:武汉大学,2016

[13] Zhuo Z, Da T L, Liu L S. A Performance Compensation Method for GPS/IN Integrated Navigation System Based on CNN-LSTM During GPS Outages[J]. Measurement: Journal of the International Measurement Confederation, 2022,188: 110 516

[14] Jiang H, Shi C, Li T, et al. Low-cost GPS/INS Integration with Accurate Measurement Modeling Using an Extended State Observer[J]. GPS Solutions,2021,25(1):1-15

[15] 唐晓霏,张士柱,厉芳婷,等. 基于TEQC的HBCORS GNSS数据质量评估方法与影响因素分析[J]. 地理空间信息,2024,22(1):106-108

[16] 罗保林,金飞,罗亮. 自适应卡尔曼滤波在GNSS沉降监测中的应用[J]. 地理空间信息,2023,21(10):73-75