大语言模型视阈下方志资源价值挖掘与开发利用新论*

赵海良

提　要：以 ChatGPT 为代表的大语言模型具有强大的自然语言理解和生成能力，其对高质量训练语料的需求，为方志资源的价值挖掘提供新的机遇；其跨学科、多用途的应用场景，为方志资源的开发利用提供新的实践路径。简要介绍当前大语言模型发展的现状及面临的主要问题，详细探讨地方志在大语言模型训练中的语料价值，重点介绍大语言模型在方志知识图谱构建、知识化检索服务等方面的开发利用实践。

关键词：大语言模型　地方志　开发利用

“大语言模型（Large Language Model，LLM），也称大模型，是指能够处理和生成自然语言文本的模型，主要基于机器学习和人工智能技术，通过学习大量的语料数据，能够生成具有连贯性和语义准确性的文本。”① 以 ChatGPT 为代表的大语言模型展现了强大的自然语言理解和生成能力，将高质量的专业领域数据与大语言模型能力融合发展，助力各行业加快数字化、智能化转型是当前大语言模型跨学科研究的最新热点。地方志作为我国特有的优秀传统文化资源，种类繁多、体量巨大，充分开发和利用方志资源对传承中华文明、挖掘历史智慧、服务社会发展具有重大的现实意义。在大语言模型被各行各业广泛应用的背景下，深入分析方志资源的语料价值，探索大语言模型在方志资源开发利用领域新的实践模式，是未来地方志事业持续创新发展的有效路径之一。

一　大语言模型发展现状及面临的主要问题

自 ChatGPT 问世以来，大语言模型迅速成为全球科技的热点，被视为近年来人工智能领域取得的最大突破。大语言模型不仅在人机对话、数据分析、任务处理等方面展现出非凡的能力，更被认为是向通用人工智能（Artificial General Intelligence），即具备人类认知和推理能力的人工智能发展迈出了关键一步。

（一）大语言模型发展现状

大语言模型的产生和发展经历了“统计语言模型、深度学习语言模型、预训练语言模型、超大规模预训练语言模型”② 四个阶段。得益于算力资源、数据量和算法的提升，当前出现的超大规模生成式预训练语言模型（GPT）标志着大语言模型发展进入一个新阶段。与传统人工智能模型需要依赖大量有标签数据的监督训练，而且一个模型只能解决一个任务，适用于单一场景不同，“生成式预训练模型主要是通过在极大规模语料（通常为上百亿词或更多）上进行无监督预训练，学习语料的规律与知识，进而能够根据不同的输入，自主生成自然流畅的类似人类语言的

*　本文为浙江省哲学社会科学规划课题“人物数据库创新研究”（19NDQN369YB）研究成果。

①　杨青：《大语言模型：原理与工程实践》，电子工业出版社，2023 年，第 12 页。

②　闫啸彤、唐晓彬：《大语言模型发展综述》，《统计学报》2024 年第 4 期。

文本”①。当前大语言模型产业蓬勃发展，国内外高科技企业及研究机构纷纷投身大语言模型研发，除 OpenAI 推出的 ChatGPT 外，国内的 DeepSeek、百度的文心一言、阿里的通义千问以及华为的盘古大模型等大语言模型也相继问世，见表 1。

表 1　部分主流大语言模型情况表

模型名称	发布公司	参数数量（亿）	中文支持性	发布时间
GPT－4	OpenAI	1750	一般	2023 年 3 月
PaLM 2	Google	340	一般	2023 年 4 月
DeepSeek	深度求索	6710	优秀	2025 年 1 月
文心一言	百度	2600	优秀	2023 年 3 月
千义通问	阿里巴巴	1100	优秀	2023 年 4 月
星火大模型	科大讯飞	1700	优秀	2023 年 5 月
豆包大模型	字节跳动	2000	优秀	2024 年 5 月
混元大模型	腾讯	3890	优秀	2023 年 9 月
盘古大模型	华为	1000	优秀	2023 年 7 月

与此同时，大语言模型的实践应用已遍及多个领域。在教育方面，大语言模型被用于生成个性化学习资料、辅助学生进行学习并提供实时反馈。② 在医疗领域，大语言模型被用于电子病历的自动化处理、疾病诊断支持和医疗信息检索，并可以在疾病预防和健康管理等方面提供相关建议。③ 在法律行业，大语言模型被用于进行法律文档审查、合同分析和案件预测，提高法律服务的准确性。④ 总体而言，当前大语言模型在各类跨领域应用均展现出强大的跨学科性和多用途性，为各行业创新发展带来新机遇。

（二）大语言模型发展面临的主要问题

大语言模型强大的推理和计算能力，主要来源于算法、算力和语料三个关键要素。⑤ 相关研究指出：“模型的推理和计算速度取决于算力，但模型推理和计算的质量，20% 由算法决定，80% 由语料质量决定，高质量的训练语料是提升模型性能的关键。”⑥ 高质量训练语料，尤其是高质量中文训练语料的短缺，是当前大语言模型发展面临的主要问题。

目前大语言模型的训练语料，其主要来源是以网页内容、新闻文章、论坛帖子、社交媒体帖子等为主的用户生成内容（UGC）。但由于用户生成内容的开放性，由此训练的大语言模型可能存在幻觉问题（生成虚假内容）、时效性问题（受到训练语料时效的影响）等。⑦ 为应对上述问

① 冯皓：《大模型在自然语言处理中的应用方法研究》，《数字通信世界》2024 年第 10 期。
② 参见苏祺：《大语言模型在二语教学中的应用效能解析》，《外语界》2024 年第 3 期。
③ 参见胡振生、杨瑞：《大语言模型在医学领域的研究与应用发展》，《人工智能》2023 年第 4 期。
④ 参见魏斌：《法律大语言模型的司法应用及其规范》，《东方法学》2024 年第 5 期。
⑤ 参见杨青：《大语言模型：原理与工程实践》，第 58 页。
⑥ 罗云鹏：《大模型发展亟需高质量“教材”相伴》，《科技日报》2024 年 1 月 15 日，第 6 版。
⑦ 参见刘亦石、周亚建：《人工智能大模型应用中的安全问题与解决策略》，《网络空间安全科学学报》2024 年第 2 期。

题，有学者提出，严谨并获得反复验证的百科全书式知识信息是大语言模型最为可靠的语料库，并以维基百科为例，认为由于其广泛性、权威性、更新性和多语言性，是当下最流行大语言模型通用语料来源。① 在国内，也有学者采用维基百科作为通用语料进行大语言模型训练，训练效果虽然优于国内的中文类百科，但由于其开放性且主要参与的编辑为境外人员，通过其训练出的模型很容易继承和放大训练数据中存在的偏见，尤其是某些政治观点、意识形态或对事件的表述②，此类模型的广泛使用，势必会对我国的文化安全造成威胁。

有报告指出："全世界高质量的大语言模型训练语料存量将在2026年耗尽，低质量的语料和图像数据的存量则分别在2030年至2050年、2030年至2060年枯竭。"③ 更加需要强调的是，"目前国际主流大语言模型，训练语料均以英文为主，全球通用的50亿大语言模型数据训练集里，中文语料占比仅为1.3%"④。中文语料不仅规模较小，其电子化和网络化程度也明显不足。受版权、隐私等限制，许多优质中文语料无法公开获取，缺乏高质量中文语料是当前国内大语言模型发展面临的最紧迫问题。

二 地方志在大语言模型训练中的语料价值

为更好地理解客观世界、掌握客观规律，大语言模型需要学习包含大量知识的数据，这些数据深受生产者主观意志的影响。得益于语言优势，暗含西方价值观的英语语料数据在大语言模型训练中被广泛使用，长期浸润在此类大语言模型营造的特定知识体系中，公众价值观念必然会受到潜移默化的影响，由此更容易认同或从属于相应的价值观念或意识形态。有报告指出："中文语料数量的短缺或许尚有解决方案，但具有中式价值观的语料不足，则会成为制约我国大模型发展的短板。其中，文言文、古汉语、电子书等反映优秀传统文化的内容，以及主流媒体发布的反映本土价值观的内容，都可视为具有中式价值观的高质量语料。"⑤ 有学者认为，"地方志内涵丰富、形态完整、特色鲜明，是具有主体性、系统性和连续性的本土知识体系，是当下建立中国自主知识体系可资取用的历史资源"⑥，基于该论断，笔者认为地方志是具有中式价值观的高质量语料。

从数据规模而言，大语言模型的训练语料必然属于大数据。事实上，方志资源的大数据特性在较早之前就被学界所关注，认为方志资源的属性与大数据的4V特性［Volume（大量）、Veracity（真实性）、Variety（多样性）、Velocity（高速）］有着很高相似度。⑦ 除此之外，大语言模型高质量训练语料还应具备"多样性、大规模性、合法性、真实性、连贯性、无偏见等特征"⑧。笔者

---

① 参见 Fan Gao, et al.: Large Language Models on Wikipedia-Style Survey Generation: an Evaluation in NLP Concepts, 2024年5月23日, https://arxiv.org/abs/2308.10410, 2024年7月26日。

② 参见张华平、李林翰、李春锦：《ChatGPT中文性能测评与风险应对》，《数据分析与知识发现》2023年第7期。

③ 张凌寒：《加快建设人工智能大模型中文训练语料数据库》，《人民论坛学术前沿》2024年第13期。

④ 中国软件测评中心：《人工智能大语言模型技术发展研究报告（2024年）》，2024年6月，http://download.people.com.cn/jiankang/nineteen17114578641.pdf，2024年7月22日。

⑤ 阿里巴巴研究：《大模型训练数据白皮书》，2024年5月6日，https://runwise.oss-accelerate.aliyuncs.com，2024年8月13日，第5页。

⑥ 陈野：《地方志资源与中国自主知识体系的建构》，《光明日报》2024年5月11日，第11版。

⑦ 参见林浩：《解析地方志资源大数据特性管窥地方志"计算"研究新领域》，《中国地方志》2019年第4期。

⑧ Alon Albalak, Yanai Elazar: A Survey on Data Selection for Language Models, 2024年3月8日, https://arxiv.org/abs/2402.16827, 2024年8月22日。

认为，地方志作为“一地之百科全书”，除具有大数据属性外，也具备高质量训练语料的特征。

在多样性方面，首先是种类的多样性，朱士嘉在《方志之名称与种类》一文中将方志分为“统志、总志、通志、郡县志”等22 种①，现代研究者沿用前人按地域分类的方法，根据志书实际编修情况，增加“市志、土司志、盐井志、特别区志、盟志、旗志、地区志、区志”② 等分类名称。其次是门类的多样性，余绍宋在《浙江省通志编纂大纲草案》中拟定的门类包括“大事纪、疆域考、地理考、民族考、社会考、田地考、物产考、艺文考等 26 个门类的正志，附志有杂记、两浙文征”③，而 2011 年 9 月启动编纂的《浙江通志》，最后成志更是有 113 卷，门类是前者的 4 倍多。最后是体裁的多样性，地方志的体裁不仅是“志”，《关于地方志编纂工作的规定》第三章第十三条规定：“地方志的体裁，一般应包含述、记、志、传、图、表、录等，以志为主体”④，可见其体裁概念范畴要比传统意义的“志”更广泛。

在大规模性方面，据不完全统计，“我国现存旧志 9000 余种、10 万余卷，约占我国现存古籍的十分之一。而改革开放至今，全国先后完成两轮修志工作，编纂出版省市县三级志书 1 万余部，地方综合年鉴 3 万余部，行业志、部门志、专业志、乡镇村志 3 万余部，整理旧志 3600 余部，编写规模庞大的地情资料书，形成了数以百亿字计的地方志成果群和地情资料库”⑤，体量之巨，规模之大是其他任何文献类型所无法比拟的。

在合法性方面，“自隋、唐确立史志官修制度以来，历代都把修志作为一种官职、官责，并颁布政令对修志进行统一规范”⑥。新中国成立后，在较长时间内没有全国性的关于地方志编修的法律法规遵循，修志工作随意性大、主观性强。2006 年，国务院《地方志工作条例》的颁布，突显了编修地方志属于“国家意志”行为，强化了地方志书作为“官书”的合法性。

在真实性方面，地方志作为“信史”，真实性是首要属性。地方志的真实性由多重机制保障，在编纂过程中，志书使用的资料须有根有据，入志资料会被多方核实、反复验证，在出版社审查阶段还要经过专家的多轮评审。

在连贯性方面，主要体现在内容的连贯性和编纂的连贯性，即“每部志书，特别是首次编写的方志，都贯古通今。所记各类事物，探本求源，发展变化的轨迹十分清楚。一个地方的志书创始后，每隔若干年总要修一次。历代统治阶级多次下诏修志，有的明文规定修或续修的年限。全国绝大多数的方志都有续修”⑦。

在无偏见方面，清代方志学家章学诚认为，“志乃史体，原属天下公物”⑧，认为修志应该“据事直书，善否自见”⑨。忠于事实，不直接分析评论，把观点倾向、是非褒贬、成败得失寓于记述之中，是地方志有别于其他文体的重要特征之一。

地方志记载自然、社会、政治、经济、文化等多方面情况，在内容上既有广度，又有深度，

---

① 参见朱士嘉：《方志之名称与种类》，《禹贡》1934 年第 1 卷第 2 期。
② 《方志百科全书》编纂委员会：《方志百科全书》，方志出版社，2017 年，第 206 页。
③ 余绍宋：《浙江省通志编纂大纲草案》，《浙江省通志馆馆刊》（创刊号），杭州古籍书店影印本，1945 年。
④ 中国地方志指导小组：《关于地方志编纂工作的规定》，1998 年 2 月 10 日。
⑤ 高京斋：《中国地方志与中华优秀传统文化》，《中国地方志》2022 年第 2 期。
⑥ 李培林：《坚持依法治志，推进地方志事业全面发展——在纪念国务院〈地方志工作条例〉颁布实施 10 周年座谈会上的讲话》，《中国地方志》2016 年第 7 期。
⑦ 袁太平主编：《湖北省第二届修志基础教程》，湖北人民出版社，2007 年，第 125 页。
⑧ 章学诚著，叶瑛校注：《文史通义校注》，中华书局，1994 年，第 544 页。
⑨ 章学诚：《答甄秀才论修志第一书》，文物出版社，1985 年，第 137 页。

将方志资源作为大语言模型的训练语料，可以从中提炼本土概念、挖掘社会经济发展与治理中的重大事件、典型案例，以及其他更多的区域、地方性知识和数据，形成具有理论性、启发性的案例库和知识库，这必将为我国大语言模型中文语料库的高质量发展提供方志力量。

三　大语言模型视阈下方志资源开发利用的实践路径

有学者将新中国地方志信息化建设的历史根据时间划分为“个别尝试（2001 年以前）、全面起步（2001—2007 年）、快速发展（2008—2014 年）、全国统一规划和全面繁荣（2015 年至今）”4 个阶段。① 笔者认为，若从方志资源开发利用阶段划分，可归纳为方志资源的数字化、方志资源的结构化和方志资源的知识化 3 个阶段。

（一）方志资源开发利用的 3 个阶段

一是方志资源的数字化阶段，主要是通过 OCR（光学字符识别）技术，将纸质方志文献转换为计算机可编辑的文字。传统的纸质方志文献受存储空间、保存环境和时间等因素限制，无法满足大规模使用需求。纸质方志文献也易受到自然灾害、人为破坏等因素影响，存在丢失和损坏的风险。通过将方志文献数字化，可以充分利用信息技术手段对其进行管理，突破传统方志资源开发利用的时间和空间限制。在此阶段，方志资源开发利用的主要技术手段是全文检索，重点关注文档层面的检索结果，只要根据关键词，检索返回方志资源中有关记述的文档或段落就认为其解决了信息的获取需求。

二是方志资源的结构化阶段，主要是通过实体识别等技术从非结构化的数字方志资源中提取和挖掘结构化的信息，进而对方志资源继续进行知识关联和聚合。从地方志利用需求层次角度而言，结构化的方志资源可以深入到记述内容层面，对方志所记内容进行多维度的关联、聚合和揭示。但目前各类方志资源的结构化工作，在深度及广度上仍存在不足，更多的是“聚焦于单一层次，或关注围绕某话题的细粒度内容梳理与结构化组织抽取，或关注粗粒度的地方志资源库构建与知识概览，缺乏不同粒度间内容的连通，难以实现横向的资源融合和纵向的地情演化追溯”②。

三是方志资源的知识化阶段，这是地方志开发利用的全新阶段。所谓知识化是指“在搜寻、分析、组织知识的能力基础上，根据用户所面临的具体问题与环境，融入用户解决问题的过程之中，为用户提供有效的知识应用和知识创新服务”③。以传统的全文检索为例，虽然可以从海量的地方志文本（或其他信息源）中获取准确的信息，但是对于检索结果通常不做深入分析，当用户信息需求较为复杂时，需要用户浏览多个结果才能获取所需信息。方志资源的知识化利用，是将大量知识存储于人工智能模型中，检索时可以直接根据用户的问题生成答案，能够更便捷地满足用户的信息需求。

自 20 世纪 90 年代初大规模开展方志资源数字化以来，全国各级地方志工作机构以及图书馆、档案馆等积累了体量巨大的数字化方志资源。以浙江为例，自开展全省数字方志资源统一归集以来，累计归集数字方志资源 3000 多册，总字数超 27 亿字。但当前，各级地方志工作机构虽都已建立以数字化方志资源为基础的数据库，但“未对资源库开展深入的特征构建、标签关联、

① 参见沈松平、汪凤娟：《新中国地方志信息化建设的历史回顾、存在问题及发展建议》，《中国地方志》2021 年第 4 期。

② 黎安润泽、牛力、郑金月：《解构与活化：基于双视角的地方志多粒度知识服务模型》，《图书馆杂志》2024 年第 8 期。

③ 张晓林：《构建数字化知识化的信息服务模式》，《津图学刊》2003 年第 6 期。

数据画像、质量评估等工作，资源利用的途径、方式较为单一，难以满足新形势下公众读志用志的个性化需求”① 的现象较为明显。

针对目前大规模方志资源结构松散，无法提供高效的知识关联、聚合与共享服务，导致开发利用率较为有限的情况，笔者认为下阶段地方志信息化建设的重点应转为结构化和知识化的开发利用，而大语言模型强大的跨学科语言处理能力对此有着广阔的应用空间。

（二）大语言模型在方志资源结构化利用中的实践

方志资源结构化的典型应用场景为方志知识图谱。知识图谱（Knowledge Graphs），也称语义网络，是“通过图形化的方式表示现实世界实体（即对象、事件、状况或概念）之间的网络与关系”②。知识图谱在众多领域得到了广泛应用，在方志资源开发利用上，知识图谱同样显示出极大潜力。有学者以现有方志文献为对象，结合知识图谱的有关技术，开展人物关系、物产、农业等相关研究，通过构建方志知识图谱，将地方志中有关记载进行结构化组织，进而更好地理解和利用其中的知识。③

1. 方志知识图谱构建的主要问题。

梳理已有各类研究不难发现，方志领域的知识图谱应用场景相对单一，主要集中于物产、人物等少数方志种类，相较于种类繁多的存量方志资源而言实属冰山一角。此外，方志领域的知识图谱构建相关研究，大多只是在元数据层面进行组织与发布，如人名、地名、物产名等实体识别，而且大部分元数据由人工标注整理，知识图谱的构建过程并没有实现自动化。究其原因，笔者认为主要受限于以下两个问题。

（1）遣词造句的古今混用问题。

学术上，习惯将中华人民共和国成立作为界限，之前所编称为旧方志，之后所编称为新方志。目前各类新方志虽说在体例上有所创新，但在内容上沿袭旧方志记述的现象较为明显，部分通志的古代部分尤甚。旧方志在遣词造句和语言习惯上与现代汉语具有明显差异。例如，旧志里常以单字代指前文出现过的人物或用单字对某个词语进行简写，如“朱绩（？—270），字公绪，丹阳郡故鄣（今安吉县）人，朱然之子。249 年，然卒，绩袭业，拜平魏将军，乐乡督”④ 中的“然”和“绩”分别代指前文的“朱然”和“朱绩”，“拜”则为“授官”之意。而在现代汉语中极少以这样的形式出现。此外，地名“丹阳郡”“故鄣”这类名词在现代语料中也很少出现。

（2）中文分词及语义的多样性问题。

中文分词问题并非方志知识图谱构建的特有问题，与英语单词之间有明显分隔符的情况不同，中文词语由字符组合而成，也没有时态、字母大小写等区分。中文的词语组合形式多样，在不同上下文语境下的含义也不尽相同，比英文存在更普遍的一词多义、歧义问题。此外，传统的知识图谱构建过程中实体识别通常依赖于预先定义的词表和类别，若某个实体未被收录到此表中，就可能造成无法准确地识别。⑤ 例如，“浙江通志”作为一本书的名称，因未被收入词表，会被拆分为“浙江”和“通志”两个独立的词语。

---

① 郑金月：《地方志工作数字转型的构想与实践——以浙江省为例》，《中国地方志》2024 年第 3 期。

② 刘峤、李杨：《知识图谱构建技术综述》，《计算机研究与发展》2016 年第 3 期。

③ 参见王永生、王昊：《融合结构和内容的方志文本人物关系抽取方法》，《数据分析与知识发现》2022 年第 6 期；李娜：《面向方志类古籍的多类型命名实体联合自动识别模型构建》，《数字人文》2021 年第 12 期；朱锁玲：《方志物产史料在地标农产品品牌发展中的价值及其开发利用研究》，《中国农史》2020 年第 4 期。

④ 浙江省地方志编纂委员会编：《浙江通志·人物传（一）》，浙江人民出版社，2021 年，第 14 页。

⑤ 参见刘峤、李杨：《知识图谱构建技术综述》，《计算机研究与发展》2016 年第 3 期。

处理以上两类问题，均需要专业的文史知识积累。因此传统的方志知识图谱构建过程需要大量的专业知识和人工劳动，相当耗时耗力，而且需要持续更新数据，对于人力、财力资源均有限的各级地方志工作机构来说，很难大规模开展。

2. 大语言模型构建方志知识图谱的优势。

随着大语言模型理解自然语言能力的突飞猛进，通过语义分析来进行实体识别和关系抽取是当下知识图谱构建领域研究的热点。对方志知识图谱的构建，大语言模型同样有着传统方式所不具有的优势。

首先，大语言模型具有强大的泛化能力，可以对未经标注的语料进行学习，相比于传统的实体关系抽取算法和模型，可以更快速有效地从方志文本内容中抽取出实体、关系等知识信息。① 尤其是针对非结构化的需要人工进行标注的方志记述内容，节省了大量数据处理时间。同时，大语言模型还可以自动对识别的实体进行分类和标注，将实体信息归集为不同类别，提高知识图谱构建效率。

其次，通过大语言模型构建知识图谱可以有效解决方志记述中数据不一致、资料前后矛盾问题。由于方志记述体量大且具有资料来源多样性的特点，相同事物的记述由于资料来源不同，可能出现歧义或者矛盾的情况。通过大语言模型构建知识图谱，可以对类似问题进行知识对齐，提高地方志的可靠性和准确性。②

最后，构建知识图谱的主要目的是知识融合和推理发现，“通过人物之间的关系知识图谱可以对目标人物的轨迹、社交、出行、网络等多模态行为进行挖掘并建立人物画像模型，并依托人物中心特征和边缘特征，实例化人物画像”③。大语言模型的文本归纳和总结能力，可以大大简化知识融合和推理步骤，不同志书来源的知识可以被有效地融合和整理，从而完善知识体系。

3. 大语言模型构建方志知识图谱的实践。

以浙江数字方志一体化平台中方志大模型根据“鲁迅”的记述进行人物实体提取为例，将会自动提取“姓名：鲁迅”“原名：周樟寿”“改名：周树人”“字：豫才”“籍贯：浙江绍兴”等结构化内容。④

可见相较于传统知识图谱构建过程中实体与关系抽取、知识对齐和链接等步骤，通过大语言模型进行知识图谱构建，不仅流程简单，同时效率和准确性也更高，可以充分利用大语言模型的语言理解和生成能力，将其作为核心引擎，实现方志知识图谱自动化、高效化的构建。

（三）大语言模型在方志资源知识化利用中的实践

笔者认为，相较于以知识图谱为代表的方志资源结构化的二次开发利用，地方志作为资料性文献，提供便捷化的查阅服务是其开发利用的首要目标。从纸质方志文献到数字化方志资源，不仅是载体的转换，也是查志用志手段的改变，由伏案翻阅志书到只需在计算机上输入关键词即可快速定位到相关页面和段落，极大提高了地方志查询利用的效率。但这种查询方式依然沿用着全文检索的传统范式，即“给定基于关键词的用户查询，搜索工具高效地从海量文档中检索到和

① 参见杨青：《大语言模型：原理与工程实践》，第 103 页。

② 参见高茂、张丽萍、侯敏：《基于 BERT 的百科知识库实体对齐》，《内蒙古师范大学学报》（自然科学）2023 年第 5 期。

③ 赵海良：《历史人物关系的量化及其可视化研究》，《第二届方志文化国际学术研讨会论文汇编》，内部资料，2019 年，第 798 页。

④ 参见浙江省地方志编纂委员会编：《浙江通志·人物传（二）》，浙江人民出版社，2021 年，第 666 页。

该查询需求有关的文档，并按照先后顺序返回给用户”①。传统全文检索最大的弊端在于对检索结果不做深入分析，只是根据关键词匹配的结果按照页码顺序机械式返回，但一本志书少则几百页，多则几千页，往往需要浏览多个结果才能找到所需的资料。例如，当需要从《茶叶志》中查询某一年度的茶叶产量时，如果以“茶叶”或以“茶叶产量”为关键词检索，基本每页都会有关键词匹配，必须逐条分析关键词命中的段落所记述的是否为某一年度的茶叶产量，虽说比传统一页页的翻阅查询已然提高了效率，但仍旧费时费力。

随着信息技术的不断发展，地方志内容形式也变得越来越多样，表格、图片甚至音频、视频被嵌入到志书的记述中，许多新编地方志已从只有文字的单一文本变为了多种内容形式并存的多模态文本。但“基于关键词的全文检索技术对于复杂文档的语义建模能力相对较弱”②，因此并不完全适用于当下内容形式逐渐多样的地方志文献。

查阅地方志的本质是为了信息的获取，而以 ChatGPT 为代表的生成式大语言模型和以全文检索为代表的检索模型是两种不同的信息资源整合和获取方式。一方面，大语言模型可以更好地理解用户查询意图，使信息检索更加准确和智能，它可以生成更精炼的摘要，帮助用户快速获取所需信息。另一方面，大语言模型的实时问答能力使用户通过简单提问即可获取所需信息，极大提升了用户交互体验。可以说，大语言模型为地方志查询利用提供了全新范式。

另一个巨大挑战是，互联网搜索引擎已经成为人们日常获取资料的首要途径。得益于体量巨大的互联网大数据，在辅以大语言模型优秀的语义理解能力，互联网搜索引擎已经变得越来越智能和快捷。以百度搜索为例，当查询“2010 年浙江省茶叶产量”时，电脑会自动进行语义分析并显示搜索结果。

但相较于地方志书，互联网搜索最大的弊端在于其数据来源于互联网用户生成的内容，权威性不足。同时“大语言模型作为黑盒模型，其内部工作机制无法对其输出内容提供合理的推理解释”③，若没有真实、可靠的语料数据支撑，其输出的准确性也难以保证。

一般情况下，根据大语言模型所用语料库的不同，可分为通用模型和专业（垂类）模型。两者最大的区别在于，通用模型训练数据通常来源于广泛的互联网数据，覆盖不同专业和行业，以确保模型具备多样性。而专业模型训练数据则是针对特定领域的数据，通常要包含大量该领域相关的专业知识（如金融、法律、医疗等）。将大语言模型的语义理解和文字总结能力与地方志相结合，即可构建出一个基于地方志资料库的垂类对话模型。

以“浙江数字方志一体化平台”中基于《浙江通志》打造的智能问答系统为例，同样查询“2010 年，浙江省产业的产量”④，可以直接通过提问的方式进行查询，其同样给出了类似互联网搜索引擎的回答，但两者结果却不一致，搜索引擎的结果为“16.6 万吨”，平台的结果为“16.3 万吨”，搜索引擎的结果并未标识其数据来源，但平台返回的答案是在志书中有明确记载的，且给出了志书原文记载的溯源页面，相比之下更加真实可靠。

在此基础上，基于地方志资料垂类模型的智能问答系统也将进一步强化地方志的“资政”作用。当前地方志发挥资政作用的掣肘，一方面在于信息大爆炸时代，地方志的资料性价值日渐式微，另一方面最重要的是“一部方志往往规模庞大，少则十万字，多则数以亿计，但分解的

① 化柏林：《文本信息分析与全文检索技术》，科技文献出版社，2008 年，第 68 页。
② 邹傲、郝文宁：《基于预训练和深度哈希的大规模文本检索研究》，《计算机科学》2021 年第 11 期。
③ 熊明辉、池骁：《论生成式大语言模型应用的安全性——以 ChatGPT 为例》，《山东社会科学》2023 年第 5 期。
④ 新编《浙江通志》资料下限为 2010 年 12 月 31 日。

每一部分都较单薄，缺乏前后的因果关系，造成资料断档、空白和脉络不清，对资料完整利用造成不小困难”①。在基于地方志资料库的智能问答系统中，用户提问方式可以具有多样性和复杂性，大语言模型利用其语义理解的能力，可以准确理解用户查询意图，从方志资料中快速找到相关内容（见表2）。大语言模型还可以对地方志记述中包含的隐性知识进行深入分析，挖掘出知识点之间的隐形联系，对志书中的有关记述进行总结提炼，应用场景更加广泛和智能。

表2　基于地方志文献的知识库问答系统应用示例表

应用场景	典型问题输入（Prompt）	来源志书
以地方志书为原始资源，进行百科知识问答	帮我介绍一下鲁迅的详细情况？	《浙江通志·人物传》
帮助决策者从海量地方志文本数据中提取有价值的信息并做分析，提高决策的质量和效率	请根据浙江省“十五计划期间”国民经济统计数据，分析一下“十五计划”的执行情况？	《经济总情志》《发展计划志》《统计志》……
通过语义级的理解，满足科学研究信息检索的需求	请帮我整理一下2000—2010年期间，西湖龙井茶的产量情况，并分析一下，天气对产量的影响？	《农业志》《茶叶志》《气象志》……
……	……	……

结　语

当下，大语言模型已经逐渐成为推动各领域创新发展的重要力量。地方志作为记录地方历史、文化、经济、社会等多方面内容的重要文献资源，正迎来与大语言模型融合发展的最佳契机。这一融合不仅可以极大提升地方志资源开发和利用效率，还能为地方志本身的更新、丰富和传承注入新的活力。大语言模型在处理和分析大量文本数据时展现出了优越的能力，可以在海量的地方志中快速识别出关键信息、规律和趋势，并通过自然语言生成技术，将这些知识以更加生动、易于理解的方式呈现给用户，这无疑为地方志数字化和智能化提供了新的可能性。另一方面，地方志作为大语言模型的训练语料，也展现出其独特价值。地方志通常具有丰富的地域特征和文化内涵，是研究地方社会发展、风俗习惯以及历史变迁的重要资料。这些独特的地方背景，使得地方志的内容不仅可以增强大语言模型的多样性和适应性，还可以将其中的知识结构和文化背景纳入大语言模型训练，生成更加符合地方特色的文本，使得大语言模型生成的内容更具本地化和人性化。本文探讨的方志资源与大语言模型融合发展模式，为地方志开发利用提供了新的实践路径。对于地方志工作者来说，借助大语言模型的力量，可以更高效地进行资料整合与分析，通过智能化的手段提高工作效率；对于研究者而言，大语言模型能够辅助进行更为深入的学术研究，打破传统研究局限，发现新的研究视角。

（作者单位：浙江省地方志工作办公室）

本文责编：周　全

① 吴道松：《地方志开发与利用的几点思考》，《福建史志》2019年第1期。