本文在浅析GPT类技术影响智能信息处理的主要方式基础上，提示多模态大模型正在给智能信息处理带来颠覆性变革，主要体现在大模型、多模态、跨语言等方面。文本生成、智能对话、机器翻译等波及智能信息处理的颠覆性应用将对信息处理的自动分析、提取、生成和决策带来巨大的效率提升，从而带来智能信息处理的颠覆性变革。目前，这一革命浪潮正以不可阻挡之势席卷而来，学界和业界须高度关注并积极应对GPT类技术应用开启的智能信息处理革命。（王静静等[[1]](#footnote-1)，）

【目的】 剖析ChatGPT的基础技术原理，探讨其对大语言模型发展产生的影响，以及对多模态大模型发展思路产生的影响。【方法】 通过分析ChatGPT的发展过程和技术原理，探讨指令微调、数据采集与标注、基于人类反馈的强化学习等模型构建方法对大语言模型产生的影响。分析当前多模态大模型构建过程中遇到的关键科学问题，并借鉴ChatGPT的技术方案，探讨多模态大模型未来的发展发展思路。【结论】 ChatGPT为预训练大模型向下游任务的发展提供了良好的参考技术路径，未来的多模态大模型构建以及下游任务实现过程中，可以充分利用高质量的指令微调等技术来显著提升多模态大模型的下游任务性能。（赵朝阳等[[2]](#footnote-2)，）

以ChatGPT及其大语言模型为代表的人工智能将对经济学研究范式产生深远影响。目前，经济学实证研究一般使用参数维数较少、经济可解释性较强的小模型。然而，经济金融系统是一个超高维动态复杂系统，受多种因素的影响，且这些因素之间的关系呈现非线性与时变性特征，小模型无法刻画其本质规律。大模型可有效减少系统偏差，更好刻画复杂经济系统的特征与运行规律；而海量数据的使用可避免模型过度拟合，使大模型具有较好的泛化能力即样本外预测能力。为支持经济学及社会科学其他领域的大模型的估计、推断与预测，需要利用人工智能技术整合各种异构、异源、异频数据，构建大规模计量经济学数据库，并加强大算力等信息技术基础设施建设。目前，ChatGPT及大模型等前沿人工智能技术仍存在局限性，如：无法像人类一样进行批判性思考或想象，只有预测能力；基于大数据的人工智能因果推断本质上是一种统计关系推断，需要引入经济理论或实验方法帮助识别真正的因果关系；人工智能技术不能改变经济学实证研究从样本推断总体性质的本质；同时，由于互联网大数据存在大量虚假信息或样本选择偏差等问题，基于人工智能所获得的结论的可靠性需要验证。（洪永淼等[[3]](#footnote-3)，2023）

[目的]简要介绍ChatGPT的主要技术创新，使用9个数据集在4个任务上测评ChatGPT的中文性能，分析ChatGPT潜在的风险以及中国应对策略建议。

[方法] 使用ChnSentiCorp数据集测试ChatGPT和WeLM模型，EPRSTMT数据集测试ChatGPT和ERNIE 3.0 Titan， 发现ChatGPT在情感分析任务上与国产大模型差距不大；使用LCSTS和TTNews数据集测试ChatGPT和WeLM模型，ChatGPT均优于WeLM；使用CMRC2018和DRCD数据集进行抽取式的机器阅读理解，C~(3)数据集进行常识性的机器阅读理解，发现ERNIE 3.0 Titan在该任务中优于ChatGPT；使用WebQA和CKBQA数据集进行中文闭卷问答测试，发现ChatGPT容易产生事实性错误，与国产模型差距较大。

[结果]ChatGPT在自然语言处理的经典任务上表现较好，在情感分析上具有85%以上的准确率，在闭卷问答上出现事实性错误的概率较高。

[局限]将判别式的任务转为生成式的过程中可能引入评估分数的误差。本文仅在零样本情况下评估ChatGPT，并不清楚其在其他情况下的表现。由于后续版本的迭代更新，ChatGPT评测结果具有一定的时效性。

[结论]ChatGPT很强大但仍然存在一些缺点，研发中国化的大模型应以国家战略需求为导向，并且注意语言模型的风险和局限。（张华平等[[4]](#footnote-4)，）

地质调查正在从“数字化”走向“智能化”，需要在大数据思维的指导下，面向非结构化数据开展机器阅读和地质知识的自动提取。地学命名实体和关系联合提取是当前研究的难点和核心。本文采用基于大规模预训练中文语言模型的BERT—BiLSTM—CRF方法开展岩石描述文本命名实体与关系联合提取。首先，通过收集数字地质填图工作中的剖面测量和路线地质观测数据，建立岩石描述语料；然后，在岩石学理论指导下分析岩石知识组成，完成岩石知识图谱命名实体与关系的模式设计，标注岩石语料；最后，开展岩石描述语料知识提取的深度学习训练和消融试验对比。试验结果显示，大规模预训练中文语言模型(BERT)对岩石描述语料知识提取具有较高的适用性。推荐的BERT—BiLSTM—CRF模型方法对岩石命名实体与关系联合提取的准确率(F1值)为91.75%，对岩石命名实体识别的准确率(F1值)为97.38%。消融试验证明基于BERT的词嵌入层对岩石描述知识提取的性能提升影响显著，双向长短时记忆网络模型层(BiLSTM Layer)能提升实体关系联合提取性能。（陈忠良等[[5]](#footnote-5)，2022）

针对属性权重部分未知且专家权重完全未知的多粒度语言大群体决策问题，提出一种基于云模型的决策方法。首先，构建一种基于信任关系的专家权重求解模型来计算专家权重；其次，将多粒度语言转换为云模型并进行聚类；然后，构建一致性优化模型来求解属性权重，从而得到各个方案的综合评价值并对方案进行排序。所构造的专家赋权模型可以有效解决大群体决策过程中决策人数众多、无法客观给出专家权重信息的问题，而且通过定义的直觉信任函数，还可以对专家之间的信任关系进行刻画，充分挖掘专家之间的信息；将多粒度语言转换为云模型，可以有效刻画语言信息的模糊性和随机性，从而避免信息的丢失和失真。（王佩等[[6]](#footnote-6)，2021）

大数据背景下，数据繁杂、数据的广度和深度达到了一定程度，给企业竞争情报的搜集、处理、分析提出了新的、艰巨的挑战。针对大数据的这些特性，需要革新数据处理的技术以适应大数据时代竞争情报的研究趋势。从数据的视角出发，形成了一套基于大数据的企业竞争情报数据共享交换规范，定义了一套基于大数据的企业竞争情报管理标记语言EIMML，设计和实现了基于大数据的企业竞争情报数据注册中心，最后提出了一种新的分布式企业竞争情报系统架构。（彭颖[[7]](#footnote-7)，2021）

大数据时代，在数据驱动下人文社会科学的研究方法在某些领域已发生改变，由定性研究为主转向定量研究为主，以语言作为研究对象的语言哲学也是如此。定量分析法之所以越来越受到重视，是因为它能够适应大数据和云计算的需求，可以提高精度以及保证处理结果的准确性和可靠性。语言哲学中定量分析法主要有基于模型技术(计算机模拟、数学模型)的方法，和数据驱动(实验哲学、语料库)的方法，这些方法使得语言哲学的研究对象发生转变，而且为人文社科研究提供新范式，并可对相关研究实践进行重构。大数据时代背景下，数据已成为比自然语言更加精准、便利、可操作、可计算的科学语言，语言哲学研究也将由此发生数据化的变革。（陈伟[[8]](#footnote-8)，2021）

文章以计算机编程类C语言进行课程教学，构建基于移动互联网大数据的学生个性化需求与学习分析模型。模型通过发现每个学生的实际需求与学习活动进程，可根据每个学生的实际需求和与学习能力为其量身设计个性多样化方式的学习途径。（甘宇雨等[[9]](#footnote-9)，2021）

本研究为基于新浪微博大数据分析热门综艺《乘风破浪的姐姐》，透过社交媒体分析网络舆论对该节目的关注倾向，探究网民对该综艺的关注热点。利用网络爬虫技术在新浪微博上爬取含有关键词"乘风破浪的姐姐"的所有博文作为研究数据，基于词频统计和LDA主题模型提取主题，从文本特征角度分析网络舆论对该综艺的关注热点。研究发现，网络舆论对该综艺的关注主要集中于"话题人物""女性价值"及"投票打榜"三个方面。（孔明[[10]](#footnote-10)，2021）

<正>近日，Google Brain的研究人员提出了稀疏激活专家模型Switch Transformer。研究人员表示，这个1.6万亿参数模型是迄今为止发布的第一个万亿级人工智能模型，其速度是Google之前开发的最大语言模型(T5-XXL)的4倍，参数规模几乎是1750亿参数的GPT-3的十倍。（[[11]](#footnote-11)，2021）

【目的/意义】为解决融媒体平台海量资源主题提取难、信息推荐精度低问题，笔者利用大数据和NLP技术对现有推荐方法加以优化，改进后的模型在精准度和匹配度方面得到显著提升。【方法/过程】首先利用MapReduce对融媒体资源数据进行过滤处理，再借助NLP技术对资源数据提取主题特征、同时对用户进行需求特征描述，然后将二者进行匹配以确定推荐内容，最后使用对比分析法检验本研究推荐方法的可行性。【结果/结论】与传统方法相比，本研究方法所推荐内容更符合用户需求，其推荐精确度和匹配度也有显著提高。推荐方法评价结果显示，随着推荐列表数量变动，曲线会有一定波动，其波动原因有待深入研究与分析。（崔金栋等[[12]](#footnote-12)，2020）

互联网技术的快速发展为传统广播电视媒介带来了新的机遇和挑战。文章在这一背景下采用了基于内容的协同过滤算法，对电视产品的文本信息分别进行jieba分词、建立了TOPSIS评价模型和余弦相似度模型，求出了爬虫的信息矩阵与每位用户观看节目的综合评价值之间的相似度；然后采用了基于用户的协同过滤算法，以电影产品为例，采用Python软件将55个电影产品的标签与爬虫得到的电影数据的标签进行连接，计算55个销售产品的标签，与用户的标签的相似度，为用户做出推荐，促进电视媒体营销的发展。（沈若男等[[13]](#footnote-13)，2020）

【目的】本文主要就大数据基础理论及系统相关研究背景、技术架构和关键技术展开介绍，并结合技术发展趋势提出未来研究和技术发展方向。【方法】本文在简要介绍大数据处理基础理论的基础上，从面向数据并行的大数据处理技术、RDF (Resource DescriptionFramework)图数据的查询与匹配、大数据分析技术三个方面简要介绍了大数据系统的关键技术。【结果】未来数据产生的速度将进一步提高，在这种应用背景下，如何在设备端进行快速的数据处理成为一种趋势。【结论】未来，我们将在继续关注大数据基础理论与系统关键技术的基础上，引入边缘计算、雾计算等场景，研究物联网环境下的大数据处理。（华强胜等[[14]](#footnote-14)，2019）

自动语音识别(ASR)技术的目的是让机器能够"听懂"人类的语音，将人类语音信息转化为可读的文字信息，是实现人机交互的关键技术，也是长期以来的研究热点。最近几年，随着深度神经网络的应用，加上海量大数据的使用和云计算的普及，语音识别取得了突飞猛进的进展，在多个行业突破了实用化的门槛，越来越多的语音技术产品进入了人们的日常生活，包括苹果的Siri、亚马逊的Alexa、讯飞语音输入法、叮咚智能音箱等都是其中的典型代表。对语音识别技术的发展情况、最近几年的关键突破性技术进行了介绍，并对语音识别技术的发展趋势做了展望。（王海坤等[[15]](#footnote-15)，2018）

由于少数民族语言有其本身的特点，不能简单地套用现有的连续语音识别的方法。本文以蒙古语为例，研讨了声学和语言模型的建立，并在日本国际电气通信基础技术研究所的连续语音识别器上实现了蒙古语的语音识别系统。本文侧重于语言模型的建立，基于蒙古语黏着性语言特点，提出用相似词聚类方法建立多类N-gram模型。实验结果显示，应用我们提出的语言模型，识别精度比用传统的词的N-gram识别法提高了5.5%.（伊·达瓦等[[16]](#footnote-16)，2010）

维吾尔语连续语音识别技术研究主要阐述维吾尔语连续语音的识别技术，主要包括声学模型和语言模型。在声学模型中，主要介绍基于隐马尔可夫模型(Hidden Markov Model，HMM)的维吾尔语连续语音识别声学建模。在语言模型中，主要对比基于文法和基于统计这两种方法的优劣。（冯丽娟等[[17]](#footnote-17)，2010）

大词汇量连续语音识别(LVCSR)技术近年来发展迅速，并在许多领域得到了广泛的应用，国内外许多大公司加大了对语音识别技术的研究，不少商业化的语音识别系统已经面世，并得到较为广泛的使用。该文综述了近年来大词汇量连续语音识别技术的研究进展，描述了汉语大词汇量连续语音识别系统，主要是基于统计方法的语音识别系统的框架与设计方法，对语音识别系统的一些关键技术和原理进行了分析，并对近年来国内外对语音识别研究发展动向进行了讨论。（倪崇嘉等[[18]](#footnote-18)，2009）

本文综述了近年来大词汇量连续语音识别中搜索空间的表示及相关搜索方法的研究进展，分析了搜索空间的表示及相关搜索方法对语音识别性能产生的影响，并对本领域的研究中存在的问题和未来的发展动向进行了讨论。（杨凤芹等[[19]](#footnote-19)，2008）

本文综述了近年来大词汇量连续语音识别技术进步和发展，描述了大词汇量连续汉语语音识别系统的设计方法．对语音识别系统中的一些关键技术和原理进行了详细地分析和讨论，并对语音识别技术进一步发展中存在的问题和近年语音识别研究发展动向进行了讨论．

（刘加[[20]](#footnote-20)，2000）

1. 王静静,叶鹰,王婉茹.GPT类技术应用开启智能信息处理之颠覆性变革[J].图书馆杂志:1-6. [↑](#footnote-ref-1)
2. 赵朝阳,朱贵波,王金桥.ChatGPT给语言大模型带来的启示和多模态大模型新的发展思路[J].数据分析与知识发现:1-12. [↑](#footnote-ref-2)
3. 洪永淼,汪寿阳.人工智能新近发展及其对经济学研究范式的影响[J].中国科学院院刊,2023,38(03):353-357. [↑](#footnote-ref-3)
4. 张华平,李林翰,李春锦.ChatGPT中文性能测评与风险应对[J].数据分析与知识发现:1-14. [↑](#footnote-ref-4)
5. 陈忠良,袁峰,李晓晖,张明明.基于BERT—BiLSTM—CRF模型的中文岩石描述文本命名实体与关系联合提取[J].地质论评,2022,68(02):742-750. [↑](#footnote-ref-5)
6. 王佩,张婧,张威威.基于云模型和多层权重求解的多粒度语言大群体决策方法[J].控制与决策,2021,36(09):2257-2266. [↑](#footnote-ref-6)
7. 彭颖.基于大数据的企业竞争情报系统模型研究[J].西南民族大学学报(自然科学版),2021,47(04):426-431. [↑](#footnote-ref-7)
8. 陈伟.定量分析：大数据背景下语言哲学研究方法论[J].浙江社会科学,2021,(06):110-116+160. [↑](#footnote-ref-8)
9. 甘宇雨,刘勇,张桂芬.基于大数据的C语言个性化学习分析模型的研究及应用[J].无线互联科技,2021,18(10):112-113. [↑](#footnote-ref-9)
10. 孔明.基于新浪微博大数据分析下的《乘风破浪的姐姐》[J].北方传媒研究,2021,(02):44-47. [↑](#footnote-ref-10)
11. .16000亿 谷歌发布人类历史首个万亿级人工智能模型[J].智能建筑与智慧城市,2021,(02):4. [↑](#footnote-ref-11)
12. 崔金栋,陈思远.融媒体信息推荐模型构建与信息推荐方法研究[J].情报科学,2020,38(07):52-58. [↑](#footnote-ref-12)
13. 沈若男,樊璐璐,王文佳,刘鹏飞.大数据时代下电视产品的营销推荐研究[J].生产力研究,2020,(03):129-132. [↑](#footnote-ref-13)
14. 华强胜,郑志高,胡振宇,钟芷漫,林昌富,赵峰,金海,石宣化.大数据基础理论与系统关键技术浅析[J].数据与计算发展前沿,2019,1(05):22-34. [↑](#footnote-ref-14)
15. 王海坤,潘嘉,刘聪.语音识别技术的研究进展与展望[J].电信科学,2018,34(02):1-11. [↑](#footnote-ref-15)
16. 伊·达瓦,匂坂芳典,中村哲.语料资源缺乏的连续语音识别方法的研究[J].自动化学报,2010,36(04):550-557. [↑](#footnote-ref-16)
17. 冯丽娟,吾守尔·斯拉木.维吾尔语连续语音识别技术研究[J].现代计算机(专业版),2010,(01):4-7. [↑](#footnote-ref-17)
18. 倪崇嘉,刘文举,徐波.汉语大词汇量连续语音识别系统研究进展[J].中文信息学报,2009,23(01):112-123+128. [↑](#footnote-ref-18)
19. 杨凤芹,孙吉贵,张长胜,张长海.大词汇量连续语音识别中搜索空间的表示及相关搜索方法的研究进展[J].计算机科学,2008,(02):191-195. [↑](#footnote-ref-19)
20. 刘加.汉语大词汇量连续语音识别系统研究进展[J].电子学报,2000,(01):85-91. [↑](#footnote-ref-20)