智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

检索增强生成推荐及其研究进展

吴国栋, 谢东辰, 黄雯婧, 郑阳, 涂立静

引用本文:
吴国栋, 谢东辰, 黄雯婧, 等. 检索增强生成推荐及其研究进展[J]. 智能系统学报, 2026, 21(3): 577-597.
WU Guodong, XIE Dongchen, HUANG Wenjing, et al. Retrieval-augmented generation for recommendation and research progress[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(3): 577-597.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202508007

DOI: 10.11992/tis.202508007
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20260312.0946.002

检索增强生成推荐及其研究进展

吴国栋, 谢东辰, 黄雯婧, 郑阳, 涂立静
(安徽农业大学 人工智能学院, 安徽 合肥 230036)

摘　要: 检索增强生成(retrieval-augmented generation, RAG)推荐, 作为一种新兴推荐范式, 已引起学术界广泛关注。本文在分析 RAG 推荐及流程基础上, 从基于内容的 RAG 推荐、协同过滤 RAG 推荐、行为序列 RAG 推荐、智能体 RAG 推荐 4 个维度, 深入探讨了现有 RAG 推荐研究的主要进展、技术特点与适用场景。同时, 指出了当前 RAG 推荐研究在检索效率与生成质量的平衡性、多源上下文信息的高效整合、知识库的实时自动更新机制以及用户隐私保护等方面存在的问题。基于上述分析, 从多源多模态信息融合、检索−生成协同优化、动态自适应机制构建以及隐私保护增强等视角, 提出了 RAG 推荐未来的主要研究方向。

关键词: 检索增强生成; 推荐; 大语言模型; 检索; 外部知识; 深度学习; 知识库; 表征学习
中图分类号: TP301　　**文献标志码:** A　　**文章编号:** 1673−4785(2026)03−0577−21

中文引用格式: 吴国栋, 谢东辰, 黄雯婧, 等. 检索增强生成推荐及其研究进展 [J]. 智能系统学报, 2026, 21(3): 577–597.
英文引用格式: WU Guodong, XIE Dongchen, HUANG Wenjing, et al. Retrieval-augmented generation for recommendation and research progress[J]. CAAI transactions on intelligent systems, 2026, 21(3): 577–597.

Retrieval-augmented generation for recommendation and research progress

WU Guodong, XIE Dongchen, HUANG Wenjing, ZHENG Yang, TU Lijing
(School of Artificial Intelligence, Anhui Agricultural University, Hefei 230036, China)

Abstract: Retrieval-augmented generation (RAG) recommendation has emerged as a new recommendation paradigm and has attracted extensive academic attention. Based on an analysis of RAG recommendation and its process, this paper examines the main progress, technical features, and applicable scenarios of existing RAG recommendation research across four dimensions: content-based RAG recommendation, collaborative filtering RAG recommendation, behavioral sequence RAG recommendation, and agent-based RAG recommendation. It also identifies key open problems in current RAG recommendation research, including the trade-off between retrieval efficiency and generation quality, the efficient integration of multi-source contextual information, real-time automatic updating of the knowledge base, and user privacy protection. Based on the above analysis, this paper proposes the main future research directions for RAG recommendation from the perspectives of multi-source and multi-modal information fusion, collaborative optimization of generation and retrieval, construction of dynamic adaptive mechanisms, and enhancement of privacy protection.

Keywords: RAG; recommendation; LLM; retrieval; external knowledge; deep learning; knowledge base; representation learning

随着深度学习的发展与大语言模型兴起, 推荐系统进入了一个新阶段。推荐系统早期主要基于协同过滤[1]方法, 深度学习的引入显著提升了推荐的准确性和效率[2]。近年来, 生成式推荐逐渐成为推荐系统研究的热点。其中, 以检索增强生成 (retrieval-augmented generation, RAG) 为代表的新兴范式, 正受到广泛关注。

传统推荐方法存在数据稀疏、冷启动、可扩展性差等问题, 深度学习推荐系统虽能建模复杂关系, 但依赖大量数据, 模型复杂且缺乏可解释性, 处理动态数据能力有限, 部署成本较高[3]。非 RAG 的纯生成式推荐系统虽然语言生成能力强, 交互自然, 但存在“知识幻觉”、推

收稿日期: 2025−08−06.　**网络出版日期:** 2026−03−12.
基金项目: 国家自然科学基金项目 (32371993); 安徽省高校自然科学研究重点项目 (2024AH050443); 安徽省自然科学基金项目 (2108085MF209); 安徽省科技重大专项项目 (202103b06020013).
通信作者: 吴国栋. E-mail: gdwu1120@qq.com.

荐延迟高、缺乏外部知识整合和可解释性等问题[4-6]。

RAG 推荐将内容信息检索与文本生成相结合，动态引入外部知识，以缓解数据稀疏与冷启动问题，有助于提升推荐准确性、个性化和可解释性，其支持上下文感知和自然语言输出，进而增强用户信任和体验。如 CHAND[7] 提出 RAG 可通过 LLM(large language model) 微调和结合外部数据源来增强推荐系统的个性化推荐能力；Shaped Blog[8] 强调，RAG 通过允许 LLM 访问其训练数据以外的外部信息，使模型能够利用海量数据来提高推荐的准确性；此外，RAG 推荐项目如 recommendation-systems-by-LLMs[9] 和 LLM-recommender-system[10] 也已被开发出来。本文在分析 RAG 推荐及其相关技术基础上，深入探讨 RAG 推荐研究进展。

1　RAG 推荐及其流程

1.1　RAG 推荐

RAG 推荐是一种结合了检索 (Retrieval) 和生成 (Generation) 技术的推荐方法[11]，其通过从大规模数据中检索与用户需求相关的信息，并利用生成模型对这些信息进行整合和优化，从而生成个性化、高质量的推荐内容。RAG 推荐的核心思想是将传统推荐系统中的检索能力与生成式模型的创造力相结合，以提供更精准、多样化和用户友好的推荐结果。

1.2　RAG 推荐流程

RAG 推荐流程通常包括知识库构建与文档准备、向量化存储、用户查询处理、文档检索、上下文增强与推荐生成、优化与反馈等部分[12-13]。其具体推荐流程如图 1 所示。

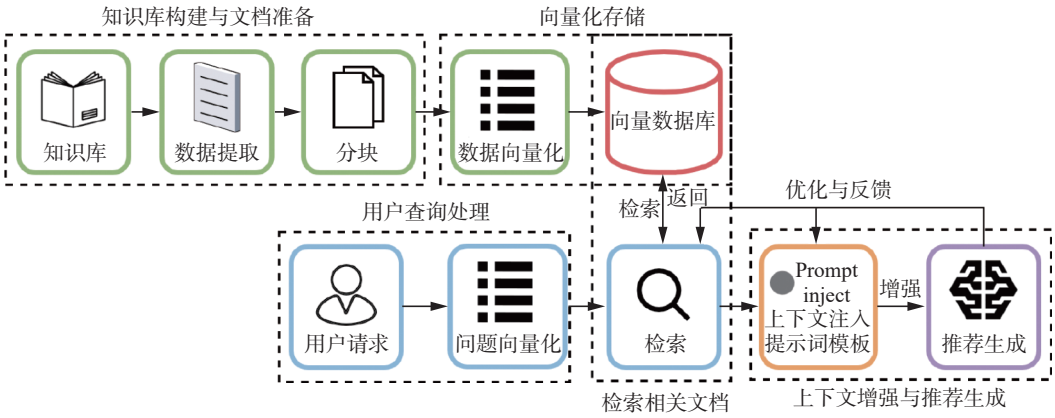

图 1　RAG 推荐系统流程

Fig. 1　RAG recommendation system flowchart

在知识库构建与文档准备阶段，系统从原始知识库中提取与推荐场景相关数据，如用户行为记录、商品信息、内容元数据等。随后，对提取的数据进行预处理，包括文本清洗、去噪、关键词提取等操作，以提升数据质量。接着，将长文本或复杂文档划分为语义完整且粒度适中的小块[14]。

向量化存储阶段使用 BERT(bidirectional encoder representations from Transformers)、OpenAI Embeddings 等预训练嵌入模型，将不同文档片段转换为高维向量表示，实现语义层面的编码。这些向量化的文本片段随后被存储到专用向量数据库如 Chroma、Faiss 中。

用户查询处理阶段，系统接收用户的自然语言输入，并使用嵌入模型将用户输入映射到与知识库一致的高维向量空间。

检索是一个从数据集合中查找并返回与特定查询相关信息的过程。文档检索阶段中，RAG 推荐通过计算用户查询与知识库中文档片段的向量相似度，返回数条最相关的文档片段作为最终检索结果。

上下文增强与推荐生成阶段，系统将检索到的文档片段整合并组织成上下文信息，并注入搭建好的提示词模板中[15]。生成模块结合用户查询和上下文信息，利用大语言模型 (如 GPT、DeepSeek 等预训练语言模型) 生成最终推荐结果。

此外，生成的推荐结果可以通过用户反馈进行优化。例如，系统可以记录用户对推荐结果的点击率、停留时间等行为数据，用于调整检索策略或优化生成模型的参数。

2　RAG 推荐相关技术

2.1　RAG 向量表征学习

2.1.1　嵌入模型的向量表征学习

1)BERT

BERT 是基于 Transformer 架构的深度预训练

语言模型, 其核心思想是在编码阶段同时捕获目标词汇左侧与右侧上下文, 以获得富含多层语义的向量表征[16]。BERT 借助堆叠的自注意力结构, 能够依据不同语境对同一词汇的表征进行实时微调, 从而更精准地刻画语义差异。在文本理解、语义匹配及个性化推荐等任务中, BERT 所生成的嵌入显著提升了上下文感知能力与表达力, 进一步改善了自然语言处理系统的性能[17-18]。

2)OpenAI Embeddings

OpenAI Embeddings 为 OpenAI 提出的嵌入模型, 以 Transformer 编码器为骨干, 旨在将文本或代码片段映射为具有语义区分度的高维向量。具体而言, 输入序列首先经分词器切分为最小语义单元, 并被赋予唯一的整数索引; 随后, 索引序列通过嵌入层转化为稠密向量, 辅以位置编码以保留词元在序列中的相对或绝对次序信息[19]。编码器利用自注意力机制捕捉序列中各位置间的依赖关系与上下文信息。池化层压缩编码器输出为固定长度向量, 形成最终嵌入表示[20]。

3)E5

E5(embeddings from bidirectional encoder representations) 也是一种基于 Transformer 架构的文本嵌入模型[21], 采用双向编码器和自注意力机制捕捉上下文信息, 通过对比学习优化语义向量表示, 拉近相关文本距离、推远无关文本, 从而提升在检索、分类等任务中的表现[22]。E5 采用参数共享的双塔架构, 分别编码查询与文档, 平均池化后通过余弦相似度进行语义匹配[23]。E5 支持多任务微调, 适用于密集、稀疏检索等多种场景, 具备跨语言泛化能力, 在检索增强生成、语义搜索等下游应用中展现出卓越性能。

2.1.2　图神经网络的向量表征学习

1) 图卷积网络的向量表征学习

图卷积网络 (graph convolutional networks, GCN)[24] 生成节点嵌入是对图结构数据进行特征学习的过程, 其输入为图邻接矩阵 $\boldsymbol{A}$ 和节点初始特征矩阵 $\boldsymbol{X} \in \mathbf{R}^{N \times D}$。GCN 通过多层堆叠, 利用加入自环 (self-loop) 的邻接矩阵及其度矩阵, 对邻居特征进行归一化加权聚合, 从而迭代节点嵌入[25-26]。

2)GraphSAGE 的向量表征学习

GraphSAGE(graph sample and aggregate) 是一种归纳式的图表示学习框架, 通过学习可训练的聚合函数, 根据其节点局部邻居信息生成嵌入[27]。GraphSAGE 中, 每个节点首先从其邻居节点中采样一个子集, 然后对这些邻居在上一层的表示进行聚合, 以获得该节点在本层的邻居表示。随后, 将该节点的上一层表示与刚刚得到的邻居表示进行拼接或相加, 经线性变换和非线性激活函数处理后, 生成该节点在当前层的新表示[28]。经过上述步骤的多层传播后, 最终输出节点嵌入。

2.2　RAG 推荐检索

2.2.1　关键词的 RAG 推荐检索

1)TF-IDF

TF-IDF(term frequency-inverse document frequency) 是一种经典的文本特征提取技术, 它通过计算词语在文档中出现频率来决定该词的重要性[29]。TF-IDF 通过结合词频 (TF) 和逆文档频率 (IDF) 两个因素, 综合评估词语的重要程度, 进而筛选出能够更好表示文档内容的关键词[30]。

2)BM25

BM25(best matching 25) 是一种信息检索排序算法, 其核心理念是在词频、文档长度和词语区分度之间实现平衡。BM25 在保留直观解释性的同时, 引入非线性饱和函数和文档长度的归一化处理, 更贴合实际的检索需求。具体而言, BM25 通过对词频增长的非线性调整, 避免了高频词对评分的过度影响。同时, 通过对文档长度归一化处理, 抑制长文档因包含更多词语而获得不公平评分的情况[31]。BM25 通过可调参数灵活控制词频饱和程度及文档归一化强度, 成为现代搜索引擎中常用且效果优异的排序方法之一。

2.2.2　近似最近邻的 RAG 推荐检索

1)BallTree

BallTree[32] 是一种面向高维数据的近邻索引算法, 基于递归构建超球体结构来优化数据空间的划分与检索效率[33]。算法以整个数据集为根节点, 计算其质心与半径来构成超球体; 随后每次选择距离质心最远的两点作为子节点中心来进行划分, 将其余点按距离分组形成两个子簇, 分别计算新质心与半径并递归分割, 直到满足终止条件为止。算法结合深度优先搜索与三角不等式剪枝, 优先探索接近目标点的子树, 并利用距离判断是否跳过兄弟子树, 从而提高效率。BallTree 以超球体代替坐标轴划分, 更适应高维数据的稀疏性与非线性分布, 支持任意距离度量, 尤其适用于数据不均或簇状结构明显的场景。

2)Annoy

Annoy(approximate nearest neighbors oh yeah)[34] 也是一种面向高维数据的近似最近邻搜索算法。它通过构建多棵随机二叉树将数据空间划分为多个子区域, 从而缩小搜索范围并提高查询效率[35]。索引构建阶段, Annoy 随机选择数据点并以其中

垂线为分割依据，递归划分数据集，直到每个子区域中的样本数低于设定阈值；查询阶段，算法会同时遍历多棵树，根据查询向量的位置逐层定位至叶子节点，并使用优先队列机制动态判断是否需拓展搜索路径，以缓解单棵树带来的分割偏差。最终，每棵树返回的候选结果将被合并去重，形成最终候选集合，进而从中选出最相似的近邻点。

3)HNSW

HNSW(hierarchical navigable small worlds)[36]是一种基于图结构的近似最近邻搜索算法，通过构建多层导航图实现高效的相似性检索[37]。算法采用分层结构，其中顶层为稀疏图，用于快速进行全局导航，底层为密集图，包含所有数据点，用于精细的局部搜索[38]。索引构建阶段，每个数据点会依据一定概率被分配到不同的层级，并在各层中与距离较近的节点建立连接，形成具有良好连通性的多层图结构；在查询阶段，搜索从顶层的入口节点开始，采用贪心策略在每一层中不断向距离查询向量更近的节点移动，并逐层向下直至底层，最终获得高质量候选结果。

4)NSG

NSG(navigating spreading-out graph)[39]也是一种基于图结构的近似最近邻搜索算法，旨在实现大规模数据集上高效相似性检索[40]。NSG 以一个预先构建的近似 k 近邻图为基础，通过选择边和修剪策略对其进行优化，生成具有良好可导航性的稀疏图结构，从而在保证连通性的同时，控制内存占用与计算复杂度。在搜索阶段，NSG 采用贪心策略从图中的一个或多个入口点迭代，根据查询向量与当前节点及其邻居的距离，不断向更接近的方向移动，并动态维护候选结果集合，直到满足终止条件为止。

2.2.3　最优子图的 RAG 推荐检索

针对任意查询q和文本图 G，在图G的所有可能子图集合中，总存在一个最具代表性且最能回应查询的子图$\hat{g}$。该子图$\hat{g}$为参数化的语言模型LLM_θ提供了最关键信息，从而引导其生成与预期高度一致的答案[41]。该方法核心是从子图集合$S(G)$中高效检索出这一最优子图$\hat{g}$，并将其内容注入模型参数θ所定义的生成过程中，以此提升最终的回答质量和相关度。

2.3　RAG 的大语言模型推荐生成

1)Decoder-only

Decoder-only 架构的大型语言模型 (LLM) 是当前自然语言处理领域主流范式，其核心设计理念源于 Transformer 解码器，通过自回归生成和因果注意力机制实现文本生成任务[42]。其架构如图 2 所示。

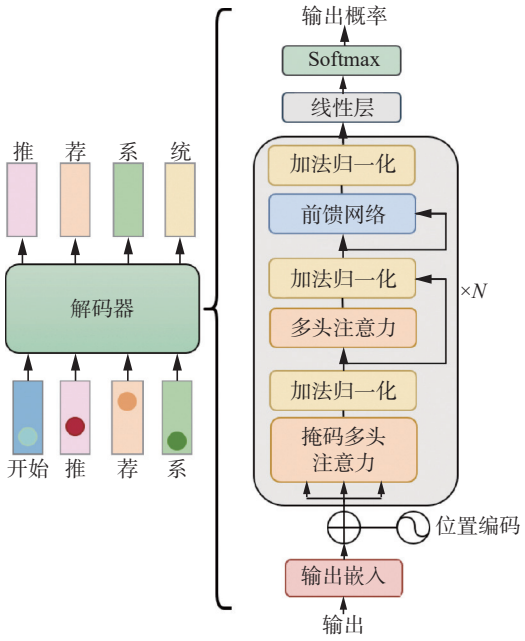

图 2　Decoder-only 大语言模型原理

Fig. 2　Schematic of decoder-only LLMs

这类 LLM 模型摒弃了 Transformer 中编码器模块，仅保留解码器，使得模型在处理输入时能够隐式完成语义理解与内容生成的统一。例如，GPT 系列模型通过单向注意力机制逐步预测下一个词元 (token)，每个词元的生成仅依赖已生成的左侧上下文，这种设计既符合人类语言生成逻辑，也避免了信息泄露问题。

从技术实现来看，Decoder-only 模型通过嵌入层将输入词元映射为向量，并叠加位置编码以保留序列顺序信息。其核心多头自注意力层采用因果掩码 (causal mask) 确保模型在生成第 t 个词元时无法“窥见”未来的词元信息，从而维持自回归生成严格单向性。前馈神经网络 (feed-forward neural network, FFN) 则对每个词元隐藏状态进行非线性变换，增强模型表达能力。模型训练阶段以互联网海量文本为语料，通过最大化下一词元预测概率持续优化十亿级参数，经预训练与微调后，能在对话交互、内容创作、知识问答等多元场景中展现出类人的语言理解与生成能力，成为大语言模型技术落地的核心架构选择[43]。

2)MoE

MoE(mixture of experts) 技术架构的大语言模型 (LLM) 通过稀疏激活机制显著提升了模型计算效率与任务适应性。其核心设计是将 Transformer 层中前馈网络 (FFN) 替换为多个专家子网

络 (Experts), 每个专家通常是独立的 MLP 结构, 通过门控网络 (gating network) 动态选择 Top-K 专家处理输入 token[44]。

训练与推理效率方面, MoE 架构通过稀疏激活实现计算资源的动态分配[45]。例如, DeepSeek-V3 总参数量 6 710 亿, 但每次推理仅激活 5.5% 的参数 (370 亿), 相比传统 Decoder-only 模型大幅降低计算成本[46]。

3　RAG 推荐

3.1　基于内容的 RAG 推荐

3.1.1　基于内容的问答 RAG 推荐

基于内容的问答 RAG 推荐通过解析用户输入的自然语言问题, 理解问答对话中用户的语义意图, 随后利用检索算法从知识库中查找相关内容片段。检索到的内容与用户上下文数据共同构成提示信息, 输入生成模型后输出最终推荐结果[47]。

例如, 医学测试推荐系统 HiRMed[48] 利用 OpenAI 嵌入模型将患者查询和医学知识文本转换为向量, 并使用 GPT-o1 语言模型处理这些向量以生成诊断路径。文献 [49] 提出的法律文档问答系统使用 BM25 和密集向量检索来提高检索的准确性。BM25 用于计算基于关键词频率的相似度得分, 密集向量搜索通过双编码器模型捕捉语义相似性; 然后, 将这些文章块传递给大语言模型, 以理解和提取与用户查询相关的信息并生成答案, 从而支持高效的问答过程。文献 [50] 提出的 Retail-GPT 通过自然对话处理用户查询, 结合 DIET 分类器与 GPT-4o 模型实现产品推荐和购物车操作, 准确理解需求并对接外部数据库获取最新商品信息, 利用 Redis 管理会话历史以增强对话连贯性。文献 [51] 提出的 CHEMRAG-BENCH 整合 PubChem、PubMed 等 6 类数据源, 构建覆盖分子属性与反应机理的专业知识库, 采用 BM25、Contriever 等 5 种检索算法组成的 CHEMRAG-TOOLKIT 及混合检索策略, 并借助 GPT-4 进行语言理解与生成。文献 [52] 提出的 ESGenius 聚焦 ESG(environmental, social, and governance) 领域, 基于 GRI、SASB 等权威框架构建可追溯知识库, 通过问题−文档映射实现语义检索, 设计含“不确定”选项的多选题并经专家双重验证, 确保每问均有源文档支持; 系统使用 LLaMA 3 进行语言处理, 并通过标准符合性检查保障输出严格遵循 ESG 规范。

上述推荐系统中, Retail-GPT 通过结合 RAG 和自然语言处理 (natural language processing, NLP) 技术进行问答式内容推荐, 其系统结构如图 3 所示。

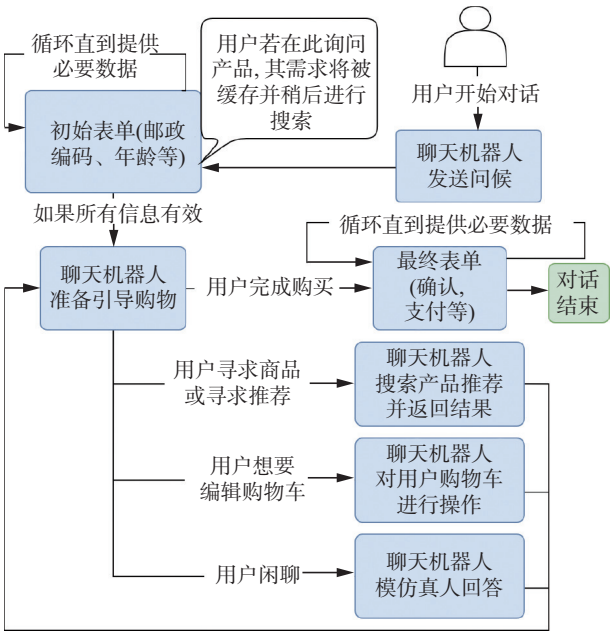

图 3　Retail-GPT 结构示意

Fig. 3　Schematic of the Retail-GPT structure

Retail-GPT 系统以用户开始对话为起点, 聊天机器人首先发送问候语, 并引导用户填写初始表单, 如邮政编码和年龄等基本信息。如果用户在此阶段询问产品, 系统会暂存该请求, 待表单信息完整后再处理。系统会持续提示用户, 直到收集到所有必要信息, 并在验证有效后进入后续流程。此后, 系统根据用户意图选择不同路径: 若用户请求产品推荐或搜索商品, 系统调用模拟外部搜索引擎的大型语言模型返回结果; 若用户需要编辑购物车, 则由系统进行自主编辑; 若用户闲聊, 机器人则以自然语言方式回应。当用户完成选购并确认支付细节后, 系统进入最终表单阶段, 在确认交易信息后结束对话。所有输入均经过安全检查, 防止不当内容及敏感信息泄露。复杂请求可交由大型语言模型子系统处理, 同时系统通过历史交互记录提供上下文记忆, 支持连贯且个性化的对话体验。

3.1.2　基于内容的情境感知 RAG 推荐

基于内容的情境感知 RAG 推荐通过检索用户当前所处情境与需求相关的高质量内容, 并将其作为上下文信息输入生成模型中, 生成更加个性化的推荐结果[53]。这种方法不仅需要考虑用户历史偏好和行为, 还要根据用户当前的地理位置、时间、查询意图等具体情境动态调整推荐内容, 从而提供更贴合用户即时需求的建议。

例如，文献 [54] 引入了基于城市流行度和季节性需求的可持续性指标 (S-Fairness)，用于重新排序检索到的上下文。其构建了包含 160 个欧洲城市旅游信息的知识库，利用 LanceDB 向量数据库和 all- MiniLM-L6-V2 模型检索信息。通过检索组件从训练集获取相关标注对，为 LLM 提供上下文，并将可持续性指标融入其中，引导 LLM 生成符合可持续发展目标的推荐。文献 [55] 从瑞典公共就业服务 API 获取并清洗 IT 职位广告数据，适配 RAG 系统。使用 Azure AI Search 检索相关职位广告，并输入 GPT-3.5 模型。通过系统提示，模型根据用户技能、期望职位和地点，生成个性化技能发展建议。文献 [56] 设计的 RAG 推荐系统，从训练集检索标注对为 LLM 提供上下文。同时引入最大边际相关性 (maximal marginal relevance, MMR) 技术，在保持相关性的同时提升结果多样性，避免语义重叠。随后将相关对作为输入提示，帮助 LLM 生成更准确的相关性标签。

以上基于内容的情境感知推荐中，文献 [57] 构建了一个涵盖景点、地理、历史等多维信息的西藏知识库，采用 TF-IDF 结合 HNSW Flat 索引，并通过 L2 距离优化检索效率。系统利用大语言模型提取用户需求，匹配相关景点，并融合外部知识生成包含地理与历史细节的推荐内容，从而实现情境感知并有效缓解幻觉问题，其系统结构如图 4 所示。

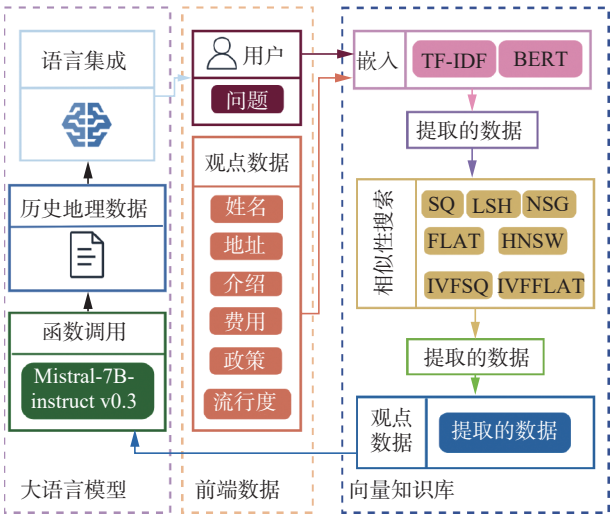

图 4　RAG 旅游推荐系统结构示意

Fig. 4　Schematic of RAG tourism recommendation system structure

首先，研究团队构建了一个详细的视点数据库，其中包含了关于西藏旅游的丰富信息，这些信息被用作外部知识源。接着，为了将文本数据转换为可以被模型理解和处理的向量形式，团队采用了两种向量化方法：TF-IDF 和 BERT。TF-IDF 方法通过计算词频和逆文档频率来衡量单词的重要性，而 BERT 方法则利用深度学习模型提取文本的语义特征。随后，为了提高检索效率，团队使用了 Flat 和 HNSWFlat 等多种索引方法对向量化的数据进行索引。在推荐过程中，当用户提出查询请求时，系统会根据用户的查询内容，通过向量数据库查询机制从外部知识库中检索与用户查询最相关的视点信息。检索到的信息以向量形式表示，这些向量随后被传递到大型语言模型中。LLM 在生成推荐内容时，不仅依赖其内部的知识库，还会结合检索到的外部知识，从而生成与用户需求更匹配、更具情境感知能力的推荐内容。

3.1.3　基于内容的知识图谱 RAG 推荐

基于内容的知识图谱 RAG 推荐是一种结合知识图谱与检索增强生成技术的推荐方法，旨在通过图结构建模和语义检索提升推荐系统的准确性与可解释性[58]。该方法利用图神经网络对知识图谱进行语义索引和信息聚合，构建结构化的知识子图，并将其嵌入大语言模型的语义空间中，从而实现高效、精准的推荐服务。

例如，文献 [59] 提出基于图检索增强生成的推荐方法，结合教育知识图谱 (educational knowledge graphs, EduKG) 和个人知识图谱 (personal knowledge graphs, PKG)，帮助学生理解知识概念。核心包括两部分内容：一是以个人知识图谱为基础的问题生成模块，通过检索个人知识图谱生成个性化问题，引导学生针对未理解的知识概念提问；二是以教育知识图谱为基础的问题回答模块，利用教育知识图谱中知识概念关系回答问题，为学生提供准确答案。文献 [60] 提出的 LlamaRec-LKG-RAG 则构建包含用户、物品及关系的异构知识图谱 (knowledge graph, KG)，并设计用户偏好模块，动态识别用户偏好路径，提取个性化知识子图。系统将用户历史交互、候选物品和知识子图整合到提示模板中，传递给 Llama-2 模型并在单次向前传播中完成最终排序。文献 [61] 提出的 K-RagRec 通过 PLM(pre-trained language models)+GNN(graph neural network) 双重编码，对知识图谱进行多跳子图语义索引并存储。系统依据项目流行度自适应选择检索，并从向量库中召回相似子图。随后，系统结合推荐提示重排序子图，用 GNN 和 MLP 将子图嵌入转换至 LLM 空间，作为软提示辅助推荐生成。K-RagRec 通过检索知识子图进行推荐，其系统结构如图 5 所示。

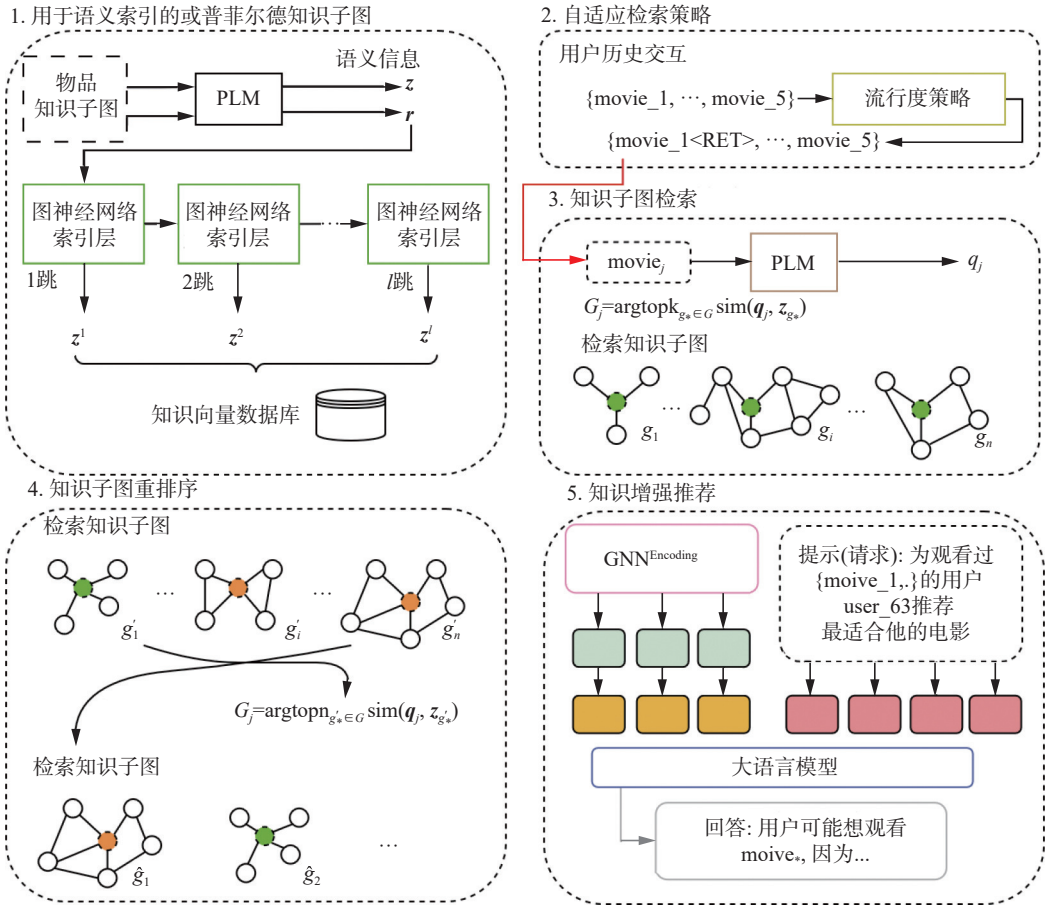

图 5　K-RagRec 结构示意

Fig. 5　Schematic of the K-RagRec structure

K-RagRec 系统首先利用预训练语言模型(pre-trained language model, PLM)捕捉实体n_0和关系e_0的语义信息, 得到:

$$\boldsymbol{z}_{n_0}=\mathrm{PLM}\left(x_{n_0}\right)\in\mathbf{R}^d \tag{1}$$

$$\boldsymbol{r}_{e_0}=\mathrm{PLM}\left(x_{e_0}\right)\in\mathbf{R}^d \tag{2}$$

式中: x_{n_0}和x_{e_0}分别为实体和关系的文本属性, d为输出维度; 再通过图神经网络$\mathrm{GNN}_{\phi_1}^{\mathrm{Indexing}}$聚合邻居信息, 得到实体$n_0$的$l$跳嵌入:

$$\boldsymbol{z}_{n_0}^{l}=\mathrm{GNN}_{\phi_1}^{\mathrm{Indexing}}\left(\left\{\boldsymbol{z}_{n_m}^{(l-1)},\boldsymbol{r}_{e_{<0,m>}}^{(l-1)}:n_m\in N(n_0)\right\}\right) \tag{3}$$

式中: $e_{<0,m>}$为实体n_0和n_m间的边, $\boldsymbol{r}_{e_{<0,m>}}^{(l-1)}$表示该边对应关系在第 l–1 层的嵌入表征, $N(n_0)$为实体 n_0 的邻居集合。由式(3)聚合得到的实体 n_0 的 l 跳表示$\boldsymbol{z}_{n_0}^{l}$用于表征其局部知识子图, 并存储于知识向量数据库中。

接着, 系统依据项目流行度判断是否检索: 若流行度低于阈值p则检索, 反之不检索。对于需检索的项目v_j, 用相同 PLM 将其文本属性转换为语义查询$\boldsymbol{q}_j=\mathrm{PLM}(x_{q_j})\in\mathbf{R}^d$, 从数据库中检索出前 k 个最相似的知识子图:

$$G_j=\mathrm{argtopk}_{g_*\in G}\mathrm{sim}\left(\boldsymbol{q}_j,\boldsymbol{z}_{g_*}\right) \tag{4}$$

式中$\mathrm{sim}(\cdot,\cdot)$为相似度度量。然后, 以推荐提示作为查询, 用 PLM 捕捉其语义信息p, 对检索到的知识子图重排序, 得到前 n 个集合:

$$\hat{G}=\mathrm{argtopn}_{g'_*\in G}\mathrm{sim}\left(\boldsymbol{p},\boldsymbol{z}_{g'_*}\right) \tag{5}$$

最后, 通过图神经网络$\mathrm{GNN}_{\phi_2}^{\mathrm{Encoding}}$编码重排序后的知识子图:

$$\boldsymbol{h}_{\hat{g}_*}=\mathrm{GNN}_{\phi_2}^{\mathrm{Encoding}}\left(\left\{\hat{g}_*:\hat{g}_*\in\hat{G}\right\}\right) \tag{6}$$

再经 MLP 投影器映射到 LLM 嵌入空间:

$$\hat{\boldsymbol{h}}_{\hat{G}}=\mathrm{MLP}_\theta\left(\left[\boldsymbol{h}_{\hat{g}_1};\boldsymbol{h}_{\hat{g}_2};\cdots;\boldsymbol{h}_{\hat{g}_N}\right]\right) \tag{7}$$

其中 $[\cdot;\cdot]$ 为连接操作。将其作为软提示附加在 LLM 的输入标记嵌入前, 生成推荐结果。

3.2　协同过滤 RAG 推荐

3.2.1　协同过滤表征学习的 RAG 推荐

协同过滤表征学习的 RAG 推荐通过对比学习等表示学习技术, 将物品的文本、图像、音频等多模态内容特征与用户历史行为进行联合建模, 生成融合内容语义和用户偏好的高质量物品嵌入向量[62]。这些嵌入随后被组织成结构化可检索知识库, 作为后续推荐的基础数据源。检索阶段, 系统基于用户查询或上下文生成查询嵌入, 快速召回语义相关且符合群体偏好的物品。这些检索

到的物品信息作为增强的外部知识被整合到生成模型的输入上下文中，使大语言模型能够结合协同过滤的群体智慧和多模态内容理解，输出个性化、可解释且与用户兴趣深度匹配的推荐结果。

如 RALLRec[63] 使用 LLM 生成物品详细描述，提取更优文本表征，并将其与简略物品表征拼接。同时，系统利用推荐模型获取物品的协同语义，通过自监督学习使协同语义与文本语义对齐，得到最终表征并用于检索，再结合语义相似度和时间戳重排序器，提升推荐效果。文献 [64] 提出的 RETURN 框架将外部数据库中的用户交互序列转换为多跳协同物品图，编码物品间的共现频率和时间间隔等协同信号，随后基于该图检索物品对共现频率以计算各物品出现概率，定位潜在扰动，再采用删除 (针对出现概率为 0 的物品) 或替换 (用共现频率高的物品替换低概率物品) 策略净化用户画像，最后通过多次随机净化并结合投票机制的稳健集成推荐策略生成最终推荐。文献 [65] 通过表示学习生成物品丰富语义表示，整合文本语义与协同信号，借助近似最近邻 (approximate nearest neighbor search, ANN) 算法高效召回相关物品，作为外部知识注入 RAG 框架，辅助生成模型提升推荐准确性与合理性。文献 [66] 通过构建多模态嵌入，将物品多模态内容特征与用户评分、互动历史等协同过滤数据融入嵌入学习，经对比学习优化，使嵌入兼具多模态信息与协同行为网络特征；检索阶段利用 ANN 在融合协同信息的嵌入空间高效查找相似物品，为 RAG 提供兼顾内容特征与群体偏好的检索结果，支持融合协同过滤的 RAG 推荐与检索。文献 [67] 提出 GCN-Retriever 机制，构建用户−项目二分图，通过 GCN 聚合邻域特征生成含多阶交互信息的嵌入，结合余弦相似度检索相似用户交互数据融入 LLM 输入，加速 RAG 过程；搭配多头部早期退出策略平衡效率与精度，助力协同过滤相关的 RAG 推荐。

RALLRec 通过用户−物品协同过滤方法进行表征学习。其系统结构如图 6 所示。

RALLRec 先通过传统推荐模型从用户−物品交互记录提取物品协同语义，其表达式为

$$\{\boldsymbol{e}^i_{\text{colla}}\}_{i=1}^{n} = \text{RecModel}(\{(u,i)\in\mathcal{V}\}) \tag{8}$$

式中：n是物品总数，$\mathcal{V}$是交互历史。接着用自监督学习对齐文本与协同语义，以两层 MLP 为投影器映射文本嵌入至低维空间，通过特定训练目标优化，得到物品对齐嵌入：

$$\boldsymbol{e}^i_{\text{ssl}} = \text{MLP}\left(\boldsymbol{e}^i_{\text{text}}\right) \tag{9}$$

之后将文本嵌入、协同嵌入和自监督学习所得嵌入经幅度归一化后拼接，形成最终物品嵌入：

$$\boldsymbol{e}_{\text{item}} = [\bar{\boldsymbol{e}}_{\text{text}}\|\bar{\boldsymbol{e}}_{\text{colla}}\|\bar{\boldsymbol{e}}_{\text{ssl}}] \tag{10}$$

式中：$\bar{\boldsymbol{e}}:=\boldsymbol{e}/|\boldsymbol{e}|$，通过点积比较检索与目标物品最相关的物品。检索阶段学习用户和物品联合表征以获取相关物品，经重排序器与最近物品融合后纳入提示，提示可用于推理或通过指令微调训练模型；生成阶段由基础 LLM 响应提示。提示构建采用对应模板，填充用户档案和行为历史，结合少量数据进行指令微调与基于相似度的检索，并按时间戳重排序列增强数据。同时，设计重排序器，为物品分配通道分数S_{c}和位置分数S_{pos}，总得分为

$$S^i_{\text{sum}} = S^i_{\text{c}} * S^i_{\text{pos}} \tag{11}$$

选取得分高的物品纳入提示并生成推荐。

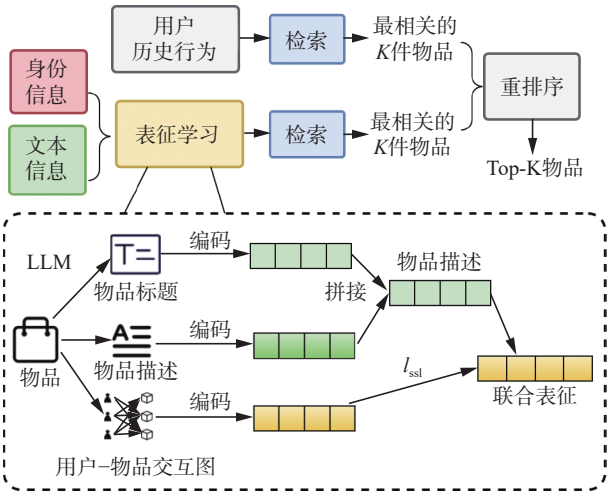

图 6　RALLRec 结构示意

Fig. 6　Schematic of RALLRec structure

3.2.2　协同过滤检索的 RAG 推荐

协同过滤检索的 RAG 推荐通过结合用户协同行为与高效检索技术，构建基于相似用户或物品的增强上下文，以优化大语言模型的推荐生成[68]。系统先利用协同信号从用户历史行为中挖掘潜在偏好模式，再检索出与目标用户最相关的协同信息 (如相似用户的评分、交互数据或社会信号)。最后，将这些检索到的协同知识以文本提示、嵌入增强或重排序等方式整合到 LLM 的输入中，使模型能够结合群体行为数据与语义理解能力，生成个性化的推荐结果。

如 CFRAG[69] 通过对比学习训练用户嵌入，以检索相似用户并引入协同信息；在此基础上，设计融合用户偏好的个性化检索器与重排序器，并利用大语言模型 (LLM) 反馈进行微调，从而获取支持个性化生成的相关文档。文献 [70] 基于余弦相似度检索相似用户，提取评分数据并设计 4 种

结构化提示策略, 将协同信号注入 LLM, 融合局部用户模式与全局统计信息生成推荐, 在效果与提示长度之间取得平衡。文献 [71] 通过混合检索整合协同过滤信息, 同时挖掘用户行为模式与物品特征; 将检索结果转化为结构化文本提示, 并在联邦学习环境下利用 LLM 对候选集进行重排序。借助 LLM 的预训练知识, 有效缓解数据稀疏性问题, 从而生成高质量的个性化推荐。文献 [72] 先将论文内容和用户兴趣映射到语义嵌入空间, 检索阶段以上下文为查询, 用 ANN 算法召回语义接近的“社会信号”和“相关项目建议”, 作为 LLM 生成推荐的依据, 确保推荐准确且具有社会语境和集体偏好。

上述提到的 CFRAG 通过协同过滤增强检索提高生成的个性化能力。CFRAG 系统结构如图 7 所示。

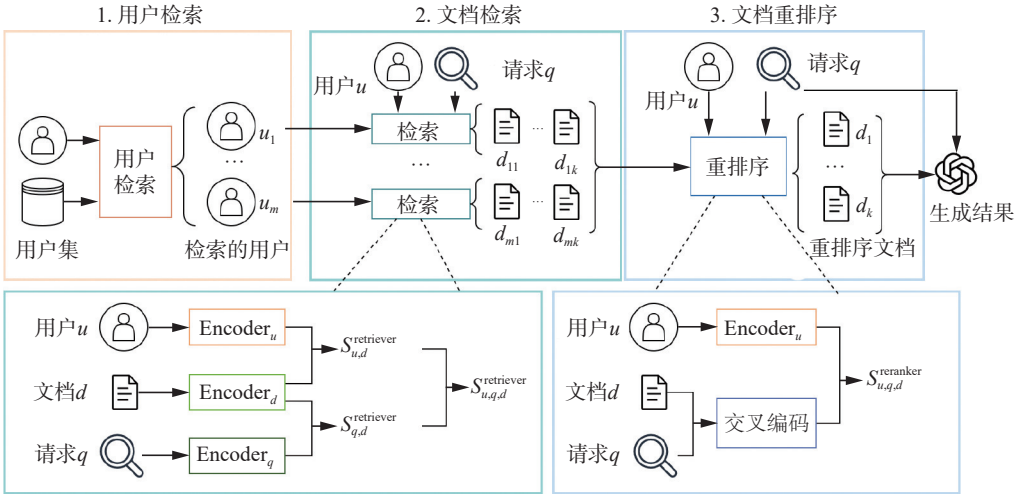

图 7　CFRAG 结构示意

Fig. 7　Schematic of CFRAG structure

实验将对比学习机制融入协同过滤系统, 增强用户行为相似度的计算效果: 对用户历史文档序列进行数据增强 (如裁剪、掩码、重排序), 生成不同视图作为正样本, 其他用户历史作为负样本, 通过 InfoNCE 损失训练用户嵌入: 训练后的用户嵌入通过余弦相似度检索 Top-m 相似用户, 其历史文档池扩展了候选文档来源。在文档检索阶段, CFRAG 设计的个性化评分函数融合语义相关性与用户偏好为

$$S^{\text{retriever}}_{u,q,d} = (1-\alpha) S^{\text{retriever}}_{q,d} + \alpha \cos\left(\text{MLP}_1(\boldsymbol{e}_u), \text{Encoder}_d(d)\right) \tag{12}$$

式中: α 为权重系数, MLP_1 将用户嵌入映射到文档嵌入空间。为优化检索效果, 系统利用 LLM 生成反馈微调检索器: 通过比较生成输出与目标输出的 ROUGE 分数构建分布 $p_{\text{LLM}}(d \mid q, y)$, 最小化其与检索器分布 $p_{\text{retriever}}(d \mid q, u)$ 的 KL 散度:

$$l_{\text{retriever}} = \text{KL}\left(p_{\text{retriever}} \parallel p_{\text{LLM}}\right) \tag{13}$$

文档重排序进一步引入用户偏好, 通过交叉编码器生成查询−文档联合表示 $\boldsymbol{h}_{q,d}$, 与用户嵌入拼接后计算得分:

$$S^{\text{reranker}}_{u,q,d} = \text{MLP}_3\left(\text{CONCAT}\left(\boldsymbol{h}_{q,d}, \text{MLP}_2(\boldsymbol{e}_u)\right)\right) \tag{14}$$

同样基于 LLM 反馈微调重排序器, 使文档排序分布逼近生成质量分布。最终, 重排序后的 Top-k 文档与查询拼接输入 LLM 生成个性化结果。

3.2.3　协同过滤的冷启动 RAG 推荐

冷启动场景下, 基于内容的 RAG 推荐系统通过检索增强生成技术实现个性化推荐。其核心流程首先对用户查询进行语义理解和扩展, 将自然语言输入转化为结构化需求表示。随后通过嵌入技术实现向量化检索, 从知识库中获取相关内容。最后结合预设的提示模板和检索结果, 利用生成模型输出个性化推荐。该流程通过检索与生成的协同优化, 在缺乏用户历史数据的情况下, 仍能基于领域知识和实时查询实现高质量的个性化推荐, 有效解决了冷启动阶段的推荐难题[73]。

如 RAMO(retrieval-augmented generation for enhancing MOOCs recommendations)[74] 系统通过加入积极副词的提示模板进行引导, 将课程数据转为嵌入存储, 检索器依据用户查询获取信息, 生成器基于提示生成推荐, 无需用户历史数据, 弥补了协同过滤的不足。文献 [75] 提出的 RAGSys 通过计算查询与样本的嵌入余弦相似度, 并结合 MMR 多样性算法与质量偏差项, 采用贪婪策略检索样本。方法借鉴协同过滤思想, 有效缓解冷启动问题, 从而提升 LLM 在少样本条件下的推理性能。文献 [76] 提出的课程推荐系统先让 LLM 生成理想课程描述并转为向量, 通过嵌入相似度检索相似课程再生成推荐, 类似协同过滤中基于

内容相似性推荐，桥接语义差距解决冷启动。

上述 RAMO 模型利用协同过滤以缓解推荐系统冷启动问题。RAMO 系统结构如图 8 所示。

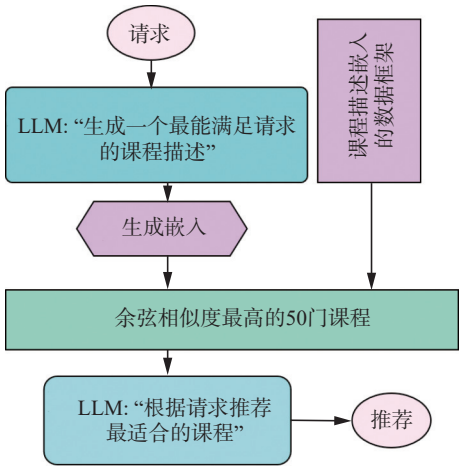

图 8　RAMO 结构

Fig. 8　RAMO structure

RAMO 推荐系统把“冷启动”拆解为“零历史、零画像、零反馈”的三零场景，并在两阶段流程中逐层注入先验与交互信号，实现从无到有、由粗到细的推荐跃迁。首先，在检索阶段，系统不再依赖传统协同过滤所需的交互矩阵，而是利用大语言模型的泛化能力，将学生一句“我想入门数据科学”直接扩展为包含学习目标、先备技能、时间预算、难度偏好的结构化画像。该画像被嵌入向量空间后，与全部课程描述做语义对齐，即使学生从未在平台留下任何点击或评分，也能通过文本语义召回首批高相关候选。其次，在生成阶段，GPT-4o 对 50 门候选课程进行重排序时，通过计算匹配度、主动生成“若你每周只能投入三小时，优先选择带自动评测的微课而非项目制长课”这类情境化解释，把潜在担忧提前暴露给学生，同时输出置信度，让学生对“推荐是否靠谱”有量化感知。系统进一步引入“追问式对话”机制：学生可继续追问“我没有编程基础，这些课还适合吗?”等问题，模型会实时缩减列表并补充“零基础友好”标签。整个流程无需历史行为，仅通过自然语言交互即可持续细化需求，把冷启动转化为“边对话、边画像、边推荐”的动态闭环，从而彻底摆脱对历史数据的依赖。

3.3　基于行为序列的 RAG 推荐

基于行为序列的 RAG 推荐通过动态整合用户历史交互序列与实时检索的外部知识，实现时序感知的个性化推荐生成。该方法首先基于用户行为序列 (如点击流、学习轨迹或移动路径) 构建时序上下文，并利用知识追踪、动态记忆库或多模态检索技术，从物品库、知识图谱或历史记录中实时检索与当前序列状态最相关的辅助信息。随后，系统检索这些时序相关的信息，将其作为增强上下文与用户当前行为序列共同输入生成模型。最终，大语言模型综合时序行为模式与检索到的动态知识，生成下一个推荐项[77]。

如 RaSeRec[78] 通过协作预训练和检索增强微调进行基于行为序列的 RAG 推荐。预训练阶段学习序列推荐和检索能力，基于用户历史交互序列训练模型预测下一项及检索相似序列。微调阶段构建包含<用户序列，目标项>对的记忆库，对于当前输入的用户序列，系统检索相似记忆，通过 RAM 模块将检索到的序列信息与当前用户表示融合，生成下一项推荐。文献 [79] 提出的 Qilin 通过整合用户在会话中的查询来源、历史请求等上下文信号，结合多模态内容特征，支持基于 RAG 的序列推荐。对于触发 DQA(deep query answering) 模块的搜索请求，系统会记录用户偏好的答案及参考结果，这些信息作为序列中的关键节点，为后续推荐提供依据。文献 [80] 提出的 PathGPT 先将历史轨迹的边缘 ID 序列通过反向地理编码转化为包含道路名称等的自然语言描述。在生成推荐序列时，系统以用户的起点、终点和约束条件为查询，检索向量数据库中语义相似的历史路径文本作为上下文，LLM 基于这些上下文生成路径序列以进行推荐。文献 [81] 提出 RAG 推荐先利用检索器从历史交互序列等数据中获取候选项目集，再结合用户的序列行为特征，由编码器−解码器 LLM 对候选集进行重排序。对于会话式序列推荐，系统会检索用户过往对话中的偏好信息作为上下文，使推荐序列能依据对话序列的发展动态调整，贴合用户实时需求。文献 [82] 提出的 TutorLLM，KT(knowledge tracing) 模块追踪学生学习序列以预测学习状态，抓取器收集课程内容构建知识库。RAG 模块依据学生学习序列阶段，检索相关资源生成推荐序列，并随学习序列推进更新，针对薄弱环节持续推荐。

如上述 TutorLLM 通过抓取器模型、知识追踪模块和 LLM 模块 3 个核心组件来进行基于行为序列的 RAG 推荐。TutorLLM 系统结构如图 9 所示。

抓取器模型通过 Jina AI 的 Reader API 实现动态抓取在线课程视频字幕、网页文本等文本内容，以此构建背景知识库。知识追踪模块采用 MLFBK(multi-features with latent relations BERT knowledge tracing) 模型，利用学生历史交互数据预测学习状态。该模型结合学生、技能与题目

ID 等基础特征, 并引入技能掌握 (基于 BKT)、能力轮廓 (基于 K-means 聚类的历史表现) 和题目难度 (按正确率映射为 1~10 等级) 等认知嵌入, 以刻画学习过程中的潜在状态。输入嵌入通过拼接多特征和潜在关系嵌入生成最终表示。随后, BERT 架构的编码器通过多头注意力机制处理序列数据, 输出学生下一步动作的预测结果。LLM 模块基于 GPT-4 API, 结合知识库和 KT 模块的实时输入生成个性化响应。当学生发起查询时, 系统首先根据 KT 预测的学习状态从抓取器构建的知识库中检索相关内容, 例如通过余弦相似度匹配相关文本片段。检索到的信息与学生的上下文结合后, 输入 LLM 生成定制化答案和学习建议。最终, 系统根据学生的实时反馈和测试表现持续优化推荐策略, 形成闭环学习支持。

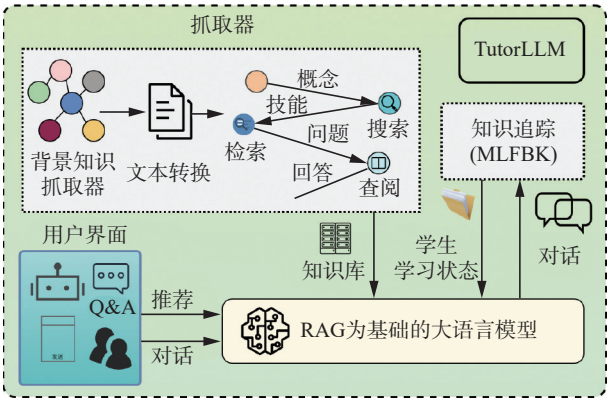

图 9　TutorLLM 结构示意

Fig. 9　Schematic diagram of TutorLLM structure

3.4　智能体 RAG 推荐

智能体 RAG 推荐是一种结合智能体 (Agent) 技术与检索增强生成的推荐方法, 旨在通过多步骤推理、动态决策和知识检索提升推荐系统的智能化与个性化。该方法利用智能体对用户输入进行理解与分类, 并根据不同的任务需求自动规划检索策略, 从多种类型的知识源中获取相关信息。随后, 系统将检索到的结果整合并注入生成过程中, 辅助大语言模型输出更准确、符合上下文的推荐结果[83]。

如 HM-RAG[84] 包含分解智能体、多源检索智能体和决策智能体。分解智能体拆分复杂查询, 多源检索智能体并行从向量、图、网络数据库获取信息, 决策智能体通过一致性投票和专家优化整合结果, 提升问答和分类准确性。文献 [85] 提出的多智能体系统先分类用户查询, 诊断智能体分阶段收集症状并融合假设生成疾病预测, RAG 则基于知识库提供检测、治疗等信息支持。智能体通过自适应提问细化症状, RAG 确保建议有兽医知识依据, 二者协作提升诊断效率。文献 [86] 提出的 Agentic-RAG 中智能体选标题、检索 JSON、生成建议并检查; Graph-RAG 智能体转文本为图结构, 查询图社区生成推荐, 均依据 NCCN(National Comprehensive Cancer Network) 指南。智能体分工处理检索与生成, RAG 保障治疗方案贴合最新临床标准, 减少误差。文献 [87] 提出的 MemInsight 中智能体挖掘实体或对话属性标注记忆, 再通过属性匹配或嵌入搜索检索信息, 增强 RAG 推荐的相关性和说服力。智能体通过自主优化其记忆结构, 为 RAG 提供高质量的结构化信息, 从而实现对用户需求的精准匹配; 该机制在推荐任务中显著提升了推荐性能。

HM-RAG 模型通过分层多智能体架构在 RAG 中实现了高效的多模态协同与动态决策。其核心创新是将复杂的查询处理流程分解为 3 个专业化的智能体模块: 分解智能体 (decomposition agent)、多源检索智能体 (multi-source retrieval agent) 和决策智能体 (decision agent)。HM-RAG 系统结构如图 10 所示。

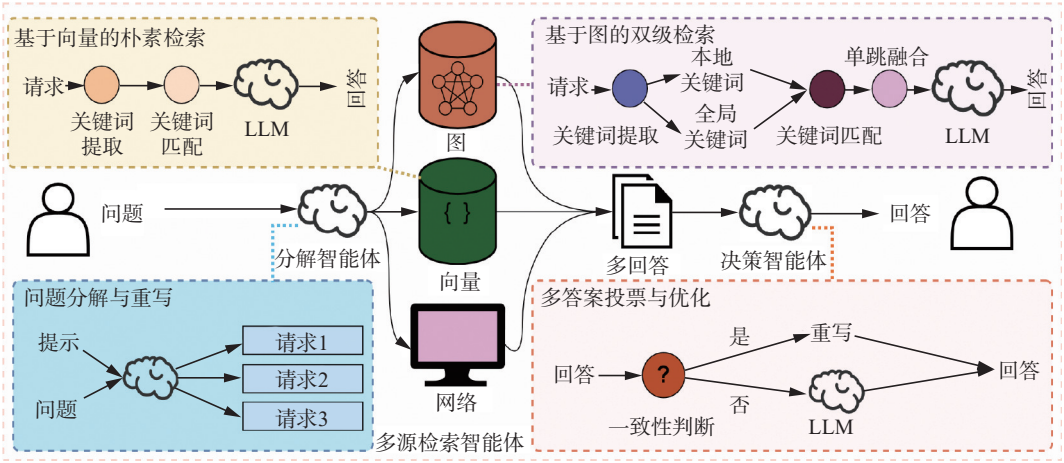

图 10　HM-RAG 结构示意

Fig. 10　Schematic of HM-RAG structure

分解智能体首先利用 LLM(如 Qwen2.5-7B) 对复杂查询进行语义解析，通过结构化提示指令将多意图问题拆分为逻辑连贯的子任务集合 $q=\{q_1,q_2,\cdots,q_n\}$，保留原始关键词并确保语义连贯性。这一步骤通过隐式的提示工程实现，无需显式公式，但依赖于 LLM 的上下文理解能力。多源检索智能体由 3 个并行模块组成，分别处理不同模态的数据。向量检索智能体通过文本嵌入方法：

$$\boldsymbol{h}_q=\mathrm{E}_{\text{text}}(q) \tag{15}$$

将查询转换为向量表示，并计算余弦相似度：

$$s_j=\frac{\boldsymbol{h}_q^{\mathrm{T}}\boldsymbol{h}_j}{\|\boldsymbol{h}_q\|\|\boldsymbol{h}_j\|} \tag{16}$$

从非结构化文本中检索 Top-k 文档R_k。图谱检索智能体则基于多模态知识图谱 (multi-modal knowledge graph, MMKG)，通过 LightRAG 动态提取与查询相关的子图：

$$G_q=\left\{(h,r,t)\mid\mathrm{LightRAG}_{\text{graph}}(q,h,r,t)>\tau\right\} \tag{17}$$

并扩展 1 跳邻居H_g以增强关系推理。网络检索智能体通过 Google Serper API 获取实时数据，补充静态知识库的不足。决策智能体通过一致性投票机制整合多源答案。使用 ROUGE-L 和 BLEU 指标评估答案间的语义一致性，若冲突则调用专家模型 (如 GPT-4) 进行精炼。ROUGE-L 基于最长公共子序列 (longest common subsequence, LCS) 衡量关键信息重叠，BLEU 则检测 n-gram 匹配精度，两者加权融合后生成最终答案A。

3.5　RAG 推荐主要应用领域

当前，RAG 推荐在电子商务、旅游休闲、专业服务和教育领域中展现出独特技术价值。

在电子商务领域，RAG 通过引入商品属性、用户评论及品牌背景等外部语义信息，有效缓解了冷启动所导致的推荐性能下降问题。相较于仅依赖协同信号的传统方法，RAG 能够生成语义连贯、内容丰富且具备解释性的推荐结果，显著提升了推荐系统的泛化能力与用户体验。

在旅游休闲领域，用户需求高度依赖时空上下文，具有强动态性与个性化特征。RAG 通过实时检索用户游记、景点详情、开放状态等多源异构信息外部知识库，实现对当前情境的精准感知与响应，从而生成时效性强、贴合实际场景的推荐内容，克服了静态推荐的局限。

在医疗、法律、金融、环境社会治理以及人力资源等专业服务领域，任务对输出结果的事实准确性、合规性及可追溯性要求极高。RAG 通过限定检索范围于权威知识源，如法规条文、临床指南或行业披露标准，确保生成内容有据可依、逻辑严谨，有效抑制幻觉现象，在高风险决策支持中展现出显著优势。

在教育领域，RAG 的核心价值在于支撑知识驱动的个性化学习。通过检索结构化的教育知识体系，包括课程标准、知识点依赖关系以及常见认知误区，系统能够动态生成契合学习者当前认知水平的问题、解释和学习路径，从而提升教学的适应性与有效性。

由此可见，RAG 在各领域推荐中有其独特的贡献：在电子商务中增强语义理解与冷启动鲁棒性，在旅游休闲中实现高灵敏度的情境适配，在专业服务中保障事实一致性与合规可信度，在教育中促进细粒度认知建模与个性化干预。这些差异化能力充分表明了 RAG 在各推荐领域中的可行性与有效性。

4　RAG 推荐数据集、评价指标与性能分析

4.1　RAG 推荐数据集与评价指标

RAG 推荐数据集涵盖多种数据类型，包括纯文本语料、图数据，以及融合文本、图像、音频等元素的多模态数据。表 1 总结了 RAG 推荐研究中部分常用数据集。

表 1　RAG 推荐常用数据集

Table 1　Commonly used datasets for RAG

数据集类型	数据集名称	数据集简介	相关链接
文本	MovieLens	数据集包含用户对电影的评分及时间戳、用户ID和电影ID等元信息，提供多个规模版本，广泛用于协同过滤与矩阵分解算法的评估。	https://grouplens.org/datasets/movielens
	WANDS	一个面向事实型问答的数据集，包含问题、答案及其来源文档，主要用于评测开放域问答系统在检索与生成环节的准确性。	https://github.com/wayfair/WANDS
	Amazon Review	数据集聚合多个商品类别的用户评论、评分及相关元数据，常用于情感分析、用户行为建模和跨域学习研究。	https://nijianmo.github.io/amazon
	TruthfulQA	用于评估语言模型生成内容真实性的数据集，包含817个对抗性构造的问题，重点测试模型避免生成常见误解或虚假陈述的能力。	https://github.com/sylinrl/TruthfulQA
	ESCI	亚马逊发布的电商搜索相关性数据集，包含用户查询、商品及其人工标注的相关性标签，适用于搜索排序与语义匹配任务。	https://github.com/amazon-science/esci-data

续表 1

数据集类型	数据集名称	数据集简介	相关链接
知识图谱	Cora	数据集包含2 700余篇机器学习领域的学术论文, 每篇以词袋形式表示并标注类别, 论文间的引用关系构成图结构, 是图神经网络中经典的节点分类基准数据集。	https://linqs-data.soe.ucsc.edu/public/lbc/cora.tgz
	Citeseer	数据集收录了3 300余篇计算机科学领域的学术论文, 提供词袋特征、类别标签及引用链接, 常用于引文网络中的图学习与半监督分类研究。	https://github.com/ZPowerZ/citeseer-dataset
多模态	MIND	数据集包含新闻标题、正文等文本内容和配图图像, 记录用户点击行为, 面向新闻推荐任务。	https://msnews.github.io
	OpenMIC-Rec	数据集含视频画面、音频信号和语音转录文本3种模态, 基于短视频平台用户交互构建, 支持跨模态推荐研究。	https://github.com/cosmir/openmic-rec
	Kwai	数据集含视频帧序列、音频流和文本描述3种模态, 结合真实用户行为日志, 适用于大规模多模态内容理解与推荐。	https://github.com/kwai-recsys/Kwai-VideoRec-Dataset
	MM-IMDb	数据集含海报图像、预告片视频、剧情摘要和用户评论文本3种模态, 面向电影语义理解与多模态推荐任务。	https://www.cs.utexas.edu/~rofuyu/mm-imdb

推荐系统的评估指标一般围绕内容相关性、排序质量、预测准确性及正负样本区分能力 4 方面展开, 分别有 RECALL、Hit Rate、BLEU(bilingual evaluation understudy); MAP(mean average precision)、NDCG(normalized discounted cumulative gain)、MRR (mean reciprocal rank); RMSE(root mean square error)、MAE(mean absolute error); AUC(area under the ROC curve)、ROC(receiver operating characteristic) 等。此外, RAG 还强调推荐在知识层面的可靠性、多模态信息间的协同一致性以及对实时上下文的感知能力, 如 Source Authority、Cross-modal Consistency、Context Relevance 等。这些指标共同构成了从传统到知识增强型推荐系统的多层次评估体系。

4.2　现有 RAG 推荐方法性能分析

现有 RAG 推荐研究围绕基于内容、协同过滤、用户行为序列以及智能体 4 大核心方向展开, 形成了丰富多样的技术路径与应用场景。基于内容的 RAG 推荐从问答交互、情境感知和知识图谱 3 个维度展开, 通过解析用户意图、结合动态情境及利用结构化知识, 提升推荐的精确性与可解释性; 协同过滤 RAG 推荐则融合表征学习、检索机制和冷启动解决方案, 借助群体行为数据增强推荐的个性化与泛化能力; 基于行为序列的 RAG 推荐聚焦于用户时序行为变化, 动态整合历史交互序列与实时检索知识, 实现时序感知的推荐生成; 智能体 RAG 推荐通过多智能体协作完成任务分解、多源检索与决策整合, 进一步提升了推荐系统的智能化水平与自适应能力。这些研究虽在各自领域展现出独特优势, 但也面临着检索质量依赖、数据融合难度大、系统复杂性高等共性挑战, 为后续研究指明了优化方向。

表 2 总结了现有 RAG 推荐的研究内容、主要优点和不足。

表 3 从准确率 (ACC)、F_1 分数 (F_1 Score)、召回率 (Recall)、BLEU 分数 (BLEU)、归一化折损累积增益 (NDCG) 以及曲线下面积 (AUC)6 个维度对上述 RAG 推荐进行系统性量化分析。其中, 指标表示该类 RAG 推荐方法在多篇文献中的平均性能趋势, 指标数值以平均值±标准差的统计格式呈现, 代表了不同推荐方法在特定评估指标上的性能中心趋势和稳健性差异。

表 2　现有 RAG 推荐研究小结

Table 2　Summary of existing RAG research studies

研究内容		主要文献	优点	不足
基于内容的RAG推荐	基于内容的问答式RAG推荐	[48−52]	①多阶段检索优化提升回答质量 ②闭环机制增强自适应能力	①多语言处理复杂 ②检索质量直接影响生成效果
	基于内容的情境感知RAG推荐	[54−57]	①结合实时情境动态调整推荐 ②推荐结果更符合实际	①多维情境融合难 ②依赖标注数据和专家支持
	基于内容的知识图谱RAG推荐	[59−61]	①结构化知识提升可解释性 ②图神经网络增强语义推理	①构建维护成本高 ②图嵌入与LLM对齐难度大

续表 2

研究内容	主要文献	优点	不足
协同过滤RAG推荐	协同过滤表征学习的RAG推荐	[63−67]	①推荐的语义性和个性化程度提高 ②推荐系统的泛化能力增强	①依赖大量标注数据和复杂模型训练 ②多模态信息融合与用户行为建模技术门槛高
	协同过滤检索的RAG推荐	[69−72]	①群体行为增强推荐相关性 ②个性化与可解释性强	①协同数据稀疏影响效果 ②大规模协同检索时隐私保护难度大
	协同过滤冷启动RAG推荐	[74−76]	①无用户历史也能个性化推荐 ②提示模板提升可解释性	①过度依赖检索准确性和生成模型能力 ②生成模型可能产生幻觉
基于行为序列的RAG推荐		[78−82]	①行为序列实时检索知识，推荐个性化与准确性高 ②动态上下文建模，增强推荐系统的泛化能力与解释性	①长序列建模较为复杂 ②实时检索对响应速度和系统性能要求较高
智能体RAG推荐		[84−87]	①任务分解与多源检索协同，提升推荐的智能化与适应性 ②多步骤推理和动态决策，增强推荐可解释性	①智能体间协调机制复杂，系统设计与部署成本较高 ②对LLM理解能力与检索质量依赖性强，影响整体推荐稳定性

表 3　现有 RAG 推荐性能分析

Table 3　Quantitative analysis of existing RAG recommendation performance

研究内容		主要文献	评价指标					
			ACC	F_1	Recall	BLEU	NDCG	AUC
基于内容的RAG推荐	基于内容的问答式RAG推荐	[48−52]	—	—	0.88±0.12	0.64±0.14	0.82±0.13	—
	基于内容的情境感知RAG推荐	[54−57]	0.53±0.16	0.40±0.21	0.52±0.18	—	0.45±0.22	0.80±0.01
	基于内容的知识图谱RAG推荐	[59−61]	—	—	0.81±0.31	—	0.70±0.15	0.87±0.21
协同过滤RAG推荐	协同过滤表征学习的RAG推荐	[63−67]	0.67±0.15	—	0.61±0.17	0.57±0.21	0.55±0.20	0.84±0.11
	协同过滤检索的RAG推荐	[69−72]	0.64±0.18	0.60±0.22	0.58±0.20	0.55±0.25	0.50±0.24	—
	协同过滤冷启动RAG推荐	[74−76]	—	—	0.53±0.22	—	0.45±0.25	0.79±0.19
基于行为序列的RAG推荐		[78−82]	0.74±0.12	0.70±0.16	0.67±0.14	0.60±0.18	—	—
智能体RAG推荐		[84−87]	—	—	0.73±0.15	—	0.75±0.16	—

从表 3 可以看出，当前 RAG 推荐主要方法中，基于知识图谱的方案凭借强大语义推理能力，在 AUC 和召回率上领先，充分体现了结构化知识对提升推荐质量的关键作用；协同过滤结合表征学习的方法 AUC 表现稳健，凸显其在用户−物品交互建模上的优势；而行为序列类 RAG 推荐则以较高的准确率和 F_1 分数脱颖而出，具备突出的刻画用户动态兴趣能力；与此同时，在召回能力方面，问答式方法位居前列，智能体驱动的 RAG 也在排序质量上展现独特潜力。综合来看，融合知识图谱、行为序列、协同信号、智能体等多维 RAG 架构，不仅在关键指标上实现基本覆盖，更代表了下一代智能推荐系统向语义理解、动态适应与自主推理深度融合的发展方向。

5　现有 RAG 推荐研究存在的主要问题

5.1　检索与生成平衡问题

检索效率与生成质量的平衡问题仍是难题之一，高效的检索算法虽然能快速返回结果，但可能牺牲一定的准确性；复杂的检索算法虽能提高准确性，却会导致系统响应速度下降。例如模型 RAG-Sequence 在检索更多文档时，在开放域问答任务中的表现逐步提高；然而，对于 RAG-Token 模型，检索 10 个以上文档虽可以提高 Recall 和

ROUGE-L 得分, 但可能会导致 BLEU-1 得分下降[11]。又例如 CaseGPT 在处理复杂的医疗诊断或法律案例时, 系统可能以牺牲深度分析为代价, 换取高效、快速的响应时间。正因如此, 系统在面对需要细致推理的场景时, 这种快速生成的结果可能不够丰富和全面[88]。

5.2 上下文内容融合问题

RAG 推荐系统在上下文融合方面存在的困难表现在: 即使检索到相关文档, 系统在将多个信息块整合成连贯回答时仍可能出现问题, 导致生成的内容缺乏逻辑性或遗漏关键信息[89]。这往往由于大语言模型在处理多文档上下文时可能无法有效提取并整合信息。多篇文献中研究人员通过对不同数据集和模型进行研究, 发现即使在上下文足够的情况下, 模型仍然会产生幻觉, 因此给出错误答案。又例如, 在处理 HotPotQA 等多跳问题的数据集时, 模型在有足够上下文时也不能保证回答的正确性, 说明了其在整合多信息块, 形成连贯回答方面仍存在不足[90]。

5.3 推荐结果的实时性问题

现有 RAG 推荐系统普遍依赖于事先构建的静态知识库, 因此难以在与用户交互过程中主动获取并学习新信息。特别是在面对新闻等快速更迭的领域时, 系统可能会因缺乏及时更新的能力, 生成过时的答案或无法反映最新的情况[91]。在当下, RAG 推荐系统中的知识库仍需要被动地定期更新才能保证生成的答案和推荐的及时性和新颖性。例如, 在电商和内容推荐中, RAG 推荐系统面对的是持续涌现的新品上架、价格调整、内容更新等动态信息。如果数据库没有被及时更新, 系统生成的推荐或回答可能会过时, 从而影响推荐的相关性、准确性以及推荐系统的整体价值[92]。

5.4 推荐过程中隐私保护问题

RAG 推荐系统在处理敏感信息或个人隐私时面临多重挑战。数据检索阶段, 系统可能需要频繁访问外部知识库中潜在的用户敏感信息。推荐内容生成阶段, 推荐系统又有可能在不经意间根据检索的结果泄露其中的隐私。在系统架构方面, 如果使用了第三方云服务或者依赖外部 API, 用户敏感数据则会在不同服务提供商之间传输、存储, 增加了隐私泄露的风险[93]。此外, 用户数据可能会被大语言模型服务商用于模型的二次训练或其他未明确告知的用途。最后, 系统运行过程中产生的日志数据、交互记录等也可能在用户未授权的情况下被长期留存和分析, 进一步加剧了隐私泄露的风险。

6 RAG 推荐未来主要研究方向

6.1 多源多模态信息的 RAG 推荐

多源多模态推荐通过深度整合多元异构数据与跨模态内容, 来构建更加智能化的推荐体系。系统将全面融合知识图谱结构化知识、社交媒体用户生成内容、物联网设备的传感器数据以及传统数据库的历史记录等多源信息, 同时整合文本、图像、视频、语音等多模态内容, 形成全方位的数据感知能力[94]。多级处理架构下, 系统实现从原始数据采集、跨模态特征对齐到智能推理决策的完整流程。技术方面, 系统可利用基于深度注意力机制的多源可信度动态评估模型, 实现对不同数据源的自动质量评价和权重分配[95]; 开发时空上下文感知引擎, 捕捉用户行为轨迹与环境特征; 并引入增量式持续学习算法, 使系统能够实时适应用户偏好的变化[96]。

6.2 生成与检索协同优化的 RAG 推荐

检索与生成的平衡可通过动态权重调整算法实现。使用基于查询复杂度的启发式规则或机器学习模型动态分配检索和生成的权重。对于结构化或明确性查询, 提高检索模块的权重以优先获取精确结果; 而对于开放性或模糊查询, 则增加生成模块的权重以生成更灵活的推荐。此外, 使用模型融合技术来进一步提升平衡效果。例如, 使用集成学习方法将多个检索模型和生成模型的结果进行融合, 引入强化学习框架, 通过奖励机制动态调整检索与生成的协作方式。最后, 实时获取用户反馈数据, 利用在线学习算法持续优化系统参数, 确保检索与生成的平衡能够适应用户需求的变化[97]。

6.3 动态自适应 RAG 推荐

动态自适应 RAG 推荐研究旨在构建更智能、灵活及个性化的推荐系统。它突破了当下 RAG 模型对静态检索和生成策略的依赖, 通过实时感知用户兴趣漂移、交互反馈及外部知识流, 在线调整检索与生成双重策略: 一方面, 依据查询复杂度与数据时效性动态切换知识源、调节召回粒度; 另一方面, 基于检索置信度和用户偏好实时改写生成风格与内容侧重[98]。同时, 系统内置持续学习框架, 利用在线梯度更新与记忆回放机制, 在极短时间内完成参数微调与策略进化, 进而实现“随用随学、按需适配”的个性化推荐闭环。

6.4 隐私保护增强的 RAG 推荐

用户隐私保护与数据利用之间的矛盾始终是

推荐系统面临的核心挑战之一。在未来的工作中，应大力推动隐私保护技术的应用，例如联邦学习、差分隐私、可信执行环境、零知识证明等，确保在收集和使用用户数据的同时又能充分保护用户隐私。除此之外，建立更透明的数据管理机制也至关重要，使用户有途径了解个人数据的收集、存储和使用的方式，并赋予用户更多的控制权，从而进一步增强用户对推荐系统的信任[99]。此外，引入区块链技术，通过其去中心化、不可篡改和透明的特性，可以为数据交换提供一个更安全的环境，确保用户数据的所有权和使用情况可追溯[100]。

6.5　记忆管理和任务规划的智能体 RAG 推荐

该方向通过智能体融合技术来进行不同功能角色之间的高效协作，实现动态记忆管理和任务规划。统一框架内的所有智能体共享上下文与决策状态，形成协同感知与响应机制，提升系统的灵活性与响应能力。动态记忆管理通过分层存储短期交互状态与长期用户行为，支持记忆的动态更新、评估与跨智能体检索，构建起可共享的集体记忆体系，确保系统在多轮交互中保持上下文连贯[101]。在此之上，任务规划模块将复杂的推荐目标细化为可执行的步骤，协调各智能体按需分工、接力推进，并根据实时反馈动态调整策略。

7　结束语

本文系统综述了 RAG 推荐这一新兴范式，其通过结合信息检索与文本生成，动态引入外部知识，有效缓解数据稀疏、冷启动、大语言模型幻觉等问题。文章阐述了 RAG 推荐的 6 步流程：知识库构建、向量化存储、查询处理、文档检索、上下文增强生成及优化反馈；分析了向量表征、检索方法和大语言模型生成架构等关键技术；从基于内容、协同过滤、行为序列推荐和智能体推荐等方面深入探讨了现有 RAG 推荐研究的主要进展；指出其在检索−生成平衡、上下文整合、实时性与隐私保护等方面存在的挑战；最后从多源/多模态融合、协同优化、动态自适应与隐私增强等方面展望了 RAG 推荐未来主要研究方向。RAG 推荐凭借“检索−生成”协同机制，为个性化推荐领域带来了新的活力，随着技术不断发展，有望在更多领域实现广泛应用并创造更大价值。

参考文献：

[1] PAPADAKIS H, PAPAGRIGORIOU A, PANAGIOTAKIS C, et al. Collaborative filtering recommender systems taxonomy[J]. Knowledge and information systems, 2022, 64(1): 35−74.

[2] STECK H, BALTRUNAS L, ELAHI E, et al. Deep learning for recommender systems: a Netflix case study [J]. AI magazine, 2021, 42(3): 7−18.

[3] ZHANG Yongfeng, CHEN Xu. Explainable recommendation: a survey and new perspectives[J]. Foundations and trends® in information retrieval, 2020, 14(1): 1−101.

[4] JOHNSON S, HYLAND-WOOD D. A primer on large language models and their limitations[EB/OL]. (2024−12−03) [2025−07−30]. https://arxiv.org/abs/2412.04503.

[5] LI L, ZHANG Y, LIU D, et al. Large language models for generative recommendation: a survey and visionary discussions[EB/OL]. (2023−09−03)[2025−07−30]. https://ar-xiv.org/abs/2309.01157.

[6] 吴国栋, 秦辉, 胡全兴, 等. 大语言模型及其个性化推荐研究[J]. 智能系统学报, 2024, 19(6): 1351−1365.
WU Guodong, QIN Hui, HU Quanxing, et al. Research on large language models and personalized recommendation[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1351−1365.

[7] CHAND S . Leveraging LLM fine-tuning and RAG for advanced recommendation engines[EB/OL]. (2024−05−29) [2025−07−30]. https://medium.com/@shireenchand/leveraging-llm-fine-tuing-and-rag-for-advanced-recommendation-engines-a3d683e39976.

[8] SKOVORODNIKOV H. RAG for RecSys: a magic formula. [EB/OL]. (2023−10−16) [2025−07−30]. https://www.shaped.ai/blog/rag-for-recsys-a-magic-formula.

[9] FATTAHI T. recommendation-systems-by-LLMs [EB/OL]. (2024−01−01)[2025−07−30]. https://github.com/taherfattahi/recommendation-systems-by-llms.

[10] ANDIKA J, AKBARZADEH N. LLM-recommender-system [EB/OL]. (2024−01−01)[2025−07−30]. https://github.com/jand-odoo/llm-recommender-system.

[11] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [EB/OL]. (2020−05−22)[2025−07−30]. https://arxiv.org/abs/2005.11401.

[12] MEI Lang, MO Siyu, YANG Zhihan, et al. A survey of multimodal retrieval-augmented generation[EB/OL]. (2025−03−26) [2025−07−30]. https://arxiv.org/abs/2504.08748.

[13] XU Chenhao, GAO Longxiang, MIAO Yuan, et al. Distributed retrieval-augmented generation[EB/OL]. (2025−05−01) [2025−07−30]. https://arxiv.org/pdf/2505.00443v1.

[14] 吴璇, 付涛. 检索增强生成技术研究综述[J]. 计算机工

程与应用, 2025, 61(20): 19−35.
WU Xuan, FU Tao. Comprehensive review of retrieval-augmented generation[J]. Computer engineering and applications, 2025, 61(20): 19−35.

[15] 田永林, 王雨桐, 王兴霞, 等. 从 RAG 到 SAGE: 现状与展望[J]. 自动化学报, 2025, 51(6): 1145−1169.
TIAN Yonglin, WANG Yutong, WANG Xingxia, et al. From retrieval-augmented generation to SAGE: the state of the art and prospects[J]. Acta automatica sinica, 2025, 51(6): 1145−1169.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017−06−12)[2025−07−30]. https://arxiv.org/abs/1706.03762.

[17] SUYUNU B, TAYLAN E, ÖZGÜR A. Linguistic laws meet protein sequences: a comparative analysis of subword tokenization methods[C]//2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Piscataway: IEEE, 2025: 4489−4496.

[18] 罗玲, 李硕凯, 何清, 等. 基于知识图谱、TF-IDF 和 BERT 模型的冬奥知识问答系统[J]. 智能系统学报, 2021, 16(4): 819−826.
LUO Ling, LI Shuokai, HE Qing, et al. Winter Olympic Q & A system based on knowledge map, TF-IDF and BERT model[J]. CAAI transactions on intelligent systems, 2021, 16(4): 819−826.

[19] XIAN J, TEOFILI T, PRADEEP R, et al. Vector search with OpenAI embeddings: lucene is all you need[C]// Proceedings of the 17th ACM International Conference on Web Search and Data Mining. New York: ACM, 2024: 1090−1093.

[20] KORADE N B, SALUNKE M B, BHOSLE A A, et al. Strengthening sentence similarity identification through OpenAI embeddings and deep learning[J]. International journal of advanced computer science and applications, 2024, 15(4): 821−829.

[21] WANG Liang, YANG Nan, HUANG Xiaolong, et al. Multilingual E5 text embeddings: a technical report [EB/OL]. (2024−02−08) [2025−07−30]. https://arxiv.org/abs/2402.05672.

[22] JIANG Ting, SONG Minghui, ZHANG Zihan, et al. E5-V: universal embeddings with multimodal large language models[EB/OL]. (2024−07−17) [2025−07−30]. https://arxiv.org/abs/2407.12580.

[23] WANG Liang, YANG Nan, HUANG Xiaolong, et al. Text embeddings by weakly-supervised contrastive pre-training[EB/OL]. (2024−02−08) [2025−07−30]. https://arxiv.org/abs/2402.05672.

[24] 王伟, 徐鑫, 杨景超. 基于图神经网络的个性化推荐系统研究[J]. 九江学院学报 (自然科学版), 2025, 40(2): 39−43, 51.
WANG Wei, XU Xin, YANG Jingchao. Personalized recommendation system based on graph neural network [J]. Journal of Jiujiang University (natural science edition), 2025, 40(2): 39−43, 51.

[25] HE Xiangnan, DENG Kuan, WANG Xiang, et al. LightGCN: simplifying and powering graph convolution network for recommendation[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020: 639−648.

[26] BERREZIGA R, BRAHIMI M, KRAIM K, et al. Combining GCN structural learning with LLM chemical knowledge for enhanced virtual screening[EB/OL]. (2025−04−24)[2025−07−30]. https://arxiv.org/abs/2504.17497.

[27] JAHIN M A, SHAHRIAR S, MRIDHA M F, et al. Soybean disease detection via interpretable hybrid CNN-GNN: integrating MobileNetV2 and GraphSAGE with cross-modal attention[EB/OL]. (2025−03−03)[2025−07−30]. https://arxiv.org/abs/2503.01284.

[28] 陈靖耀, 李敬华, 于彤. 基于图神经网络增强句嵌入的中医文献多标签分类方法研究[J]. 世界科学技术−中医药现代化, 2025, 27(2): 420−430.
CHEN Jingyao, LI Jinghua, YU Tong. A study on multi-label classification methods for traditional Chinese medicine literature based on sentence embedding enhanced by graph neural networks[J]. World science and technology-modernization of traditional Chinese medicine, 2025, 27(2): 420−430.

[29] CAI Jingwen, LECKNER S, BJÖRKLUND J. From precision to perception: user-centred evaluation of keyword extraction algorithms for Internet-scale contextual advertising[EB/OL]. (2025−04−30)[2025−07−30]. https://arxiv.org/abs/2504.21667.

[30] 白云天, 郝文宁, 靳大尉. 基于检索增强生成的开放域问答方法研究[J]. 计算机科学, 2025, 52(S1): 36−42.
BAI Yuntian, HAO Wenning, JIN Dawei. Study on open-domain question answering methods based on retrieval-augmented generation[J]. Computer science, 2025, 52(S1): 36−42.

[31] CHEN X, WISEMAN S. BM25 Query Augmentation Learned End-to-End[EB/OL]. (2023−05−23)[2025−07−30]. https://ar-xiv.org/abs/2305.14087.

[32] SLATTON G. Ball-tree: A ball-tree implementation for K-NN [EB/OL]. (2026−02−01) [2026−03−06]. https://github.com/ grantslatton/ball-tree.

[33] DOLATSHAH M, HADIAN A, MINAEI-BIDGOLI B. Ball*-tree: Efficient spatial indexing for constrained nearest-neighbor search in metric spaces[EB/OL]. (2015−11−02) [2025−07−30]. https://arxiv.org/abs/1511.00628.

[34] Spotify. Annoy: Approximate Nearest Neighbors in C++/Python optimized for memory usage and loading/saving to disk[EB/OL]. (2025−10−06) [2026−03−06]. https://github.com/spotify/annoy.

[35] SINGH P, AHUJA K. Chameleon2++: an efficient and scalable variant of chameleon clustering[EB/OL]. (2025−01−05) [2025−07−30]. https://arxiv.org/abs/2501.02612.

[36] MALKOV Y A. Hnswlib: Header-only C++/Python library for fast approximate nearest neighbors [EB/OL]. (2025−10−18) [2026−03−06]. https://github.com/nmslib/hnswlib.

[37] RAINES N, SAMARTH M, LAX K, et al. Semantic Vector Search using an HNSW Index for Twitter Data [EB/OL]. (2023−05−08)[2025−07−30]. https://journals.orclever.com/oprd/article/view/516.

[38] MALKOV Y A, YASHUNIN D A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 42(4): 824−836.

[39] ZJULearning. NSG: Fast approximate nearest neighbor search with navigating spreading-out graph [EB/OL]. (2024−05−24) [2026−03−06]. https://github.com/ZJU-Learning/nsg.

[40] FU C, XIANG C, WANG C, et al. Fast Approximate Nearest Neighbor Search With The Navigating Spreading-out Graph[EB/OL]. (2017−07−01) [2025−07−30]. https://arxiv.org/abs/1707.00143.

[41] HU Yuntong, LEI Zhihan, ZHANG Zheng, et al. GRAG: graph retrieval-augmented generation[EB/OL]. (2024−05−26) [2025−07−30]. https://arxiv.org/abs/2405.16506.

[42] TAY Y, DEHGHANI M, TRAN V Q, et al. UL2: unifying language learning paradigms[EB/OL]. (2024−05−26) [2025−07−30]. https://arxiv.org/abs/2405.16506.

[43] 李博, 莫先. 大语言模型在推荐系统中的应用[J]. 计算机科学, 2025, 52(S1): 19−25.
LI Bo, MO Xian. Application of large language models in recommendation system[J]. Computer science, 2025, 52(S1): 19−25.

[44] DAI Damai, DENG Chengqi, ZHAO Chenggang, et al. DeepSeekMoE: towards ultimate expert specialization in mixture-of-experts language models[EB/OL]. (2024−01−11)[2025−07−30]. https://arxiv.org/abs/2401.06066.

[45] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity[EB/OL]. (2021−01−11)[2025−07−30]. https://arxiv.org/abs/2101.03961.

[46] DEEPSEEK-AI, LIU Aixin, FENG Bei, et al. DeepSeek-V3 technical report[EB/OL]. (2024−12−27)[2025−07−30].19437. https://arxiv.org/abs/2412.19437.

[47] 王宇哲. 基于内容的电影推荐算法研究[J]. 信息系统工程, 2023(12): 117−120.
WANG Yuzhe. Research on content-based movie recommendation algorithm[J]. China CIO news, 2023(12): 117−120.

[48] YANG Yahe, HUANG Chengyue. Tree-based RAG-agent recommendation system: a case study in medical test data[EB/OL]. (2025−01−06)[2025−07−30]. https://arxiv.org/abs/2501.02727.

[49] BA T N, THE V D, QUANG T P, et al. Vietnamese legal information retrieval in question-answering system [EB/OL]. (2025−01−06)[2025−07−30]. https://arxiv.org/abs/2409.13699.

[50] DE FREITAS B A T, DE ALENCAR LOTUFO R. Retail-GPT: leveraging retrieval augmented generation (RAG) for building E-commerce chat assistants[EB/OL]. (2024−08−15) [2025−07−30]. https://arxiv.org/abs/2408.08925.

[51] ZHONG Xianrui, JIN Bowen, OUYANG Siru, et al. Benchmarking retrieval-augmented generation for chemistry[EB/OL]. (2024−05−12) [2025−07−30]. https://arxiv.org/abs/2505.07671.

[52] HE Chaoyue, ZHOU Xin, WU Yi, et al. ESGenius: benchmarking LLMs on environmental, social, and governance (ESG) and sustainability knowledge[EB/OL]. (2025−01−02) [2025−07−30]. https://arxiv.org/abs/2506.01646.

[53] 王鹏哲. 基于情境感知的农业知识智能推荐模型研究[D]. 南宁: 广西大学, 2024: 1−90.
WANG Pengzhe. Research on intelligent recommendation model of agricultural knowledge based on situational awareness[D]. Nanning: Guangxi University, 2024: 1−90.

[54] BANERJEE A, SATISH A, WÖRNDL W. Enhancing tourism recommender systems for sustainable city trips using retrieval-augmented generation[C]//Recommender Systems for Sustainability and Social Good. Cham: Springer, 2025: 19−34.

[55] ANDERSSON M, ENQVIST T. Bridging the skills gap: applying an LLM and RAG architecture for recommend-

ing competence development in the IT JobMarket[EB/OL]. (2024−07−09) [2025−07−30]. https://www.diva-portal.org/smash/record.jsf pid=diva2%3A1883256&dswid=4999.

[56] SACHDEV J, D ROSARIO S, PHATAK A, et al. Automated query-product relevance labeling using large language models for E-commerce search[C]//Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval. New York: ACM, 2024: 32−40.

[57] QI Jinhu, YAN Shuai, ZHANG Yibo, et al. RAG-optimized Tibetan tourism LLMs: enhancing accuracy and personalization[C]//Proceedings of the 2024 7th International Conference on Artificial Intelligence and Pattern Recognition. New York: ACM, 2024: 1185−1192.

[58] 王光, 姜皓. 融合多视图对比学习和知识图谱的推荐算法[J]. 计算机系统应用, 2025, 34(6): 118−127.
WANG Guang, JIANG Hao. Recommendation algorithm incorporating multi-view contrastive learning and knowledge graph[J]. Computer systems & applications, 2025, 34(6): 118−127.

[59] ABDELMAGIED M, CHATTI M A, JOARDER S, et al. Leveraging graph retrieval-augmented generation to support learners' understanding of knowledge concepts in MOOCs[EB/OL]. (2025−04−15) [2025−07−30]. https://arxiv.org/abs/2505.10074.

[60] AZIZI V, KOOCHAKI F. LlamaRec-LKG-RAG: a single-pass, learnable knowledge graph-RAG framework for LLM-based ranking[EB/OL]. (2025−05−15) [2025−07−30]. https://arxiv.org/abs/2506.07449.

[61] WANG Shijie, FAN Wenqi, FENG Yue, et al. Knowledge graph retrieval-augmented generation for LLM-based recommendation[EB/OL]. (2025−01−04) [2025−07−30]. https://arxiv.org/abs/2501.02226.

[62] 马赫, 王海荣, 周北京, 等. 基于表示学习的实体对齐方法综述[J]. 计算机工程与科学, 2023, 45(3): 554−564.
MA He, WANG Hairong, ZHOU Beijing, et al. Overview of the entity alignment methods based representation learning[J]. Computer engineering & science, 2023, 45(3): 554−564.

[63] XU Jian, LUO Sichun, CHEN Xiangyu, et al. RALLRec: improving retrieval augmented large language model recommendation with representation learning[C]//Companion Proceedings of the ACM on Web Conference 2025. New York: ACM, 2025: 1436−1440.

[64] NING Liangbo, FAN Wenqi, LI Qing. Retrieval-augmented purifier for robust LLM-empowered recommendation[EB/OL]. (2025−04−03)[2025−07−30]. https://arxiv.org/abs/2504.02458.

[65] LUO Sichun, XU Jian, ZHANG Xiaojie, et al. RALLRec+: retrieval augmented large language model recommendation with reasoning[EB/OL]. (2025−04−03) [2025−07−30]. https://arxiv.org/abs/2503.20430.

[66] ORAMAS S, FERRARO A, SARASUA A, et al. Talking to your recs: Multimodal embeddings for recommendation and retrieval [EB/OL]. (2024−10−31) [2025−07−30]. https://ceur-ws.org/ Vol-3787.

[67] ZHOU Huixue, GU Hengrui, LIU Xi, et al. The efficiency vs. accuracy trade-off: optimizing RAG-enhanced LLM recommender systems using multi-head early exit[EB/OL]. (2025−01−04) [2025−07−30]. https://arxiv.org/abs/2501.02173.

[68] 李孟浩, 赵学健, 余云峰, 等. 推荐算法研究进展[J]. 小型微型计算机系统, 2022, 43(3): 544−554.
LI Menghao, ZHAO Xuejian, YU Yunfeng, et al. Survey on research progress of recommendation algorithms [J]. Journal of Chinese computer systems, 2022, 43(3): 544−554.

[69] SHI Teng, XU Jun, ZHANG Xiao, et al. Retrieval augmented generation with collaborative filtering for personalized text generation[EB/OL]. (2025−04−08) [2025−07−30]. https://arxiv.org/abs/2504.05731.

[70] POURYOUSEF S. What LLMs miss in recommendations: bridging the gap with retrieval-augmented collaborative signals[EB/OL]. (2025−05−27) [2025−07−30]. https://arxiv.org/abs/2505.20730.

[71] ZENG Huimin, YUE Zhenrui, JIANG Qian, et al. Federated recommendation via hybrid retrieval augmented generation[C]//2024 IEEE International Conference on Big Data (BigData). Piscataway: IEEE, 2025: 8078−8087.

[72] WANG Ruotong, ZHOU Xinyi, QIU Lin, et al. Social-RAG: retrieving from group interactions to socially ground AI generation[EB/OL]. (2024−11−04) [2025−07−30]. https://arxiv.org/abs/2411.02353.

[73] 毛骞, 谢维成, 乔逸天, 等. 推荐系统冷启动问题解决方法研究综述[J]. 计算机科学与探索, 2024, 18(5): 1197−1210.
MAO Qian, XIE Weicheng, QIAO Yitian, et al. Survey on solving cold start problem in recommendation systems[J]. Journal of frontiers of computer science & technology, 2024, 18(5): 1197−1210.

[74] RAO Jiarui, LIN Jionghao. RAMO: retrieval-augmented generation for enhancing MOOCs recommendations[EB/OL]. (2025−07−06) [2025−07−30]. https://arx-

iv.org/abs/2407.04925.

[75] CONTAL E, MCGOLDRICK G. RAGSys: item-cold-start recommender as RAG system[EB/OL]. (2024−08−15)[2025−07−30]. https://arxiv.org/abs/2405.17587.

[76] VAN DEVENTER H, MILLS M, EVRARD A. From interests to insights: an LLM approach to course recommendations using natural language queries[EB/OL]. (2024−12−26) [2025- 07−30]. https://arxiv.org/abs/2412.19312.

[77] 徐凤如, 李博涵, 胥帅. 基于深度学习与大语言模型的序列推荐研究进展[J]. 计算机科学与探索, 2025, 19(2): 344−366.
XU Fengru, LI Bohan, XU Shuai. Research progress on sequence recommendation based on deep learning and large language model[J]. Journal of frontiers of computer science and technology, 2025, 19(2): 344−366.

[78] ZHAO Xinping, HU Baotian, ZHONG Yan, et al. RaSeRec: retrieval-augmented sequential recommendation[EB/OL]. (2024−12−24) [2025−07−30]. https://arxiv.org/abs/2412.18378.

[79] CHEN Jia, DONG Qian, LI Haitao, et al. Qilin: a multimodal information retrieval dataset with APP-level user sessions[EB/OL]. (2025−03−01) [2025−07−30]. https://arxiv.org/abs/2503.00501.

[80] MARCELYN S C, GAO Y, ZHANG Y, et al. PathGPT: Leveraging Large Language Models for Personalized Route Generation[EB/OL]. (2025−04−08)[2025−07−30]. https://arxiv.org/abs/2504.05846.

[81] DELDJOO Y, HE Zhankui, MCAULEY J, et al. A review of modern recommender systems using generative models (gen-RecSys)[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2024: 6448−6458.

[82] LI Zhaoxing, YAZDANPANAH V, WANG Jindi, et al. TutorLLM: customizing learning recommendations with knowledge tracing and retrieval-augmented generation [EB/OL]. (2025−04−27) [2025−07−30]. https://arxiv.org/abs/2502.15709.

[83] 赵莹萍, 梁锦名, 陈贝章, 等. 多智能体大模型在农业中的应用研究与展望[J]. 智慧农业 (中英文), 2025, 7(5): 37−51.
ZHAO Yingping, LIANG Jinming, CHEN Beizhang, et al. Applications research progress and prospects of multi-agent large language models in agricultural[J]. Smart agriculture, 2025, 7(5): 37−51.

[84] LIU Pei, LIU Xin, YAO Ruoyu, et al. HM-RAG: hierarchical multi-agent multimodal retrieval augmented generation[EB/OL]. (2025−04−27)[2025−07−30]. https://arxiv.org/abs/2504.12330.

[85] MAIRITTHA T, SAWANGLOK T, RADEN P, et al. When pigs get sick: multi-agent AI for swine disease detection[EB/OL]. (2025−03−19)[2025−07−30]. https://arxiv.org/abs/2503.15204.

[86] MOHAMMED A M, MANSOOR I, BLYTHE S, et al. Developing an artificial intelligence tool for personalized breast cancer treatment plans based on the NCCN guidelines[EB/OL]. (2025−01−06)[2025−07−30]. https://arxiv.org/abs/2502.15698.

[87] SALAMA R, CAI J, YUAN M, et al. MemInsight: autonomous memory augmentation for LLM agents [EB/OL]. (2025−03−27) [2025−07−30]. https://arxiv.org/abs/2503.21760.

[88] YANG Rui. CaseGPT: a case reasoning framework based on language models and retrieval-augmented generation[EB/OL]. (2024−07−04) [2025−07−30]. https://arxiv.org/abs/2407.07913.

[89] JOREN H, ZHANG Jianyi, FERNG C S, et al. Sufficient context: a new lens on retrieval augmented generation systems[EB/OL]. (2024−11−09)[2025−07−30]. https://arxiv.org/abs/2411.06037.

[90] YANG Zhilin, QI Peng, ZHANG Saizheng, et al. HotpotQA: a dataset for diverse, explainable multi-hop question answering[EB/OL]. (2018−09−25)[2025−07−30]. https://arxiv.org/abs/1809.09600.

[91] FAN Yuxin, WANG Yuxiang, LIU Lipeng, et al. Research on the online update method for retrieval-augmented generation (RAG) model with incremental learning[EB/OL]. (2025−01−13) [2025−07−30]. https://arxiv.org/abs/2501.07063.

[92] 王栋. 大数据技术在互联网产品实时推荐算法中的应用与优化研究[J]. 软件, 2025, 46(3): 126−128.
WANG Dong. Research on the application and optimization of big data technology in the real time recommendation algorithm of Internet products[J]. Software, 2025, 46(3): 126−128.

[93] HIMEUR Y, SOHAIL S S, BENSAALI F, et al. Latest trends of security and privacy in recommender systems: a comprehensive review and future perspectives[J]. Computers & security, 2022, 118: 102746.

[94] ZENG Shenglai, ZHANG Jiankun, HE Pengfei, et al. The good and the bad: exploring privacy issues in retrieval-augmented generation (RAG)[EB/OL]. (2024−02−23)[2025−07−30]. https://arxiv.org/abs/2402.16893.

[95] 暴琳, 朱志宇, 孙晓燕, 等. 面向多源异构数据的个性化搜索和推荐算法综述[J]. 控制理论与应用, 2024, 41(2): 189−209.
BAO Lin, ZHU Zhiyu, SUN Xiaoyan, et al. Review on

personalized search and recommendation algorithms for multi-source heterogeneous data[J]. Control theory & applications, 2024, 41(2): 189−209.

[96] 张强, 刘丰. MDKG: 基于多模态知识图谱的 RAG 框架[J]. 计算机应用文摘, 2025, 41(2): 182−184, 188.
ZHANG Qiang, LIU Feng. MDKG: RAG framework based on multimodal knowledge graph[J]. Abstracts of computer applications, 2025, 41(2): 182−184, 188.

[97] LIN X V, CHEN X, CHEN M, et al. Ra-dit: Retrieval-aug- mented dual instruction tuning[C]//The Twelfth International Conference on Learning Representations. Vienna: IGLR, 2023: 1−18.

[98] VU T, IYYER M, WANG Xuezhi, et al. FreshLLMs: refreshing large language models with search engine augmentation[EB/OL]. (2023−11−22)[2025−07−30]. https://arxiv.org/abs/2310.03214.

[99] Yeon I, Choi J W. 3D room geometry inference from multichannel room impulse response using deep neural-network[EB/OL]. (2024−01−19)[2025−07−30]. https://ar-xiv.org/abs/2401.10453.

[100] BURGOS A, ALCHIERI E. Privacy-preserving smart contracts for permissioned blockchains: a zk-SNARK-based recipe part-1[EB/OL]. (2025−05−27)[2025−07−30]. https://arxiv.org/abs/2501.03391.

[101] 贾子琦, 王健宗, 张旭龙, 等. 基于大模型的具身智能任务规划研究: 从单智能体到多智能体[J]. 大数据, 2025, 11(2): 73−90.
JIA Ziqi, WANG Jianzong, ZHANG Xulong, et al. Large language model-based embodied intelligence task planning: from single-agent to multi-agent[J]. Big data research, 2025, 11(2): 73−90.

作者简介:

吴国栋, 副教授, 博士, 主要研究方向为人工智能、推荐系统。主持安徽省科技重大专项项目 1 项, 安徽省自然基金面上项目 1 项, 省级自然科学研究重点项目 1 项、一般项目 1 项。发表学术论文 40 余篇。E-mail: gdwu1120@qq.com。

谢东辰, 硕士研究生, 主要研究方向为推荐系统。E-mail: 764361338@qq.com。

黄雯婧, 硕士研究生, 主要研究方向为推荐系统。E-mail: 799558789@qq.com。