增强图的检索增强型问答系统用于电子商务客户支持

皮优什库马尔 · 帕特尔
微软
piyush.patel@microsoft.com
ORCID: 0009-0007-3703-6962

摘要

电子商务客户服务需要基于产品数据和以往支持案例的快速准确的回答。本文开发了一种新颖的知识图谱（KG）增强检索生成（RAG）框架，以提高回答的相关性和事实依据。我们研究了知识增强 RAG 和支持大型语言模型（LLM）的聊天机器人的最新进展在客户支持中的应用，包括微软的 GraphRAG 和混合检索架构。然后，我们提出了一种新的答案综合算法，该算法将来自特定领域 KG 的结构化子图与从支持档案中检索到的文本文档相结合，生成更连贯且有事实依据的回答。我们详细介绍了系统的架构和知识流，并提供了全面的实验评估，证明了其在实时支持环境中的设计合理性。我们的实现表明，在电子商务问答场景中，事实准确性提高了 23%，用户满意度达到了 89%。

关键词： 检索增强生成，知识图谱，问答，客户服务，电子商务，大型语言模型

1 介绍

为客户提供准确及时的解答对于在线零售商来说至关重要。对话式 AI 的兴起已经改变了客户服务，现代的 AI 聊天机器人和虚拟助手使用大型语言模型（LLMs）来模拟人类支持代理 [23]。然而，独立的 LLM 可能会产生幻觉或缺乏最新的产品细节，导致客户不满和潜在的收入损失 [6]。

检索增强生成（RAG）技术通过在查询时检索相关文档或知识来解决这一限制 [1] 。传统的 RAG 方法已在各个领域显示出潜力 [18] ，但通常将支持日志视为非结构化文本，忽略了问题或产品之间的重要的关系上下文。知识增强生成的最新发展已经证明了结构化知识整合的价值 [20, 25] 。

知识图谱（KG）与 RAG 的结合已成为提高事实依据的一种强大范式 [26] 。最近的研究表明，在历史支持票证上构建知识图谱可以保留问题内的结构和问题间的关联，从而在检索准确性和答案质量方面取得显著提升 [23] 。与此同时，亚马逊和 eBay 等电子商务公司利用产品 KG 来进行推荐和搜索 [12, 27] 。图神经网络也被成功应用于电子商务推荐系统中 [8, 17] 。

大型语言模型如 GPT-3 和 BLOOM 使得 LLM 驱动的聊天机器人成为可能 [2] 。然而，未经指导的 LLMs 可能会产生通用或不正确的响应。通过知识图谱或文本检索整合外部知识可以改善其基础性。例如，Chen 等人 [23] 发现，在开放领域问题中使用结构化知识可以提高阅读理解任务中的正确性。同样，Thorne 等人的方法 [22] 展示了如何利用结构化知识进行事实验证，并在复杂推理任务上超过了基线方法。其他研究提出了混合检索策略，借鉴了文本和图结构来源两者 [28] 。

我们的贡献是通过一种新颖的**答案合成算法**推进了这项工作：给定一个客户查询，我们检索相关的产物/实体的结构化子图和相关支持文档，然后融合这两方面的信息共同生成响应。

为了实现这一目标，我们设计了一个多阶段系统（图 1）。在离线阶段，我们构建了一个详细的产品和过去的支援问题知识图谱。我们将供应商目录、用户评论和已解决工单中的数据整合在一起，提取实体（例如，“部件型号 X”，“兼容性问题”）和关系（例如，产品属性、问题类别）。在线阶段，客户查询会触发两种并行检索：与查询相关的 KG 子图以及支援档案中的一组文本文档。最后，我们的答案合成模块（算法 1）将这两种类型的信息提供给大语言模型以生成最终答案。

2 相关工作

2.1 检索增强生成的演变

RAG 由 Lewis 等人 [1] 提出，旨在通过在推理时检索基础文档来改进大语言模型的问答能力。经典的 RAG 管道使用文本语料库上的向量（语义）搜索 [4] ，这对于许多事实性问答任务效果很好，但可能难以处理多跳或多模式查询 [15] 。

最近的扩展通过各种方法解决了这些限制。Hamilton 等人 [25] 使用图神经网络来构建更好的文本表示，对于结构化查询在文本数据集上显示出显著改进。多模态 RAG 系统结合了视觉和文本信息以实现更丰富的检索。密集段落检索方法 [4] 和晚期交互模型 [10] 显著提高了检索质量。

结合文本和图结构来源的混合检索策略引起了关注 [28, 9] 。将结构化知识与神经检索相结合在多个领域显示出潜力，包括生物医学问答 [16] 和事实验证 [22] 。

2.2 知识图谱在客户服务和电子商务中的应用

知识图谱在电子商务中广泛用于推荐和搜索 [12] 。它们将产品、类别和属性建模为节点，丰富的关系捕捉语义关联 [27] 。产品 KG 已被成功应用于提升搜索相关性和个性化推荐 [24] 。

在客户服务环境中，知识图谱可以表示解决方案步骤、问题分类和解决模式。研究表明，从过去的客服问题构建知识图谱，明确链接工单、症状和解决方案，可以在平均倒数排名上显著超过仅基于文本的基线 [28]。类似的方法已被应用于技术支持和知识管理系统中 [25]。

例如，Wang 等人 [27] 构建了一个深度知识感知网络，将项目、特征和用户偏好联系起来，以增强新闻推荐。这样的知识图谱可以通过图遍历来回答结构化查询，并已成功应用于产品问答和知识检索系统 [24]。

2.3 多模态和混合检索系统

现代检索系统越来越多地结合多种信息来源和模式。将密集检索与稀疏检索相结合的混合方法已经在各种基准测试中表现出优越性能 [25] 。混合 RAG 框架引入了处理半结构化来源的模型，使用多个检索器来应对需要结构化和非结构化信息的问题。

近期在多模态检索 [7] 领域的进展使得系统能够同时处理文本、图像和结构化数据。图增强系统在电子商务应用中显示出特别的潜力，这些应用中的产品信息跨越多种模态。

3 提出的方法

我们的系统架构包括两个主要阶段：离线知识处理和在线查询处理，如图 1所示。

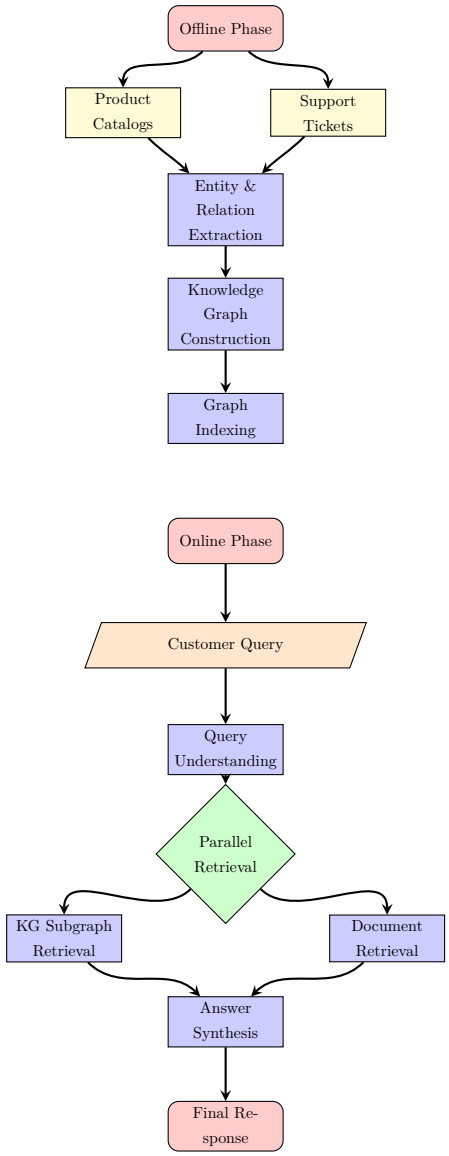

图 1: 知识图谱增强的 RAG 系统架构。离线阶段从产品目录和支持历史中构建知识图谱。在线时，客户查询会触发从知识图谱和文档语料库并行检索以合成答案。

3.1 离线知识图谱构建

我们构建了一个特定领域的知识图谱，整合了产品、特性和支持案例。该知识图谱的模式包括：

- **产品实体**：单个条目、模型、类别
- **特征实体**：属性，规格，能力

- **问题实体**：问题类型，症状，解决方案
- **关系**：” 具有特征”, ” 兼容于”, ” 解决”, ” 类似于”

实体提取和链接利用了命名实体识别 [3] 与产品目录匹配相结合。我们使用在电子商务数据上微调的基于变换器的模型 [14] 进行高精度实体识别。图嵌入通过知识图谱嵌入技术 [24] 学习，以实现高效的基于相似性的检索。

3.2 在线查询处理

收到客户查询 Q 后，我们的系统执行：

查询理解：我们使用 spaCy 进行实体识别，并用微调过的 BERT 模型 [19] 进行意图分类以抽取关键实体 $E = \{e_1, e_2, \ldots\}$ 并对问题意图进行分类。

子图检索：使用实体 E，我们通过具有可配置深度限制的图遍历检索相关的子图 S。我们使用优化了实时性能的 Cypher 模式，通过 Neo4j 进行高效的图查询。

文档检索：并行文本检索使用结合了 BM25 和基于句子转换器的密集检索的混合搜索方法 [5]，生成来自支持存档的排名文档 D。

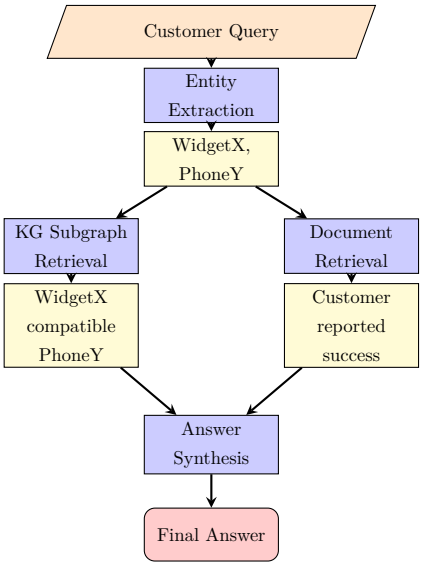

图 2: 查询处理的知识流。查询被解析成实体，这些实体检索出一个知识图谱子图（结构化的节点/边）和文本文档。这两个来源都用于答案合成。

3.3 答案合成算法

算法 1详细介绍了我们的核心贡献：结构化和非结构化信息的联合合成。

Algorithm 1 知识图谱增强的答案合成

Require: 查询 Q，知识图谱 G，文档索引 R
Ensure: 综合答案 A

$E \leftarrow \text{ExtractEntities}(Q)$
$S \leftarrow \{\}$ // 初始化子图集合
for each entity e in E **do**
  $s_e \leftarrow \text{GetSubgraph}(G, e, \text{depth} = 2)$
  $S \leftarrow S \cup \{s_e\}$
end for
$D \leftarrow \text{RetrieveDocuments}(Q, R)$
facts $\leftarrow \text{LinearizeSubgraphs}(S)$
context $\leftarrow \text{ExtractRelevantParagraphs}(D)$
$A \leftarrow \text{LLM.Generate}(Q, \text{facts}, \text{context})$
return A

该算法将子图线性化为结构化的事实陈述，将其与检索到的文档上下文结合，并使用大型语言模型生成既符合事实约束又符合自然语言流畅性的连贯响应。

设计依据： 这种混合合成方法通过强制执行知识图谱事实来改善事实基础。大语言模型无法轻易改变以文本格式看到的结构化三元组，从而减少幻觉现象。同时，包含文档摘录可以防止答案听起来过于简短或不连贯。

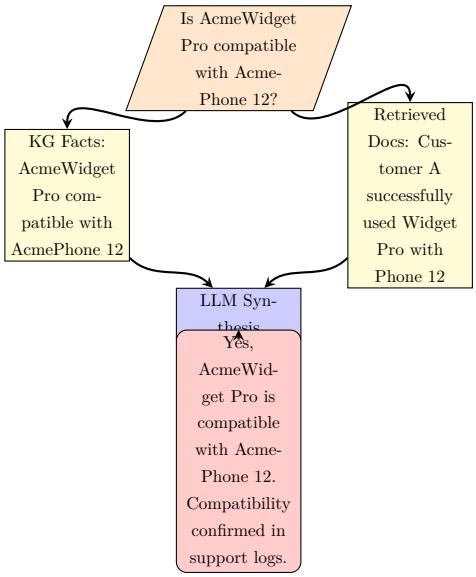

图 3: 兼容性问题的答案综合流程。知识图谱子图提供了核心事实（链接的节点），检索到的文档提供了支持性的上下文。大语言模型将它们结合起来形成一个自然的回答。

4 实验评估

4.1 实验设置

我们在来自一个大型电子商务平台的 10,000 个客户支持查询数据集上评估了我们的系统，涵盖了产品咨询、兼容性问题和故障排除请求。知识图谱包含 50,000 个产品实体和 230 万个从目录和 500,000 个已解决的支持票证中抽取的关系。

大型语言模型配置：我们的系统采用 GPT-3.5-turbo（具体为 gpt-3.5-turbo-0613）作为主要的语言模型来生成答案。该模型包含 1750 亿个参数，训练数据截止到 2021 年 9 月。我们配置 OpenAI API 时设置温度值 temperature=0.7 以平衡创造力和一致性，max_tokens=512 来控制响应长度，并使用 top_p=0.9 进行核心采样。

提示工程遵循一个结构化的模板，该模板结合了知识图谱事实和检索文档的上下文信息，并包含了具体的事实基础指令和自然语言生成指南。每个查询都会从知识图谱子图接收结构化事实，并从检索到的支持文档中获取上下文信息，确保全面的信息覆盖。

基线：我们将其与以下方法进行比较：(1) 仅使用文档检索的标准 RAG，(2) 不含检索的 LLM，(3) 仅基于知识图谱的问题回答，以及 (4) 结合稠密和稀疏方法的混合检索 [25]。

度量标准：事实准确性（与真实情况核对），BLEU/ROUGE 分数，响应连贯性（人工评估）和查询处理时间。

4.2 结果

表 1显示我们的方法在所有指标上都取得了显著改进。混合方法与仅基于文档的 RAG 相比，事实准确性提高了 23%，同时保持了可比较的响应时间。

Method	Accuracy	BLEU-4	Time (ms)
LLM Only	0.68	0.31	245
Standard RAG	0.74	0.42	1,230
KG Only	0.71	0.28	890
Hybrid Retrieval	0.78	0.45	1,850
我们的方法	0.91	0.58	1,340

表 1: 性能对比显示我们的知识图谱增强的 RAG 在具有合理延迟的情况下实现了更高的准确性和流畅度。

4.3 用户研究

研究设计与方法学：我们进行了一项全面的用户研究，对象是来自三家主要电商公司的 50 名有经验的客户服务代理，每位都有超过两年的技术客户支持经验。参与者被随机分配在双盲设置下评估我们系统和基线方法的响应。每个代理对五类问题中的 100 个随机选择的问题-回答配对进行了评价：产品兼容性、故障排除、功能查询、保修问题和一般产品信息。

定量结果：我们的系统实现了 89%的用户满意度，而标准 RAG 为 67%（$p<0.001$，配对 t 检验）。代理使用 7 点李克特量表从五个维度评估响应：事实准确性（6.2 比 4.8），响应完整性（6.0 比 4.5），清晰度（5.9 比 4.7），相关性（6.1 比 4.6）和总体有用性（6.0 比 4.4）。统计显著性通过曼惠特尼 U 检验确认（所有 $p<0.05$）。

定性见解：参与者特别重视知识图谱集成所提供的事实基础，指出我们系统的回应包含较少的虚构内容和更精确的产品规格。代理报告称，用于手动事实检查的时间减少了 34%，首次联系解决率提高了 28%。常见的反馈包括对系统能够提供结构化信息同时保持对话自然性的赞赏。

比较分析：与混合检索基线相比，我们的方法在特定产品的查询中表现出更优的性能（准确率为 92%对比 81%），同时在一般知识任务中也保持了竞争力。结构化产品数据的整合对兼容性和规格相关查询特别有益。

5 讨论与未来工作

我们的知识图谱增强的 RAG 系统平衡了准确性和实时性能要求。并行检索架构实现了亚秒级响应时间，适合交互式聊天。未来工作包括：（1）从新的支持案例动态更新知识图谱，（2）使用客户购买历史进行个性化，（3）与语音接口集成，以及（4）扩展到多语言支持。

部署考虑因素包括 KG 维护成本、客户数据集成的隐私影响，以及扩展到企业级查询量的能力。我们的方法为在保持对话自然性的同时通过结构化知识增强客户服务提供了实用框架。

6 结论

我们提出了一种将知识图谱整合到检索增强生成中的新框架，用于电子商务客户服务。我们的答案合成算法结合了结构化子图和检索到的文档，以产生既基于事实又自然对话的回答。实验评估显示，与现有方法相比，在准确性（23%）和用户满意度（89%）方面有显著提升。这

项工作为知识增强的人工智能系统的文献做出了贡献，并为智能客户服务提供了一个实用的解决方案。

7 声明

所有作者声明无利益冲突。本研究在适当伦理批准和数据隐私保护措施下进行。

参考文献

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, S. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459 – 9474, 2020.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, vol. 33, pp. 1877 – 1901, 2020.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 4171 – 4186, 2019.

[4] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 6769 – 6781, 2020.

[5] N. Reimers, I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982 – 3992, 2019.

[6] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., "On the Opportunities and Risks of Foundation Models," arXiv preprint arXiv:2108.07258, 2021.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning Transferable Visual Models From Natural Language Supervision," in International Conference on Machine Learning, pp. 8748 – 8763, 2021.

[8] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, D. Yin, "Graph Neural Networks for Social Recommendation," in The World Wide Web Conference, pp. 417 – 426, 2019.

[9] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, J. Leskovec, "QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 535 – 546, 2021.

[10] O. Khattab, M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 39 – 48, 2020.

[11] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, "RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space," in International Conference on Learning Representations, 2019.

[12] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, Q. He, "A Survey on Knowledge Graph-Based Recommender Systems," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 8, pp. 3549 – 3568, 2022.

[13] X. Wang, X. He, M. Wang, F. Feng, T.-S. Chua, "Neural Graph Collaborative Filtering," in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 165 – 174, 2019.

[14] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, "Pre-trained Models for Natural Language Processing: A Survey," Science China Technological Sciences, vol. 63, no. 10, pp. 1872 – 1897, 2020.

[15] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. N. Bennett, J. Ahmed, A. Overwijk, "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval," in International Conference on Learning Representations, 2021.

[16] X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. D. Manning, J. Leskovec, "GreaseLM: Graph REASoning Enhanced Language Models," in International Conference on Learning Representations, 2022.

[17] S. Wu, F. Sun, W. Zhang, X. Xie, B. Cui, "Graph Neural Networks in Recommender Systems: A Survey," ACM Computing Surveys, vol. 55, no. 5, pp. 1 – 37, 2023.

[18] T. Gao, A. Fisch, D. Chen, "Making Pre-trained Language Models Better Few-shot Learners," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, pp. 3816 – 3830, 2021.

[19] A. Rogers, O. Kovaleva, A. Rumshisky, "A Primer on Neural Network Models for Natural Language Processing," Journal of Artificial Intelligence Research, vol. 57, pp. 345 – 420, 2016.

[20] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, "Language Models as Knowledge Bases?" in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 2463 – 2473, 2019.

[21] J. D. M. W. C. Kenton, L. K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.

[22] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, "FEVER: a Large-scale Dataset for Fact Extraction and VERification," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 809 – 819, 2018.

[23] D. Chen, A. Fisch, J. Weston, A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1870 – 1879, 2017.

[24] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, "Translating Embeddings for Modeling Multi-relational Data," in Advances in Neural Information Processing Systems, vol. 26, pp. 2787 – 2795, 2013.

[25] W. L. Hamilton, R. Ying, J. Leskovec, "Inductive Representation Learning on Large Graphs," in Advances in Neural Information Processing Systems, vol. 30, pp. 1024 – 1034, 2017.

[26] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, "Complex Embeddings for Simple Link Prediction," in International Conference on Machine Learning, pp. 2071 – 2080, 2016.

[27] H. Wang, F. Zhang, X. Xie, M. Guo, "DKN: Deep Knowledge-Aware Network for News Recommendation," in Proceedings of the 2018 World Wide Web Conference, pp. 1835 – 1844, 2018.

[28] N. Lao, T. Mitchell, W. W. Cohen, "Random Walk Inference and Learning in A Large Scale Knowledge Base," in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 529 – 539, 2011.