大模型（LLMs）RAG 优化策略 —— RAG-Fusion篇

来自： AiGC面试宝典

宁静致远 2024年03月19日 22:30

一、RAG 有哪些优点？

1. **向量搜索融合**：RAG 通过将向量搜索功能与生成模型相结合，引入了一种新颖的范式。这种融合使大型语言模型（LLM）能够生成更丰富、更具上下文意识的输出。
1.
2. **减少幻觉现象**：RAG 显著降低了 LLM 产生幻觉的倾向，使生成的文本更加基于数据。
3. **个人和专业效用**：从个人应用（如浏览笔记）到更专业的集成，RAG 在提高生产力和内容质量方面展示了其多功能性，同时基于可信的数据来源。

二、RAG 存在哪些局限性？

1. **当前搜索技术的限制**：RAG 受到限制的方面与我们的检索式基于词汇和向量的搜索技术相同。
2. **人类搜索效率低下**：人类在向搜索系统输入他们想要的内容时并不擅长，如打字错误、含糊的查询或词汇有限，这常常导致错过那些超出显而易见的顶部搜索结果的大量信息。虽然 RAG 有所帮助，但它并没有完全解决这个问题。
1.
3. **搜索的过度简化**：我们普遍的搜索范式是将查询线性映射到答案，缺乏理解人类查询的多维性。这种线性模型通常无法捕捉更复杂用户查询的细微差别和上下文，导致结果相关性较低。

三、为什么 需要 RAG-Fusion？

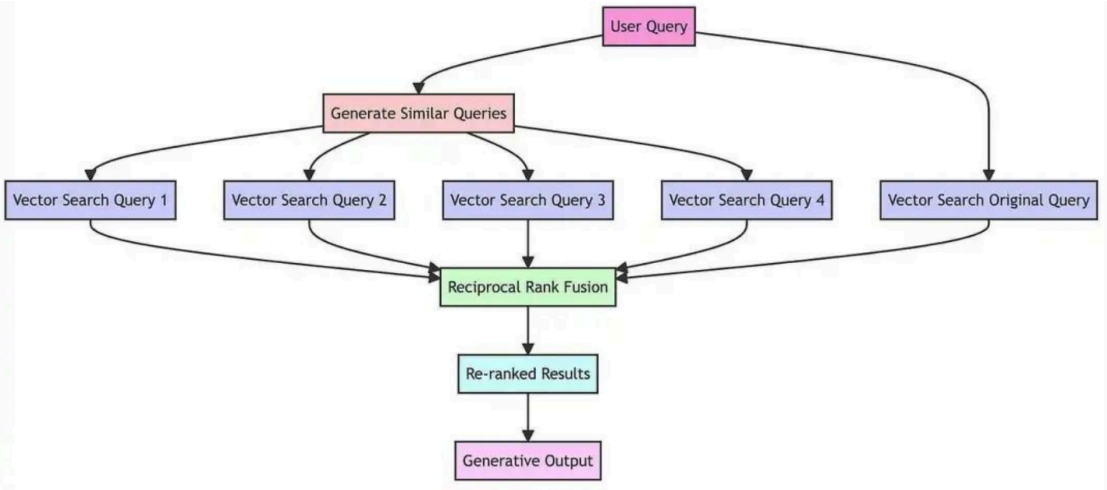
它解决了 RAG 固有的限制，**通过生成多个用户查询并重新排序结果。利用逆向排名融合和自定义向量评分加权进行综合、准确的搜索。**

RAG-Fusion 旨在弥合用户明确询问与他们意图询问之间的差距，更接近于发现通常隐藏的变革性知识。

四、说一下 RAG-Fusion 核心技术？

RAG-Fusion 的基础三元组与 RAG 相似，核心技术包括：

1.
    1. 通用编程语言，通常是 Python。
    2. 专用的向量搜索数据库，如 Elasticsearch 或 Pinecone，用于驱动文档检索。
    3. 强大的大型语言模型，如 ChatGPT，用于创造文本。



然而，与 RAG 不同的是，RAG-Fusion 通过几个额外的步骤区分自己——**查询生成和结果重新排序**。

五、说一下 RAG-Fusion 工作流程？

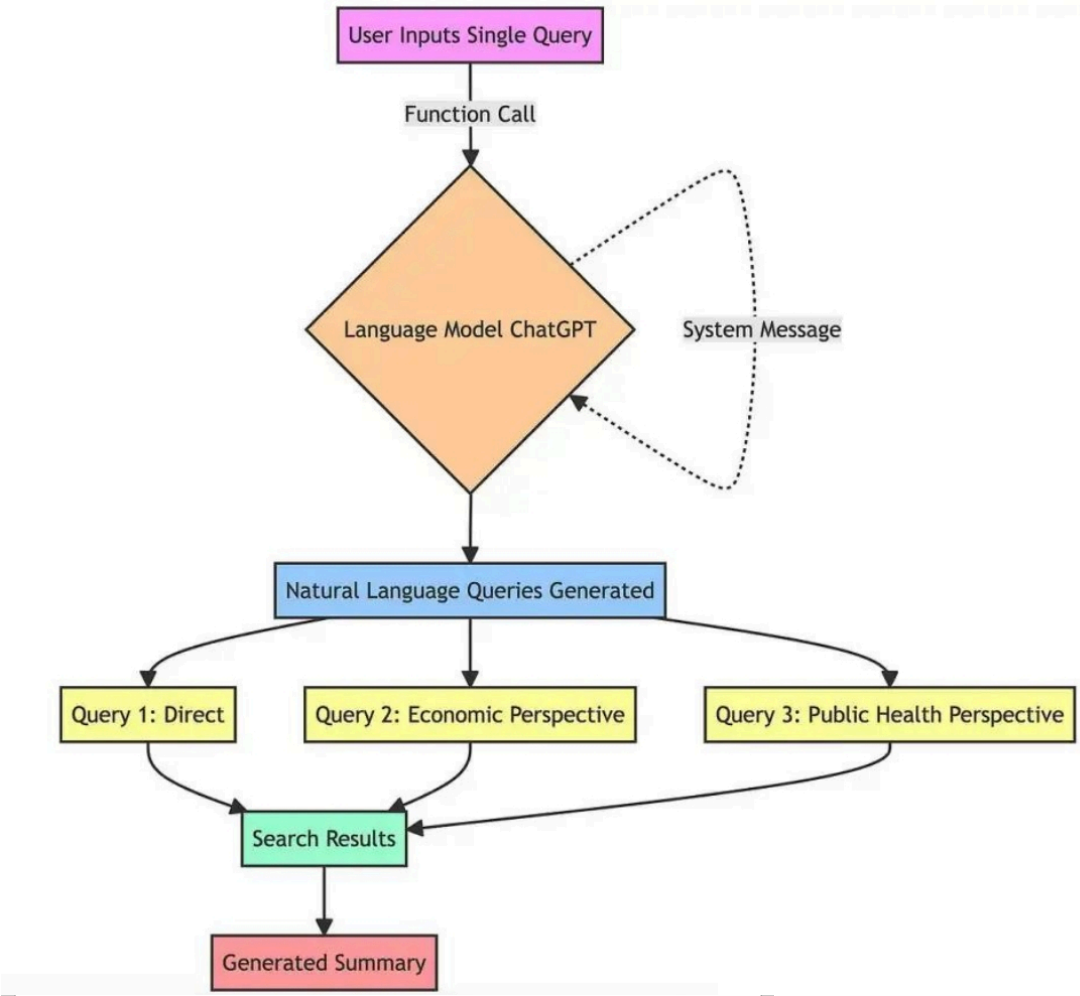
5.1 多查询生成

- 为什么要生成多个查询？

在传统的搜索系统中，用户通常输入一个查询来查找信息。虽然这种方法直接简单，但它有局限性。

单一查询可能无法完全捕捉用户感兴趣的全部范围，或者可能过于狭窄而无法产生全面的结果。因此，从不同角度生成多个查询就显得尤为重要。

5.2 多查询生成 技术实现（提示工程）？



利用提示工程和自然语言模型拓宽搜索视野，提升结果质量。利用提示工程生成多个查询至关重要，这些查询不仅与原始查询相似，还提供不同的视角或角度。

5.3 多查询生成 工作原理？

1. **调用语言模型**：该函数调用一个语言模型（在本例中为 chatGPT）。该方法期望一个特定的指令集，通常描述为“系统消息”，以指导模型。

例如，这里的系统消息指导模型充当“AI 助手”。

1. **自然语言查询**：模型接着基于原始查询生成多个查询。
2. **多样性和覆盖范围**：这些查询不是随机变化。它们是经过精心生成的，以提供原始问题的不同视角。

例如，如果原始查询是关于“气候变化的影响”，那么生成的查询可能包括像“气候变化的经济后果”、“气候变化与公共健康”等角度。

这种方法确保了搜索过程考虑了更广泛的信息范围，从而提高生成总结的质量和深度。

5.4 逆向排名融合（RRF）

5.4.1 为什么选择 RRF？

逆向排名融合（RRF）是一种将多个搜索结果列表的排名结合起来产生单一统一排名的技术。该技术由滑铁卢大学（加拿大）和谷歌合作开发，根据其作者的说法，“产生的结果比任何单个系统更好，也比标准”重新排名方法更好。

$$RRFscore(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)},$$

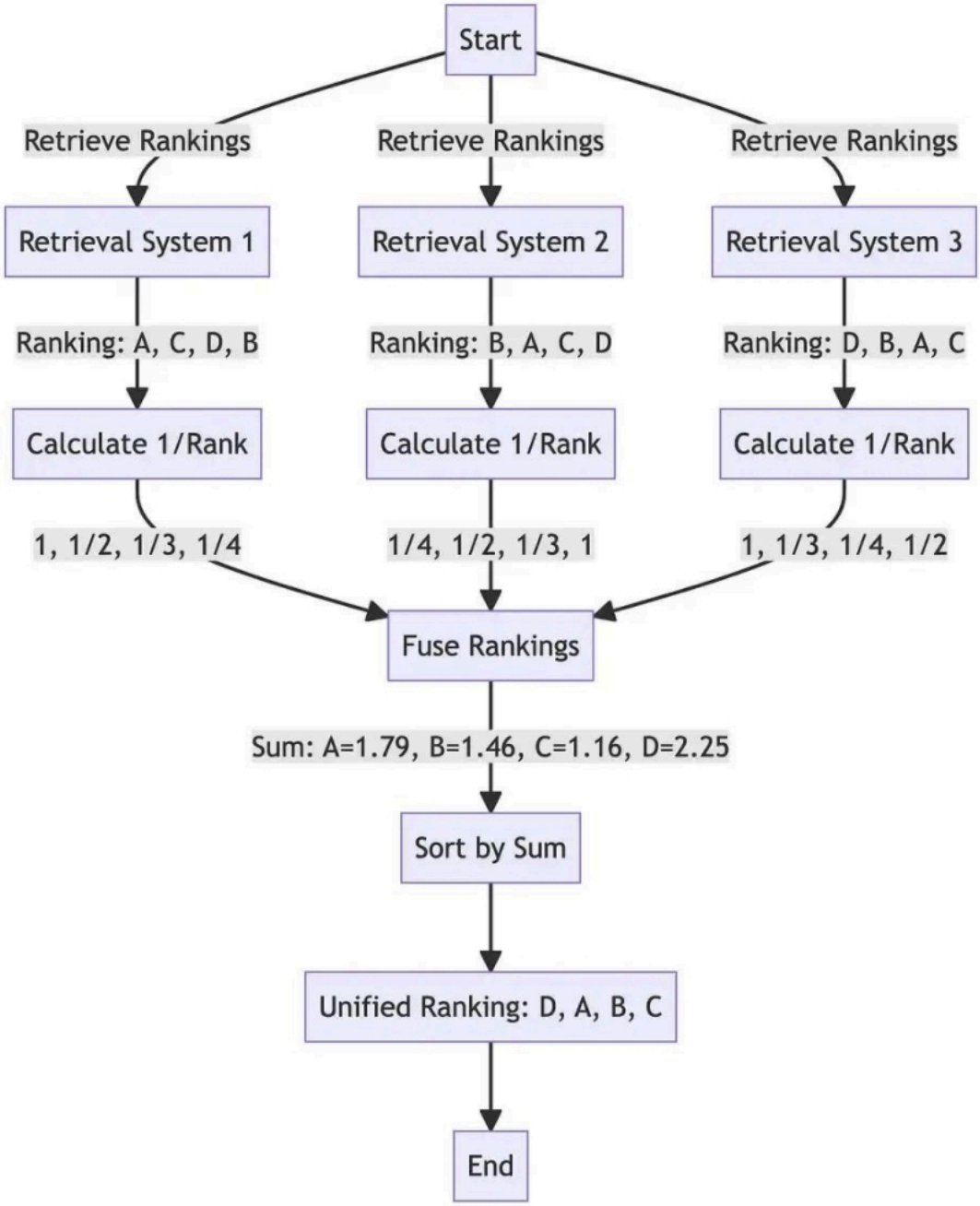
RRF algorithm where k=60. Image from — Reciprocal Rank Fusion outperforms Condorcet and individual Rank Learning Methods

通过结合不同查询的排名，我们增加了最相关文档出现在最终列表顶部的机会。RRF 特别有效，因为它不依赖于搜索引擎分配的绝对分数，而是依赖于相对排名，使其非常适合结合可能具有不同规模或分数分布的查询结果。

通常情况下，RRF 被用于混合词汇和向量结果。虽然这种方法有助于弥补向量搜索在查找特定术语（例如缩写）时的不足，但我对结果并不印象深刻，**这些结果往往更像是多个结果集的拼凑，因为同一个查询的词汇和向量搜索很少出现相同的结果。**

可以将 RRF 看作是那种坚持在做决定前听取每个人意见的人。只不过在这种情况下，它不仅不烦人，而且有帮助。众多观点越多，结果越准确。

5.4.2 RRF 技术实现?



运用 RRF 根据多组搜索结果的位置重新排序文档。逆向排名融合位置重新排序系统

1. 函数 reciprocal_rank_fusion 接收一个搜索结果的字典，其中每个键是一个查询，相应的值是根据该查询的相关性排名的文档 ID 列表。
2. **RRF 算法然后基于其在不同列表中的排名为每个文档计算一个新分数，并根据这些分数排序以创建最终的重新排名列表。**
3. 计算完融合分数后，函数按照这些分数的降序对文档进行排序，以获得最终的重新排名列表，然后返回该列表。

5.4.3 生成性输出 用户意图保留

使用多个查询的一个挑战是可能稀释用户的原始意图。为了缓解这一点，我们指示模型在提示工程中更重视原始查询。

5.4.4 生成性输出 用户意图保留 技术实现

最后，将重新排名的文档和所有查询输入到 LLM 提示中，以生成典型的 RAG 方式的生成性输出，如请求回应或摘要。

通过将这些技术和技巧层叠起来，RAG Fusion 提供了一种强大而细腻的文本生成方法。它利用搜索技术和生成性人工智能的最佳特性，产生高质量、可靠的输出。

六、RAG-Fusion 的优势和不足

6.1 RAG-Fusion 优势

1. **更优质的源材料**：使用 RAG Fusion 时，你的搜索深度不仅仅是“增强”——而是被放大。重新排名的相关文档列表意味着你不只是在信息表面刮刮而已，而是潜入观点的海洋。结构化输出易于阅读，直观上可信赖，这在对人工智能生成内容持怀疑态度的世界中至关重要。
2. **增强用户意图**：对齐 RAG Fusion 的核心设计是作为一个富有同情心的人工智能，揭示用户努力表达但可能无法清晰表述的内容。采用多查询策略捕捉用户信息需求的多面性表现，因此提供全面的输出，并与用户意图产生共鸣。
3. **结构化、富有洞见的输出**：通过汲取多样化的信息源，模型制作出组织良好且富有洞见的答案，预测后续问题并主动解答。
4. **自动纠正用户查询**：该系统不仅解释，还优化用户查询。通过生成多个查询变体，RAG Fusion 执行隐含的拼写和语法检查，从而提高搜索结果的准确性。
5. **处理复杂查询**：人类语言在表达复杂或专业思想时常常出现障碍。该系统作为语言催化剂，生成可能包含所需专业术语或术语的变体，用于更集中和相关的搜索结果。它还可以将更长、更复杂的查询分解成向量搜索可以处理的更小、更易管理的部分。
6. **搜索中的意外发现**：考虑“未知的未知”——直到遇到你才知道需要的信息。通过采用更广泛的查询范围，系统促进了发现意外信息的可能性。虽然这些信息并非明确寻求，但对用户来说却可能是一个欧雷卡时刻。这使 RAG Fusion 区别于其他传统搜索模型。

6.2 RAG-Fusion 挑战

1. **过于冗长的风险**：RAG-Fusion 的深度有时可能导致信息泛滥。输出可能过于详细，令人不堪重负。可以将 RAG-Fusion 比作那个解释过多的朋友——信息丰富，但有时你可能需要他们直接了当一些。

2. **平衡上下文窗口**：多查询输入和多样化文档集的引入可能会使语言模型的上下文窗口受到压力。想象一个舞台上挤满了演员，使得剧情难以跟进。对于上下文限制较紧的模型，这可能导致输出不连贯甚至被截断。
3. **伦理和用户体验考虑**：拥有巨大力量的同时也伴随着巨大的责任。对于 RAG Fusion 来说，操作用户查询以改善结果的能力似乎正踏入某种道德灰区。在改善搜索结果的同时平衡用户意图的完整性至关重要，我对于实施这个解决方案时你应该考虑的一些事情有所思考：
    1. 伦理顾虑：
        1. 用户自主性：操作用户查询有时可能偏离原始意图。考虑我们向人工智能让渡多少控制权以及代价是什么非常重要。
        2. 透明度：不仅仅是关于更好的结果；如果用户的查询被调整，他们应当意识到这一点。这种透明度对于维护信任和尊重用户意图至关重要。
    2. 用户体验（UX）增强：
        1. 保留原始查询：RAG Fusion 优先考虑初始用户查询，确保其在生成过程中的重要性。这作为防止误解的保障。
        2. 过程可见性：展示生成的查询以及最终结果，为用户提供搜索范围和深度的透明视图。这有助于建立信任和理解。
    3. UX/UI 实施建议：
        1. 用户控制：提供用户切换 RAG Fusion 的选项，允许他们在手动控制和增强的人工智能辅助之间选择。
        2. 指导和清晰度：关于 RAG Fusion 工作方式的工具提示或简要说明可以帮助设定明确的用户期望。