大模型面经—RAG工程实践经验总结

原创 喜欢卷卷的瓦力 瓦力算法学研所 2024年10月24日 10:00 上海

◈ 技术总结专栏 ◈

作者：喜欢卷卷的瓦力

RAG工程经验面经总结。

本篇属于RAG系列，RAG系列的文章可以参考下面的内容。

大模型面经——以医疗领域为例，整理RAG基础与实际应用中的痛点

RAG工程如何评测?

虽然RAG工程整体有很多论文、算法和方法论，但在实际使用过程中，当数据量大了RAG很容易出现不可控的问题， 本篇就针对实践过程中遇到的问题总结面经进行分享，看看能不能给大家提供一些帮助。下面是一个快捷目录。

下面是答案。

一. RAG如何去优化索引结构?

1. 优化被检索的embedding

1）微调被检索的embedding

目的：让被检索的内容与query之间的相关性更加紧密

特别是术语更新较快且比较罕见的领域，可以针对性地进行微调。

2）动态embedding

目的：基于上下文动态调整embedding

当然这只是个发论文的思路，工程落地的时候这块还是有待验证的。

3）检索后处理流程优化

目的：直接把所有检索结果给大模型可能会超出上下文窗口限制，内容过多噪声也可能比较多。

优化方法：

- ReRank
- Prompt 压缩
- RAG 管道优化
- 混合搜索
- 递归检索与查询引擎
- StepBack-prompt 方法
- 子查询
- HyDE 方法

2. 优化query的chunk大小

chunk大小非常关键，决定了从向量存储中检索的文档的长度。小块可能导致文档缺失一些关键信息，而大块可能引入无关的噪音。找到最佳块大小是要找到正确的平衡。

目前来说一般是按不同块大小划分验证集做实验，直接用验证集效果说话。

3. 结合不同粒度信息进行混合检索

虽然向量搜索有助于检索与给定查询相关的语义相关块，但有时在匹配特定关键词方面缺乏精度。根据用例，有时可能需要精确匹配。

混合检索就是结合embedding搜索和关键词搜索。

二. 当混合检索以及基于不同大小的chunk去检索效果都不太好的时候，如何优化?

这种情况就要针对具体的case关注知识库里是否有答案了。

如果有答案但是没检索出来，那么大概率可能答案被错误分割开了，那么可以结合一些小模型（BERT等）拿来做上下句预测；

另外也可以分析 query 和 doc 的特点：字相关还是语义相关，一般建议是先用推荐系统经典的ES做召回，然后才用模型做精排

三. 如何通过rerank去提升RAG效果的，有哪些方案?

背景：当检索时，前K个结果不一定按最相关的方式排序。它们都是相关的， 但在这些相关内容中，最相关的可能并不是第1或第2个，而是排名靠后的。rerank就是将最相关的信息重新定位到排名靠后的检索结果。

这里推荐一些思路：
Diversity Ranker 根据文档的多样性进行重新排序；
LostInTheMiddleRanker 中提出LLM 会着重把注意力放在文本开头和结尾的位置，那就把最需要让 LLM 关注的 documents 放在开头和结尾的位置。

另外还有一些经典的框架LlamaIndex、LangChain 和 HayStack都可以参考和直接用。

其实主要的思路都大同小异，实际工作中还是主要会结合具体的case来优化，大家有更多的问题和经验也可以一起分享讨论。

参考文献

[1] Retrieval-Augmented Generation for Large Language Models: A Survey(arxiv.org/pdf/2312.10997)

[2] 论文分享|RAG理论-第一篇-概述 - 知乎(https://zhuanlan.zhihu.com/p/678616587)

[3] 提升RAG性能的关键技术：从数据清理到混合检索的全方位讨论 - 知乎(https://zhuanlan.zhihu.com/p/676463769)

想要获取技术资料的同学欢迎关注公众号，进群一起交流~