從局部到全局
一場由圖表驅動的檢索增強生成技術 GraphRAG

RAG 是一種利用外部知識庫，來增強大型語言模型 (LLM) 回答問題能力的技術。它首先從數據庫中檢索與查詢相關的訊息，然後將這些訊息添加到 LLM 的上下文窗口中，最後生成答案。然而傳統的 RAG 技術在處理全局性問題時效率較低，比如，當用戶詢問：「去年 MyLLM 技術團隊的成果？」這時傳統 RAG 可能就有點頭大了，因為它對整個文本庫的把握不夠深入回答全局性問題就會有點吃力啦！

為了克服這一挑戰，微軟提出了 GraphRAG 方案。

GraphRAG 是一種結合知識圖譜生成、RAG 及全局性摘要（QFS）的技術，旨在為用戶解決全局問題。

GraphRag 在對數據建立索引時，流程如下：

1. 構建知識圖譜： 使用 LLM 從源文檔中提取實體、關係和相關訊息，並構建知識圖譜。

2. 生成社區摘要： 利用社區檢測算法將知識圖譜劃分為多個緊密聯繫的社區，並為每個社區生成摘要。

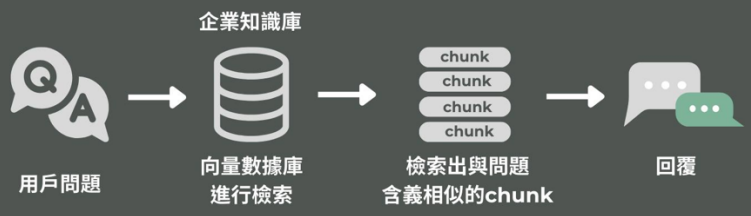
3. 生成答案： 根據用戶查詢，並行地從相關的社區摘要中生成答案片段，最後將這些片段整合為最終答案。

GraphRag 結合了 RAG 的高效檢索能力和 QFS 的全局摘要能力，可以有效地回答全局性的問題！不過，目前，GraphRag 的評估僅限於特定類型的全局性問題和數據集，且若資料源太龐大，可能會花費很多 API Request 去調用模型，或許正因為如此微軟才會開源 GraphRag，希望讓更廣泛的社群能夠使用和受益於這項技術，同時促進創新，讓系統更快地改進和優化。然而，來源並沒有明確說明微軟是否希望依靠社群力量來優化 GraphRAG，也沒有提到任何關於速度和成本的問題！

GraphRag

結合知識圖譜的檢索增強生成技術

傳統 Rag 有什麼缺點呢？



檢索增強生成 (RAG) 是一種透過從外部知識來源檢索相關資訊，使大型語言模型 (LLM) 能夠回答關於私有和先前未見過的文檔集的問題的既定方法。然而，RAG 無法處理針對整個文字語料庫的**全局性問題**，例如：**「去年MyLLM技術團隊的成果？」**，因為這是一個**查詢導向摘要 (QFS) 任務**，而非一個明確的檢索任務。

GRAPHRAG

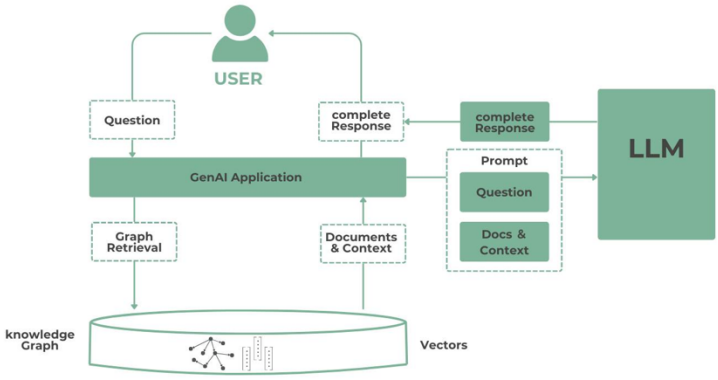
典型 *RAG* 核心是向量搜索：接收到文本，並從候選的數據庫中，結合問題，返回相似概念的文本
而 *GraphRAG*還是 *RAG* 只是在檢索路徑中添加「知識圖譜」。

預處理：使用大模型構建

（提前整理訊息）

GRAPHRAG

結合知識圖譜生成、檢索增強生成（RAG）和查詢重點摘要（QFS）的技術，
用於支持用戶對整個文本語料庫進行理解和分析。



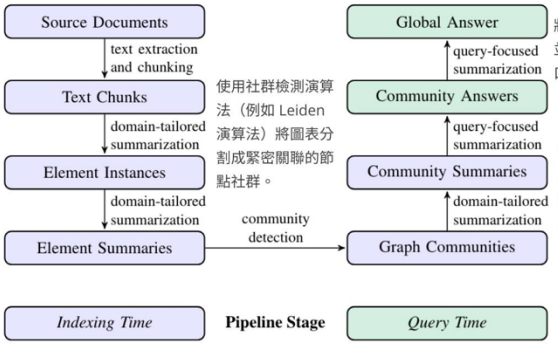
生成式AI應用架構圖

GRAPHRAG

文檔分塊：將輸入文檔分割成較小的文字區塊。

元素實例化：使用 LLM 從每個文字區塊中識別和提取圖表元素（實體、關係、斷言）。

元素摘要：將每個圖表元素的所有實例級摘要合併為單個描述性文字區塊。



使用社群檢測演算法（例如 Leiden 演算法）將圖表分割成緊密關聯的節點社群。

將生成的社區答案按相關性排序，並迭代地添加到 LLM 的上下文窗口中，直到達到token限制。

使用上一步生成的社區摘要生成最終答案。

社群摘要：為每個社群生成摘要，提供圖表索引和輸入文檔結構的全局視圖。

索引和查詢流程圖

參考資料：

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., T& Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130 https://arxiv.org/abs/2404.16130

圖表 RAG 的開源網址：https://www.microsoft.com/en-us/research/project/graphrag/。