# Um Relatório Técnico e Proposta de Avanço para Sistemas GraphRAG

**Nicolas Mady Corrêa Gomes**[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)

nicolas.gomes@icomp.ufam.edu.br

***Abstract.*** *The Retrieval-Augmented Generation (RAG) has shown itself to be efficient in the solution of some limitations on Large Language Models (LLMs). On one side, the traditional RAG separates the knowledge base into independent parts, ignoring semantic connections between them. On the other side, there arrives the RAG based on graph (GraphRAG) as an extension of the traditional RAG, mantaining semantic relations between supposedly isolated parts. This article starts by presenting a brief introduction to the RAG fundamentals, then it is shown a comparative analysis between four current GraphRAG methodologies (GraphRAG Microsoft, G-RAG, Align-GRAG e G-Retriever) and spots of gray literature are identified, like retrieval noise and the gap in the representation between graph and language. Finally, a research agenda focused on graph denoising, reranking with structural conscience, and hybrid fusion architecture is proposed.*

***Resumo.*** *A Geração Aumentada por Recuperação (RAG) tem se mostrado eficiente na solução de certas limitações de Modelos de Linguagem Grandes (LLMs). Por um lado, a RAG tradicional separa a base de conhecimento em partes independentes, ignorando conexões semânticas entre elas. Por outro lado, surge a RAG baseada em grafo (GraphRAG) como uma extensão da RAG tradicional, matendo relações semânticas entre partes supostamente isoladas. Este artigo começa apresentando uma breve introdução aos fundamentos da RAG, em seguida é mostrada uma análise comparativa entre quatro metodologias atuais de GraphRAG (GraphRAG Microsoft, G-RAG, Align-GRAG e G-Retriever) e são identificados pontos de literatura cinzenta, tais como ruído na recuperação e a lacuna na representação entre grafo e linguagem. Por fim, propõe-se uma agenda de pesquisa focada em denoising de grafos,reranking com consciência estrutural e arquiteturas de fusão híbrida.*

## 1. Introdução

A RAG estabeleceu-se como uma solução robusta para limitações intrínsecas aos LLMs, com destaque à propensão a alucinações e à dependência de dados estáticos. A RAG tradicional já aumenta a confiabilidade ao basear as respostas dos modelos em conhecimentos externos específicos. Contudo, a abordagem tradicional, ao fragmentar documentos em *chunks* independentes, sem considerar conexões semânticas, impede uma visão global do conhecimento e, por conseguinte, deixa de fornecer respostas mais precisas.

Para contornar esse problema, foi criada a RAG baseada em grafos (GraphRAG), empregando estruturas que preservam relações entre entidades. Este trabalho tem dois

objetivos: (1) analisar e comparar quatro das principais abordagens contemporâneas de GraphRAG, detalhando suas arquiteturas; e (2) identificar lacunas críticas na literatura atual a fim de propor novas direções de pesquisa com o objetivo de desenvolver sistemas mais robustos, precisos, e com o mínimo de alucinação percebida nos LLMs.

## 2. Fundamentos do GraphRAG

Para facilitar o entendimento das implementações, é preciso definir o conceito de **Grafo de Atributos Textuais (TAG)**: uma estrutura onde vértices ($V$) e arestas ($E$) são enriquecidos com atributos textuais ($x_n, x_e$), favorecendo-os para o processamento por LLMs.

O fluxo geral de um sistema GraphRAG segue as seguintes fases:

1. **Construção e Indexação:** Processamento de documentos para extração de entidades e relações (nós e arestas).
2. **Recuperação (Retrieval):** Busca de subgrafos semanticamente relevantes em vez de textos isolados.
3. **Pós-processamento:** Filtragem de ruído, reclassificação (*reranking*) ou alinhamento.
4. **Geração:** Textualização do conhecimento estruturado para ser processado pelo LLM final.

## 3. Análise das Metodologias do Estado da Arte

É importante frisar que o estudo e aplicação do GraphRAG possui vários frontes. A seguir, foram analisadas quatro abordagens que se destacam por inovar em diferentes estágios do pipeline.

### 3.1. GraphRAG (Microsoft)

Focado em *Global Sensemaking*, o método da Microsoft busca responder perguntas holísticas sobre todo o corpus. Utiliza detecção de comunidades (algoritmo de Leiden) para criar resumos hierárquicos. A resposta é gerada via um processo de *Map-Reduce* sobre os resumos das comunidades, permitindo sínteses temáticas que o RAG vetorial não consegue realizar [Edge et al. 2024].

### 3.2. G-RAG (UCLA/Google)

Atua como um *reranker* especializado. Emprega uma arquitetura de duplo grafo: um Grafo de Representação de Significado Abstrato (AMR) e um Grafo de Documentos. Sua inovação reside em usar conexões inter-documentos para identificar relevância por associação, capturando documentos que possuem conexões fracas com a consulta, mas fortes com outros documentos relevantes [Research Team and Google 2024].

### 3.3. Align-GRAG (USTC/Huawei)

Foca na fase de pós-recuperação para resolver a ”lacuna de representação”. Utiliza um mecanismo de alinhamento duplo guiado por uma ”cadeia de raciocínio”gerada por LLM. Realiza o alinhamento de nós (para poda de ruído) e o alinhamento de representação (aprendizado contrastivo) para projetar grafos e texto em um espaço semântico compartilhado [Lab Team and Huawei 2024].

### 3.4. G-Retriever (NUS/Notre Dame)

Otimiza a recuperação em grafos grandes. Reformula a recuperação como um problema de *Prize-Collecting Steiner Tree* (PCST). Os nós recebem "prêmios"por relevância e arestas têm "custos". O algoritmo busca um subgrafo conectado que maximize prêmios e minimize custos, mitigando alucinações e controlando o tamanho do contexto [He et al. 2024].

## 4. Comparação e Sinergias

A Tabela 1 sintetiza a comparação das metodologias analisadas, destacando características técnicas e problemas abordados.

**Tabela 1. Comparação de Metodologias de GraphRAG**

| Característica | GraphRAG (Microsoft) | G-RAG (UCLA/Google) | Align-GRAG (USTC/Huawei) | G-Retriever (NUS/Notre Dame) |
|---|---|---|---|---|
| **Fase do Pipeline de Foco** | Indexação por Sumarização e Geração via Map-Reduce | Reranking da lista de documentos | Refinamento do subgrafo recuperado | Recuperação (Retrieval) |
| **Problema Principal** | Resposta a perguntas abertas e compreensão holística. | Qualidade e relevância dos documentos recuperados. | Ruído no subgrafo e lacuna de representação grafo-linguagem. | Escalabilidade da QA em grafos grandes e alucinações. |
| **Técnica Chave** | Detecção de comunidades hierárquicas e sumarização. | Arquitetura de duplo grafo (AMR + Documentos) para relações inter-documentos. | Alinhamento duplo guiado por cadeia de raciocínio gerada por LLM. | Otimização via *Prize-Collecting Steiner Tree* (PCST). |
| **Tipo de Grafo** | Grafo de conhecimento auto-construído. | Grafo AMR e Grafo de Documentos. | Grafo de conhecimento genérico/textual. | Grafo de conhecimento ou textual genérico. |

Nota-se que cada metodologia busca resolver um gargalo diferente. A combinação da recuperação eficiente do **G-Retriever** com o refinamento pós-recuperação do **Align-GRAG** poderia compor o pipeline de um sistema ideal [Zhang et al. 2024].

## 5. Limitações e Proposta de Pesquisa

Apesar dos avanços mencionados, foram identificados desafios críticos que limitam a implantação do GraphRAG em larga escala.

### 5.1. Lacunas Identificadas

- **Ruído e Redundância:** A recuperação em grafos densos frequentemente traz nós adjacentes irrelevantes. Além disso, grafos gerados automaticamente por LLMs quase sempre contêm entidades duplicadas e relações espúrias.
- **Lacuna de Representação:** Embeddings de grafos (topológicos) e de texto (sequenciais) são fundamentalmente desalinhados. Abordagens atuais que apenas concatenam essas informações perdem riqueza semântica.
- **Paradoxo do Reranking:** LLMs em configuração *zero-shot* mostram desempenho inferior como rerankers de grafos, falhando em capturar o contexto relacional ao avaliar nós isoladamente.

### 5.2. Direções de Pesquisa Propostas

Objetivando empurrar a fronteira do estado da arte, propõe-se três eixos de pesquisa:

1. **Otimização Semântica do Grafo (Denoising):** Investigar técnicas de resolução de entidades e ”Reflexão de Triplas”(usando *LLM-as-a-judge*) para limpar grafos gerados automaticamente, reduzindo a redundância antes da indexação.
2. **Ajuste Fino de Rerankers com Consciência de Grafo:** Desenvolver metodologias de *fine-tuning* onde LLMs são treinados para avaliar a relevância de subgrafos inteiros, utilizando ”cadeias de raciocínio”de modelos maiores como rótulos de supervisão.
3. **Modelos de Geração Híbrida:** Criar arquiteturas que utilizem aprendizado contrastivo para alinhar os espaços vetoriais de grafo e texto, além de mecanismos de atenção cruzada (*cross-attention*) para fusão dinâmica de informações.

## 6. Conclusão

Ao término deste trabalho, (1) foram apresentadas soluções do Estado da Arte na área de GraphRAG, como o GraphRAG da Microsoft e o G-Retriever, (2) e observou-se que o campo ainda apresenta lacunas que carecem de tratamento adequado, como ruído estrutural e alinhamento de representações. A proposta de uma agenda de pesquisa visa voltar a atenção para a investigação de soluções para preencher tais lacunas, construindo um caminho para sistemas que raciocinem de forma confiável sobre o conhecimento complexo.

## Referências


Edge, D. et al. (2024). From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

He, X. et al. (2024). G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *arXiv preprint arXiv:2402.07630*.

Lab Team, U. and Huawei (2024). Align-grag: Aligning graph and textual representations. *IEEE Transactions on Knowledge Discovery*.

Research Team, U. and Google (2024). Augmenting rag with graph-based reranking. *Proceedings of NLP Conference*.

Zhang, Q. et al. (2024). Graph retrieval-augmented generation: A survey. *arXiv preprint*.