LEVERAGING MICROSOFT'S GRAPHRAG FOR MEDICAL LITERATURE REVIEWS: A GAME-CHANGER IN INFORMATION RETRIEVAL

With the rapid advancements in the field of AI and language models, the ability to retrieve and analyse vast amounts of information has become crucial. Retrieval-Augmented Generation (RAG) techniques were a step forward, allowing models to combine retrieval with generation, improving the precision of their responses. However, as we embarked on a journey of medical literature review on a set of research articles for a use case, we encountered limitations with traditional RAG models, which led us to explore **Microsoft's GraphRAG**—an innovative framework that integrates knowledge graphs with retrieval-augmented generation for superior information retrieval.

The use case is to build a solution for Pharmacovigilance teams who spend significant time in manually reviewing large volumes of scientific publications, to resolve the pressing issue of both slow process and increased the risk of human error. The extensive time required for these activities detracts from the team's ability to focus on more strategic tasks such as analysis and decision-making. Moreover, the constant influx of new data exacerbates the workload, leading to potential bottlenecks in ensuring timely and accurate safety reporting.

This article documents our experience using Microsoft's GraphRAG for reviewing medical research papers, specifically focusing on answering crucial questions such as identifying the **primary author**, determining whether **subjects are human**, and extracting the **casualties (side effects)** of drugs used in these studies.

What is GraphRAG?

GraphRAG, short for **Graph Retrieval-Augmented Generation**, is Microsoft's answer to overcoming the limitations of traditional RAG models. While RAG retrieves documents and generates responses based on the information in the document, **GraphRAG** introduces a structured approach, using **graph-based representations** to manage complex queries more effectively.

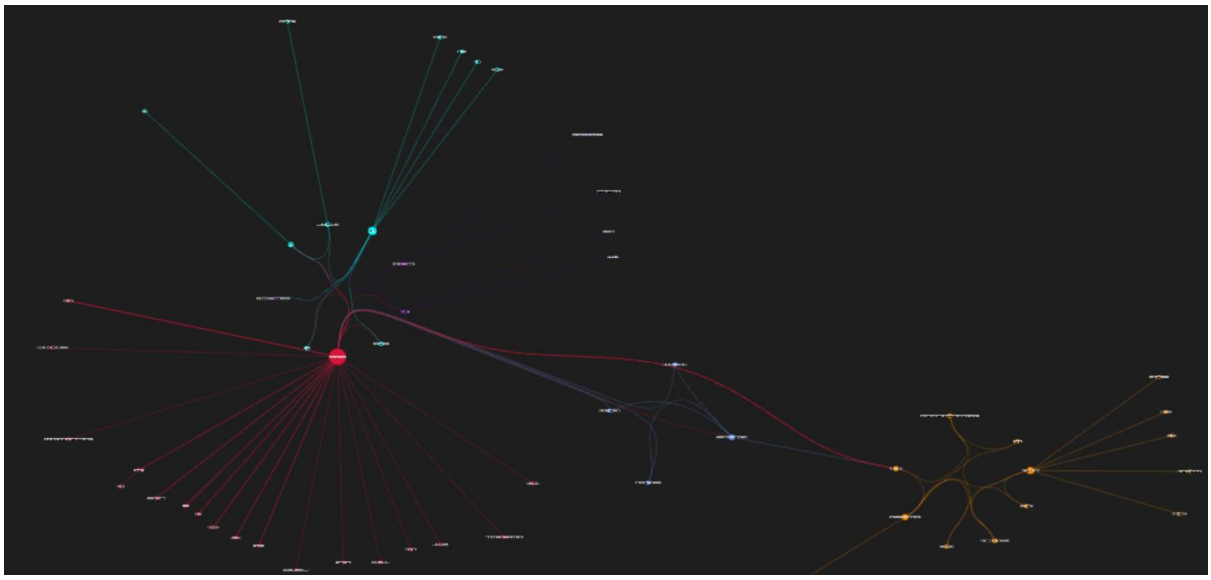

Fig 1.1: Knowledge Graph for a specific article, representing entities and their relationships.

GraphRAG leverages knowledge graphs by breaking down documents into entities and relationships, creating a detailed graph structure of the text. These graphs provide deeper insights by representing the connections between different data points, such as people, organizations, and concepts, allowing a model to understand and generate more contextually rich responses.

The core components of GraphRAG include:

1. **Entities** – Represent distinct concepts or individuals, such as authors or drugs.
2. **Relationships** – Define how entities interact with each other.
3. **Communities** – Cluster entities with similar or interconnected data points.
4. **Community Reports** – Summarized insights for each community.
5. **Text Units** – Chunks of the original text that are joined to entities and relationships for more effective retrieval.

Disadvantages of RAG:

In our initial efforts using the **RAG** system for my literature review, the retrieval was relatively straightforward. However, it became evident that RAG struggled with:

- **Ambiguous queries**: When tasked with questions like "What are the casualties (side effects) of the drug?" or "Who is the primary author?", RAG's simple retrieval mechanism often provided incomplete or irrelevant results.
- **Lack of reasoning**: Traditional RAG does not reason over relationships or connections in the text; it simply pulls the most relevant information, which leads to surface-level understanding without nuanced insights.
- **Global understanding**: For queries that required summarization or understanding of the entire dataset (e.g., "Who is the primary author and where are they from?"), RAG often missed the mark, failing to incorporate broader contextual information.

How GraphRAG Improved the Results:

GraphRAG transformed the review process, especially when it came to extracting more complex information like drug casualties or identifying authors from research articles. The steps it follows to create a structured knowledge graph helped in numerous ways:

1. **Entity and Relationship Extraction**:

   Instead of retrieving only pieces of the text, GraphRAG first **identifies entities** like authors, drugs, and patients. These entities are then linked by **relationships**—for example, connecting drugs to their side effects. This structured representation allowed to pinpoint the exact side effects associated with a particular drug much more efficiently.

Fig 2.1: Entities and their relationship in Knowledge Graph

In Figure 2.1, we observe a connection between the entity **"ALLOPURINOL"** and the entity **"DRESS (Drug Reaction with Eosinophilia and Systemic Symptoms)."** The established relationship is articulated as **"Allopurinol caused direct toxicity, contributing to the severity of DRESS in the patient."** This exemplifies how GraphRAG effectively creates entities and defines relationships, enhancing our understanding of their interactions within the data.

2. **Building Knowledge Graphs**:
   GraphRAG went further by building **community structures** from the data. These communities helped group related entities, such as drug side effects and the trials they were part of. Using algorithms like the **Leiden algorithm** for community detection, GraphRAG grouped related nodes into clusters, making it easier to retrieve comprehensive and accurate answers.


```
{"id":"1","title":"Community 1","level":0,"raw_community":"1","relationship_ids":["5c41f96be13e49dba649454297834546","7ea4afbf8a264f29af29950ce98105ba",
{"id":"2","title":"Community 2","level":0,"raw_community":"2","relationship_ids":["553b285bba60460ab1ed8341ae61282b","cec95bf17e7e4c939b56c9c6f402a29f",
{"id":"0","title":"Community 0","level":0,"raw_community":"0","relationship_ids":["3e95dacfe57b4d57b5da4310ef2e157f","1f1545308e9347af91fd03b94aadc21f",
{"id":"4","title":"Community 4","level":0,"raw_community":"4","relationship_ids":["aff21f1da1654e7babdcf3fb0e4a75fc","dc2cc9016e3f49dbac7232f05cce794d",
{"id":"3","title":"Community 3","level":0,"raw_community":"3","relationship_ids":["af1d0fec22114a3398b8016f5225f9ed","b07a7f088364459098cd8511ff27a4c8",
{"id":"5","title":"Community 5","level":1,"raw_community":"5","relationship_ids":["5c41f96be13e49dba649454297834546","7ea4afbf8a264f29af29950ce98105ba",
{"id":"6","title":"Community 6","level":1,"raw_community":"6","relationship_ids":["18b839da898e4026b81727d759d95c6a","eeef6ae5c464400c8755900b4f1ac37a",
```


Fig 2.2: Communities and Relationships in GraphRAG

3. **Summarization from Community Reports**:
   GraphRAG also generates **Community Reports**, which are natural language summaries based on the clusters of data. This not only improved the accuracy of the retrieved answers but also allowed to see a broad view of the dataset. For example, when looking for the **casualties (side effects)** of a drug, it could easily retrieve the summary of all reports on that particular drug.

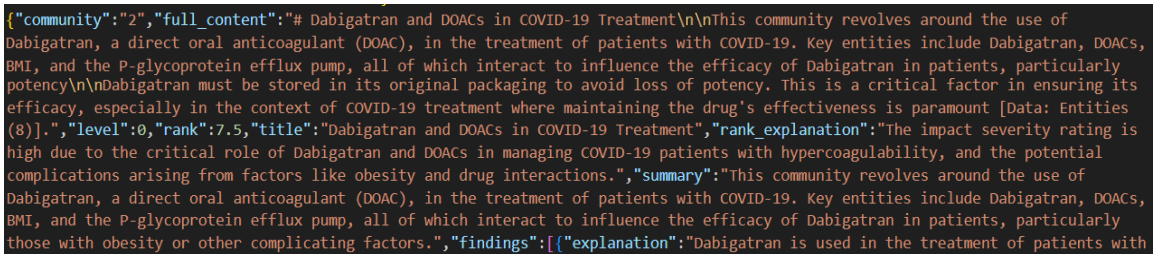
{"community":"2","full_content":"# Dabigatran and DOACs in COVID-19 Treatment\n\nThis community revolves around the use of Dabigatran, a direct oral anticoagulant (DOAC), in the treatment of patients with COVID-19. Key entities include Dabigatran, DOACs, BMI, and the P-glycoprotein efflux pump, all of which interact to influence the efficacy of Dabigatran in patients, particularly potency\n\nDabigatran must be stored in its original packaging to avoid loss of potency. This is a critical factor in ensuring its efficacy, especially in the context of COVID-19 treatment where maintaining the drug's effectiveness is paramount [Data: Entities (8)].","level":0,"rank":7.5,"title":"Dabigatran and DOACs in COVID-19 Treatment","rank_explanation":"The impact severity rating is high due to the critical role of Dabigatran and DOACs in managing COVID-19 patients with hypercoagulability, and the potential complications arising from factors like obesity and drug interactions.","summary":"This community revolves around the use of Dabigatran, a direct oral anticoagulant (DOAC), in the treatment of patients with COVID-19. Key entities include Dabigatran, DOACs, BMI, and the P-glycoprotein efflux pump, all of which interact to influence the efficacy of Dabigatran in patients, particularly those with obesity or other complicating factors.","findings":[{"explanation":"Dabigatran is used in the treatment of patients with

Fig 2.3: Community Reports in GraphRAG

4. **Improved Contextual Retrieval**:
   By joining **text units** to entity and relationship IDs, GraphRAG ensures that even when the source documents are chunked into small sections, the context is not lost. This was a significant improvement over RAG, where sometimes` qw2334 context was fragmented or omitted.


{"text_unit_ids":"42eb66c45c5538a45f83e341c289bb2c","entity_ids":["b45241d70f0e43fca764df95b2b81f77","077d2820ae1845bcbb1803379a3d1eae",
{"text_unit_ids":"45c30a6b9ef64376d1de7c27ddb3a4ca","entity_ids":["b45241d70f0e43fca764df95b2b81f77","077d2820ae1845bcbb1803379a3d1eae",
{"text_unit_ids":"8abe38e92a327d7d210ec57b0fe5a0f3","entity_ids":["b45241d70f0e43fca764df95b2b81f77","4119fd06010c494caa07f439b333f4c5",
{"text_unit_ids":"4df9743dc5eb4a5b2f9dea6cf8349287","entity_ids":["f7e11b0e297a44a896dc67928368f600","3d0dcbc8971b415ea18065edc4d8c8ef",
{"text_unit_ids":"74f95bebd9e7aad7f8901e0b1aa350e3","entity_ids":["f7e11b0e297a44a896dc67928368f600","96aad7cb4b7d40e9b7e13b94a67af206",
{"text_unit_ids":"9363095f00a49e101316fe0a1df25f9f","entity_ids":["f7e11b0e297a44a896dc67928368f600","1fd3fa8bb5a2408790042ab9573779ee",
{"text_unit_ids":"170f4f4a32da9860f0207033de0c8b97","entity_ids":["1fd3fa8bb5a2408790042ab9573779ee","273daeec8cad41e6b3e450447db58ee7",
{"text_unit_ids":"3600bacdeda7968fac925c1542961629","entity_ids":["4a67211867e5464ba5126315a122a8a","04dbbb2283b845baaeac0eaf0c34c9da",
{"text_unit_ids":"6998e8378c5cda8723d99bc885ae843e","entity_ids":["04dbbb2283b845baaeac0eaf0c34c9da","147c038aef3e4422acbbc5f7938c4ab8",
{"text_unit_ids":"5fe34a96e7661001559aa2f843c7935b","entity_ids":["17ed1d92075643579a712cc6c29e8ddb","fa3c4204421c48609e52c8de2da4c654",

Fig 2.4: Joining text units to entity IDs


{"id":"8abe38e92a327d7d210ec57b0fe5a0f3","relationship_ids":["fa14b16c17e3417dba5a4b473ea5b18d","bb9e01bc171d4326a29afda59ece8d17",
{"id":"42eb66c45c5538a45f83e341c289bb2c","relationship_ids":["7cc3356d38de4328a51a5cbcb187dac3","bef16fb5fd7344cca5e295b13ef3e0cd",
{"id":"9363095f00a49e101316fe0a1df25f9f","relationship_ids":["7ea4afbf8a264f29af29950ce98105ba","553b285bba60460ab1ed8341ae61282b",
{"id":"74f95bebd9e7aad7f8901e0b1aa350e3","relationship_ids":["91ff849d12b24574b0691dbddf44968b","d73c1f2fb3094d8dace42ad2a76e9a52",
{"id":"4df9743dc5eb4a5b2f9dea6cf8349287","relationship_ids":["ceadf262ef834e9ab146b20650912cae","7f65feab75424b53b24470d305ba331a",
{"id":"45c30a6b9ef64376d1de7c27ddb3a4ca","relationship_ids":["0fbcca3f17c649a08aea64b5a7d9ef36","482027a59f32484c9c44fd700615c1b6",
{"id":"170f4f4a32da9860f0207033de0c8b97","relationship_ids":["cec95bf17e7e4c939b56c9c6f402a29f","599164aead034bc19446efacc77554d2",
{"id":"3600bacdeda7968fac925c1542961629","relationship_ids":["aff21f1da1654e7babdcf3fb0e4a75fc","dc2cc9016e3f49dbac7232f05cce794d",
{"id":"6998e8378c5cda8723d99bc885ae843e","relationship_ids":["6ea0cef05f694dcea455478f40674e45","353d91abc68648639d65a549e59b5cf3",
{"id":"5fe34a96e7661001559aa2f843c7935b","relationship_ids":["86505bca739d4bccaaa1a8e0f3baffdc","a5e0d1644eb547ba9a5c3211aac4631a",

Fig 2.5: Joining text units to relationship IDs

Comparison of RAG vs. GraphRAG on Drug Casualty Extraction:

Query	RAG Response	Microsoft GraphRAG Response
What is address of Primary Author?	USA	University of North Carolina School of Medicine, Chapel Hill, USA
Are the patients human, if yes them mention them?	No human patients are mentioned in the provided text.	82-year-old female patient with a one-month history of worsening dyspnea on exertion and a one-week history of scant hemoptysis and fatigue, diagnosed with hydralazine-induced ANCA-associated vasculitis (AAV) leading to diffuse alveolar hemorrhage (DAH) [Data: Sources (2, 3, 4)].
What are the casualties (side effects) of the drug Hydralazine?	Hydralazine can cause anti-neutrophil cytoplasmic antibody (ANCA)-associated glomerulonephritis (AAGN) with a rapidly progressive clinical course mixed with extrarenal symptoms such as arthralgia, pleuropulmonary disease, and cutaneous vasculitis.	Hydralazine is associated with several severe adverse effects, primarily due to its potential to induce autoimmune reactions. The key adverse effects include: 1. **ANCA-Associated Vasculitis (AAV)**: Hydralazine can induce a specific form of AAV,

		known as hydralazine-induced AAV. This condition can lead to severe health issues, including pulmonary-renal syndrome (PRS) and diffuse alveolar hemorrhage (DAH) [Data: Entities (8, 32, 104); Relationships (21, 22, 20, 17)]. 2. **Pulmonary-Renal Syndrome (PRS)**: This severe condition involves both glomerulonephritis and DAH, associated with a high mortality risk. Hydralazine-induced AAV can lead to PRS, further complicating the patient's health status [Data: Entities (10, 109); Relationships (58)]. 3. **Diffuse Alveolar Hemorrhage (DAH)**: DAH is a serious complication characterized by bleeding into the alveoli of the lungs, requiring immediate medical intervention. Hydralazine-induced AAV can result in DAH, making it a life-threatening condition [Data: Entities (7, 11); Relationships (16, 52, 64)]. 4. **Acute Kidney Injury (AKI)**: Hydralazine-induced AAV can occur in the setting of AKI, indicating a severe impact on kidney function [Data: Entities (47, 46); Relationships (59, 65)].

Here is how GraphRAG Extracted Casualties Better than Basic RAG:

The key to GraphRAG's success over RAG lies in its structured approach. In traditional RAG systems, retrieval is based purely on pulling relevant text from a document without understanding the **relationships** between entities. As a result, RAG might pull information that is not directly relevant to the query, or worse, it may miss out on critical context.

GraphRAG, however, creates a **knowledge graph** from the document. Here's why this works better:

- **Entity Creation**: GraphRAG identifies specific entities like "Drug X" and "Side Effect Y" within the document, creating nodes for them.
- **Relationships**: It then forms connections, such as "Drug X" causes "Side Effect Y" in "Patients," allowing a much richer and more connected understanding of the content.
- **Summarization**: Using the community reports, GraphRAG can generate natural language summaries that describe the relationships in a structured and coherent manner, making it easier to retrieve detailed information about drug casualties.

In contrast, **RAG** would have simply retrieved fragments of the text related to "Drug X" and "Side Effect Y" without linking them meaningfully. The difference is stark when querying complex datasets like medical research articles, where understanding relationships is critical.

Conclusion:

Microsoft's **GraphRAG** represents a leap forward in retrieval-augmented generation by integrating knowledge graphs and community detection algorithms. For the medical literature review work, it could provide precise answers to complex questions that were beyond the capability of traditional RAG. By creating a graph-based representation of the text, it allowed for more accurate and comprehensive retrieval, transforming the way information was extracted like **drug side effects** and **author details** from research articles.

To know more about our work, please reach out to us for further queries at manasa.kasivajjula@deepforrest.ai and follow us on LinkedIn https://www.linkedin.com/company/deepforrest-ai/?viewAsMember=true