Literature Research

Dennis Shushack

Literature Research on Knowledge Graphs & LLM

GraphRAG by Microsoft

Sophisticated system designed by Microsoft research for very large datasets. Transforms raw data (unstructured) into structured and semi-structured hierarchies and communities.

Traditional Graph Systems

Normal RAG suffers from:

- Contextual depth
- Scalability in evolving datasets.

We can incorporate GraphRAG, when we want the LLM to provide answers that reflect deeper links and relationships within the data. This is useful if we want to handle large and evolving datasets efficiently. The secret sauce of GraphRAG is **community construction:**

1. **Entity Detection**: GraphRAG scans the dataset to identify and categorize entities.
2. **Relationship Mapping:** Examine the connections between these entities, mapping out the relationships that bind them.
3. **Community Clustering:** Group these entities into communities, that represent closely knit clusters of related information. i.e., a group of companies working in a certain space, etc.

The indexing Pipeline: e2e Knowledge Graph creation pipeline. (Indexing Dataflow, Medium)

Phase 1: Compose Text Units

In a first step some initial processing happens for the raw input data. A Text Unit is a chunk of text that is used for the graph extraction and is also employed as a source-reference. The size of the text chunk can be defined by the users (default 300 tokens). This chunk size determines the granularity of the Text Units. This helps capture enough detail inside the text without loosing the context. What is the impact of Chunk Size:

- **Larger Chunks:**
  - *Faster Processing Times:* number of text segments to process is reduced. Less time is needed to retrieve and process the information from multiple chunks.
  - *Context Preservation:* Larger chunks retain more context, which can beneficial for tasks that require a broader range of context.
  - *Potential loss of granularity:* Larger chunks keep more context, however might dilute the specific focus on fine-grained details i.e., subtle relationships. Also these provide less meaningful reference texts.

- **Smaller Chunks:**
  - *Improved Data Completeness:* Smaller chunks provide finer granularity, allowing the system to focus on more specific pieces of information.
  - *Detailed and Focused References:* Because smaller chunks concentrate on smaller sections of text, the extracted information is often more precise and relevant. This is especially important for entity extraction.
  - *Increased processing:* More chunks lead to more retrieval and processing operations (time & resources).

Grouping Configuration

Next, the GraphRag system aligns chunks to the document boundaries. As such, we end up with a **1-to-Many** chunk-document relationship. We thus have one chunk per document. This preserves the document's contextual integrity. For shorter documents we can adjust this to **Many-to-Many.**

Each chunk is then **text embedded** (an embedding is created for each of our text chunks).

The output for our document is as follows:

1. chunk: Chunk of the text in a specific size i.e., 200 tokens
2. chunk_id: Id of the chunk
3. document ids: [list of corresponding document ids.
4. n_tokens: the token size

Phase 2: Graph Extraction

Transforming TextUnits into a Structured Knowledge Graph
This transforms the raw TextUnits into a graph that we can query. In this phase we extract the primitives from the textunits:

- **Entities:** Represent people, places, events, or some other entity-model provided.
- **Relationships:** A relation between two entities. These are generated from the *covariates.*

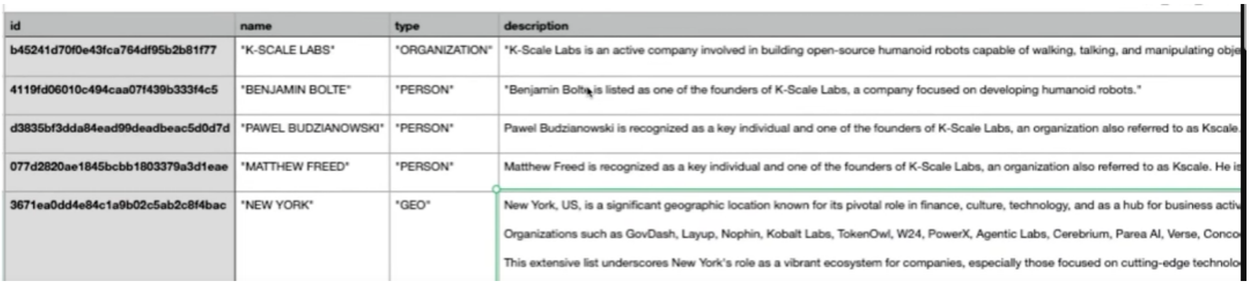
These are extracted using LLMs. Each textunit is processed in order to extract the entities & relationships out of the raw text. We end up with a sub graph for each textunit:

- TextUnit containing a list of e**ntities** with a *name*, *type*, and *description*
- list of **relationships** with a *source*, *target*, and *description.*

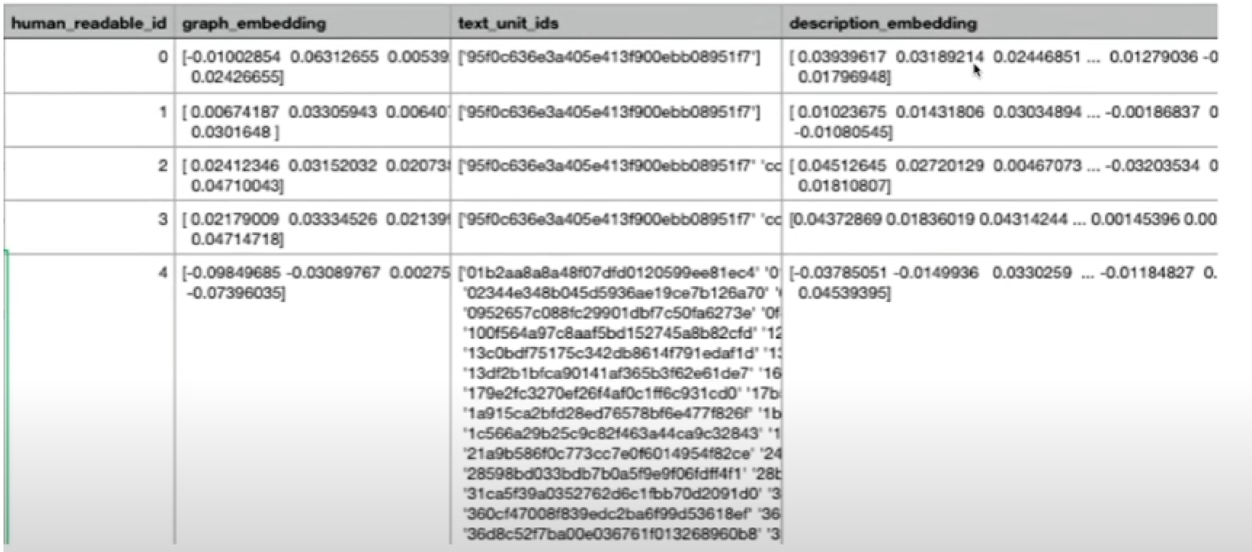
The sub-graphs are then merged together. Entities with the same *name* and *type* are merged by creating an array of their descriptions (Entity Resolution). The same happens for the relationships. Relationships with the same source and target are merged by creating an array of their descriptions.

The next step is entity & Relationship Summarization. We have a graph with entities & relationships, each containing a list of descriptions. We now summarize these lists into a single description per entity & relationship. This is achieved by prompting the LLM to create a summary of the descriptions. In the end each entity and relationship contains a single description.

Claims can be extracted as well but are left out by default.
This is an Entity:



id	name	type	description
b45241d70f0e43fca764df95b2b81f77	"K-SCALE LABS"	"ORGANIZATION"	"K-Scale Labs is an active company involved in building open-source humanoid robots capable of walking, talking, and manipulating obje
4119fd06010c494caa07f439b333f4c5	"BENJAMIN BOLTE"	"PERSON"	"Benjamin Bolte is listed as one of the founders of K-Scale Labs, a company focused on developing humanoid robots."
d3835bf3dda84ead99deadbeac5d0d7d	"PAWEL BUDZIANOWSKI"	"PERSON"	Pawel Budzianowski is recognized as a key individual and one of the founders of K-Scale Labs, an organization also referred to as Kscale.
077d2820ae1845bcbb1803379a3d1eae	"MATTHEW FREED"	"PERSON"	Matthew Freed is recognized as a key individual and one of the founders of K-Scale Labs, an organization also referred to as Kscale. He is
3671ea0dd4e84c1a9b02c5ab2c8f4bac	"NEW YORK"	"GEO"	New York, US, is a significant geographic location known for its pivotal role in finance, culture, technology, and as a hub for business activ Organizations such as GovDash, Layup, Nophin, Kobalt Labs, TokenOwl, W24, PowerX, Agentic Labs, Cerebrium, Parea AI, Verse, Conco This extensive list underscores New York's role as a vibrant ecosystem for companies, especially those focused on cutting-edge technolo



human_readable_id	graph_embedding	text_unit_ids	description_embedding
0	[-0.01002854 0.06312655 0.00539 0.02426655]	['95f0c636e3a405e413f900ebb08951f7']	[0.03939617 0.03189214 0.02446851 ... 0.01279036 -0 0.01796948]
1	[0.00674187 0.03305943 0.00640 0.0301648]	['95f0c636e3a405e413f900ebb08951f7']	[0.01023675 0.01431806 0.03034894 ... -0.00186837 0 -0.01080545]
2	[0.02412346 0.03152032 0.02073 0.04710043]	['95f0c636e3a405e413f900ebb08951f7' 'cc	[0.04512645 0.02720129 0.00467073 ... -0.03203534 0 0.01810807]
3	[0.02179009 0.03334526 0.02139 0.04714718]	['95f0c636e3a405e413f900ebb08951f7' 'cc	[0.04372869 0.01836019 0.04314244 ... 0.00145396 0.00
4	[-0.09849685 -0.03089767 0.00275 -0.07396035]	['01b2aa8a8a48f07dfd0120599ee81ec4' '0 '02344e348b045d5936ae19ce7b126a70' ' '0952657c088fc29901dbf7c50fa6273e' '0f '100f564a97c8aaf5bd152745a8b82cfd' '12 '13c0bdf75175c342db8614f791edaf1d' '13 '13df2b1bfca90141af365b3f62e61de7' '16 '179e2fc3270ef26f4af0c1ff6c931cd0' '17b '1a915ca2bfd28ed76578bf6e477f826f' '1b '1c566a29b25c9c82f463a44ca9c32843' '1 '21a9b586f0c773cc7e0f6014954f82ce' '24 '28598bd033bdb7b0a5f9e9f06fdff4f1' '28b '31ca5f39a0352762d6c1fbb70d2091d0' '3 '360cf47008f839edc2ba6f99d53618ef' '36 '36d8c52f7ba00e036761f013268960b8' '3	[-0.03785051 -0.0149936 0.0330259 ... -0.01184827 0. 0.04539395]