Analyzing the Impact of Knowledge Graph Hubs on Language Model Retrieval Quality

Miakov Timofey (timyakov@edu.hse.ru),
Shustrov Dmitrii (dmshustrov@edu.hse.ru),
Izmailov Ruslan (rnizmailov@edu.hse.ru).

In retrieval-augmented language models, external knowledge is used to enrich generation and understanding. *Knowledge graphs* (KGs) serve as structured sources for such external information. However, not all nodes in a KG contribute equally - some entities act as **hubs**, highly connected nodes that may dominate retrieval results. This project investigates how the presence of such hubs affects the quality of information retrieval and, consequently, the downstream performance of language models in knowledge-intensive tasks. The goal is to systematically analyze hub influence, develop hub-aware retrieval strategies, and propose adjustments to improve RAG method for QA benchmarks.

---

1. Introduction

This project explores the use of node class-aware retrieval in large-scale knowledge graphs to enhance the performance and efficiency of retrieval-augmented generation (RAG) for question answering. This research is important because large knowledge graphs often contain highly connected hub nodes that introduce noise and ambiguity during retrieval. By classifying nodes into hubs, semi-hubs, and peripheral nodes and tailoring retrieval strategies accordingly, the system can retrieve more relevant, specific context, leading to more accurate answers. This approach helps improve both the quality and scalability of knowledge-intensive NLP applications.

2. Literature Review

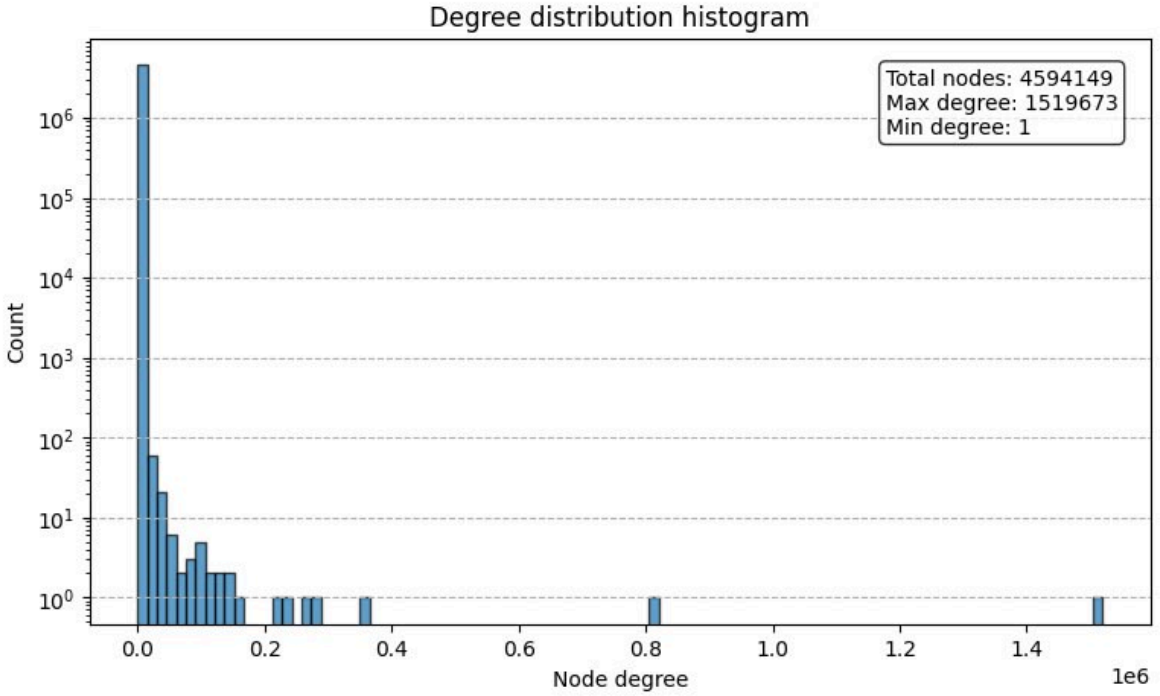
Retrieval-Augmented Generation (RAG) has emerged as a pivotal approach for enhancing language models by integrating external knowledge bases, effectively mitigating issues related to model hallucinations and limited factual accuracy. Recent research, such as Lewis et al. (2020), demonstrates that leveraging external retrieval mechanisms enables language models to access factual, contextually relevant information dynamically, thereby improving response accuracy and relevance. However, traditional RAG pipelines predominantly utilize dense or sparse vector representations for retrieval, which sometimes leads to suboptimal precision, particularly in complex information landscapes. To address these limitations, researchers have proposed integrating structured knowledge representations, notably through the use of knowledge graphs, as complementary retrieval sources.

In response, Graph Retrieval-Augmented Generation (GraphRAG) pipelines have been developed, combining the strengths of structured knowledge graph data with the flexibility of neural retrieval mechanisms. GraphRAG frameworks leverage entities and relational structures within knowledge graphs to refine retrieval processes, emphasizing "graph hubs"—nodes with high centrality and connectivity—to guide retrieval and generation tasks (Izacard et al., 2021; Yasunaga et al., 2021). The incorporation of graph hubs into retrieval pipelines notably enhances the quality of retrieved information, enabling the language model to generate more coherent and contextually accurate responses, especially in domains requiring precise fact retrieval.

3. Main part

3.1. Dataset and Knowledge Graph Selection

The *Wikidata5M* dataset is a large-scale knowledge graph dataset derived from Wikidata, containing approximately 5 million entities that are the texts of wikipedia articles and 20 million triplets. It provides structured information in the form of (subject, predicate, object) relations, covering diverse domains such as people, places, organizations, and scientific concepts. *Wikidata5M* includes textual descriptions of entities, making it suitable for RAG and GraphRAG approaches.



Img 1. The degree distribution histogram of the *Wikidata5M* exhibits characteristics of a *power-law distribution*, which is common in large-scale real-world networks such as knowledge graphs, the web, and social networks.

In the Wikidata5M dataset, hub nodes - entities with extremely high degrees - represent highly connected articles. These hubs create contention in retrieval-augmented generation

(RAG) pipelines because they are linked to a vast number of other nodes, leading to ambiguous or overly generic retrieval results during question answering. This hubs affects RAG pipeline in two main ways:

- Reduced Precision: Queries involving hub nodes often retrieve large, unfocused neighborhoods, diluting relevant information with noise.
- Inefficient Ranking: It becomes harder for the retriever to surface the most relevant facts, increasing the burden on the generator and reducing answer quality.

This research aims to improve retrieval efficiency and answer quality in large-scale knowledge graph-based systems by categorizing nodes into hubs, semi-hubs, and peripheral classes

3.2. Retrieval Experiment Setup and Evaluation

The experiment will evaluate the impact of node class-aware retrieval on the performance of a retrieval-augmented generation (RAG) pipeline for question answering. We will use an open-source large language model from the Hugging *Face Transformers* library and *Langchain* as framework for RAG system building.

The experiment pipeline consists of three main stages:

1. *Preprocessing*: The Wikidata5M graph will be parsed to compute degree-based metrics. Nodes will be categorized into hubs, semi-hubs, and peripheral nodes using scalable methods (e.g., approximate degree centrality and sampling-based PageRank).
2. *Indexing & Retrieval Setup*: Documents or triples associated with each node class will be indexed separately using a vector store such as FAISS. Experiments will be conducted with different vector stores setup: with all hubs, only with semi-hubs and without hubs at all.
3. *Question Answering with RAG*: For a benchmark set of natural language questions, the pipeline will retrieve context from the graph-aware index and pass it to the LLM for answer generation.

Analysis will include both quantitative and qualitative evaluations. Quantitatively, we will measure standard QA metrics such as F1 score, exact match, ROUGE and BERTScore (semantic comparison between embeddings of the model response and the reference one). Also we apply LLM based evaluation of answers and references. Comparisons will be made between baseline (class-agnostic) and node-aware configurations to assess performance gains.

3. Expected Results

This project is expected to improve retrieval precision and answer accuracy in RAG-based question answering by using node class-aware strategies. By handling hubs, semi-hubs, and peripheral nodes differently, we aim to reduce irrelevant retrieval, enhance contextual relevance, and improve efficiency. The approach should demonstrate better performance and scalability compared to class-agnostic methods.