2025 Semantic Web - Postgraduate Students

Reproducible Knowledge Graph Summarization Using Embedding - Based Semantic Clustering

Project Motivation

Knowledge Graphs (KGs) such as Wikidata are widely used but difficult to explore due to their size and complexity. Query-log–driven summarization methods provide compact, user-centric views of KGs by grouping entities into semantic categories and building quotient graph summaries.

Existing approaches rely on Large Language Models (LLMs) to generate semantic categories, which introduces high cost, non-deterministic behaviour, limited reproducibility, and challenges for systematic evaluation. This project investigates embedding-based semantic clustering using Sentence-BERT as a scalable and reproducible alternative.

Project Objectives

1. **Extract entities from SPARQL query logs.**
2. **Group entities using Sentence-BERT embeddings and clustering algorithms.**
3. **Construct quotient graph summaries from these groups.(SEMANTIC WEB PROJECT COMPLETED)**
4. (FOR THESIS) Compare embedding-based summaries with type-hierarchy and LLM-based approaches.
5. (FOR THESIS) Evaluate semantic quality, efficiency, and temporal stability.

Research Questions

1. Can embedding-based clustering match or outperform LLM-based categories in semantic coherence?
2. How stable are embedding-based summaries when tracking user interests over time?
3. What are the trade-offs between interpretability, scalability, and reproducibility?
4. Which clustering methods are best suited for knowledge graph summarization?

Project Tasks

1. Review background literature on Knowledge Graphs, SPARQL query logs, quotient graph summarization, and semantic embeddings.

2. Implement a pipeline to extract entities and labels from query logs.
3. Encode entity labels using Sentence-BERT and apply clustering techniques such as k-Means and HDBSCAN.
4. Construct quotient graph summaries based on clustering results.
5. Evaluate the summaries using intrinsic clustering metrics, efficiency measures, and optional human evaluation.
6. Analyse the temporal evolution of summaries across multiple query log snapshots.
7. Document the methodology, results, and limitations in a final report.

Expected Outcomes

1. A fully reproducible, embedding-based knowledge graph summarization system.
2. Empirical comparison between embedding-based, type-based, and LLM-based summarization methods.
3. Insights into the stability and scalability of semantic clustering for large knowledge graphs.
4. Well-documented code and experimental results suitable for further research or publication.

Skills Developed

1. Python programming and machine learning experimentation.
2. Natural language processing with transformer-based models.
3. Knowledge graph querying and analysis using SPARQL.
4. Unsupervised learning and clustering evaluation.
5. Research writing and reproducible experimentation.

Possible Extensions

1. Supervised or semi-supervised clustering approaches.
2. Domain-adaptive fine-tuning of Sentence-BERT.
3. Cross-lingual knowledge graph summarization.
4. Explainability techniques for embedding-based clusters.
5. Interactive visualization of quotient graph summaries.