Data Fusion Knowledge Graph: A Visual Analytics System for Uncertainty-Aware Data Integration

Grand Valley State University
College of Computing

Name: **Caleb Cheruiyot** Project Investigator: **Dr. Jiaxin Du**

Abstract

This project presents a knowledge graph system that integrates heterogeneous data with dynamic uncertainty modeling. Using AI-driven automation, intelligent search, and interactive visualization, this system enhances data analysis and decision-making. Tested on real-world datasets, it improves scalability and supports research in data-driven domains.

Introduction

Integrating diverse datasets is critical for advanced analytics but often involves slow, error-prone manual processes. Our knowledge graph system addresses these challenges by automating data extraction, modeling uncertainty, and providing interactive visualization tools.

Key Features:

- **Ontology-driven integration :** Standardizes diverse datasets for seamless fusion.
- **AI-powered automation:** Extracts information with reduced manual effort.
- **Uncertainty modeling:** Represents confidence levels in relationships.
- **Interactive visualization:** Enables efficient exploration and querying of data.

This system empowers researchers by improving accuracy and accessibility in data-rich environments.

Implementation Pipeline

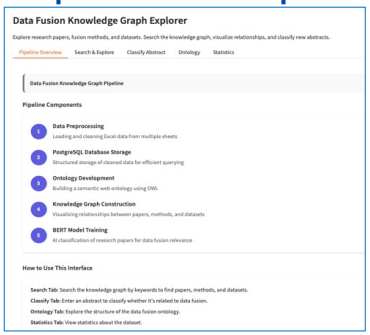

Fig 1: Pipeline Overview

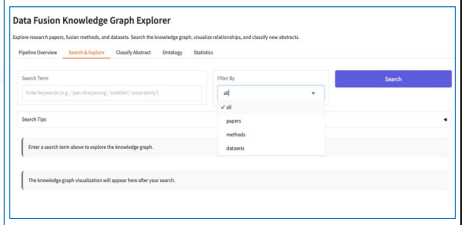

Fig 2: Search View

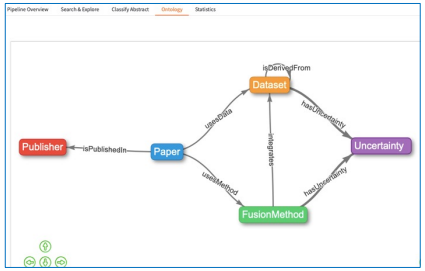

Fig 3: Data Fusion Ontology

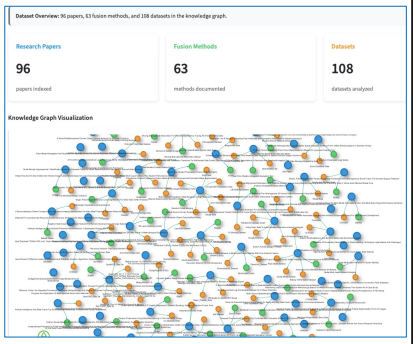

Fig 4: Data Fusion Knowledge Graph

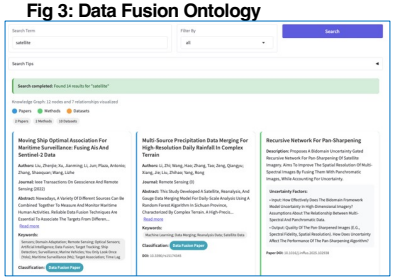

Fig 5: Results Searching by Keyword

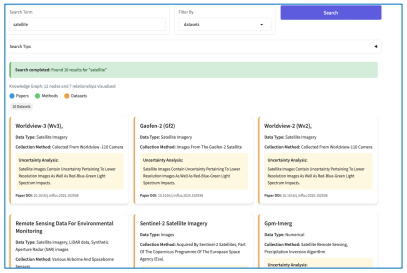

Fig 6: Results using Filter

Results Findings

- **Enhanced Automation & Accuracy:** The system significantly improves data retrieval automation and enhances decision-making accuracy through dynamic uncertainty modeling.
- **Unified Knowledge Fusion:** Successfully bridges research papers and datasets into a cohesive knowledge graph, facilitating cross-domain insights.
- **User-Friendly Interface:** Achieves high query success rates and fast response times, ensuring a seamless user experience.

Technical Details

- **Data Preprocessing:** Research papers and datasets are cleaned and preprocessed using **Python Pandas**.
- **Database:** Processed data is stored in a **PostgreSQL database** for efficient querying.
- **Ontology Development:** Semantic relationships are defined using an **OWL ontology layer**.
- **Knowledge Graph:** Built with **NetworkX**, linking entities and their relationships.
- **AI Classification:** A **BERT-based classifier** tags fusion-related papers.
- **User Interface:** A **Gradio UI** enables intuitive search, visualization, and graph queries.

Future Work

- **Real-Time Updates:** Integrate real-time scholarly databases such as Semantic Scholar and arXiv APIs to keep the knowledge graph current.
- **Ontology Reasoning:** Use OWL and SPARQL for inferring new knowledge and enabling complex queries.

Acknowledgment

We extend our sincere gratitude to **Timothy Mugambi, Nancy Odhiambo,** and **Braiden Betway** for their invaluable contributions in the research project. This work was supported by seed funding from the Grand Valley State University College of Computing.