Winter 12-11-2024

Big data-guided Knowledge Graph and its Visualization from Multi-Text Formats and Value-Added Interpretation

Vishnu Tripathi

Azad Singh

Shastri Nimmagadda

Neel Mani

Big data-guided Knowledge Graph and its Visualization from Multi-Text Formats and Value-Added Interpretation

Vishnu Tripathi
CAIR, DSVV
Uttarakhand, India
vishnu.tripathi@aicentre.org

Azad Singh
CAIR, DSVV
Uttarakhand, India
azad.singh@aicentre.org

Shastri Nimmagadda
Southern Cross University
Gold Coast, QLD, Australia
shastri.nimmagadda@aicentre.org

Neel Mani
CAIR, DSVV
Uttarakhand, India
neel.mani@dsvv.ac.in

Abstract

The research presents a novel method of creating and visualizing semantic graphs from multi-formatted big textual data using Large Language Models (LLMs). The proposed system leverages advanced natural language processing techniques with LLM capabilities to extract key concepts and relationships from text data visualization. The approach constructs comprehensive knowledge graphs and enables interactive investigation through URL redirection. Our approach involves multi-tiered processing: robust text extraction from various formats (PDF, DOCX, TXT), text segmentation into manageable chunks, and LLM-based analysis to identify key concepts and semantic relationships. The system integrates semantic and contextual relationships into a comprehensive Knowledge Graph, with added values, linking URLs to specific pages within the original inputs. The methodology automates knowledge extraction by enhancing the Knowledge Graph with detailed relationships and robust interactive features, representing a holistic approach. The paper discusses the system architecture, implementation details, potential applications, and future research directions and development.

Keywords: Big Data**,** Semantic Graph Construction, Knowledge Graph Visualization, Large Language Models, Natural Language Processing, URL Redirection, Value Creation

Introduction and Purpose of Architectural Design

In natural language processing (NLP) and text summarization, the representation and manipulation of textual data through advanced graphical models have garnered significant attention. Semantic graphs play a crucial role in natural language processing (NLP) by providing a structured representation of interconnected concepts and relationships, aiding in understanding human language (Bharambe et al. 2022; Prasad et al. 2022). Knowledge graphs (KGs) have gained significant attention in academia and industry for their ability to represent semantic relations between entities, proving particularly relevant for NLP applications (Schneider et al., 2022). These graphs are powerful tools for managing big data, enabling tasks such as knowledge management, information discovery, semantic search, and content-based recommendation systems (Prasad et al., 2022). These semantic hypergraphs are a powerful tool for encapsulating and summarizing the semantic content of large text datasets (Menezes and Roth, 2019). Accurately measuring the similarity between conceptual graphs is crucial for various NLP tasks, including information retrieval and text clustering (Hui and Jianjun, 2018). An extension to accommodate graphs with unbounded node degrees is also proposed, expanding the applicability of DAG automata in processing complex graph structures (Chiang et al., 2018).

In pursuit of advancing the field of intelligence extraction and visualization, we present an innovative methodology for construing and visualizing Knowledge Graphs from texts, incorporating redirection capabilities. Our approach leverages advanced natural language processing (NLP) techniques and the LLM to manage and interpret text from documents, ensuring comprehensive understanding and organization of information. A critical component of this research is the construction of semantic graphs from the dependency parsing of individual sentences. As depicted in Figure 1, our framework comprises multi-tiered tiers: robust text extraction from multi-format inputs (PDF, DOCX, TXT), segmentation of text into manageable chunks, and the analysis of these chunks using the LLM to identify key concepts and their semantic relationships. We integrate semantic and contextual relationships into a comprehensive Knowledge Graph by concatenating data frames and enhancing the data frame with URLs linking to specific pages within the original multi-format inputs. This enables direct access to the source material. This multi-tiered methodology automates knowledge extraction and organization, enriches the Knowledge Graph with detailed relationships, and provides interactive features, representing a significant advancement in the field and a powerful tool for researchers, analysts, and knowledge workers. Future enhancements will explore integrating additional data sources and expanding interactive features for more sophisticated queries and analyses.

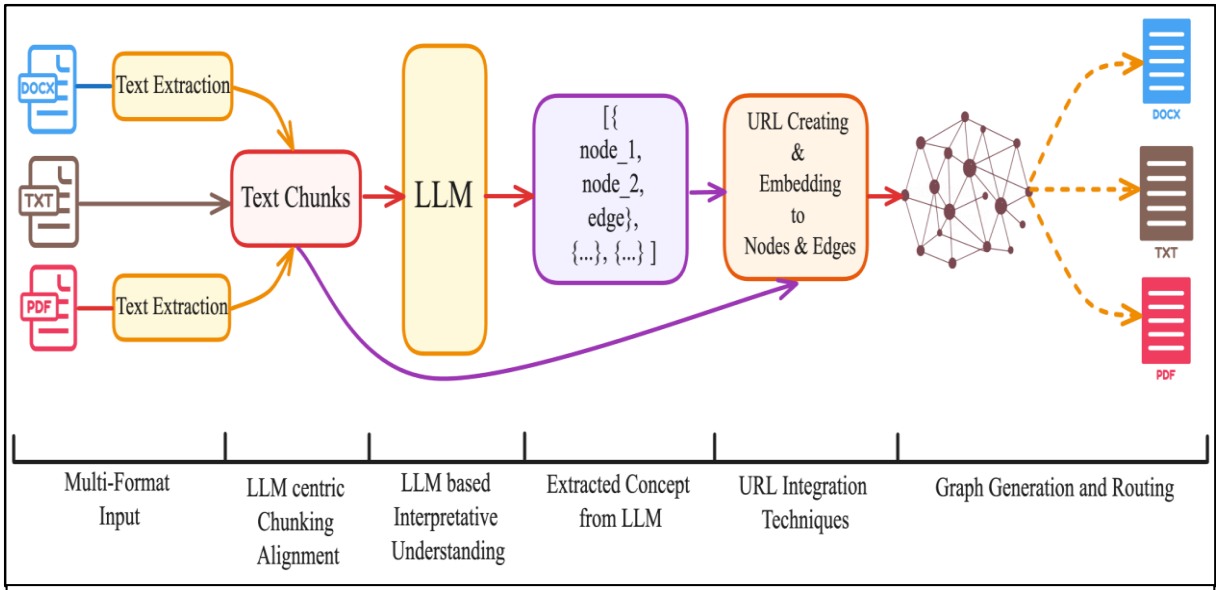

Figure 1: Large Language Model Driven Document Processing, Knowledge Graph Generation, and Adaptive Output Routing

In recent years, the need for advanced techniques in document analysis has grown significantly due to the vast amount of information stored in textual formats. One promising approach is using semantic graphs for the visual analysis of documents. Semantic graphs represent the semantic relationships within the text as directed graphs, providing a robust framework for data mining tasks such as exploratory data

analysis, data description, and summarization (Rusu et al. 2009a; Rusu, 2009b). This technique leverages natural language processing (NLP) to transform textual data into a structured format that can be more easily analyzed and understood (Rusu et al. 2009). Despite the advantages, constructing and visualizing semantic graphs from PDF texts presents several challenges. These challenges stem from the inherent complexity of natural language, the need for accurate entity extraction, and the preservation of semantic relationships in lower-dimensional spaces. Various methods have been proposed to address these issues. For instance, the Semaphore model introduces an unsupervised probabilistic approach that aims to maintain the manifold structure through neighbourhood regularization, which enhances the quality of semantic visualizations (Le and Lauw, 2016).

Additionally, the Doc2SoarGraph framework by (Zhu et al., 2023) integrates discrete reasoning capabilities to manage the differences and correlations among various document elements effectively. Moreover, Graph Style Sheets have been developed to facilitate filtering, grouping, and styling information elements through declarative transformation rules (Pietriga, 2006). These advancements contribute to the effectiveness of semantic graphs in various applications, including the Semantic Publishing Challenge, where text segmentation and entity extraction workflows are critical for linking detected entities to resources in existing open datasets (Sateli and Witte, 2015).

Significance and Motivation of the Architectural Design

LLMs are trained on unstructured Big Data and machine learning models to understand Big textual Data and generate human-interpreted language text. In addition to visualization, these models can analyze massive textual data that describe multiple languages in various formats. The LLMs can rapidly generate content, review, summarize, and easily translate text, creating interpretation values without human effort. For improving research students' academic skills and career development, the LLMs play a major role in detailing semantics and schematic values for professional documents.

Related Work

Various techniques for visualizing semantic graphs have been explored in the literature. One approach involves using tools and platforms designed explicitly for visualizing semantic web data, aiding in better understanding and analysis (Gupta and Malik, 2022). Another technique focuses on creating abstract semantic graphs highlighting path patterns and catering to users who need a deep understanding of the graph's structure (Leal, 2018). A novel Graph-based visual semantic entanglement Network has also been proposed to model visual features through a knowledge graph, outperforming existing methods in zero-shot learning tasks (Hu et al., 2020). These techniques offer diverse advantages, such as improved analysis, path pattern highlighting, and enhanced semantic modelling. Yet, they also come with limitations like complexity in visualization and potential trade-offs in displaying detailed content versus structural insights. Semantic graphs have shown significant potential impact across healthcare, finance, and marketing industries. In healthcare, semantic data integration using knowledge graphs aids in interrelating medical ontologies for disease prediction systems (Maghawry et al., 2023). This integration provides a reliable knowledge base visualization for intelligent healthcare systems, enhancing the analysis of therapeutic interventions. Besides, in the marketing sector, semantic graphs facilitate data transformation in motion, allowing for missing node predictions through machine learning and graph algorithms (Reimer, 2022). Semantic graphs present a dynamic and efficient approach to managing text data in various industries, leading to improved decision-making and problem-solving capabilities.

Evaluating redirecting capabilities in enhancing the visualization of complex semantic graphs is crucial for improving layout quality and performance. Various force-directed algorithms are utilized for visualizing graphs, focusing on mitigating challenges related to scalability and performance (Cheong and Si, 2018; Hussain et al., 2014). These algorithms aim to create aesthetically pleasing layouts by emphasizing power nodes, establishing local neighbourhood clusters, and applying semantic filtration to avoid cluttered views (Hussain et al., 2014). By considering these factors, researchers can assess the effectiveness of redirection capabilities in improving the visualization of intricate semantic graphs. Future developments in natural language processing (NLP) are poised to integrate redirecting capabilities for semantic graph visualization, enhancing tasks like semantic parsing and information retrieval (Collarana, 2017). By leveraging semantic graphs, NLP systems can optimize various tasks, improve search functionalities, and enable a deeper understanding of human language through ontologies (Bharambe et al., 2022). The integration of redirecting capabilities will allow for more interactive and dynamic visualization of semantic graphs, aiding in tasks like compositional generalization and weakly-supervised learning in semantic parsing (Petit and Corro, 2023). This advancement will enable users to interact with and explore complex semantic structures more

intuitively, potentially revolutionizing how information is processed and understood in NLP applications.

Research Method and Design of Overall Architecture

Qualitative research uses descriptive, exploratory, and explanatory methods in language modelling contexts. The techniques track information extraction, entity and dimension recognition, relation extraction, and event detection. Figure 1 represents the detailed workflow for extracting concepts and their semantic relationships from text chunks within large documents and enabling interactive redirection. It has many stages: Multi-Format Input, LLM-centric chunking Alignment, LLM-based Interpretative Understanding, Extracted Concept from LLM, URL Integration Techniques, Graph Generation, and routing.

Multi-Format Input

In the domain of graph generation and redirection, they involve handling and processing text inputs from various formats, such as DOCX, TXT, and PDF. The text is parsed for DOCX files using (*python-docx — pypi.org*, n.d.; *python-docx — python-docx 1.1.2 documentation*, n.d.) to preserve formatting like headings and styles. PDF files are handled using (*PyPDF2 — pypi.org*, n.d.; *Welcome to PyPDF2 — PyPDF2 documentation*, n.d.) for extracting text while managing layout complexities.TXT files are processed by directly reading text content. Character encoding considerations ensure text integrity, and pre-processing steps like cleaning non-textual elements are applied to maintain accuracy and consistency across all formats, preparing the extracted text for subsequent analysis and graph generation.

Model-Centric Chunking Alignment

Chunking alignment is facilitated using the long-chain library, configured with Chunk Size and Chunk Overlap parameters. This setup ensures efficient segmentation of text data into manageable chunks for subsequent LLM-based analysis. The LLM-centric format of the segmented text includes attributes like text, source, page, and chunk ID, as shown in Figure 2, essential for maintaining contextual integrity and facilitating interpretative understanding through natural language processing techniques.

	text	source	page	chunk_id
0	Foams Polymers and elastomers \nMetals Ceram...	MaterialScience/2_materials_charts_2009.pdf	0	7b7e124ebeb74e579126fba3ff5855ed
1	Foams Polymers Metals \nT echnica \nl ceramic...	MaterialScience/2_materials_charts_2009.pdf	0	8a9a7d186f314c048c045be7227fa40e
2	© Granta Design, January 2009 ...	MaterialScience/2_materials_charts_2009.pdf	2	2778cd05b06d4425b8a2531d3f7b0658
3	Table 1 Stiffness-limited design at minimum ma...	MaterialScience/2_materials_charts_2009.pdf	2	be76883da3a5437aa6664816159ed951
4	© Granta Design, January 2009 ...	MaterialScience/2_materials_charts_2009.pdf	3	aac6a6dfa51b40fba8dcd7b64ecb81c0

Figure 2: Language Model Centric Format

Model-Based Interpretative Insights

The segmented text chunks obtained from the previous processing phase are analyzed using an examined Large Language Model (LLM), such as GPT-4 or BERT, and a local model, such as LLAMA3 (Touvron et al., 2023), DSVV fine-tune model (*dsvv-cair/alpaca-cleaned-llama-30bbf16· Hugging Face — huggingface.co*, n.d.) and so on. The LLM performs a comprehensive contextual analysis of each text chunk, leveraging its advanced natural language processing (NLP) capabilities. Through this analysis, the LLM interprets the semantic content, identifying and extracting key concepts, entities, and their interrelationships. Techniques such as entity recognition, sentiment analysis, and topic modelling are utilized to derive valuable insights from the text. These insights are then systematically structured into nodes and edges, forming the foundational elements for subsequent graph representation. Using an LLM ensures high interpretative accuracy, capturing nuanced relationships and contextual dependencies within the text. This stage is critical in transforming unstructured text into a structured semantic graph, facilitating enhanced interpretability and navigability of the information.

Extracting Conceptual Value from Big Data-Guided Large Language Model

A pre-trained Large Language Model (LLM) is employed in this stage to extract and interpret critical concepts, entities, and relationships from the segmented text chunks, as shown in Figure 3. The process begins by analyzing each chunk to identify meaningful insights, then structuring it into nodes (representing entities, nodes, ideas, and events in the spatial domain) and edges (representing the relationships between these concepts). This information is organized into a coherent graph structure that reflects the semantic relationships derived from the text. To ensure accuracy, the graph is validated against the original text and refined by merging similar nodes and eliminating redundant edges. This method leverages advanced natural language processing capabilities to create a detailed and navigable semantic graph, enhancing the interpretability and accessibility of the extracted information and developing instances of semantic graphs.


```
[
  {
    "node_1": "School of Mechanical Engineering",
    "node_2": "Department of Mechanical Engineering",
    "edge": "The School of Mechanical Engineering and the Department of Mechanical Engineering are both components
of a larger educational institution."
  },
  {
    "node_1": "Unit",
    "node_2": "I",
    "edge": "In the context provided, 'Unit' refers to an organizational unit or course within the School of Mechan
ical Engineering, and 'I' is a specific number or label assigned to this unit."
  },
```


Figure 3: Language Model Output

Calculating Contextual Proximity

Contextual proximity is the assumption that concepts appearing close to each other in a text corpus are related. To calculate this, we first transform the data frame by collapsing node_1 and node_2 into a single column and then perform a self-join using chunk_id as the key, pairing nodes that co-occur within the same text chunk. We remove self-loops by dropping rows where node_1 is equal to node_2. The resulting data frame includes a count of co-occurrences and a list of chunk_ids where these occur. By merging this with a data frame capturing semantic relations, we create a comprehensive network graph. As shown in Figure 4, we combine the source of the specific node from the specific PDF documents and use the page column in the data frame, enhancing traceability. This approach provides a detailed and accessible visualization of the relationships between concepts within the text.

	node_1	node_2	page	source	edge	chunk_id	count
16	k'	mpa	355.0	MaterialScienceBooks/ASHBY99_Materials_Selecti...	relationship between the thermal expansion coe...	570b09af6f4b45e0b5d7b5866da18c9d,570b09af6f4b4...	2
44	$/kg	mpa	355.0	MaterialScienceBooks/ASHBY99_Materials_Selecti...	relation between the price of materials discus...	570b09af6f4b45e0b5d7b5866da18c9d,570b09af6f4b4...	2
64	$ut shapes (stamping)	shapes	277.0	MaterialScienceBooks/ASHBY99_Materials_Selecti...	production of $ut shapes through sheet-forming...	ccafac73faf8439197ca41de867bb9a4,ccafac73faf84...	3
65	$ut shapes (stamping)	sheet-forming processes	277.0	MaterialScienceBooks/ASHBY99_Materials_Selecti...	production of $ut shapes through sheet-forming...	ccafac73faf8439197ca41de867bb9a4,ccafac73faf84...	2

Figure 4: Contextual Proximity

URL Integration Techniques

This stage involves embedding URLs into the nodes and edges of the semantic graph, enabling seamless redirection to original content sources. The process begins by associating each extracted concept and its relationship with its corresponding section in the original document. Unique URLs are generated for these sections using techniques such as anchor tags for HTML documents or page-specific links for PDFs. These URLs are then embedded into the graph's nodes and edges, ensuring that each element in

the semantic graph can link back to its original context. This integration enhances user experience by allowing direct access to source materials, facilitating in-depth information exploration, and the verification method ensures that the semantic graph is a visual representation of relationships and a functional navigational tool that bridges the gap between summarized concepts and detailed original content.

Graph Generation and Routing

This stage transforms the structured nodes and edges into an interactive semantic graph visually representing the extracted concepts and their relationships. The process begins with graph generation tools like NetworkX or Graphviz to create a visual graph layout. Each node represents a concept, and each edge represents a relationship plotted to ensure clarity and coherence. Interactive elements are integrated into the graph, allowing users to click on nodes and edges to access embedded URLs, which redirect to the corresponding sections in the original documents. Routing capabilities are incorporated to enhance navigation, enabling users to trace paths between related concepts efficiently. This approach ensures that the semantic graph is a static representation and a dynamic tool that facilitates exploration and deeper understanding of the textual data. The final output is an intuitive, interactive graph that significantly improves the interpretability and usability of the information extracted from the documents.

Findings - Extraction of Structured Information, LLM Value Creation and Evaluation

Creating value from wide-ranging multimedia data is the focus of the current research. The authors have realized the complexity of the data, processing intricacies and interpretation ambiguities in real-world business contexts. Increased volumes of data, varieties and variabilities accumulated on different media platforms highlight the importance of sophisticated information extraction from unstructured data. Given the heterogeneity of unstructured data volumes, we present a Big Data-trained large language model framework for extracting structured information from unstructured data sources. The LLM framework generates text in a desired format, maintaining consistency and validating the semantic graph outputs. Information extraction, entity and dimension recognition, relation extraction and event detection are other valuable features interpreted from the LLM outputs. The framework is adaptable to corroborating various graph-structured data with model intrinsic knowledge, usability, and retrieval of new insights from big textual data.

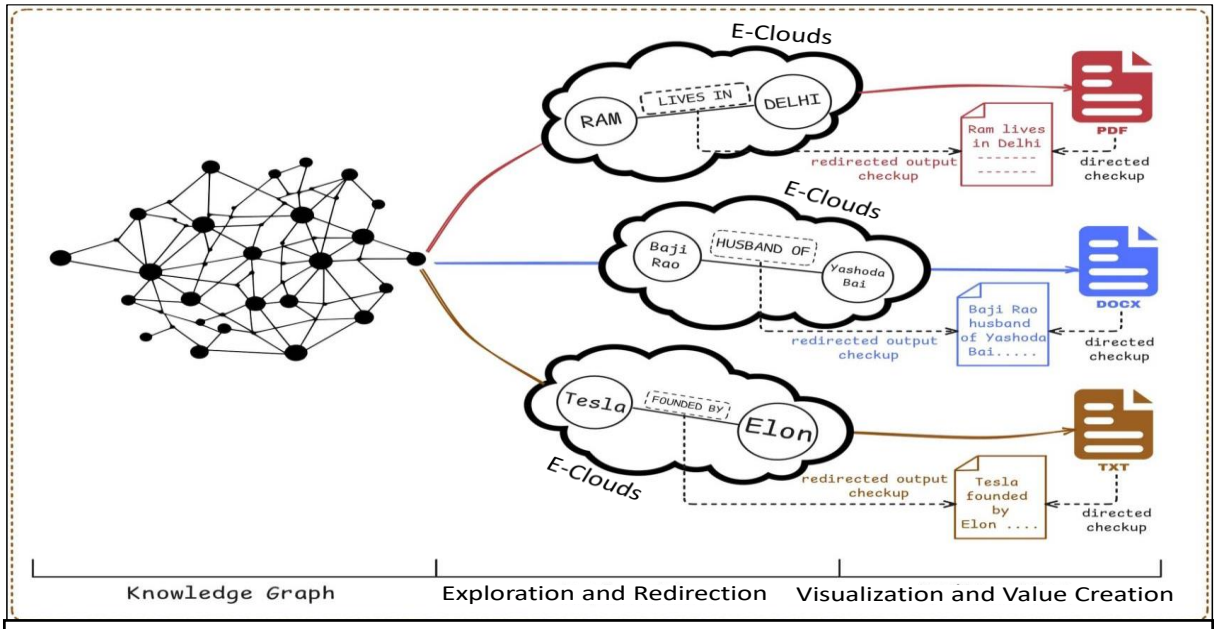

Figure 5: Knowledge Graph and LLM-guided Value Creation

Figure 5 presents the LLM output model, showing the realized events, which are represented through e-Clouds. The model depicts the knowledge graph assimilating the structured textual information. These facts are further validated through exploration and redirection features for final value creation. Several textual views provide values of the LLM model in Figure 5.

LLM guided Value Added Applications

Big Data guided semantic graph construction, visualization and value creation have promising applications, bridging numerous fields of study and industries. Some of the focused research applications include:

1. **Academic Literature Review:** The LLM framework can describe hundreds of academic papers during a literature review. The concepts of bibliographic drawings, methods employed, and findings of literature reviews can address textual data presentation concerns. The language models could significantly accelerate literature reviews and examine gaps in the existing knowledge.
2. **Legal Document Analysis:** In legal research, LLMs can be applied to examine case laws, statutes, and legal commentaries. The resulting semantic graphs can help lawyers and legal scholars present case information and present in hierarchical and relational databases. The legal entries highlight the evolution of legal terms and better understand the respective changes in legislation incorporated in the LLM models. The graphs can semantically extract many legal interrelations.
3. **Medical Research Integration:** In healthcare and medical research, the LLM framework can be applied to combine data from medical journals, clinical trials, and medical records. New connections between diseases, therapies, and their results may be extracted, which could make clinical research and practice easy.
4. **Policy Analysis and Development:** Government agencies and NGOs could use the LLM framework to evaluate new strategic insights from papers, opinion polls, and related research studies. The semantic graphs can guide decision-makers in comprehending policy dynamics, including relative stakeholder policies and related values.
5. **Patenting and Forecasting Research Trends:** The LLM may assist researchers in research and development (R&D) departments in analyzing patent-related documents, developing semantic maps that show the evolution of a specific technology, potential scopes and opportunities of value-added new technologies, and mapping such technologies across different industries.
6. **Marketing Promotions:** The LLMs are valuable tools for various marketing teams to quickly analyse customer sentiments and generate campaign ideas or political pitches during election times.

Research Implications and Limitations

LLMs can contribute to acquiring data and interpreting relationships, events, and anomalies between different textual data patterns, such as alphanumeric characters within words, sentences, and paragraphs. The language models can generate new text that is consistent and coherent with the concepts and contexts of the domain in which the text is supported. The LLMs can also translate the text into different formats. The LLMs have significance in empowering knowledge discovery and creating value from large volumes of unstructured data and information.

The research audience for creating Knowledge Graphs (KGs) is part of Natural Language Processing (NLP) teams and information retrieval researchers focused on extracting structured knowledge from big textual data. It engages data scientists and machine learning experts working on document parsing, knowledge extraction, and knowledge representation from a massive amount of unstructured material. Domain-specific healthcare, finance, and education researchers would benefit from automated KG generation to structure vast amounts of document-based data. This research is relevant to a broad audience seeking to enhance document understanding, find similarities between documents, and explore relationships, content organization, and semantic search across various fields.

This paper presents a process for creating Knowledge Graphs (KGs) that visualize the relationships between several documents. It makes it possible to convert unstructured data into structured knowledge and identify connections and links between different data sources. This paper provides meaningful information for semantic search, information retrieval, and multi-document knowledge discovery while improving the capacity to validate the connectivity of multiple documents, integrate the extracted data, and organize information from different sources. This paper offers a method for building interconnected Knowledge Graphs (KGs), making it easier to analyze and apply information in fields that require bringing together and connecting large amounts of data with a semantic meaning.

Conclusion and Future Work

This paper presents a novel approach for constructing and visualizing semantic graphs from multi-format texts using LLMs. The method integrates advanced NLP techniques with LLM capabilities to

extract concepts, build knowledge graphs, and enable interactive exploration through URL redirection. This system generates rich semantic representations that capture nuanced relationships within complex textual data visualizations.

Future work could focus on expanding input format support, exploring advanced LLM architectures, developing sophisticated visualization techniques, integrating with external knowledge bases, and adding values by user studies in various domains. These advancements would contribute to the evolution of semantic graph construction and visualization, transforming how we interact and derive insights from large volumes of textual information.

Acknowledgements

We are grateful to all individuals and organizations that helped make this research possible and to our colleagues and mentors at the CAIR for their help during the study process at the DSVV, Uttarakhand, India. We are thankful to the Vice Chancellor of the DSVV for permitting us to present and publish the work at the proceedings of the SIG BD, ICIS 2024.

References

Bharambe, U., Narvekar, C., and Andugula, P. 2022. Ontology and knowledge graphs for semantic analysis in natural language processing. In *Graph learning and network science for natural language processing* (pp. 105–130). CRC Press.

Cheong, S.-H., and Si, Y. W. 2018. Snapshot visualization of complex graphs with force directed algorithms. In *2018 ieee international conference on big knowledge (icbk)* (pp. 139–145).

Chiang, D., Drewes, F., Gildea, D., Lopez, A., & Satta, G. (2018). Weighted dag automata for semantic graphs. *Computational linguistics*, *44*(1), 119–186.

Collarana, D. 2017. A semantic integration approach for building knowledge graphs on-demand. In *International conference on web engineering* (pp. 575–583). *dsvv-cair/alpaca-cleaned-llama-30b-bf16 · Hugging Face — huggingface.co*. (n.d.). https://huggingface.co/dsvv-cair/alpaca-cleaned-llama-30b-bf16. ([Accessed 01-07-2024])

Gupta, R., and Malik, S. K. 2022. Visualizing semantic web data using various tools focusing rdf, owl and sparql. In *2022 11th international conference on system modeling & advancement in research trends (smart)* (pp. 1456–1460).

Hu, Y., Wen, G., Chapman, A., Yang, P., Luo, M., Xu, Y. Hall, W. 2020. Semantic graph-enhanced visual network for zero-shot learning. *arXiv preprint arXiv:2006.04648*, *4648*(2020), 1–10.

Hui, Z., Liyan, X., & Jianjun, C. (2018). The intensional semantic conceptual graph matching algorithm based on conceptual sub-graph weight self-adjustment. *International Journal of Computational Science and Engineering*, *16*(1), 53–62.

Hussain, A., Latif, K., Rextin, A. T., Hayat, A., and Alam, M. 2014. Scalable visualization of semantic nets using power-law graphs. *Applied Mathematics & Information Sciences*, *8*(1), 355.

langchain_text_splitters.character.RecursiveCharacterTextSplitter — #x2014; #x1F99C;#x1F517; LangChain 0.2.6 — api.python.langchain.com. (n.d.). https://api.python.langchain.com/en/latest/character/langchain_text_splitters.character.RecursiveCharacterTextSplitter.html. ([Accessed 01-07-2024])

Le, T. M., and Lauw, H. W. 2016. Semantic visualization with neighborhood graph regularization. *Journal of Artificial Intelligence Research*, *55*, 1091–1133.

Leal, J. P. (2018). Path patterns visualization in semantic graphs. In *7th symposium on languages, applications and technologies (slate 2018)*.

Maghawry, N., Ghoniemy, S., Shaaban, E., and Emara, K. 2023. An automatic generation of heterogeneous knowledge graph for global disease support: A demonstration of a cancer use case. *Big Data and Cognitive Computing*, *7*(1), 21.

Menezes, T., and Roth, C. 2019. Semantic hypergraphs. *arXiv preprint arXiv:1908.10784*.

Naik, N., Nimmagadda, S., Purohit, S., Reiners, T. and Mani, N. 2023. "Information System Articulation Development - Managing Veracity Attributes and Quantifying Relationship with Readability of Textual Data" (2023). *ICIS 2023 Proceedings*. 21. https://aisel.aisnet.org/icis2023/dab_sc/dab_sc/21.

Nimmagadda, S. L., Zhu, D., and Reiners, T. 2018. On Managing Contextual Knowledge of Digital Document Ecosystems, characterized by Alphanumeric Textual Data. *Procedia Computer Science*, *159*, 1135-1144. https://doi.org/10.1016/j.procs.2019.09.282.

Petit, A., & Corro, C. 2023. On graph-based reentrancy-free semantic parsing. *Transactions of the Association for Computational Linguistics*, *11*, 703–722.

Pietriga, E. 2006. Semantic web data visualization with graph style sheets. In *Proceedings of the 2006 acm symposium on software visualization* (pp. 177–178).

Prasad, J., Mulla, R., Naikwade, N., Kumar, B. S., and Shanmugasundaram, S. 2022. Ontology and knowledge graphs for natural language processing. In *Graph learning and network science for natural language processing* (pp. 131–146). CRC Press.

PyPDF2 — pypi.org. (n.d.). https://pypi.org/project/PyPDF2/. ([Accessed 01-07-2024]) *python-docx — pypi.org*. (n.d.). https://pypi.org/project/python-docx/. ([Accessed 01-07-2024]) *python-docx — python-docx 1.1.2 documentation*. (n.d.). Retrieved from https://python-docx.readthedocs.io/en/latest/

Reimer, T. (2022). Semantic transformation utilizing graphs. In *2022 international conference on advanced enterprise information system (aeis)* (pp. 138–145).

Rusu, D., Fortuna, B., Mladenic, D., Grobelnik, M., and Sipoš, R. 2009. Document visualization based on semantic graphs. In *2009 13th international conference information visualisation* (pp. 292–297).

Rusu, D., Fortuna, B., Mladenić, D., Grobelnik, M., and Sipoš, R. 2009. Visual analysis of documents with semantic graphs. In *Proceedings of the acm sigkdd workshop on visual analytics and knowledge discovery: Integrating automated analysis with interactive exploration* (pp. 66–73).

Sateli, B., & Witte, R. (2015). Automatic construction of a semantic knowledge base from ceur workshop proceedings. In *Semantic web evaluation challenges: Second semwebeval challenge at eswc 2015, portorož, slovenia, may 31-june 4, 2015, revised selected papers* (pp. 129–141).

Schneider, P., Schopf, T., Vladika, J., Galkin, M., Simperl, E., and Matthes, F. 2022. A decade of knowledge graphs in natural language processing: A survey. *arXiv preprint arXiv:2210.00105*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Welcome to pypdf2 — pypdf2 documentation. (n.d.). Retrieved from https://pypdf2.readthedocs.io/en/3.x/

Zhu, F., Wang, C., Feng, F., Ren, Z., Li, M., and Chua, T. S. (023. Doc2soargraph: Discrete reasoning over visually-rich table-text documents with semantic-oriented hierarchical graphs. *arXiv preprint arXiv:2305.01938*.