Exploration of Connections in Biomedical Research through Knowledge Graphs with Neo4j

1st Ali Barfi Bafghi
Mathematics & Systems Engineering
Florida Institute of Technology
Melbourne, FL
abarfibafghi2022@my.fit.edu

2nd Joshua Breininger
Computer Engineering and Sciences
Florida Institute of Technology
Melbourne, FL
jbreininger2018@my.fit.edu

3rd Dhruthi Sridhar Murthy
Computer Engineering and Sciences
Florida Insitute of Technology)
Melbourne, FL
dsridharmurt2022@my.fit.edu

4th Aristotelis Dougales
Computer Engineering and Sciences
Florida Institute of Technology)
Melbourne, FL
adougalis2022@my.fit.edu

5th Malakai Spann
Computer Engineering and Sciences
Florida Institute of Technology
Melbourne, FL
mspann2019@my.fit.edu

6th Candice Chambers
Electrical Engineering and Computer Science
Florida Institute of Technology
Melbourne, FL
chambersc2017@my.fit.edu

7th Thu Thu Hlaing
Mathematics & Systems Engineering
Florida Institute of Technology
Melbourne, FL
thlaing2022@my.fit.edu

***Abstract*—A challenging aspect of studying large amounts of information and data is that it is frequently not intuitive to search through. Data points are related by some kind of shared piece of data, typically keys, and what those relationships represent are often left implied. While experts in big data management and database systems may possess an intuitive understanding of the database contents and their interconnections, those such as researchers and other interested parties may find navigating this information challenging. As such, finding a way to represent the data in an understandable format can allow for an easier way of study and research. In this paper, we investigate this process using SemMedDB as an illustrative example—a medical literature database comprising specific words, associated predicates, and contextual sentences This database could be a useful source of information for medical researchers, however the data is stored in massive tables with an outdated schema, creating a barrier to entry. To help represent the database in an understandable format we selected three important tables - entity, predicate, and sentence. Subsequently each data base was uploaded to the database graphing software Neo4J, in order to visualize the data and the relationships between the three tables and in allow the database to be more accessible for research. Sample queries were then used to show the underlying structure of the database and to reveal how they can be used to access information from the database and display it in a visual manner. With the aid of the graphical database, researching SemMedDB is expected to be more straightforward and intuitive. The creation of a visualization opens avenues for extensive future research and analysis opportunities.**

***Index Terms*—Semantic Relations, SemMedDB, SemRep, Knowledge Graph, Neo4J**

INTRODUCTION

This research focuses on harnessing the vast repository of biomedical knowledge embedded in literature to facilitate knowledge discovery and exploration. The study utilizes the Semantic MEDLINE Database (SemMedDB), a comprehensive resource containing information automatically extracted from medical literature. In addressing the challenge of managing the overwhelming volume of published biomedical literature, we employ advanced natural language processing (NLP) techniques, particularly leveraging the Unified Medical Language System (UMLS) and the SemRep NLP system.

The Unified Medical Language System (UMLS), established by the National Library of Medicine, provides a robust foundation by encompassing a suite of files and software designed to offer an extensive array of health and biomedical vocabularies. SemRep, a natural language processing system developed by the U.S. National Library of Medicine, is a pivotal component of our research. It is designed to extract semantic relations from biomedical texts and is built upon the rich biomedical domain knowledge contained in the UMLS. SemRep extracts semantic relationships in the form of subject-predicate-object triples, mapping distinct concepts from the UMLS Meta thesaurus and utilizing the Semantic Network to intricately link these entities together.

Our investigation involves data from the SemMed database, a repository rich with information gathered automatically from medical literature. The database comprises tables, including the Sentence table containing information about the sentences in the literature, the Entity table detailing specific entities/nouns, and the Predication table representing the verbs applied to entity nouns, indicating relationships. This structured approach allows us to uncover complex relationships and connections within the biomedical knowledge space.

Despite the immense potential, the scale of data in the SemMed database posed challenges for local storage, prompting a reduction in data size for the purposes of our local project. This reduction involved careful selection and processing of data to ensure it remains manageable for analysis without compromising the integrity of the insights derived.

In the subsequent sections of this paper, we delve into the methodology employed, including the schema definition, data integrity establishment, text normalization, and feature engineering. We explore how graph algorithms and visualization tools, such as Neo4j Bloom, facilitate exploratory analysis and interactive exploration of the biomedical knowledge graph constructed from SemMedDB data. The insights extracted from this knowledge graph are interpreted, documented, and reported, with a focus on continuous refinement based on feedback. The paper concludes by discussing the deployment of the solution and the maintenance of the database for regular updates, aiming to communicate valuable insights to both the research community and healthcare professionals.

Through this research, we present a comprehensive approach to leveraging SemMedDB and SemRep for knowledge discovery, with the potential to bridge connections between various research topics, clinical experiments, and case studies, thereby contributing to advancements in biomedical research. Toward that end, the study aims to address the following Research Questions:

1) **RQ1**: How can we employ the use of knowledge and relational graphs to explore connections among concepts, treatments, and diseases discussed in various published biomedical literature?
2) **RQ2**: How can the creation of these knowledge graphs be used to optimize the retrieval of information in biomedical research and the communication of complex biomedical data to the general public?

LITERATURE REVIEW

In various academic disciplines, a common issue revolves around handling the overwhelming volume of published literature, leading to a focus on text-based information management research. A given example in the biomedical domain includes MedLine; a bibliographic database that is rich with information to the point where query results will often include hundreds of citations [1]. One of the ways to mitigate this issue is to employ natural language processing techniques to mine and extract key biomedical concepts, as well as their relationships using the Unified Medical Language System database. [5].

Unified Medical Language System

The Unified Medical Language System (UMLS), established by The National Library of Medicine in 1990 and consistently updated each year [3], encompasses a suite of files and software designed to offer an extensive array of health and biomedical vocabularies. This aims to enhance the interpretability of these concepts for computer systems, healthcare professionals, and the general public [2]. The UMLS is structured into three key knowledge sources: the Metathesaurus, the Semantic Network, and the Specialist Lexicon, each serving distinct purposes [9]. The Metathesaurus functions as a repository of terms, vocabularies, and relationships; the Semantic Network consists of 135 broad categories and 54 semantic relationships; and the Lexicon is a comprehensive syntactic collection of biomedical vocabularies [9].

SemRep

SemRep is a natural language processing system that was developed by the U.S. National Library of Medicine with the purpose of extracting semantic relations from biomedical texts. It is primarily built upon the biomedical domain knowledge contained in the Unified Medical Language System (UMLS) [4]. SemRep extracts semantic relationships from text in the form of subject-predicate-object triples. This process involves mapping distinct concepts from the UMLS Metathesaurus, utilizing MetaMap to designate them as the subject and object within the triple. The predication, drawn from the Semantic Network, then intricately links these entities together [8]. For example, if given the sentence shown in (1), Sem Rep will produce the following predication (2) [7]:

(1) .. fish oils can protect against coronary heart disease
(2) Fish Oils [Pharmacologic Substance] **PREVENTS** Coronary heart disease [Disease or Syndrome]

SemRep is able to map fish oil and heart disease in their respective categories when referring back to the biomedical concepts provided in the Metathesaurus, while also replacing the original verb ”protect against” with an appropriate predication of ”prevents” provided by the Semantic Network. While, the primary focus of SemRep has been on literature from PubMed, but it has also been applied to clinical narratives [6]. Previous literature has shown that SemRep has successful results when applied to clinical notes due to the fact that while clinical notes are full of medical concepts, there are not always clear relationships that can be detected simply from the notes. SemRep adeptly addresses this challenge by articulating the semantic relations between medical entities, thereby bridging a crucial gap. SemRep has access to a wide variety of predicates relating to choose from whether it is for clinical medicine (e.g TREATS, DIAGNOSES, PROCESS OF), molecular interactions (e.g., INTERACTS_WITH, INHIBITS, STIMULATES), disease etiology (e.g., ASSOCIATED_WITH, CAUSES, PREDISPOSES), pharmacogenomics (e.g., AFFECTS, AUGMENTS, DISRUPTS), or orstatic relations (ISA, PART_OF, LOCATION_OF) [5]. Any one of these could appropriately determine the relationship between the medical entities in the clinical notes. Health professionals can also leverage this information for various applications, such as analyzing the connections between specific medications and disorders [6].

Other work done with SemRep, as well as the inspiration for this project, was the use of knowledge graphs to display the information collected by SemRep from biomedical literature. Adopting a knowledge-rich abstraction approach, which builds upon the semantic predication offered by SemRep, presents a compelling alternative for interpreting and digesting literature [1]. Within this abstraction approach, SemRep efficiently condenses the abstract of a research article into semantic predications, which are then transformed and consolidated into a list. This list is subsequently visualized in the form of a knowledge graph, revealing all the interconnections and relationships among the discussed concepts within the abstract [1]. Departing from the practice of creating knowledge graphs for individual research papers, our proposed approach involves constructing a comprehensive knowledge graph that encompasses medical concepts derived from diverse literature sources, provided to us by Semantic Medline Data Base (SemMedDB). This broader perspective not only facilitates the identification of new connections but could also serves as a bridge between various research topics, clinical experiments, and case studies.

Proposed Approach

Data Description

SemMed contains data gathered automatically from medical literature. The SemMed database contains many tables, of which we used only three. The overarching connective table is the sentence table, which contains information about the actual sentence from the piece of literature which originates the information in the other tables. Each sentence has its own ID, a PubMed identifier for the publication from which it came and a "type" column that contains an identifier stating if it is in the title of the publication or the abstract. There then is a number representing the location of the sentence in the title or abstract, with indexes representing the actual character start and end locations of the sentence. Finally, a string of the actual sentence itself is stored. The next table, the Entity table, contains information about specific entities/nouns in the sentences. This makes it naturally the largest table being worked with. Each entity has an ID, the ID to the sentence it originates from, a CUI identifier, a name, semantic type, the text that maps to the entity, the character start and end indexes from the source, and a confidence score. Some examples of an entity are "diagnosis," "oral bacteria," and various medicines. The last table examined in the project is the prediction table, which represents the verbs being applied to the entity nouns. Possible predictions could be stating that a noun "cures" or is "cured by." This would expose possible information about interactions medicines have, implications on what medicines might cure, what diseases might be curable with what medicine, and more. Each predication contains an ID, an ID to its source sentence much like the entity table, the PubMed ID of the publication, the predicate itself, the name, CUI, semantic type, and the novelty of the subject the predicate is referring to. There then are the same attributes from the subject to the object, forming a small object predicate subject sentence.

There are a total of 44 GB of 1,887,317,669 rows in the Entity table, 3 GB of 126,268,045 rows in the predication table, and 16 GB of 253,029,872 sentences. This amount of data was too much for the purposes of the local project without the ability yet to store on a server, so it was reduced heavily.

Methodology

The data was briefly cleaned before being stored, then uploaded to Neo4J. Empty string and integer fields were filled with "Not Available" and "-1" respectively to avoid empty data. Additional column-based corrections were applied due to a mismatch in database version and the schema provided for the SemMed database. The sentence column appeared to actually be stored in a different column, the "SENT_END_INDEX" column, with what was the sentence column containing some unknown data with very few rows containing anything. There also was a column called "NORMALIZED_SECTION_HEADER" which similarly contained very little data - both of these columns were dropped as they provided little use.

The data was uploaded to the community version of Neo4J using python code to send the queries. Previously, all original files were converted to parquet format, compressed, and stored in AWS S3 buckets due to their size. They were gathered from the buckets, and each row for each table was taken, the columns extracted, and sent using the Cypher driver for Neo4J. Any rows where the ID's were invalid during the cleaning process were not sent. Neo4J is a graph database, and thus, relationships were necessary to build. The main relations in the data were provided via the sentence foreign keys in the predication and entity tables, which point to their respective sentence with the sentence IDs. We approached this by creating Neo4J relations between the three tables, where the entity has the "SUBJECT OF" relationship to sentences, and predication "PREDICATES" an entity, making three one way types of relationships between items in the overall database.

Results and Discussion

It was discovered during our procedure that uploading the complete data set to our local Neo4j database was not practical. The data must be uploaded to a dedicated server that hosts the database because external clients cannot access the local server. Instead, using a small subset of the actual data, we aim to confirm that the software automates the process of creating nodes and relations as predicted. Each dataset's initial partition was used to generate and carry out the Cypher commands that establish nodes and the connections between different kinds of nodes. Following every stage, we'll confirm our findings. After that, our procedure focused on uploading 400,000 nodes from each data set—Predicate, Sentences, and Entity. The entity dataset took the longest to upload, as can be seen in I.

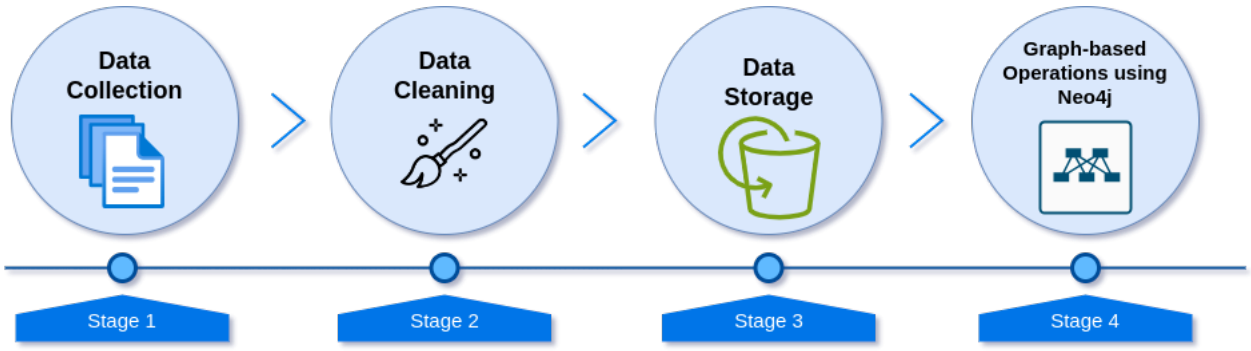


Fig. 1. Procedural Workflow of our Proposed Methodology.

Dataset	Number of Nodes	Time Taken (s)
Predication	400,000	4341.93
Sentences	400,000	3443.85
Entity	400,000	5324.15

TABLE I
TABLE SHOWING THE TIME TAKEN TO UPLOAD THE TABLE FOR EACH DATA SET

Query Analysis

Initially, we collect a list of distinct subject semantic types by querying every predication node that is accessible. Subsequently, we design a query that returns all nodes (prediction, sentence, and entity) that are connected to a single sentence by choosing two semantic types. We had to perform some background work and search the database for groups of prediction nodes that connected to a single sentence node in order for this to function. We concentrated on the predication nodes because they have more of a ”some-to-one” relationship with the sentence nodes than the entity nodes, which have a ”many-to-one” relation (where ”many” is an understatement). The semantic types ”hlca,” which stands for ”Health Care Activity,” and ”menp,” which stands for ”Mental Process,” were selected as two appropriate candidates for this query.

Finally, we construct and execute a query that returns all nodes related to the same sentence and connected to predications nodes with a subject semantic type of ”menp” or ”hlca.” In just 166 milliseconds, this query was completed.

All entity nodes were connected to the sentence node using the ”ENTITY_OF” relation, all predication nodes were connected to the sentence node using the ”PREDICATION_OF” relation, and all predication nodes were connected to all entity nodes using the ”PREDICATES” relation. The end result was a single interconnected graph. The phrase that this graph illustrates is, ”In conclusion, systematic SBN contact was uncommon in this population-based sample but positively influenced women’s perceptions of care, particularly in relation to the provision of support.” As a result, this sentence emphasizes the relationship between the mental process and the semantic types of healthcare activity. Shown in Fig. 1 is how the data looks with their relationships as a sample. Multiple entities and predicates are connected to singular sentences, creating connected bundles describing a sentence.

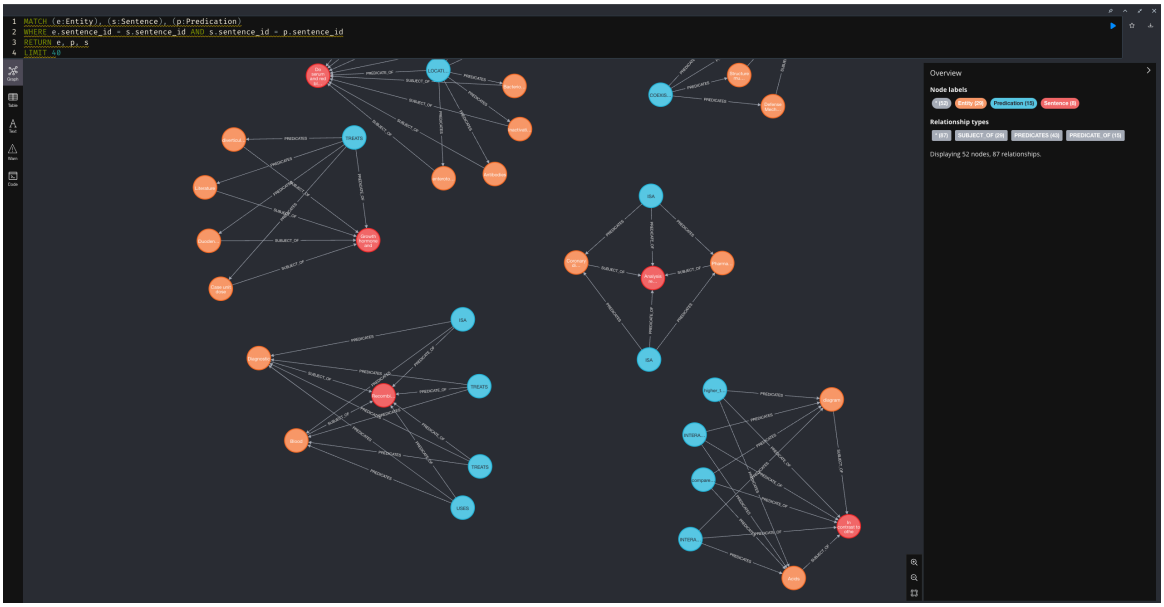
Research Question 1

We can construct comprehensive knowledge graphs using SemRepDB, which contains structured knowledge extracted from biomedical literature. These graphs represent complex relationships between biomedical concepts, treatments, and diseases. By using queries, we simplify the exploration of these intricate connections. These queries are designed to navigate the vast amount of information in SemRepDB, effectively making it easier for the general public and biomedical professionals to use relational graphs.

The ease of use brought by queries is crucial. They enable users to quickly identify specific relationships or trends within the biomedical domain. For example, a query could be structured to reveal how a particular treatment is linked to different diseases or to highlight the evolution of a specific biomedical concept over time.

Research Question 2

Neo4j is a popular graph database management system. As stated earlier, its benefits are particularly well-suited for applications with complex relationships and a high degree of interconnectedness. Because of this, providing a graph database-based platform for hosting SemMed data improves the ability to navigate relationships, and Neo4j is built to efficiently query complex relationship patterns as shown in Fig. 2 . This chance offers numerous industries valuable insights into how simple it is to analyze the various connections between a multitude of articles.



]

Fig. 2. Graphical Representation of Data and Relationships

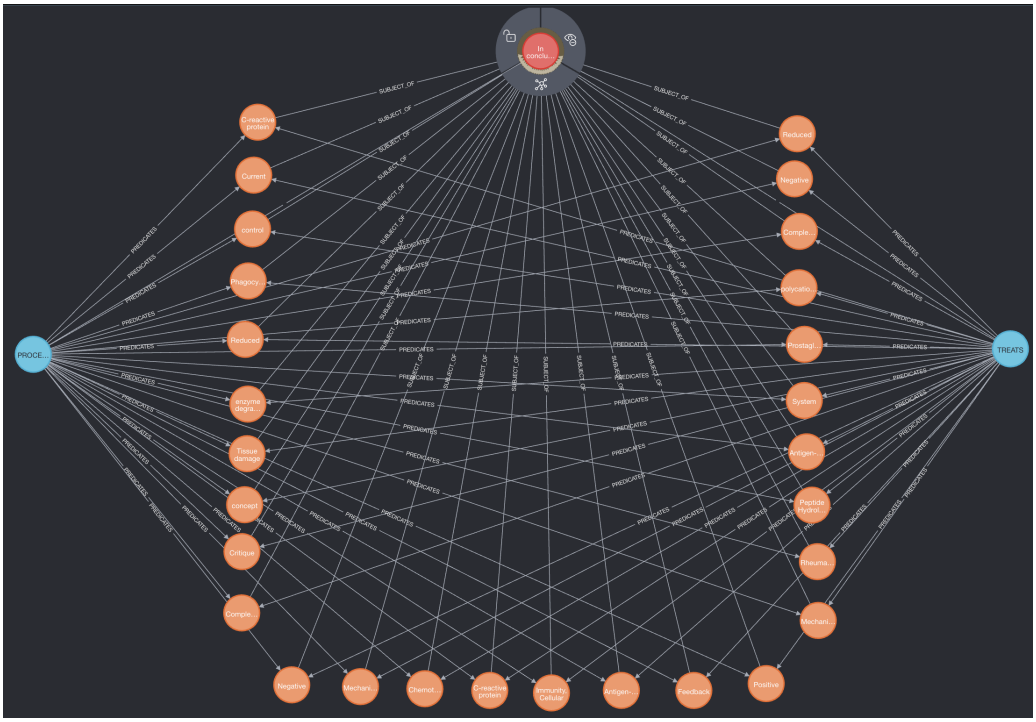


Fig. 3. Graphical Representation of a Complex Query in Neo4j

CONCLUSIONS

In the project we investigated Neo4J and how one can upload an interconnected database to it for the purpose of highlighting the relationships between the items in the database. We uploaded data from SemMedDB, a database containing literature information, specifically nouns, predicates, and the actual sentences harvested to Neo4J, connecting the three types of data via their shared keys. We found that uploading large amounts of data to Neo4J is a difficult and long process, however once it was uploaded, powerful queries could be used to search the data, and a visualization of the words and sentences was understandable, which may be useful for research into the database. The database was uploaded to a local system rather than the end goal of a server, which could serve as a more useful platform in the future. Future work could include that transition, and the creation of an API or a web application so researchers could easily use the visualizations created in the project. This project helps reveal how valuable the management of massive amounts of data can be, especially when used to make perceivable and understandable representations of that data. Through the study and implementation of Big Data, one can utilize large amounts of data to research and gather insight in a practical manner.

REFERENCES

[1] M Fiszman, T. Rindflesch, H Kilicoglu. "Abstraction Summarization for Managing the Biomedical Research Literature". Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics (2004).

[2] H Kilicoglu et al. "Constructing a semantic predication gold standard from the biomedical literature". BMC Bioinformatics (2011).

[3] A Burgun, O Bodenreider. "Comparing terms, concepts, and semantic classes in Word-Net and the Unified Medical Language System". Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations (2001), pp. 77–82.

[4] A. McCray. "Representing biomedical knowledge in the UMLS Semantic Network".High-Performance Medical Libraries: Advances in Information Management for the Virtual Era (1993), pp. 45–55.

[5] H. Kilicoglu, G. Rosemblat, M Fiszman. "Broad-coverage biomedical relation extraction with SemRep". BMC Bioinformatics (2020).

[6] Y Liu et al. "Using SemRep to label semantic relations extracted from clinical text". AMIA Annu Symp Proc (2012), pp. 587–595.

[7] R. Graciela et al. "A methodology for extending domain coverage in SemRep". Journal of Biomedical Informatics Vol. 46 (2013), pp. 1099–1107.

[8] Aronson AR. 2001 Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA Annu Symp Proc, pp 17-21.

[9] National Institutes of Health. (n.d.). Unified Medical Language System (UMLS). U.S. National Library of Medicine. https://www.nlm.nih.gov/research/umls/index.html