SemMedDB-neo4j: A Graph Database of Biomedical Semantic Relations

Dimitar Hristovski[a], Andrej Kastrin[a], Halil Kilicoglu[b,c]

[a] *Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia;*
[b] *National Library of Medicine, Bethesda, MD, USA;*
[c] *University of Illinois at Urbana-Champaign, School of Information Sciences*

ABSTRACT

Semantic relations from the biomedical literature are increasingly used in knowledge management applications such as *literature-based discovery (LBD)* as well as in clinical decision-making support. *SemMedDB* is a repository of semantic predications extracted from all PubMed by *SemRep*. SemMedDB has been so far distributed only in MySQL format. However, for many biomedical applications, the domain knowledge is more naturally represented as a graph of concepts and semantic relationships between them. In this work, we describe our recent conversion of SemMedDB to a neo4j graph database (***SemMedDB-neo4j***). It contains additional aggregated and publication history data, which is especially suitable for discovery and evaluation in LBD.

AVAILABILITY

SemMedDB-neo4j is available at:

http://lbd.mf.uni-lj.si/semmeddb-neo4j

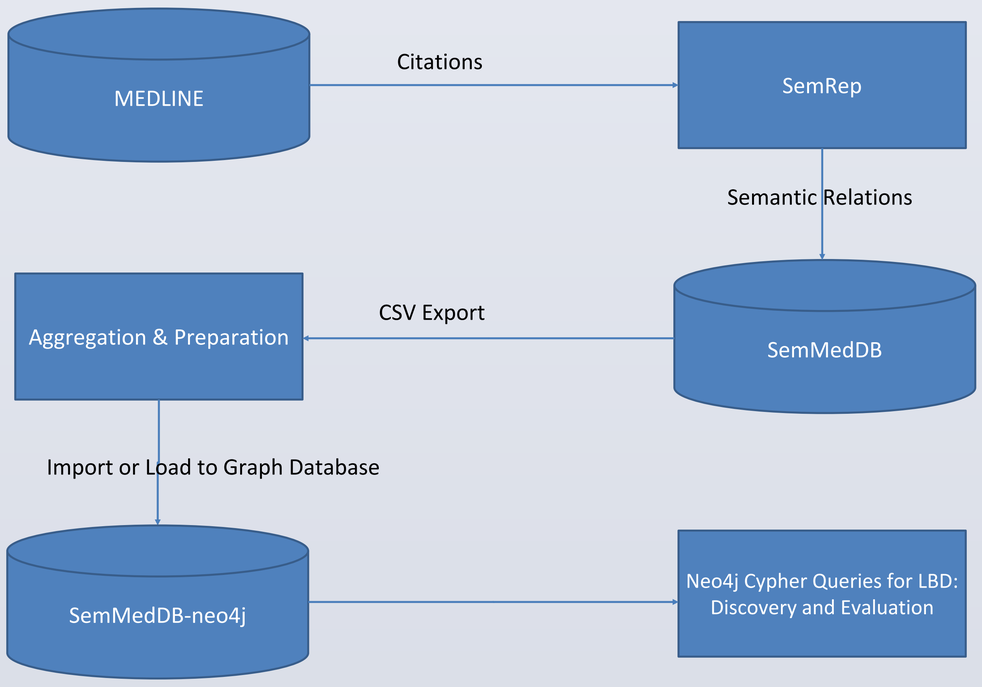
INTRODUCTION AND BACKGROUND

SemMedDB, a distribution of semantic relations extracted from full MEDLINE with SemRep (a natural language processing tool)

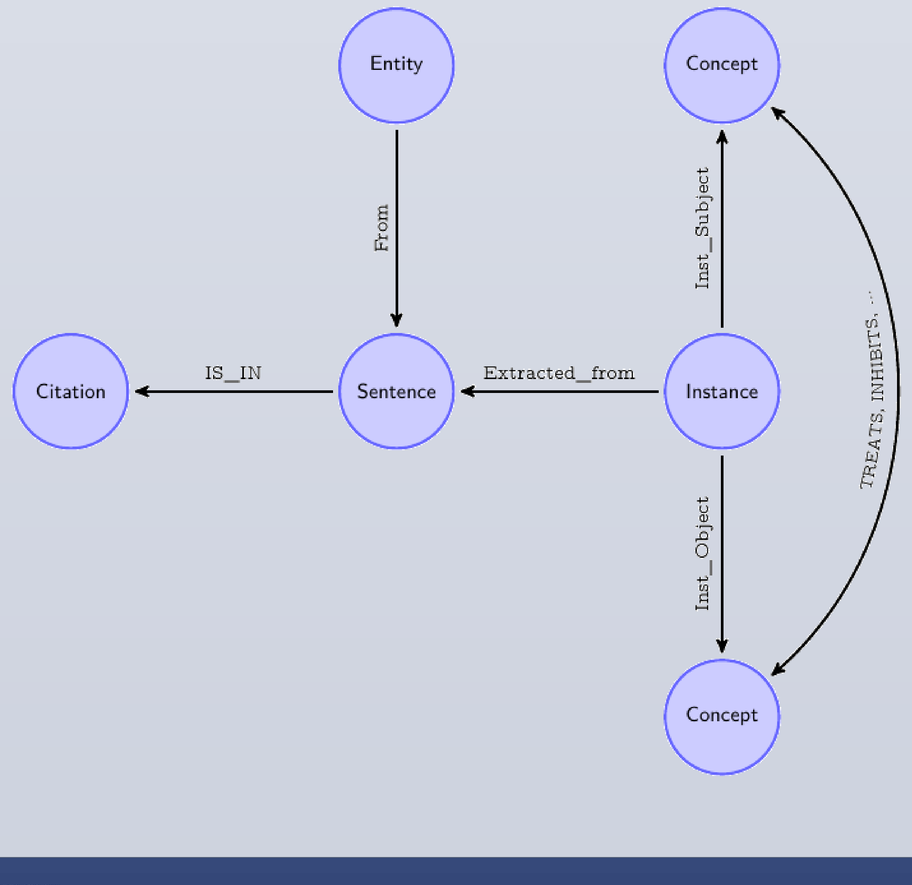
- Example: From "dexamethasone is a potent inducer of multidrug resistance-associated protein expression in rat hepatocytes" SemRep extracts:
  - Dexamethasone STIMULATES Multidrug Resistence-Associated Proteins
  - Multidrug Resistance-Associated Proteins PART_OF Rats
  - Hepatocytes PART_OF Rats

Neo4j graph database used for Implementation

PROCESSING WORKFLOW



The Graph Data Model of SemMedDB-neo4j



The Graph Data Model of SemMedDB-neo4j (cont.)

Nodes types (neo4j labels)

- `Citation` (corresponds to a PubMed bibliographic record)
- `Sentence` (corresponds to a sentence in an abstract or a title of a bibliographic record)
- `Instance` (corresponds to a semantic predication extracted from a particular sentence)
- `Concept` (corresponds to UMLS biomedical concepts that refer to arguments (i.e., subject and object) of the semantic predication). Might have additional label(s) for UMLS semantic type(s) (e.g., `dsyn` for "Disease or Syndrome" or `phsu` for "Pharmacologic Substance")
- `Entity` (corresponds to an entity mention in a sentence)

Relations between the nodes

- `IS_IN` (a relation between a `Sentence` and a `Citation` nodes),
- `Extracted_From` (a relation between a predication `Instance` and a `Sentence` nodes),
- `Inst_Subject` (a relation between an `Instance` and the `Concept` which is the subject of that instance), and
- `Inst_Object` (a relation between an `Instance` and the `Concept` which is the object of that instance).
- All semantic predicates (e.g., `TREATS`, `INHIBITS`) are represented as aggregated relations between `Concept` nodes; for example, there is only one `INHIBITS` relation between a particular concept pair, but there could be many instances of that relation (`Instance` nodes).

Properties (in addition to SemMedDB)

Related to aggregated and publication history data

EXAMPLE CYPHER QUERIES

A generic implementation of the "inhibit the cause of the disease" LBD discovery pattern for finding novel treatments:

```
MATCH (drug:Concept:phsu)-[r1:INHIBITS]->(y:Concept)-[r2:CAUSES]->(disease:Concept:dsyn)
  WHERE NOT (drug)-(:TREATS)->(disease)
  RETURN drug, r1, y, r2, disease;
```

How many actual treatments could be predicted with the above LBD pattern before the treatments were asserted in the literature?

```
MATCH (drug:Concept:phsu)-[r3:TREATS]->(disease:Concept:dsyn),
      (drug)-[r1:INHIBITS]->(y:Concept)-[r2:CAUSES]->(disease)
  WHERE r1.min_pyear<r3.min_pyear AND r2.min_pyear<r3.min_pyear
  RETURN count(r3);
```

A generic query returning linked `Citation` to `Sentence` to `Instance` to (`Subject` and `Object`), and corresponding aggregated semantic relations:

```
MATCH (c:Citation)<-[r_in:IS_IN]-(s:Sentence)<-[r_extr:Extracted_From]-(i:Instance)
       -[r_sub:Inst_Subject]->(sub:Concept), (i)-[r_obj:Inst_Object]->(obj:Concept),
      (sub)-[rel]->(obj)
  RETURN c, r_in, s, r_extr, i, r_sub,sub, r_obj, obj,rel;
```

RESULTS

Graph Database Construction

- **284,472,714** nodes (excluding "entity" nodes)
- **408,532,660** relations (excluding "entity" relations)
- **29,137,782** MEDLINE citation records processed
- **67,599,778** semantic relation instances extracted with SemRep and aggregated into
- **18,283,847** semantic relations between
- **285,675** biomedical concept nodes

CONCLUSIONS

SemMedDB-neo4j is a graph database distribution of SemMedDB, with additional aggregated and temporal data, which makes it especially suitable for knowledge management applications such as literature-based discovery (LBD).

Acknowledgements

This research has been partly supported by the Slovenian Research Agency and the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

Contact

Dimitar Hristovski e-mail: dimitar.hristovski@gmail.com