Implementing Semantics-Based Cross-domain Collaboration Recommendation in Biomedicine with a Graph Database

Dimitar Hristovski
Faculty of Medicine
Ljubljana, Slovenia
dimitar.hristovski@gmail.com

Andrej Kastrin
Faculty of Information Studies
Novo mesto, Slovenia
andrej.kastrin@guest.arnes.si

Thomas C. Rindflesch
National Library of Medicine
Bethesda, MD, USA
trindflesch@mail.nih.gov

***Abstract*— We describe a novel approach for cross-domain recommendation for research collaboration. We first constructed a large Neo4j graph database representing authors, their expertise, current collaborations, and general biomedical knowledge. This information comes from MEDLINE and from semantic relations extracted with SemRep. Then, by using an extended literature-based discovery paradigm, implemented with the Cypher graph query language, we recommend novel collaborations, which include author pairs, along with novel topics for collaboration and motivation for that collaboration.**

Keywords- Research collaboration; Recommendation system; Literature-based discovery; Semantic MEDLINE; Graph database; Neo4j.

I. INTRODUCTION

Nowadays, high quality science requires collaboration, as demonstrated by studies reporting that higher levels of collaboration correlate with higher productivity [1]. Current systems for recommending scientific collaboration are largely based on statistical analysis of co-occuring terms (e.g., ArnetMiner [2]); they provide a list of potential collaborators, but do not give motivation for the recommendations. Our methodology enhances previous work by providing a list of potential collaborators and topics for collaboration, in addition to compelling motivation for the collaboration. This innovative approach is based on a semantic implementation of literature-based discovery (LBD) methodology.

LBD is a methodology for automatically generating research hypothesis by uncovering hidden, previously unknown relationships from existing knowledge [3]. For example, suppose a researcher has studied the effect of substance X on gene Y. Further suppose that a different researcher has found a relationship between gene Y and disease Z. The use of LBD may suggest a relationship between X and Z, indicating that substance X may potentially treat disease Z. For a recent review of LBD tools and approaches see [4].

The relationships on which this project is based are semantic predications. A semantic predication is a formal structure representing part of the meaning of a sentence. For example, "Metformin TREATS Diabetes" represents part of the meaning of "Metformin is commonly used as first-line medication for management of diabetes." A semantic predication consists of a predicate ("TREATS" in this example) and arguments ("Metformin" and "Diabetes"). We used predications extracted by SemRep [5] from all of MEDLINE (titles and abstracts). SemRep is a rule-based, symbolic natural language processing system that extracts 30 predicate types expressing assertions in clinical medicine (e.g., TREATS, ADMINISTERED TO), substance interactions (e.g., INTERACTS WITH, STIMULATES), genetic etiology of disease (e.g., CAUSED, PREDISPOSES), and pharmacogenomics (e.g., AUGMENTS, DISRUPTS). The extracted predications are stored in a MySQL database (SemMedDB) which is publicly available [6]. The expressiveness inherent in semantic predications enhances the value of our system over that of the majority of LBD systems. Such systems are largely based on simple co-occurrence of phrases or concepts, which does not express the meaning of the relationship between the co-occurrences.

This work is a continuation and extension of our previous work. In [7] we described the basic cross-domain collaboration recommendation methodology, and in [8] we explained how to implement LBD with Neo4j graph database [9]. In this paper, we describe the implementation of the cross-domain collaboration recommendation methodology with the Neo4j graph database and its query language Cypher [9].

The paper is structured as follows. In Section II, we present the methods used to construct the graph database and the prediction algorithm, in Section III we present the results, and in Section IV we preset the conclusions .

II. METHODS

We first construct a large network and load it into the Neo4j graph database. We have used the Neo4j graph database because our data can be naturally expressed as a large graph and because Neo4j is well suited for storing and working with graphs. The network (graph) consists of two major types of nodes: authors and biomedical concepts. We extract the authors from the full MEDLINE bibliographic database. We extract the biomedical concepts from the set of arguments (subjects or objects) of semantic relations extracted from all MEDLINE titles and abstracts with SemRep. Each biomedical concept has a subtype, such as Disease or Syndrome or Pharmacologic Substance. We call the node subtypes semantic types and they come from Unified Medical Language System (UMLS). We use 126 semantic types. Our network contains several types of arcs and edges. `co_author` edges link any two authors that have been co-authors in at least one paper. We use this edge type to determine which authors already know each other. `writes_about` arcs link authors to biomedical concepts.

 ISBN: 978-1-61208-486-2

These arcs are derived from the semantic relations extracted from the articles written by the authors. We use the writes_about arcs to represent the expertise of the authors. Finally, we have 30 types of semantic relations extracted with SemRep that link the nodes representing biomedical concepts. These relations represent current biomedical knowledge.

The large network of nodes, arcs and edges described is the foundation on which the algorithm for recommending research collaboration operates. We implement the algorithm with the Cypher query language and it operates as follows. For a given input author (last and first names), we first compile the author's topic (concept) profile, which represents both the authors interests and expertise, by following the `writes_about` arcs as described above. The concepts from the author's profile are input to the LBD phase. For each input concept we perform an open LBD discovery. Here the methodology of discovery patterns can be used to improve the precision of the LBD process [10]. LBD proposes target concepts as novel collaboration (research) topics that are not yet published in the literature. For all target concepts found by LBD, we find authors who have these concepts in their profiles and eliminate those authors who are already co-authors with the starting author. The output is a list of the remaining authors as potential collaborators and topic(s) for collaboration. Shown in Figure 1 is a generic implementation with a Cypher query, which can be made more specific as needed.

```
MATCH (author1:author)-[:WRITES_ABOUT]->
(X:Concept) -[Rel_XY]-> (Y:Concept)
-[Rel_YZ]-> (Z:Concept) <-[:WRITES_ABOUT]-
(author2:author)
WHERE NOT (X)-[RelXZ]->(Z)
AND NOT (author1)-[CO_AUTHOR]-(author2)
RETURN author1, X, Rel_XY, Y, Rel_YZ, Z,
author2;
```

Figure 1. Generic implementation of the collaboration recommendation algorithm with a Cypher query.

We provide an illustration for this discovery process (Figure 2). In this example we use the "inhibit the cause of the disease" LBD discovery pattern [11] which states that a drug X maybe treats disease Z (new hypothesis) if the drug X inhibits gene(s) Y which causes disease Z. We have all the necessary information for this discovery pattern in our network. Also from the network we can find that authors A and B are experts for drug X and disease Z, respectively, because they write about these topics. The explanation goes as follows: If author B wants to find a novel way to treat disease Z she should collaborate with author A, because she is an expert for drug X which might be beneficial for disease Z.

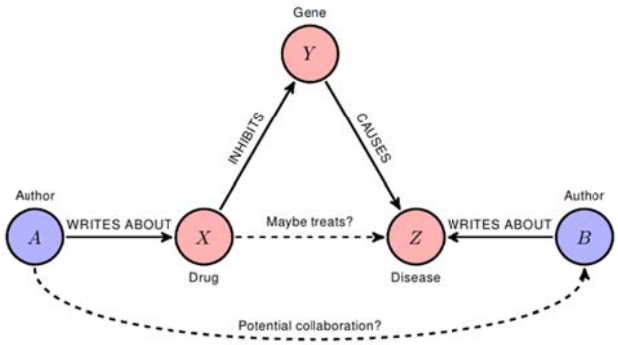

Figure 2. Illustration of the recommendation process. Based on the current biomedical knowledge (solid arcs) we recommend novel collaboration (dashed arcs).

Shown in Figure 3 is a Cypher query for the collaboration recommendation based on "inhibit the cause of the disease" discovery pattern.

```
MATCH (author1:author) -[:WRITES_ABOUT]->
(drug:phsu)-[:INHIBITS]->(gene:gngm)
-[:CAUSES]-> (disease:dsyn)
<-[:WRITES_ABOUT]- (author2:author)
WHERE NOT (drug)-[:TREATS]->(disease)
AND NOT (author1)-[CO_AUTHOR]-(author2)
RETURN author1, drug, gene, disease,
author2;
```

Figure 3. Implementation of the collaboration recommendation algorithm with a Cypher query based on the "inhibit the cause of the disease" discovery pattern.

III. RESULTS

The network we constructed consists of 69333420 semantic relations extracted from 23657386 MEDLINE bibliographic records using SemRep. The characteristics of the network are as follows: There are 9516106 author nodes and 269047 biomedical concept nodes. Regarding edges, there are 181664746 `co_author` edges between the authors, 189294999 `writes_about` arcs between the authors and biomedical concepts, and 69333420 arcs that come from SemRep semantic relations between the biomedical concepts.

We applied the collaboration recommendation algorithm using the LBD discovery pattern "inhibit the cause of the disease," which returned 275661539 unique pairs of authors. 1817 distinct drugs, 3218 distinct genes, and 8698 distinct diseases were included in the topics for collaboration; these recommendations need to be evaluated from the biomedical point of view.

Future work includes development of a user interface and visualization module. Author name ambiguity currently introduces considerable noise into the discovery process, and we need to address disambiguation in this area.

IV. CONCLUSIONS

Using a graph database such as Neo4j for storing the large network data structure needed for semantics-based cross-domain collaboration recommendation is more natural

 ISBN: 978-1-61208-486-2

and efficient than using a relational database. Implementing collaboration recommendation algorithms is conceptually easier and more simple when using a graph query language such as Cypher when compared to standard SQL.

ACKNOWLEDGMENT

This work was supported by the Slovenian Research Agency and by the Intramural Research Program of the U.S. National Institutes of Health, National Library of Medicine.

REFERENCES

[1] J. Katz and B. R. Martin, "What is research collaboration," Research Policy, vol. 26, no. 1, pp 1-18, 1997.

[2] J. Tang, J. Zhang, L.Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining – KDD 08. New York, NY: ACM Press, pp. 990-998, 2008.

[3] D. R. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," Perspectives in Biology and Medicine, vol. 30, no. 1, pp. 7-18, 1986.

[4] D. Hristovski, T. Rindflesch, and B. Peterlin, "Using literaturebased discovery to identify novel therapeutic approaches," Cardiovascular & Hematological Agents in Medicinal Chemistry, vol. 11, no. 1, pp. 14-24, 2013.

[5] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text," Journal of Biomedical Informatics, vol. 36, no. 6, pp. 462-477, 2003.

[6] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, "SemMedDB: A PubMed-scale repository of biomedical semantic predications," Bioinformatics, vol. 28, no. 23, pp. 3158-3160, 2012.

[7] D. Hristovski, A. Kastrin, and T. C. Rindflesch, "Semantics-bsed cross-domain collaboration recommendation in the life sciences: Preliminary results," Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 805-806, 2015.

[8] D. Hristovski, A. Kastrin, D. Dinevski, and T. C. Rindflesch, "Towards implementing semantic literature-based discovery with a graph database," Proceedings of the GraphSM 2015, The Second International Workshop on Large-scale Graph Storage and Management, pp. 180-184, 2015.

[9] Neo4j website. Available at: http://neo4j.com. Last accessed June 20th 2016.

[10] D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin, "Exploiting semantic relations for literature-based discovery," AMIA Annual Symposium proceedings, pp. 349-353, 2006.

[11] C. B. Ahlers, D. Hristovski, H. Kilicoglu, and T. C. Rindflesch, "Using the literature-based discovery paradigm to investigate drug mechanisms," AMIA Annual Symposium proceedings, pp. 6-10, 2007.

 ISBN: 978-1-61208-486-2