Discovering Information from an Integrated Graph Database

Erik M. van Mulligen[1,*], Wytze J. Vlietstra[1], Rein Vos[1,2], Jan A. Kors[1]

[1]Erasmus University Medical Center, Rotterdam, The Netherlands
{e.vanmulligen, w.vlietstra, j.kors}@erasmusmc.nl
[2]Maastricht University, The Netherlands
{rein.vos}@maastrichtuniversity.nl

Abstract. The information explosion in science has become a different problem, not the sheer amount per se, but the multiplicity and heterogeneity of massive sets of data sources. Relations mined from these heterogeneous sources, namely texts, database records, and ontologies have been mapped to Resource Description Framework (RDF) triples in an integrated database. The subject and object resources are expressed as references to concepts in a biomedical ontology consisting of the Unified Medical Language System (UMLS), UniProt and EntrezGene and for the predicate resource to a predicate thesaurus. All RDF triples have been stored in a graph database, including provenance. For evaluation we used an actual formal PRISMA literature study identifying 61 cerebral spinal fluid biomarkers and 200 blood biomarkers for migraine. These biomarkers sets could be retrieved with weighted mean average precision values of 0.32 and 0.59, respectively, and can be used as a first reference for further refinements.

Keywords: knowledge based discovery, graph databases

Introduction

Discovering new information from PubMed and from other biomedical databases is a time consuming and tedious process [1]. Retrieving and combining information from these databases has to be performed manually and requires an understanding of the different information models. In this paper we present a method to harmonize the information from all these biomedical databases as RDF triples and integrate them within a graph database. To investigate whether such a harmonized and integrated approach is beneficial we evaluated this against a formal literature review, performed by a collaborative expert group in neurology research, that identified from (full text) literature migraine biomarkers in cerebral spinal fluid and blood [2].

Background

Many have recognized the potential of computers to support the discovery process of new biomedical information. The pioneer in this field, Swanson, recognized the potential of relating disconnected fields of knowledge in biomedicine, in particular by discovering new associations between, as he called it, A and C terms, consisting of single words or short phrases (2-3 words). He developed a program named ArrowSmith to automatically find B terms that co-occur with A and C terms in Medline titles3. If the A and C terms were never co-mentioned in a title, a new potential discovery was identified. Using this approach he was able to discover a connection between Raynaud's disease (A) and fish oil (C) through blood coagulation (B), and between migraine (A) and magnesium (C) via blood clotting (B). These hypotheses were later on proven correct in experimental studies [4].

The value of this approach has been recognized by many scientists and a series of new research projects were started to improve on this. One method, explored by Blake and Pratt, was to use concepts as defined by the UMLS instead of separate terms [5,6]. In the UMLS thesaurus, different terms that denote the same unit of thought have been normalized to a single concept. Weeber et al. were the first to mine concepts from both Medline titles as well as abstracts, by mapping terms to the UMLS thesaurus with the MetaMap concept recognizer [7]. Weeber et al. were also succesful in applying their system for a new discovery in drug research, suggesting thalidomide as a treatment for chronic hepatitis C, among others [8].

Swanson manually selected the B terms that he considered most relevant for further exploration, but he did put also much effort in bringing together very different datasets, covering different research fields in medicine. Many researchers have followed up Swanson's work and have worked on approaches to algorithmically select the B terms. The concept-based approaches using UMLS have explored the use of the semantic types of the B concepts. Blake and Pratt used this approach to discard several semantic types and reported an 81% decrease of the number of B terms5. Srinivasan et al. applied a similar approach to filter out B terms based on semantic types [9]. If the relevant semantic types were precisely known, the set of terms could be reduced by as much as 91%; if only the obviously irrelevant semantic types were removed, the number of terms was reduced by an average of 31%. Gordon and Lindsay evaluated several ranking algorithms borrowed from the information retrieval field when they re-analyzed Swanson's fish oil-Raynaud's Disease discovery, such as Term Frequency-Inverse Document Frequency (TF-IDF) [10]. They reported reproduction of 10 of the 12 relevant B-terms for Swanson's discovery in a list of 35 terms. Torvik and Smalheiser applied an ensemble algorithm to rank the B-terms that combined eight weighted variables, such as "B-term occurs in more than one paper within literature sets A and C", "B-term maps to at least one UMLS semantic category", "B-term first appears recently within Medline as a whole", etc. [11]. While Swanson originally used a fixed order approach of first filtering uninformative terms using a stop word list, subsequently term categorization, and finally manual selection of B-terms, this ensemble algorithm contains all steps of Swanson's fixed order approach, but has the advantage of not losing potentially relevant B terms in any of the intermediate steps.

Yetsigen-Yildiz et al. compared statistics to rank the B-terms [12]. Two of them were frequency-based, with the TF-IDF and the association rules as tested by Hristovski [13] et al., and two were probability based; the Z-score, which creates literature subsets, and the mutual information score. The association rules were not evaluated against the Swanson sets, but they were analyzed on their predictions from a subset of Medline's future published discoveries.

Hristovski et al. were the first to test the added value of incorporating relation predicates into a literature-based discovery process [14]. They applied the UMLS semantic network and the SemRep text mining system to identify relationships between terms [15]. Predicates were used to identify discovery patterns: specific combinations of two predicates between three terms, which when combined would constitute a functional, biologically relevant association. Although the inclusion of predicates was considered to offer clear advantages, the lack of accuracy of the relationship extraction hampered practical application.

With the ANNI discovery system the co-occurrences between a concept and other concepts in all Medline abstracts were computed and stored in a so-called concept profile [16]. The strength of a relationship between two concepts is expressed as a matching score between their concept profiles. Concepts can be grouped based on their semantic type and their concept profiles can be matched based on various algorithms: mutual information measure, log-likelihood, and dot product [17]. The matching strategy takes into account all the B concepts contained in the concept profile, filters the resulting C concepts on the required semantic type(s) and ranks the result on matching score. This approach has been used by Jelier et al. in a study to match the concept profiles for genes from DNA microarray data with concepts that denote gene functions [18]. The same approach has been used by Van Haagen et al. to predict protein-protein interactions by computing the matching score between protein concept profiles at certain time intervals in Medline [19]. An extension of this approach has been developed by using ANNI in mapping disease-disease relationships for knowledge discovery in multi-morbidity research on somatic and psychiatric diseases [20].

Previous work mainly focused on discovering direct relations based on mainly one source of information (literature). In this paper we describe a novel approach to extend discoveries of associations between a source and target concept to associations that involve a series of intermediate concepts (paths) combining information mined from literature, conventional biological databases, semantically enriched information (RDF) sources, and biomedical ontologies and thesauri. The formalization of this information from different heterogeneous sources into RDF triples and the integration of these triples in a graph database seems to be logical next step to support information discovery tasks with multiple intermediate nodes and offers more possibilities to rank the various discovered connections using graph statistics. We evaluated this approach of using a graph database based on heterogeneous sources for information discovery by comparing discovered associations with the results of an actual formal literature review.

Methods

Our approach semantically integrates triples extracted from Medline abstracts as provided in Semantic Medline [21] with relations obtained from the UMLS 2012AA and databases such as UniProt, EntrezGene, Comparative Toxicogenomics Database, and RDF triples from the datasets contained in Linked Open Drug Data [22] (DrugBank, DailyMed, and SIDER) into a graph database. Furthermore, the relations between The subject and object resources and the predicates in the Semantic Medline triples were already expressed in terms of our ontology and predicate thesaurus. For UniProt, EntrezGene and the Comparative Toxicogenomics Database the process of making triples included the mapping of the implicit relations of the database schema to explicit predicates and the mapping of the subject and object to a RDF resource, i.c. a UMLS, UniProt or EntrezGene identifier. For each UMLS concept in our ontology we have all the different identifiers and all terms used to refer to the concept. Mapping the information of a database record to a concept in our ontology was obtained either by matching it to one of its identifiers or by matching it to one of its terms. The term matching was performed by applying our Peregrine [23] text mining pipeline.

From these different sources we identified 2,669,792 individual concepts, together with about 71 million relations between them. The relations are based on the relationships defined in the UMLS Semantic Network, the relationships defined in the UMLS MetaThesaurus (MRREL table), and the predicates defined by Halil and used within Semantic Medline [21]. We harmonized the set by looking at trivial synonyms in this set and mapped these to 171 different predicate identifiers. Each subject and object from a RDF triple are related with an "isa" relation to one or more semantic types as defined in UMLS. Semantic types on their turn are linked with an "isa" relation to a semantic group.

The resulting semantically mapped relations, commonly referred to as triples, have been stored in a graph database. The graph database has been implemented in the Neo4J graph database, version 1.8.324. In order to add the triples with their provenance we developed an import program that uses the Neo4J using the java API of Neo4J. We implemented a REST API on top of Neo4J that implements the notion of concepts, labels, semantic types, semantic groups, and provenance. Each RDF triple is represented in the graph database by making a labeled node using the preferred term of the ontology concept for both subject and object and the predicate name as labeled edge to represent the relation between subject and object node. A subject and object node can be linked with multiple semantic predicate labels and the provenance information implemented as a reference to a text or a database record as the source of the relation can be added to each edge. Semantic predicates contain a direction and for both directions labels are provided, typically the active and passive form of a verb. Neo4J has built a path finder algorithm that find paths between nodes in the graph. We extended this functionality with the use of provenance information in scoring the various paths. An example of the mapping of the database of UniProt is provided in Table 1.

Table 1. A mapping of some of the UniProt record fields for 14-3-3 protein beta/alpha to an RDF triple. This protein is mapped to the subject resource http://www.uniprot.org/uniprot/P31946. The UniProt Keyword is manually mapped to the closest RDF thesaurus predicate. The field contents have been mapped to ontology concepts using the Peregrine concept identification pipeline.

UniProt Keyword	UniProt Annotation	RDF Predicate	RDF Object resource
gene	YWHAB	gene_product_encoded_by_gene	http://www.ncbi.nlm.nih.gov/gene/7529
GO - Molecular function	enzyme binding	gene_product_has_biochemical_function	https://uts-ws.nlm.nih.gov/rest/content/current/CUI/C1149286
GO - Molecular function	histone deacetylase binding	gene_product_has_biochemical_function	https://uts-ws.nlm.nih.gov/rest/content/current/CUI/C1323310
GO - Biological process	activation of MAPKK activity	gene_product_plays_role_in_biological_process	https://uts-ws.nlm.nih.gov/rest/content/current/CUI/C1155556
GO - Biological process	epidermal growth factor	gene_product_plays_role_in_biological_process	https://uts-ws.nlm.nih.gov/rest/content

	receptor signaling pathway		/current/CUI/C1155379
Keywords - Biological process	Host-virus interaction	gene_product_plays_role_in_biological_process	https://uts-ws.nlm.nih.gov/rest/content/current/CUI/C0599952
Subcellular location	Cytoplasm	location_of	https://uts-ws.nlm.nih.gov/rest/content/current/CUI/C0010834
Keywords - Cellular component	Cytoplasm	part_of	https://uts-ws.nlm.nih.gov/rest/content/current/CUI/C0010834
Keywords - Cellular component	perinuclear region of cytoplasm	part_of	https://uts-ws.nlm.nih.gov/rest/content/current/CUI/C2253855
Keywords - Coding sequence diversity	Polymorphism	gene_product_has_abnormality	https://uts-ws.nlm.nih.gov/rest/content/current/CUI/C0032529
Organism	Homo sapiens (Human)	conceptual_part_of	https://uts-ws.nlm.nih.gov/rest/content/current/CUI/C0086418

The challenge of integrating UniProt entries lies in mapping the annotation fields to the corresponding ontology concepts. We used our concept identification pipeline Peregrine to identify concepts in the free text UniProt annotation fields [22]. The mapping of the implicit relations defined in the UniProt schema to the proper semantic predicates is a one-time manually effort and requires understanding of the biological meaning of the data. This mapping process has been performed for all integrated databases. Once created a mapping can be applied to each update of the database.

Discovering connections

Around Neo4J's basic functionality we provided a web service that implements functionality necessary for our discovery task. In particular, for inferencing we implemented a path-finding algorithm that extends Neo4J's functionality. This simple, path-finding type of inferencing is not following the main, logic-based inferencing approaches such as implemented with OWL-DL and formal reasoners. The extension of Neo4J's path-finding function allows one to specify a set of semantic predicates that restricts the set of triples that can be explored to find a path between the source and target concepts. The paths lengths are currently limited to a maximum of five triples to avoid computational explosion. The path function can be modified and can take into account additional information and graph statistics that may influence the selection of triples, e.g., the amount of provenance information (the sources that support the relation), the variety of databases that support the triple.

Results

The graph database has been used in a number of application domains. To evaluate the use of this graph database with triples obtained from PubMed abstracts and biomedical databases we evaluated the identification of biomarker compounds marking the imminence of a migraine attack with those reported in a literature review study. We obtained the set of 61 compounds that have been reported to be measurable in the cerebral spinal fluid, and a set of 200 compounds reported to be measurable in the serum of migraine patients. The latter set was obtained at the same literature review study but not yet submitted for publication. Both sets were manually constructed by a manual review process of a corpus of articles retrieved with PubMed, EMBASE, and Web of Science. The objective of our study was to test whether a graph database could be used to identify a set of linking concepts, similar to the linking B-terms, between these compounds and migraine. The question was whether this set of linking concepts with their interconnectivity could be used to identify (1) the original set of compounds, and (2) new compounds of interest.

Table 2. Overview of semantic types in migraine subgraph.

Biologically Active Substance	Chemical Viewed Structurally
Neuroreactive Substance or Biogenic Amine	Organic Chemical
Hormone	Nucleic Acid, Nucleoside, or Nucleotide
Enzyme	Organophosphorus Compound
Vitamin	Amino Acid, Peptide, or Protein
Immunologic Factor	Carbohydrate
Receptor	Lipid
Disease or Syndrome	Steroid
Mental or Behavioral Dysfunction	Eicosanoid
Body Part, Organ, or Organ Component	Element, Ion, or Isotope
Tissue	Physiologic Function
Cell	Organism Function
Cell Component	Organ or Tissue Function
Gene or Genome	Cell Function
	Molecular Function

The two sets of compounds were fed to the graph database to obtain the paths between these compounds and migraine. These paths were analyzed for characteristics (number of publications, range of publication dates, path length, etc.). Additional compounds that were not part of the initial set have been viewed as potentially new discovered compounds.

The final result of this study was a set of concepts found in the paths linking migraine to these sets of compounds. A selection of this set of linking B-concepts was made on basis of a subset of the semantic types. The Signs and Symptoms semantic type was excluded based on discussion with the migraine researchers. Furthermore, Pharmacological Substances and Antibiotics, and concepts which were both a Pharmacological Substance, Organic Chemical, or Steroids, or Nucleic acids and amino acids, as well as Antibiotics were explicitly excluded, because the migraine researchers were only interested in endogenous compounds associated with migraine, and not in chemotherapeutic treatments. The final list of semantic types is shown in Table 2. Using this selected B-concept set we used the number of different connections between a compound and the B-concept set for reconstructing the initial given set of compounds and secondly to identify potential new compounds. Several ranking statistics were evaluated and overall there was only very little difference. From the cerebral spinal fluid set of 61 compounds directly connected to migraine one could not be identified and from the serum set of 200 compounds directly connected to migraine 23 could not be identified using this approach. A weighted mean average precision of 0.32 was computed for the cerebral spinal fluid set and 0.59 for the blood set. Or stated differently, 78% of the unique reference compounds (222 compounds) were found in the top 4% (=1500) results, which means that about one out of ten results was a reference compound.

Error Analysis

We performed an error analysis on our reference set, by examining compounds from the top 100 of our results that are not included in our set of reference compounds. The results of this analysis are shown in Table 3. From the 24 compounds that were not in the list, 13 were excluded from the list because they were categorized as inorganic chemical or as a pharmaceutical preparation and excluded initially from the analysis and 11 were only remotely connected to migraine and therefore excluded from the result list. From the 49 compounds low on the list 6 were ranked low because of the ranking mechanism, 28 compounds were connected but due to many connections outside the migraine cluster ranked low, and 15 compounds were only connected via an ontological relationship (a "isa" relationship with a class) and where therefore ranked low.

Table 3. Error categories for missed compounds.

Reason for not being found	n
Not in list	24
Does not fit inclusion criteria	13
Too far away in graph	11
Low on list	49
Not always excluded	6
Other	28
Ontologically connected only	15

Discussion

As mentioned in the introduction, the ranking and filtering of the B-terms determines to a large extent the success of the knowledge discovery method. A similar issue can be raised about the ranking and relevance of the connecting paths that our method constructs in a multi-source graph database. With increasing path lengths, at some point each pair of concepts in the graph database will be connected. It will therefore be important to investigate approaches that can differentiate between useful and sound discovery paths and those that are noisy and redundant. The platform is powerful in its potential to implement discovery patterns that combine a rich feature set consisting of semantic types, semantic groups, semantic predicates, connectivity, and amount of provenance stemming from different sources.

From our experiments and our interactions with the expert group thus far it became clear that adding more ontological grounding to the semantic predicates would be helpful. Similar to semantic types and groups, which denote the specific properties of concepts, we can imagine that representation of the predicates by concepts with references to specific types of predicates - e.g., transitive, intransitive, causative, factitive, etc. Predicate types would impose specific properties to the predicates, such as transitive inference, that could be relevant for the discovery process.

For this application we did not restrict the discovery connection paths on basis of the combination of a particular semantic groups or types of concepts with a set of particular predicates. Our first experience is that such a selection might help in finding more relevant connections. The flexibility of the graph database to support various types of selections has been used in a different application in the field of adverse drug reactions and in food safety. We will further investigate in how far these selections are depending on an application and can be formalized in a guideline on how to use a graph database for discovery.

The compounds in the top of the result list that were not part of the reference set have not been analyzed yet, but a quick scan learned that there could be potential interesting compounds that are worth further investigation. This approach can also potentially "use" the high connectivity of a compound with a reference set to discover new potential interesting compounds. The verification of this is, however, a tedious and costly process.

Conclusion

The graph database that we constructed combines information extracted from biomedical texts with information obtained from biological databases. We have shown in this paper that relations from texts and structured databases can be effectively combined in a single graph database. However, this approach is a first step to use large integrated datasets to support the discovery process. Research will be required to better understand the importance of graph statistics for the discovery process. What we present in this paper is a first step that can be used as a reference for further work.

The information discovery approach illustrated in this paper shows that relevant compounds as identified in an actual formal literature review can be retrieved with a fairly high recall. Furthermore, our approach shows that the connectivity to a set of other concepts has potential. The flexibility of the graph database enables the application of

the approach to other discovery applications and evaluate other approaches to combine graph statistics and filters on semantic groups and predicates.

References

1. Lu Z.: Pubmed and beyond: a survey of web tools for searching biomedical literature. Database, Oxford (2011)
2. van Dongen R.M., Zielman R., Noga M., Dekkers O.M., Hankemeier T., van den Maagdenberg A.M., Terwindt G.M., Ferrari M.D.: Migraine biomarkers in cerebrospinal fluid: A systematic review and meta-analysis. Cephalalgia (2016)
3. Swanson D.R., Smalheiser N.R.: An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence, 91, 183-203 (1997)
4. Swanson D.R.: Medical literature as potential source of new knowledge. Bull. Med. Libr. Assoc. 78, 1, 29-37 (1990)
5. Blake C., Pratt W.: Automatically identifying candidate treatments from existing medical literature. AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases. 9-13 (2002)
6. Bodenreider O.: The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Res., 32 (Database issue), D267-D270 (2004)
7. Weeber M., Klein H., de Jong-van den Berg L.T.W., Vos R.: Using concepts in literature-based discovery: Simulating Swansons Raynaud-Fish Oil and Migraine-Magnesium Discoveries. Journal of the American Society for Information Science and Technology 52, 7, 548-557 (2001)
8. Weeber M., Vos R., Klein H., Aronson A.R., Molema G.: Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. J. Am. Med. Inform. Assoc. 10, 3, 252-259 (2003)
9. Srinivasan P. Text mining: generating hypotheses from Medline. Journal of the American Society for Information Science and Technology 55, 5, 396-413 (2004)
10. Lindsay R.K., Gordon M.D.: Literature-based discovery by lexical statistics. (1999)
11. Torvik V.I., Smalheiser N.R.: A quantitative model for linking two disparate sets of articles in Medline. Bioinformatics 23, 13, 1658-1665 (2007)
12. Yetisgen-Yildiz M., Pratt W.: A new evaluation methodology for literature based discovery systems. J. Biomed. Inform. 42, 4, 633-643 (2009)
13. Hristovski D., Stae J., Peterlin B., Dzeroski S.: Supporting Discovery in Medicine by Association Rule Mining in Medline and UMLS. Medinfo 10 (Pt2), 1344-1348 (2003)
14. Hristovski D., Friedman C., Rindflesch T.C., Peterlin B.: Exploiting semantic relations for literature-based discovery. AMIA annual symposium proceedings. 349 (2003)
15. Ahlers C.B., Fiszman M., Demner-Fushman D., Lang F., Rindflesh T.C.: Extracting semantic predication from MEDLINE citations for pharmacogenomics. In: Pacific Symposium on Biocomputing 209–220 (2007)
16. Jelier R., Schuemie M.J., Veldhoven A., Dorssers L.C., Jenster G., Kors J.A.: Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol. 9, 6, R96 (2008)
17. Jelier R., Schuemie M.J., Roes P.J., van Mulligen E.M., Kors J.A.: Literature-based concept profiles for gene annotation: The issue of weighting. Int. J. of Med. Inform. 77, 5, 354–362 (2008)
18. Jelier R., Jenster G., Dorssers L., Wouter B., Hendriksen P., Mons B., Delwel R., Kors J.A.: Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. BMC Bioinformatics 8, 14 (2007)
19. van Haagen H.H., 't Hoen P.A., de Morrée A., van Roon-Mom W.M., Peters D.J., Roos M., Mons B., van Ommen G.J., Schuemie M.J.: In silico discovery and experimental validation of new protein-protein interactions. Proteomics 11, 5, 843-853 (2011)
20. Vos R., Aarts S., van Mulligen E.M., Metsemakers J., van Boxtel M.P., Verhey F., van den Akker M.J.: Finding potentially new multimorbidity patterns of psychiatric and somatic diseases: exploring the use of literature-based discovery in primary care research. J. Am. Med. Inform. Assoc. 21, 1, 139-145 (2014)
21. Kilicoglu, H., Shin D., Fiszman M., Rosemblat G., Rindflesch T.C.: SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics 28, 23, 3158-3160 (2012)
22. Linking Open Drug Data (LODD), http://www.w3.org/wiki/HCLSIG/LODD
23. Peregrine, https://trac.nbic.nl/data-mining/
24. Neo4J Developers: Neo4J, Graph NoSQL Database 2012, http://neo4j.org/