A Multi-Source Linked Open Data Fusion Method for Gene Disorder Drug Relationship Querying

Guozheng Rao[1, 4, [0000-0002-6261-576X]], Li Zhang[2, *], Xiaowang Zhang[1, 4], Wenwen Li[1], Fang Li[3], Cui Tao[3]

[1] College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[2] School of Economics and Management, Tianjin University of Science and Technology, Tianjin 300222, China
[3] The University of Texas School of Biomedical Informatics, 7000 Fannin St Suite 600, Houston, TX 77030, United States
[4] College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
{rgz, zhangli, xiaowangzhang, lww1204}@tju.edu.cn, {Fang.Li, Cui.Tao}@uth.tmc.edu

Abstract. Biomedical data are gradually increasing. More and more research focused on Relationship querying between genes, drugs, and disorders. Generally, genes, drugs, and disorders information were stored in different heterogeneous datasets. These datasets were stored in different places and different formats, such as RDF/XML, SQL relational databases, and text, etc. The challenge is the fusion of multi-source and cross-platform biomedical datasets for most application based on gene disorder drug Relationship querying. To tackle these problems, we propose a novel multi-source linked open data fusion method for gene disorder drug Relationship querying. In this method, a variety of biomedical datasets are converted into RDF triple data; and then multi-source datasets are formed into a storage system with data fusion method. After fusion, the system can query the relationships among various entities from different datasets. The experiment results demonstrate that our method significantly has advantages in integrating multi-source heterogeneous biomedical datasets with high efficiency and reliability. The SPARQL query experiment is carried on 4 different datasets by using 9 kinds of common query questions proposed in this paper. The results show that most of the query results came from different datasets. the method can be used to fusion more other different biomedical datasets.

Keywords: Linked Open Data, Gene Disorder Drug Relationship querying, Data Fusion.

1 Introduction

Large amounts of semantic data are available in RDF format in many fields, such as life science[1]. These datasets cover many fields such as medicine and biology. Most biomedical researchers hope to find more research results through these biomedical

***Corresponding author：Li Zhang, zhangli@tju.edu.cn**

datasets. Relationship querying among disorders, gene, and drugs is important to multiple biomedical research, such as extracting disease biomarkers[2], identifying disease mechanisms[3], and predicting health benefits (efficacy) of a drug over a placebo[4], which could further facilitate precision medicine and clinical decision support.
However, genes, drugs, and disorders datasets were built by different research organization and instruction for different biomedical application. Furthermore, these datasets were stored in different places and different formats, such as RDF/XML, SQL relational databases, and text, etc. How to integrate multi-source and cross-platform biomedical datasets is a big challenge for gene disorder drug Relationship querying. To tackle these problems, semantic web technology is used. For example, Bio2RDF[5] was built for a more complete biological database for associated bioscience data. A total of 35 datasets have been integrated.
Most of these datasets are in a specific field. However, more and more genes, drugs, and disorders datasets will be emerged. Furthermore, there are many new genes, drugs, and disorders will be found in them. How to realize the fusion of multi-source data as needed is an important issue. Another key issue is how to realize a fast query of large-scale data across multiple datasets for a specific application.

2 Multi-source Data Fusion

A comprehensive fusion of multi-source data scheme is presented: normalize the RDF vocabulary and replace the original multiple old vocabularies, and normalize the URI of the same entity and replace the old URI. The steps mainly include:

1. Data collection: There are three methods to collect data. The data can be imported from local datasets, crawled through the crawler, or imported via SPARQL.
2. Map to schema: The purpose of this step is to generate a united RDF vocabulary. The tool R2R [6] is used as a mapping language that translates concepts from all the datasets into an application's target vocabulary.
3. Unifying URI. The purpose of this step is to normalize the URI of the same entity in different data sources, avoiding different aliases of the same entity. The tool used is SilkLink Discovery Framework.
4. Quality evaluation and data fusion: LDIF[7] (Linked Data Integration Framework) is an open-source linked data integration framework., *sieve* of LDIF is used to evaluate the quality of data fusion.
5. Output: in LDIF, the format of the data output can be written to a file or stored in the corresponding memory system.

In this paper, we first preprocess each dataset:
1) convert PharmGKB to plain text format;
2) import it into relational database;
3) convert it to RDF data with D2R tool.
KEGG is preprocessed in the same way; SemMedDB is a relational database that needs to be converted to RDF data with the D2R tool.
After preprocessing, we merge the multi datasets using the Algorithm I.

```
1: while(getTriple(?s, rdfs:label, ?o) || getTriple(?s, pharmgkb:name, ?o)
2:  || getTriple(?s, kegg:name, ?o) || getTriple(?s, sem:name, ?o))
3:      the predicate is replaced with myprop:Label
4:
5:  while(getTriple(?s, a, kegg:drug) || getTriple(?s, a, pharmgkb:
6:  PharmGKB_Drugs || getTriple(?s, a, sem:drug))
7:      the object is replaced with myclass:Drug
8：
9:  while(getTriple(?s, a, kegg:gene) || getTriple(?s, a, pharmgkb:
10: PharmGKB_Genes || getTriple(?s, a, sem:gene) || getTriple(?s, a,
11: uniprot:Gene))
11:     the object is replaced with myclass:Gene
12:
13: while(getTriple(?s, a, kegg:disorder) || getTriple(?s, a, pharmgkb:
14: PharmGKB_Disorders || getTriple(?s, a, sem:disorder))
15:     the object is replaced with myclass:Disorder
```

Fig. 1. Algorithm I- Datasets Fusion (Uniprot, PharmGKB, SemMedDB, KEGG_GENE, KEGG_PATH)

Algorithm I describes the operation of the dataset fusion.

1. Merge the name and name-like properties in each dataset into the myprop: Label type;
2. Normalize drugs in each dataset into customized myclass:Drug;
3. Normalize the gene in each dataset to the customized myclass:Gene;
4. Normalize the disorder in each dataset to the customized myclass:Disorder;

3 Experiment

3.1 Datasets

The following are several biomedical datasets used in the experiment to verify the method. PharmGKB[8] is a database of pharmacogenetics and pharmacogenomics. UniProt [9] is a comprehensive resources for protein sequences and annotation data. The KEGG[10] database is now used as a reference knowledge base for the integration of molecular data sets for genome sequencing. SemMedDB is a repository of semantic predications from MEDLINE citations (titles and abstracts)[11]. In this paper, the Gene-Disorder-Drug relationship extracted from SemMedDB is used and converted to RDF for experiments.

3.2 Query Design

For the gene-drug-disorder relationship querying, nine kinds of relational queries are designed in Table 1. These queries contain all the relationships between gene-drug-disorder.

Table 1. Comparison Table of Predicate Relationships

No.	Query pattern
Q1	Query all genes involved in a specific gene
Q2	Query all disorders caused by a specific gene
Q3	Query all drugs targeted by a specific gene
Q4	Query all disorders involved in a specific disorder
Q5	Query all genes caused a specific disorder
Q6	Query all drugs treated a specific disorder
Q7	Query all drugs involved in a specific drug
Q8	Query all disorders treated by a specific drug
Q9	Query all genes targeted a specific drug

It is necessary to know the possible paths from a disorder to a drug to query the relevant drugs from a disorder.

4 Results

Table 2. Statistical analysis of query results

No.	SemMedDB	PharmGKB	KEGG	Uniprot
Q1	62.22%	25.56%	—	12.22%
Q2	63.48%	36.52%	—	—
Q3	80%	20%	—	—
Q4	100%	—	—	—
Q5	—	100%	—	—
Q6	76.60%	23.40%	—	—
Q7	83.56%	—	16.44%	—
Q8	74.42%	25.58%	—	—
Q9	100%	—	—	—

For the nine relationships between genes, disorders, and drugs, nine queries (*Q1*-*Q9*) were designed. Table 2 is shown the source and respective proportion of each query results. The data is mainly from SemMedDB and PharmGKB, and some of the results are from KEGG and Uniprot. Except the *Q4*, *Q5*, *Q9*, all the results are from multi datasets. It can help to get more valued results to analyze the relationship among gene-disorder-drug.

5 Conclusions

In this paper, a multi-source linked open data fusion method for gene-disorder-drug Relationship querying is proposed. A variety of biomedical data are converted into RDF triple data; and then multi datasets are formed into a storage system with data fusion method. After fusion, the system can discover the relationships among various entities. The SPARQL query experiments are carried out by using 9 kinds of common query questions proposed in this paper. The experiments results show that most of the

query results came from different datasets. The multi-source medical linked data storage method that supports federal query among multi datasets.

Acknowledgment

This research is partially supported by the National Natural Science Foundation of China (NSFC) (61373165, 61672377). The authors also appreciate the support from the National Library of Medicine of the National Institutes of Health under Award Number R01LM011829.

Reference

1. Cong, Q., Feng, Z., Li, F., Zhang, L., Tao, C.: Constructing Biomedical Knowledge Graph Based on SemMedDB and Linked Open Data. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, Madrid, Spain (2019).
2. Vlietstra, W.J., Zielman, R., van Dongen, R.M., Schultes, E.A., Wiesman, F., Vos, R., van Mulligen, E.M., Kors, J.A.: Automated extraction of potential migraine biomarkers using a semantic graph. J. Biomed. Inform. 71, 178–189 (2017).
3. Hofmann-Apitius, M., Ball, G., Gebel, S., Bagewadi, S., De Bono, B., Schneider, R., Page, M., Kodamullil, A.T., Younesi, E., Ebeling, C., Tegnér, J., Canard, L.: Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders. Int. J. Mol. Sci. 16, 29179–29206 (2015).
4. Gao, J., Rao, G., Zhang, L.: PSPS: A pharmacological substances prediction system based on biomedical literature data. In: The 2nd International Workshop on the Semantics of Mental Health. pp. 1–6. IEEE, Xian,China (2019).
5. Belleau, F., Tourigny, N., Good, B., Morissette, J.: Bio2RDF: a semantic web atlas of post genomic knowledge about human and mouse. In: International Workshop on Data Integration in the Life Sciences. pp. 153–160. Springer, Berlin, Heidelberg (2008).
6. Bizer, C., Schultz, A.: The R2R Framework: Publishing and Discovering Mappings on the Web. COLD. 665, (2010).
7. Schultz, A., Matteini, A., Isele, R., Mendes, P.N., Bizer, C., Becker, C.: Ldif-a framework for large-scale linked data integration. In: 21st International World Wide Web Conference Developers Track. CEUR-WS.org, Lyon, France (2012).
8. Whirl-Carrillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B., Klein, T.E.: Pharmacogenomics knowledge for personalized medicine. Clin. Pharmacol. Ther. 92, 414–417 (2012).
9. Consortium, U., others: UniProt: the universal protein knowledgebase. Nucleic Acids Res. 46, 2699 (2018).
10. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M.: KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, 457–462 (2015).
11. Rindflesch, T.C., Kilicoglu, H., Fiszman, M., Rosemblat, G., Shin, D.: Semantic MEDLINE: An advanced information management application for biomedicine. Inf. Serv. Use. 31, 15–21 (2011).