



Master's Project Documentation

Structural Summary of RDF Graph

Abstract

This document covers the master's project scope, implementation extracts, and descriptions

Krupali Patel
kjp59@njit.edu

Table of Contents

Abstract	2
Terms and Definition Description	3
Dataset Used	5
Mondial Database	5
Structural Summary	6
Definition:	6
Importance:	6
Extraction of Structure:	6
API	6
Software	7
Method	8
Implementation	8
Results	9
Setbacks	9
Conclusion	9
References	10
Figure 1 RDF graph from (Ananya Dass)	3
Figure 2 ER Diagram of Mondial Dataset (Informatics)	5
Figure 3 Relational table generated from RDF data	7

Abstract

The Masters project focuses on implementing an application using Apache Jena API with MySQL workbench and Java, aiming to extract Pattern Graphs from the Structured Graph based on the Keyword Search applied to Linked Data. Based on exploiting semantic result clustering to support keyword search on linked data, the project involves the use of the concepts like Linked Data, RDF Data Graphs, Keyword Search, Keyword Search on RDF Data Graphs, and Personalization.

I implemented a technique for extracting the structural summary of a RDF graph. The project also involves on applying the selected algorithm to compute answers of keyword queries on the structural summary of an RDF data graph. Different datasets are examined as a target including the Mondial dataset, the DBLP dataset and the Jamendo dataset.

I also studied different algorithms for Keyword Search on Graph Data. In particular, I studied the algorithms BANKS (Gaurav Bhalotia), BLIKS (Hao Hey, 2007), and EASE (Guoliang Li). (Haixun Wang) research paper is a nice survey of Algorithms for keyword search on Graph Data.

Terms and Definition Description

For the project implementation, the concepts revolve highly upon the below mentioned terms.

Linked Data: It revolves around the concept of Sematic Web, which is a web of data. There are a number of Web Technologies (RDF, OWL, SPARQL, etc) that provide to a user an environment where they can query data, and draw inferenced from the same. This information should be available in standard formats, and relationships among the data should be made available to create a Web of Data. Linked Data is the term used to infer to these interrelated data sets.

RDF Data Graphs: RDF (Resource Description Framework) is a standard model for data interchange on the Web.

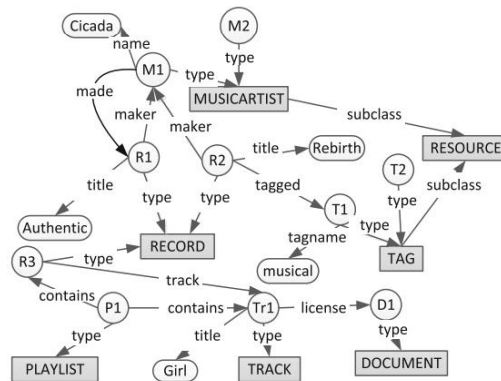


Figure 1 RDF graph from (Ananya Dass)

RDF provides a framework for representing information about web resources in a graph form. The RDF vocabulary includes elements that can be classified into Classes, Properties, Type, Values, Entities, and Relationships. An RDF graph is a quadruple $G = (V, E, L, I)$ where V : finite set of vertices. E is a finite set of directed edges, L is a finite set of lables, and I assigns lables to class and value vertices and to relationship and property edges. (Ananya Dass)

Structural Summary: It is a graph that summarizes an RDF graph by distinguishing between the various classes, entities, values of a graph. And finding the type, relationship, subclass, and property between them. It summarizes the entire RDF graph by consolidating it for query optimization.

Pattern Graph: Pattern Graph is a non-cyclic graph that is formed for every keyword in query Q and the connections between them. These graphs compute the subgraph of the structural summary of RDF graph. They are also known as result pattern graphs.

Keyword Search: Searches for similar keywords from the web of data to retrieve user relevant results. Doesn't take an input a relation query.

Keyword Search on Graph Data: Takes an input a keyword, and applies it on the structured graph to retrieve pattern graphs that match relevant keywords that are taken as the input. Relational query is not entered as the input to process on the Graph data.

Personalization: It is referred to the concept in which the user is provided results based not only on the keyword of interest entered, but it is rather tailored to an individual's interest based on the information extraneous to the inputted query. This section will be incorporated in the later stage of the project after the pattern graphs have been fetched from the underlying structured graph data based on the keyword of interest entered.

Dataset Used

Mondial Database

The MONDIAL dataset has been formed by integrating information from a number of data sources, covering key information about countries and geographic features. It consists of various nodes which indicate geographic data like cities, rivers, country, states, islands, mountains, deserts, organizations, etc. The data sources covered in the Mondial database are listed below:

- CIA World Factbook,
- A predecessor of Global Statistics which has been collected by Johan van der Heijden.
- Additional textual sources for coordinates,
- The International Atlas by Kümmerly & Frey, Rand McNally, and Westermann,
- And some geographical data of the Karlsruhe TERRA database. (Göttingen, n.d.)

The RDF model of Mondial has been used in the project. It consists of URI's (Uniform Resource Indicators) in the form of triples that are used to indicate relationships between structured, and semi-structured data.

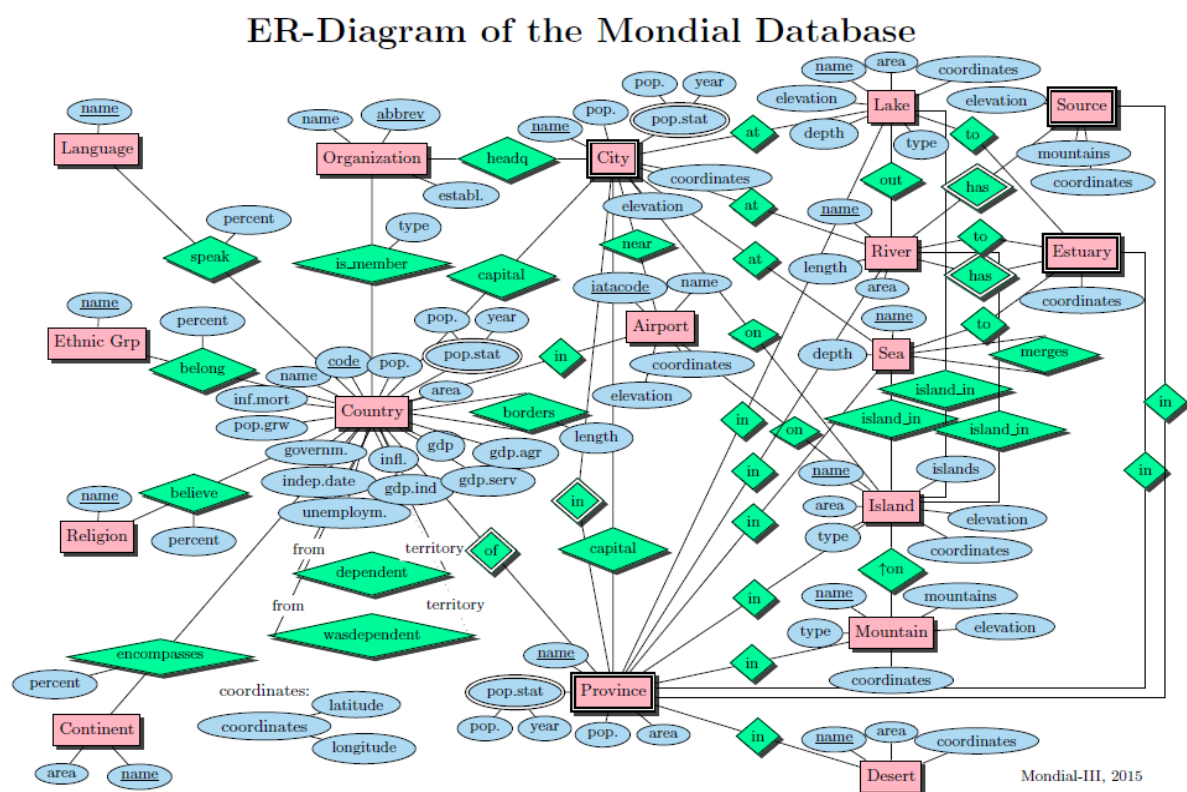


Figure 2 ER Diagram of Mondial Dataset (Informatics)

Structural Summary

Definition:

Structural Summary of an RDF Graph G is a special type of graph which summarizes the data graph showing vertices and edges corresponding to the class vertices and property, relationship and subclass edges in G. (Ananya Dass)

A structural summary extracts the following information from a RDF dataset. The entities, classes, properties, relationships, type edges, and values.

The edges possible in an RDF Graph are:

1. Type: occurs between an entity and a class
2. Property: occurs between an entity and a value
3. Relationship: occurs between two entities
4. Subclass: occurs between two classes

The vertices in an RDF Graph can be:

1. Entity: A node is an entity when the corresponding edge is a type between a class, and a property between a value. The nodes at the vertices of a relationship edge are also entities.
2. Class: A class node occurs where there is a type edge corresponding to an entity.
3. Value: A value node exists when there is a property corresponding to an entity.

Importance:

Structural Summary of an RDF Graph makes it possible to optimize result calculation for a pattern graph. The results obtained on finding a non-cyclic graph for matching constructs of every keyword in the query Q and the connections between them can be optimized using Structural Summary. These subgraphs are called result pattern graphs. The results obtained on finding result pattern graphs on a structural summary of an RDF graph vs the entire RDF graph in itself is the same. Structural summaries however provides efficient query optimization.

Extraction of Structure:

API

In the project MySQL is used to extract the Structural Summary of an RDF Graph. The Mondial dataset is in a RDF format. I used Java Programming to import the RDF data into a Relational Database table using the Apache Jena API.

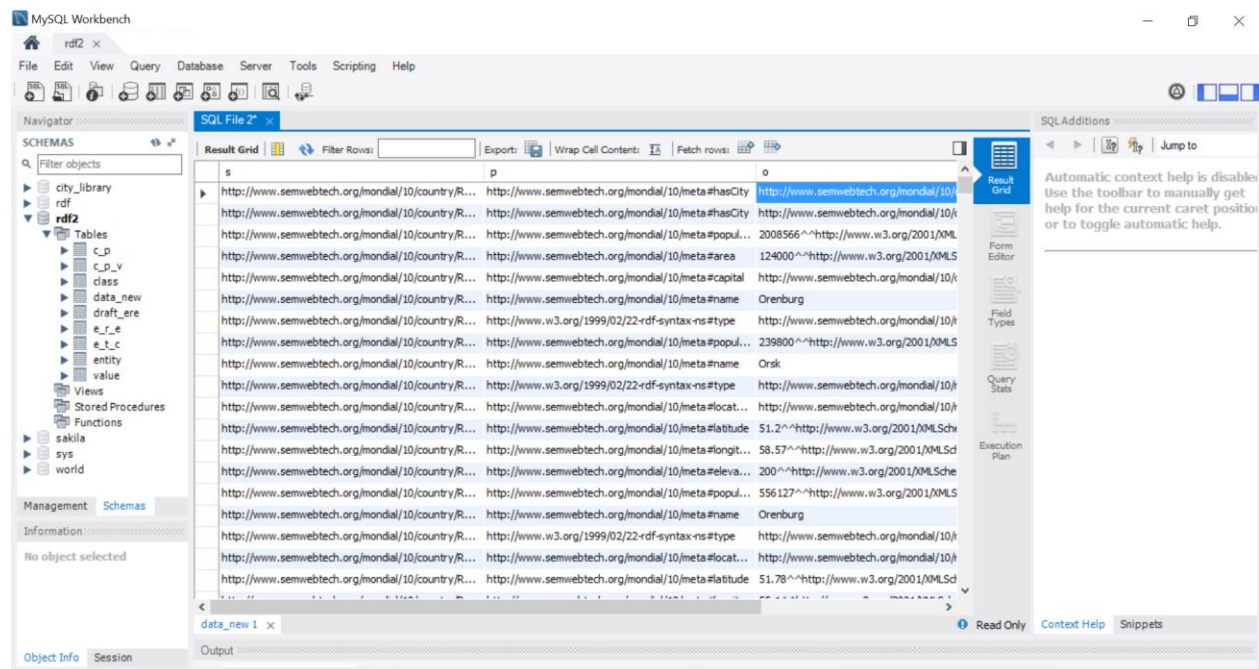
Apache Jena is an open source Semantic Web framework for Java. It provides an API to extract data from and write to RDF graph. These RDF graphs can be queried using the SPARQL (SPARQL protocol and RDF query language). (Apache Jena, n.d.) (Jena Framework, n.d.)

Jena Supports serialization of RDF graphs to:

- A relational database
- RDF/XML
- Turtle
- Notation 3

Software

Eclipse Neon was used for development in the Masters project. The development portion of the project involved importing RDF data into MySQL workbench stored in a relational database table. I used two programs for the same. The first program parsed the RDF data and extracted the URI triples. The other program involved adding the parsed triples in a relational database table under three columns. These columns were named s, p, o and were abbreviations for subject, predicate, object respectively. These columns corresponded to the triples of the RDF data.



s	p	o
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#hasCity	http://www.semwebtech.org/mondial/10/...
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#hasCity	http://www.semwebtech.org/mondial/10/...
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#popul...	2008566^http://www.w3.org/2001/XMLSchema
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#area	124000^http://www.w3.org/2001/XMLSchema
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#capital	http://www.semwebtech.org/mondial/10/...
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#name	Orenburg
http://www.semwebtech.org/mondial/10/countryR...	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.semwebtech.org/mondial/10/...
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#popul...	239800^http://www.w3.org/2001/XMLSchema
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#name	Orsk
http://www.semwebtech.org/mondial/10/countryR...	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.semwebtech.org/mondial/10/...
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#locat...	http://www.semwebtech.org/mondial/10/...
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#latitude	51.2^http://www.w3.org/2001/XMLSchema
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#longit...	58.57^http://www.w3.org/2001/XMLSchema
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#eleva...	200^http://www.w3.org/2001/XMLSchema
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#popul...	556127^http://www.w3.org/2001/XMLSchema
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#name	Orenburg
http://www.semwebtech.org/mondial/10/countryR...	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.semwebtech.org/mondial/10/...
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#locat...	http://www.semwebtech.org/mondial/10/...
http://www.semwebtech.org/mondial/10/countryR...	http://www.semwebtech.org/mondial/10/meta#latitude	51.78^http://www.w3.org/2001/XMLSchema

Figure 3 Relational table generated from RDF data

Method

The following table describes the relation of RDF vertices and edges in relation to the subject, predicate, and object columns of a relational database table.

Subject	Predicate	Object
Entity	Type	Class
Entity	Property	Value
Entity	Relationship	Entity
Class	Subclass	Class

MySQL queries are applied to the relational table to extract the summary and determine the relational database tables with the following columns:

1. Class, Property
2. Class, Property, Value
3. Class
4. Entity, Relationship, Entity
5. Entity, Type, Class
6. Value

Implementation

To determine the above mentioned tables MySQL queries are applied to the relational database table containing the RDF data in the form of 3 columns, Subject, Predicate, and Object.

I first extracted the values from the Object column. These are the values are not URI's (Uniform Resource Indicators). The edge corresponding to values are property edges in the predicate column. The values in the subject table correspond to entities. The results obtained are stored in the values table.

The type edges can be extracted by finding the '#type' ending in the URI's of the property column. The subjects of the type predicate are entities, and the objects are classes. The results obtained are added in the entity, type, class table.

The entities obtained in the previous two tables are added in the entity table. This table is used to obtain the entity, relationship, entity table.

The values obtained in an entity, type, class are used to extract the class table.

Results

The number of triples in the Mondial database are: 691396

The number of tuples in the value table are: 6045

The number of tuples in the entity, type, class are: 293329

The number of tuples in the class, property, value are: 7354

The number of tuples in the class table are: 73

The number of tuples in the entity table are: 163353

Setbacks

The entity, relationship, entity table is difficult to calculate due to the number of joins involved in the database.

Conclusion

The project focuses on extracting the Structural Summary, then analyzing it and applying the Pattern Graph on it. The later steps involve comparing the effectiveness of each algorithm on the pattern graph obtained. The optimal algorithm will be implemented for the data. Personalization will be applied on later stages of the project.

References

- Ananya Dass, C. A. (n.d.). Exploiting Sematic Result Clustering to Support Keyword Search on Linked Data. *WISE 2014, Part I, LNCS 8786, pp. 448-463, 2014.*
- Apache Jena*. (n.d.). Retrieved from <https://jena.apache.org/>
- Gaurav Bhalotia, A. H. (n.d.). Keyword Searching and Browsing in Databases using BANKS. *ICDE 2002: 431-440.*
- Göttingen, I. f.-A.-U. (n.d.). *Mondial* . Retrieved from <https://www.dbis.informatik.uni-goettingen.de/Mondial/>
- Guoliang Li, B. C. (n.d.). EASE: An effective 3 in 1 keyword search method for unstructured, semi-structured and structured data. . *SIGMOD Conference 2008:903-914.*
- Haixun Wang, C. C. (n.d.). A survey of Algorithms for Keyword Search on Graph Data. *Managing and Mining Graph Data 2010: 249 - 273.*
- Hao Hey, H. W. (2007). BLINKS: Ranked Keyword Searches on Graphs. *SIGMOD*.
- Informatics, G.-A.-U. G.-I. (n.d.). <https://www.dbis.informatik.uni-goettingen.de/Mondial/mondial-ER.pdf>. Retrieved from <https://www.dbis.informatik.uni-goettingen.de/Mondial/mondial-ER.pdf>
- Jena Framework*. (n.d.). Retrieved from [https://en.wikipedia.org/wiki/Jena_\(framework\)](https://en.wikipedia.org/wiki/Jena_(framework))