RichRDF: A Tool for Enriching Food, Energy, and Water Datasets with Semantically Related Facts and Images

Mohamed Gharibi, Praveen Rao, and Nouf Alrasheed

University of Missouri-Kansas City (UMKC), Kansas City MO, USA
mggvf@mail.umkc.edu, raopr@umkc.edu, nalrasheed@mail.umkc.edu

Abstract. Food, energy, and water (FEW) are the key resources to sustain human life and economic growth on Earth. While there is a plethora of information related to FEW systems online, there is a lack of reliable knowledge management tools that enable easy consumption of such information. In this paper, we present a web-based tool called RichRDF with the goal of enriching exiting FEW systems with semantically related facts and images. The main features of RichRDF include (1) an entity extraction algorithm that extracts meaningful subjects from Resource Description Framework (RDF) statements using natural language processing (NLP) techniques, (2) a reliable approach to add semantic similarity scores and relationships between different RDF subjects based on ConceptNet, (3) an efficient way to use the numbers of WordNet synsets to request the associated images from ImageNet, and (4) a user friendly interface that allows users to load and convert FEW datasets to RDF and then query the RDF datasets using an existing SPARQL engine. A video highlighting the key features of RichRDF is available at https://youtu.be/vyHgh4LgKCo.

Keywords: Knowledge Graphs, Food, Energy, and Water, RDF.

1 Introduction

Food, energy, and water (FEW) are the interdependent components that are undoubtedly imperative for our lives on Earth. Tremendous stress in these resources are expected by 2050 due to population growth, natural disasters, and human activities, which emphasize the need to improve available FEW resources. The United Nations has classified FEW components as a high priority within their sustainable development goals [1].

Meanwhile, there are several federal agencies such as the United States Department of Agriculture (USDA) and the National Drought Mitigation Center (NDMC) that provide massive amounts of data related to FEW systems. However, the available data exist in CSV, XML, and JSON formats that are not readily consumable in the world of Linked Data (LD).[1]

[1] http://linkeddata.org

Today, billions of RDF triples are available on the Web for developing new Semantic Web applications [1]. These triples are expressed as (*subject, predicate, object)* to represent entities and their relationships within a knowledge base. These relationships can be indicated by using different ontologies such as FOAF and DBpedia [2]. A specific Internationalized Resource Identifier (IRI) can be added as the context to RDF triples. Such statements are called as RDF quads. Here is an example of an RDF triple capturing the relationship between Oswego and the United States [2]:

```
<http://dbpedia.org/resource/Oswego>
<http://dbpedia.org/ontology/country>
<http://dbpedia.org/resource/United_States> .
```

The first term in the triple represents the subject, which is 'Oswego.' The second term is the predicate, which was provided by the DBpedia ontology, is 'country', and the last term of the triple represents the object, the 'United States.' Such triples can be generated from files in different formats including JSON, CSV, and TSV. Converting such files to RDF triples will allow users to express semantic relationships between subjects and objects and to structure information using RDF graphs. Moreover, these RDF triples will provide several benefits than processing raw files including the ease of integration and use, modeling of semantic similarity between entities, and the ability to query the knowledge base using a SPARQL engine.

While there are several tools that can produce RDF datasets (e.g., Karma [3]), they do allow us to easily add new assertions in the form of triples or quads to a dataset. Moreover, the lack of reliable knowledge graphs serving FEW systems has motivated us to build our own knowledge graph that helps in decision-making, enriching FEW datasets by providing extra knowledge and images based on the semantic similarities between the dataset entities, improving knowledge discovery, simplifying access, and providing better search results. Our system, RichRDF, employs RDF, Web Ontology Language (OWL), and SPARQL to construct and query the FEW knowledge graph.

2 RichRDF

The overall architecture of RichRDF is shown in Fig. 1. RichRDF has four stages of execution to ensure that the system can run under different conditions before it starts each stage so that the total processing time is minimized.

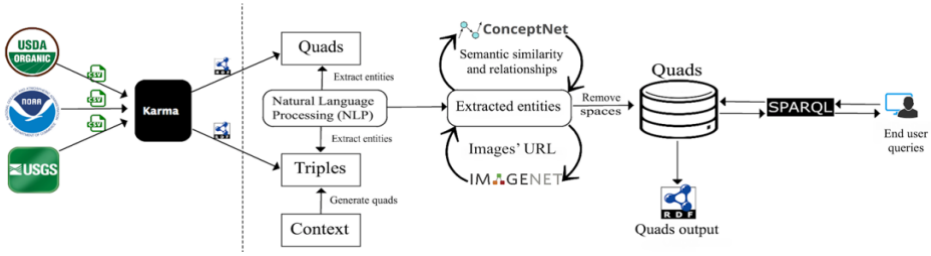

Fig. 1: Architecture of RichRDF

We assume that an input CSV file is converted into RDF using an integration tool such as Karma. The first stage of RichRDF checks the structure of the file uploaded by

a user. If the file contains RDF quads, then the file will be ready for the next stage of processing. Otherwise, the user is asked to provide a context IRI, which will be added as the context of each RDF triple to produce RDF quads. In the second stage, RichRDF runs NLP techniques [4] on the subjects, to extract the meaningful entities in order to use these entities later on for further processing. The main goal of entity extraction is to identify a real entity as the main subject, since subjects may be a word, a couple of words, long text, or a number like an ID. The extracted entities will subsequently represent an entire subject.

RichRDF processes two quads at a time, therefore, there are several possibilities during the entity extraction. Consider two subjects that end with the following strings "`CHEESE,COTTAGE,CRMD,W/FRU`" and "`BUTTER,PDR,1.5OZ,PREP,W/1/1.HYD`" After the entity extraction stage, "`CHEESE,COTTAGE`" denote the entities extracted from the first subject and "`BUTTER`" from the second subject. We use the "isA" relation to link the original subjects with the extracted entities. The third stage of RichRDF uses the extracted entities on ConceptNet [5] to obtain the relationship between these subjects. If a relationship exists between the first and the second subject, another request will be generated to fetch the semantic similarity score. We use RDF reification/blank nodes to represent the similarity score between the original subjects. The semantic similarity score and the relationship between subjects can be exploited during information retrieval and NLP, and it also expands the search to ontology keywords. Furthermore, these scores can be used in advanced machine learning techniques to understand the data in a better way and to build better models [6].

At the last stage, RichRDF queries WordNet using the extracted entities to generate synset groups of words [7]. Using these synset groups, we generate the offset ID numbers based on their relationship with the subjects. The offset ID numbers are used to look up images on ImageNet [8]. For every offset ID, we request ImageNet to provide us with the Uniform Resource Locators (URLs) of all the images that are associated with the subject ID number. As a result, hundreds of URLs will be returned. Instead of adding all these URLs using blank nodes, we split them into two categories based on their content. The first category is a single blank node with the relationship "`IURLs_subjectName`" that contains a link to a page containing hundreds of images related to this subject. The second category contains the pure images that represent the subject only. We add the second category as multiple blank nodes using the relationship "`subjectName_Images`". Finally, the user will be able to download the output at this stage or RichRDF can provide the user with another service to query the output using a SPARQL engine with a user-friendly interface. The user will be able to download the output of the queries at any time.

Performance Evaluation. We report the performance evaluation results of RichRDF to provide insights on its speed. Fig. 2 shows the best-case and worst-case time taken for RichRDF while running an input file containing 1,000 triples. This file was processed in four different rounds where we added a different feature in each round. Table 1 shows the time taken for different numbers of triples for various datasets.

We would like to mention that the execution time depends on the triples in the input RDF dataset. Richer the dataset, i.e., containing commonly used entities such as food types, fruits, brand name, objects names, etc., more would be the time taken to process

all the relationships, semantic scores, and obtain the relevant images. In the future, we plan to leverage concurrency and parallelism to speed up RichRDF.

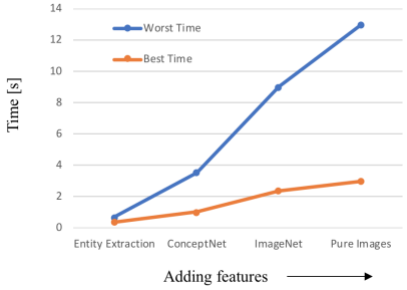

Fig. 2: Four execution rounds of RichRDF

Table 1: Number of triples with the average execution time

Type	Triples	Time required
Food	1,000	5.05 seconds
Energy	10,000	46.13 seconds
Water	100,000	8.43 minutes

To conclude, RichRDF is a new tool for modeling FEW datasets using Semantic Web technologies to enable easy consumption and analysis of FEW information for intelligent decision making. In the future, we would like to automatically publish the RDF data produced by RichRDF on Linked Data. The source code of RichRDF is available at `https://github.com/UMKC-BigDataLab/RichRDF`.

Acknowledgments: The first author (M. G.) would like to thank the support of UMKC SGS Travel Grant.

References

1. P. Rao, A. Katib, D. Barron.: A knowledge ecosystem for the food, energy, and water system. In KDD 2016 Workshop on Data Science for Food, Energy and Water, pp. 1-4, San Francisco, 2016.
2. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives.: DBpedia: A nucleus for a web of open data. In the Semantic Web Lecture Notes in Computer Science, pp. 722-735. Springer, Berlin, 2007.
3. C. Knoblock, P. Szekely.: Exploiting semantics for big data integration. In AI Magazine, Vol. 36, no. 1, 2015.
4. C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky.: The Stanford CoreNLP natural language processing toolkit. In Proc. of 52nd annual meeting of the Association for Computational Linguistics, pp. 55-60, 2014.
5. R. Speer, C. Havasi.: ConceptNet 5: A large semantic network for relational knowledge. The People's Web Meets NLP, pp 161-176. Springer-Verlag Berlin, 2013.
6. M. Pham, S. Alse, C. Knoblock, P. Szekely.: Semantic labeling: A domain-independent approach. In International Semantic Web Conference (ISWC), pp. 446-462, Kobe, 2016.
7. M. Hsu, M. Tsai, H. Chen.: Combining WordNet and ConceptNet for automatic query expansion: A learning approach. Information Retrieval Technology, vol. 4993, pp 213–224, Springer, 2008.
8. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, L. Fei-Fei.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision, Vol. 115, pp. 211-252, Springer, 2015.