Kalpa Publications in Computing

Volume 10, 2019, Pages 38–42

Selected Student Contributions and
Workshop Papers of LuxLogAI 2018

Inducing Schema.org Markup from Natural Language Context

Gautam Kishore Shahi[1], Durgesh Nandini[1], and Sushma Kumari[2]

[1]Dipartimento di Ingegneria e Scienza dell'Informazione (DISI),
University of Trento, Italy
[2]Deparment of Computer Science & Engineering, Ramgarh Engineering College,
Jharkhand, India
(gautamshahi16, sushma031996)@gmail.com, durgeshnandini16@yahoo.in

Abstract

Schema.org creates, supports and maintain schemas for structured data on the web pages. For a non-technical author, it is difficult to publish contents in a structured format. This work presents an automated way of inducing Schema.org markup from natural language context of web-pages by applying knowledge base creation technique. As a dataset, Web Data Commons was used, and the scope for the experimental part was limited to RDFa. The approach was implemented using the Knowledge Graph building techniques - Knowledge Vault and KnowMore.

1 Introduction

On the web, many websites nowadays come with Schema.org[1] annotations. As of 2016, more than 31% of Alexa ranked websites contain data elements annotated with Schema.org [1]. Among these, were popular community-driven sites such as Blogspot, Wordpress, and Wikipedia. For sites like these, the quality of content is dependent on the diligence of the individual author, of which (i) there are many, and, (ii) most are hobbyists. The average content is lacking in detail and completeness (among others). As such, we assume the same should especially hold for semantic annotations within web pages, particularly, since the average author may not be aware well enough about Schema.org or other vocabularies suitable for these kinds of annotations.

Schema.org is a lightweight web vocabulary with a rather flat class hierarchy. At its core, it currently contains roughly 600 classes, 10 data types, and 1000 attributes and relations of general purpose for data on the web. Beyond that, it also comes with five domain-specific extensions, covering automotive, bibliographic, Internet of Things, health and metamodelling related domains. As with other Resource Description Framework (RDF) [2] vocabularies, Schema.org, some particular formats may be used for embedding annotations within web resource. Today, three of the most widely used formats are: Microdata[2](58% of annotated pages), RDFa[3] (36%

[1]https://schema.org/
[2]https://www.w3.org/tr/microdata/
[3]https://www.w3.org/tr/xhtml-rdfa-primer/

of annotated pages) and JSON-LD[4] (7% of annotated pages). RDFa uses simple XHTML attributes to categorises the data, it was in abundance, and it is a standard for Semantic Web data. Microdata, however, is not a standardised serialisation for RDF and does not have a prefix mechanism for comfortably using multiple vocabularies at once. JSON-LD, on the other hand, often, is found as blobs embedded at the roots of documents. Hence, for this work, we will limit our scope to embedded RDFa only.

Remaining sections of the paper are organized as follows: Section 2 discusses the methodology, Section 3 discusses implementation and experimental results, Section 4 discusses conclusion and future work.

2 Methodology

This section describes the proposed methodology for inducing Schema.org to form the local context of a web-page.

- **Data Preparation** Data was comprised of three parts: (i) HTML documents containing Schema.org markup in RDFa, The HTML contents should have been curated by non-professional authors, such that one may assume the contents to be lacking in quality of the semantic markup. To attain such documents, in our selection we turn to blogs and wikis as the source for our HTML. (ii) RDF Triples parsed from the HTML documents, stemming from the RDFa markup. (iii) Information about the location of individual RDF triples within the respective document object model (DOM) tree, For instance, encoded via XPath (i.e. location's information pointing to containing HTML element).

- **Adapting RDFa Parser** To attain location information of an RDF triple, the RDFa parser needed to be adapted, such that it returned the location information for any RDF triple it was parsing.

- **Induction Approach** For the initial stage, fusion-based technique was chosen for creating Knowledge Graph by inducing the natural language context. The Knowledge Vault [3] and KnowMore [4] approach has been applied.The Knowledge Vault works in automatic knowledge base constructions to combine noisy extractions from the web together with prior knowledge obtained from the existing Knowledge Base. This approach was analogous to speech recognition, which combines noisy acoustic signals with prior derived from language model. Knowledge Vault can overcome errors occurring due to the extraction process as well as an errors in the sources themselves.

  KnowMore represents a novel data fusion approach addressing the issues uncovered by Knowledge Base creation, through a combination of entity matching and fusion technique geared towards specific challenges associated with markup.

- **Creation of a test and train dataset** From the collected data, we applied K-fold cross validation was applied to obtain a fair result from the knowledge graph building technique, K=5 was considered for validation.

[4]https://www.w3.org/tr/json-ld/

- **Comparative Analysis** We apply our approach of Schema.org markup induction to the prepared test data. We can now measure the extent to which the gold standard and the induced markup are in agreement. For instance, by measuring the similarity of types annotated according to the gold standard and the types induced from context.

3 Implementation & Experimental Result

The Common Crawl[5] triggered the area of making crawl data available to public. It has a collection of crawled data from all the Web. As an extension, in 2012 another data hub called Web Data Commons [5] came up with structured data extracted from the Common Crawl. The Web Data Commons[6] had billions of triples, stored in the form of N-Quads, a set of s, p, o (Subject, Predicate and Object and the URL from triples have been generated). For our experiments, the Web Data Commons was preferred as the data source due to its vast collection of structured data. The dataset was available in four formats- RDFa, Microdata, Microformats and Embedded JSON- LD. The data was available in WARC[7] format; therefore, we extracted the file to N-Quads format and stored it.

We collected 600 RDFa WARC files, the total size amounting to 60 GB. To maintain diversity in our domain, we chose the dataset by sampling. We randomly sampled the RDFa files and chose 60 WARC files out of 600 files. The size of the data used was 6.29 GB which was roughly 10% of the size of the RDFa data available on the Web Data Commons, the selected dataset contains around 20.6 millions of triples and is stored data in SPARQL is an RDF query language SPARQL [6] Endpoint, for instance, Blazegraph[8]. Using the SPARQL query, we selected the list of URLs which contains more than ten distinct properties and ten distinct resources.

For the selected 60 WARC files, we used 10000 URLs containing around 60000 Triples and stored it in a JSON File. For further processing, from the JSON file, each URL was passed to an HTML reader to download the HTML content. During the process, there were several exceptions that had to be handled to obtain a maximum list of HTML contents, the most common exceptions being the 'Http File not found' and the 'HttpConnectionPool'. As a result, we could successfully parse 6000 out of 10000 URLs. The HTML contents of each URI was then to a parser to extract RDFa data to get the output as triples. We used two kinds of data storage, one was SPARQL Endpoint to store the triples generated by the RDFa parser and another was MySQL database to store URL and HTML content and the exceptions occurred. Figure 1 represents the Data Processing steps

The collected RDF triples were used for inducing Schema.org markup from Natural Language context using the KnowMore and Knowledge Vault approach.

The comparative analysis for KnowMore and Knowledge Vault and obtained Accuracy and F1 score has been shown in Table 1. The results represent that the proposed approach can be implemented on a large scale data set of RDF triples and a novel method can be introduced to improve the accuracy.

[5]http://commoncrawl.org/

[6]http://webdatacommons.org/

[7]The WARC file format is proposed by the Internet Archive foundation as successor to the ARC file format – http://archive-access.sourceforge.net/warc/.

[8]https://www.blazegraph.com/

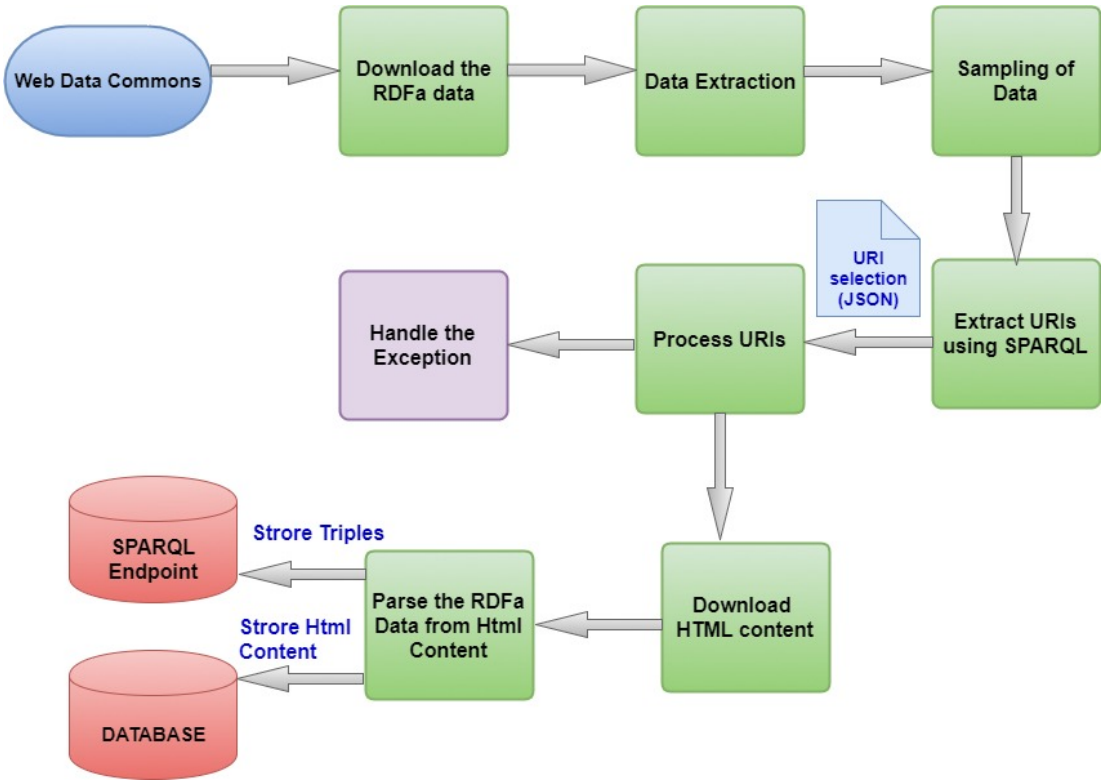

Figure 1: Data Processing Diagram

Table 1: Experimental Reuslts

Approach	Accuracy	F1 score
KnowMore	76.96	0.91
Knowledge Vault	63.24	0.82

4 Conclusion and Future Work

In the paper, we presented our current work on inducing Schema.org markup from Natural Language Context. The RDFa format from the Web Data Commons corpus was used as a data set and the approach was implemented using the Knowledge Base creation methodology (Knowledge Vault and KnowMore). As a result, a satisfactory accuracy was obtained which represents the validity of our approach for inducing Schema.org markup.

As an extension of this work, we propose implementing the methodology on larger data sets and comparing results with other Knowledge Base creation methodologies. A further extension would be coming up with a novel approach for inducing Schema.org from the local contents of the web page using machine learning techniques.

References

[1] Ramanathan V Guha, Dan Brickley, and Steve Macbeth. Schema. org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51, 2016.

[2] Graham Klyne and Jeremy J Carroll. Resource description framework (rdf): Concepts and abstract syntax. 2006.

[3] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610. ACM, 2014.

[4] Ran Yu, Ujwal Gadiraju, Besnik Fetahu, Oliver Lehmberg, Dominique Ritze, and Stefan Dietze. Knowmore-knowledge base augmentation with structured web markup. *Semantic Web Journal, IOS Press*, 2017.

[5] Hannes Mühleisen and Christian Bizer. Web data commons-extracting structured data from two large web corpora. *LDOW*, 937:133–145, 2012.

[6] Eric Prud, Andy Seaborne, et al. Sparql query language for rdf. 2006.