Usability and User Experience, Vol. 39, 2022, 482–489
https://doi.org/10.54941/ahfe1001745

A Framework for Data Mining of Structured Semantic Markup Extracted From Educational Resources on University Websites

Lorena Recalde, Rosa Navarrete, and Luis Rosero Correa

Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Quito, Ecuador

ABSTRACT

Currently, there is a vast availability of educational resources on the Web, mainly published by universities. Users search them using search engines, but their obtained results are inaccurate, affecting the quality of user experience. The embedded structured semantic markup in the HTML content is a mechanism to enrich the meaning in search results. This research proposes a framework that enables analyzing the top-ranking universities' websites to explore the degree of adoption of the structured semantic markup that uses Schema.org vocabulary notated in JSON-LD format. The dataset for analysis is collected through Web Scraping techniques, and data mining strategies are used to describe and organize the educational resources obtained. Once the framework has been evaluated and deployed, the obtained results have shown an elemental use of structured semantic markup on educational resources and imprecise use of the vocabulary available in Schema to describe them. The research has proved the framework's validity and its capability to be extended to analyze other areas of interest.

Keywords: User experience, Data mining, Structured semantic markup, Educational resources, JSON-LD

INTRODUCTION

The Coronavirus pandemic has forced education at all levels to change from face-to-face mode to online learning (Daniel, 2020). In keeping with that purpose, Universities are releasing a significant number of educational resources on the Web to support virtual education. Final users, who need these educational resources, explore the Web through search engines such as Google, Yahoo, Yandex, or Bing; but, the search results they obtain lack accuracy and are not necessarily adequate to their requirements (Navarrete et al. 2019). To improve the user experience, embedding structured semantic markup into the HTML of web pages may deliver more appropriate content in response to searches. Search engines can interpret this markup to understand the resources being published and, consequently, improve the correctness of search results (Bakhouyi et al. 2019). For example, Google uses the structured semantic markup to show rich fragments, Rich Snippets, or even Knowledge Graphs in user searches (Ohshima and Toyama, 2018).

This research proposes a framework that enables a systematic analysis of the structured semantic markup of the educational content published by top-ranking universities. Then, by using Web Scraping techniques, analyzes these universities' websites in search of educational resources and reviews if the structured markup is embedded. Finally, it uses data mining techniques to describe and organize the educational resources obtained. The contribution of this work is two-fold. The first contribution is the analysis of structured semantic markup in universities' websites that use Schema vocabulary and JSON-LD format to find how this technology is used. This analysis is relevant since previous research has not explicitly focused on the educational field or has not used a specific dataset within this context. The second contribution is a three-layer framework that allows accomplishing this type of analysis of embedded semantic markup from a data collection phase to obtaining results and indicators on the data. The remainder of the paper is organized as follows: Background Section summarizes the context of the present work; Framework Section presents the framework layers and components; Results Section details the obtained findings; finally, the last section outlines the conclusions and future work.

BACKGROUND

Several works have been carried out regarding the use of structured markup in the educational field; for instance, in (Ambite et al. 2019), to describe the content of educational resources related to Data Science, Machine Learning techniques are employed in the labeling task. In (Bakhouyi et al. 2019), the use of semantic Web technologies is proposed to develop an intelligent Web where machines can understand information to improve interoperability between e-Learning systems such as Moodle. (Georgescu, 2019) proposes a system for semantic indexing of documents related to cybersecurity by using natural language processing (NLP) techniques. The purpose is to facilitate the documentation process of cybersecurity topics. In (Tavakoli et al. 2020), the authors propose a prototype of an open educational resources (OER) recommender system, based on the metadata they contain, that aims to support the development of students' skills to fill the job market needs related to data science.

As can be seen, some works propose using structured semantic markup to improve the user experience with more accurate results for their Internet searches. However, according to the related literature, structured semantic markup in Web content is not appropriately used. In (Navarrete et al. 2019), a study on the use of structured semantic markup with Microdata and JSON-LD formats to describe educational resources is presented. This study shows that employing structured markup to describe Web resources is not a very common practice, especially in the educational field. The fact that structured semantic markup is not correctly used to describe educational resources on the Web is worrying considering that education is increasingly Web-oriented. Given this background, we propose a framework that allows

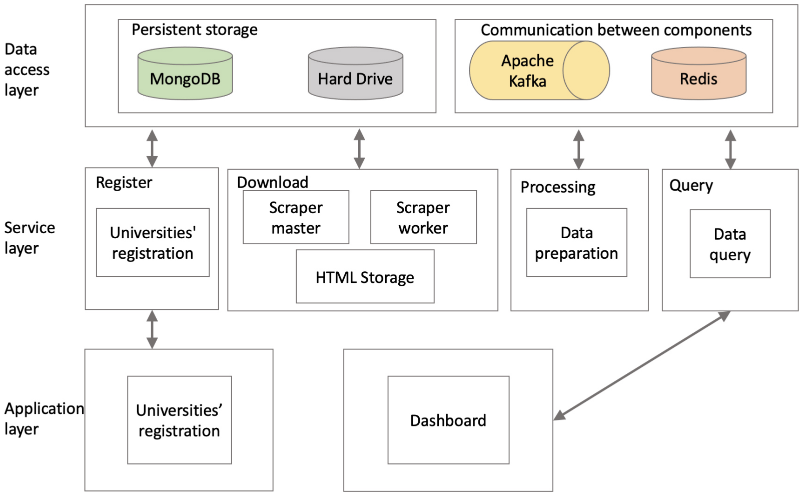

Figure 1: Overview diagram of the framework.

an analysis of universities' websites in the top international ranking and focuses on how they use structured semantic markup with the syntax JSON-LD and the Schema vocabulary[1].

FRAMEWORK

The framework proposed in this research is divided into three layers which in turn group several components according to the role they play within the framework (Figure 1). The layers defined for the framework are the following.

- Application Layer: It contains the components that allow interaction with the framework. These components are the URL entry application and dashboard.
- Service layer: It groups the components that execute tasks corresponding to reading, writing, and processing data. The components that make up this layer are URL entry service, Downloader service (scraper master, scraper worker, and HTML storage service), data processing service, and Data query service.
- Data access layer: It groups those components whose function is to provide a space and a form of data storage. The components that make up this layer are MongoDB, Redis, Apache Kafka, and HDD.

Application Layer

URL entry application. It consists of a Web application that allows to enter the URLs of the Web sites of the universities from which we want to download the data. Therefore, prior to data collection, the sources from which

[1] JSON-LD is a W3C recommendation. It adds a script element used as a data block separately from the existing markup. Schema.org defines the vocabulary. It was created in 2011 by the major search engines, Google, Bing, Yahoo, and Yandex. It provides terms for describing a wide variety of entities and integrates other vocabulary and standards (Navarrete et al. 2020).

the extraction will be made need to be defined as an initial step. Considering that the universities positioned in the top international rankings must be at the forefront of technology issues, it was determined that the websites of each of these universities would be a good source for obtaining data. The list of the top 150 universities in the top ranking was taken from the QS World University Rankings (*https://www.topuniversities.com/university-rankings/world-university-rankings/2021*). For our study, we considered the 100 universities from which the most data was downloaded. This list was entered into an Excel file. The first column has the name of the university, and the second has the URL of the university's website. When the file is ready, we use the URL entry application (Web form) to upload the file and send it to the URL input service.

Dashboard. This component belongs to the application layer and interacts with the data query service. The dashboard presents the data resulting from the framework activities. It allows to analyze the data interactively by applying filters. The Results Section shows the role of this element of the application layer.

Services Layer and Data Access Layer

URL input service. This component interacts with the application layer because it processes the file loaded through the URL entry component. Once the file is received and verified its correct structure, the data is processed and saved in a collection called universities in MongoDB. It has the fields i) id to identify the university, ii) name and iii) URL of the university, and iv) a flag that controls if the data download for the given URL has been executed or not.

Download service. The collections in MongoDB are used by the download service for reading and storing data. First, to download data, the *universities collection* is required. The universities HTML data is obtained by using Web Scraping (Python); two components are responsible here: the scraper master and the scraper workers. The scrapers use the *scrapedPages collection* of MongoDB to particularly save a list of the structured markup in JSON-LD format when it is obtained from the HTML of the Web page.

The scraper master starts the Web Scraping process and opens the scraper workers. First, the scraper master queries the *universities collection*. For each university, a Web Scraping service is generated within the scraper master to *1)* retrieve the URL of the home page of the university website; *2)* request of the web page; *3)* check the validity of the response to the request; *4)* retrieve the HTML code of the web page; *5)* extract data from HTML; *6)* store the extracted data in the *scrapedPages collection* in MongoDB; *7)* send data to a topic in Apache Kafka for the storage of the Web pages' HTML in its corresponding directory on hard disk (HTML storage component); *8)* check if links to other pages extracted from HTML have been processed before, and; *9)* send data for the processing of the second phase of Web Scraping. The scraper worker corresponds to the second phase of the Web Scraping process. It uses the URLs obtained by the scraper master. This second phase of Web Scraping was created to process several instances in *parallel*. Then,

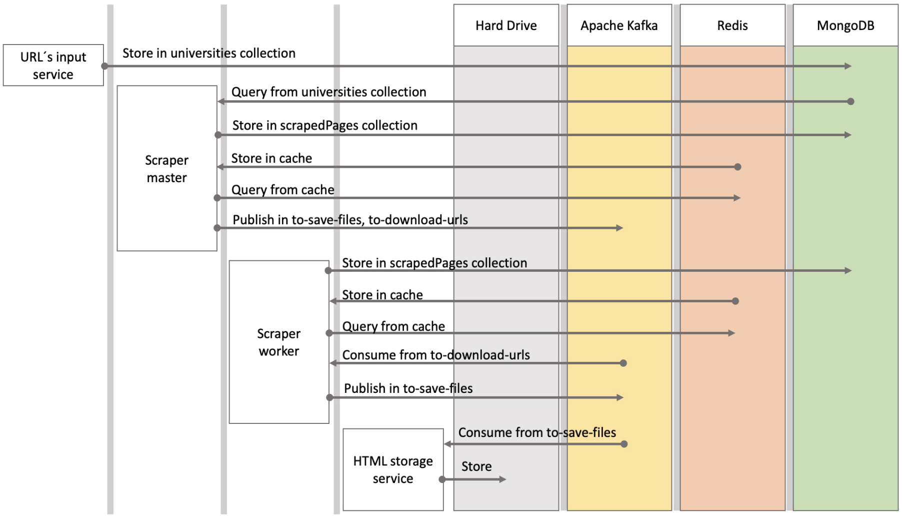
Hard Drive
Apache Kafka
Redis
MongoDB
URL´s input service
Store in universities collection
Scraper master
Query from universities collection
Store in scrapedPages collection
Store in cache
Query from cache
Publish in to-save-files, to-download-urls
Scraper worker
Store in scrapedPages collection
Store in cache
Query from cache
Consume from to-download-urls
Publish in to-save-files
HTML storage service
Consume from to-save-files
Store

Figure 2: Download service components and their interaction with the data access layer.

data download process is optimized. The activities carried out by the scraper workers are the same as the master (from 1 to 8), except for *9)* for which it iterates over Web Scraping process by sending data within the same service until a stop condition is met.

When URLs are retrieved, the scrapers check if they have already been processed to avoid duplicate downloads. This validation is done by querying Redis for the existence of the URL. Then, only in case the URL does not exist in Redis it is sent to processing. In addition, the HTML storage component stores the HTML of the pages on the hard disk, creating a directory by each university. The HTML storage service is directly related to Apache Kafka. Inside the storage service, a loop is executed that queries and retrieves data from a Kafka topic. The recovered data is used to validate the existence of the directory in which the file is going to be saved. If the directory does not exist, it is created. Once the destination directory is defined, the file is written to save it on disk. Figure 2 shows the interaction of the components of the service layer with the components of the access layer.

Data processing service. It belongs to the services layer of the framework and allows performing data cleaning and processing tasks to prepare them for analysis. It has a communication scheme with MongoDB to store preprocessed data.

Data query service. This component is responsible for communicating with the MongoDB database to perform queries and provide data to the dashboard in the application layer. Once the service receives the request, it creates the structure of a general query; then, it performs validations to determine if the queries to be made will include all the data or if filters must be applied.

RESULTS

The results showed the use of embedded structured markup with the Schema vocabulary and the JSON-LD format in published educational resources.

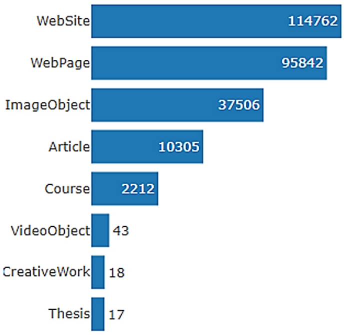

Figure 3: Valid educational resources obtained.

For the framework setup, one hundred universities' websites were considered (top ranking of world universities provided by QS World University Rankings). Once the framework was deployed, it was possible to download 1,019,268 Web pages, of which 195,098, corresponding to 19.1% of the total pages, were considered valid since they made use of the Schema vocabulary and the JSON-LD format for the description of resources. The remaining 80.9% corresponded to 824,170 pages that do not use Schema or JSON-LD, so they were considered invalid for analysis. The dashboard in the application service showed the three university websites from which the largest number of pages were downloaded: University of Southern California with 39950 pages, University of Oxford with 39791 pages, and Harvard University with 34761 pages. The valid Web pages (195,098 documents) were processed (data processing service), and we obtained 645,613 resources described using Schema and JSON-LD. Also, 47 Schema vocabulary classes were found, from which, we considered *CreativeWork*, *WebSite*, *Article*, *Course*, *Book*, *WebPage*, *ImageObject*, *VideoObject*, and *Thesis* as valid for the description of educational resources. Therefore, the number of educational resources found was 260705, which corresponds to 40.4%. Figure 3 shows the distribution of values per educational resource.

Finally, Table 1 presents the details of the most used properties when describing the resources associated with the classes selected for this analysis. The URL property was the one most frequently used in the description of educational resources.

Discussion. The results showed that effectively, the structured markup with JSON-LD and Schema vocabulary is used in educational resources, but not to the extent that would be expected. Schema classes that are commonly used to describe educational resources were found; but not those properties that better represent the semantic for this type of content. Indeed, three axes help define an educational resource more precisely: educational value, license, and accessibility (Navarrete et al. 2019). Each of these axes has its own properties that are more specific to an educational context. Their use was not evidenced in the resources found on the universities' websites analyzed.

Table 1. Most used properties for the description of educational resources.

Property	Number of resources	Percentage
url	245948	15,029%
name	214718	13,120%
inLanguage	204371	12,488%
potentialAction	167489	10,234%
description	125014	7,639%
datePublished	103003	6,294%
dateModified	102989	6,293%
isPartOf	102465	6,261%
breadcrumb	59719	3,649%

CONCLUSION

In this work, a framework to understand if universities publish educational resources on their websites using embedded structured markup was proposed. The three-tier structure of the framework presents components which interact with each other to fulfill a specific task. The results obtained after the data analysis showed that the universities do not use the embedded structured markup with the Schema vocabulary and the JSON-LD format to the extent that would be expected. In fact, even though 88% of the universities use structured markup with Schema and JSON-LD, from the total number of Web pages downloaded, only 19.1% contained the said structured markup. Schema classes associated with educational resources such as *WebSite*, *WebPage*, *ImageObject*, *Article*, *Course*, *VideoObject*, *CreativeWork*, and *Thesis* were found, but specific properties such as *educationalAlignment* or *educationalUse* were not. As future work, we plan minor modifications in the framework structure to focus on other contexts different from the educational one. Moreover, we can adapt the framework to other vocabularies and formats. So, it can be concluded that the framework may be extensible and easily generalizable.

REFERENCES

Ambite, J.L., Gordon, J., Fierro, L., Burns, G. and Mathew, J. (2019). Linking educational resources on data science. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *July*. 33(01), pp. 9404–9409.

Bakhouyi, A., Dehbi, R., Banane, M. and Talea, M. (2019). A semantic web solution for enhancing the interoperability of e-learning systems by using next generation of SCORM specifications. In: *International Conference on Advanced Intelligent Systems for Sustainable Development, July*. Springer, Cham, pp. 56–67.

Daniel, S.J. (2020). Education and the COVID-19 pandemic. *Prospects*, 49(1), pp. 91–96.

Georgescu, T.M. (2019). Machine learning based system for semantic indexing documents related to cybersecurity. *Academy of Economic Studies. Economy Informatics*, 19(1), pp. 5–13.

Navarrete, R., Recalde, L., Montenegro, C. and Luján-Mora, S. (2019). Analyzing embedded semantic with JSON-LD and Microdata for educational resources in large scale web datasets. In: *2019 International Conference on Computational Science and Computational Intelligence (CSCI) December*. IEEE, pp. 1133–1138.

Navarrete, R., Montenegro, C. and Recalde, L. (2020). Systematic Mapping on Embedded Semantic Markup Validated with Data Mining Techniques. In: *International Conference on Applied Human Factors and Ergonomics, July*. Springer, Cham, pp. 384–391.

Ohshima, T. and Toyama, M. (2018). SDC: structured data collection by yourself. In: *Proceedings of the 8th International Conference on Information Systems and Technologies, March*. pp. 1–8.

Tavakoli, M., Faraji, A., Mol, S.T. and Kismihók, G. (2020). OER recommendations to support career development. In: *2020 IEEE Frontiers in Education Conference (FIE), October*. IEEE, pp. 1–5.