View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repositorio Universidad de Zaragoza

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 155-158 recibido 21-03-22018 revisado 13-04-2018 aceptado 03-05-2018

Open Data for Public Administration: Exploitation and semantic organization of institutional web content

Datos Abiertos para la Administración Pública: Explotación y organización semántica del contenido web institucional

Paula Peña, Rocío Aznar, Rosa Montañés, Rafael del Hoyo
Grupo de Big Data y Sistemas Cognitivos
ITAINNOVA (Instituto Tecnológico de Aragón)
C/ María de Luna, nº 7. 50018 Zaragoza
{ppena,raznar,rmontanes,rdelhoyo}@itainnova.es

Abstract: The project presented has been financed by Government of Aragon and is part of the 'Open Data' initiative promoted by that organization. Given the amount of unstructured information related to the Government of Aragon currently published on the Internet, with slightly or no standardization and decentralized, it emerges the need to gather it systematically to be offered to all interested collectives from a single access point in a public and structured way. Within this context, 'Aragon Open Data' project aims to collect, organize, store and maintain updated, Administration's web information by means of human language and semantic technologies. Firstly, crawling is performed over websites in order to retrieve textual data over which Natural Language Processing (NLP) and ontology-based techniques are applied. Thereafter, results are stored into NoSQL databases, allowing future open access and simple data exploitation. NLP techniques used in the project involve named-entities recognition and classification (NERC) and texts semantic classification and summarization.
Keywords: Open Data, ontologies, natural language processing, crawlers

Resumen: El proyecto presentado, financiado por el Gobierno de Aragón, se enmarca dentro de la iniciativa de 'Open Data' promovida por dicho organismo. Dada la cantidad de información no estructurada relacionada con el Gobierno de Aragón, publicada en Internet de forma no estandarizada y descentralizada, surge la necesidad de recopilarla sistemáticamente para ser ofrecida a los colectivos de interés desde un único punto de acceso, pública y estructuradamente. En este contexto el objetivo del proyecto 'Aragón Open Data'es extraer, organizar, almacenar y mantener actualizada la información web de la administración, mediante el uso de tecnologías semánticas y del lenguaje. Concretamente, se realiza un crawling exhaustivo de páginas web para extraer los datos textuales sobre los cuáles se aplican técnicas basadas en ontologías y de procesamiento de lenguaje natural (PLN). Finalmente se almacenan los resultados en bases de datos NoSQL, permitiendo su futura explotación de manera sencilla, abierta y transparente al ciudadano. Las técnicas de PLN utilizadas en el proyecto incluyen el reconocimiento y clasificación de entidades nombradas (NERC) y la clasificación semántica y resumen de textos.
Palabras clave: Open Data, ontologías, procesamiento del lenguaje natural, crawlers

1 Introduction

Nowadays, Open Data is a worldwide movement whose philosophy aims to provide data openness and availability to citizens. Many countries have introduce an Open Data Initiative using government data[1]. Particularly, in Europe, Open Data has positioned itself as a focus of interest among policymakers for

[1] "Open Data Barometer Global Report". Available on https://opendatabarometer.org

ISSN 1135-5948. DOI 10.26342/2018-61-21
© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

over a decade[2].

Aragón Open Data is a project framed by the agreement of July 17, 2012 of the Government of Aragon guided by the core idea of opening public data to the general public in order to be accessed, used and shared in a standard and valuable way. Its Internet portal `http://opendata.aragon.es` was publicly presented with the final purpose of creating economic value within the ICT sector through the reuse of public information, increasing Administration's transparency, promoting innovation, improving information systems of the Administration, adopting technical standards in the information society field and generating data interoperability between public sector websites.

Aragonese public administration presents a complex casuistry in their data generation processes that is reflected in the proliferation of a large number of websites, subdomains, portals and downloadable documents in heterogeneous formats under its root domain: `aragon.es`. This circumstances make it difficult to easily access and make use of the information by users and Government services as well, generating the popular sense of certain lack of transparency from the public administration.

In order to deal with these issues, a solution is required to allow data and institutional information that is currently dispersed, non-homogeneous, uncontrolled and non-exploitable, is converted into structured data that can be analyzed together, be accessible, be browsed by related concepts and be exploited and served to third parties (other institutional websites, media, developers or citizens). For this reason, an open data solution has been implemented. It retrieves all the institutional information published over the Internet by means of web crawling or *spider* techniques within the existing domains of the Government of Aragon. Data analysis and structuring is performed by NLP techniques, such as semantic classification, NERC or summarization, combined with the use of an ontology designed and implemented within the context of the project, this is, the Interoperable Information Scheme of Aragon (EI2A). Likewise, the data obtained throughout this process will be used to verify and, if necessary, enrich the operation of the ontology, leading to a continuous update of the system knowledge. Therefore, EI2A emerges with the main idea of generating a framework in which the open data and regional government data in general, can begin to be automated in a much deeper way.

[2] "The Re-use of Public Sector Information Regulations". Available on http://www.legislation.gov.uk/uksi/2015/1415/made

2 Proposed Approach

In this section, the proposed approach followed to develop the system of capture and exploitation of the institutional web content is described. The functional design of the system is visually displayed in Figure 1. It is focused on three main modules:

1. Textual information retrieval from websites under `aragon.es` domain.
2. Natural language processing (NLP) techniques application on extracted data.
3. Results storage into NoSQL databases, conforming EI2A structure.

This process is executed periodically, allowing exploitation and query of updated results in real time.

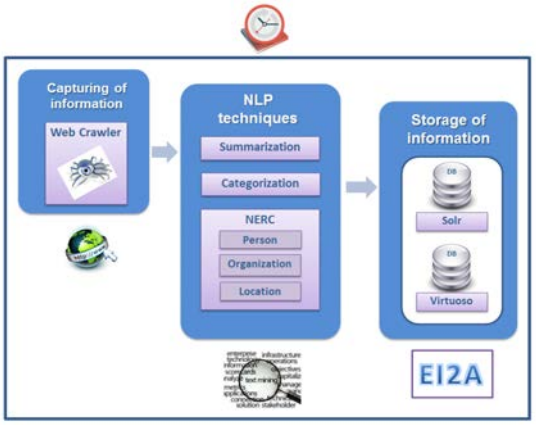

Figure 1: Functional Design

These functionalities have been implemented through *Moriarty*[3] (Peña et al., 2016), an ITAINNOVA's framework that allows the development of advanced software solutions for Big Data and Artificial Intelligence. In the following subsections a detailed explanation of each one of these three modules is provided.

[3] "Moriarty". Information available on: http://www.ita.es/moriarty/

2.1 Information Retrieval

A set of information sources as websites, subdomains or portals, are considered the seed of this approach, from which a list of URLs is created with the depth of the analysis, maximum number of pages to analyze and number of crawling-threads desired. Afterwards, *web crawling* techniques are applied in order to extract all the institutional textual information possible from these sources.

Once information has been extracted, a cleaning task is performed applying customized meta-data removing rules, which results in a text prepared for the later application of NLP techniques. Elimination of headers, footnotes or indexes are some of the functionalities applied in this phase.

Since web information could change frequently and new pages may appear, web crawling is executed periodically, analyzing the new webs that appear or reprocessing the webs that have changed.

2.2 NLP techniques

Main NLP techniques used over the textual data, as shown in Figure 1, involve text summarization, thesaurus-based semantic classification and the named entities recognition and classification (NERC). Before applying these high level NLP techniques, it is necessary to preprocess text, performing some common task such as lowercase transformations, lemmatization or stopwords filtering.

These NLP techniques provide valuable results of general interest to the user, contributing to the performance of the project objective. Summary task offers a synthesis of the textual information with the most relevant sentences by means of graph-based ranking algorithms (Erkan and Radev, 2004), the NERC task identifies implicit information of the texts about the people, organizations and places that are named in them, thanks to the use of neural networks algorithms (Chiu and Nichols, 2015).

The information extracted from the application of NLP techniques and the storage of their results has followed a legal methodology. In particular, due to legal aspects, people extracted from NERC technique, who do not belong to the organizational chart of Government of Aragon, is anonymized in the summary task by asterisks.

2.3 Storage and structuring of information: *EI2A ontology*

Extracted knowledge of web textual content is stored in a structured and controlled manner, both into Solr database and Virtuoso through triplets according to the EI2A scheme for further exploitation.

Based on the philosophy of the Semantic Web, well-known ontologies, schemes and vocabularies endorsed by European directives (INSPIRE)[4] and International Consortium (W3C)[5] such as *The Organization Ontology*, *Simple Knowledge Organization System*, *RDF Schema*, *XML Schema*, *Dublin Core Metadata Terms*, *Schema.org*, *Friend Of A Friend*, *ISA Program Person Core Vocabulary*, *ISA Program Location Core Vocabulary*, *WGS84 Geo Positioning*, *The Event Ontology* and *OWL-Time* have been reused to model EI2A ontology. In the process of construction of the EI2A ontology methodological guides (Noy and McGuinness, 2005; Fernández-López, Gómez-Pérez, and Juristo, 1997) have been followed.

The Organization Ontology provides concepts and relations to support the representation of organizational structure (notion of an organization, decomposition into sub-organizations and units, purpose and classification of organizations); reporting structure (membership and reporting structure within an organization; roles, posts, and relationships between people and organizations); location information (sites or buldings, locations within sites) and organizational history (merger, neaming). A Government of Aragon domain-especific extension has been added to model the nature of an organic unit or office in Aragon (level of administration, public or private character, classification of a public entity, etc.). EI2A ontology model has been enriched with aspects and metadata of the DIR[6] and ENI[7].

ISA Programme Person Core Vocabulary and *ISA Programme Location Core Vocabulary* are reused for describing a natural person and any place in terms of its name, address or geometry. Other institutional data of common interest identified and modeled are focused on concepts related to documents, geolocation, territory, event, temporality and

[4] http://www.idee.es/europeo-inspire
[5] https://www.w3.org/
[6] http://administracionelectronica.gob.es/ctt/dir3
[7] http://administracionelectronica.gob.es/ctt/eni

URIs.

In order to apply EI2A scheme on real data, the ontology has been populated. On one hand, EI2A has been populated with information related to the organizational chart of Government of Aragón extracted from spreadsheets. In this way, semantic information is added to indicate the nature of a person's membership of an organization, that is to say, that a person belongs to a unit or department with a specific role in a valid time interval. On the other hand, motivated by the need of automatic way to extract, structure and standardize information from the huge amount of textual content available on the institutional websites, EI2A ontology has also been populated with new instances of concepts and relations provided by the NERC process. For example, information related to a recognized entity (person, organization and/or location) has been cited on a web that is classified under a categorization of themes of the Government of Aragon is specified semantically, in addition to add data related to the summary, the url and the date of capture of textual web content. A browser through the ontological model EI2A is illustrated in Figure 2.

Figure 2: Browser for EI2A elements

3 Conclusions and Future Work

Despite dealing with texts with a great diversity of domains and formats, the work carried out manages to integrate a generic system capable of fulfilling the expectations presented at the beginning. In addition, the results obtained are significantly satisfactory.

In this context, the viability of the proposed project has been verified and new aspects have been detected in which it is necessary to continue exploring. To this end, the need to deploy the solution over the public Aragonese infrastructures is raised, in order to develop on top of this system natural language recognition services with the challenge to understand questions asked by a user and know what needs to be answered (for example, semantic search engine and assistant BOT), to investigate new services in the line of extracting knowledge from the unstructured information that the Government of Aragon has, and to continue expanding and evolving the EI2A schema with the definition of new concepts and relationships based on the information processed as a consequence of the indicated actions.

Acknowledgements

This work has been partly funded by the Department of Innovation, Research and University of the Government of Aragón in the context of the Aragón Open Data project. Special thanks to General Direction of Electronic Administration and the Information Society, Iciar Alonso and Julián Moyano for their collaboration. Also, this work has been partly financed by the FSE Operative Programme for Aragon (2014–2020).

References

Chiu, J. P. and E. Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

Erkan, G. and D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Fernández-López, M., A. Gómez-Pérez, and N. Juristo. 1997. Methontology: from ontological art towards ontological engineering.

Noy, N. F. and D. L. McGuinness. 2005. Desarrollo de ontologías-101: guía para crear tu primera ontología. Available on http://ocw.uc3m.es/ingenieria-informatica/sistemas-avanzados-de-recuperacion-de-informacion/ejercicios.

Peña, P., R. del Hoyo, J. Vea-Murguía, V. Rodrigálvarez, J. I. Calvo, and J. M. Martín. 2016. Moriarty: Improving 'time to market' in big data and artificial intelligence applications. *International Journal of Design & Nature and Ecodynamics*, 11:230–238.