Generating structured Profiles of Linked Data Graphs

Besnik Fetahu[1], Stefan Dietze[1], Bernardo Pereira Nunes[1,3], Davide Taibi[2] and Marco Antonio Casanova[3]

[1] L3S Research Center, Leibniz University Hanover, Germany
{fetahu, nunes, dietze}@L3S.de
[2] Italian National Research Council, Institute for Educational Technologies, Italy
davide.taibi@itd.cnr.it
[3] Department of Informatics - PUC-Rio - Rio de Janeiro, RJ - Brazil
{bnunes, casanova}@inf.puc-rio.br

Abstract. While there exists an increasingly large number of Linked Data, metadata about the content covered by individual datasets is sparse. In this paper, we introduce a processing pipeline to automatically assess, annotate and index available linked datasets. Given a minimal description of a dataset from the DataHub, the process produces a structured RDF-based description that includes information about its main topics. Additionally, the generated descriptions embed datasets into an interlinked graph of datasets based on shared topic vocabularies. We adopt and integrate techniques for Named Entity Recognition and automated data validation, providing a consistent workflow for dataset profiling and annotation. Finally, we validate the results obtained with our tool.

Keywords: Linked Data, Annotation, Datasets, Metadata

1 Introduction

The emergence of the Web of Data, in particularly Linked Data [1], has led to a vast amount of data being available on the Web. The DataHub[1], which serves as the central registry for open Web data, currently contains over 6000 datasets, 338 of which are (at the time of writing) part of the Linked Open Data group[2].

While datasets are highly heterogeneous with respect to represented resource types, currentness, quality or topic coverage, only brief and insufficient structured information about datasets are available. In the case of DataHub, only simple tags, few structured metadata about the size, endpoints or used schemas and a brief textual descriptions are available. This causes significant problems for data consumers (e.g. educational service providers or developers) to identify useful and trust-worthy data for different scenarios.

Nevertheless, earlier works address related issues [2, 3], such as schema alignment and extraction of shared resource annotations across datasets. However, they do not yet facilitate the extraction of reliable dataset metadata with respect

[1] `http://www.datahub.io`
[2] `http://datahub.io/group/lodcloud`

to represented topics. In order to address these limitations, we present an approach that automatically and incrementally indexes datasets by interlinking and annotating arbitrary datasets with relevant topics in the form of DBpedia entities and categories. By incrementally computing topic relevance scores for individual datasets, we gradually create a knowledge base of dataset meta-information. To improve scalability the process exploits representative sample sets of resources. Moreover, to ensure high annotation accuracy a semi-automated evaluation approach is proposed.

2 Semi-Automatic Dataset Annotation

Our dataset profiling platform automatically extracts top-ranked topic annotations (DBpedia categories) and captures these together with a relevance score for each dataset description. All dataset descriptions are captured using the VoID schema[3].

2.1 Entity Recognition

The analysis of sampled resources for a set of datasets consists of an annotation process using Named Entity Recognition (NER) and disambiguation tools (DBpedia Spotlight[4]). From each resource we extract the textual content assigned to the following properties: {`rdfs:label`, `rdfs:comment`, `teach:courseTitle`, `teach:courseDescription`, `skos:prefLabel`, `dcterms:description`, `dcterms:alternative`, `dcterms:title`, `bibo:abstract`, `bibo:body`, `cnrb:titolo`, `cnrd:descrizione`, `foaf:name`, `rdf:value`}; and perform contextual, that is resource-wise, NER. This establishes a common descriptive layer of top-ranked entities for each dataset extracted from DBpedia.

As the NER process can pose a bottleneck, we introduce an *incremental annotation* extraction process to alleviate this issue. This process avoids annotating resources similar to previously annotated ones by reusing already obtained annotations. Thus, for a predefined threshold similarity τ, from a pool of existing annotations $\mathcal{A}$, we assign an annotation to a resource if the similarity (resource-annotation) computed by the Jaccard's index is above threshold τ:

$$\forall a \in \mathcal{A} : \; J(r, a) = \frac{|r \cap a|}{|r \cup a|} \tag{1}$$

where $a \in \mathcal{A}$ represents already extracted annotations, while r is a resource instance which is analysed using the *incremental annotation* process.

2.2 Category Annotation

From the extracted annotations (DBpedia entities) $\mathcal{A}$, we analyse the set of assigned categories for each annotation. Such information is extracted from the DBpedia graph via the property `dcterms:subject` representing the topic covered by an entity. Furthermore, we leverage the hierarchical category organisation (as defined by SKOS schema: `skos:broader` and `skos:related`) assigned to entities within DBpedia.

[3] `http://www.w3.org/TR/void/`
[4] `http://spotlight.dbpedia.org`

However, such information extracted about categories is only useful when ranked according to their relevance for each dataset. Hence, we compute a normalised *relevance score* for each category assigned to a dataset by taking into (i) entities assigned to a category intra- and inter-datasets; and (ii) number of entities assigned to a dataset and over all datasets, see Equation 2:

$$score(t) = \frac{\Phi(t, D)}{\Phi(\cdot, D)} + \frac{\Phi(t, \cdot)}{\Phi(\cdot, \cdot)}, \quad \forall t \in \mathcal{T} \wedge D \in \mathcal{D} \tag{2}$$

where $\Phi(\cdot, \cdot)$ represents the number of entities associated with a topic t and for a dataset D, in case of void arguments, it outputs the number of entities in a dataset or over all datasets.

2.3 Automated Annotation Validation & Filtering Approach

Validation and filtering of extracted annotations is necessary, due to noise inherited from NER&NED results. The approach we propose for filtering out noisy annotations takes into account the contextual support given for an annotation from the resource instance it is extracted from. Therefore, we compute a *confidence score* which measures the similarity between an annotation and a resource using Jaccard's index similar to Equation 1, based on values extracted from properties `dbpedia-owl:abstract` and `rdfs:comment`, and the set of analysed properties listed in Section 2.1, respectively.

Whereas, in the validation phase we consider only entities that have a *confidence score* above some pre-define threshold and use human evaluators to assess the relevance of an extracted annotation with respect to the resource context.

3 Results and Evaluation

Our current implementation focuses on educationally relevant datasets as collected in a dedicated group on the DataHub[5] from which we selected a subset of 17 datasets based on their accessibility. Our topic annotation used representative, randomly selected samples of resources from each datasets, with approximately 100 instances for each resource type. Steps included NER, category extraction and threshold-based filtering using our *relevance & confidence scores*.

From the extracted categories based on the resulting annotations, we incorporated only the top-*50* categories being the most representative ones for a dataset based on the computed *normalised-score*. Results obtained from this processing are stored as part of a VoID[6]-based dataset catalog currently being provided as part of the LinkedUp project[7]; a catalogue providing access to such extensive information can be accessed under the following url[8].

The evaluation of annotation accuracy was measured based on two datasets: (a) annotation accuracy without any filtering (see Section 2.3); and (b) annotation accuracy after filtering, where only annotations with scores above some

[5] `http://datahub.io/groups/linkededucation`
[6] `http://www.w3.org/TR/void/`
[7] `http://www.linkedup-project.eu`
[8] `http://data.linkededucation.org`

threshold (in our case ≥ 0.15) are considered. The accuracy was measured for **1000** extracted annotations, picked randomly from $\mathcal{A}$. For (a) the accuracy was **71**%, whereas for (b) after filtering annotations below threshold $\tau \geq 0.15$. We observed an increase in accuracy of almost +**10**%.

Our demo application[9] focuses mainly on representation, profiling and search functionalities of the analysed datasets based on the structured descriptions. Figure 1 shows a screenshot of the exploratory search functionality of datasets using extracted annotations and categories. The user interface provides the following:

- Exploratory search of datasets based on extracted annotations & categories
- Interlinking of datasets based on most representative categories
- List of ranked categories for each dataset

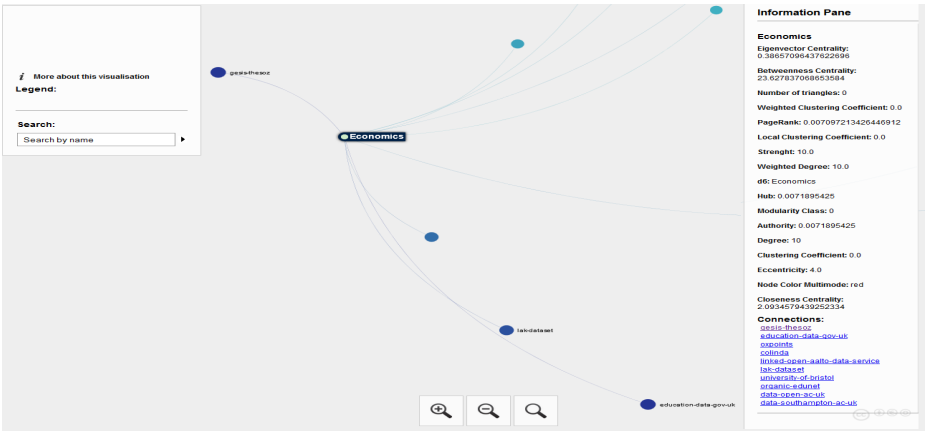

Fig. 1. Screenshot of the profiling of Linked Data demo, with an example category interlinking different datasets shown on the right hand side panel.

4 Future Work

Our current processing pipeline is able to extract topic annotations for arbitrary Linked Data with only minimal manual intervention. Having applied it to a small subset of available datasets, our future work aims at the automatic profiling of all available LOD datasets, towards providing a more descriptive catalog of Linked Datasets.

Acknowledgements. This work was partly funded by the LinkedUp (GA No:317620) and DURAARK (GA No:600908) projects under the FP7 programme of the European Commission.

References

1. C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.
2. M. d'Aquin, A. Adamou, and S. Dietze. Assessing the educational linked data landscape. In *WebSci*, pages 43–46. ACM, 2013.
3. D. Taibi, B. Fetahu, and S. Dietze. Towards integration of web data into a coherent educational data graph. In *WWW (Companion Volume)*, pages 419–424, 2013.

[9] `http://l3s.de/~fetahu/iswc_demo/`