DOI: 10.26314/GARR-Conf19-proceedings-01

Analysing Knowledge Domains that Emerge from Linked Open Data

Luigi Asprino[1,2], Paolo Ciancarini [2], Valentina Presutti[1,2]

[1] STLab, ISTC-CNR, Roma [2]Università di Bologna

Abstract. The aim of the Semantic Web initiative is to create a Web in which data is represented through symbols with a shared syntax and formal semantics. Several analyses have sought to investigate how these formal languages are used in practice, but very few of them analysed the Semantic Web per knowledge domain. In this paper we present a novel approach for analysing Semantic per knowledge domain in bottom-up fashion leveraging on topic modeling and natural language processing techniques.

Keywords. Semantic Web, Linked Open Data, Empirical Semantics

Introduction

The aim of the Semantic Web initiative is to create a Web in which data is represented through symbols with a shared syntax and semantics (Berners-Lee et al., 2001). This vision enacts machines to autonomously exchanging, analysing and using data found on the Web for their tasks, thus making the Web a huge knowledge base for intelligent agents. In the last 20 years the research, industry, and public administration communities have contributed to make real this vision by giving birth to the Linked Open Data: a huge network of ~200 billions[1] linked facts formally and uniformly represented in RDF and OWL.

[1]This estimation is obtained from metadata provided by www.lod-cloud.net

The collection of Linked Open Data (LOD) datasets forms the largest publicly accessible Knowledge Graph (KG) that is available on the Web today. Over the years, many studies have analysed these datasets, often focusing on the structure and dimension of the data, as well as providing statistics that shed light on the composition of the datasets. In most cases, such observations have been based on relatively small samples of the LOD KG. Moreover, it is not always clear how representative the chosen samples are. This is especially the case when observations are based on one dataset (e.g., DBpedia), or a handful of datasets, draw from the much larger LOD KG.

Linked Open Data and Semantic Web ontologies are encoded using RDF facts and/or OWL axioms, thereby exhibiting a formal semantics. Several analyses have sought to investigate how these formal languages are used in practice: how are certain formal constructs (e.g., owl:sameAs identity) used, and to what extent are LOD guidelines (e.g., limiting the use of blank nodes) followed in practice? Very few of them analysed LOD per knowledge domain, namely: how are certain formal constructs used in linguistic or government knowledge domain? To what extent are LOD guidelines followed in practice in encyclopedic or geographical knowledge domain? Most of proposed approaches for performing this kind of analyses rely on metadata provided by \url{lod-cloud.net} that

specifies provenance and knowledge domain of LOD datasets (e.g Schmachtenberg et al., 2014). However, most of LOD datasets are not provided with suitable metadata for such an analysis:

- The vocabulary used for specifying the knowledge domain of LOD datasets is poor (it consists of 9 labels: Cross-domain, Geography, Government, Life Sciences, Linguistics, Media, Publications, Social Networking, User-Generated).
- ~25% of LOD datasets do not indicate their knowledge domain, while the others indicate a single label for whole datasets.

Moreover, most of these labels are given in a top-down fashion with the risk that label and dataset may result partly uncorrelated.

This paper presents a novel approach for analysing LOD per knowledge domain that relies on topic modeling and natural language processing techniques. Topic modeling is a text-mining technique for discovering the topics that occur in a collection of documents. Intuitively, given a document about a particular topic (e.g. ``Religion''), one would expect words related to that topic to appear in the document more or less frequently than others (e.g. ``God'' will appear more often than ``Pizza'' in document about religion). The ``topics'' produced by topic modeling techniques are lists (one for each topic) of weighted words. Specifically, a topic t is a list of pairs (wi, pi) where wi is a word and pi is the likelihood that wi appears in a document about the topic t. A document typically concerns multiple topics in different proportions: for example in a document that is 10% about science and 90% about religion, there would probably be about 9 times more words about religion than those about science.

1. Proposed Approach

The collection of documents to analyse with topic modeling techniques is created as follows. For each LOD dataset d, we create a virtual document by concatenating natural language descriptions associated with entities within d. A LOD entity can be associated with two natural language descriptions:

- a label, a short text content used for naming the entity which is indicated by the property rdfs:label;
- a comment, a description of the resource in natural language, often providing examples of the concept being defined which is indicated the property rdfs:comment.

Then, a topic modeling library is used in order to extract the topics that emerge from virtual documents. The emerging topics will be inspected and manually aligned with a well-known taxonomy of knowledge domains (e.g WordNet's taxonomy). Finally, the extracted topic model will be used to associate virtual documents (hence, LOD datasets) with the extracted topics. As a result, LOD datasets will be annotated (in a bottom-up fashion) with topics aligned to a well-known taxonomy of knowledge domains, thus enabling an analysis of LOD datasets per knowledge domain.

2. Current status of the work

As input dataset for our analysis we used the crawl provided by LOD Laundromat (Beek

et al., 2014) (a project aimed at crawling data dumps published as part of the LOD cloud). For each dataset crawled by LOD Laundromat, we computed a virtual document. This process took 11.5 hours for analysing 640M triples on a m1.xxl instance (64 GB RAM, 16 virtual CPUs) provided by the GARR's cloud platform. The resulting dataset is provided in TSV format and available at the following link . In the next months the dataset will be analysed with the latent semantic analysis (LSA)\footnote{We are using the implementation of LSA provided by Gensim . LSA is a topic modelling technique that assumes that words with related meaning will occur in similar context. A word-document matrix containing word counts per document is constructed from the dataset (rows of the matrix represent unique words and columns represent documents). Then, a mathematical technique called singular value decomposition (SVD) is used to reduce the number of rows while preserving the similarity structure among columns.
The resulting rows with their associated scores for each document will constitute the extracted topics.

References

Tim Berners-Lee, James Hendler, and Ora Lassila. (2001), The Semantic Web, Scientific american (284.5), pp 34–43.

Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. (2014) Adoption of the Linked Data Best Practices in Different Topical Domains, Proceedings of the 13th International Semantic Web Conference (ISWC 2014), Part I, pp 245– 260.

Wouter Beek, Laurens Rietveld, Hamid R. Bazoobandi, Jan Wielemaker, and Stefan Schlobach. (2014), LOD Laundromat: A Uniform Way of Publishing Other People's Dirty Data, Proceedings of the 13th International Semantic Web Conference (ISWC 2014), Part I, pp 213–228.

https://wiki.dbpedia.org

http://lodlaundromat.org

https://w3id.org/edwin/garr2019

https://radimrehurek.com/gensim

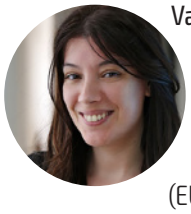
Authors

Luigi Asprino - luigi.asprino@istc.cnr.it

Luigi Asprino is a research assistant at the Institute of Cognitive Science and Technologies of National Research Council in Italy. He received a PhD in Computer Science and Engineering in 2019 from the University of Bologna. He has worked in national and european projects: MARIO (EU), MARE (EU) ArCo (IT), EcoDigit (IT). He has been involved in the organisation ESWC 2017 and he has served as program committee member and reviewer for many international conferences such as: ESWC, ISWC, WOP, EKAW, IJCAI and AAAI. His research interests include Knowledge Representation and Reasoning, Natural Language Processing and Machine Learning.

Paolo Ciancarini - paolo.ciancarini@unibo.it

Paolo Ciancarini is Professor of Computer Science at the Univ. of Bologna. He got a Phd in Informatics at the University of Pisa.
In Bologna he lectures on Software Engineering and Software Architecture, and is member of the Faculty of the PhD School in Computer Science. He currently is the President of the Italian Association of University Professors in Computer Science. He is the author of over 200 scientific papers and books. He is married, has two children, and is a passionate chess player and book collector.



Valentina Presutti - valentina.presutti@cnr.it

Valentina Presutti coordinates the Semantic Technology Laboratory of the National Research Council (CNR) in Rome. She received her Ph.D in Computer Science in 2006 at University of Bologna (Italy). She has coordinated, and worked as researcher in, many national and european projects on behalf of CNR – some examples: IKS (EU), MARIO (EU), NeOn (EU), ArCo (IT), EcoDigit (IT). She co-directs the International Semantic Web Research Summer School (ISWS). She serves in the editorial board of top journals such as Journal of Web Semantics, Journal of the Association for Information Science and Technology, Data Intelligence Journal, Intelligenza Artificiale. And she has been involved in the organisation of top semantic web conferences such as ISWC and ESWC and she is senior PC for IJCAI. She is one of the creators of the ontologydesignpatterns.org initiative and of the workshop series on Ontology Design and Patterns (WOP). She has 100+ publications in international journals/conferences/workshops on topics such as semantic web, knowledge extraction, and ontology design. She teaches Social Robotics and Programming as adjunct professor at the University of Bologna, and collaborates as scientific and technological expert for private as well as public organizations. Her research interests stand at the crossing between semantic web and artificial intelligence and include knowledge graphs, empirical analysis of the semantic web, social robotics.