On the Graph Structure of the Web of Data

Alberto Nogales Moyano, Alcalá University, Alcalá de Henares, Spain
Miguel Ángel Sicilia, Alcalá University, Alcalá de Henares, Spain
Elena García Barriocanal, Alcalá University, Alcalá de Henares, Spain

ABSTRACT

This article describes how the Web of Data has emerged as the realization of a machine readable web relying on the resource description framework language as a way to provide richer semantics to datasets. While the web of data is based on similar principles as the original web, being interlinked in the principal mechanism to relate information, the differences in the structure of the information is evident. Several studies have analysed the graph structure of the web, yielding important insights that were used in relevant applications. However, those findings cannot be transposed to the Web of Data, due to fundamental differences in the production, link creation and usage. This article reports on a study of the graph structure of the Web of Data using methods and techniques from similar studies for the Web. Results show that the Web of Data also complies with the theory of the bow-tie. Other characteristics are the low distance between nodes or the closeness and degree centrality are low. Regarding the datasets, the biggest one is Open Data Euskadi but the one with more connections to other datasets is Dbpedia.

KEYWORDS

Linked Open Data, Semantic Web, Web, Web of Data

1. INTRODUCTION

The “Linked Data approach” is the label for a set of techniques of publishing structured data on-line so that it can be interlinked and become more useful. Currently, Linked Data relies on standard Web technologies as the Hypertext Transfer Protocol\(^1\) (HTTP), to retrieve data from the World Wide Web\(^2\) (WWW) and the mechanisms of Uniform Resource Identifiers\(^3\) (URI) for identification. It combines this with a rich representation of datasets based on Resource Description Framework\(^4\) (RDF). The Web of Linked Data can thus be defined as a way to share open and structured data on the Web. Based on four principles exposed by Tim Berners-Lee in 2006\(^5\), the objective is that of making the data more readable to computers by publishing it in RDF and using URIs to identify each resource. Finally, data can be retrieved using the standardized language and protocol SPARQL\(^6\) and RDF Query Language. These approaches can be framed in the field of the Semantic Web, (Albert, Berners-Lee, & Fichetti, 1999) or (Berners-Lee, Hendler, & Lassila, 1999). A general overview of what Linked Data is can be found in (Bizer, Heath & Berners-Lee, 2009).

The adoption of Linked Data best practices, described in (Heath & Bizer, 2011), has resulted in the publication of datasets by data providers in several domains. Datasets use RDF as language and expose triples to describe the information. Each triple has the structure of subject, predicate and object. Subjects are URIs representing resources. Objects could be URIs or particular values. Predicates are also represented by URIs and are the way to relate a subject with its object. The entities

DOI: 10.4018/IJSWIS.2018040104

Copyright © 2018, IGi Global. Copying or distributing in print or electronic forms without written permission of IGi Global is prohibited.
represented by the URIs can be looked up by using the HTTP protocol. In the case that the subject and the object of a triple belong to different datasets a link between them is created, it is what is often called a RDF link. Taking into account the different datasets and its RDF links, the Web of Linked Data (LOD) can be seen as a graph. Here the datasets will represent the nodes and the RDF links the edges between them. This structure of graph allows applications to navigate between them and discover new information. By adding new datasets, the Web of Linked Data will evolve changing its structure. Since 2007, it has increased its size from a dozen of datasets to more than a thousand. A recent review can be found in (Schmachtenberg, Bizer, & Paulheim, 2014).

The techniques used to analyse the Web as a graph can also be applied to the Web of Linked Data. By studying the link structure of LOD, different measures used in Network Analysis can be applied, (Wasserman & Faust,1994): how to crawl information from it, importance of datasets by their sizes or understanding its behaviour looking at the evolution.

This paper is reporting on the main empirical findings on a comprehensive analysis of the graph structure of the Web of Linked Data. Following approaches used successfully in the past to analyse that structure for the Web of documents. This led to important insights helping to innovation in data retrieval strategies or an increased understanding of the social structure of Web communities. First of all, some metrics about the datasets in LOD and its usage are obtained. Then Social Network Analysis (SNA) techniques are applied, providing a general picture of the Web of Linked Data.

Results show that the structure of the Web is very compact which a low distance between nodes. The nodes also have a reasonable number of edges and most of them are close to the rest of the graph. Finally, it can be concluded that the structure of the Web fits with the theory of the bow-tie.

The rest of this paper is structured as follows. Section 2 provides a brief overview of existing studies on the graph structure of the Web and its potential usefulness to be applied to the Web of Linked Data. Section 3 outlines the main objectives of the study presented here. Then, materials and methods are described in Section 4 and results are discussed in Section 5. Finally, conclusions and outlook are presented in Section 6.

2. BACKGROUND

SNA has been successfully performed in order to know the structure and also to measure the classical Web of Documents, also the Web of Data. The following papers have as aim give measures or the structure of a set of any of these networks.

Talking about the Web in general, the first important study about the structure of the Web was reported by (Broder et al., 2000). They analysed 200 million of pages and 1.5 billion links. It concluded that the World Wide Web can be grouped in four big sets with the shape of a bow-tie. Based on this, (Metaxas, 2012), gives a theory about why the Web of Data has the shape of a Bowtie, presenting an algorithm that comprises the regions of a Bowtie in a directed graph. The evolution of the Web is studied in (Hall &Tiropanis, 2012), in the paper is found the definition of Web Science. It has also been reviewed, (Meusel, Vigna, Lehmborg, & Bizer, 2014), confirming that the structure has not changed a lot and it maintains the shape of a Bowtie. The same authors, in (Meusel, Vigna, Lehmborg, & Bizer, 2015), has analysed the same structure in three different levels of aggregation: page, host and pay-level domain (PLD). Measures like average degree, connectivity, connected components or in-degrees and out-degrees have been studied. There are also several papers where SNA metrics have been applied in the World Wide Web. In (Di Fatta, Caputo, Evangelista, & Dominici, 2016), the theories of small world are used in it. A metric like betweenness centrality based on Random Shortest Paths (RSP) framework is applied to evaluate the visit rate of pages in the Web of Documents, (Kivimäki, Lebichot, Saramäki, & Saerens, 2016). The two papers, (Broder et al., 2000) and (Meusel, Vigna, Lehmborg, & Bizer, 2014), will be used as the main point of departure for this research. In the paper, will be applied some of the techniques exposed in these papers but the structure that will be analysed is the Web of Linked Data.
Nested Optional Join for Efficient Evaluation of SPARQL Nested Optional Graph Patterns
[www.igi-global.com/chapter/nested-optional-join-efficient-evaluation/41658?camid=4v1a](www.igi-global.com/chapter/nested-optional-join-efficient-evaluation/41658?camid=4v1a)

Harnessing Semantic Features for Large-Scale Content-Based Hashtag Recommendations on Microblogging Platforms
[www.igi-global.com/article/harnessing-semantic-features-for-large-scale-content-based-hashtag-recommendations-on-microblogging-platforms/172423?camid=4v1a](www.igi-global.com/article/harnessing-semantic-features-for-large-scale-content-based-hashtag-recommendations-on-microblogging-platforms/172423?camid=4v1a)