Information Sciences 421 (2017) 15–29

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A Structural Analysis of topic ontologies

Eduardo Xamena[a,b,c,*], Nélida Beatriz Brignole[b,d,e], Ana Gabriela Maguitman[c,d]

[a] Facultad de Ciencias Exactas - UNSa - Universidad Nacional de Salta - Av. Bolivia 5150, Salta, Argentina
[b] LIDeCC - Laboratorio de Investigación y Desarrollo en Computación Científica, Argentina
[c] Knowledge Management and Information Retrieval Research Group, ICIC CONICET-UNS, Argentina
[d] DCIC-UNS - Departamento de Ciencias e Ingeniería de la Computación - Universidad Nacional del Sur - San Andrés 800, Bahía Blanca, Argentina
[e] Planta Piloto de Ingeniería Química, UNS-CONICET - Cno la Carrindanga km 7, Bahía Blanca, Argentina

ARTICLE INFO

Article history:
Received 17 December 2016
Revised 11 August 2017
Accepted 26 August 2017
Available online 31 August 2017

Keywords:
Topic ontologies
Topological analysis
Complex networks

ABSTRACT

DMOZ is the largest human-edited topic ontology available on the Web. This article studies the structural properties of the DMOZ graph. A number of global and local properties of this graph and the subgraphs resulting from isolating edges of different types are examined by means of metrics commonly used in complex network analysis. In particular, we investigate the presence of various features that characterize small-world networks. This analysis is complemented by examining other characteristics of the graphs such as connectivity and centrality measures. The connectivity and centrality patterns are further studied by means of visualizations of the graphs' k-core decomposition and a selection of strongly connected components. Several non-trivial regularities that are also encountered in other artificial and natural complex networks provide a general picture of this large human-edited topic ontology. This analysis is of major pragmatic interest as it allows a better understanding of notions such as navigability among topics, hierarchical structure and topic cohesiveness, which are of great importance in the design of topic ontologies.

1. Introduction

Ontologies are structures commonly used to capture knowledge about certain areas by providing relevant topics or concepts and relations between them. A general topic ontology such as DMOZ (Directory Mozilla), historically known as ODP (Open Directory Project), is a complex structure that reflects the collective knowledge of the ontology editors about a broad range of topics. Revealing information about structural aspects of this ontology can provide useful insights on the nature of topic connectivity, topic importance, topic relevance and topic similarity, among other useful concepts, conferring a unique opportunity to address the important cognitive problem of understanding "the landscape of topics" as realized by a large number of human editors.

The DMOZ is a collaborative classification of websites. The topical structure made up from this classification can be represented as a big graph or ontology. In the DMOZ topic ontology, topics are represented as nodes in a tree-structured hierarchy, with "is-a" connections determined by topic-inclusion relations. In addition, DMOZ admits cross-links representing "symbolic" connections to allow for topics with multiple parents. Finally, another type of relation represented by "related" links allows to connect related topics that are not involved in a class-inclusion relation. While the tree-structured hierarchy

[*] Corresponding author.
E-mail address: examena@plapiqui.edu.ar (E. Xamena).

http://dx.doi.org/10.1016/j.ins.2017.08.081

imposes strong constraints on the general organization of the DMOZ ontology, the “symbolic” and “related” connections loose up these constraints and offer the possibility of integrating the taxonomical component of DMOZ with more general components, resulting in less restricted connectivity patterns when analyzed as a whole.

Network analysis constitutes a powerful tool for inferring several properties on datasets arising from a wide range of areas. From the topological structure of the web [9] to the analysis of the economy of a country [16], network properties reveal many important features of the represented models. These properties have important implications on the robustness, navigability, and cohesiveness of the networks. Large volumes of linked data can be analyzed from a Complex Network perspective, with application to information retrieval. A structural study of a big corpus, made up of interconnected documents or other kinds of information entities, could help on finding useful information about the semantic relations existing among these entities. Graph representations have proved to be an effective and efficient way for structural semantic similarity calculations [28]. The structure of semantic networks constructed from word associations has been widely studied in cognitive science [7,31,39], with application in several areas such as the assistance of people with the anomic aphasia disease [34].

The study presented here focuses on analyzing the network topology of the DMOZ graph in its pure form. It also analyses the network topology of the subgraphs of DMOZ corresponding to edges of the three different types involved in this ontology, namely taxonomical, symbolic and related edges. The analysis is carried out by computing various complex network metrics, such as node degree, local clustering coefficient, average shortest path length, and diameter of the network, allowing to draw interesting conclusions about non-trivial regularities present in the analyzed graphs. To the best of the authors’ knowledge this article provides the first large-scale analysis of a topic ontology graph from a complex network perspective.

2. Background

In this section some graph-theoretic concepts are briefly described, in particular those that relate to the analysis carried out on the DMOZ structure. Then, we describe various measures and tools that have been adopted to complete the analysis reported in this article.

2.1. DMOZ as a graph

The DMOZ project is a large directory of websites organized by topics. The main component of this directory is a hierarchical structure, the DMOZ taxonomy. Websites are added to the directory by assigning them to existing topics from the taxonomy. Besides its hierarchical structure, DMOZ contains other kinds of links between topics, as is the case of “symbolic” and “related” links. “Symbolic” links correspond to alternative classifications that escape from the taxonomy, and have to be included in the directory. “Related” links are used to connect topics that are associated according to some criterion whenever such relation is not expressed as a taxonomic or symbolic relation. More formally, the structure of DMOZ can be represented as a directed graph $G = (N; E)$ with a set of nodes N and a set of edges E. Each node in N represents a topic containing documents, and every edge of E connects two nodes of N. The set of edges E is made up of three classes of links between topics:

- class T, corresponding to the hierarchical component of the ontology,
- class S, reflecting the non-hierarchical “symbolic” cross links, and
- class R, representing the “related” cross links, also organized in a non-hierarchical fashion.

These three types of links give rise to the *T-subgraph, S-subgraph* and *R-subgraph*, respectively. Each of these subgraphs will be analyzed as independent networks as well as jointly.

Fig. 1 illustrates a portion of the structure of the DMOZ ontology graph, showing the three types of links.

2.2. Structural analysis of graphs

This section reviews some concepts, measures and algorithms that we have adopted to analyze the most salient properties of the DMOZ ontology graph.

2.2.1. Connectivity and centrality measures

Several connectivity and centrality measures commonly used for complex network analysis can be applied to the DMOZ graph, offering a means to assess topic importance and relevance among topics. Next, we describe the connectivity and centrality measures used in this work.

- **Graph density**: The density of a graph is the proportion of edges actually present in a graph with respect to the number of possible links that could be established between the nodes of the graph. This measure is computed on a graph G = (N,E) as follows [14]:

$$Density(G) = \frac{|E|}{|N|(|N|-1)}.$$

- **Diameter**: The diameter of a graph is characterized by the largest distance between any two nodes, where the distance between a pair of nodes is the length of the shortest path between them [22].