Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 37 (2014) 56 – 63

Procedia Computer Science

The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2014)

A semantic-based variables selection for ontology learning Taking Jaccard alignment as case

Djellali Choukri[a,b]

[a]*LANCI UQAM, Local W-5353 Case postale 8888, Succ. Centre-Ville Montréal (Québec) H2X 3Y7, Canada*
[b]*LATECE UQAM, 201, PK 4470, Président Kennedy Montréal (Québec) H2X 3Y7, Canada*

Abstract

In the past decade, research on numerical schemes on ontology learning has been quite intensive. Several learning approaches have been proposed to help developers during the maintenance process. Most of the proposed approaches do not process the curse of dimensionality and the semantic contained in the information structure. A novel semantic-based method for ontology learning, which can provide improvement in both alignment and learning, is described. Good comparisons with the experimental studies demonstrate the multidisciplinary applications of our approach.

Peer-review under responsibility of the Program Chairs of EUSPN-2014 and ICTH 2014.

Keywords: Machine Learning, variables selection, clustering, Semantic Web, Ontology, alignment.

1. Introduction

Shared understanding and better enabling computers and people to work in cooperation are the goals of the Semantic Web. The following definition was adopted by the Semantic Web community « *The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation* » Tim Berners-Lee, James Hendler, Ora Lassila. The Semantic Web, May 2001.

The Semantic Web, also known as Web 3.0, is described as a multilayer architecture, where the ontology is in the middle of this architecture. Ontology is proving to be the best solution for knowledge sharing. However, ontology is a dynamic structure due to dynamic conditions resulting from changes in the conceptual domain, conceptualization and specification.

djellali.choukri@courrier.uqam.ca

Peer-review under responsibility of the Program Chairs of EUSPN-2014 and ICTH 2014.
doi:10.1016/j.procs.2014.08.012

Several learning approaches focused on the extraction of ontologies by applying the techniques from Natural Language Processing, Information Retrieval and Machine Learning. Most of them are sensitive to noise, presentation order and Bellman's curse of dimensionality.
In this study, we propose a new conceptual model for ontology learning based on variables selection, clustering and an alignment process.

The paper is organized as follows: In Section 2, we present the current state of the art in ontology learning, our research questions and the problematic of ontology learning. The conceptual architecture of our approach is given in Section 3. Before we conclude, we give in Section 4 a short evaluation with benchmarking models for our conceptual model. Then, a conclusion (Section 5) and future work (Section 6) end the paper.

2. State of the art, Problem and Research Questions

« *Ontology learning can be defined as the set of methods and techniques used for building ontology from scratch, enriching, or adapting an existing ontology in a semiautomatic fashion using several sources*» [1].

Figure (1) illustrates a classification of several learning approaches that use structured, semi-structured and unstructured text [3].

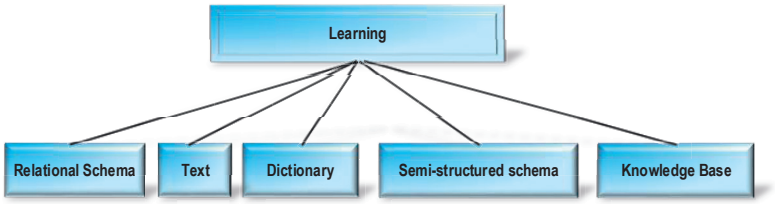

Fig. 1. Classification of different approaches to ontology learning

The best known approaches to ontology learning from unstructured text focus on the extraction of relevant pattern by applying techniques from Natural Language Processing, Information Retrieval and Machine Learning. These approaches include: ontology pruning [14], conceptual grouping [15], formal concept analysis (FCA) [16], association rules [17], pattern extraction [18] and conceptual learning [2].

Most of the previous approaches provide limited support for all activities of the maintenance process. They neglect the overall structure contained in the ontology and provide only limited support to generate ontology. The precision and recall do not meet the demands of actual applications due to semantic heterogeneity. Axiomatic learning is unexplored and they are therefore not suitable for problems with heterogeneity. In addition, the evaluation of ontology remains a significant problem and the choice of an appropriate method depends on the used criteria and model domain.

3. The architecture of our learning system

The process of learning starts with capturing a set of terms from the available documents as shown in Figure (2). Before creating the indexation model, it is necessary to remove all occurrences of noise. The removal of punctuation, negative dictionary and stemming techniques are used to remove noise. In order to index the textual document, we used the Vector Space Model (also known as VSM) [12]. Each document is indexed by its terms in a vector and each term is weighted by means of the TF-IDF function (Term Frequency Inverse Document Frequency) [10]. The representation model generates a very high dimensionality even after pre-treatment and cleaning.
In order to acquire a semantic indexation and to reduce the space of representation (Bellman's curse of dimensionality), we used a Wrapper Model based on the Truncated Singular Value Decomposition (also known as TSVD) [4].