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].

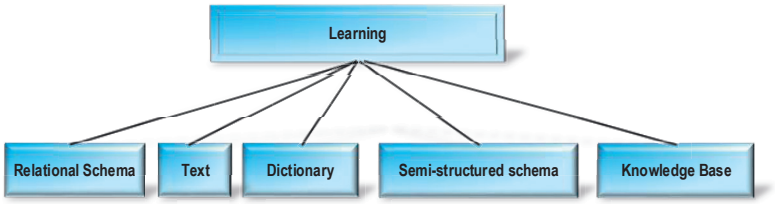

Fig. 1. Classification of different approaches to ontology learning

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].

In order to automatically identify the number of clusters, we used a clustering model based on the Fuzzy Adaptive Resonance Theory [9], [13]. This dynamic model allows the neural network to automatically adjust its size depending on the dynamics of the environment (dynamic knowledge, complex shapes, variables distributions, incremental acquisition, etc) (stability-plasticity dilemma). All clusters are described by keywords (labels) representing their content. An alignment process based on Jaccard distance [6] is used to identify the correspondence between the ontological artifacts and descriptive labels. It creates alignment rules that define how to transform the entities by defining all types of possible associations between ontological artefacts and labels. The update process is used to describe an explicit and formal conceptualization for the domain model.
The CRISP-DM-OWL[1] ontology used in this project is integrated into a hybrid system DM [19], describing the artifacts and the basic rules.

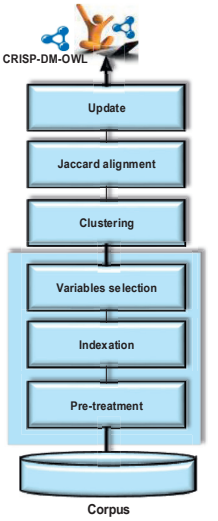

Fig. 2. The process of learning

4. Experimentation

4.1. Pretreatment

The training corpus consists of a set of IEEE abstracts divided in several categories. The average length of the document in terms of words is 182.53 in the training set and 178.14 in the test set. The number of documents in each category is highly unbalanced. Thirty percent of the data are selected to test the model (no theoretical justification for this percentage). Table (1) shows in detail the statistical distribution of words in the data sets ($\overline{L}$ is the average length of the document and σ_L the standard deviation of document length).

Table 1. The Statistical Distribution of Words.

Data Sets	$\overline{L}$	σ_L
Learning	182.53	60.65
Testing	178.14	60.55

We used the Glasgow list [8] as a stop words list in our experiments. This list is widely used as English standard stop word; it covers a large number (351 stop words). In order to seek the lexical root of a term, we used Porter algorithm [11].

4.2. Variable selection

The TSVD process calculates an approximation of lower rank that takes advantage of the correlation of terms as shown in Figure (3).

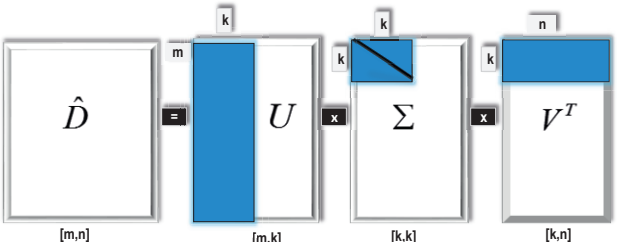

Fig. 3. TSVD process.

[1] http://www.elmanahel.ca/ontology/crisp-dm-owl.owl

$$\hat{D} = U_k \Sigma_k V_k^T \; ; \; \Sigma_k = diag\ [\sigma_1, \sigma_2, \ldots\ldots\ldots\ \sigma_k] \tag{1}$$

Where :

$U_k U_k^T = I_m \wedge V_k^T V_k = I_n$; $I_m\,(I_n)$: identity matrix of size m(n) (respectively); $\hat{D}_{ij} = tfidf_{ij}$.

We used the energy ratio (ER) algorithm [5] as a criterion to find the number of the relevant variables. We can express the variance accounted for by the i[th] singular vector as defined by formula (2):

$$\text{var}_i = \frac{\sigma_i^2}{\sum_{j=1}^{j=r} \sigma_j^2}, \; r = rank\ (\hat{D}) = \left|\{\sigma_i\}\right|, \; \sigma_i \succ 0 \tag{2}$$

Hence, the TSVD wrapper framework to select the most relevant variable subset $S_1 = \Phi$ given the set of singular values $S = \{diag\ (\Sigma_k)\} = \{\sigma_1, \sigma_2, \ldots \sigma_r\}$ is defined by the following formula:

$$\Psi^{k+1} = ArgMax_{s_1 \subset S} \left(\frac{\sum_{i=1}^{i=k} \sigma_i^2}{\sum_{j=1}^{j=r} \sigma_j^2} \right) \tag{3}$$

As shown in figure (4) (a), more than 91.13% of the variance in the Data Set was explained by the first 721 singular values and the cumulative variance of the last 379 singular values does not exceed the contribution of the first singular values. In addition, the first 721 singular values are much greater than the last singular values ($\sigma_1 \geq \sigma_2 \geq, \ldots \sigma_r \succ 0$).

The figure (4) (b) shows the dispersion of the variance with polar coordinates. This representation is particularly useful because the relationship between the singular value and variance is easily expressed in terms of angle and distance. In this two-dimensional system, each point is determined by the polar coordinates, i.e., the radial and angular coordinates. In the light of the results it is undoubtedly to see that the variables related to the small singular values are almost irrelevant and do not affect the measures of similarity between documents, i.e., they explain a small amount of the variance. As a result, we generated reduced projection space by keeping only the first 721 singular values.

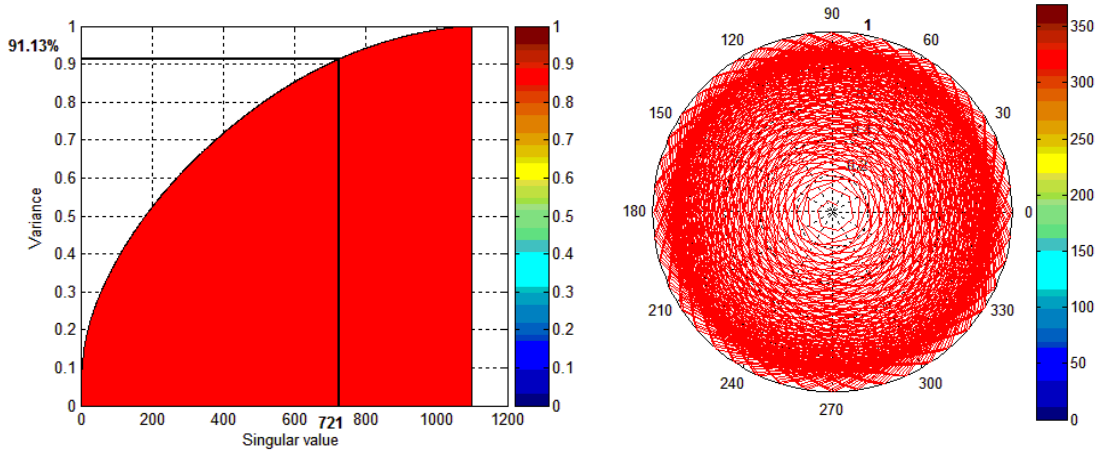

Fig. 4. (a) Variance vs. Singular Value; (b) Polar coordinates of the variance.

4.3. Clustering

We apply mean square error algorithm as a stopping criterion for the learning step. This criterion is defined as follows :

$$E_p = \sum_{j=1}^{n_L} (e_{1j}^{|L|})^2 = \sum_{j=1}^{n_L} (d_j^{|L|} - y_j^{|L|}) \quad (4)$$

$d_j^{|L|}, y_j^{|L|}$: are respectively the actual output and the desired output.

The convergence speed of our clustering model is based on typical initializations [13]. This initialization scheme reduces the computation time and improves the convergence speed to achieve the neighbourhood vicinity of the response.

Figure (5) illustrate the learning phase of the neural network. The learning is performed independently in each iteration (epoch). All inputs are presented in neural network and each output is calculated individually. The objective is to adjust the weight matrix to find the best clustering. The horizontal axis represents the number of iterations and the vertical axis shows the squared error after the patterns submission.

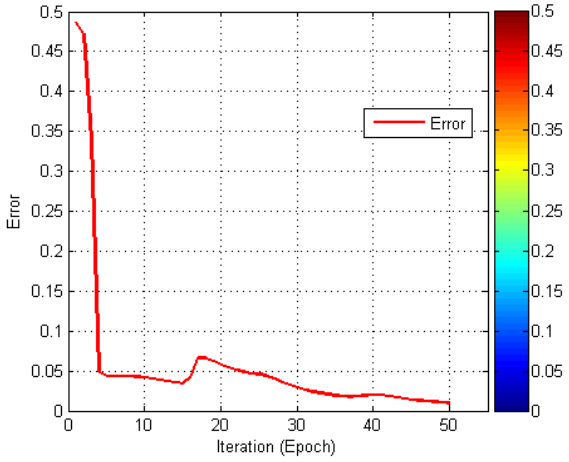

Fig. 5. Learning phase.

When the connectionist network learns by changing the synaptic weights, the recognition of the output vector generated by the output layer should correspond to the desired output. The recognition is used to discover hidden patterns and to discern the content in clusters that have a high density in the indexing space. Clusters are considered attractors automatically generated to feed the update process.

4.4. Alignment

In order to identify the correspondence between the ontological artifacts and descriptive labels, we used Jaccard index (also known as the Jaccard similarity coefficient) defined as the size of the intersection divided by the size of the union [6].

$$Cof_{sim_J}(C, Label_k) = J_\delta(Label_k, C) = \frac{|Label_k \cap C|}{|Label_k \cup C|} = \frac{|\{c^i\} \cap \{Label_k^i\}|}{|\{c^i\} \cup \{Label_k^i\}|} \quad (5)$$

$$C \in \Gamma^m, Label_k \in \Gamma_{OWL}^n$$

Γ : the set of alphabets used to build chains of descriptive labels.

Γ_{OWL} : the set of alphabets representing the artefacts of the CRISP-DM-OWL ontology.

Figure (6) shows the use of Jaccard index to compute the alignment rules. The horizontal axis represents the Jaccard alignment and the vertical axis indicates the calculated similarity values between labels generated by the clustering algorithm and the set of artifacts in the CRISP-DM-OWL ontology.

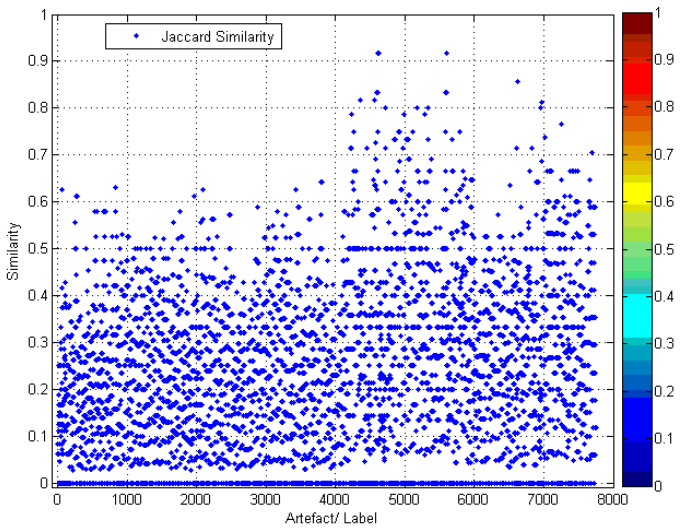

Fig. 6. The Jaccard similarity between concepts and descriptive labels.

The Jaccard alignment is rigidly based on points of overlap between two strings. If there are occurrences of noise that does not belong to the strings, the total number of mutation also increases. Hence, the choice of an appropriate distance measure to meet the application needs is crucial task and an attention should be paid to the selection of an appropriate measure for an alignment process.

4.5. Update

The update process establishes a model of representation in a formal language by involving the following steps:

- Choose a representation language for encoding ontology: OWL-DL is computable code used to represent the updated ontology in a structured and formal model. This language facilitates the interpretation of the content by providing additional vocabulary and a formal semantics.
- Describe the specifications of the ontology according to the chosen editor: maintenance of ontology is performed using the plug-in OWL[2]. This plug-in is used to modify the ontology, access to reasoning tools based on description logic and to acquire taxonomic structure.

4.6. Evaluation

Consistency, subsumption and instantiation can accentuate the main features of the conceptual hierarchy as well as the ontological inference. Description logic is perfectly suited to this situation. It has a formal semantics based on logic and equipped by decidable decisions. As shown in figure (7), the descriptive inference system used to verify the consistency, soundness and completeness of ontology is based on the inference engine RacerPro[3] (**R**enamed **A**Box and **C**oncept **E**xpression **R**easoner) [7].

[2]http://protege.stanford.edu/
[3]http://racer.sts.tuhh.de/

The DIG protocol (XML standard) [20] is used to connect the Data Mining applications. The allocation of the terminological knowledge base allows users to query the conceptual model.

It should be noted that the enriched ontology $O = (C, H_C, R_C, H_R, I, R_I, A)$ is considered in two parts:

- The extensional part C, R_C ou $\mathcal{T}$-Box : representation and manipulation of concepts and roles in terminological level.
- The intensional part I, R_I ou $\mathcal{A}$-Box : representation and manipulation of individuals in a factual level [21].

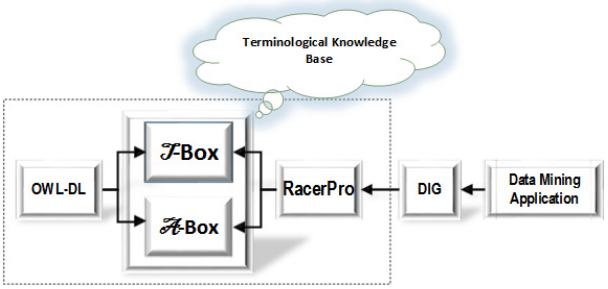

Fig. 7. The Descriptive Inference System.

The descriptive inference system is used to verify the ontology consistency, terminological knowledge base unsatisfiability, avoid definitions of terminology that contain cycles and fixed points, etc.

Our evaluation approach is independent of the conceptualization of the domain model and considers the main features of the ontology structure and its population (concepts, instances, axiom, relationship, etc.).

5. Conclusion

In this paper, we have introduced a new architecture providing semantic variables selection for ontology learning. The Wrapper Model based on truncated singular value decomposition uses a small space to represent the semantic relations between terms. Hence, it reduces the noise in the indexation and improves the accuracy of clustering. The clustering model chosen in our system does not depend on the order of on-line presentation (plasticity-elasticity). Thus, it eliminates the laborious process of knowledge engineering involved in the process of knowledge extraction. In order to identify the correspondence between the ontological artifacts and labels, we used an alignment process based on Jaccard distance. The evaluation step provides a symbolic reasoning used to verify the ontology consistency and unexpected relationships between the ontological artefacts.

6. Future work

The purpose of our next work is to aggregate several individual alignments. In this context, the aggregation process combines the decisions to obtain a single alignment rule.

References

1. Hazman, Maryam El-Beltagy, Samhaa R Rafea, Ahmedmhaa R. Survey of Ontology Learning Approaches. International Journal of Computer Applications 2011; 22:0975 – 8887.
2. Gomez-Perez A, Manzano-Macho D. A survey of ontology learning methods and techniques. OntoWeb Deliverable D 2003; 1-5.
3. Maedche A, Staab S. Ontology learning for the semantic web. Intelligent Systems, IEEE 2003 16; 72-79.
4. Djellali C. Truncated singular value decomposition for semantic-based data retrieval. Third International Conference on Communications and Information Technology (ICCIT), 2013; 61-66.
5. Jyh-Jong Wei, Chuang-Jan Chang,Nai-Kuan Chou, Gwo-Jen Jan. ECG data compression using truncated singular value decomposition. IEEE Transactions on Information Technology in Biomedicine 2001 5; 290-299.
6. Shibata N, Kajikawa Y. How to measure the semantic similarities between scientific papers and patents in order to discover research fronts: A case study of solar cells. Technology Management for Global Economic Growth (PICMET), 2010;1-6.
7. Haarslev V, Hidde K, Mölr R, Wessel M. The RacerPro knowledge representation and reasoning system. Semantic Web 2011; 1-7.
8. Zaman A , Matsakis P and Brown C. Evaluation of stop word lists in text retrieval using Latent Semantic Indexing. In Digital Information Management (ICDIM), 2011 Sixth International Conference 2011; 133-136.
9. Isawa H, Matsushita H and Nishio Y. Fuzzy Adaptive Resonance Theory Combining Overlapped Category in consideration of connections. In Neural Networks, 2008. IJCNN 2008. IEEE International Joint Conference 2008; 3595,3600.
10. Baena-Garcia M, Carmona-Cejudo J M, Astillo G C,and Morales-Bueno R. TF-SIDF: Term frequency, sketched inverse document frequency. In Intelligent Systems Design and Applications (ISDA) 2011 pages 1044-1049.
11. Issac B, Jap W J. Implementing spam detection using Bayesian and Porter Stemmer keyword stripping approaches. In TENCON Conference 2009; 1-5.
12. Salton G, Wong A, and Yang C S. A vector space model for automatic indexing. Communications of the ACM,1975;613-620.
13. Djellali C. Enhancing text clustering model based on truncated singular value decomposition, fuzzy art and cross validation. the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2013; 1078-1083
14. Lv Yanhui. An approach to ontologies integration. In Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference 2011; 1262-1266.
15. Starr R, Oliveira P. Conceptual Maps as the First Step in an Ontology Construction Method. In Enterprise Distributed Object Computing Conference Workshops (EDOCW) IEEE, 2010; 199-206.
16. Ning Liu, Guanyu Li, and Li Sun. Using Formal Concept Analysis for Maritime Ontology Building. In Information Technology and Applications (IFITA), 2010 159-162.
17. Tseng Ming-Cheng, Lin Wen-Yang, and Jeng Rong. Incrememtal Maintenance of Ontology-Exploiting Association Rules. In Machine Learning and Cybernetics, 2007; 2280-2285.
18. ellandi, S Nasoni, A Tommasi, and C Zavattari. Ontology-Driven Relation Extraction by Pattern Discovery. In Information, Process, and Knowledge Management, 2010; 1-6.
19. Shen Yanfen. A formal ontology for Data Mining : principles, design and evolution Thesis UQTR, 2007.
20. S Bechhofer, R Moller, and P Crowther. The DIG description logic interface: DIG/1.1, 2003.
21. Napoli, Amedeo. An introduction to description logics, 1997.