Ontology Summarization through Simple Pruning Measures

Isaac Lera, Carlos Juiz and Ramon Puigjaner
Dept. Matemàtiques i Informàtica, Universitat de les Illes Balears, 07122, Palma de Mallorca, Spain

Keywords: Ontology Summarization, Context.

Abstract: This paper addresses the problem of synthesizing an ontology by defining prunning measures based on OWL axioms. From a deep structural and axiomatic analysis of current ontologies, we have defined a set of basic measures of selection of important elements that it has a linear computational cost.

1 INTRODUCTION

A semantic model as DBpedia consists of 1 billion triples and GeoNames, of 146 million triples. Managing this volume of linked statements, using inference reasoners, SPARQL endpoints, or simple data tasks such as removing or adding, implies accesses and updates of thousands RDF triples. Resource management is critical to achieve suitable response times and an effective data consumption. Data synthesis, independently of the representation model, allows to generate solutions to multiples areas such as the maximizing of data exchange, on social understanding, the simplify of domain/context management, on caching techniques, on information retrieving, on ontology mapping techniques, and so on where it may be useful the manipulation of fewer elements.

In our opinion, the motivation of a summarization depends on the final application of the results. Some works focus on application use (Alani et al., 2006) which it is based on the elements involved with user queries; other approaches simplify a large number of hierarchies (Stuckenschmidt and Schlicht, 2009) or some show the most key concepts (Peroni et al., 2008). The evaluation is task motivation and is often compared with a gold standard (human judgements) (Li and Motta, 2010).

Our approach is designed to achieve an effective response time in performance terms. Our pruning measures are chosen in function of centrality, coverage, and richness of OWL axioms. For evaluation results, we have used the classical group of ontologies (biosphere, financial, music, aktors portal) (Peroni et al., 2008; Li and Motta, 2010), and also we have included the *conference* group of ontologies from Ontology Alignment Evaluation Initiative (OAEI).

2 RELATED WORK AND ANALYSIS

Zhang et *al.* (Zhang et al., 2007) defined an algorithm where an ontology is transformed in a RDF sentence graph and it sets the final size given a summarize length. The selection is based on the notion of centrality on social networks. The value of centrality is determined by some formulas. The first one counts the degree of centrality by means of the number of connections: incoming and outgoing links. The second one determines its relative centrality in terms of the shortest path with other nodes. The third measure consist on two alternatives to eigenvector centrality: PageRank and HITS. Finally, a re-ranking process sets the salient elements using domain filters of the previous measures.

Peroni et *al.* (Peroni et al., 2008) presented two versions of an algorithm to make easy the ontology domain understanding through the classes more representatives. It is based on the idea of the Eleanor Rosch where people use often basic terms to describe things instead of abstract ones. Authors defined a set of measures such as: *name simplicity* avoiding compound nouns; *basic level* indicates the centrality of a label in the taxonomy; the *density* is computed by a weighted aggregation on the number of direct subclasses, properties and instances; and, the *coverage* represents the subsumption of a class with the others in the taxonomy.

Stuckenschmidt et *al.* (Stuckenschmidt and Schlicht, 2009) proposed a division of modules which contains coherent hierarchical data and it has the maximum level of interpretation. By means of a graph, they determined the weight of dependencies among classes using social network theories.

Lera I., Juiz C. and Puigjaner R..
Ontology Summarization through Simple Pruning Measures.
DOI: 10.5220/0004127303390342
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 339-342
ISBN: 978-989-8565-30-3

And finally, regarding with the evaluation, Li et *al.* (Li and Motta, 2010) analysed evaluation methods in semantic representations. They used other areas and methods to classify three types of evaluations: application-driven ontology, gold standard based ontology, and corpus coverage ontology. Moreover, they provided a list of evaluation measures: recall-based, sentence-rank-based, and content-based. They applied these observations in two previous approaches (Peroni et al., 2008; Zhang et al., 2007) using financial and biosphere ontologies.

In order to introduce our approach, we have considered noteworthy the relativity of this subject: the summarization depends entirely on the person who do it or on the final nature that it has the application. Thus, we decided to carry on a brief questionnaire where the volunteers did not know nothing about ontologies and the aim of the same, but most of them were university students and the rest was research staff.

We realize some expected results. The most simple is a structure, the most easier is the selection of elements. A taxonomy is more simple visually than a graph and a simple text, and the centrality of the elements by means of other concepts determine its significance. Thus, the organization/layout is the key of the selection. When all the elements are isolated in a layout people choose those words that are more simple and more frequents independently of the context. Furthermore, people answered that the context is vital for select elements, but the position inside the representation is not crucial.

Roughly speaking, human criterion for determining the importance of elements in complex structure representations such as ontologies is not objective. In addition, in this type of graphs, the cardinality and complex OWL axioms are not included. Perhaps other type of representations could be more useful but it is not our study. In addition, as the notion of 'importance' is relative, the importance of a elements may be not set by the position. It is more useful to take into account the frequency or the sparseness ratio. Finally, the evaluation of ontology summarization approaches should avoid the human judgements and it should be based on quantitative measures from computational issues such as: performance indices, data-driven applications and reasoning tasks.

3 OUR APPROACH

Our design try to achieve some basic notions. (i) The representation and its interpretation sets the size of the summarization, the final size depends on the original source. For that, text summarization is variable according with human criteria. As we have checked, each person decides that concepts are more useful. In this way, our design extracts a number undetermined of classes and their respective properties in function of the whole representation. With this design, the evaluation is transparent to the decision of setting a size. As we see experimentally, this indetermination remains constant under a certain percentage of elements. (ii) The ratio between utility and summarization computational cost should be low. It seems more useful to use all ontology elements that the summarization when the second process has a bigger computational cost. This fact impacts in the design of the pruning measures. (iii) According with (Peroni et al., 2008), the pruning measures should mark the elements that are information-rich, that is known by the notion of density. (iv) The previous fact implies that those concepts have a great number of triples and for example, it could be useful in caching techniques. Thus, ontology elements are clustered by their number of triplets, which it means that they are more mentioned.

In contrast with (Peroni et al., 2008), our pruning measures does not consider the label of a concept since the interpretation of labels only is used in human evaluations. Moreover, the number of compound nouns is around 56% considering only concepts. We analysed 693 labels concepts form 13 ontologies which are part of Ontology Alignment Evaluation Initiative (OAEI)[1] in the dataset of *conference* 2010. In our opinion, this group of ontologies can be considered as 'real ontologies' since they present a wide set of OWL axioms and they are not simple taxonomies.

Peroni and Zhang mentioned the notion of density/coverage and centrality/re-ranking weight the classes. Basically, in the first case, the number of subclasses, properties and instances are weighted and divided by the ratio of distance. The final score of a classes depends on three weighted components: it measures this density, another one the popularity hit-based on Yahoo queries, and the last one component considers the label string. In the second approach, the notion of centrality is based on a ratio of input-output degree of vertexes, plus an algorithm to calculate the shortest-path and another based on eigenvector. They apply these algorithms on RDF sentences. In conclusion, OWL axioms are not considered for selecting classes, only the 'centrality' and 'citation' of classes is the common criteria. At the same time, other factors as queries are used to select classes. In the fig. 1, it is represented biosphere ontology where some classes have some extra-circles ac-

[1] http://oaei.ontologymatching.org/

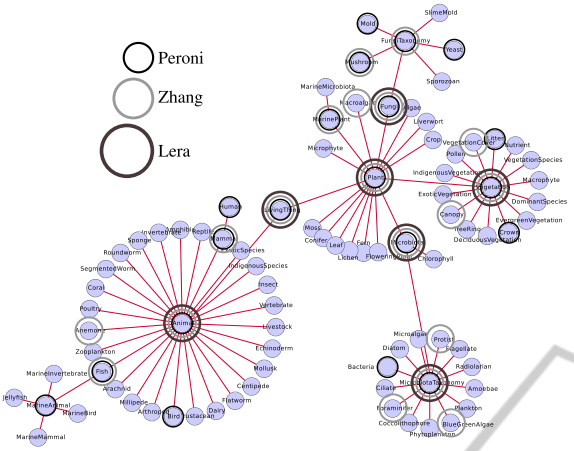

Figure 1: Biosphere selection of classes according with Peroni, Zhang and our approach.

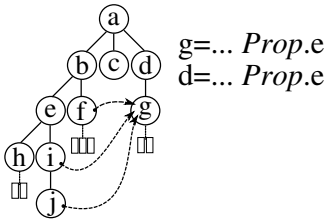

Figure 2: Sketch of the structure of an ontology.

Table 1: Criteria values from the fig. 2.

	classes									
	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**
relative depth	5	4	1	2	3	1	1	1	2	1
direct subclasses	3	2	0	1	2	0	0	0	1	0
inc. range							3			
inc. restrictions					2					
individuals						3	2	2		

cording to Peroni, Zhang and our approach. Thus, there is not ontological difference among some similar classes however they are chosen. For example, in the case of Zhang, it happens with *anemone* class, around *animal* class. The rest of *animal* subclasses are similar to *anemone* but they are not chosen. In the case of Peroni approach, there are: *bird, bacteria, crown* classes, and so on. This type of classes are chosen for criteria applications, basically, specific queries. Although, biosphere ontology is only a taxonomy without individuals, object or data properties, restrictions, etc.

Based on our notions, we synthesize the representation in a group of classes, called structural predominant classes (SPC). Structural also refers the use of complex ontology constructors, besides of traditional hierarchical or properties axioms. We have analysed the set of constructors of RDF, RDFs, and OWL to discover some relationship about centrality, citation, and inference. At the same time, we have studied the correlation among them but there is not relation since each measure is independent and variable. Thus, we decide to have a list of criteria independently among them based on frequencies. Furthermore, due to the low number of restrictions: cardinality, compositions, disjointness, etc. we group the restrictions in only one category.

Following the previous considerations we select the next criteria:

- According with centrality notion:
  - the relative depth, it is the maximum depth of its subclasses,
  - the number of direct subclasses,
- According with citation notion:
  - the number of relationships with a range on it,
- the number of restrictions on it,
- and the number of individuals.

These factors are independents and they have a linear computational cost, so a class could be candidate in various factors and all classes could be tagged as SPC. In the figure 2, we show an example where it is represented an ontology with hierarchical relationships, object properties, individuals -rectangles- and restrictions. At the same time, the table 1 has the value of each class for each criterion of that ontology.

The selection of classes in function of these criteria is filtered in base a common percentile. Using previous example, with a percentile of a 40% we have: *a, b* and *e* by relative depth; *a, b,* and *e*, by direct subclasses; *g* by incidence of range; *e* by incidence of restrictions; *f, g*, and *h* by individuals. The total number of classes is *a, b, e, g, f*, and *h* classes.

As a simple case, `edas` ontology is represented in the fig. 4. This ontology comes from a data case of a OAEI benchmark. The visualization contains two layouts to display SPC. On the left (*a*) is represented by a circular layout. The classes are circles, with a darker colour are the SPC. At-a-glance we observe the greatest number of relationships belonging a SPC group. On right section (*b*), it is a force direct layout. Notice that individuals and restrictions axioms are not represented, thus isolated nodes or no-central positions can be SPC. This image was created using a plug-in specially developed for this research.

This percentile is the only value that is necessary. However, we have analysed 16 ontologies from OAEI *conference* group, and the number of selected classes remains constant. In the fig. 3 the number of SPC does not exceed 40% of the total. Mean value of classes analysed has been 49.41.

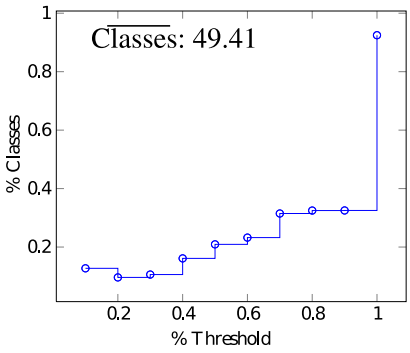

Figure 3: Blue line the number of SPC according a threshold percentile value, using the 16 *conference* ontologies of OAEI.

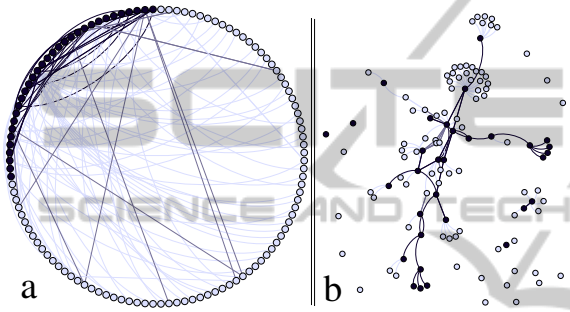

Figure 4: Two layouts of `edas` ontology where SPC are tagged.

4 CONCLUSIONS

We have designed a set of rules to summarize ontology elements, specifically classes. These rules consider all OWL constructors instead of traditional synthesis based on taxonomies and centrality nodes. Moreover, all these rules have a linear computational cost which makes easy to deploy the algorithm in devices with low computational capacities that it is where summarization techniques are more useful.

ACKNOWLEDGEMENTS

This work is partially supported by the project *TIN2011-23889* from Spanish Ministry of Science, Culture and Sport.

REFERENCES

Alani, H., Harris, S., and O'Neil, B. (2006). Winnowing ontologies based on application use. In *Proceedings of the 3rd European conference on The Semantic Web: research and applications*, ESWC'06, pages 185–199, Berlin, Heidelberg. Springer-Verlag.

Li, N. and Motta, E. (2010). Evaluations of user-driven ontology summarization. In *Proceedings of the 17th international conference on Knowledge engineering and management by the masses*, EKAW'10, pages 544–553, Berlin, Heidelberg. Springer-Verlag.

Peroni, S., Motta, E., and D'Aquin, M. (2008). Identifying key concepts in an ontology, through the integration of cognitive principles with statistical and topological measures. In *Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web*, ASWC '08, pages 242–256, Berlin, Heidelberg. Springer-Verlag.

Stuckenschmidt, H. and Schlicht, A. (2009). Structure-based partitioning of large ontologies. In Stuckenschmidt, H., Parent, C., and Spaccapietra, S., editors, *Modular Ontologies*, volume 5445 of *Lecture Notes in Computer Science*, pages 187–210. Springer.

Zhang, X., Cheng, G., and Qu, Y. (2007). Ontology summarization based on rdf sentence graph. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 707–716, New York, NY, USA. ACM.