Two Protégé plug-ins for supporting document-based ontology engineering and ontological annotation at document-level

Viktoria Pammer and Peter Scheir and Stefanie Lindstaedt

May 1, 2007

Abstract

We present two plug-ins for Protégé OWL. The first, Discovery Tab, supports document-based ontology engineering with relevant term extraction, clustering and related functionality. The second, Annotation Tab, provides a facility to annotate documents manually and automatically (based on a training set).

1 Introduction and motivation

Knowledge is available in the minds of human experts. One way of externalizing this knowledge is in the form of documents written in natural language. Document-based ontology engineering aims at using information available in documents to create formal knowledge representations. Done manually, it saves time of the domain experts but not of the ontology engineers. Automatically done it also saves time for the ontology engineers.

Among many things, an ontology can be used as a kind of controlled vocabulary for annotation of entities. Annotation is again a time-consuming task. It can be partly automated however, except for the creation of a training set.

2 Supporting document-based ontology engineering

The Discovery Tab offers functionality to discover knowledge from text-documents. Documents can be grouped to form clusters of different topics, which gives an overview of the domain covered by the documents. Clusters of documents can also be labelled. Relevant terms can be extracted from a set of documents using statistical and natural language processing methods. Terms can be grouped by synonymity. A number of other functionalities like phrase or relation extraction and hierarchical clustering are possible but are not currently implemented.

Within Discovery Tab, standard text mining methods have been used. For relevant term extraction, various pre-processing techniques, such as stop-word lists and stemming, are employed. Extracted terms are weighted by TFIDF. Grouping according to synonymity is currently available only for English and uses WordNet¹ and Apache's Wordnet package². The grouping of documents is a mixture of standard clustering algorithms like Hierarchical Agglomerative Clustering (for seeding) and k-means.

One possible workflow for creating ontological elements from a set of documents could be: Group documents into topic-clusters. This gives an overview over topically different subsets of the domain. Within each cluster, extract relevant terms. To arrange results more clearly, group these terms according to synonymity. Finally, from each term, create either classes or properties.

3 Ontological annotation at document-level

The Annotation Tab offers two functionalities: First, to annotate documents with concepts (actually classes and instances). Second, to automatically suggest concepts for a document if a training set of annotated documents is given. At its backend the Annotation Tab uses a text classification algorithm (k-Nearest Neighbor). This means, that it suggests a set of concepts for a document based on a training set of previously annotated documents. The plug-in presented here shares the backend and a common approach to annotation with the work presented in [2]. However, the frontend as well as the modeling of annotation has been largely revised.

In our approach to semantic annotation we do not aim at identifying concepts at word level. Instead, we follow the tagging metaphor and annotate whole documents with the concepts they deal with. This pragmatic approach follows a motto of which Hendler reminds the Semantic Web community in [1]: "A little semantic goes a long way". In this case, we think that for lots of applications it is not only sufficient but totally sensible to annotate at document level.

4 Conclusion

The interest of our work for a wider scientific community lies first in having grouped together a number of standard text-processing methodologies within one tool. Second it lies in the fact that this tool is embedded in a standard ontology engineering environment, namely Protégé, for which a number of complementary plug-ins (e.g. for rules, querying, visualization etc.) are available. Altogether, this is one further step towards a more and more complete environment in which to implement ontologies.

References

- James Hendler. The dark side of the semantic web. *IEEE Intelligent Systems*, 22(1):2–4, 2007.
- [2] Peter Scheir, Philip Hofmair, Michael Granitzer, and Stefanie N. Lindstaedt. The ontologymapper plug-in: Supporting semantic annotation of text-documents by classification. In *Proceedings of the SEMANTICS 2006*, 2006.

¹http://wordnet.princeton.edu/

²http://lucene.apache.org/java/2_0_0/api/org/apache/lucene/wordnet

< ressue Protépis 3.2.1 (files)D.Aprojektelaposdlelwop1kdmtLevaluierumpldata_task3kressues.pprj. OWL / ROF. Files)								
Ele Eat Broket GML Code Tools Window Help								
Metadata (system_goals own) OvALClasses Properties Annotation The Olscovetry Tab								
CLASS BROWSER	Documents	Term Extractor Document Clusterer						
For Project: • rescue	D: \projekte\aposole\wp1\dmt\dmt_eva	Cluster						
Class Hierarchy 🔒 👻 🔆 💥 🖛	D: 'projekte'aposole'wp1'.dmt'.dmt_eva D: 'projekte'aposole'wp1'.dmt'.dmt_eva	con - concept - icting (6 documents)						
owt:Thing	D:/projekte/aposole/wp1/dmt/dmt_eva	models - customers - furniture (6 documents)						
Actor	D:/projekte/aposole/w/p1/dmt/dmt_eva	goals - soft - linking (3 documents) web - policy - semantics (2 documents)						
Axiom_1	D: \projekte\aposole\wp1\dmt\dmt_eva D: \projekte\aposole\wp1\dmt\dmt_eva	view - procy - admentica (a documenta)						
🕨 🛑 Goal	D: projekte/aposole/wp1/dmt/dmt_eva							
Issues_in_System_Goal_Modelling	D: projekte/aposole/wp1/chnt/chnt_evs							
Model Notation	D: \projekte\aposdle\wyp1\dmt\dmt_eva							
Notation Physical Person	D: \projekte\aposole\wp1\dmt\dmt_eva D: \projekte\aposole\wp1\dmt\dmt_eva							
Product	D: projekte/aposole/wp1/dmt/dmt_eve							
Project	D:\projekte\aposcle\wp1\cimt\cimt_eva	Attended Decements						
Property	D. projekte aposole wyp1 klmt klmt_evs	-Attached Documents						
Resource	D: \projekte\aposole\wp1\dmt\dmt_eva D: \projekte\aposole\wp1\dmt\dmt_eva	Driprojekte/aposslet/wp1/aht/aht_evaluierung/alata_task114.pdf Driprojekte/aposslet/wp1/aht/aht_evaluierung/alata_task118.pdf						
Role	b. projence opositie imprementaria_ere	a. In shore allocant u. k., and an all hand land, a shore land.						
System								
- Tool								
- 88								
PROPERTY BROWSER								
For Project: • rescue								
Object Datatype Annotation All		Cluster Terms of Cluster Files						
0 = 0		Term						
Diject properties		Prove ▲						
playsRolein		dependable						
describes		security						
informs								
isComposedOf		hypezze						
usesNotation		rep 						
hasGoal		··· D service						
validates_modelToModel								
ealsWith								
dependsOn		≥ application						
		Name Cluster						
	•	Group Terms Sementically						
▼ 86	Add Documents	Add Term as Class						
Super Properties	Remove Selected Document(s)	Add Term as Property						
	Start Clustering	Save Term List						
þ								

Figure 1: Screenshot of clustering functionality embedded in Discovery Tab

< ressue Protépis 3.2.1 (file;30:3projekte/aposdle/wp1kdmt_evaluierumpldata_task3kressue.pprj, OWL / ROF. Files)									
The East Project QNL Code Tools Window Help									
Metadata (system, goals ow) 🔴 OVA Classes 🔳 Properties 🛉 Individuals 🚍 Forms 🔶 Annotation Tab									
CLASS BROWSER	Cualified-Mappings-Store	rQu	alified Assignments						
For Project: • rescue	D:/projekte/aposole/wp1/dmt/dmt_evaluierung/data_task3/vescue.n3		npe	Ele	ment(s)				
Class Hierarchy	Load Mappings		Context_Model						
Class Hierarchy	Save Mappings As								
🔻 😑 Issues_in_System_Boundaries 🗇	file: //D:/projekte/aposdle/wp1/dmt/dmt_evaluierung/data_task3/[*								
Context_Model	file://ID:/projekte/aposole/wp1/dmt/dmt_evaluierung/data_task3/f								
System_Boundary	file://D./projekte/sposde/wp1/ktmL/ant_evablerungkista_task3ft								
Strategic_Dependency_Model	ffie:///D./projekte/sposde/wp1kimt/dmt_evaluerungkista_task3/ti ffie://D./projekte/sposde/wp1kimt/dmt_evaluerungkista_task3/ti								
Strategic_Rationale_Model System_Goal_Model	file:///D:/projekte/aposole/wp1/clmt/clmt_evaluierung/data_task3/f								
System_doal_wode	file:///D:/projekte/aposolle/wp1/dmt/dmt_evaluierung/data_task3/f								
Activity_Model	file:///D:/projekte/aposolle/wp1idmt/dmt_evaluierungktata_task3/f								
Context_Model	file:///D:/projekte/aposolle/wp1/dmt/dmt_evaluierungklata_task3/f file://D:/projekte/aposolle/wp1/dmt/dmt_evaluierungklata_task3/f								
Eye_Star_Model	file://D:/projekte/aposole/wp1/dmt/dmt_evaluierungklata_task3/f								
Strategic_Dependency_Model	file:///D:/projekte/aposolle/wp1/dmt/dmt_evaluierung/data_task3/f								
Strategic_Rationale_Model	file:///D:/projekte/aposolle/wp1/dmt/dmt_evaluierung/data_task3/f file://D:/projekte/aposolle/wp1/dmt/dmt_evaluierung/data_task3/f								
System_Goal_Model	file:///D:/projekte/aposole/wp1/dmb/dmt_evaluerung/data_task3/li file:///D:/projekte/aposole/wp1/dmt/dmt_evaluerung/data_task3/li								
Notation (2)	file:///D:/projekte/eposdle/wp1/kimt/kimt_evoluierung/kieto_tesk3/f 👻								
- 38 38									
- 00 00	Remove Selected Document(s)								
INSTANCE BROWSER	Train Classifier		Remove Element(s)						
For Class: 🔴 Notation	Workspace	Workspace // Workspace Assignments							
Asserted Instances 👻 🗣 🗮 🌣	file:///D:/projekte/aposolle/wp1/dmt/dmt_evaluierung/data_task4/IN3C		ype Element(s)	F	tating	ок			
Context_Diagram	file://ID:/projekte/aposolle/wp1idmt/dmt_evaluierungktata_task4/IN3C		Context_Model Context_Diagram		0,277				
Eye_Star_Notation			Notation		0,12				
	Add Document(s) To Workspace								
	Remove Selected Document(s)								
- 88	Classify Selected			Assign Instance(s)	Assign Class(es)				

Figure 2: Screenshot of annotation functionality embedded in Annotation Tab