Coupling and cohesion of modularized ontologies

Miloš Savić Zoran Budimac Gordana Rakić Mirjana Ivanović

University of Novi Sad
Faculty of Sciences
Department of Mathematics and Informatics

August, 2014.

Content

Introduction

Ontology

- Formal representation of shared and reusable knowledge
- Description of concepts (classes), individuals (objects), roles (properties, relationships) in a domain
- Backbone of the Semantic Web, Web Ontology Language (OWL)

Ontology design

- **Monolithic**: all captured concepts, roles, axioms and assertions of the domain are gathered in one monolithic ontology (one file ontologies)
- **Modular**: ontology that consists of multiple ontologies (ontology modules)
  - OWL2 import construction: an OWL 2 ontology can import other ontologies in order to gain access to their entities, expressions, and axioms
  - Better reasoning performance, efficient ontology management, change and reuse

Coupling and cohesion

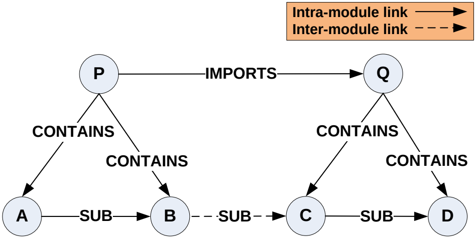
- **Ontology module coupling**: the degree of relatedness between ontology modules
  - low coupling: concepts in a module are not strongly related to concepts in other modules.
- **Ontology module cohesion**: the degree of relatedness of classes in a module
  - high cohesion: concepts in a module tend to be strongly related to other concepts in the module.
- Good, modular ontology design: “Low coupling, high cohesion“ principle

Graph representation of modularized ontologies

- Ontologies describe relations between ontological entities → ontologies can be represented by graphs (networks)
- Graph representation of modularized ontologies
  - Nodes: ontological entities (concepts, individuals, etc.) + **ontologies**
  - Links: relations between ontological entities + **relations between ontologies and ontological entities** + **relations between ontologies**
  - Different types of links
    - Horizontal and vertical links
    - Inter-module and intra-module links
    - Directed and undirected links

```
Ontology(<P>
    Import(<Q>)
    SubClassOf( <P#B> <Q#C> )
    SubClassOf( <P#A> <P#B> )
)

Ontology(<Q>
    SubClassOf( <Q#C> <Q#D> )
)
```



Ontology coupling/cohesion measurement

- Yao et al. [2005] ontology cohesion metrics suite
  - Taxonomy tree considered as the backbone of ontology
  - Number of root classes (NRC), Number of leaf classes (NLC), average depth of leaf nodes in inheritance tree (ADL)
  - Higher values of NRC, NLC and ADL indicate greater separation of concepts
  - **Metrics not suitable for ontology modules**
- Orme et al. [2006] ontology coupling metrics suite
  - NEC (number of external classes used in ontology), REC (number of references to external classes), RI (number of referenced ontologies)
  - (+) Metrics that can be computed locally, without construction of ontology graph
  - (-) Metrics that cover efferent (aggregation) coupling but not afferent (reuse) coupling.

Structural and semantic stable ontology metrics

- Vrandečić and Sure [2007]:
  - Structural metrics - complexity of underlying ontology graph
  - Semantic stable metrics - apply ontological inference (normalization) and then compute metrics
    - Naming of anonymous entities, materialization of subsumption hierarchy, instantiation to the deepest concept, etc.
- **Ontology normalization may break predefined modularization**
  - MSH: $A_1 \sqsubseteq A_2 \sqsubseteq \ldots \sqsubseteq A_n \sqsubseteq A_1 \longmapsto A_1 \equiv A_1 \equiv \ldots \equiv A_n$

```
(Ontology <A>
   Import(<B>)
   SubClassOf(<A#a> <B#b>)
)
```

```
(Ontology <B>
   SubClassOf(<B#b> <A#a>)
)
```

Cohesion and coupling metrics for ontology modules

- Metrics suite introduced by Oh et al. [2011]
- Two coupling metrics
  - NSHR, Number of separated hierarchical relations (inter-module subsumption links)
  - NSNR, Number of separated non-hierarchical relations (other inter-module links)
- One cohesion metric
  - Strength of (indirect) connection between two concepts inversely proportional to the length of the shortest path connecting them
  - Cohesion: sum of strength of intra-module links normalized by the number of all possible intra-module links.

Integrated approach to evaluation of cohesion/coupling of modularized ontologies

- Approach that includes:
  - Domain-independent coupling metrics used in complex network analysis
  - Complex network analysis techniques, hybrid ontology metric set and statistical tests to identify coupling patterns
  - Graph clustering evaluation metrics as ontology cohesion metrics
- Supported by ONGRAM tool
  - Integration of OWL2 into SSQSA (Set of Software Quality Static Analyzers) framework
  - Performs ontology normalization that preserve predefined modularization
    - Complex class expression represented by one node which references all named concepts contained in the expression
    - Property domain and property range relations are assembled

Coupling metrics and patterns

- Module graph: subgraph of ontology graph induced by ontologies and IMPORTS links
- Coupling metrics computed on module graph
  - Centrality metrics used in complex network analysis
    - Local: in-, out- and total degree
    - Global: betweenness, closeness, page rank, etc.
- Coupling patterns
  - Each ontology module characterized by a numeric metric vector M that contains metrics of internal complexity, coupling and cohesion metrics
  - Connected component analysis
    - Strongly and weakly connected components
    - Non-parametric statistical tests to compare metric vectors of different components
  - Assortativity index generalized to M
    - Ex. "Modules having high LOC tend to references modules having high in-degree".

Cohesion metrics and patterns

- "low coupling, high cohesion" principle → elements of a module form a cluster (community) in ontology graph
- Graph clustering evaluation metrics as ontology cohesion metrics
  - Cut based metrics
    - Cut(M) - the number of inter-cluster links connecting elements from M to elements in other clusters
    - Conductance, Expansion and Cut-ratio
  - Out degree fraction (ODF) based metrics
    - ODF(a) - the fraction of inter-cluster links emanating from node a
    - AvgODF, MaxODF, FlakeODF
- Radicchi et al. [2004] notion of community (cluster)
  - Strong community

$$(\forall i \in C) out^{\text{intra-cluster}}(i) > out^{\text{inter-cluster}}(i)$$

  - Weak community

$$\sum_{i \in C} out^{\text{intra-cluster}}(i) > \sum_{i \in C} out^{\text{inter-cluster}}(i)$$

  - Conductance(C) $< 0.5 \rightarrow C$ is weak, FlakeODF(C) $= 1 \rightarrow C$ is strong

Case studies

- NASA SWEET ontologies (Semantic Web for Earth and Environmental Terminology)
  - Collection of ontologies for describing satellite data about Earth
- NIF ontologies (Neuroscience Information Framework)
  - Vocabularies covering major domains of neuroscience (anatomy, cell, etc.), concepts for describing experimental techniques and instruments in neuroscientific studies

	SWEET	NIF
Number of compilation units (ontology modules)	204	37
Total LOC (excluding empty lines)	21449	122482
Parse time (sec.)	8	18
Ontology size (MB)	3.03	23.7
eCST representation size (MB)	25.1	153
The number of nodes in ontology graph	10873	28549
The number of links in ontology graph	26725	88103

Coupling patterns

Properties of ontology module graphs:

	SWEET	NIF
The number of nodes	204	56
The number of external nodes	0	19
The number of links	1340	85
The number of external links	0	19
Small-world coefficient	5.295	1.857
Clustering coefficient	0.234	0.164
The number of WCCs	1	1
The number of SCCs	3	1
Size of the largest SCC	**125 (61.27%)**	**2 (3.57%)**

Two different types of ontology design

- SWEET: large cyclic dependencies between modules
- NIF: hierarchical (layered) design

Coupling patterns (SWEET)

SWEET possesses strongly connected core: LSCC encompasses modules that have significantly higher degree of reuse (DEG-IN), coupling (DEG-TOT) and centrality/importance (BET, PR) than the rest of modules.

No statistically significant differences in volume (LOC, HVOL), internal complexity (EXPR), degree of aggregation (DEG-OUT), richness (CR, RR), strength of efferent coupling (NEC, REC) and cohesiveness (CON, EXP).

Metric	Avg(S)	Avg(R)	U	P-value	P_s	Diff. type
LOC	106.832	102.468	6036	0.007	0.611	small
EXPR	5.264	4.063	5524	0.153	0.559	small
HVOL	2905.051	2877.333	6064	0.006	0.614	small
HDIF	20.502	17.531	6501	< 0.001	0.658	medium
DEG-IN	9.336	2.189	9203	< 0.001	0.932	large
DEG-OUT	5.768	7.835	5002	0.874	0.507	insignificant
DEG-TOT	15.096	10.013	8053	< 0.001	0.815	large
PR	0.007	0.002	9095	< 0.001	0.921	large
BET	870.802	20.299	8904	< 0.001	0.902	large
NCLASS	34.056	26.696	5896	0.02	0.597	small
NINST	9.24	13.797	5051	0.781	0.512	insignificant
AP	1.74	1.269	4971	0.934	0.503	insignificant
CR	0.114	0.085	5189	0.54	0.526	insignificant
RR	0.234	0.226	5141	0.619	0.521	insignificant
NEC	5.12	4.62	5086	0.718	0.515	insignificant
REC	9.488	8.646	4982	0.913	0.505	insignificant
CON	0.209	0.238	5615	0.099	0.569	small
EXP	0.305	0.349	5604	0.105	0.567	small

Cohesiveness of modules

	SWEET	NIF
Radicchi strong modules	25 (12.25%)	15 (40.54%)
Radicchi weak modules	196 (96.08%)	28 (75.68%)
Poorly cohesive modules	8 (3.92%)	9 (24.32%)

Poorly cohesive modules in SWEET:

Module	LOC	EXPRC	IN	OUT	PR	BET	CON
stateSpaceConfiguration.owl	106	0	2	2	0.0016	11	0.75
stateTimeFrequency.owl	72	0	3	7	0.0021	260	0.75
realmEarthReference.owl	144	0	1	10	0.001	0	0.73
quanTimeAverage.owl	89	1	4	8	0.0012	450	0.72
realmAtmoWeather.owl	61	4	1	7	0.001	0	0.65
stateSpace.owl	70	0	1	5	0.001	0	0.63
reprSpaceDirection.owl	97	0	10	2	0.0045	15	0.61
phenOcean.owl	15	1	3	2	0.0013	13	0.6
M	105.1	4.8	6.5	6.5	0.0049	541.4	0.22

Correlations between coupling/cohesion metrics

(a) SWEET

	IN-DEG	OUT-DEG	PR	BET	NEC	REC	CON
OUT-DEG	-0.101						
PR	0.954	-0.201					
BET	0.753	0.254	0.721				
NEC	-0.005	0.697	-0.117	0.269			
REC	-0.042	0.658	-0.144	0.213	0.946		
CON	-0.256	0.46	-0.277	-0.07	0.443	0.515	
EXP	-0.26	0.482	-0.284	-0.061	0.443	0.503	0.982

(b) NIF

	IN-DEG	OUT-DEG	PR	BET	NEC	REC	CON
OUT-DEG	0.195						
PR	0.74	0.131					
BET	0.647	0.587	0.611				
NEC	0.546	0.314	0.382	0.397			
REC	0.552	0.261	0.376	0.352	0.963		
CON	0.327	0.11	0.217	0.058	0.785	0.828	
EXP	0.39	0.162	0.297	0.17	0.83	0.852	0.977

In both cases studies we have moderate to strong correlations between:

- Cohesion metrics (CON, EXP)
- Strength of efferent coupling metrics (NEC, REC)
- Cohesion and strength of efferent coupling metrics - SWEET (moderate), NIF (strong)
- Centrality/importance metrics (BET, PR)
- Reuse (DEG-IN) and centrality/importance metrics

Conclusion

- Overview of coupling/cohesion measurement for modularized ontologies
  - Ontology metrics tend to assume monolithic ontology design
  - Semantic normalizations may break predefined modularization
- Integrated approach to the evaluation of coupling/cohesiveness of modularized ontologies
  - Complex network analysis (CNA), hybrid ontology metrics set and statistical tests
  - Characterization of ontology modularization: identification of coupling/cohesion patterns
  - CNA centrality metrics as ontology module coupling metrics
  - Graph clustering evaluation metrics as ontology module cohesiveness metrics
  - Supported by ONGRAM tool which is part of SSQSA framework

Coupling and cohesion of modularized ontologies

Miloš Savić Zoran Budimac Gordana Rakić Mirjana Ivanović

University of Novi Sad
Faculty of Sciences
Department of Mathematics and Informatics

August, 2014.