Terminological Cluster Trees for Disjointness Axiom Discovery

Giuseppe Rizzo, Claudia d'Amato, Nicola Fanizzi, Floriana Esposito
Dipartimento di Informatica - Università degli Studi di Bari Aldo Moro

14th Extended Semantic Web Conference, Portoroz – 1st June 2017

Outline

1. Introduction & Motivation
2. The approach
3. Empirical Evaluation
4. Conclusions and Outlooks

Introduction & Motivation

Disjointness axioms often missing in Semantic Web knowledge bases

- counterintuitive inference

Example

$\mathcal{K}$={JournalPaper $\sqsubseteq$ Paper, ConferencePaper $\sqsubseteq$ Paper, ConferencePaper(a) }

$\mathcal{K} \models$ JournalPaper(a)?

Answer: No

Contributions

- Goal: discovery disjointness axioms from Semantic Web knowledge bases through inductive method
- Contribution: a framework for eliciting disjointness axioms
  - solving a clustering problem via Terminological Cluster Trees providing a concept description for each cluster
  - Assumption: two concepts disjoint when there is no overlap between their extensions

The approach

Definition

Given

- a knowledge base $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$
- a set of individuals $I \subseteq Ind(\mathcal{A})$

Find

- n pairwise disjoint clusters $\{C_1, \ldots, C_n\}$
- for each $i = 1, \ldots, n$, a concept description D_i that describes C_i, such that:
  - $\forall a \in C_i : \mathcal{K} \models D_i(a)$
  - $\forall b \in C_j, j \neq i : \mathcal{K} \models \neg D_i(b)$.

Terminological Cluster Tree

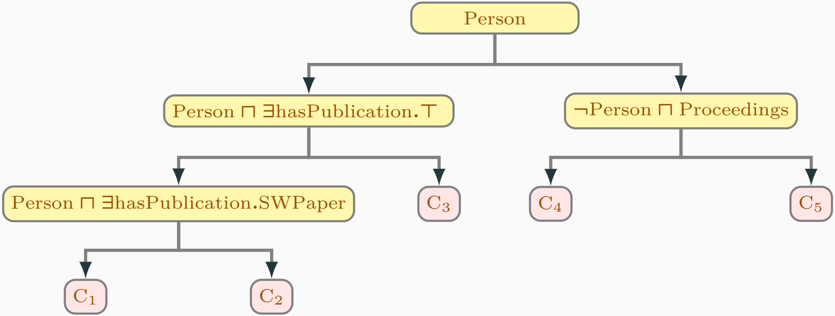
Terminological cluster tree (TCT)

A binary logical tree where

- a node stands for a cluster of individuals C
- each internal node contains a concept D (defined over the signature of $\mathcal{K}$)
- each edge departing from an internal node corresponds to the outcome of the membership test of individuals with respect to D.

Example of TCT

TCT for describing individuals involved in the Semantic Web research community



Inducing a TCT

Given the set of individuals I

- various concept descriptions are generated
  - The sets of positive and negative instances w.r.t. each specialization are determined
  - The medoids of the set of positive (resp. negative) instances are computed
- Best concept: the one minimizing the overlap between the sets of positive and negative instances
  - maximization of the distance between the medoids
- Individuals of the cluster sorted to either set according to the closeness w.r.t. the medoids
- Stop condition: distance between medoids below a threshold ν

Distance measure between individuals

Distance Function (adapted from [d'Amato et al.@ESWC2008]):

$$d_n^{\mathcal{C}} : \mathrm{Ind}(\mathcal{A}) \times \mathrm{Ind}(\mathcal{A}) \to [0, 1]$$

$$d_n^{\mathcal{C}}(a, b) = \left[\sum_{i=1}^{m} w_i \left[1 - \pi_i(a)\pi_i(b) \right]^n \right]^{1/n}$$

Context: a set of atomic concepts $\mathcal{C} = \{B_1, B_2, \ldots, B_m\}$

Projection Function:

$$\forall\ a \in \mathrm{Ind}(\mathcal{A}) \qquad \pi_i(a) = \begin{cases} 1 & \text{if } \mathcal{K} \models B_i(a) \\ 0 & \text{if } \mathcal{K} \models \neg B_i(a) \\ 0.5 & \text{otherwise} \end{cases}$$

Deriving Disjointness Axioms

Given a TCT T:

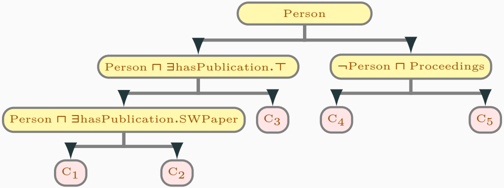
Step I:

- Traversing the tree to collect the concept descriptions describing the clusters
- A set of concepts CS is returned

Step II:

- A set of candidate axioms A is generated from CS:
  - an axiom $D \sqsubseteq \neg E$ ($D, E \in CS$) is generated if
    - $D \not\equiv E$ (or $D \not\sqsubseteq E$ or viceversa)
    - $E \sqsubseteq \neg D$ has not been generated

Example of TCT - Deriving axioms



$$
\begin{aligned}
Cs = \{ \quad & \text{Person}, \\
& \text{Person} \sqcap \exists \text{hasPublication}.\top, \\
& \neg(\text{Person} \sqcap \exists \text{hasPublication}.\top) \\
& \text{Person}\exists \text{hasPublication}.\text{SWPaper} \\
& \neg\text{Person} \sqcap \text{Proceedings} \cdots \}
\end{aligned}
$$

Axiom:

$\text{Person} \sqcap \exists \text{hasPublication}.\text{SWPaper} \sqsubseteq \neg(\neg\text{Person} \sqcap \text{Proceedings})$

Empirical Evaluation

Evaluation: Goals

- Experiment I: Re-discover a target axiom (existing in $\mathcal{K}$)
- Experiment II: Re-discover randomly selected target axioms introduced according to the Strong Disjointness Assumption
  - two sibling concepts in a subsumption hierarchy should be considered as disjoint [Schlobach@ESWC2005]

Empirical Evaluation: Ontologies

Ontology	#Concepts	#Roles	#Individuals	#Disj. Ax.s
BioPax	74	70	323	85
NTN	47	27	676	40
Financial	60	16	1000	113
GeoSkills	596	23	2567	378
Monetary	323	247	2466	236
DBPedia3.9	251	132	16606	11

Experiment I - Settings

- A copy of each ontology is created removing a target axiom
- Threshold $\nu = 0.9, 0.8, 0.7$
- Metrics # discovered axioms and #cases of inconsistency

Ontology	C	D
BioPax	bioSource	xref
NTN	Man	Pop $\sqcup$ Woman $\sqcup$ Supernat. Being
Financial	PermanentOrder	Account $\sqcup$ Region
GeoSkills	Educational Level	Educational Pathway
Monetary	ISO3166-CC	ISO31813-Market IC $\sqcup$ ISO4217-CC
DBpedia3.9	Activity	Person

Experiment I - Outcomes

Ontology	TCT 0.9		TCT 0.8		TCT 0.7	
	#inc.	#ax's	#inc.	#ax's	#inc.	#ax's
BioPax	2	53	2	53	3	52
NTN	10	70	9	73	10	75
Financial	0	125	0	126	0	127
GeoSkills	2	345	1	347	4	347
Monetary	0	432	0	432	0	433
DBPedia3.9	45	45	44	44	43	43

- limited number of inconsistency
- significant number of discovered axioms

Experiment II -Settings

- A copy of each ontology is created removing 20%, 50%, 70% of the disjointness axioms
- The copy is used to induce TCT
- Threshold $\nu = 0.9, 0.8, 0.7$
- Metrics: rate of rediscovered target axioms, #cases of inconsistency and # discovered axioms
- # Run: 10 times
- Comparison against the methods bases on negative association rule mining (NAR) and Pearson correlation coefficient (PCC)

2nd Experiment - Outcomes

Ontology	f	TCT 0.9		TCT 0.8		TCT 0.7		PCC		NAR	
		#inc.	#ax's	#inc.	#ax's	#inc.	#ax's	#inc.	#ax's	#inc.	#ax's
BioPax	20%	235	3859	357	4235	365	4256	257	280	352	2990
	50%	125	3576	357	4176	432	4115				
	70%	125	3432	235	3875	417	4154				
NTN	20%	312	3128	343	3126	354	3124	32	957	376	3766
	50%	234	3023	234	3034	235	3034				
	70%	156	2987	176	2679	123	2675				
Financial	20%	76	165	87	325	96	276	124	1112	542	5366
	50%	37	143	56	307	53	259				
	70%	33	143	43	276	40	221				
GeoSkills	20%	234	14289	357	14297	432	14345	456	13384	456	13299
	50%	231	14123	356	14154	417	14256				
	70%	234	14122	358	14154	377	14187				
Monetary	20%	535	13456	573	13453	623	13460	543	13384	423	13456
	50%	315	13236	432	13236	532	13236				
	70%	247	13127	231	13127	312	13127				
DBPedia3.9	20%	1345	29730	1432	30143	1432	30567	1243	30470	1243	30365
	50%	1346	29730	1431	30143	1433	30567				
	70%	1343	19730	1432	30143	1432	30567				

Experiment II - Outcomes

Ontology	f	TCT – standard mode		
		TCT 0.9	TCT 0.8	TCT 0.7
BioPax	20%	0.90 ± 0.12	0.76 ± 0.13	0.74 ± 0.13
	50%	0.85 ± 0.13	0.74 ± 0.13	0.74 ± 0.13
	70%	0.85 ± 0.13	0.74 ± 0.12	0.74 ± 0.14
NTN	20%	0.99 ± 0.08	0.95 ± 0.06	0.95 ± 0.08
	50%	0.97 ± 0.03	0.93 ± 0.10	0.93 ± 0.01
	70%	0.90 ± 0.10	0.89 ± 0.11	0.89 ± 0.10
Financial	20%	0.99 ± 0.08	0.99 ± 0.08	0.99 ± 0.08
	50%	0.97 ± 0.03	0.97 ± 0.03	0.97 ± 0.03
	70%	0.95 ± 0.05	0.95 ± 0.05	0.95 ± 0.05
GeoSkills	20%	0.99 ± 0.08	0.99 ± 0.08	0.99 ± 0.08
	50%	0.92 ± 0.10	1.00 ± 0.00	1.00 ± 0.00
	70%	0.92 ± 0.10	0.92 ± 0.10	0.92 ± 0.10
Monetary	20%	0.99 ± 0.08	1.00 ± 0.00	1.00 ± 0.00
	50%	0.94 ± 0.13	1.00 ± 0.00	1.00 ± 0.00
	70%	0.94 ± 0.13	0.91 ± 0.14	0.91 ± 0.13
DBPedia3.9	20%	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	50%	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	70%	0.96 ± 0.08	0.90 ± 0.08	0.90 ± 0.08

Conclusions and Outlooks

Conclusions and Outlooks

- An approach to discover disjointness axioms
  - exploiting the knowledge base
  - generating axioms that involves complex concept descriptions
- Extensions:
  - Post-pruning strategy
  - Other distance measures

Questions?

Refinement Operators

Downward refinement operators specializing a concept C

$\boldsymbol{\rho}_1$ $\mathrm{C}' = \mathrm{C} \sqcap (\neg)\mathrm{A}$;

$\boldsymbol{\rho}_2$ $\mathrm{C}' = \mathrm{C} \sqcap (\neg)(\exists)\mathrm{R}.\top$;

$\boldsymbol{\rho}_3$ $\mathrm{C}' = \mathrm{C} \sqcap (\neg)(\forall)\mathrm{R}.\top$;

$\boldsymbol{\rho}_4$ $\exists \mathrm{R}.\mathrm{C}'_\mathrm{i} \in \boldsymbol{\rho}(\exists \mathrm{R}.\mathrm{C}_\mathrm{i}) \wedge \mathrm{C}'_\mathrm{i} \in \boldsymbol{\rho}(\mathrm{C}_\mathrm{i})$;

$\boldsymbol{\rho}_5$ $\forall \mathrm{R}.\mathrm{C}'_\mathrm{i} \in \boldsymbol{\rho}(\forall \mathrm{R}.\mathrm{C}_\mathrm{i}) \wedge \mathrm{C}'_\mathrm{i} \in \boldsymbol{\rho}(\mathrm{C}_\mathrm{i})$.

Example of axioms

Successfully discovered axioms

- $\text{ExternalReferenceUtilityClass} \sqcap \exists \text{TAXONREF}.\top$
  disjoint with
  xref
- Activity
  disjoint with
  $\text{Person} \sqcap \exists \text{nationality}.\text{United\_states}$
- $\text{Person} \sqcap \text{hasSex}.\text{Male}$ ($\equiv$ Man)
  disjoint with
  $\text{SupernaturalBeing} \sqcap \text{God}$ ($\equiv$ God)

Not discovered axioms

- Actor disjoint with Artefact

(concepts with few instances)