Discovering Alignments in Ontologies of Linked Data

Rahul Parundekar, Craig A. Knoblock, and Jose Luis Ambite

Winner of Best Paper: Research Track at ISWC 2012

USC **Viterbi** — School of Engineering

INFORMATION SCIENCES INSTITUTE · · · · · agent of innovation · · · · ·

Motivation

Data Integration on the Web

- There is a need for the interoperability of information from different sources on the Web
- This involves joining data at the
  - Object Level
  - Schema Level

The Semantic Web Community is Linking Different Sources at the Object Level

Example: Geospatial Domain

However, Not Many Sources have Links at the Schema Level

Source 1 / Source 2 — Schema/Ontology Level: Populated Place — NO LINKS!! — City. Object Level: Los Angeles — owl:sameAs — City of Los Angeles

Problem is Non-Trivial

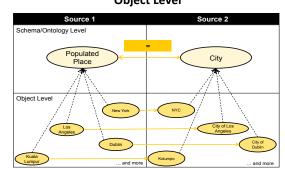- **Sources may have poor ontologies**

	DBpedia	GeoNames
Description	Semantic version of Wikipedia	Geographical Database
#Classes	359 (Well defined hierarchy)	1 (Single Class for all Objects)

DBpedia: Rich, Descriptive. GeoNames: Impoverished

- **Need expressive alignments (not just 1-to-1 class matches)**

Approach

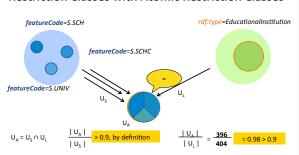### Align Concepts when Supported by Evidence at the Object Level

Source 1 / Source 2 — Schema/Ontology Level: Populated Place = City. Object Level: New York — NYC; Los Angeles — City of Los Angeles; Dublin — City of Dublin; Kuala Lumpur — Kolumpo; … and more

Create *New* Concepts to Overcome the Rudimentary Ontology Problem

1. Atomic Restriction Classes*

   Set of all Objects in GeoNames → **All Places** *i.e. Set of all objects with* $featureClass=P$

2. Conjunctive Restriction Classes

   All Objects in the US *i.e. with* $countryCode=US$ — **All Schools** *i.e. Objects with* $featureCode=S.SCH$ — **All Schools in the US** *i.e.* $featureCode=S.SCH$ *&* $countryCode=US$

3. Disjunctive Restriction Classes

   Set of All Schools, Colleges and Universities *i.e. Union of Set of Objects with* $featureCode=S.SCH, S.SCHC,$ *or* $S.UNIV$

   * Value Restrictions in OWL-DL

Align Atomic & Conjunctive Restriction Classes by Comparing the Overlap of the Set of Objects

$featureClass=P$; $rdf{:}type=PopulatedPlace$

r_1 — $Img(r_1)$ i.e. "linked" objects in $Source_2$; r_2 ; =

$$\frac{|\ Img(r_1) \cap r_2\ |}{|\ Img(r_1)\ |} > 0.9 \qquad \frac{|\ Img(r_1) \cap r_2\ |}{|\ r_2\ |} > 0.9$$

A relaxed threshold of 0.9, which was decided empirically, is used to overcome missing instances and incorrect alignments

Find Concept Coverings by Aligning Disjunctive Restriction Classes with Atomic Restriction Classes

$featureCode=S.SCH$; $featureCode=S.SCHC$; $featureCode=S.UNIV$; $rdf{:}type=EducationalInstitution$

U_S ; U_A ; U_L ; =

$$U_A = U_S \cap U_L \qquad \frac{|\ U_A\ |}{|\ U_S\ |} > 0.9, \text{ by definition} \qquad \frac{|\ U_A\ |}{|\ U_L\ |} = \frac{396}{404} = 0.98 > 0.9$$

Results

Example Alignments of Atomic Restriction Classes

Restriction Class from GeoNames	Restriction Class from DBpedia	Rel	P	R	$\|Img(r_1) \cap r_2\|$
We find alignments with concepts that hadn't existed before – e.g. Concept of Places					
featureClass=P	rdf:type=PopulatedPlace	=	99.6	90.5	70658
We also find alignments with geographical regions like countries, administrative divisions, etc. – e.g. alignment of the Concepts for the country Spain					
countryCode=ES	country=Spain	=	94.5	99.9	3917
We find the real relationship between concepts as against the perceived one					
featureCode=T.MT	rdf:type=Mountain		96.8	78.4	1721

Example Alignments of Conjunctive Restriction Classes

Restriction Class from GeoNames	Restriction Class from DBpedia	Rel	P	R	$\|Img(r_1) \cap r_2\|$
We find alignments with conjunctions on either sides – e.g. Concepts of 'Places in the US' are Equal					
featureClass=P & countrycode=US	rdf:type=PopulatedPlace & country=United_States	=	97.2	96.7	26061
We also find alignments with conjunctive restriction classes that have related properties – e.g. Places in North Dakota have 701 as area code for phone numbers					
featureClass=P & parentADM1= North_Dakota	areaCode=701	=	98.1	96.5	361
In some highly specialized classes, we notice that the meaning of a concept shifts slightly with limited evidence - e.g. Populated Places in Senegal are aligned to Towns rather than PopulatedPlaces					
featureClass=P & countryCode=SN	rdf:type=Town & country=Senegal	=	92.6	100	25

Example Concept Coverings

Larger Restriction Class	Union of Smaller Restriction Classes	Rel	R	Over -lap	Outliers
We find concept coverings with disjunctive restriction classes – e.g. Educational Institution concept in Dbpedia covers concepts of Schools, Colleges and Universities					
rdf:type= dbpedia: Educational_Institution	featureCode= {S.SCH, S.SCHC, S.UNIV}	=	98.0	396 / 404	S.BLDG, S.HSP, S.MUS, etc.
We are able to flag outliers, which need to be corrected					
rdf:type= dbpedia:Airport	featureCode= {S.AIRB, S.AIRP}	=	99.2	1981 / 1996	**S.AIRF**, S.FRMT, S.SCH, **T.HLL**, etc.
Another example: We are able to find all terms used for the country Netherlands					
countryCode=NL	country= {dbpedia:The_Netherlands, dbpedia:Flag_of_the_Netherlands.svg, dbpedia:Netherlands}	=	98.0	1939 / 1978	dbpedia: Kingdom_of_t he_Netherlan ds

Detection of Outliers:

Objects that contradict the overwhelming evidence of the alignment may have wrong links or values

Restriction Class from GeoNames	Restriction Class from DBpedia	Rel	R	$\|Img(r_1) \cap r_2\|$	Outliers
countryCode=ES	country=Spain	=	99.9	3917 / 3918	*countryCode=IT (1/7635)*
geonames:featureCode= {S.SCH, S.SCHC, S.UNIV}	rdf:type= dbpedia: Educational_Institution	=	98.0	396 / 404	*featureCode= S.HSP (1/31), S.BLDG (3/122), S.EST (1/13), S.MUS (1/43), S.LIBR (1/7)*