LSDIS

Large Scale Distributed Information Systems

University of Georgia
Computer Science Department

OntoQA: Metric-Based Ontology Quality Analysis

Samir Tartir, I. Budak Arpinar,
Michael Moore, Amit P. Sheth,
Boanerges Aleman-Meza

IEEE Workshop on Knowledge Acquisition
from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources

Houston, Texas, November 27, 2005

The Semantic Web

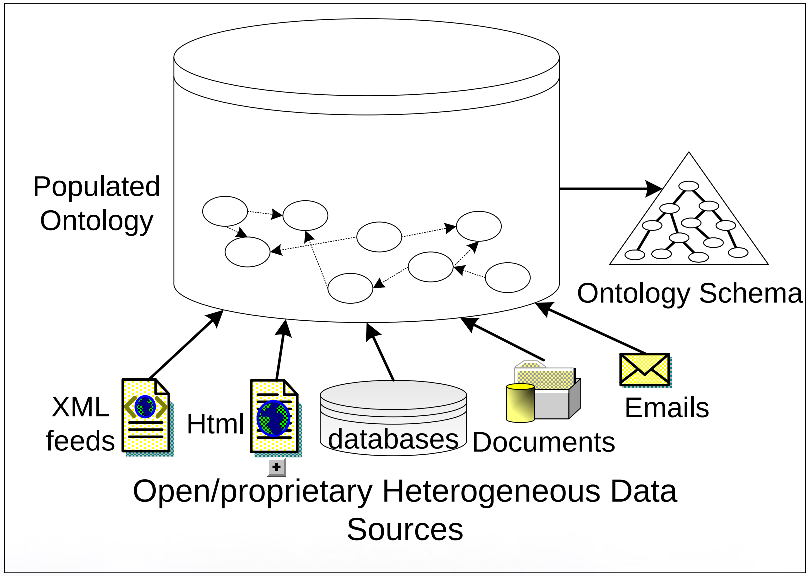
- Current web is intended for human use
- Semantic web is for humans and computers
- Semantic web uses ontologies as a knowledge-sharing vehicle.
- Many ontologies currently exist: GO, OBO, SWETO, TAP, GlycO, PropreO, etc.

Motivation

- Having several ontologies to choose from, users often face the problem of selecting the best ontology that is suitable for their needs.

OntoQA

- Metric-Based Ontology Quality Analysis
- Describes ontology schemas and instancebases (IBs) through different sets of metrics
- OntoQA is implemented as a part of SemDis project.



Contributions

- Defining the quality of ontologies in terms of:
  - Schema
  - Instances
    - IB Metrics
    - Class-extent metrics

- Providing metrics to quantitatively describe each group

I. Schema Metrics

- Schema metrics address the design of the ontology schema.

- Schema quality could be hard to measure: domain expert consensus, subjectivity etc.

- Three metrics:
  - Relationship richness
  - Attribute richness
  - Inheritance richness

I.1 Relationship Richness

- How close or far is the schema structure to a taxonomy?
- Diversity of relations is a good indication of schema richness.

$$RR = \frac{|P|}{|IsA| + |P|}$$

|P|: Number of non-IsA relationships

|IsA|: Number of IsA relationships

I.2 Attribute Richness

- How much information do classes contain?

$$AR = \frac{|A|}{|C|}$$

|A|: Number of literal attributes

|C|: Number of classes

I.3 Inheritance Richness (Fan-out)

- General (e.g. spanning various domains) vs. specific

$$IR_S = \frac{\sum_{C_i \in C} \left| H^C\left(C_j, C_i\right) \right|}{\left| C \right|}$$

$|H^c(c_j, c_i)|$: Number of subclasses of Class C_i

$|C|$: Number of classes

II. Instance Metrics

- Deal with the size and distribution of the instance data.

- Instance metrics are grouped into two subcategories:
    1. **IB metrics**: describe the IB as a whole
    2. **Class metrics**: describe the way each class that is defined in the schema is being utilized in the IB

II.1.a Class Richness

- How much does the IB utilizes classes defined in the schema?
- How many classes (in the schema) are actually populated?

$$CR = \frac{|C`|}{|C|}$$

|C'|: Number of used classes

|C|: Number of defined classes

II.1.b Average Population

- How well is the IB “filled”?

$$P = \frac{|I|}{|C|}$$

|I|: Number of instances

|C|: Number of defined classes

II.1.c Cohesion

- Is IB graph connected or disconnected?

$$Coh = |CC|$$

|CC|: Number of connected components

II.2.a Importance

- How much focus was paid to each class during instance population?

$$Imp_{C_i} = \frac{|C_i(I)|}{|I|}$$

$|C_i(I)|$: Number of instances defined for class C_i

$|I|$: Number of instances

II.2.b Connectivity

- What classes are central and what are on the boundary?

$$Conn_{C_i} = \left| \{ I_j : \exists P(I_i, I_j) \wedge I_i \in C_i(I) \wedge I_j \in C_j(I), \forall C_j \in C \} \right|$$

$P(I_i,I_j)$: Relationships between instances I_i and I_j.

$C_i(I)$: Instances of class C_i.

C: Defined classes.

II.2.c Fullness

- Is the number of instances close to the expected?

$$F = \frac{|C_i(I)|}{|C_i`(I)|}$$

$|C_i(I)|$: Number of instances of class C_i.

$|C_{i'}(I)|$: Number of expected instances of class C_i.

II.2.d Relationship Richness

- How well does the IB utilize relationships defined in the schema?

$$RR_{C_i} = \frac{\left|\{ Distinct(P(I_i, I_j)) : I_i \in C_i(I), I_j \in C_j(I), \forall C_j \in C \}\right|}{\left|P(C_i, C_j)\right|}$$

$P(I_i,I_j)$: Relationships between instances I_i and I_j.

$C_i(I)$: Instances of class C_i.

$C_j(I)$: Instances of class C_j.

C: Defined classes

$P(C_i,C_j)$: Relationships between instances C_i and C_j.

II.2.e Inheritance Richness

- Is the class general or specific?

$$IR_{C_i} = \frac{\sum_{C_j \in C'} \left| H^C\left(C_k, C_j\right) \right|}{\left| C' \right|}$$

C’: Classes belonging to the subtree rooted at Ci

$|H^c(c_k, c_j)|$: Number of subclasses of Class Ci

Implementation

- Written in Java
- Processes ontology schema and IB files written in OWL, RDF, or RDFS.
- Uses the Sesame to process the ontology schema and IB files.

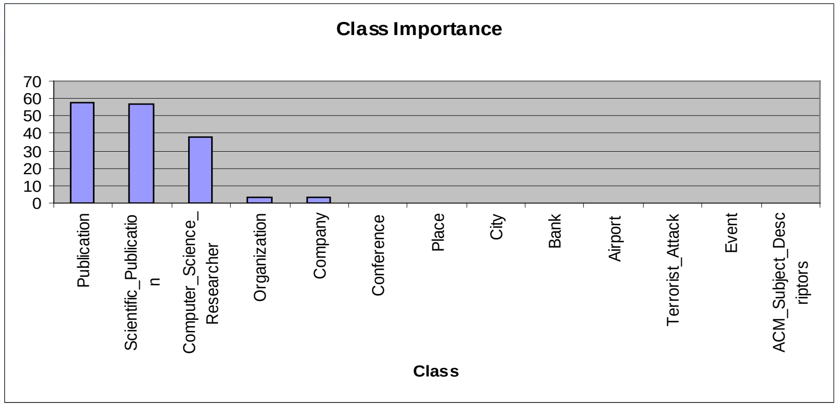
Testing

- SWETO: LSDIS' general-purpose ontology that covers domains including publications, affiliations, geography and terrorism.
- TAP: Stanford's general-purpose ontology. It is divided into 43 domains. Some of these domains are publications, sports and geography.
- GlycO: LSDIS' ontology for the Glycan Expression
- OBO: Open Biomedical Ontologies

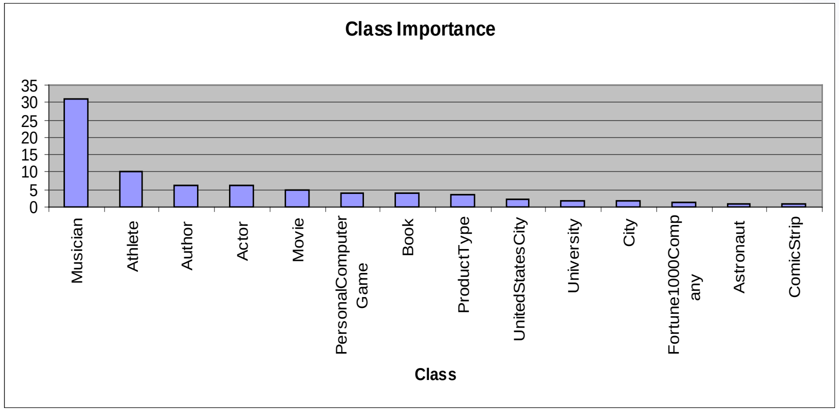
Results – Class Metrics

Ontology	# of Classes	# of Instances	Inheritance Richness	Class Richness	Average Population
SWETO	44	1,003,021	0.9	56.8%	22,795.9
TAP	3,230	71,487	1.2	9.4%	22.1
GlycO	356	387	1.3	18.0%	1.1
PropreO	244	0	1.0	0.0%	0.0

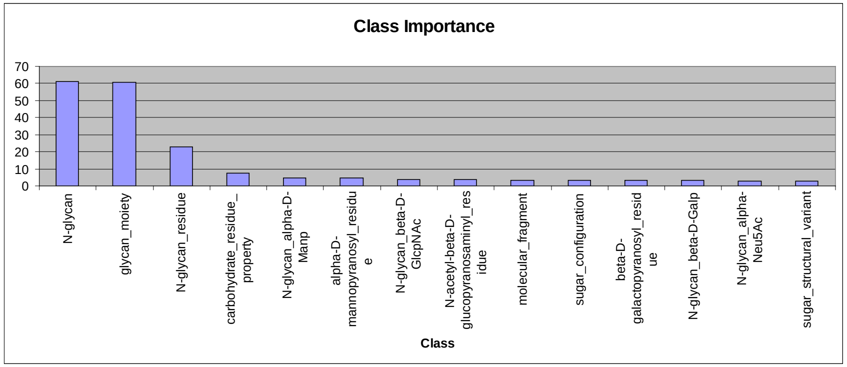
Results – Class Importance



SWETO

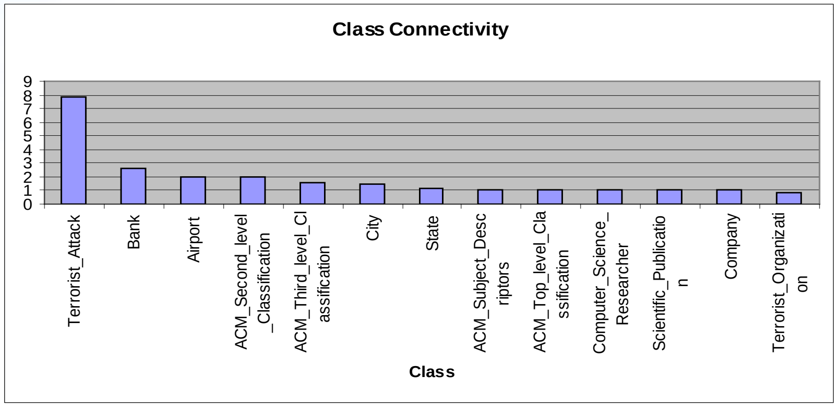


TAP

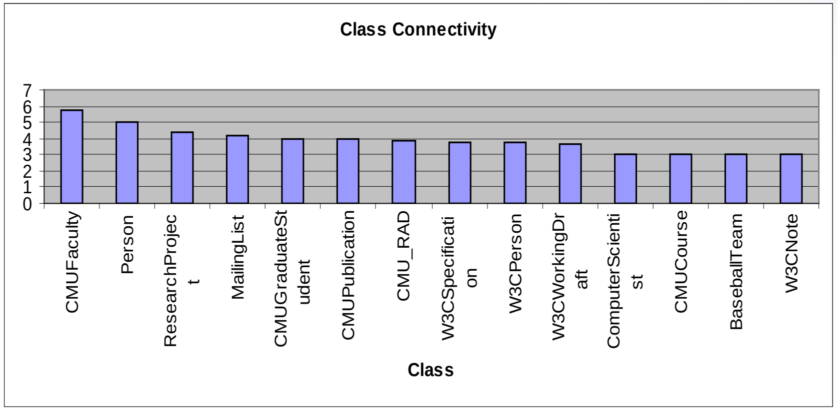


GlycO

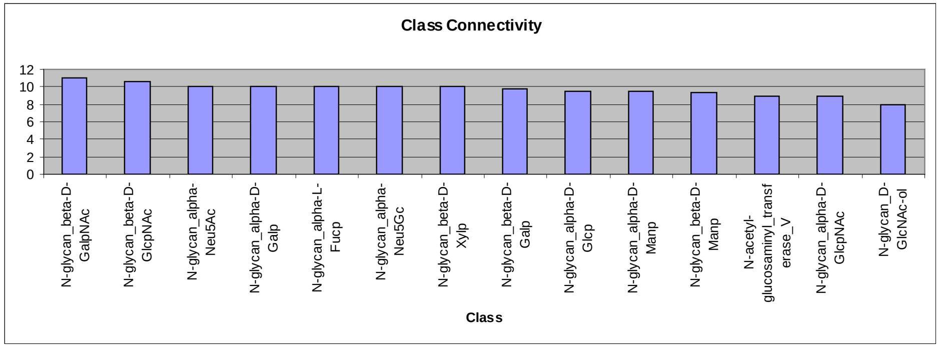
Results – Class Connectivity



SWETO



TAP



GlycO

BioMedical Ontologies

Ontology	No. of Terms (Instances)	Average No. of Subterms	Connectivity
Protein-protein Interaction	195	4.6	1.1
MGED	228	5.1	0.3
Biological Imaging Methods	260	5.2	1.0
Physico-chemical Process	550	2.7	1.3
Cereal Plant Trait	692	3.7	1.1
BRENDA	2,222	3.3	1.2
Human Disease	19,137	5.5	1.0
Gene Ontology	20,002	4.1	1.4

Conclusions

- More ontologies are introduced as the semantic web is gaining momentum.
- There is no easy way for users to choose the most suitable ontology for their applications.
- OntoQA offers 3 categories of metrics to describe the quality and nature of an ontology.

Future Work

- Calculation of domain dependent metrics that makes use of some standard ontology in a certain domain.

- Making OntoQA a web service where users can enter their ontology files paths and use OntoQA to measure the quality of the ontology.

Questions