SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation

By
Thong Chanchaem

Contents

Introduction

What is … ?

- **Semantic Web** : a vision of a future web of machine understandable document and data.
- **XML, RDF and OWL** : semantic web format.

Introduction

OWL [2] is a set of XML elements and attributes, with standardized meaning, that are used to define term and their relationship.

OWL extends RDF Schema:

OWL	Class, equivalentProperty, sameIndividualAS …
RDF Schema	SubClassOf, resource, ID ...

Introduction (OWL example)

Looking for a Camera with 75-300mm. zoom lens size, an aperture of 4.5-5.6

Relevant document?

```
.
.
<SLR rdf:ID="Olympus-OM-10"
xmlns="http://www.camera.org#">
<lens>
  <Lens>
        <focal-length>75-300mm zoom</focal-
length>
       <f-stop>4.5-5.6</f-stop>
  </Lens>
</lens>
<body>
.
.
```

Camera.OWL

```
.
.
<owl:Class rdf:ID="SLR">
     <rdfs:subClassOf rdf:resource ="#Camera"/>
 </owl:Class>
.
.
<owl:DatatypeProperty rdf:ID="focal-length">
    <owl:equivalentProperty rdf:resource="#size"/>
    <rdfs:domain rdf:resource="#Lens"/>
    <rdfs:range rdf:resource="&xsd;#string"/>
</owl:DatatypeProperty>
.
.
<owl:DatatypeProperty rdf:ID="f-stop">
    <owl:equivalentProperty rdf:resource="#aperture"/>
    <rdfs:domain rdf:resource="#Lens"/>
    <rdfs:range rdf:resource="&xsd;#string"/>
</owl:DatatypeProperty>
```

Note: SLR single lens reflex

Introduction

TAP KB : a knowledge base that contains a board range of lexical and taxonomic about popular object like: music, movie, author, place, etc.

- Browse the TAP KB
- Example of Places.rdf file
- Tap Activity Based Search

Goal

- To perform automated semantic tagging of large corpora.
- To introduce a new disambiguation algorithm to resolve ambiguities in a natural language corpus.
- To introduce the platform which different tagging applications can share.

How they do that ?

- **SemTag** : an application written on the platform that perform automated semantic tagging of large corpora.
- **Seeker** : a platform for large-scale text analytics.
- **TBD** : a new algorithm for Taxonomy-Based Disambiguation.

SemTag

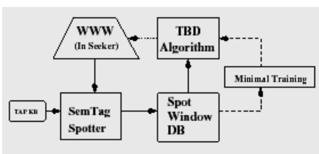
"The Chicago Bulls announced yesterday that Michael Jordan will …"

The <resource ref=http://tap.stanford.edu/BasketballTeam_Bulls">Chicago Bulls</resource>announced yesterday that <resource ref="http://tap.stanford.edu/AthleteJordan,_Michael">Michael Jordan</resource> will …"

SemTag

- SemTag uses TAP KB to build a web scale ontology
- SemTag uses the concept of label bureau from PICS to obtain semantic annotation from the third party.

SemTag Archtecture



Two fundamental categories of ambiguities

- Some labels appear at multiple locations in the TAP ontology.
- Some entities have labels that occur in contexts that have no representative in the taxonomy.

Term definitions

- O *(ontology)* is defined by four elements
  - C (Class)
  - S $\subseteq$ C × C (subClass relation)
  - I $\subseteq$ I × C (instances relation)
  - T $\subseteq$ I × C (type relation)
- T *(Taxonomy)* is defined by three elements
  - V (a set of nodes)
  - r $\in$ V (a root)
  - p : V $\rightarrow$ V

Algorithm Sim

Sim(c, v)
    Let $b = \underset{u \in \pi(v)}{\operatorname{argmin}} \{ f_u(c) \}$
    if $b = r$ return 0
    else return 1

Algorithm TBD

TBD(c, u)
    Let u be the nearest ancestor of v with a measurement.
    if $| 0.5 - m_u^a | > | 0.5 - m_u^s |$
        if $m_u^a > 0.5$
            return 1
        else
            return 0
    else
        if $m_u^s > 0.5$
            return Sim(c, u)
        else
            return 1 - Sim(c, u)

Results of SemTag

They applied SemTag to set of 267 million pages producing 270G of dump data corresponding to 550 million labels in context.

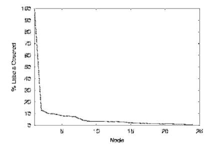
Approximately 79% are judged to be on-topic, resulting in a final set of about 434 million spots, with accuracy around 82%.

Nodes of TAP

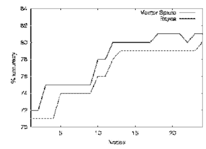
Node	Fraction of spots
Class	100.00%
UnitedStatesCity	12.97%
ProfessionalType	10.21%
Country	9.66%
Musician	8.14%
City	7.86%
ProductType	7.31%
Fortune1000Company	4.41%
TechnologyBrand	3.45%
PersonalComputerGame	3.45%
University	3.45%
Book	3.17%
Movie	3.03%
UnitedStatesState	2.90%
Actor	2.07%
OperatingSystem	1.93%
MusicalInstrumentBrand	1.66%
ComedyTVShow	1.38%
Author	1.38%
ConsumerElectronicsCorporation	1.10%
Athlete	1.10%
ComicStrip	0.97%
HomeAndGardenBrand	0.83%
SportingGoodsBrand	0.83%

Nodes of TAP with percentage of spots occurring in corresponding subtree.

Results of SemTag



Percentage of spots influenced by hand classified data



Accuracy of the two algorithms employed in SemTag

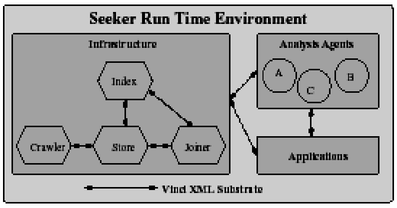
Design goal for Seeker

- Composibility
- Modularity
- Extensibility
- Scalability
- Robustness

Infrastructure Components

- The Data Store
- The Index
- The Joiner

Architecture of the Seeker system



Advantage

- Other application can obtain semantic annotation from web-available database.
- They use both human and computer judgment to solve ambiguous data in their TBD algorithm.

Disadvantage

- The system requires a large amount of storage space to store data.

Future SemTag

- They will use some techniques to bootstrap from TAP to build much larger and richer ontologies in the future.
- Currently, SemTag uses RDF but in the future, SemTag will use advanced language as OWL.

Critical review

- This system writes the resulting annotations to the database which other mechanisms can obtain the data from. Can this concept work with any dynamic pages?
- They use only 11 volunteers to exam the selections. Is it enough? And Does the background of volunteers influence of the judgment of label selecting?

Quiz

- What is the different between SemTag and Seeker?
- Why OWL is more advanced language than RDF?
- What does TBD do?
- Which ontology is used in the system?
- What makes SemTag and Seeker different form other applications?

References

[1] http://tap.stanford.edu

[2] http://www.xfront.com/owl-quick-intro/sld001.htm

[3] http://www.w3.org/

[4] **SemTag and Seeker : Bootstrapping the semantic web via automated semantic annotation**