Semantic Web Technology Evaluation Ontology (SWETO): A Test Bed for Evaluating Tools and Benchmarking Applications

Boanerges Aleman-Meza

Amit P. Sheth
Wright State University - Main Campus, amit.sheth@wright.edu

I. Budak Arpinar

Chris Halaschek

Follow this and additional works at: http://corescholar.libraries.wright.edu/knoesis

Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

Repository Citation
http://corescholar.libraries.wright.edu/knoesis/741

This Presentation is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu.
Semantic Web Technology Evaluation Ontology (SWETO): A test bed for evaluating tools and benchmarking semantic applications


Boanerges Aleman-Meza, Amit Sheth, I. Budak Arpinar, Chris Halaschek
Large Scale Distributed Information Systems Lab
University of Georgia

This material is based upon work supported by the National Science Foundation under NSF-ITR-IDM Award # 0325464 and NSF-ITR-IDM Award # 0219649.
Motivation for SWETO

- Many new techniques and software tools from emerging Semantic Web (SW) community
  - Need a common infrastructure for testing

- Need of an open and freely available ontology with a very large knowledge base
  - Scalability testing as the most important objective
  - Quality and comparability as other criteria
SWETO Objectives

- Develop a broad and deep ontology populated with real facts/data from real world heterogeneous sources
  - the instances in the knowledge base should be highly interconnected
- Serve as a test-bed for advanced semantic applications (i.e. business intelligence, national security, etc.)
- Address the requirements of a research benchmark for semantic analytics, and the semantic techniques of:
  - ontology creation
  - semi-automatic extraction
  - entity disambiguation
The big picture

Open/proprietary Heterogeneous Data Sources

- documents
- databases
- Html pages
- XML feeds
- emails

Massive Metadata Store

SWETO Web Service

Knowledge Discovery Algorithms

Browsing

API

Ontology / knowledge base

Trusted Sources

populates

Semi-structured data
Utilized Semagix Freedom for SWETO ontology design and population
Development Framework

- Utilized Semagix Freedom for ontology design and population
- With Freedom, knowledge extractors were created to extract entities from various data sources
Development Framework

- Data sources:
  - Selected sources which were highly reliable Web sites that provide entities in a
    - semi-structured format
    - unstructured data with parse-able structures (e.g., html pages with tables)
    - dynamic web sites with database back-ends
  - Considered the types and quantity of implicit/explicit relationships
    - preferred sources in which instances were interconnected
  - considered sources whose entities would have rich metadata
  - Public and open sources were preferred
    - due to the desire to make SWETO openly available
Development Framework

- As the sources are processed by the extractors, entities are extracted and stored in appropriate classes in an ontology.

- Due to heterogeneous data sources, entity disambiguation is a crucial step:
  - Freedom’s disambiguation techniques automatically resolved entity ambiguities in 97% of the cases, leaving the rest for human disambiguation (and may be ignored).
Development Framework

- Utilize Freedom’s API for exporting both the ontology and its instances in either RDF [5] or OWL [2] syntax

- Extractors are scheduled to rerun for keeping the ontology updated
Current Status

- V.1 population includes over 800,000 entities and over 1,500,000 explicit relationships among them.
- Continue to populate the ontology with diverse sources thereby extending it in multiple domains, new larger release due soon.
- Significant information for provenance/trust support [UMBC partnership].
## Current Status – Classes

<table>
<thead>
<tr>
<th>Subset of classes in the ontology</th>
<th># Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cities, countries, and states</td>
<td>2,902</td>
</tr>
<tr>
<td>Airports</td>
<td>1,515</td>
</tr>
<tr>
<td>Companies, and banks</td>
<td>30,948</td>
</tr>
<tr>
<td>Terrorist attacks, and organizations</td>
<td>1,511</td>
</tr>
<tr>
<td>Persons and researchers</td>
<td>307,417</td>
</tr>
<tr>
<td>Scientific publications</td>
<td>463,270</td>
</tr>
<tr>
<td>Journals, conferences, and books</td>
<td>4,256</td>
</tr>
<tr>
<td>TOTAL (as of January 2004)</td>
<td>811,819</td>
</tr>
</tbody>
</table>
## Current Status – Relationships

<table>
<thead>
<tr>
<th>Subset of relationships</th>
<th># Explicit relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>located in</td>
<td>30,809</td>
</tr>
<tr>
<td>responsible for (event)</td>
<td>1,425</td>
</tr>
<tr>
<td>Listed author in</td>
<td>1,045,719</td>
</tr>
<tr>
<td>(paper) published in</td>
<td>467,367</td>
</tr>
</tbody>
</table>
## Current Status – Disambiguation

<table>
<thead>
<tr>
<th>Disambiguation type</th>
<th># Times used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic (Freedom)</td>
<td>248,151</td>
</tr>
<tr>
<td>Manual</td>
<td>210</td>
</tr>
<tr>
<td>Unresolved (Removed)</td>
<td>591</td>
</tr>
</tbody>
</table>
Browsing of the Schema
Evaluation/Usage 1: Industry

- Evaluation of Fast Semantic Enhancement

(in Marianas SDK)

[Hammond, Sheth, Kochut 2002]
Blue-chip bonanza continues

Dow above 9,000 as HP Home Depot lead advance; Microsoft upgrade helps techs.
August 22, 2002: 11:44 AM EDT

By Alexandra Twin, CNN/Money Staff Writer

New York (CNN/Money) - An upgrade of software leader Microsoft and strength in blue chips including Hewlett-Packard and Home Depot were among the factors pushing stocks higher at midday Thursday, with the Dow Jones industrial average spending time above the 9,000 level.

Around 11:40 a.m. ET, the Dow Jones industrial average gained 65.06 to 9,022.09, continuing a more than 1,300-point resurgence since July 23. The Nasdaq composite gained 9.12 to 1,418.37.

The Standard & Poor's 500 index rose 9.61 to 958.97.

Hewlett-Packard ( HPQ: up $0.33 to $15.03, Research, Estimates) said a report shows its share of the printer market grew in the second quarter, although another report showed that its share of the computer server market declined in Europe, the Middle East and Africa.

Home Depot ( HD: up $1.07 to $33.75, Research, Estimates) was up for the third straight day after topping fiscal second-quarter earnings estimates on Tuesday.

Tech stocks managed a turnaround. Software continued to rise after Salomon Smith Barney upgraded No. 1 software maker Microsoft ( MSFT: up $0.55 to $52.83, Research, Estimates) to "outperform" from "neutral" and raised its price target to $59 from $56. Business software makers Oracle ( ORCL: up $0.18 to $10.94, Research, Estimates), PeopleSoft ( PSFT: up $1.17 to $20.67, Research, Estimates) and BEA Systems ( BEAS: up $0.28 to $7.12, Research, Estimates) all rose in tandem.
Application 2: Web of Belief (WOB) by UMBC

- Web Of Belief (WOB) framework that maintains trust and provenance for SWETO
- L. Ding, P. Kolari, A. Joshi, T. Finin, Y. Yesha (UMBC)

Presented at: “Trust on the Web Track” (also at Developers Day)
Ongoing work

- Quality measures of the ontology

- Access to the ontology
  - Web service
  - Filtering, views and versioning

- On-the-fly semantic annotation
Future plans for benchmarking

- Semantic Search, Browsing and Personalization
- Semantic Portals
  - i.e., *SEMPL* automatically identifies entities
- Semantic Analytics
  - Discovery of *semantic associations* [{\(p\)-operator}]
  - Example apps: CIRAS (Semagix), PISTA
Approach to developing Semantic Analytics Application Benchmarking
Conclusions & Future Work

- Using Semagix Freedom, we have created a broad and deep Semantic Web Evaluation Ontology (SWETO)
  - Public access under Creative Commonsense license;
  - Looking for usage, feedback (of all kinds: schema, population, quality), and partners (for developing benchmarks)
Conclusions & Future Work

- More extraction of entities focusing on partners’ needs
- Also plan to further investigate the use of semantic similarity for entity disambiguation
- Ontology lifecycle support
SWETO Project Homepage

- [http://lsdis.cs.uga.edu/Projects/Semdis/SWETO/](http://lsdis.cs.uga.edu/Projects/Semdis/SWETO/)
  - Google or other search engine: “SWETO”
  - Project description, papers, presentations

- Acknowledgements: This work is partially funded by NSF-ITR-IDM Award #0325464 and NSF-ITR-IDM Award #0219649.
References


