3-28-2006

Semantic Web Applications in Financial Industry, Government, Health Care and Life Sciences

Amit P. Sheth

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

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/30

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 applications in Financial Industry, Government, Health care and Life Sciences

SWEG 2006, March 2006

Amit Sheth
LSDIS Lab, Department of Computer Science, University of Georgia
http://lsdis.cs.uga.edu
SW research @ LSDIS

- Ontology design and population
- Automatic Metadata Extraction
- Semantic Annotations (text, medical docs, scientific data)
- Semantic Computations: (Inference), Rules, Complex Relationships, Knowledge Discovery, Semantic Associations
- Semantic Visualization
- Active Semantic Documents
- Semantic Web Services/Processes
- Semantic Applications: Bioinformatics, Health Care, Intelligence/Gov., (Commercial: Risk & Compliance, Content Aggregators)
- Semantics Enabled Networking
Part II: Semantic Web Applications in Government

- Passenger Threat Analysis
- Need to Know -> Demo*
- Financial Irregularity

*on the Web: Google “SemDis”, go to: NeedToKnow
Financial Irregularity

Aim

– Ability to automate the detection of financial inconsistency and irregularity

Problem

– Need to create a unified and logically rigorous terminology of financial domain
– Need to integrate data from multiple disparate structured and semi-structured sources
– Need to create, store, update and execute analytic formulas on financial data
Financial Irregularity

Approach

– Creation of financial domain ontology, populated from trusted sources
– Creation of multiple extractors to disambiguate data and form relevant relationships
– Creation of framework for mathematical formula/rule specification and semantic querying of ontology
Financial Irregularity

Solution

– Developed ontology schema for financial domain using modeling capabilities of Semagix Freedom toolkit
– Extracted, merged, and linked financial data from multiple sources using the extraction and disambiguation capabilities of Semagix Freedom toolkit
– Utilized MathML, a Mathematical Markup Language, to represent mathematical formulas and rules
– Extended MathML to include ability to represent RDF subgraphs of paths through the financial ontology
Financial Irregularity
Financial Irregularity

Subset of Financial Domain Ontology
Financial Irregularity

Graphical User Interface

Ontology Calculator

Traversals

Formula Construction

Rule Construction

Ontology Hierarchy

Location
Report
Identification
Agent
Finance
Event
Financial Irregularity

Creation of financial asset variable “bank account value”
Financial Irregularity

Creation of financial asset variable “bank account value”

\[
\begin{align*}
&\text{MathML-S} \\
&\text{(Semantics on MathML)}
\end{align*}
\]
Financial Irregularity

Creation of financial liability variable “loan value”
Financial Irregularity

Creation of financial formula “solvency ratio”
Financial Irregularity

Creation of financial rule “solvency ratio check”
Financial Irregularity

Result display of "solvency ratio check" rule execution
Semantic Visualization
Semantic Visualization

Aim

- Provide a comprehensive visualization and interactive search and analytics interface for exploiting Semantic Web capabilities

Problem

- Need for intuitive visualization of highly expressive ontologies (e.g., complex carbohydrate molecules)
- Need for intuitive visual display of semantic analytics showing "connections between the dots" between heterogeneous documents and multi-modal content
- Need for graphical tracking and association of activities to discover semantic associations between events using thematic and topological relations
Semantic Visualization

Solution

- **OntoVista** is an ontology visualization tool, with unique capabilities related to complex (representationally rich) biological and biochemical ontologies.

- **Semantic Analytics Visualization (SAV)** is a 3D visualization tool for Semantic Analytics. It has the capability for visualizing ontologies and meta-data including annotated web documents, images, and digital media such as audio and video clips in a synthetic three-dimensional semi-immersive environment.

- **Semantic EventTracker (SET)** is a highly interactive visualization tool for tracking and associating activities (events) in a Spatially Enriched Virtual Environment (SEVE).
GlycO – A domain ontology for glycans
OntoVista representation of Glycan Molecule (with monosaccharide residue composition)
Pathway representation in GlycO

Pathways do not need to be explicitly defined in GlycO. The residue-, glycan-, enzyme- and reaction descriptions contain the knowledge necessary to infer pathways.
Zooming in a little …

Reaction R05987 catalyzed by enzyme 2.4.1.145 adds_glycosyl_residue N-glycan_b-D-GlcpNAc_13
Semantic Analytics Visualization representation of entities and relationships

Entities
- blue rectangles

Relationships
- arrows between entities - a yellow rectangle above the arrow is the property's label
Overview of Virtual Environment

- GraphViz’s "Dot" layout of instances and their relationships in the foreground.

- In the background, the document nodes are shown as red 3D ovals.
Interaction

- Remote object selection using ray casting.
- A laser beam extends from the user's hand to infinity.
- The first object that is penetrated by the laser is selected.
“Detail” of Selection, “Overview” still visible

- After a selection of a property (shown at the center of the figure), all entities and properties become semi-transparent but the selected property and the attached entities.
- Additionally, all documents become semi-transparent but the common documents attached to the entities.
Layout using “dot”

- "Dot" layout of instances and their relationships
- (no documents are shown for clarity)
Layout using “neato”

- "Neato" layout of instances and their relationships
- no documents are shown for clarity
Space Partitioning

**Foreground**
- visualization of entities and their properties in the foreground.

**Background**
- documents are visualized in the background.
Semantic EventTracker representation of geospatial and temporal dimensions for semantic associations

- Visualization of association unfolding over time
- Integration of associated multimedia content
- Separate Temporal, Geospatial, and Thematic ontologies describe data
Part III: A Healthcare Application

Thanks to our collaborators the Athens Heart Center & Dr. Wingeth ©UGARF and Amit Sheth (except when attributed to someone else).
Active Semantic Document

A document (typically in XML) with
- Lexical and Semantic annotations (tied to ontologies)
- Actionable information (rules over semantic annotations)

Application: Active Semantic Patient Record for Cardiology Practice.
- 3 populated ontologies
- EMRs in XML
Practice Ontology
Practice Ontology
Drug Ontology Hierarchy (showing is-a relationships)
Drug Ontology showing neighborhood of *PrescriptionDrug* concept
First version of Procedure/Diagnosis/ICD9/CPT Ontology
Active Semantic Doc with 3 Ontologies


Problem List:
1. Hypertension (365.04) [E]
2. Cholecystectomy (576.0) [E]
3. Chest Pain [E]

Chief Complaint: Evaluation of abnormal EKG status post abnormal Echo. Evaluation of aortic stenosis status post arterial examination. Cardiac clearance for aneurysm removal. Follow up of recent hospitalization at Barrow Community Hospital for acute myocardial infarction.

History of Present Illness: He was evaluated at Athens Regional Medical Center emergency room by Dr. Harry Wingate. He is here today for cardiac clearance for aneurysm removal. The patient reports chronic moderate burning and cramping chest pain located across the chest, which radiates to the arms. He reports that his chest pain is aggravated by movement. The chest pain is relieved by breathing deeply. Patient's history is positive for the following cardiovascular risk factors: diabetes and family history of coronary artery disease.

Current Medications
- Actos 30 mg, 1 tab [E]
- Coumadin tablets 10 mg, 1 tab [F E]
- Viagra 50 mg, 1 tab [F E]
- Zyrtec 5 mg, 1 tab [E]
- Zyvox 2 mg/ml, 1 inj [A E]

Allergies: LINEZOLID

Impressions:
1. Abdominal aortic aneurysm, advanced secondary to by a positive nuclear scan.
2. Abnormal cardiac study associated with chest tightness appears to be secondary to a noncardiac cause as evidenced by arterial scan of lower extremities.
3. Normal cardiac cath.
Explore neighborhood for drug *Tasmar*

<table>
<thead>
<tr>
<th>Current Medications</th>
<th>Intropin injection 40mg/ml, 1 inj qd E</th>
<th>Intropin injection 40mg/ml, 1 inj qd E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tasmar tablets 200mg, 1 tab qd E</td>
<td>Adal 100mg/5ml, 1 susp qd E</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tasmar tablets 200mg, 1 tab qd E</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tylenol extended release 650mg, 1 tab qd E</td>
</tr>
</tbody>
</table>

**Pharmacy:** Carson's Commerce Drug Company  
**Phone:** 706-754-4128  
**Phone:** 706-336-3111

**Allergies:** AMPICILLIN, MP DYE, PENCILLIN, TYLENOL

**Past Medical History:** No past trauma

**Family History:** Mr. Prabhakar has a positive family history of coronary artery disease. The patient's spouse is deceased.

**Psychosocial:** Mr. Prabhakar resides in an apartment home. He is lives with his daughter to help. He has excellent social support.

**Life History:**

**Review of Systems:**

**General:** Patient reports daily child associated with chest pain. These symptoms have not been worked up by his primary care physician. Patient reports recent unintentional weight gain. This problem has not been worked up by his primary care physician. Patient reports recent onset of severe trouble falling asleep. This problem has been working up by his primary care physician.

**HEENT:** The patient reports migraine headache.


**Hematologic:** Patient reports history of blood transfusion as a result of anemia. Patient reports he did experience a transfusion reaction.

**Skin:** Patient reports frequent of mild pruritus associated with weakness. This problem has been worked up by the patient's dermatologist. Patient reports a walnut-sized keratoses that is located over the entire body.

**Psychiatric:** Patient admits to a history of panic attacks that is currently managed by common mental health. His symptoms are felt to be not under control.

**Vital Signs:** Height: 170 Weight: 280 lbs BP: 120/80 Pulse: 80

**Physical Examination:**

**General:** The patient appears the stated age.
Explore neighborhood for drug *Tasmar*

Semantic browsing and querying-- perform decision support (how many patients are using this class of drug, …)
More on ontologies, Languages and Rules

• Schema
• Population (knowledge source)
• Freshness

• Use of W3C standards (XML, RDF, OWL, RQL/SPARQL, SWRL)
On-line demo of Active Semantic Electronic Medical Record
(deployed at Athens Heart Center)

For on line demo: Google: Active Semantic Documents
• **Extreme User Friendliness**
  - Electronic Health Record is the focus of all activities, no extra search, no switching of windows

• **Error Prevention**

• **Decision Support**

• **Better Patient Support and Insurance management**
Part IV: Biological Applications

Funded by NIH-NCRR

Acknowledgement: NCRR funded *Bioinformatics of Glycan Expression*, collaborators, partners at CCRC (Dr. William S. York) and Satya S. Sahoo, Christopher Thomas, Cartic Ramakrishan.
Computation, data and semantics in life sciences

• “The development of a predictive biology will likely be one of the major creative enterprises of the 21st century.” Roger Brent, 1999

• “The future will be the study of the genes and proteins of organisms in the context of their informational pathways or networks.” L. Hood, 2000

• "Biological research is going to move from being hypothesis-driven to being data-driven." Robert Robbins

• We’ll see over the next decade complete transformation (of life science industry) to very database-intensive as opposed to wet-lab intensive.” Debra Goldfarb

We will show how semantics is a key enabler for achieving the above predictions and visions.
Semantic GlycoInformatics - Ontologies

- **GlycO**: A domain ontology for glycan structures, glycan functions and enzymes (embodying knowledge of the structure and metabolisms of glycans)
  - Contains 600+ classes and 100+ properties – describe structural features of glycans; unique population strategy

- **ProPreO**: a comprehensive process Ontology modeling experimental proteomics
  - Contains 330 classes, 6 million+ instances
  - Models three phases of experimental proteomics
GlycO

Large Scale Distributed Information Systems
GlycO – A domain ontology for glycans
Ontology population workflow

**GLYCAN: G00020**

**Entry**  
G00020  
Glycan

**Composition**  
(GlcNAc)5  (Man)3  (Asn)1

**Mass**  
1502.4  (Asn)

**Structure**

```
  GlcNAC{1}----2 Man
    \   /       / \
     6       7   8
  Man{1}______GlcNAC{1}______GlcNAC{1}_______Asn
       \                     /
        \                    
         \                   
          \                 
           \               
            \             
             \           
              \         
               \      
                \    
                 \  
                  \ 
                   \ 
                    \ 
                     GlcNAC
                      G00020
```

**Class**  
Glycoprotein; N-Glycan

**Reaction**

[R05987]  [R05991]

**Pathway**  
PATH: map00510  N-Glycan biosynthesis

**Enzyme**

2.4.1.145  2.4.1.155

**Ortholog**

KO: K00738  alpha-1,3-mannosylglycoprotein
    beta-1,4-N-acetylgalactosaminyltransferase
KO: K00744  alpha-1,3(5)-mannosylglycoprotein
    beta-1,5-N-acetylgalactosaminyltransferase

**Other DBs**

CCSD: 4362 5050 7635 30020 33040 34937 35723 42024 43245

**LinkDB**

[All DBs]

**KCF data**

[Show]
N-Glycosylation Process (NGP)

Cell Culture
- extract

Glycoprotein Fraction
- proteolysis

Glycopeptides Fraction
- Separation technique I

Glycopeptides Fraction
- PNGase

Peptide Fraction
- Separation technique II

Peptide Fraction
- Mass spectrometry

- ms data
  - Data reduction
  - ms peaklist

- ms/ms data
  - Data reduction
  - ms/ms peaklist

Glycopeptide identification and quantification
- N-dimensional array
  - Data correlation
  - Peptide list

Signal integration

Peptide identification
- binning
- Peptide list

Data reduction
Phase II: Ontology Population

- Populate ProPreO with all experimental datasets?
- Two levels of ontology population for ProPreO:
  - Level 1: Populate the ontology with instances that are stable across experimental runs
    Ex: Human Tryptic peptides - 1.5 million+ instances in ProPreO
  - Level 2: Use of URIs to point to actual experimental datasets
Web Services based Workflow = Web Process

WORKFLOW

Web Service 1

WS1

WS 2

WS 3

WS 4

Web Service 2

Web Service 3

Web Service 4

LINUX

MAC

Windows XP

Solaris
Semantic Annotation of Scientific Data

830.9570 194.9604 2
580.2985 0.3592
688.3214 0.2526
779.4759 38.4939
784.3607 21.7736
1543.7476 1.3822
1544.7595 2.9977
1562.8113 37.4790
1660.7776 476.5043

ms/ms peaklist data

<ms/ms_peak_list>
<parameter
    instrument=micromass_QTOF_2_quadropole_time_of_flight_mass_spectrometer
    mode = "ms/ms"/>
<parent_ion_mass>830.9570</parent_ion_mass>
<total_abundance>194.9604</total_abundance>
<z>2</z>
<mass_spec_peak m/z = 580.2985 abundance = 0.3592/>
<mass_spec_peak m/z = 688.3214 abundance = 0.2526/>
<mass_spec_peak m/z = 779.4759 abundance = 38.4939/>
<mass_spec_peak m/z = 784.3607 abundance = 21.7736/>
<mass_spec_peak m/z = 1543.7476 abundance = 1.3822/>
<mass_spec_peak m/z = 1544.7595 abundance = 2.9977/>
<mass_spec_peak m/z = 1562.8113 abundance = 37.4790/>
<mass_spec_peak m/z = 1660.7776 abundance = 476.5043/>
<ms/ms_peak_list>

Annotated ms/ms peaklist data
Semantic annotation of Scientific Data

Annotated ms/ms peaklist data
Summary, Observations, Conclusions

- Ontology Schema: relatively simple in business/industry, highly complex in science
- Ontology Population: could have millions of assertions, or unique features when modeling complex life science domains
- Ontology population could be largely automated if access to high quality/curated data/knowledge is available; ontology population involves disambiguation and results in richer representation than extracted sources, rules based population
- Ontology freshness (and validation—not just schema correctness but knowledge—how it reflects the changing world)
Summary, Observations, Conclusions

- Quite a few applications: semantic search, semantic integration, semantic analytics (AML, need to know, financial irregularity), decision support and validation (e.g., error prevention in healthcare), knowledge discovery, process/pathway discovery, ...
More information at