7-2008

Semantic Web: Promising Technologies, Current Applications & Future Directions

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

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: Promising Technologies, Current Applications & Future Directions

Invited and Colloquia talks at: Swinburne Institute of Technology –Melbourne (July 18), University of Adelaide-Adelaide (July 23), University of Melbourne- Melbourne (July 31), Victoria University- Melbourne Australia, 2008

Amit P. Sheth
amit.sheth@wright.edu
Kno.e.sis Center, Comp. Sc & Engg
Wright State University, Dayton OH, USA

Thanks Kno.e.sis team and collaborators
Outline

- Semantic Web – key capabilities and technologies
- Real-world Applications demonstrating benefit of semantic web technologies
- Exciting on-going research
Evolution of the Web

Web of pages
- text, manually created links
- extensive navigation

Web of databases
- dynamically generated pages
- web query interfaces

Web of resources
- data = service = data, mashups
- ubiquitous computing

Web of people
- social networks, user-created casual content
- Twine, GeneRIF, Connotea

2007
- "ask the Web": using semantics to leverage text + data + services
- Powerset

Web as an oracle / assistant / partner
of Semantic Web
• **Ontology**: Agreement with a common vocabulary/nomenclature, conceptual models and domain Knowledge

• Schema + Knowledge base

• Agreement is what enables **interoperability**

• Formal description - Machine processability is what leads to **automation**
• Semantic Annotation (Metadata Extraction): Associating meaning with data, or labeling data so it is more meaningful to the system and people.
• Can be manual, semi-automatic (automatic with human verification), automatic.
• **Reasoning/Computation**: semantics enabled search, integration, answering complex queries, connections and analyses (paths, sub graphs), pattern finding, mining, hypothesis validation, discovery, visualization
Different foci

• TBL – focus on data: Data Web ("In a way, the Semantic Web is a bit like having all the databases out there as one big database.")

• Others focus on reasoning and intelligent processing
From Syntax to Semantics

Types of Metadata and Annotations

**Ontology**
(Example: Anatomy, Diagnostics, ...)

**Semantic Metadata**
(Example ontology-driven metadata:
Region: Upper Abdomen
Organ: Liver
Pathological_Structure: Abscess, Abscess located_in Liver)

**Structural Metadata**
(document structure: DTDs, XSL clustering and similarity processing: concept extraction)

**Syntactic Metadata**
(language, format, document length, creation date, source, audio bit rate, encryption, affiliation, date last reviewed, authorization, ...)

**Data**
(Structured, semi-structured and unstructured)

Deep semantics

Shallow semantics

Expressiveness, Reasoning

Semantics for deeper understanding, meaningful analysis and actionable information
a little bit about ontologies
<table>
<thead>
<tr>
<th>Ontology Description</th>
<th>Namespace</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>MESH</td>
<td>mesh.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Microarray experimental conditions</td>
<td>MGEDOntology.owl</td>
<td>OWL</td>
</tr>
<tr>
<td>Molecular function</td>
<td>gene ontology.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Molecule role (INOH Protein name/family name ontology)</td>
<td>molecule role.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Mosquito gross anatomy</td>
<td>mosquito anatomy.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Mouse adult gross anatomy</td>
<td>adult mouse anatomy.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Mouse gross anatomy and development</td>
<td>EMAP.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Mouse pathology</td>
<td>mouse pathology.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Multiple alignment</td>
<td>mao.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>NCI organismal classification</td>
<td>taxonomy.dat</td>
<td>plain text</td>
</tr>
<tr>
<td>NCI Thesaurus</td>
<td>EVS ftp site</td>
<td>OWL</td>
</tr>
<tr>
<td>NMR-instrument specific component of metabolomics investigations</td>
<td>nmr.owl</td>
<td>OWL</td>
</tr>
<tr>
<td>OBO relationship types</td>
<td>relationship.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Pathway ontology</td>
<td>pathway.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>PATO</td>
<td>quality.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Physico-chemical methods and properties</td>
<td>fix.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Physico-chemical process</td>
<td>rex.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Plant environmental conditions</td>
<td>environment ontology.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Plant growth and developmental stage</td>
<td>po_temporal.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Plant structure</td>
<td>po_anatomy.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Plasmodium life cycle</td>
<td>PLO.obo</td>
<td>GO</td>
</tr>
<tr>
<td>Protein covalent bond</td>
<td>[none]</td>
<td></td>
</tr>
<tr>
<td>Protein domain</td>
<td>InterPro FTP directory</td>
<td>XML</td>
</tr>
<tr>
<td>Protein modification</td>
<td>psi-mod.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Protein-protein interaction</td>
<td>psi-mi.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Proteomics data and process provenance</td>
<td>ProPreO.owl</td>
<td>OWL</td>
</tr>
<tr>
<td>Sequence types and features</td>
<td>so.obo</td>
<td>OBO</td>
</tr>
<tr>
<td>Systems Biology</td>
<td>sbo.obo</td>
<td>OBO</td>
</tr>
</tbody>
</table>

Open Biomedical Ontologies, http://obo.sourceforge.net/
From simple ontologies
Drug Ontology Hierarchy
(showing is-a relationships)
to complex ontologies
N-Glycosylation metabolic pathway

GNT-I attaches GlcNAc at position 2

\[
\text{UDP-N-acetyl-D-glucosamine} + \alpha-D-\text{Manosyl-1,3-(R1)-\beta-D-mannosyl-R2} \leftrightarrow \text{UDP} + \text{N-Acetyl-}\beta-D-\text{glucosaminyl-1,2-\alpha-D-mannosyl-1,3-(R1)-\beta-D-mannosyl-R2}
\]

GNT-V attaches GlcNAc at position 6

\[
\text{UDP-N-acetyl-D-glucosamine} + \text{G00020} \leftrightarrow \text{UDP} + \text{G00021}
\]
A little bit about semantic metadata extractions and annotations
Create/extract as much (semantics) metadata automatically as possible; Use ontologies to improve and enhance extraction
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 66.08 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.
Semantic Web in Action

Supporting Clinical Decision Making
1. Supporting Clinical Decision Making

- Status: In use today
- Where: Athens Heart Center
- What: Use of Semantic Web technologies for clinical decision support
Active Semantic Electronic Medical Records (ASEMR)

Goals:

- Increase efficiency with decision support
  - formulary, billing, reimbursement
  - real time chart completion
  - automated linking with billing
- Reduce Errors, Improve Patient Satisfaction & Reporting
  - drug interactions, allergy, insurance
- Improve Profitability

Technologies:

- Ontologies, semantic annotations & rules
- Service Oriented Architecture

Thanks -- Dr. Agrawal, Dr. Wingeth, and others. ISWC2006 paper
Click to Launch
Further Opportunity: Clinical and Biomedical Data

Collection of (heterogeneous) Documents
- text
- XML
- HTML pages
- Databases

Scientific Literature
- PubMed
  - 300 Documents Published Online each day
- Elsevier iConsult

Health Information Services
- User-contributed Content (Informal)
  - GeneRifs

NCBI Public Datasets
- Genome, Protein DBs
  - new sequences daily

Clinical Data
- Personal health history

Laboratory Data
- Lab tests, RTPCR, Mass spec

Search, browsing, complex query, integration, workflow, analysis, hypothesis validation, decision support.
Semantic Web in Action

Querying Integrated Data Sources
2. Querying Integrated Data Sources

• **Status**: Completed research

• **Where**: NIH

• **What**: Querying Integrated Data Sources
  – Enriching data with ontologies for integration, querying, and automation
  – Ontologies beyond vocabularies: the power of relationships
Link between glycosyltransferase activity and congenital muscular dystrophy?

Adapted from: Olivier Bodenreider, presentation at HCLS Workshop, WWW07
On the World Wide Web

Knowledge Enabled Information and Services Science

Adapted from: Olivier Bodenreider, presentation at HCLS Workshop, WWW07

Congenital muscular dystrophy, type 1D

has-associated_disease

Gene Ontology

Function

Acetylglucosaminyltransferase activity

Process

N-acetylglucosamine metabolic process
carbohydrate biosynthetic process
glycosphingolipid biosynthetic process
muscle maintenance
protein amino acid glycosylation

Evidence

TAS PubMed
TAS PubMed
TAS PubMed
TAS PubMed

(GeneID: 9215)

Adapted from: Olivier Bodenreider, presentation at HCLS Workshop, WWW07
With the Semantically Enhanced Data

SELECT DISTINCT ?t ?g ?d {
  ?t is_a GO:0016757.
  ?g has_molecular_function ?t.
  ?g has_associated_phenotype ?b2.
  ?b2 has_textual_description ?d.
  FILTER (?d, "muscular dystrophy", "i") .
  FILTER (?d, "congenital", "i") .
}

From medinfo paper.
Adapted from: Olivier Bodenreider, presentation at HCLS Workshop, WWW07
Semantic Web in Action
Industry Examples
• Zemanta
• Twine
• Digger
• Calais – Reuters Thompson
• Powerset
• Talis
Emerging Research Areas
Fact Extraction and Schema Creation

Knowledge Extraction from Community-Generated Content
Search helps us find relevant pages/articles
But: It doesn’t answer questions.
• Fact Extraction is the first step towards answering questions.
• Famous new company that does fact extraction from Wikipedia is Powerset http://www.powerset.com
Fact Extraction From Community Content

• Problem: without a guiding schema, extracted predicates are just terms
• ➞ useful for humans, but not for machines
• Expert-created schemas are expensive and usually very restricted
Solution: Have a community-generated schema

- Wikipedia hierarchy for terms and concepts
Hierarchy Creation

Seed Query: "cognition"

Wikigraph-Based expansion

Knowledge Enabled Information and Services Science
Fact Extraction From Community Content

- Solution: Have a community-generated schema
- Wikipedia hierarchy for terms and concepts
  - See Automatic Domain Model creation
- Wikipedia Infoboxes for relationship types
Learn Patterns that indicate Relationships

Aboriginal Tent Embassy

The Aboriginal Tent Embassy is a controversial semi-permanent assemblage claiming to represent the political rights of Australian Aborigines. It is made of a large group of activists, signs, and tents that reside on the lawn of Old Parliament House in Canberra, the Australian capital. It is not considered an official embassy by the Australian government.

- in Sydney, New South Wales, Australia
- Sydney is the most populous city in Australia
- Canberra, the Australian capital city
- Canberra is the capital city of the Commonwealth of Australia
- Canberra, the Australian capital
Add relationships

- Geographical Location
  - Country
    - Cuba
    - USA
    - Australia
  - Settlement
    - City
    - Town
    - Village
  - has subclass
  - has instance
  - capital_city
    - Canberra
    - Sidney
Fact Extraction From Community Content

• The accumulation of many pattern occurrences give the necessary support

  – Canberra, Australia → minimal positive support

  – The Australian capital of Canberra → additional major support
Summary

- Create Domain models from seed queries or seed concepts

- Connect the concepts in the created domain models with valid relationships
  - Learn pertinent patterns for relationships
  - Find evidence for relationships in text
    - Wikipedia
    - WWW
Discovering Undiscovered Knowledge

Connecting the Dots
How are Harry Potter and Dan Brown related?
Motivation

• Undiscovered Public Knowledge [Swanson 89]
  – Hidden connections in text

• Our objective: build mechanisms to reveal these connections

• Our approach:
  o Populate existing ontology schemas via information extraction from text
  o Use the extracted information to
    o Support browsing
    o Text retrieval
    o Knowledge discovery
Discovering the Undiscovered Knowledge

- **Swanson's discoveries** – Associations between Migraine and Magnesium [Hearst99]
  - stress is associated with migraines
  - stress can lead to loss of magnesium
  - calcium channel blockers prevent some migraines
  - magnesium is a natural calcium channel blocker
  - spreading cortical depression (SCD) is implicated in some migraines
  - high levels of magnesium inhibit SCD
  - migraine patients have high platelet aggregability
  - magnesium can suppress platelet aggregability

- Data sets generated using these entities (marked red above) as boolean keyword queries against PubMed

- Bidirectional breadth-first search used to find paths in resulting RDF
Background Knowledge Used

- **UMLS** – A high level schema of the biomedical domain
  - 136 classes and 49 relationships
  - Synonyms of all relationship – using variant lookup (tools from NLM)
  - 49 relationship + their synonyms = ~350 mostly verbs

- **MeSH**
  - 22,000+ topics organized as a forest of 16 trees
  - Used to query PubMed

- **PubMed**
  - Over 16 million abstract
  - Abstracts annotated with one or more MeSH terms
An excessive endogenous or exogenous stimulation by estrogen induces adenomatous hyperplasia of the endometrium.

- Entities (MeSH terms) in sentences occur in modified forms
  - "adenomatous" modifies "hyperplasia"
  - "An excessive endogenous or exogenous stimulation" modifies "estrogen"
- Entities can also occur as composites of 2 or more other entities
  - "adenomatous hyperplasia" and "endometrium" occur as "adenomatous hyperplasia of the endometrium"
Endometrial Neoplasms

Influencing phenotypes

Neoplasms

PTEN

Epidermal Growth Factor Receptor (EGFR)

Cell

- Modulation of PTEN inactivation may also occur at the transcription level influencing the specific phenotypes seen in Cowden and Bannayan-Riley-Ruvalcaba syndromes.

- PTEN protein could inhibit cell invasion even in the presence of the constitutively active epidermal growth factor receptor (EGFR).

- EGF and hypoxia induce CXCR4 in non-small cell lung cancer, a process regulated by the PI3-kinase/PTEN/AKT/mTOR signaling pathway and activation of HIF-1alpha.
Utilizing Extracted Knowledge

Supporting browsing, querying and knowledge discovery

– Semantic Browser
– Query semi-structured representations
  • SPARQL
  • Hypothesis-Driven Retrieval
– Discovery complex connection patterns
  • Knowledge Discovery operators
Example - Evaluating Hypotheses

Keyword query: Migraine[MH] + Magnesium[MH]

Supporting Document sets retrieved
Example - Semantic Browser

Click to Launch
Web 2.0

Man Meets Machine
Putting the man back in Semantics

Semantic Web focuses on artificial agents

“Web 2.0 is made of people” (Ross Mayfield)

“Web 2.0 is about systems that harness collective intelligence.” (Tim O’Reilly)

The relationship web combines the skills of humans and machines
Semantic Web focuses on artificial agents

“Web 2.0 is made of people” (Ross Mayfield)

“Web 2.0 is about systems that harness collective intelligence.” (Tim O’Reilly)

The relationship web combines the skills of humans and machines
Going places …

- **Formal**
  - Artificial Intelligence
  - Personal Assistants
  - Knowledge Bases
  - Intelligent Agents
  - Semantic Webs
  - Knowledge Management

- **Social, Informal**
  - Search Engines
  - Content Portals
  - Database
  - Web Sites
  - P2P File-sharing
  - PIM’s

- **Implicit**
  - Taxonomies
  - Ontologies
  - Knowledge Bases
  - Enterprise Portals
  - Enterprise Minds
  - Group Minds
  - Lifelogs

- **Powerful**
  - The Global Brain
  - Smart Marketplaces
  - Decentralized Communities
  - Knowledge Networks
  - Semantic Weblogs
  - Wikis

- **Social, Informal**
  - Conferencing
  - IM

Permission to re-use with attribution to: Nova Spivack www.mindingtheplanet.net

Knowledge Enabled Information and Services Science
A Community’s Pulse

• Wealth of information available in blogs, social networks, chats etc.
• Free medium of self-expression makes mass opinions / interests available
• Polling for popular culture opinions is easier
• Social Production undeniably affects markets
• Results of analysis more effectively tailored to specific audience: geo-specific retail ads, demographic interests in music
Buzz on MySpace

Mining artist popularity from chatter on MySpace
- Lists close to listeners preferences
  vs.
- Bill Boards

<table>
<thead>
<tr>
<th></th>
<th>BB</th>
<th>User Comments: May 07</th>
<th>User Comments: Jun 07</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rihanna</td>
<td>Rihanna</td>
<td>Winehouse</td>
<td>Rihanna</td>
</tr>
<tr>
<td>Biffy Clyro</td>
<td>Twang</td>
<td>Maroon 5</td>
<td>Winehouse</td>
</tr>
<tr>
<td>Twang</td>
<td>McCartney</td>
<td>Biffy Clyro</td>
<td>Maroon 5</td>
</tr>
<tr>
<td>Maroon 5</td>
<td>Winehouse</td>
<td>Twang</td>
<td>McCartney</td>
</tr>
<tr>
<td>McCartney</td>
<td>Biffy Clyro</td>
<td>Rascal</td>
<td>Biffy Clyro</td>
</tr>
<tr>
<td>Winehouse</td>
<td>Twang</td>
<td>Rascal</td>
<td>Rascal</td>
</tr>
<tr>
<td>Rascal</td>
<td></td>
<td></td>
<td>Twang</td>
</tr>
</tbody>
</table>
The How

Metadata Extraction from Comments
- Artist, Track name in comments are common words
- “Keep your smile on Lil.” (Artist: Lilly Allen, Track: Smile)
- Necessitate a combination of linguistic, statistical, domain knowledge and domain specific rules to do well

Detecting and discarding Spam
- Accurate popularity estimates

Transliterating Slang
- I say: “Your music is wicked”
- What I really mean: “Your music is good”

Hypercube: Demographics' of users who post, non-spam positive and negative sentiment comment counts
- Lets one ask questions like “Who is the most popular artists among the 19 year olds in New York?”
Opportunities

- Casual Text more and more pervasive
- Extracting Semantic Metadata a whole different problem
  - What works for a news article, scientific literature does not work well for content that does not follow rules of edited text
- Need to systematically understand differences in these types of text in order to improve enablers like entity extraction
Event Web and the Semantic Sensor Web

Time, Space and Theme
NY's 'Halo 3' launch was no riot, but it was close

By Caroline McCarthy
Staff Writer, CNET News.com
Published: September 24, 2007, 11:35 PM PDT

reporter's notebook NEW YORK--Late Monday night, George Clooney waltzed into a midtown Manhattan hotel, with the camera flashes of the paparazzi following him into the lobby.

A block away at the Best Buy store on Fifth Avenue and 44th Street, those waiting for the launch of Microsoft's Halo 3 video game couldn't have cared less.
Events and STT Dimensions

E1: Soldier
- located_at
- lives_at
- assigned_to
- participates_in

E2: Soldier
- located_at
- lives_at

E3: Soldier

E4: Address
- located_at

E5: Battle
- occurred_at

E6: Address
- located_at

E7: Battle
- occurred_at
- participates_in

E8: Military_Unit
- assigned_to
- participates_in

Georeferenced Coordinate Space (Spatial Regions)

Residency
Battle Participation

Named Places | Spatial Occurrents | Dynamic Entities
How do we determine if the three images depict ...

• the **same time and same place**?
• the **same entity**?
• a **serious threat**?
Sensor Data Pyramid

- Raw Sensor (Phenomenological) Data
- Feature Metadata
- Entity Metadata
- Ontology Metadata

Expressiveness

Knowledge

Information

Data
What is Sensor Web Enablement?

- All sensors reporting position
- All connected to the Web
- All with metadata registered

- All readable remotely
- Some controllable remotely

http://www.opengeospatial.org/projects/groups/sensorweb
Semantic Sensor Web

Semantic annotation of SWE

```xml
<swe:component rdfa:about=""time_1"
  rdfa:instanceof=""time:Instant">
  <swe:Time rdfa:property=""xs:dateTime">
    2008-03-08T05:00:00
  </swe:Time>
</swe:component>

<swe:value name="satellite-data"
  rdfa:about=""Dayton"
  rdfa:instanceof=""geo:City">
  001100011001111 ... 
</swe:value>
```

Ontological Knowledge
space, time and theme
Conclusion
Take Home Points

Semantics - from documents, to entities, to relationships
  – Richer, meaningful representations offer more insight, powerful reasoning capabilities

Semantics and Web technologies for integration of information from disparate sources, often created for very different purposes with lesser human involvement

Semantic Web is highly interdisciplinary – uses IR, AI, KR, DB, DC, ...

Increasing mesh of Semantics, Services, People for better exploitation of resources (data, sensors, services, people)
Kno.e.sis Labs (3rd floor, Joshi)

- Semantic Sciences Lab (Dr Sheth)
- Bioinformatics Lab (Dr Raymer)
- Semantic Web Lab (Dr Sheth + Dr. S.Wang)
- Service Research Lab (Dr Sheth)
- Metadata and Languages Lab (Dr Prasad)
- Data Mining Lab (Dr Dong)

Joint Proposals With Each

Sensor Networking Bin Wang
Kno.e.sis Members – a subset

Meena
Casual text analysis

Ajith
Web 2.0, Services

Karthik
Web 2.0, Services

Cory
Semantic Sensor Web

Cartic
Relationship Extraction in biomedical text

Topher
Social content analysis

Delroy

Matt
Geo-Spatial informatics

Pablo
Relationship extraction, semantic browsing

Satya
Bio-Informatics, Provenance

Prateek

Knowledge Enabled Information and Services Science
Projects: http://knoesis.org/research/

Demos at: http://knoesis.wright.edu/library/demos/

Publications: http://knoesis.wright.edu/library

Rest: http://knoesis.org

Thanks to our key sponsors: National Science Foundation, National Institute of Health, AFRL and industry partners.