Kno.e.sis Center -- Labs

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)

Sensor Networking Bin Wang

Knowledge Enabled Information and Services Science
Kno.e.sis Members – a subset

Ajith
Web 2.0, Services

Karthik
Web 2.0, Services

Cory
Semantic Sensor Web

Meena
Casual text analysis

Topher
Social content analysis

Cartic
Relationship Extraction in biomedical text

Pablo
Relationship extraction, semantic browsing

Satya
Bio-Informatics, Provenance

Matt
Geo-Spatial informatics

Delroy

Prateek

Knowledge Enabled Information and Services Science
Evolution of the Web

1990s
Web of databases
- dynamically generated pages
- web query interfaces

2005
Web of pages
- text, manually created links
- extensive navigation

2010
Web of services
- data = service = data, mashups
- ubiquitous computing

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

Computing for Human Experience
- Web as an oracle / assistant / partner
  - “ask to the Web”
  - using semantics to leverage
    text + data + services + people
Relationship Web:
Trailblazing , Analytics and Computing for Human Experience
27th International Conference on Conceptual Modeling (ER 2008)
Barcelona, Oct 20-23 2008

Amit Sheth
Kno.e.sis Center, Wright State University, Dayton, OH

This talk also represents work of several members of Kno.e.sis team, esp. the Semantic Discovery and Semantic Sensor Data. http://knoesis.wright.edu

Thanks, K. Godam, C. Henson, M. Perry, C. Ramakrishnan, C. Thomas.
Also thanks to sponsors: NSF (SemDis and STT), NIH, AFRL, IBM, HP, Microsoft.
“(Human mind works) by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain.”

Dr. Vannevar Bush, As We May Think, 1945
“An object by itself is intensely uninteresting”.
– Grady Booch, Object Oriented Design with Applications, 1991

Keywords | Entities | Relationships, Events
---|---|---
Search (data) | Integration (information) | Analysis, Insight (knowledge)
Semantics and Relationships on a Web scale

Increasing depth and sophistication in dealing with semantics
– From searching to integration to analysis, insight, discovery and decision making
– Relationships
  • Heart of semantics
  • Allows to continue progress from syntax and structure to semantics

Relationship Web takes you away from “which document” could have data/information I need, to interconnecting Web of information embedded in the resources to give knowledge, insights and answers I seek.

Knowledge Enabled Information and Services Science

- Domain Models
- Structured text (biomedical literature)
- Informal Text (Social Network chatter)
- Multimedia Content and Web data
- Web Services

- Search
- Integration
- Analysis
- Discovery
- Question Answering

- Metadata Extraction
- Patterns / Inference / Reasoning
- Meta data /
- Semantic Annotations
- Relationship Web

- Metadata Extraction
- Structured text (biomedical literature)
- Informal Text (Social Network chatter)
- Multimedia Content and Web data
- Web Services

Knowledge Enabled Information and Services Science
Issues - Relationships

Understanding and **modeling** relationships

Identifying Relationship (**extraction**)  
Discovering and Exploring Relationships (**reasoning**)  
Hypothesizing and Validating Complex Relationships

**Using/exploiting Relationships**  
for Semantic Applications  
(in search, querying, analysis, insight, discovery)
Data, Information, and Insight

What: Which thing or which particular one
Who: What or which person or persons
Where: At or in what place
When: At what time
How: In what manner or way; by what means
Why: For what purpose, reason, or cause; with what intention, justification, or motive

© Ramesh Jain
EventWeb [Jain]
RelationshipWeb [Sheth]

A Web in which each node is an event or object, and connected to other nodes using

– Linguistic Relationships
– Referential Links: hypertext links to related information (HREF); links with metadata (MREF), links referring to model (model reference in SAWSDL, SA-REST)
– Structural Links: showing spatial and temporal relationships
– Causal Links: establishing causality
– Relational Links: giving similarity or any other relationship

Adapted/extended from Ramesh Jain
Supporting Relationships in the Real World

Objects and event represent or model real world
– More complex relationships
  • Living organisms: Saprotropism, Antagonism, Exploitation, Predation, Ammensalism or Antibiosis, Symbiosis
  • Relationships between humans: ...
Karthik Gomadam

Ph.D Student

is_advised_by

Professor

is_advised_by

Assistant Professor

Attended Google IO

Moscone Center, SFO May 28-29, 2008

Location

has_location

Research Paper

published_in

Journal

Conference

published_in

Research Paper

presents

Journal

published_in

Conference

has_location

Location

Kno.e.sis

Directs

Kno.e.sis

Event

Karthik Gomadam

is_advised_by

Amit Sheth

Ph.D Student

is_advised_by

Researcher

publishes

Research Paper

published_in

Journal

published_in

Conference

has_location

Location

Knowledge Enabled Information and Services Science
Implicit Relationships

– Statistical representation of interactions between entities, co-occurrence of terms in the same cluster, tag cloud...

Linking of one document to another via a hyperlink...

– Two documents’ belonging to categories that are siblings in a concept hierarchy.
Explicit Linguistic Relationships

“He beat Randy Johnson” converted into the triple

Dontrelle Willis $\rightarrow$ beat $\rightarrow$ Randy Johnson
Identifying & Representing Relationships

Formal Relationships

– Subsumption, partonomy, ....

Domain specific vs domain-independent relationships

– Example of domain independent relationships: time, space/location

Complex entity and relationships (example later)
Why is This a Hard Problem?

Are objects/entities equivalent/equal(same)?
How (well) are they related?

– Implicit vs explicit:
  • Statistical representation of interactions between entities, co-occurrence of terms in the same cluster, tag cloud...
  • formal/assertional vs social consensus based
  • powerful (beyond FOL): partial, probabilistic and fuzzy match

– Degrees of relatedness and relevance: semantic similarity, semantic proximity, semantic distance, ...
  • [differentiation, disjointedness]
  • related in a “context”
Why is This a Hard Problem?

Semantic ambiguity
When information (knowledge) is
- Incomplete
- inconsistent
- Approximate

- Even is-a link involves different notions: identify, unity, essense (Guarino and Wetley 2002)

Faceted Search and Semantic Analytics – Some Early Work where relationships played key role

InfoQuilt (1996-2000)
MREF (1996-1998)
OBSERVER (1996-2000)
InfoQuilt (1996-2000): Using metadata PatchQuilt and user models/ontologies to support queries and analytics over globally distributed heterogeneous media repositories.
<TITLE> A Scenic Sunset at Lake Tahoe </TITLE>
<p>
Lake Tahoe is a popular tourist spot and
<a href="http://www1.server.edu/lake_tahoe.txt">some interesting facts</a> are available here. The scenic beauty of Lake Tahoe can be viewed in this photograph:
<center>
<img src="http://www2.server.edu/lake_tahoe.img">
</center>

Traditionally, correlation is achieved by using physical links
Done manually by user publishing the HTML document
MREF (1996-1998)
Metadata Reference Link -- complementing HREF

Creating “logical web” through
Media independent metadata based on correlation
Metadata Reference Link (<A MREF ...>)

<A HREF="URL">Document Description</A>
physical link between document (components)

<A MREF KEYWORDS=<list-of-keywords>; THRESH=<real>>Document Description</A>

<A MREF ATTRIBUTES(<list-of-attribute-value-pairs>)>Document Description</A>
Correlation Based on Content-Based Metadata

Some interesting information on dams is available here. “information on dams” defined by MREF to keywords and metadata (may be used for a query)

water.gif (Data)

Metadata Storage

Content Dependent Metadata

height, width and size

Content based Metadata

Major component (RGB)

Blue

Knowledge Enabled Information and Services Science
Abstraction Layers

ONTOLOGY NAMESPACE

METADATA

DATA

MREF in RDF

METADATA

DATA

ONTOLOGY NAMESPACE
Model for Logical Correlation using Ontological Terms and Metadata

Framework for Representing MREFs

Serialization (one implementation choice)

Domain Specific Correlation

Potential locations for a future shopping mall identified by all regions having a population greater than 500 and area greater than 50 sq meters having an urban land cover and moderate relief <A MREF ATTRIBUTES(population < 500; area < 50 & region-type = ‘block’ & land-cover = ‘urban’ & relief = ‘moderate’)>can be viewed here</A>
Domain Specific Correlation

=> media-independent relationships between domain specific metadata: population, area, land cover, relief

=> correlation between image and structured data at a higher domain specific level as opposed to physical “link-chasing” in the WWW
Repositories and the Media Types

- Population
- Area
- Boundaries
- Land cover
- Relief

Regions (SQL) → Census DB → TIGER/Line DB

Image Features (IP routines) → Map DB

Knowledge Enabled Information and Services Science
Query Results

Geographical Information about the Region

County = Clarke
Block = 731

Census Information about the Region

Area = 34
Population = 73

Spatial Location of Region  Land Cover of Region

No more regions satisfy the given constraints

CLEANUP AND QUERY AGAIN
Complex Relationships

Some relationships may not be manually asserted, but according to statistical analyses of text, experimental data, etc.

– allow association of provenance data with classes, instances, relationship types and direct relationships or statements
Complex Relationships

Relationships (mappings) are not always simple mathematical / string transformations

Examples of complex relationships

- Associations / paths between classes
- Graph based / form fitting functions
- Probabilistic relational
A Simple Relationship?

Graph based / form fitting functions

Smoking

Cancer
Complex Relationships
Cause-Effects & Knowledge Discovery

VOLCANO
LOCATION
PYROCLASTIC FLOW
ASH RAIN
DESTROYS
WEATHER
COOLS TEMP
PLANT
DESTROYS
PEOPLE
KILLS

ENVIRON.
LOCATION
BUILDING
DESTROYS
Knowledge Discovery - Example

Earthquake Sources

Nuclear Test Sources

Nuclear Test May Cause Earthquakes

Is it really true?

Complex Relationship: How do you model this?
Complex Relationships – Several Challenges

– Probabilistic relations

Number of earthquakes with magnitude > 7 almost constant. So if at all, then nuclear tests only cause earthquakes with magnitude < 7
Entity, Relationship, Event Extraction and Semantic Annotation

From content with structure
- Web pages
- Deep Web

From well-formed text (edited for rules of grammar)
From informal or casual text
- social networking sites
From digital media
Extracting Semantic Metadata from Semistructured and Structured Sources

Semagix Freedom for building ontology-driven information system
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 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)
Semantic Annotation (Extraction + Enhancement)

Value-added Voquette Semantic Tagging

- Private companies
- Type of company
- Industry affiliation
- Sector
- Exchange
- Company Execs
- Competitors
Braves refuse to offer Galarraga arbitration

Posted: Thursday December 07, 2000 6:16 PM

ATLANTA (AP) -- The Braves refused to offer salary arbitration to Andres Galarraga on Thursday, apparently ending the first baseman's career in Atlanta.

Atlanta did offer arbitration to six of its former players who became free agents: pitchers Andy Ashby, Terry Mulholland, John Burkett and Scott Kamieniecki; first baseman Wally Joyner and outfielder Bobby Bonilla.

Ashby and Kamieniecki, both with two years remaining in their four-year contract. Free agents who were not offered arbitration and will become unrestricted were John Smoltz and Tony Caminiti.

Galarraga's contract expired at the end of the season and he wasn't offered arbitration, which expires on December 7.

After missing the 1999 season because of cancer, Galarraga missed 25 games. He finished with 87 RBIs, 22 doubles and a .274 average.

Free agents not offered arbitration by their former teams include Curt Schilling, Scott Garrelts and Dan Plesac. Free agents not offered arbitration by the Braves include Craig Biggio, John Cerutti and David Wells.

The Braves made an offer Wednesday morning, but Galarraga said it was too low. Galarraga is seeking a two-year contract.

Players offered arbitration have until Dec. 19 to accept or reject the offers and can negotiate with their former teams through Jan. 8.
Ontology-Directed Metadata Extraction (Semi-Structured Data)

Web Page

Enhanced Metadata Asset

Extraction Agent
Some real-world or commercial semantic applications
Taalee’s Semantic (Facted) Search (1999-2001): Highly customizable, precise and fresh A/V search

Uniform Metadata for Content from Multiple Sources, Can be sorted by any field

Knowledge Enabled Information and Services Science
What else can a context do?

Rocker looks for mercy from foreign arbitrator

John Rocker arrived in New York Monday, seeking to overturn his suspension for making racist remarks about 'foreigners' in New York. Rocker was heckled by New York fans in the lobby of the MLB building and reportedly yelled back at them. Commissioner Bud Selig was the first man to appear at the hearing, and afterwards discussed the hearings.

Produced by: FoxSports.com
Posted Date: 2000-02-09 12:00:00
League: Professional
Teams: Atlanta Braves
Players: John Rocker

Atlanta not in story, Taalee Knowledge experts added this.
Players name automatically extracted from story by Taalee

Baseball Stats from MajorLeagueBaseball.com
Rocker Stats
Braves Stats
Braves Schedule

Category specific sponsorship added by Taalee
What can a context do?  
(a commercial perspective)
Semantic Targeting

What else can a context do?
(a commercial perspective)

Semantic Enrichment

Knowledge Enabled Information and Services Science
VideoAnywhere and Taalee Semantic Querying and Browsing (1998-2001)

BLENDED BROWSING & QUERYING INTERFACE

ATTRIBUTE & KEYWORD QUERYING

uniform view of worldwide distributed assets of similar type

SEMANTIC BROWSING

Targeted e-shopping/e-commerce

Your Search Results
We found no matches for keywords: "Alien+Resurrection+". Below are results for: "alien resurrection". If you prefer, you may try another search.

Top matches for this search:
- Alien Resurrection (1997) -- Sigourney Weaver, Winona Ryder; VHS
- Alien Resurrection (1997) -- Sigourney Weaver, Winona Ryder; DVD
- Alien Resurrection (1997) -- Sigourney Weaver, Winona Ryder; VHS Widescreen

Full Results: 5 items are shown below.

Alien Resurrection -- VHS
Sigourney Weaver, Winona Ryder / Subtitled in Spanish
Our Price: $12.99 ~ You Save: $1.99 (13%)
Read more about this title...
Precisely targeted through the use of Structured Metadata and integration from multiple sources.
Keyword, Attribute and Content Based Access

Blended Semantic Browsing and Querying (Intelligence Analyst Workbench) 2001-2003

Search Terms: bin Laden, Osama

Precise Matches:
- Blair predicts Bin Laden killing
- Carmin bin Laden
- Karl Rove: ‘We'll get Bin Laden’
- Newslink: Bin Laden link
- Security: Bin Laden
- Watchdog: Bin Laden

Kuwait fears Bin Laden funding link
The BBC's Frank Gardner reports on alleged links between some Kuwaiti charities and Osama Bin Laden's al-Qaeda network.

Item Details:
- People: bin Laden, Osama
- Organization: Al Qaeda
- Location: Kuwait, Middle East
- Media: video
- Length: 01:14

View Complete Story

Click to play: 112K
Semantic Browsing/Directory (2001-...)

- Links to news on companies that compete against Commerce One

- Crucial news on Commerce One's competitors (Ariba) can be accessed easily and automatically
Semantic Directory

System recognizes ENTITY & CATEGORY

Relevant portion of the Directory is automatically presented.
Users can explore Semantically related Information.
Semantic Relationships

Looking at relationships for Tiger Woods as Players in Golf

Tiger Woods is associated in Golf with

**Location**
- Los Angeles, California*
- Columbus, Ohio*
- Augusta, Georgia*
- Ponte Vedra Beach, Florida*
- Orlando, Florida*

**Tournament**
- Byron Nelson Classic*
- Masters*
- GTE Byron Nelson Classic*
- Memorial Tournament*
- The Memorial*
- Pebble Beach National Pro-am*
- Buick Invitational*
- Los Angeles Open*
- Phoenix Open*
- AT&T Pebble Beach National Pro-am*
- The Nissan Open*
- The Players Championship*
- The Mci Classic*
- The Motorola Western Open*
- MCI Classic*
- Bay Hill Invitational*
- Nissan Open*
- National Car Rental Golf Classic*
- NEC World Series Of Golf*
- Players Championship*

**Course Name**
- Riviera Country Club*
- Augusta National*

**Players**
- Hal Sutton*
- Stuart Appleby*
- Davis Love III*
- Phil Mickelson*
- Fred Couples*
- Jack Nicklaus*
- Steve Lowery*
- Mike Weir*
- Notah Begay*
- Michael Campbell*
- Retief Goosen*
- David Duval*
- Lee Janzen*
- Vijay Singh*
- David Sutherland*
- Sergio Garcia*
- Matt Gogel*
- Greg Norman*
Knowledge Enabled Information and Services Science

Focused relevant content organized by topic (semantic categorization)

Related relevant content not explicitly asked for (semantic associations)

Automatic Content Aggregation from multiple content providers and feeds

Automatic 3rd party content integration

Competitive research inferred automatically

Ahmed Yaseer:
- Appears on Watchlist ‘FBI’
- Works for Company ‘WorldCom’
- Member of organization ‘Hamas’
Semi-structured Government Data

In-structure text, Semi-structured Data

Scores the entity based on the content and entity relationships

Example of Fraud prevention application used in financial services
demostration
How Background Knowledge helps Informal Content Analysis
A Community’s Pulse is often Informal

• 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
  • geo-specific retail ads, demographic interests in music
Background Knowledge Improves Content Analysis

- Metadata creation
  - Example – music comments
    - Spot Artist, Track names and associated sentiments
  - Example comment
    - “Keep your smile on Lil.”
- Smile here is a track from Artist Lilly Allen’s album
- Background knowledge from MusicBrainz taxonomy provides evidence
  - Annotate ‘smile’ as Track
  - ‘Lil’ as Lilly Allen
- Background knowledge from Urban Dictionary for understanding Slang
  - I say: “Your music is wicked”
  - What I really mean: “Your music is good”
Results - Pulse of a Music Community

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

<table>
<thead>
<tr>
<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>Rihanna</td>
</tr>
<tr>
<td>Biffy Clyro</td>
<td>Winehouse</td>
<td>Winehouse</td>
</tr>
<tr>
<td>Twang</td>
<td>Maroon 5</td>
<td>Maroon 5</td>
</tr>
<tr>
<td>Maroon 5</td>
<td>Mccartney</td>
<td>Mccartney</td>
</tr>
<tr>
<td>McCartney</td>
<td>Biffy Clyro</td>
<td>Biffy Clyro</td>
</tr>
<tr>
<td>Winehouse</td>
<td>Twang</td>
<td>Rascal</td>
</tr>
<tr>
<td>Rascal</td>
<td>Rascal</td>
<td>Twang</td>
</tr>
</tbody>
</table>
### Mass Spectrometry (MS) Data

<table>
<thead>
<tr>
<th>Parent Ion m/z</th>
<th>Fragment Ion m/z</th>
<th>Fragment Ion Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>7830.9570</td>
<td>194.9604</td>
<td>2</td>
</tr>
<tr>
<td>580.2985</td>
<td>0.3592</td>
<td></td>
</tr>
<tr>
<td>688.3214</td>
<td>0.2526</td>
<td></td>
</tr>
<tr>
<td>779.4759</td>
<td>38.4939</td>
<td></td>
</tr>
<tr>
<td>784.3607</td>
<td>21.7736</td>
<td></td>
</tr>
<tr>
<td>1543.7476</td>
<td>1.3822</td>
<td></td>
</tr>
<tr>
<td>1544.7595</td>
<td>2.9977</td>
<td></td>
</tr>
<tr>
<td>1562.8113</td>
<td>37.4790</td>
<td></td>
</tr>
<tr>
<td>1660.7776</td>
<td>476.5043</td>
<td></td>
</tr>
</tbody>
</table>

**ms/ms peaklist data**

**Parent ion charge**

**Parent ion abundance**

**Fragment ion m/z**

**Fragment ion abundance**
Semantic Sensor Web

Semantic annotation of SWE

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

<swe:value name="satellite-data"
  rdfa:about="Dayton"
  rdfa:instanceof="geo:City">
  0011000111001111 ...
</swe:value>
```

Ontological Knowledge

space, time and theme
<swe:component name="time">
  <swe:Time definition="urn:ogc:def:phenomenon:time" uom="urn:ogc:def:unit:date-time">
    <sa:swe rdfa:about="?time" rdfa:instanceof="time:Instant">
      <sa:sml rdfa:property="xs:date-time"/>
    </sa:swe>
  </swe:Time>
</swe:component>

<swe:component name="measured_air_temperature">
    <sa:swe rdfa:about="?measured_air_temperature" rdfa:instanceof="senso:TemperatureObservation">
      <sa:swe rdfa:property="weather:fahrenheit"/>
      <sa:swe rdfa:rel="senso:occurred_when" resource="?time"/>
      <sa:swe rdfa:rel="senso:observed_by" resource="senso:buckeye_sensor"/>
    </sa:swe>
  </swe:Quantity>
</swe:component>

<swe:value name="weather-data">
  2008-03-08T05:00:00,29.1
</swe:value>
Semantic Sensor ML – Adding Ontological Metadata

Mike Botts, "SensorML and Sensor Web Enablement," Earth System Science Center, UAB Huntsville
Semantic Temporal Query

Model-references from SML to OWL-Time ontology concepts provides the ability to perform semantic temporal queries

Supported semantic query operators include:

- **contains**: user-specified interval falls wholly within a sensor reading interval (also called *inside*)
- **within**: sensor reading interval falls wholly within the user-specified interval (inverse of *contains* or *inside*)
- **overlaps**: user-specified interval overlaps the sensor reading interval

Example SPARQL query defining the temporal operator ‘within’

```sparql
SELECT ?interval
WHERE {
  ?interval time-entry:endTime ?e .
  ?e time-entry:isBefore ?e_datetime .

  FILTER (
    xsd:dateTime("2005-11-10T01:00:00.00") < xsd:dateTime(?b_datetime) &&
    xsd:dateTime("2008-11-10T01:00:00.00") > xsd:dateTime(?e_datetime)
  ) .
} .
ORDER BY ASC(?b_datetime)
```
demo of Semantic Sensor Web
http://knoesis.wright.edu/research/semsci/application_domain/sem_sensor/
Relationship/Fact Extraction from Text

Knowledge Engineering approach

– Manually crafted rules
  • Over lexical items <person> works for <organization>
  • Over syntactic structures – parse trees

– GATE
Relationship/Fact Extraction from Text

Machine learning approaches
– Supervised
– Semi-supervised
– Unsupervised
Schema-Driven Extraction of Relationships from Biomedical Text

Method – Parse Sentences in PubMed

[1254239-1] An excessive endogenous or exogenous stimulation by estrogen induces adenomatous hyperplasia of the endometrium.

SS-Tagger (University of Tokyo)
SS-Parser (University of Tokyo)

(TOP (S (NP (NP (DT An) (JJ excessive) (ADJP (JJ endogenous) (CC or) (JJ exogenous) ) (NN stimulation) ) (PP (IN by) (NP (NN estrogen) ) ) ) (VP (VBZ induces) (NP (NP (JJ adenomatous) (NN hyperplasia) ) (PP (IN of) (NP (DT the) (NN 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”
An excessive endogenous or exogenous stimulation by estrogen induces adenomatous endometrial hyperplasia of the endometrium.
An excessive endogenous or exogenous stimulation

[Diagram showing relationships between entities such as estrogen, modified_entity1, modified_entity2, composite_entity1, endometrium, adenomatous hyperplasia, and hyperplasia, with annotations like hasModifier and hasPart.]
Blazing Semantic Trails in Biomedical Literature


“The physician, puzzled by her patient's reactions, strikes the trail established in studying an earlier similar case, and runs rapidly through analogous case histories, with side references to the classics for the pertinent anatomy and histology.

The chemist, struggling with the synthesis of an organic compound, has all the chemical literature before him in his laboratory, with trails following the analogies of compounds, and side trails to their physical and chemical behavior.”

**File Name:** 3416116  
**Abstract:** The rates of growth of 29 hepatic metastases from 15 patients with primary colorectal carcinoma were studied using serial computed tomography (CT). Eleven metastases were found by the surgeon at laparotomy (overt metastases); the remaining eighteen were not evident to the surgeon at laparotomy, but were detected by CT scan during the immediate postoperative period postoperative (occult metastases). An estimate of tumour volume doubling time was obtained from a semi-logarithmic plot of tumour cell number against time. The mean doubling time for the overt metastases (s.m.) compared with 86 the occult metastases. The rate of surgery was estimated growth curve assuming (age of the overt metastases s.m.) The corresponding age of the 2.3 +/- 0.4 years.

<table>
<thead>
<tr>
<th>Relations</th>
<th>Classes</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>affects</td>
<td>cell</td>
<td>animal</td>
</tr>
<tr>
<td>associated with</td>
<td>molecular dysfunction</td>
<td>disease or syndrome</td>
</tr>
<tr>
<td>complicates</td>
<td>experimental model of disease</td>
<td>avian leukaemia</td>
</tr>
<tr>
<td>degree of</td>
<td>mental or behavioral dysfunction</td>
<td>ehrlich tumor carcinoma</td>
</tr>
<tr>
<td>issue in</td>
<td>neoplastic process</td>
<td>experiment arthritis</td>
</tr>
<tr>
<td>occurs in</td>
<td>neoplastic process</td>
<td>experiment autoimmune encephalitis</td>
</tr>
<tr>
<td>precedes</td>
<td>pathologic function</td>
<td>experiment myasthenia gravis</td>
</tr>
<tr>
<td>produces</td>
<td>experiment autoimmune neuritis</td>
<td>experiment autoimmune disease</td>
</tr>
</tbody>
</table>

**Previous Traversals**
demo of Semantic Browser
http://knoesis.wright.edu/research/semweb/projects/textMining/iswc2008/
**PMID-15886201**

Identification of the substrates for BRCA1-dependent ubiquitination activity is important for understanding how mutation of BRCA1 is associated with loss of tumor suppression activity. The currently identified substrates include histone proteins, p53, Fas-associated protein D2, and centrosomal proteins including NPM1 and (gamma)-tubulin (24, 51–53). Among these, only the modification of (gamma)-tubulin by BRCA1/BARD1 has been shown to affect the biology of breast cells. It has been shown that failure to ubiquitinate (gamma)-tubulin results in centrosome amplification (24). The BRCA1/BARD1 proteins are known to regulate multiple processes in the cell, including transcription, DNA repair, and centrosome dynamics (5, 25–29). Although the ubiquitination of (gamma)-tubulin may in part explain the BRCA1-dependent regulation of centrosome dynamics, it was unclear whether the BRCA1-dependent ubiquitination activity also regulates the transcription and DNA repair function of BRCA1.

We had proposed that the BRCA1-dependent ubiquitination activity may function in DNA repair by modification of RNAPII that transcribes DNA near a lesion (44, 15). This proposed role for BRCA1 in transcription-coupled repair could be important following UV damage or double strand breaks. One prediction of this model was that BRCA1/BARD1 ubiquitination activity would be targeted to the elongating, hyperphosphorylated form of RNAPII. Actively transcribing RNAPII is phosphorylated on Ser-5 proximal to the promoter and on Ser-2 further downstream (23). Thus, the principal form of RNAPII that elongates through a gene is the Ser2* form, which we now show is not a substrate for BRCA1/BARD1. The model that BRCA1-dependent ubiquitination directly links transcription elongation to repair is thus not supported. Instead, we found that Ser-5 phosphorylation of RNAPII is a generalized response to UV irradiation, and BRCA1-dependent ubiquitination modifies the RNAPII. It has been observed that transcriptionally engaged RNAPII does become phosphorylated on Ser-5 by the action of extracellular signal-regulated kinases 1 and 2 (69). The data are most consistent with a model whereby DNA damage causes phosphorylation of a subpopulation of RNAPII, followed by ubiquitination by BRCA1/BARD1 and subsequent degradation at the proteasome.

In these experiments we found that overexpression of BRCA1 in cells could stimulate the damage-induced ubiquitination of RNAPII. When we inhibited BRCA1 expression by transfection of short interfering RNA specific for BRCA1, we did not observe a decrease in ubiquitination of RNAPII. We interpret these results to indicate that one or more other abumin ligase can execute this function. Several other factors have been implicated in the ubiquitination of RNAPII, including cyclins.

**PMID-10037099**

In the mouse two-stage skin carcinogenesis model, tumor promotion is a distinct, rate-limiting step that determines the formation of premalignant tumors. As discussed above, the role of tumor promoters in human cancer is more complex because human exposure tends to involve sporadic low doses of complex mixtures of carcinogens, co-carcinogens, and tumor-promoting agents. Nonetheless, studies of rodent tumor models of liver, bladder, colon, and breast cancers—and analyses of human tumor formation—suggest that processes analogous to tumor promotion by TPA on the mouse skin are a common feature of carcinogenesis (1). Thus, epigenetic changes in cell signaling, such as altered growth factor production and receptor expression, and elevated synthesis of inflammatory and mitogenic factors, such as cytokines and eicosanoids, are key targets for inhibiting tumor promotion.

**Tumor Progression**

As noted earlier, tumor progression involves the accumulation of additional genetic alterations in an initiated cell clone and generally gives a growth advantage to the progressing clone. In progression, a focal lesion consisting of a population of initiated and promoted cells ultimately becomes an invasive malignant tumor. One frequently observed genetic alteration that appears to contribute to malignant progression is mutation in the p53 [also known as TP53] tumor suppressor gene (62). The p53 gene product is a transcription factor that regulates the expression of a number of DNA-damage and cell cycle-regulatory genes and genes regulating apoptosis. By enhancing transcription of these critical genes, p53 regulates the cellular response to DNA damage (63). p53 also plays a role in maintaining genomic stability (64). Genomic instability, a hallmark of spontaneous malignant progression, is characterized by sequential chromosomal aberrations, such as duplications, deletions, and loss of
Semantic Trail

(b)
“Everything's connected, all along the line. Cause and effect. That's the beauty of it. Our job is to trace the connections and reveal them.”
Jack in Terry Gilliam's 1985 film - “Brazil”
How are Harry Potter and Dan Brown related?
Semantic Trails can be built over a Web of Semantic (Meta)Data extracted (manually, semi-automatically and automatically) and gleaned from

- Structured data (e.g., NCBI databases)
- Semi-structured data (e.g., XML based and semantic metadata standards for domain specific data representations and exchanges)
- Unstructured data (e.g., Pubmed and other biomedical literature)

and

- Various modalities (experimental data, medical images, etc.)
Discovering Complex Connection Patterns

Discovering informative subgraphs
Given a pair of end-points (entities) produce a subgraph with relationships connecting them such that the subgraph is small enough to be visualized and contains relevant “interesting” connections

We defined an interestingness measure based on the ontology schema

– In future biomedical domain the scientist will control this with the help of a browsable ontology

– Our interestingness measure takes into account
  • Specificity of the relationships and entity classes involved
  • Rarity of relationships etc.

Heuristics

Sports Schema
- League
- Team
- Athlete
- Coach

Business Schema
- Owner
- Company
- Trustee
- Product

Entertainment Schema
- Person
- Movie
- Director
- Actor

Legend
- rdf:type
- rdf:Property

RDF Schema
- Person
- City
- lives_in
- council_member_of

Instance Base
- $e_1$
- $e_2$
- $e_3$
- $e_4$
- $e_5$
- $e_6$
- $e_7$
- $e_8$

Knowledge Enabled Information and Services Science
Discovery Algorithm

- Bidirectional lock-step growth from S and T
- Choice of next node based on interestingness measure
- Stop when there are enough connections between the frontiers
- This is treated as the candidate graph
Model the Candidate graph as an electrical circuit

- S is the source and T the sink
- Edge weight derived from the ontology schema are treated as conductance values
- Using Ohm’s and Kirchoff’s laws we find maximum current flow paths through the candidate graph from S to T
- At each step adding this path to the output graph to be displayed we repeat this process till a certain number of predefined nodes is reached
Discovery Algorithm

Results
– Arnold Schwarzenegger, Edward Kennedy

Other related work
– Semantic Associations
Hypothesis Driven Retrieval Of Scientific Text

Knowledge Enabled Information and Services Science
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

Powerful mechanism to integrate content
– Describes the Real-World occurrences
– Can have video, images, text, audio all of the same event
– Search and Index based on events and STT relations
Events and STT dimensions

Many relationship types

**Spatial:**
- What events happened near this event?
- What entities/organizations are located nearby?

**Temporal:**
- What events happened before/after/during this event?

**Thematic:**
- What is happening?
- Who is involved?
Events and STT dimensions

Going further
Can we use
To answer
  – Why? / How?
Use integrated STT analysis to explore cause and effect
Example Scenario: Sensor Data Fusion and Analysis

High-level Sensor (S-H)

Low-level Sensor (S-L)

- How do we determine if A-H = A-L? (Same time? Same place?)
- How do we determine if E-H = E-L? (Same entity?)
- How do we determine if E-H or E-L constitutes a threat?
Data Pyramid

Sensor Data Pyramid

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

Expressiveness

Data

Information

Semantics/Understanding/Insight
Sensor Data Architecture

Knowledge
- Object-Event Relations
- Spatiotemporal Associations
- Provenance Pathways

Information
- Entity Metadata
- Feature Metadata

Data
- Raw Phenomenological Data

Analysis Processes
- Semantic Analysis
- Entity Detection
- Feature Extraction
- Fusion
- Collection

Annotation Processes
- SML-S
- SML-S
- SML-S
- SML-S
- O&M
- TML

Ontologies
- Object-Event Ontology
- Space-Time Ontology

Sensors (RF, EO, IR, HIS, acoustic)
Current Research Towards STT Relationship Analysis

Modeling Spatial and Temporal data using SW standards (RDF(S))\(^1\)

– Upper-level ontology integrating thematic and spatial dimensions
– Use Temporal RDF\(^3\) to encode temporal properties of relationships
– Demonstrate expressiveness with various query operators built upon thematic contexts

Current Research Towards STT Relationship Analysis

Graph Pattern queries over spatial and temporal RDF data²

– Extended ORDBMS to store and query spatial and temporal RDF
– User-defined functions for graph pattern queries involving spatial variables and spatial and temporal predicates
– Implementation of temporal RDFS inferencing

Upper-Level Ontology Modeling Theme and Space

Occurrent: Events – happen and then don’t exist
Continuant: Concrete and Abstract Entities – persist over time

- **Occurrent**
- **Continuant**
- **Named Place**: Those entities with static spatial behavior (e.g. building)
- **Dynamic Entity**: Those entities with dynamic spatial behavior (e.g. person)

### Spatial Relationships
- **occurred_at**: Links Spatial Occurents to their geographic locations
- **located_at**: Links Named Places to their geographic locations

### Spatial Region
- **Spatial Region**: Records exact spatial location (geometry objects, coordinate system info)

---

**Spatio-Temporal-Thematic Query Processing @ Kno.e.sis**

**Knowledge Enabled Information and Services Science**
Temporal RDF Graph: Platoon Membership

E1: Soldier
assigned_to [1, 10] → E2: Platoon
assigned_to [5, 15] → E4: Soldier
assigned_to [5, 15] → E5: Soldier

Time interval represents valid time of the relationship

E1 is assigned to E2 from time 1 to 10 and then assigned to E3 from time 11 to 20

Also need to handle inferencing:

ORDBMS Implementation: DB Structures

**Fig. 2.** Storage structures for RDF data. Existing tables of Oracle Semantic Data Store are shown on the right, and our additional tables for efficiently searching spatial and temporal data are shown on the left.

Unlike thematic relationships which are explicitly stated in the RDF graph, many spatial and temporal relationships (e.g., distance) are implicit and require additional computation.
Scenario (Biochemical Threat Detection): Analysts must examine soldiers’ symptoms to detect possible biochemical attack

Query specifies

```sql
select a from table (spatial_eval (‘(?a has_symptom ?b) (Chemical_X induces ?b) (?a fought_in ?c)’, ?c, ‘(?d member_of Enemy_Group_Y) (?d spotted_at ?e)’, ?e, ‘geo_distance(distance=2 units=mile)’));
```
Going places ...

Degree of social connectivity

Formal

Implicit

Powerful

Social, Informal

Permission to re-use with attribution to: Nova Spivack www.mindingtheplanet.net
looking into my crystal ball
“The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.

Machines that fit the human environment instead of forcing humans to enter theirs will make using a computer as refreshing as a walk in the woods.”

Mark Weiser, *The Computer for the 21st Century* (Ubicomp vision)
imagine
imagine when

farm helper

tags blight, corn
meets

Farm Helper

Knowledge Enabled Information and Services Science
with this

Latitude: 38° 57’36” N
Longitude: 95° 15’12” W
Date: 10-9-2007
Time: 1345h
that is sent to

Geocoder

Sensor Data Resource

Weather Resource

Structured Data Resource

Services Resource

Agri DB

Soil Survey

Lat-Long

Location

Date

Time

Farm Helper

Weather data

Soil Information

Pest Information

...
and
Six billion brains
imagination today
impacts our experience tomorrow
COMPUTING FOR HUMAN EXPERIENCE
Computing For Human Experience

Prof. Amit P. Sheth,
LexisNexis Eminent Scholar and
Director, kno.e.sis center
Wright State University
http://knoesis.org