Semantic Content Annotation and Applications:
Semantic Web Now and Challenges Ahead

Invited Speaker
2001 IRC International Conference on Internet Information Retrieval, Korea

Amit Sheth
Large Scale Distributed Information Systems (LSDIS) Lab
University Of Georgia; http://lsdis.cs.uga.edu


October 29, 2001

© Amit Sheth
“The Web of data (and connections) with meaning in the sense that a computer program can learn enough about what the data means to process it. . .

Imagine what computers can understand when there is a vast tangle of interconnected terms and data that can automatically be followed.”

(Tim Berners-Lee, Weaving the Web, 1999)
Semantics for the Web

Another important step

- It would be great if computer can understand what data means; and it would be even better if computers can understand human’s information request and provide relevant data.

- Complementary view of Semantic Web: The concept that Web-accessible content can be organized and utilized semantically, rather than though syntactic and structural methods.
Information System Research – march towards Semantics (a personal perspective)

**Semantics** (Ontology, Context, Relationships, KB)

**Generation III**
- Information brokering
- 1997...
- InfoQuilt
- ADEPT
- VideoAnywhere
- Semantic Web, some DL-II projects, Taalee Semantic Engine, Oingo

**Generation II**
- Mediators
- 1990s
- VisualHarness
- InfoHarness
- InfoSleuth, KMed, DL-I projects
- Infoscopes, HERMES, SIMS, Garlic, TSIMMIS, Harvest, RUFUS, ...

**Generation I**
- Federated DB/multidatabases
- 1980s
- Mermaid
- DDTS
- Multibase, MRDSM, ADDS, IISS, Omnibase, ...

**Metadata** (Domain model)

**Data** (Schema, “semantic data modeling”)
Semantics and Semantic Web

- “meaning or relationship of meanings, or relating to meaning” (Webster)
- is concerned with the relationship between the linguistic symbols and their meaning [Kromrey 1994]
- meaning and use of data (Information System)

Semantics has been studied well before the Web. It is worth looking into multidisciplinary investigation on Semantics.
Challenge

Semantics brings information closer to human thinking and decision making. WWW suddenly forces us to simultaneously deal with the complexity of modeling, reasoning and perceptions to support semantics, with the huge scale and heterogeneity of all imaginable kind needed to deal with WWW. Semantic Web promises to allow for programs/agents to automatically understands what data is about, and act upon this understanding *to better meet human’s needs*. What and how much of “semantic” activities can be automated?
Building Blocks (and Debates)

- **Ontologies**
  - basis for modeling domains and knowledge sharing; ontological commitment; concepts/shared vocabulary and relationships
  - Types of relationships
  - How do experts come to agreements
  - integration vs interoperation
  - Articulations (inter-ontological relationships)

- **Ontology Specification and Content/Data/Resource Annotation/Markup**
  - Languages: RDF, DAML+OIL,..
  - Automatic classification (relevant ontologies)
  - Semantic (ontology-driven, or domain-specific) metadata extraction
  - Resource modeling, Web services
Building Blocks (and Debates)

- **Inferencing**
  - Higher Order Logics, Full First Order Logic, Description Logics
  - Datalog and Logic Programming, Problem Solving Methods

- **Query Languages or Information Requests**

- **Semantic Applications**
  - Semantic -- Search, Browsing, Personalization ...
  - Semantic Processes: travel booking
  - Knowledge Discovery and Learning
Part 1:

Quick and Partial Overview of Commercial State of the Art
Metadata Creation and Semanticization

- Automatic Content Classification/Categorization
- Metadata Creation/Extraction: Types of metadata created
Most traditional Content Management

**Traditional Text Categorization**

**Customer Training Set**

**Statistical/AI Techniques**

**Place in a taxonomy**

**Routing/Distribution**

**Classification of Article 4715**

**Standard Metadata**

**Article Source:** iSyndicate

**Posted Date:** 11/20/2000

---

**FINANCIAL TIMES**

*World business newspaper.*

**France Telecom to control Equant**

David Owain in Paris and Andrew Ward in London

November 20, 2000

France Telecom is taking control of Equant of the Netherlands in a deal designed to create a world-leading data and IP services provider. The complex transaction will see the French company combine its Global One corporate data business with the Dutch-based network service provider.

In return, France Telecom will receive £3.6bn new Equant shares. The French operator will also acquire the 34 per cent Equant stake held by Societe Internationale de Telecommunications Aeronautiques Foundation.

This will be in exchange for 91.9m existing France Telecom shares, valued by the French company at about £3bn.

It will also invest £1bn in cash in the activities of the new entity in exchange for 21.4m new convertible preferred shares. These will be converted five years after completion into 13m new Equant ordinary shares.

The net result will give France Telecom 54.3 per cent of Equant's voting rights.

The deal yesterday received a mixed reception, with Equant shares closing ahead 1.72 per cent at 34.30 and France Telecom down 4.07 per cent at 103.70.

"France Telecom is in growth mode," said Robert Macatta, telecoms analyst at Schroder Salomon Smith Barney. "The market is currently..."
Voquette’s Categorization and Automatic Metadata Creation

Knowledge-base & Catalog-by-Example (AI/ML) Techniques

Article 4715 Metadata
Feed Source: iSyndicate
Posted Date: 11/20/2000
Company Name: France Telecom, Equant
Ticker Symbol: FTE, ENT
Exchange: NYSE
Topic: Company News

Classification of Article 4715

Syntactic metadata

Semantic metadata

Voquette Training Set & KB

Customer Training Set & KB

Automated Content Enrichment (ACE)

FTE
Company Analysis
Conference Calls
Earnings
Stock Analysis

ENT
Company Analysis
Conference Calls
Earnings
Stock Analysis

NYSE
Member Companies
Market News
IPOs

Precise Personalization/
Syndication/Filtering

Routing/Distribution

Map to another taxonomy

Voquette Training Set & KB

Customer Training Set & KB

Financial Times
World business newspaper.
France Telecom to control Equant

France Telecom is taking control of Equant of the Netherlands in a deal designed to create a world-leading data and IP services provider.

The deal, announced last week, will see the French company combine its Global One corporate data business with the Dutch-based network service provider.

In return, France Telecom will receive 60.9m new Equant shares. The plan will undergo a procedure in May 2002 and a French court has approved the proposed transaction.

This will be in exchange for 30.9m existing France Telecom shares, valued by the French company at about 1.5bn.

France Telecom will also receive 3.4% of new Equant shares, valued at 1.5bn.

The deal should also allow the French company to enter the European market for the first time.

The deal is expected to be completed by the 2002 second quarter.

France Telecom is in growth mode,” said Robert Macaia, telecoms analyst at Schneider Salomon Smith Barney. “The market is currently

Article Number
4715

Posted Date: 11/20/2000
Company Name: France Telecom, Equant
Ticker Symbol: FTE, ENT
Exchange: NYSE
Topic: Company News

Feed Source: iSyndicate

France Telecom is taking control of Equant of the Netherlands in a deal designed to create a world-leading data and IP services provider.

The deal, announced last week, will see the French company combine its Global One corporate data business with the Dutch-based network service provider.

In return, France Telecom will receive 60.9m new Equant shares. The plan will undergo a procedure in May 2002 and a French court has approved the proposed transaction.

This will be in exchange for 30.9m existing France Telecom shares, valued by the French company at about 1.5bn.

France Telecom will also receive 3.4% of new Equant shares, valued at 1.5bn.

The deal should also allow the French company to enter the European market for the first time.

The deal is expected to be completed by the 2002 second quarter.

France Telecom is in growth mode,” said Robert Macaia, telecoms analyst at Schneider Salomon Smith Barney. “The market is currently
Braves refuse to offer Galarraga arbitration

Posted: Thursday December 07, 2000 6:15 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 Greg Mathis.

Ashby accepted the offer, but Kamieniecki, Joyner, Mulholland, Ashby and Galarraga rejected it. Only Mathis is unsigned.

Galarraga, who is 36, was set to become a free agent for the second time after his contract expired at the end of the season. He had been offered only a one-year contract.

After missing the 1999 season because of cancer, Galarraga, who was 10th among American League first basemen with a .290 average last season, is back and healthy.

Free agents not offered arbitration by their former teams include five pitchers and seven position players.

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.
Information Extraction for Metadata Creation

Key challenge:
Create/extract as much (semantics) metadata automatically as possible
Voquette (Taalee) Extraction and Knowledgebase Enhancement
Metadata and Semantic Technology enabled Applications
Voquette’s Semantic Search

Context and Domain Specific Attributes

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

Delightful, relevant information, exceptional targeting opportunity
Creating a Web of related information

What can a context do?
Links to news on companies that compete against Commerce One

Crucial news on Commerce One’s competitors (Ariba) can be accessed easily and automatically
What else can a context do? (a commercial perspective)

Semantic Targeting

Semantic Enrichment

---

Semantic Enrichment

---

Semantic Targeting
Semantic/Interactive Targeting

Precisely targeted through the use of Structured Metadata and integration from multiple sources.
Semantic Content

- Content which does contain the words the user asked for
- Content which does not contain the words the user asked for, but is about what he asked for.
- Content the user did not think to ask for, but which he needs to know.

Extractor Agents
Value-added Metadata
Semantic Associations
Example of Value Added Metadata for a Web page

- Search for ‘Jamal Anderson’ in ‘Football’
- Click on first result for Jamal Anderson
- The original source HTML page contains no mention of Team name and League name. They are value-additions to the metadata to facilitate easier search.

View metadata. Note that Team name and League name are also included in the metadata.
Semantic Associations using Simple Relationships
supported by Voquette (Taalee) Semantic Engine

- Related Stock News
  - COMPANIES in Same or Related INDUSTRY

- Technology Products
  - Important to INDUSTRY or COMPANY

- Industry News

- Regulations
  - Impacting INDUSTRY or Filed By COMPANY

- Competition
  - COMPANIES in INDUSTRY with Competing PRODUCTS

- EPA

- SEC
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
Competitive research inferred automatically
Automatic 3rd party content integration
Voquette Semantic Application – Equity Dashboard

Voquette Equity Research Dashboard - Microsoft Internet Explorer

Automatic Content Aggregation from multiple content providers and feeds
Related relevant content not explicitly asked for (semantic associations)
Competitive research inferred automatically
Automatic 3rd party content integration

Voquette Semantic Application – Equity Dashboard
Knowledge Inferencing Workflow

- Syntax Metadata
- Semantic Metadata

Human-assisted inference

- Same entity
- led by
  - Islamic Jihad Group led by Dr. Ayman al-Zawahiri

- Ahmed Ibrahim al-Naggar, one of 12 handed over to Egypt from Albania last summer, wore the red clothing of a condemned man because he had been sentenced to death in an earlier trial for plotting an attack on Cairo's Khan al-Khalili market, a major tourist attraction.

- Most of the defendants were said to belong to the armed Islamic movement Al-Jihad, and some of them are known to be closely associated with Osama Bin Laden, the Saudi dissident accused by the United States of masterminding attacks on American embassies in east Africa last year.

- Eleven others, convicted of conspiring to overthrow the government, received life sentences with hard labour.

- Some of the men have known links with Osama Bin Laden.
Voquette Semantic Technology System
Architecture

CACS provides automatic classification (w.r.t. WorldModel)
from unstructured text and extracts contextually relevant metadata

WorldModel specifies enterprise’s normalized view of information (ontology)

Knowledgebase represents the real-world instantiation (entities and relationships) of the WorldModel

Distributed agents that automatically extract/mine knowledge from trusted sources

Toolkit to design and maintain the Knowledgebase

Knowledgebase represents the real-world instantiation (entities and relationships) of the WorldModel

Fast main-memory based query engine with APIs and XML output

Voquette Semantic Technology System
Architecture
Interesting Semantic Web companies

- Voquette/Taalee
- Applied Semantics
- Ontoprise
- Mondeca
- H5 Technologies

see business.semanticweb.org
Part 2:

Some Research Investigations
Semantic Web And Beyond...

Knowledge Discovery

Semantic Web

Simple Relationships
(is-a, instance-of, part-of, value-of, less than)

Information Interoperability/Integration

Complex, user defined, fuzzy Relationships
(cause, affect)
Complex Relationships

- Arise in several contexts, especially involving multiple ontologies (hence mappings)
  - information interoperability where related resources subscribe to different but related ontologies
  - information requestor and resource modelers choose to use different ontologies
  - information requests to support analysis, knowledge discovery, decision making, learning that requires linking multiple domains with different ontologies

Developing all encompassing, unified ontology is not shown to be practical. Preexisting classifications/metadata standards/taxonomies are hard to ignore.
Complex Relationships and Knowledge discovery

VOLCANO

LOCATION
PYROCLASTIC FLOW
ASH RAIN
BUILDING
WEATHER
PLANT
PEOPLE
LOCATION

ENVIRON.

DESTROYS
COOLS TEMP
DESTROYS
KILLS

AFFECTS
InfoQuilt System Core capabilities

- Ability to handle heterogeneous, static or dynamic content – wrappers & extractors, with resource modeling (completeness, data characteristics, binding patterns)
- Information Extraction: Semi-Automatically or Automatically create domain-specific or contextually relevant metadata
- Domain modeling with complex (user defined, inter-ontology) relationships, domain rules and FD
- User defined Functions (esp. for fuzzy/approximate matching) and Simulation
- Post processing result analysis (e.g., chart creator)
A nuclear test could have caused an earthquake if the earthquake occurred some time after the nuclear test was conducted and in a nearby region.

\[
\text{NuclearTest Causes Earthquake} \\
\leq \text{dateDifference}(\text{NuclearTest.eventDate, Earthquake.eventDate}) < 30 \\
\text{AND } \text{distance}(\text{NuclearTest.latitude, NuclearTest.longitude, Earthquake.latitude, Earthquake.longitude}) < 10000
\]
IScape (Information Scape)

- A computing paradigm that allows users to query and analyze the data available from a diverse autonomous sources, gain better understanding of the domains and their interactions as well as discover and study relationships.
IScape … a simple example

- user’s request
  - for semantically related information (regardless of all forms of heterogeneity)
  - specified in terms of components of knowledge base (domain model, relationships, functions, simulations)

“Find all earthquakes with epicenter less than 5000 mile from the location at latitude 60.790 North and longitude 97.570 East and find all tsunamis that they might have caused”

Next - KD using ISacpes
Knowledge Discovery - Example

Earthquake Sources

Nuclear Test Sources

Nuclear Test May Cause Earthquakes

Is it really true?

Complex Relationship:
How do you model this?
Knowledge Discovery - Example

When was the first recorded nuclear test conducted?

1950

Find the total number of earthquakes with a magnitude 5.8 or higher on the Richter scale per year starting from 1900

Increase in number of earthquakes since 1945
For each group of earthquakes with magnitudes in the ranges 5.8-6, 6-7, 7-8, 8-9, and >9 on the Richter scale per year starting from 1900, find number of earthquakes.

Number of earthquakes with magnitude > 7 almost constant. So nuclear tests probably only cause earthquakes with magnitude < 7.
Find nuclear tests and earthquakes that may have occurred as a result of the test
Ontologies

Terms/Concepts (Attributes)
- latitude
- longitude
- damage
- damagePhoto
- numberOfDeaths
- magnitude
- magnitude > 0
- magnitude < 10

Hierarchies
- Natural Disaster
- Man-made Disaster
- Volcano
- Earthquake
- NuclearTest

Functional Dependencies (FDs)
- site => latitude, longitude
- bodyWaveMagnitude
- conductedBy
- explosiveYield
- bodyWaveMagnitude > 0
- bodyWaveMagnitude < 10

Domain Rules
Knowledge Builder
IScape Builder

Specify all the ontologies and relationships to be used in IScape

Statement: Study the effect of nuclear tests on earthquakes
Description: List earthquakes caused just after a nuclear test and in a nearby location

Ontologies:
- Tsunami
- Earthquake
- NuclearTest
- Vegetation
- DailyWeather

Ontologies in IScape:
- Earthquake
- NuclearTest

Relationships:
- NuclearTestCausesEarth
- EarthquakeCausesTsunami
- EarthquakeAffectsEnvironment

Relationships in IScape:
- NuclearTestCausesEarth
- EarthquakeCausesTsunami
- EarthquakeAffectsEnvironment
IScape Execution

Knowledge Agent

User Agent

Planning Optimizing Agent

Broker

Correlation Agent

Resource Agent

Extractor / Wrapper

Extractor / Wrapper

Database Wrapper

Database Wrapper

Other Wrapper

Others

Knowledge

IScape

Final Results

IScape

Final Results

Plan

Plan

Data retrieved

Query

Query

IScape

Final Results
Planning and Optimization

- IScapes are specified in terms of ontologies
- Source selection
- Execution plans that are executable
- Plans that retrieve more complete information
- Integrate data from sources
- Optimization using domain and resource characteristics
“Find all nuclear tests conducted by India or Pakistan after January 1, 1995 with seismic body wave magnitude > 4.5 and find all earthquakes that could have been caused due to these tests.”
IScape Processing Monitor (IScape 1)

**Processing Log:**

<table>
<thead>
<tr>
<th>Message Id</th>
<th>Time Stamp</th>
<th>Message From</th>
<th>Brief Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>0282:01:14,484</td>
<td>1282:01:14,671</td>
<td>User Agent</td>
<td>Started processing</td>
</tr>
<tr>
<td>0282:01:15,046</td>
<td>1282:01:15,109</td>
<td>Broker Agent</td>
<td>Started processing</td>
</tr>
<tr>
<td>0282:01:15,109</td>
<td>1282:01:15,187</td>
<td>Planning Agent</td>
<td>Applied Domain Rules to IScape's constraint</td>
</tr>
<tr>
<td>0282:01:15,187</td>
<td>1282:01:15,234</td>
<td>Planning Agent</td>
<td>Checking IScape's constraint and relationship...</td>
</tr>
<tr>
<td>0282:01:15,234</td>
<td>1282:01:15,453</td>
<td>Planning Agent</td>
<td>Selecting sources for Earthquake</td>
</tr>
<tr>
<td>0282:01:15,453</td>
<td>1282:01:15,578</td>
<td>Planning Agent</td>
<td>Selecting sources for Nuclear Test</td>
</tr>
<tr>
<td>0282:01:15,578</td>
<td>1282:01:15,843</td>
<td>Planning Agent</td>
<td>Plan created by the planner, Returning it to Br...</td>
</tr>
<tr>
<td>0282:01:15,843</td>
<td>1282:01:16,390</td>
<td>Brooker Agent</td>
<td>Received plan from planner</td>
</tr>
<tr>
<td>0282:01:16,390</td>
<td>1282:01:16,500</td>
<td>Correlation Agent</td>
<td>Executing IScape</td>
</tr>
<tr>
<td>0282:01:16,500</td>
<td>1282:01:16,748</td>
<td>TestSitesDB Res. Agent</td>
<td>Queried TestSitesDB</td>
</tr>
<tr>
<td>0282:01:16,748</td>
<td>1282:01:16,843</td>
<td>SignificantEarthquakesDB Res. Agent</td>
<td>Queried SignificantEarthquakesDB</td>
</tr>
<tr>
<td>0282:01:16,843</td>
<td>1282:01:16,904</td>
<td>Correlation Agent</td>
<td>Computed Union</td>
</tr>
<tr>
<td>0282:01:16,904</td>
<td>1282:01:17,016</td>
<td>Correlation Agent</td>
<td>Computed Union</td>
</tr>
<tr>
<td>0282:01:17,016</td>
<td>1282:01:17,250</td>
<td>Correlation Agent</td>
<td>Computed Union</td>
</tr>
<tr>
<td>0282:01:17,250</td>
<td>1282:01:17,359</td>
<td>Correlation Agent</td>
<td>Computed Union</td>
</tr>
<tr>
<td>0282:01:17,359</td>
<td>1282:01:17,750</td>
<td>Correlation Agent</td>
<td>Computed Union</td>
</tr>
<tr>
<td>0282:01:17,750</td>
<td>1282:01:17,785</td>
<td>Correlation Agent</td>
<td>Computed Union</td>
</tr>
<tr>
<td>0282:01:17,785</td>
<td>1282:01:17,906</td>
<td>Brooker Agent</td>
<td>Computed Union</td>
</tr>
<tr>
<td>0282:01:17,906</td>
<td>1282:01:17,937</td>
<td>User Agent</td>
<td>Computed Union</td>
</tr>
</tbody>
</table>
Future

- SemWeb related standardization activities (DAML+OIL)
- More effective ontology creation and maintenance
- Ontology Interoperation
- Heuristic, Inexact/Fuzzy vs Formal/Classical Reasoning
- Query Languages/Information Request Specifications
- P2P - more dynamic/volatile/heterogeneous
Thank You!

Lab: http://lsdis.cs.uga.edu
Speaker: http://lsdis.cs.uga.edu/~amit

A course on Semantic Web: