Outline

1. Foundations of Semantic Web
2. Ontology
3. Introduction to RDF & OWL
4. Semantic Web Layer Cake
5. Semantic Search
6. Applications of Semantic Web
Outline

1. Foundations of Semantic Web
2. Ontology
3. Introduction to RDF & OWL
4. Semantic Web Layer Cake
5. Semantic Search
6. Applications of Semantic Web
Web of named relationships amongst named objects (Tim-Berners Lee).

Mustafa Nural is a Researcher working on the Semantic Web, which is a W3C Activity.
Foundations of Semantic Web

- Hypertext: a set of nodes and links.
Foundations of Semantic Web

- Not Machine Readable
  - There is very little machine-readable information there.
  - The meaning of the documents is clear to those with a grasp of (normally) English, and the significance of the links is only evident from the context around the link.
Foundations of Semantic Web

• Current Web
  • Current Web represents information using:
    • Natural language (e.g., English)
    • Graphics, multimedia
    • Page layout
  • Okay for human understanding
  • Difficult for machine processing
Foundations of Semantic Web

- **Analogy**
  - "Stay off the couch now, Ginger! You hear me? Ginger, stay off of the couch!"

- **What Dogs Understand**
  - "Blah blah blah blah blah GINGER blah blah blah GINGER blah blah blah blah blah"

- **Semantic Web Dog might understand:**
  - "Blah blah COUCH blah GINGER blah blah blah GINGER blah blah blah COUCH"

- **What Computers Understand**
  - “Blah blah blah blah blah blah blah <A HREF=...> blah blah blah . . ."
Foundations of Semantic Web

• Enabling machine processing
  • Two approaches:
    • Smarter machines
    • Smarter data
Foundations of Semantic Web

- Appr. 1: Smarter machines
  - Teach computers to understand the meaning of Web data
    - Natural language processing
    - Image recognition
    - Etc.
  - The Artificial Intelligence (AI) approach
Foundations of Semantic Web

- Smarter machines
  - Not the Semantic Web approach

```
Usefulness

"Smarter Machines"
(More Artificial Intelligence)

WARNING: HARD PROBLEMS!
```
Approach 2: Smarter data
- Make data easier for machines to understand
  - Express meaning in a machine processable format
  - Example: metadata
- The Semantic Web approach
Foundations of Semantic Web

- Smarter data
  - The Semantic Web approach
Foundations of Semantic Web

- The Current Web
  - Minimal machine-processable information -- dumb links
Foundations of Semantic Web

- The Semantic Web - An extension of the current Web
- More machine-processable information
Foundations of Semantic Web

- How Google works?
  - Links into page determine importance
  - "Importance" is cumulative

- Links are machine processable
- Links have (Minimal) semantics
- Amazing results from minimal semantics
Foundations of Semantic Web

- Why is machine processing difficult?
  - Identifying the key problems:
    - Ambiguity
    - Complexity of information formats
  - Solving the ambiguity problem
    - URIs
    - Ontologies
Problem 1: Ambiguity

- Mustafa Nural owns VIN #2775534.
  - Which “Mustafa Nural”?
  - Vehicle #2775534?
  - Vinyl siding order #2775534?

Need to identify things:
- Unambiguously, in a
- Uniform
- Web-friendly way
Foundations of Semantic Web

• Kinds of things to identify
  • Three kinds of things in the universe:
    • Web resources
    • Non-Web resources
      • Physical objects
        • Cars, people, houses, etc.
      • Abstract concepts
        • Sizes, colors, verbs, "love", etc.
        • "Creator" (e.g., the creator of a document)
        • "Location"
        • "Airline reservation"
        • "Airline reservation service"
Foundations of Semantic Web

- Unambiguously identifying Web resources
- Solution (trivial): URLs
  - http://www.example.org/index.html
Unambiguously identifying physical objects

Many human systems:
- Vehicle Identification Numbers (VIN)
- Product serial numbers
- UPC product codes
- Employee numbers
- Etc.

Problems:
- Too many formats
- Most are not global in scope

Solution: Convert to URIs
- http://www.example.com/employeeid/85740
Foundations of Semantic Web

- Unambiguously identifying abstract concepts
  - Solution: Use URIs
  - Problem: Which URIs?
    - Need to agree on common vocabulary
  - Solution: Ontology
Outline

1. Foundations of Semantic Web
2. **Ontology**
3. Introduction to RDF & OWL
4. Semantic Web Layer Cake
5. Semantic Search
6. Applications of Semantic Web
"Formal description of concepts and their relationships"

In other words:

**Vocabulary of terms**
- "book", "publication", "greyhound", "dog"

**And their relationships**
- "book is-a-kind-of publication"
- "greyhound is-a-kind-of dog"
Ontology

Vocabulary + Structure = Taxonomy

Taxonomy + Relationships, Constraints, Rules = Ontology

Ontology + Instances = Knowledge Base
Dublin Core

One well-known ontology

Defines 15 basic terms for documents and publishing:

- "title", "creator", "subject", "publisher"

Each term unambiguously identified by URI

- http://purl.org/dc/elements/1.1/creator
One global ontology?

- No. Not realistic.
  - Multiple ontologies will co-exist
  - Often specialized for problem domain

But:

- Can be merged later
- "Popularity contest"
Example of unambiguous identification

To say: "Web page foo.html was created by John Smith"

Need to unambiguously identify 3 things:

- Web page: http://www.example.org/foo.html
- "was created by": http://purl.org/dc/elements/1.1/creator
- "John Smith": http://www.example.org/staffid/85740
Ontology

- **Entities and relations**
  - Documents describe real objects and imaginary concepts, and give particular relationships between them.
    - A document might describe a person.
    - The title document to a house describes a house and also the ownership relation with a person.
    - A program could search for a house and negotiate transfer of ownership of the house to a new owner.
1. Foundations of Semantic Web
2. Ontology
3. Introduction to RDF & OWL
4. Semantic Web Layer Cake
5. Semantic Search
6. Applications of Semantic Web
Why is machine processing difficult? (cont’d)

Identifying the key problems:
- Ambiguity
  Solution: URI’s
- Complexity of information formats
Complexity of information formats

- Web pages use complex information formats:
  - English grammar, page layout, etc.
  - Easy for human to parse / understand
  - Hard for machine to parse / "understand"

- Example: "Time flies like an arrow"
  - How to parse?
  - Which is Subject? Verb? Object?

- Need a common, machine-processable information format
Important characteristics for a machine-processable format

- Scalable (the whole Web!)
- General
  - Allow any info to be expressed
- Extremely flexible:
  - Allow new data to be added
    - From any source
    - Without breaking existing data/systems
  - Allow any kind of query
  - Easily combine/join data in new ways
- Solution: RDF
Introduction to RDF & OWL

- Enabling standard: RDF
  - RDF: Resource Description Framework
    - Resources: things that can be named with URIs
    - Description: statements about the properties of these resources
  - RDF aims to build a Web of overlapping metadata vocabularies
  - Use URIs to define metadata vocabularies
  - Build graphs using these vocabularies to say things
Introduction to RDF & OWL

- RDF
  - W3C Recommendation
  - Language for making statements about things
  - Primarily for metadata
    - Author, title, subject, date-of-last-access
  - Can be used for any kind of statements
  - Has XML syntax: "RDF/XML"
Introduction to RDF & OWL

- **RDF Triples**
  - **All info expressed as triples:**
    - `<subject> <verb> <object>`
    - `<subject> <property> <value>`
Introduction to RDF & OWL

Example triple (Not RDF/XML syntax):

- http://www.example.org/foo.htm (Subject)
- http://purl.org/dc/elements/1.1/creator (Verb/Property)
- http://www.example.org/staffid/85740 (Object/Value)

Meaning:

"Web page foo.html was created by John Smith"
Introduction to RDF & OWL

Another example

```xml
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:love="http://love.example.org/terms/" >
    <rdf:Description rdf:about="http://aaronsw.com/">
        <love:reallyLikes rdf:resource="http://www.w3.org/People/Berners-Lee/Weaving/" />
    </rdf:Description>
</rdf:RDF>
```

Difficult to create by humans...
Introduction to RDF & OWL

• Joining triples to create a graph
  • Triples can be viewed as links in a graph
  • Equivalent of "joining" in relational database
  • Joining is automatic in RDF, because:
    • Nodes are URIs (unambiguous)
Introduction to RDF & OWL

Joining triples to create a graph

m:homePage

m:attending

p:GivenName

p:hasEmail
Introduction to RDF & OWL

- Joining data from multiple sources
  - Trivial: Same URI => same node.
  - How about extracted data?
Point vs. general solutions

- Any specific problem can be solved by a **point solution**
- *Many* conceptually similar problems, different in details
  - Approach doesn't scale well
  - N*N solutions required?
  - Inflexible: Point solutions don't facilitate **new** uses
- Conclusion: Need general solution
Introduction to RDF & OWL

Application Integration: XML Versus RDF

N*N complexity

N*1 complexity
Introduction to RDF & OWL

- What is missing from RDF?
- A Schema Support
  - Enables Reasoning
- Use RDF-S (RDF Schema)
  - Limited to Subclass Hierarchy
Web Ontology Language

- OWL is based on RDF and has three increasingly general levels: OWL Lite, OWL-DL, and OWL Full.

- OWL adds many new features to RDF:
  - Functional properties
  - Inverse functional properties (database keys)
  - Local domain and range constraints
  - General cardinality constraints
  - Inverse properties
  - Symmetric and transitive properties
Introduction to RDF & OWL

- Class Constructors
  - OWL classes can be constructed from other classes in a variety of ways:
    - Intersection (Boolean AND)
    - Union (Boolean OR)
    - Complement (Boolean NOT)
    - Restriction
  - Class construction is the basis for description logic.
Introduction to RDF & OWL

• Description Logic Example
  • Concepts are generally defined in terms of other concepts. For example:
    • The iridocorneal endothelial syndrome (ICE) is a disease characterized by corneal endothelium proliferation and migration, iris atrophy, corneal oedema and/or pigmentary iris nevi..
  • ICE-Syndrome class is the intersection of:
    • The set of all diseases
    • The set of things that have at least one of the four symptoms
Example of Description Logic
OWL Semantics

- An OWL ontology defines a theory of the world. States of the world that are consistent with the theory are called *interpretations* of the theory.
- A fact that is true in every model is said to be *entailed* by the theory. Logical inference in OWL is defined by entailment.
- Entailment can be counter-intuitive, especially when it entails that two resources are the same.
Introduction to RDF & OWL

OWL Semantics (cnt’d)

- OWL assumes an open world, while databases assume a closed world.
Open World vs. Closed World

- The advantage of the open world assumption is that it is more compatible with the web where one need not know all of the facts, and new facts are continually being added.
- The disadvantage is that operations (such as queries) are much more computationally complex.
- Another disadvantage is that one cannot have defaults or any inference based on the lack of information.
Closed World & Open World

- Statement: "Mary" "is a citizen of" "France"
- Question: Is Paul a citizen of France?
- "Closed world" (SQL or XML) answer: No.
- "Open world" (OWL) answer: unknown.
Introduction to RDF & OWL

- **Computational Complexity**
  - The various languages are progressively more complex.
  - Operations (such as queries) in XML and RDF require polynomial time in the worst case.
  - OWL-Lite operations are much more difficult, requiring exponential time in the worst case.
  - OWL-DL is even more difficult than OWL-Lite. One can only show that an operation can be completed in a finite amount of time.
  - OWL Full is the most difficult of all. It is undecidable.
In spite of these negative results, OWL is quite reasonable in practice.
The reason for this phenomenon is that the hard cases are not randomly distributed, but rather concentrated in a small region of the problem space.
Reasoners are used to infer information that is not explicitly contained within the ontology.

You may also hear them being referred to as Classifiers.

Standard reasoner services are:

- Consistency Checking
- Subsumption Checking
- Equivalence Checking
- Instantiation Checking

Reasoners can be used at runtime in applications as a querying mechanism (esp useful for smaller ontologies).
Outline

1. Foundations of Semantic Web
2. Ontology
3. Introduction to RDF & OWL
4. **Semantic Web Layer Cake**
5. Semantic Search
6. Applications of Semantic Web
Semantic Web future

- Self-desc. doc.
- Data
- RDF + rdfschema
- XML + NS + xmschema
- Unicode
- URI
- Ontology vocabulary
- Logic
- Proof
- Rules
- Trust
- Digital Signature
Logic and proofs

• Current semantic Web research
  • Good: systems can understand basic concepts (subclass, inverse etc.)
  • Better: if we could state any logical principles we wanted to.
    • Logical statements (rules) that allow the computer to make inferences and deductions.
Logic

- I am an employee of MemberCo.
- MemberCo is a member of W3C.
- MemberCo has GET access to http://www.w3.org/Member/.
- I (therefore) have access to http://www.w3.org/Member/.
Example (deduction)

- If someone sells more than 100 products then they are a member of Super Salesman club.
  - John sold 102 things; therefore John is a member of the Super Salesman club.
- More complex rules and inference engines explored.
Different people can write logic statements
Machines can follow semantic links to prove facts
Prove John is a Super Salesman
  • Sales: John sold 55 widgets + 47 sprockets
  • Widgets + sprockets: company products
  • $55 + 47 = 102$
  • $102 > 100$
  • Super Salesman rule
  • Proved: John is a Super Salesman
A Web of information processors (e.g. P2P)
Proof

- MemberCo's document employList lists me as an employee.
- W3C's member list includes MemberCo.
- The ACLs for http://www.w3.org/Member/ assert that employees of members have GET access.
Trust

- Useless if anyone can say whatever they want
- Digital signatures provide proof that a certain person wrote (or agrees with) a document or statement
- Digitally sign all RDF statements
- Tell programs whom to trust
Trust

- MemberCo's document employList is signed by a private key that W3C trusts to make such assertions.
- W3C's member list is trusted by the access control mechanism.
- The ACLs for http://www.w3.org/Member/ were set by an agent trusted by the access control mechanism.
Web of trust

- I trust my best friend Robert
- Robert trusts quite a number of people, and so on...
- Robert can trust Wendy a whole lot, but Sally only a little
Outline

1. Foundations of Semantic Web
2. Semantic Web Layer Cake
3. Introduction to RDF & OWL
4. **Semantic Search**
5. Applications of Semantic Web
Semantic Match vs. Key Word Match

SEARCH: Going rate for leasing a billboard near Triborough Bridge

Top hits from popular search engines miss the mark...

• Keywords may Match
• BUT WRONG Content returned
• And right content MISSED
Going rate for leasing a billboard near Triborough Bridge
Remarkably...With some Location semantics

We can quickly find Hi-Res examples of area of interest

But NOT the information we need
Simon and Garfunkel's "The 59th Street Bridge Song" was rated highly by the Billboard magazine in the 60's...

Analysis can detect semantic types to improve search precision and recall.
"...We were offered $250,000/year in 2001 for an outdoor sign in Hunts Point overlooking the Bruckner expressway. ..."

Analysis can detect semantic types to improve search precision and recall
Report Date 27 April, 2003. From a laptop computer captured in Bermuda it was learned that an American named John Odom, who worked for First Financial Inc., Bermuda from 1990-1992, travels using a Canadian passport in the name Phillip Barrage. On this same computer was found the name Phillip Gorman, who worked for First Financial Inc. from 1987-1993. Records on this computer reveal that Phillip Gorman entered the USA in March, 1993 and uses the alias Glen Bolduc.

- Analyzed by a collection of text analytics
- Detected Semantic Entities and Relations Highlighted
- Represented in UIMA Common Analysis Structure (CAS)
UIMA: Unstructured Information Management Architecture

- Open Software Architecture and Emerging Standard
  - Platform independent standard for interoperable text and multi-modal analytics
  - Under Development: UIMA Standards Technical Committee Initiated under OASIS

- Software Framework Implementation
  - SDK Available on IBM Alphaworks
  - Tools, Utilities, Runtime, Extensive Documentation
    - Creation, Integration, Discovery, Deployment of analytics
    - Java, C++, Perl, Python (others possible)
    - Supports co-located and service-oriented deployments (eg., SOAP)
    - x-Language High-Performances APIs to common data structure (CAS)
  - Embeddable on Systems Middleware (e.g., ActiveMQ, WebSphere, DB2)

- Apache UIMA open-source project
  - http://incubator.apache.org/uima/
We index and search over tokens AND the semantics annotations

“first” is an ambiguous

We are looking for the terms but with particular semantics detected by the UIMA analytics and indexed in the semantic search engine

The JuruXML Query Language Exploits the results of Analysis

KeyWord Query: “first”
Semantic Search Query:
<organization> first </organization> (i)
Semantic Search Query: <person> <ceo_of> <organization> first </organization> </ceo_of> </person>

We are looking for a person who is a ceo of an organization which has “first” in its name

“first” as it appears in the name of an organization

(i) XMLFragment Syntax
The Semantic Search 4-Step Program

1. UIMA Corpus-Analysis
   - Detect the Semantic Content in the Corpus
   - Build the Semantic Signatures

2. Semantic Search Index
   - Index the text and the Semantic Signatures

3. User Query
   - Automatically generate Semantic Search queries
   - For back-end

4. Semantic Search Engine (e.g., OmniFind)
   - SIAPI: Efficiently retrieve documents matching Semantic Signature
Concluding Remarks

- **Raising the Search Bar**
  - Known-Item Search must evolve into Knowledge Gathering and Synthesis
  - Semantic Search can improve precision and recall
  - Graceful degradation: Worst-Case should be keyword search

- **Semantics Analysis is Key**
  - Perfect, consistent or massive manual semantic annotation NOT likely
  - Automated annotation is essential
    - Many annotators must emerge
    - Must be easy to discover, combine, aggregate and deploy
    - UIMA is an enabling Integration Platform

- **Approximate Semantics**
  - Universal semantic consensus won’t happen
  - But approximations can work to better search applications
  - Improve precision, recall and density across artifact boundaries
Outline

1. Foundations of Semantic Web
2. Semantic Web Layer Cake
3. Introduction to RDF & OWL
4. Semantic Search
5. Applications of Semantic Web
Outline

1. Foundations of Semantic Web
2. Semantic Web Layer Cake
3. Introduction to RDF & OWL
4. Semantic Search
5. Applications of Semantic Web
Projects on Semantic Web

- **Dbpedia**: It is an effort to publish structured data extracted from Wikipedia: the data is published in RDF and made available on the Web. Thus allowing Semantic Web agents to have advanced querying over the Wikipedia-derived dataset and facilitating interlinking, re-use and extension in other data-sources.

- **FOAF**: A popular application of the semantic web is Friend of a Friend (or FoaF), which describes relationships among people and other agents in terms of RDF.
SIMILE: Semantic Interoperability of Metadata and Information in unlike Environments. SIMILE is a joint project, conducted by the MIT Libraries and MIT CSAIL, which seeks to enhance interoperability among digital assets, schemata/vocabularies/ontologies, metadata, and services.

Power Set:
1. Discover Facts: For most people, places and things, Powerset shows a summary of Factz from across Wikipedia.
2. Unlock Meaning: Powerset finds articles related to the meaning of your query. And sometimes even direct answers.
3. Scan Summaries: Powerset summarizes pages for easy browsing, with tools that follow you as you read and explore.
Projects (Cont’d)

- **Cognition**: Cognition’s unique Semantic Map is the most comprehensive and complete map of the English language available today. It can be used in support of the Semantic Web for semantic search, search tools, business analytics, machine translation, document search, context search, and much more. (http://www.cognition.com)
Semantic Web Ping Service

1. The Semantic Web Ping Service is a notification service for the semantic web that tracks the creation and modification of RDF based data sources on the Web.
2. It provides Web Services for loosely coupled monitoring of RDF data.
3. In addition, it provides a breakdown of RDF data sources tracked by vocabulary that includes: FOAF, RDF, and OWL.
Services (Cont)

- **Piggy Bank**
  1. Another freely downloadable tool is the Piggy Bank plug-in to Firefox.
  2. It works by extracting or translating web scripts into RDF information and storing this information on the user’s computer.
  3. Piggy Bank was developed by the Simile Project, which also provides RDFizers, tools that can be used to translate specific types of information, for example weather reports for US zip codes, into RDF. Efforts like these could ease a search between the web of today and its semantic successor.
Conclusion

- Will the Semantic Web initiative succeed? While many people believe in it (and in fact are investing in it), the outcome is still open. The success or failure will not be decided by scientific progress.
- In fact the functionality of the Semantic Web can be realized, to a certain degree, even with today's technology.
- The key question is whether the usefulness of the technology will be demonstrated quickly and powerfully enough to created momentum, perhaps recreating something similar to the World Wide Web.