Semantic and Data Mining Technologies

Simon See, Ph.D.,
Introduction to Semantic Web and Business Use Cases
Lots of Scientific Resources
NAR 2009 – over 1170 databases

Growth of GenBank
(1982 - 2004)
Paul writes workflows for identifying biological pathways implicated in resistance to Trypanosomiasis in cattle.

Paul meets Jo. Jo is investigating Whipworm in mouse.

Jo reuses one of Paul’s workflow without change.

Jo identifies the biological pathways involved in sex dependence in the mouse model, believed to be involved in the ability of mice to expel the parasite.

Previously a manual two year study by Jo had failed to do this.
Workflows

- Access to distributed and local resources
- Automation of data flow
- Iteration over data sets
- Interactive
- Agile software development
- Experimental protocols
- Declarative mashups?
The Data Playground

1) The Playground Panel with A) Data Objects that feed into BioMoby objects, B) BioMoby Services which can be executed through a right-click option and C) the ‘record’ button for recording user activity.

2) The Taverna tree of Web Services which can be dragged into the Playground Panel and connected to Data Objects.

3) Semantic Discovery panel that shows compatible services based on input or output.

4) The Data Editor panel for inputting Data.

5) The Data Viewer panel for viewing data that features the Taverna data rendering capabilities.

Andrew Gibson
Semantic Technology Stack

- **Basic Technologies**
  - **URI**
    - Uniform Resource Identifier
  - **RDF**
    - Resource Description Framework
  - **RDFS**
    - RDF Schema
  - **OWL**
    - Web ontology language
  - **SPARQL**
    - Protocol and Query Language
Extraction, Modeling, Reasoning & Discovery Workflow

Transform & Edit Tools
- Entity Extraction & Transform
- Ontology Engineering
- Categorization
- Custom Scripting

Load, Query & Inference
- RDF/OWL Data Management
- SQL & SPARQL Query
- Inferencing
- Semantic Rules
- Security
- Semantic Indexing
- Versioning

Applications & Analysis Tools
- BI Analytics
- Graph Visualization
- Social Network Analysis
- Metadata Registry
- Faceted Search
- SPARQL Endpoint

Partner Tools
- Partner/Oracle Tools

Data Sources
- Transaction Systems
- Unstructured Content
- RSS, email
- Other Data Formats

Partner/Oracle Tools
Why Organizations use Oracle RDF Database

- Oracle database is the leading commercial database with native RDF/OWL data management
- Scalable & secure platform for wide-range of semantic applications
- Readily scales to ultra-large repositories (10s billions of triples)
- Choice of SQL or SPARQL query
- Leverages Oracle Partitioning. RAC supported
- Growing ecosystem of 3rd party tools partners

**Key Capabilities:**

**Load / Storage**
- Native RDF graph data store
- Manages billions of triples
- Fast batch, bulk and incremental load

**Query**
- SQL: SEM_Match
- SPARQL: via Jena plug-in
- Ontology assisted query of RDBMS data

**Reasoning**
- Forward chaining model
- OWLprime, OWL 2 RL, RDFS
- User defined rule base
Architectural Overview

- **SPARQL Endpoints**: Joseki / Sesame
- **3rd Party Tools**: Topbraid Composer

**Program Interface**
- **Java API support**: Jena / Sesame
- **Java Programs**: JDBC

**Programming functionality**
- **LOAD**: Bulk-Load, Incr. DML
- **INFER**: OWL subsets, RDF/S, User-def.
- **QUERY (SQL-based SPARQL)**: Query RDF/OWL data and ontologies, Ontology-assisted Query of Enterprise Data

**Oracle DB**
- **RDF/OWL data and ontologies**: RDF/OWL, OWL, RDF/S, user-defined
- **Rulebases**: OWL, RDF/S, user-defined
- **Inferred RDF/OWL data**: Inferred RDF/OWL data

**Security**: fine-grained

**Versioning**: Workspaces

**Enterprise (Relational) data**

**Visualizer** (cytoscope)

**Tools**
- **Reasoners**: Pellet
- **NLP Info. Extractor**: Calais, GATE

**3rd Party Callouts**
- **Java API support**
- **SPARQL**: Jena / Sesame
- **Java Programs**: JDBC
- **SQL Interface**
  - SQLplus
  - PL/SQL
  - SQLd
- **SQL**
- **Tools (cytoscope)**
- **3rd Party Tools**: Topbraid Composer

**RDF/OWL data and ontologies**

**User-def.**

**Inferred RDF/OWL data**

**Ontology-assisted Query of Enterprise Data**

**Rulebases**: OWL, RDF/S, user-defined

**Semantic Indexes**

**Enterprise (Relational) data**
Core Entities in Oracle Database Semantic Store


• SDO_RDF_TRIPLE_S ➔ A new object type for RDF.

• Application Table ➔ Contains col of object type sdo_rdf_triple_s to allow loading and accessing RDF triples, and storing ancillary values.

• Model ➔ A model holds an RDF graph (set of S-P-O triples) and is associated with an sdo_rdf_triple_s column in an application table.

• Rulebase ➔ A rulebase is a set of rules used for inferencing.

• Entailments ➔ An entailment stores triples derived via inferencing.
Load / Query / Inference

- **Load ➔**
  - Bulk load
  - Incremental load

- **Query ➔**
  - SPARQL (direct or SQL-based)
  - simple data access

- **Inference ➔**
  - using OWL 2 RL, RDFS, etc.
  - using User-defined rules
Data Integration Platform in Health Informatics

Enterprise Information Consumers (EICs)
- Patient Care
- Workforce Management
- Business Intelligence
- Clinical Analytics

Model
- Virtual
- Physical

Relate

Design-Time Metadata

Access
- Integration Server (Semantic Knowledge base)

Deploy

Run-Time Metadata

Access

Integration Server (Semantic Knowledge base)
National Intelligence: Text Mining

1. Unstructured Data (Text)
   - Blogs, open source, newsfeed
   - Signal Intelligence, message traffic
   - Analyst Reports (Content Mgmt)

2. Model
   - Entity Extraction Engine: Feature/term/relation Extraction, categorization (Insight, Lymba, Calais, Gate)

3. Structured Data
   - Oracle’s founder Larry Ellison wins 2010 America’s Cup Race ...
   - Knowledge Base (RDF Store)
     - Ontologies + Rules
     - XML/OWL/N3
     - Triple Structure: Subj – Pred - Obj
     - 10’s of billions of triples

4. Mining & Discovery
   - SPARQL/SQL
   - Analyst Reports (Content Mgmt)
   - Explore
   - Browsing, Presentation, Reporting, Visualization, Query Tools (e.g. i2, Centrifuge, Visual Analytics)
<table>
<thead>
<tr>
<th>Problem</th>
<th>Algorithm</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>Multiple Regression (GLM) Support Vector Machine</td>
<td>Classical statistical technique Wide / narrow data / text</td>
</tr>
<tr>
<td>Anomaly Detection</td>
<td>One Class SVM</td>
<td>Lack examples</td>
</tr>
<tr>
<td>Attribute Importance</td>
<td>Minimum Description Length (MDL)</td>
<td>Attribute reduction Identify useful data Reduce data noise</td>
</tr>
<tr>
<td>Association Rules</td>
<td>Apriori</td>
<td>Market basket analysis Link analysis</td>
</tr>
<tr>
<td>Clustering</td>
<td>Hierarchical K-Means Hierarchical O-Cluster</td>
<td>Product grouping Text mining Gene and protein analysis</td>
</tr>
<tr>
<td>Feature Extraction</td>
<td>NMF</td>
<td>Text analysis Feature reduction</td>
</tr>
</tbody>
</table>
In-Database Data Mining

Traditional Analytics

- Data Import
- Data Mining Model “Scoring”
- Data Preparation and Transformation
- Data Mining Model Building
- Data Prep & Transformation
- Data Extraction

Oracle Data Mining

Results
- Faster time for “Data” to “Insights”
- Lower TCO—Eliminates
  - Data Movement
  - Data Duplication
- Maintains Security

Savings

Model “Scoring”
Data remains in the Database
Embedded data preparation

Cutting edge machine learning algorithms inside the SQL kernel of Database
SQL—Most powerful language for data preparation and
Data remains in the Database

Copyright 2010 Oracle Corporation
11g Statistics & SQL Analytics

- Ranking functions
  - rank, dense_rank, cume_dist, percent_rank, ntile

- Window Aggregate functions (moving and cumulative)
  - Avg, sum, min, max, count, variance, stddev, first_value, last_value

- LAG/LEAD functions
  - Direct inter-row reference using offsets

- Reporting Aggregate functions
  - Sum, avg, min, max, variance, stddev, count, ratio_to_report

- Statistical Aggregates
  - Correlation, linear regression family, covariance

- Linear regression
  - Fitting of an ordinary-least-squares regression line to a set of number pairs.
  - Frequently combined with the COVAR_POP, COVAR_SAMP, and CORR functions

- Descriptive Statistics
  - DBMS_STAT_FUNCS: summarizes numerical columns of a table and returns count, min, max, range, mean, median, stats_mode, variance, standard deviation, quantile values, +/- n sigma values, top/bottom 5 values

- Correlations
  - Pearson's correlation coefficients, Spearman's and Kendall's (both nonparametric).

- Cross Tabs
  - Enhanced with % statistics: chi squared, phi coefficient, Cramer's V, contingency coefficient, Cohen's kappa

- Hypothesis Testing
  - Student t-test, F-test, Binomial test, Wilcoxon Signed Ranks test, Chi-square, Mann Whitney test, Kolmogorov-Smirnov test, One-way ANOVA

- Distribution Fitting
  - Kolmogorov-Smirnov Test, Anderson-Darling Test, Chi-Squared Test, Normal, Uniform, Weibull, Exponential

Note: Statistics and SQL Analytics are included in Oracle Database Standard Edition

Copyright 2010 Oracle Corporation
ODM provides likelihood of expense reporting fraud ....and other important questions.
Oracle’s Partners for Semantic Technologies
Integrated Tools and Solution Providers:

<table>
<thead>
<tr>
<th>Ontology Engineering</th>
<th>Reasoners</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>TopQuadrant</td>
<td>Clarkparsia, LLC</td>
<td>Teranode</td>
</tr>
<tr>
<td>protege</td>
<td></td>
<td>Metatomix</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query Tool Interfaces</th>
<th>Standards</th>
<th>NLP Entity Extractors</th>
<th>SI / Consulting</th>
</tr>
</thead>
<tbody>
<tr>
<td>jena</td>
<td>openRDF.org</td>
<td>Calais Linguamatics</td>
<td>Northrop Grumman</td>
</tr>
<tr>
<td>Sesame</td>
<td>W3C</td>
<td>Gate</td>
<td>Raytheon</td>
</tr>
<tr>
<td>Joseki</td>
<td>RDF</td>
<td>General architecture</td>
<td>McDonal Bradley</td>
</tr>
</tbody>
</table>

SI / Consulting:
- Boeing
- Cognia
- Accenture
- Orbis Technologies, Inc.
Some Oracle Database Semantics Customers

- Life Sciences
  - Lilly
  - Pfizer
  - Swiss Institute of Bioinformatics

- Defense/Intelligence
  - National Geospatial-Intelligence Agency

- Clinical Medicine & Research
  - The University of Texas Health Science Center at Houston
  - Cleveland Clinic

- Education
  - The University of Michigan

- Telecomm & Networking
  - Hutchinson 3G Austria

- Publishing
  - Thomson Reuters

- Education
  - Stanford

- Telecomm & Networking
  - Cisco

- Publishing
  - Westlaw
Summary: Oracle Database 11g Release 2

Oracle database is the leading commercial database with native support for W3C standards compliant RDF/OWL data store w/ comprehensive capabilities for ....

Reasoning and Discovery supporting std. ontologies persistent, native & 3rd party inference, and user-defined rules

Scalability to evolve schemas dynamically and grow to 10’s billions of triples, incremental & parallel inference

Data Integration to link structured & unstructured content, Loosely couple business silos

Security to protect data on a “need to know” basis

Integrated querying & manageability SPARQL & SQL for RDF/OWL, relational, XML, text, location, & multimedia data
An Analytical Database Changes—Everything!

Less data movement = faster analytics, ... and faster analytics = better BI throughout enterprise
For More Information

search.oracle.com

Semantic Technologies

or

oracle.com
Additional Information

- Preview of the new Oracle Data Miner 11g R2 “work flow” New GUI
- Oracle Data Mining 11gR2 presentation at Oracle Open World 2009
- Oracle Data Mining Blog
- Funny YouTube video that features Oracle Data Mining
- Oracle Data Mining on the Amazon Cloud
- Oracle Data Mining 11gR2 data sheet
- Oracle Data Mining 11gR2 white paper
- New TechCast (audio and video recording): ODM overview and several demos
- Fraud and Anomaly Detection using Oracle Data Mining 11g presentation
- Algorithm technical summary with links to Documentation
- Getting Started w/ ODM page w/ instructions to download
  - Oracle Data Miner graphical user interface (GUI),
  - ODM Step-by-Step Tutorial
  - Demo datasets
- ODM Discussion Forum on OTN (great for posting questions/answers)
- ODM 11g Sample Code (examples of ODM SQL and Java APIs applied in several use cases; great for developers)
- Oracle’s 50+ SQL based statistical functions (t-test, ANOVA, Pearson’s, etc.)
ORACLE®