Semantic, Statistical, and Logical Reasoning with Massive Knowledgebases

Rangan Sukumar (Email: sukumarsr@ornl.gov)

ORNL Team: Matt Lee, Seung-Hwan Lim, Regina Ferrell, Larry Roberts, Arjun Shankar

Team of interns that have contributed to the work:

Tyler Brown¹(HERE), Keela Ainsworth¹(HERE), Larry Roberts²(SULI), Yifu Zhao³(ORSS), Nathaniel Bond⁴(SULI), Carlos del-Castillo-Negrette⁵(SULI), Katie Senter⁶(RAMS), Seokyong Hong⁷(Go!) and Gautam Ganesh⁸(HERE)

¹. University of Tennessee, Knoxville, 2. Tennessee Tech University, Cookeville, 3. Denison University, Ohio, 4. Cedarville University, 5. Yale University, 6. Penn State University, 7. North Carolina State University, University of Texas, Dallas

ORNL is managed by UT-Battelle
for the US Department of Energy
The Lifecycle of Data-Driven Discovery

- **Interrogation**
  - Querying and Retrieval
  - e.g. Google, Databases

- **Association**
  - Data-fusion
  - e.g. Mashups

- **Modeling, Simulation, & Validation**
  - Predictive Modeling
  - e.g. Climate Change Prediction

- **Better Data Collection**

**Domain Scientist’s View**

**Data Scientist’s View**

---

The Process of Data-Driven Discovery

- **Science of scalable predictive functions**
  - Pattern Discovery
  - e.g. Hypothesis generation
  - Pattern Recognition
  - e.g. Classification, Clustering

- **Science of data (Data-aware)**
  - e.g. Deep learning, Feature extraction, Meta-tagging

- **Data science (Infrastructure-aware)**
  - Shared-storage, shared-memory, shared-nothing

---

Data-Driven Discovery: Challenges at Scale

The Lifecycle of Data-Driven Discovery

Querying and Retrieval
- e.g. Google, Databases

Interrogation

Association
- e.g. Mashups

Better Data Collection

Predictive Modeling
- e.g. Climate Change Prediction

Modeling, Simulation, & Validation

What did we learn from “Big Data”?

- Knowledge discovery is a “greedy” and “never ending” thirst.
  - Big Data produces “Bigger Data”
  - “Lifecycle” management vs. “Project” management

- Big Data comes with Big Expectations
  - Data sets are expected to answer more than one question.
  - “We have lots of data – we do not know what questions to ask?”

- If Big Data => Smarter decisions, we need “Smarter Methods”
  - Discover “newer” insights in context with evolving new knowledge
  - Methods that can work well when there is more noise than signal.

Our Team: Graph Computing Solutions

Graph Computing…

- Supports discovery by interrogation, association and predictive modeling from structured and unstructured data

- Supports discovery with evolving knowledge and incremental domain hints

- Supports exploratory and confirmatory analysis
  - Data and meta-data integrated analytics
  - Flexible data structure seamless to growth while avoiding analytical artifacts

Patents:


Today’s Talk: “Eureka ! With Urika”

Objectives:

– Seamless semantic, logical and statistical reasoning with knowledgebases.

– Smart at finding what we know now, smarter at helping us find what we do not know yet.

504 compute cores, 5.4 TBs of distributed memory, and 576 TBs of local storage

64 Threadstorm processors, 2 TBs of shared memory connected to 125 TB of Lustre file storage
What is a Knowledgebase?

Dictionary Definition

1. a store of information or data that is available to draw on.

2. the underlying set of facts, assumptions, and rules that a computer system has available to solve a problem.

Computer Science Jargon

A knowledgebase is a technology used to store complex structured and unstructured information used by a computer system.
Example of a Knowledgebase: ConceptNet

MITs Open Common-sense Reasoning Framework

~3.4 M words, ~57 K types of relationships, ~10 million assertions
Examples of Knowledgebases: Semantic MEDLINE

Big Data from the fields of medicine, nursing, dentistry, veterinary medicine, health care systems, and preclinical science.

26,000 sources feed PubMed indices

**SEMANTIC MEDLINE: 70 million predications (133 node types and 69 edge types) from PubMed archive (more than 23.5 million citations, as of April 1st, 2014)**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acne</td>
<td>PROCESS_OF</td>
<td>Patients</td>
</tr>
<tr>
<td>Microbiode</td>
<td>TREATS</td>
<td>Patients</td>
</tr>
<tr>
<td>Acine_Vulgaris</td>
<td>PROCESS_OF</td>
<td>Patients</td>
</tr>
<tr>
<td>Biopsy_of_skin_procedure</td>
<td>USES</td>
<td>Isotretinoin</td>
</tr>
<tr>
<td>beta_defensin-2</td>
<td>ISA</td>
<td>Microbiode</td>
</tr>
<tr>
<td>beta_defensin-2</td>
<td>PART_OF</td>
<td>Human</td>
</tr>
<tr>
<td>Ribonuclease</td>
<td>AUGMENTS</td>
<td>Acine_Vulgaris</td>
</tr>
<tr>
<td>beta_defensin-2</td>
<td>PART_OF</td>
<td>Human</td>
</tr>
<tr>
<td>Acne</td>
<td>PROCESS_OF</td>
<td>Patients</td>
</tr>
<tr>
<td>doxycycline</td>
<td>TREATS</td>
<td>Patients</td>
</tr>
<tr>
<td>Isotretinoin</td>
<td>INTERACTS_WITH</td>
<td>Microbiode</td>
</tr>
<tr>
<td>doxycycline</td>
<td>TREATS</td>
<td>Acine</td>
</tr>
<tr>
<td>Microbiode</td>
<td>AFFECTS</td>
<td>Acine_Vulgaris</td>
</tr>
<tr>
<td>Microbiode</td>
<td>AUGMENTS</td>
<td>Acine</td>
</tr>
<tr>
<td>Microbiode</td>
<td>TREATS</td>
<td>Acine</td>
</tr>
<tr>
<td>Immunotherapy</td>
<td>TREATS</td>
<td>Malignant_Neoplasia</td>
</tr>
<tr>
<td>zolendronic_acid</td>
<td>AUGMENTS</td>
<td>Renal_Responses</td>
</tr>
<tr>
<td>zolendronic_acid</td>
<td>TREATS</td>
<td>Cancer_Patient</td>
</tr>
<tr>
<td>monocyte</td>
<td>LOCATION_OF</td>
<td>Phagocytes</td>
</tr>
<tr>
<td>Chemokine_C_C Motif</td>
<td>ISA</td>
<td>Chemokine_receptor</td>
</tr>
</tbody>
</table>

Made available by the National Library of Medicine (NLM)
What is Reasoning?

Semantic:

Are giraffes vegetarian?
Giraffes only eat leaves.
Leaves are parts of trees, which are plants.
Plants and parts of plants are disjoint from animals and parts of animals.
Vegetarians only eat things which are not animals or parts of animals.

Statistical

Is the vaccine working? How confident are you with that answer?

<table>
<thead>
<tr>
<th></th>
<th>Vaccine</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polio</td>
<td>82</td>
<td>162</td>
</tr>
<tr>
<td>No Polio</td>
<td>200,663</td>
<td>201,067</td>
</tr>
</tbody>
</table>

(Salk, 1954)

Logical

Inductive

The grass got wet numerous times when it rained, therefore: the grass always gets wet when it rains (Scientist)

Deductive

When it rains, things outside get wet. The grass is outside, therefore: when it rains, the grass gets wet (Mathematician, Philosopher)

Abductive

When it rains, the grass gets wet. The grass is wet. Therefore, it might have rained." (Detective, Doctor)
What does reasoning with a knowledgebase mean?

**Semantic:**

Infer logical consequences from a set of asserted facts or axioms.

**Statistical**

Involves connecting one concept to another while ensuring support/evidence in data is significant and did not happen by chance.

**Logical**

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inductive</td>
<td>Determines whether the truth of a <em>conclusion</em> can be determined for that <em>rule</em>, based solely on the truth of the premises</td>
</tr>
<tr>
<td>Deductive</td>
<td>Hypothesizes a <em>rule</em> after numerous examples are taken to be a <em>conclusion</em></td>
</tr>
<tr>
<td>Abductive</td>
<td>Given a true <em>conclusion</em> and a <em>rule</em>, it attempts to select some possible <em>premises</em> that, if true also, can support the <em>conclusion</em>, though not uniquely</td>
</tr>
</tbody>
</table>
Inspiring Motivation: Swanson’s Story from 1987


---

Today: There are 133,193 connections between migraine and magnesium.

SEMANTIC MEDLINE: 70 million predications (133 node types and 69 edge types)) from PubMed archive (more than 23.5 million citations, as of April 1st, 2014)

Given a knowledgebase and new clinical data/experiments, can we
(i) Find “novel” patterns of interest?
(ii) Rank and evaluate the patterns for significance?
Discovery using Knowledgebases: Open Problems

1. **Big Data**: Store, process, retrieve and reason with massive-scale datasets for newer discoveries.

2. **Signal-to-Noise Ratio**: Separate signal from noise when noise > signal.

3. **Hypothesis Generation**: Given knowledgebase + new data, formulate interesting ‘questions’.

4. **Significance Ranking**: Given different knowledge nuggets, find significant associations or predict future connections.

We need tools to enable discoveries to augment limited human bandwidth.
Leadership Computing @ ORNL takes care of the Big Data Problem

Store, process, retrieve and reason with massive-scale datasets for newer discoveries

<table>
<thead>
<tr>
<th>Leadership Computing</th>
<th>Titan</th>
<th>Apollo</th>
<th>CADES (Cloud)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery Approach</td>
<td>Modeling and Simulation</td>
<td>Association</td>
<td>Querying, Prediction</td>
</tr>
<tr>
<td>Architecture</td>
<td>Shared-compute</td>
<td>Shared-memory</td>
<td>Shared-storage</td>
</tr>
<tr>
<td>Scalability</td>
<td>Compute (# of cores)</td>
<td>Horizontal (# of datasets)</td>
<td>Vertical (# of rows)</td>
</tr>
<tr>
<td>Algebra</td>
<td>Linear</td>
<td>Relationship</td>
<td>Set-theoretic</td>
</tr>
<tr>
<td>Challenge (Pros)</td>
<td>Resolution</td>
<td>Heterogeneity</td>
<td>Cost</td>
</tr>
<tr>
<td>Challenge (Cons)</td>
<td>Dimensionality</td>
<td>Custom Solution</td>
<td>Flexibility</td>
</tr>
<tr>
<td>Leadership</td>
<td>#2 in the world (2013)</td>
<td>1 of 15 installs (2013)</td>
<td>--</td>
</tr>
<tr>
<td>User-interface</td>
<td>OpenMP, MPI, CUDA</td>
<td>SPARQL</td>
<td>SQL</td>
</tr>
</tbody>
</table>

Shared-memory architectures emerged as the best option for knowledge discovery heterogeneous graphs (pattern search and pattern mining):

- Compared to CMU Pegasus (2010) – ten times speed-up.
- Compared to Berkeley GraphX (2014) on a select few algorithms – 2 to 5 times speed-up.
- Compared to Desktops – 1000 times larger size for similar latency.
- First of its kind handling “heterogeneous-graphs” with near real-time latency.


S. M. Lee, S.R. Sukumar and S- H. Lim, “Graph mining meets the Semantic Web”, in the Proc. of the Workshop on Data Engineering meets the Semantic Web in conjunction with International Conference on Data Engineering, Korea, 2015.
The Eureka! Eureka! Project ‘App’ Store

Framework of Knowledge Discovery for a future beyond the Big Data Era

PLUS
Programmatic Python Login for Urika-like SPARQL Endpoints

FELT
Flexible, Extract, Transform and Load Toolkit

EAGLE-C
EAGLE ‘Is a’ algorithmic Graph Library for Exploratory Analysis

GRAPH-IC
Graph-Interaction Console

PAUSE
Predictive Analytics using SPARQL-Endpoints

KENODES
Knowledge Extraction using Network-Oriented Discovery Enabling System

Code Development
Graph Creation
Scalable Algorithms
Interactive Visualization
Reasoning + Inference
Hypothesis Creation

Some parts are open-source @ https://github.com/ssrangan/gm-sparql
Apps @ Work: Graph-theoretic Algorithms

• Some of the popular graph-theoretic algorithms implemented and tested so far
  • Summary metrics (~ 20 for both homogenous and heterogeneous graphs)
  • Degree (Diversity Degree)
  • Triangles (Count, Equilateral, Isosceles, Scalene)
  • N-gons
  • Shortest-path
  • PageRank (General, Personalized, BadRank, TrustRank)
  • Connected Components
  • Radius
  • Eccentricity
  • Degree-stratified clustering co-efficient
  • Peer-pressure clustering
  • Recommender systems
  • Label Propagation
Apps @ Work: Relevance Summarization

What is the underlying concept associated with words – (airplane, rainbow, star, thunder)?

Concept cloud for “rainbow”

Concept cloud for “aeroplane”

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>aerial_bomb</td>
<td>AtLocation</td>
<td>airplane</td>
</tr>
<tr>
<td>aerial_bomb</td>
<td>AtLocation</td>
<td>plane</td>
</tr>
<tr>
<td>aeroplane</td>
<td>AtLocation</td>
<td>at_airport</td>
</tr>
<tr>
<td>aeroplane</td>
<td>AtLocation</td>
<td>cloud</td>
</tr>
<tr>
<td>Aeroplane</td>
<td>CapableOf</td>
<td>also_be_spell_airplane</td>
</tr>
<tr>
<td>Rainbow</td>
<td>IsA</td>
<td>organisation</td>
</tr>
<tr>
<td>Rainbow</td>
<td>AtLocation</td>
<td>sky</td>
</tr>
<tr>
<td>Rainbow</td>
<td>PartOf</td>
<td>sky</td>
</tr>
<tr>
<td>Rainbow</td>
<td>IsA</td>
<td>music_group</td>
</tr>
<tr>
<td>Rainbow</td>
<td>IsA</td>
<td>band</td>
</tr>
<tr>
<td>rainbow</td>
<td>IsA</td>
<td>music_album</td>
</tr>
<tr>
<td>sky</td>
<td>AtLocation</td>
<td>rainbow</td>
</tr>
<tr>
<td>sky</td>
<td>HasA</td>
<td>airplane</td>
</tr>
<tr>
<td>sky</td>
<td>RelatedTo</td>
<td>airplane</td>
</tr>
<tr>
<td>air</td>
<td>AtLocation</td>
<td>rainbow</td>
</tr>
</tbody>
</table>
What is the underlying concept associated with words – (airplane, rainbow, star, thunder)?
Given the word “sky” what are the topics that come to mind?

Top 10 most relevant words

Blue
Cloud
Air
Rain
Thunder
Heaven
Sun
Fly

天空的定义：50,000种关系的集合

Apps @ Work: Semantic Coherence

WordNet Proximity

Conceptual Proximity

Phonetic and Phonemic Proximity

Quantifies relationship between two or more words with respect to a corpus

Extracting Novel and Useful Associations

Where are the new connections?
What are the “important” new connections?

New longer paths

New triangles and polygons
Finding all paths between two nodes is straightforward mathematically, but millions of paths exist between any two terms and not all paths are meaningful!

How do we only see useful ones? How do we rank paths for significance and saliency?
Reasoning Apps @ Work: Information Foraging

Approach #1: Rank nodes based on heuristics and then explore knowledgebase using random walks
Reasoning Apps @ Work: Information Foraging

Approach #2: Context-aware exploration

(1) Anchor nodes approach

Start Node: Smoking

Context Node: mutation

End Node: Lung Cancer

(2) k-hops in context approach

Start Node: Smoking

Multiple context nodes

End Node: Lung Cancer

Smoking_cigarette, smoking_tabacco, smoking_habit,

(3) Context similarity

Start Node: Smoking

End Node: Lung Cancer

Lung cancer stage I, stage I lung cancer, ...
## Apps @ Work: Back to Migraine and Magnesium

### Results: Eureka! Eureka!

<table>
<thead>
<tr>
<th>Magnesium</th>
<th>Rev_INHIBITS</th>
<th>Tantalum</th>
<th>AUGMENTS</th>
<th>Osseointegration</th>
<th>NEG_COEXISTS_WITH</th>
<th>Bone_Regeneratio</th>
<th>Rev_COEXISTS_WITH</th>
<th>Platelet_function</th>
<th>Rev_NEG_MANIFESTATION_OF</th>
<th>Migraine_Disorders</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnesium</td>
<td>Rev_STIMULATES</td>
<td>BW35</td>
<td>NEG_ASSOCIATION_WITH</td>
<td>Renal_Insufficiency</td>
<td>Rev_PRECEDES</td>
<td>Cervix_carcinoma</td>
<td>Rev_NEG_PREDISPOSES</td>
<td>Combined_Oral_Contraceptives</td>
<td>NEG_COMPICATI</td>
<td>Migraine_Disorders</td>
<td>0.355147</td>
</tr>
<tr>
<td>Magnesium</td>
<td>NEG_COMPICATE</td>
<td>Malaria</td>
<td>ASSOCIATED_WITH</td>
<td>home_binding</td>
<td>Rev_NEG_AUGMENTS</td>
<td>insuline_receptor_related_receptor_INS</td>
<td>RR</td>
<td>Melatonin</td>
<td>NEG_DISRUPTS</td>
<td>Migraine_Disorders</td>
<td>0.351549</td>
</tr>
<tr>
<td>Magnesium</td>
<td>Rev_NEG_USES</td>
<td>kidney_preservation</td>
<td>DIAGNOSES</td>
<td>Unilateralagenesis</td>
<td>of_kidney</td>
<td>Rev_COMPICATE</td>
<td>Congenital_obstruktion</td>
<td>CAUSES</td>
<td>Varicosity</td>
<td>NEG_COMPICATI</td>
<td>Migraine_Disorders</td>
</tr>
<tr>
<td>Magnesium</td>
<td>Rev_NEG_USES</td>
<td>kidney_preservation</td>
<td>TREATS</td>
<td>Hypertension</td>
<td>Re</td>
<td>novascular</td>
<td>ASSOCIATED_WIT</td>
<td>H</td>
<td>noradrenergic</td>
<td>Rev_ASSOCIATED</td>
<td>Varicosity</td>
</tr>
<tr>
<td>Magnesium</td>
<td>Rev_STIMULATES</td>
<td>heiacidtin</td>
<td>Rev_USES</td>
<td>Percutaneous</td>
<td>insertion</td>
<td>of_nephrostomy</td>
<td>tube</td>
<td>AFFECTS</td>
<td>Cervix_carcinoma</td>
<td>Rev_NEG_PREDISPOSES</td>
<td>Combined_Oral_Contraceptives</td>
</tr>
<tr>
<td>Magnesium</td>
<td>NEG_PREVENTS</td>
<td>Brain_Injuries</td>
<td>NEG_AFFECTS</td>
<td>Blood_Flow_Velocity</td>
<td>Rev_CAUSES</td>
<td>Dihydroergotamine</td>
<td>same_as</td>
<td>Triptans</td>
<td>DISRUPTS</td>
<td>Migraine_Disorders</td>
<td>0.288334</td>
</tr>
<tr>
<td>Magnesium</td>
<td>NEG_PREVENTS</td>
<td>Brain_Injuries</td>
<td>NEG_AFFECTS</td>
<td>Blood_Flow_Velocity</td>
<td>Rev_CAUSES</td>
<td>Dihydroergotamine</td>
<td>same_as</td>
<td>Triptans</td>
<td>AUGMENTS</td>
<td>Migraine_Disorders</td>
<td>0.288062</td>
</tr>
<tr>
<td>Magnesium</td>
<td>NEG_CAUSES</td>
<td>Hypomagnesemia</td>
<td>Rev_NEG_COEXIST</td>
<td>Rinal_Osteodystrophy</td>
<td>Rev_COMPICATE</td>
<td>Repair_of_bladder</td>
<td>CAUSES</td>
<td>Interstitial_Cystitis</td>
<td>Rev_COMPICATE</td>
<td>Migraine_Disorders</td>
<td>0.280192</td>
</tr>
<tr>
<td>Magnesium</td>
<td>NEG_COMPICATE</td>
<td>Malaria</td>
<td>Rev_TREATS</td>
<td>Heparitin_Sulfate</td>
<td>NEG_PART_OF</td>
<td>Herpesvirus_1</td>
<td>Interacts</td>
<td>Varicosity</td>
<td>NEG_COMPICATI</td>
<td>Migraine_Disorders</td>
<td>0.276062</td>
</tr>
</tbody>
</table>
Apps @ Work: Back to Migraine and Magnesium

Eureka ! Eureka !
Summary

1. **Big Data**: Store, process, retrieve and reason with massive-scale datasets for newer discoveries.

2. **Signal-to-Noise Ratio**: Separate signal from noise when noise > signal

3. **Hypothesis Generation**: Given knowledgebase + new data, formulate interesting ‘questions’.

4. **Significance Ranking**: Given different knowledge nuggets, find significant associations or predict future connections.

Preliminary results are very promising!! Still more work to do!
Thank you

• How are we doing on time?
  – Demo on a “Ebola Use Case”?