Ontology Summit 2014
Track D: Tackling the Variety Problem in Big Data – Summary

Ken Baclawski

Anne Thessen

Track D Co-Champions

The Potential of Big Data

- Could address important social and commercial needs.
- Traditional techniques are inadequate
  - Cost, feasibility, ethical concerns
- New techniques are much better, but
  - Still have ethical issues
  - Privacy has become a prominent issue
  - The variety problem is typically handled with ad hoc techniques

Emergence of Big Data

- Recognition of the problem in early 1980s: originally called the “information onslaught” [1]
- Different fields experience the problem at different times.
- The transition from information scarcity to information abundance in a field appears to occur abruptly.
- Organizations have difficulty coping with the transition.
- Many organizations are still based on the information scarcity model.

Characteristics of Big Data

- Large amounts of data (volume)
- Rapid pace of data acquisition (velocity)
- Complexity of the data (variety)
  - Complexity of individual data items
  - Both structured and unstructured data
  - Complexity of data from one source
  - Multiple sources

Technology for Big Data

- Development of new storage and indexing strategies for handling volume and velocity
  - "Map Reduce" was developed in 1994. [2]
- Development of techniques for handling variety
  - Schema mapping
  - Controlled vocabularies
  - Knowledge representations
  - Ontologies and semantic technologies
- Connection between these two?
  - Surprisingly little collaboration and communication.
  - A notable exception is the early work starting in 1992 on representing biological research papers. [3]

Session Speakers

- Eric Chan - **Enabling OODA Loop with Information Technology**
- Nathan Wilson - **The Semantic Underpinnings of EOL TraitBank**
- Ruth Duerr - **Semantics and the SSIII Project**
- Mark Fox - **Variety in Big Data: A Cities Perspective**
- Malcolm Chisolm - **Data Governance to Manage Variety in Big Data**
- Dan Brickley - **Schema.org, FOAF and Linked Data: Lessons for Web-scale vocabulary deployment**
- Rosario Uceda-Sosa - **Big Data, Open Data and the Smart City**

Speakers

- Eric Chan (Oracle)
  - Tool development addressing many aspects of the variety problem including provenance, metadata, and hypothesis generation and workflow (OODA loop)
- Dan Brickley (Google)
  - Examines the Variety problem on the Web and provides an overview of some lessons learned from schema.org, FOAF and Linked Data deployment of RDF-based vocabularies.

Speakers

- Malcolm Chisolm (AskGet.com)
  - Data governance is “a collection of disciplines that ensure data is managed adequately in an enterprise.” Malcolm will discuss what is involved in data governance and its implications for managing Variety in Big Data.
- Ruth Duerr (National Snow and Ice Data Center)
  - Experiences with using ontologies to address very heterogeneous sources of data for snow and ice, including satellite data, model output, point observations, social science data (interviews with Alaskan community members), and many other kinds of data.

Speakers

- Nathan Wilson (Encyclopedia of Life)
  - Experiences with using semantic technology in the tree of life, a notoriously messy ontology
- Mark Fox (University of Toronto)
  - More than half the world population live in cities and the proportion is growing, so cities are an enormous source of data. However, it is not just the amount of data that is daunting, but the enormous variety not only within a single city but also among the thousands of different cities.
- Rosario Uceda-Sosa (IBM)
  - Continues the examination of the Variety problem for city data and proposes some solutions.

Why is semantics important?

- Misunderstanding the data can result in invalid or misrepresented analyses.
  - A recognized problem well before Big Data.
  - John P. A. Ioannidis claimed that most published scientific research findings are false. [4] [5]
    - Many forms of bias can invalidate results
    - Bias and observer effects are difficult to recognize
    - Standard experimental procedures do not eliminate bias.
- Could Big Data exacerbate this problem?

What can ontologies do to help?

- Background knowledge of the domain
- The structure of the data
- Annotation of data and metadata
- Provenance of the data
  - Transformations
  - Analyses
  - Interpretations
- Data processing workflows
- Privacy concerns
- Hypothesis generation and their workflows

Still more opportunities for helping

- Data governance (Malcolm Chisholm)
- Data integration (Mark Fox and Rosario Uceda-Sosa)
- Schema.org, FOAF, Linked Data (Dan Brickley)
- ...

Harvest from data partners

- Rather than build it yourself, make use of collaborators.
  - You still have a lot of work to do converting, formalizing the input and integrating the sources.
  - But the result can be very high quality and it has a builtin user community.
- Nathan Wilson uses input from the EOL community.
- Mark Fox and Rosario Uceda-Sosa gave an example of harvesting data and metadata from city data sources.
- Ruth Duerr combined input from native communities in the Arctic.

Modular Development

- It is much easier to reuse a smaller ontology
  - Part of the Track A theme on reuse. One can combine several of them to create an ontology that satisfies most of your requirements.
- Ruth Duerr used this technique to develop an ontology for sea ice.

Formalize existing informal models

- Eric Chan presented a formalization of the OODA loop and then extended and generalized it.
- Ruth Duerr presented a formalization of existing techniques for describing sea ice conditions and indigenous knowledge.
- Mark Fox and Rosario Uceda-Sosa presented formalizations of existing informal data models for city data.

Develop ontology with extension points

- Eric Chan presented a framework for observation and decision making.
  - The framework has explicit extension points to allow ease of reuse.

Involve communities

- Similar to the Harvest from data partners use case
- Community involvement is important even if the communities do not directly contribute to the ontology.
- Ruth Duerr described this technique in her presentation on the ontology of sea ice.

Governance framework

- This framework uses relatively deep inference using axioms and rules to ensure data quality.
- The OOR initiative emphasized governance and business rules for ensuring quality.
- Malcolm Chisholm discussed the important characteristics of governance frameworks.

Bridge axioms

- Use axioms both at the data and metadata levels to bridge the gap between the semantics of data from different sources.
- Mark Fox and Rosario Uceda-Sosa both used this technique in their work on smart cities.

Other Use Cases

- Vocabulary pipeline
  - Dan Brickley Presentation
- Pattern matching
  - Dan Brickley Presentation
- Information ecosystem
  - Rosario Uceda-Sosa Presentation

Some Challenges

- Little collaboration between the communities
- Big Data focus on volume and velocity, assuming someone else will handle variety
- Tool incompatibility
- Incompatibility between statistical and logical techniques (hybrid reasoning gap)

Bibliography

1. D. Kerr, K. Braithwaite, N. Metropolis, D. Sharp, G.-C. Rota (Ed). Science, Computers and the Information Onslaught. Academic Press. (1984)
2. K. Baclawski and J.E. Smith. High-performance, distributed information retrieval. Northeastern University, College of Computer Science. (1994)
3. K. Baclawski. Data/knowledge bases for biological papers and techniques. Northeastern University, College of Computer Science. NSF grant. (1992)
4. Ioannidis JPA, Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124 (2005)
5. Critical Data Conference: Secondary Use of Big Data from Critical Care http://criticaldata.mit.edu/ (2014)
6. K. Baclawski and T. Niu. Ontologies for Bioinformatics. (2005)
7. K. Baclawski. Bayesian network development. In International Workshop on Software Methodologies, Tools and Techniques, pages 18-48. Keynote address. (September, 2004)