Open Knowledge Network: Recap & Some Related Work at NSF

Chaitan Baru
Senior Advisor for Data Science
CISE Directorate
National Science Foundation

Recap: Meeting #0

- January 2015
  - NITRD Big Data Strategic Initiative workshop, http://workshops.cs.georgetown.edu/BDSI-2015/.
  - Invited keynote by Andrew Moore, CMU
    - Introduced the idea of an initiative to “enumerate all entities” as a common platform for next generation, knowledge-based applications (taking a leaf from what is happening at large Web companies)
- May-June 2015
  - Discussions among Andrew Moore, RV Guha (ex-Google), George Strawn (NITRD), Phil Bourne (NIH), Chaitan Baru (NSF)
  - About an Open Knowledge Network: semantically-linked concepts, data
    - Can we create a **simple but powerful**, common platform (Semantic Web redux)

Meeting #1

- July 2016:
  - *Entities, Facts, Questions, Answers: Building the Foundations for Semantic Information Processing* meeting at White House OSTP
  - Attendees: Academia (CMU, UCSC, U.Mass, Georgetown); Industry (MSR, Amazon, IBM, ex-Google); Govt (NSF, NIH, NIST, DOD)
- ➔ Open Knowledge Network
  - Can we catalyze a community to bootstrap a new national infrastructure for a comprehensive distributed shared network of knowledge that builds upon simple (bottom up) standards and capable infrastructure to enable a new generation of knowledge-rich intelligent applications

OKN: Motivation

- **Natural interfaces** to **large knowledge structures** have the potential to impact science, education and business to an extent comparable to the WWW.
- The first wave has appeared in consumer services, e.g, **Siri, Cortana, Alexa**.
  - But limited in their scope to specific business areas, and proprietary, not open
- An **open effort** could help expand to 1000's of **new topic areas,** and many more **useful classes of questions**—even with current technologies.
  - A "Siri for science"
- Requires **convergence** among technology areas and domain sciences

Meeting #2

- February 2017:
  - *TOKeN: The Open Knowledge Network*, February 27th, Amazon offices, Sunnyvale, CA
  - About 45 attendees. Many from industry; several from academia.
- Topics discussed
  - Representation: Can we build more functional capabilities (e.g., representation of time, provenance, etc.) on a *core triples-based* model?
  - Serving: Can we provision a (cloud-based) service for hosting the OKN with acceptable performance; where users are able to upload, collaborate, download data
  - Competitions: Can progress be made via competitions? If so, for which tasks? (Kaggle is willing to host competitions)

Examples of Related NSF Efforts

- NSF has supported significant research on
  - creation of knowledge bases (representation, performance)
  - creation of ontologies
  - knowledge extraction
  - knowledge aggregation
  - reasoning …

Example NSF projects - 1

- **Knowledge Graph Mining for Financial Risk Analytics**, PI: Mohammed Zaki, 2017
  - a "financial risk" knowledge graph from textual and semantic features mined from the publicly available annual and quarterly reports filed with the SEC; and textual data from news articles and credit assessment reports.

- **Developing the Next Generation of Community Financial CyberInfrastructure for Monitoring and Modeling Financial Eco-Systems and for Managing Systemic Risk**, PI: Louiqa Raschid, 2013
  - Financial entity identification data challenges 2016, 2017
  - In collaboration with NIST and OFR, https://ir.nist.gov/dsfin
  - Creation of multiple open source graph datasets using SEC filings—in collaboration with IBM Almaden.

Example NSF projects - 2

- **From Data to Knowledge: Extracting and Utilizing Concept Graphs in Online Environments**, PI: Cornelia Caragea, 2016
  - Explore construction of scholarly knowledge graphs by combining evidence from multiple resources, in an open information extraction framework;
  - Design and develop novel algorithms for detection and analysis of interesting and previously unknown connections between concepts, to enforce knowledge discovery on the Scholarly Web;
  - Investigate the utility of scholarly knowledge graphs in a question answering system

Example NSF projects – 3

- **Scalable Probabilistic Inference for Large Knowledge Bases**, PI: Dan Suciu, 2016
  - Use of database technology to support construction of knowledge bases/graphs
- **Efficient Query Processing over Large Probabilistic Knowledge Bases**, PI: Daisy Zhe Wang, 2015
  - Infer missing knowledge from large-scale knowledge bases
- **Fusion of Heterogeneous Networks for Synergistic Knowledge Discovery**, PI: Philip Yu, 2015
  - Effective transfer of relevant knowledge across “partially aligned” networks—depends upon the relatedness of the different networks, and also the target applications/uses

Example NSF projects - 4

- **Constructing Knowledge Bases by Extracting Entity-Relations and Meanings from Natural Language via "Universal Schema”**, PI: Andrew McCallum, 2015
  - Automated knowledge base (KB) construction from natural language

- **Knowledge Graph Query Processing and Benchmarking**, PI: Xifeng Yan
  - Provide a standardized way to fairly and comprehensively evaluate different knowledge graph query algorithms;
  - Improve understanding of existing query engines;
  - Advance the area by providing a common benchmarking framework

Example NSF projects - 5

- **Using Knowledge Resources to Improve Information Retrieval, PI: Jamie Callan, 2014**
  - Examines how to use knowledge bases to improve IR tasks such as *ad hoc* search
  - Some of the work was performed in conjunction with Allen Institute for Artificial Intelligence's Semantic Scholar search engine.
  - Link documents and queries to the KB through entities…which improves the representation of the query and document, leading to more accurate ranking.
  - **KG4IR**: The First Workshop on Knowledge Graphs and Semantics for Text Retrieval and Analysis, in conjunction with ACM SIGIR 2017, Tokyo, Japan, August 11, 2017

Scientific data and ontologies

- Many efforts across sciences in development and use of ontologies:
  - E.g., biomedicine, biology, ecology, astronomy, hydrology, some areas of engineering…
- More recent efforts in other domains:
  - e.g. materials science, social science, education research, …

Examples from NSF EarthCube projects

- The Geoscience Standard Names (GSN) ontology, http://geoscienceontology.org
  - Contains 13,000 standardized modeling variable names using conventions based on very general principles.
  - Developed and used in NSF EarthCube project related to software descriptions, model interoperability, and modeling services

- The Linked Earth Ontology, http://linked.earth/ontology/
  - Describes paleoclimate datasets
  - Being used for crowdsourcing data curation for PAGES2K data—climate in the last 2K years: http://www.pages-igbp.org/ini/wg/2k-network/intro

OKN going forward…

- How to make progress in monotonically increasing, small steps…
- How to build this out as “infrastructure”…with continuity and persistence…