Semantic Technologies for Data Analysis in Health Care

Robert Piro Boris Motik Yavor Nenov Ian Horrocks
Peter Hendler Scott Kimberly Mike Rossman

University of Oxford

Kaiser Permanente

IHTSDO

Kobe, October 2016

Project Funding

DBOnto

- EPSRC (Government) funded "platform" at University of Oxford
- Funds exploratory projects with industry collaborators

Kaiser Permanente

- US "Health Maintenance Organisation" (HMO)
- Largest 'managed care' organisation in the US with 10.2M members
- Active in 8 US regions with 195 000 employees
- Turn over 56.4bn US$ and net income 3.1bn US$

Project Scope

Quality Measures in US Health Care

- HMOs must deliver anually quality of care info
- NCQA[1] maintains specifications for quality measures, e.g., HEDIS
- A quality measure is a percentage of a selected population, e.g.:

$$\frac{\#\text{diabetic patients with eye exams}}{\#\text{diabetic patients}} \times 100\%$$

- Accredits HMOs for public health care schemes (≈20% market)

Challenges with HEDIS

- HEDIS is a very complex specification (example later)
- Quality measures require complex analysis of the data
- Data needs to be assembled from heterogeneous data sources

[1] The National Committee of Quality Assurance

Benefits of Semantic Technology

Declarative Approach[2]

- Datalog with intuitive if-then-statments
- Close to natural language
- Tool support through explanations/query browser etc.

RDF as data format

- Predestined for data integration
- Flexible schema eases incremental development steps
- Leads naturally to a 'named perspective'

Benefits

- Shortened development cycles
- Improved maintenance
- Improved accuracy

[2] Requires optimised efficient evaluation by reasoner (here RDFox)

Project Scope

Tasks

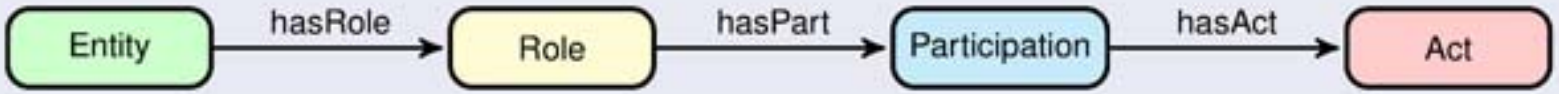
- Develop a data schema
- Encode the HEDIS specifications
- Translate data into RDF
- Evaluate & reconcile data

Data integrity was not an issue!

Semantic Technology Approach

The Data Model

- RDF is used to integrate data from the heterogeneous data sources
- Schema ontology describes a flexible but largely uniform schema
- Schema ontology designed according HL7 RIM standard
  - Ontology desing pattern: 'Enities in Roles Participating in Acts'
  - Familiar to Health-IT experts



Why not OWL / OWL 2 RL?

Example (Quote from the Definition of a Diabetic Patient)

[Diabetics are those patients] who met any of the following criteria during the measurement year [2013] or the year prior to the measurement year [2012] (count services that occur over both years):

- *At least two outpatient visits (Outpatient Value Set), observation visits (Observation Value Set) or nonacute inpatient visits (Nonacute Inpatient Value Set) on different dates of service, with a diagnosis of diabetes (Diabetes Value Set). Visit types need not be the same for the two visits.*
- ...

Why Datalog?

- non-treeshaped rules
- value comparisons
- large amount of data
- aggregation / negation

Datalog properly subsumes OWL 2 RL (PTIME complete)

Computation of the measures

RDFox

- in-memory RDF-store
- low memory footprint
- C++/Java/Python API
- SPARQL endpoint
- equality reasoning
- parallel Datalog engine
- highly scalable
- aggregates/stratified NAF
- explanations
- incremental reasoning

www.rdfox.org

Computation of the measures

A rule is an if-then-statment $B_1, \ldots, B_n \rightarrow H$

- Data is enriched with all consequences of RDF-data and rules
- Recursive evaluation of rules until a fixed point is reached
- SPARQL counting queries used to finally compute quality measures

Semantic Technology Approach

End-to-End Evaluation Process

- Commodity Hardware: 8 Intel Xeon @2.7GHz and 64GB RAM
- Data Translation: 10GB of patient data provided in 100 Million Records
- Translation with a Scala (Java) application: time 45min on 8 cores; resulting RDF-graph 293M triples
- Data Import with RDFox: 11min on 8 cores using 18GB (28% RAM)
- Computation of the HEDIS CDC measures and executing the counting queries: 19min on 8 cores

KAISER PERMANENTE

Conclusion

Summing up

- Applied Semantic Technologies to Data Analysis Problem
- Encoded one of the most complicated specifications
- Reduced development time, improved maintenance
- Increased accuracy, can justify results

Lessons Learnt

- OWL 2 is not enough!
- Invest time in designing a schema ontology!
- Choose UIDs (IRI) sensibly!

THANK YOU

Example (Expansion of the 'Patient-branch' in Clinical Visit)