Building Data Warehouses with Semantic Web Data

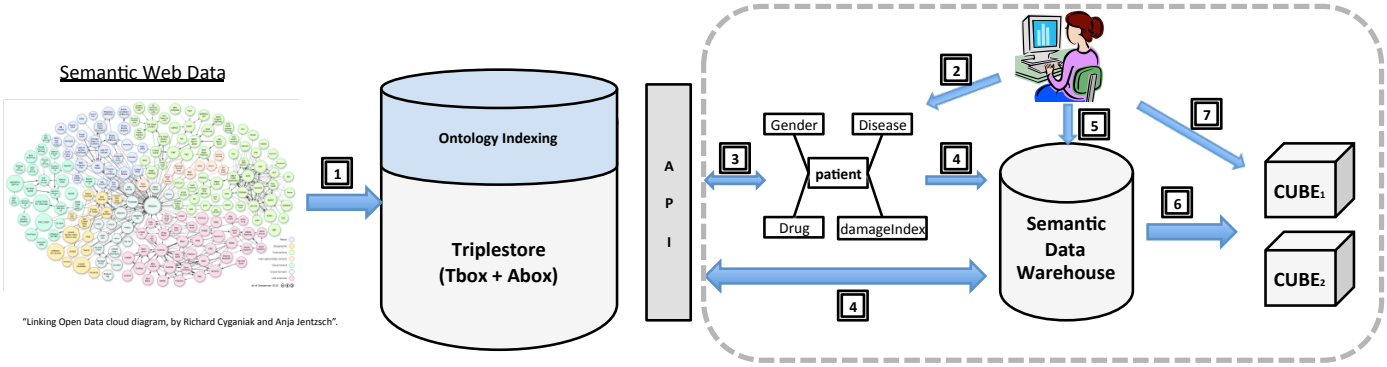
Victoria Nebot, Rafael Berlanga
Universitat Jaume I, Spain

1. MOTIVATION

- Lots of Semantic Web content being generated:
  - Triples *(subject, predicate, object)*
  - Graph-like structure
  - Open world assumption
  - Dynamic
  - Formal semantics -> implicit knowledge

2. GOAL

- Analyze Semantic Web data using multidimensional techniques (fact/dimension view)



FACT EXTRACTOR

- Explicit graph-based representation of Tbox and Abox.
- **Fact**: group of "combinable" instances characterizing some subject instance (i.e. instances having compatible contexts both at the conceptual and instance level).
  - Contexts(c_1, c_2, c_{sub}) are least common reachable concepts/instances. E.g. *Contexts(Diagnosis, Rheumatology, Patient)=Visit,* C*ontexts(DIAG2,RHEX2,PTN_XY21)= VISIT2*
  - Contexts graph (CG): composed by the subject of analysis (root), dimensions, measures, senses and contexts of senses.
  - Process CG in depth to recursivelly combine instances under context node instances.

DIMENSION EXTRACTOR

- **Dimension** D_i : subset of ontology taxonomy constructed from the instances of D_i in the fact table.
  - *Sig*(D_i): set of most specific concepts of the instances of D_i.
  - M_{Di} : upper module of ontology for *Sig*(D_i)
  - Select "good" nodes from M_{Di} to favor dense regions of the taxonomy (*share*) and good classification nodes (*entropy*).

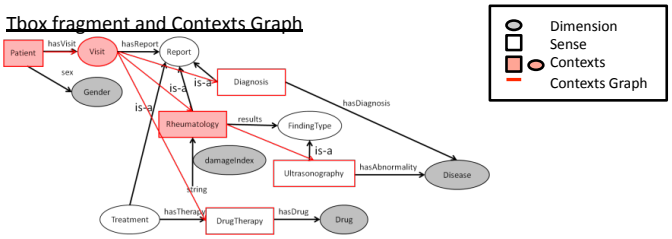
$$ratio\_share(n) = \frac{ideal\_share(n)}{share(n)} \qquad share(n) = \prod_{n_i \in ancs(n)} \frac{1}{|children(n_i)|}$$

$$entropy(n) = \sum_{n_i \in children(n)} P_{sig(n,n_i)} * \log(P_{sig}(n,n_i)) \qquad P_{sig}(n,n_i) = \frac{\text{cov}eredSig(n_i)}{\text{cov}eredSig(n)}$$

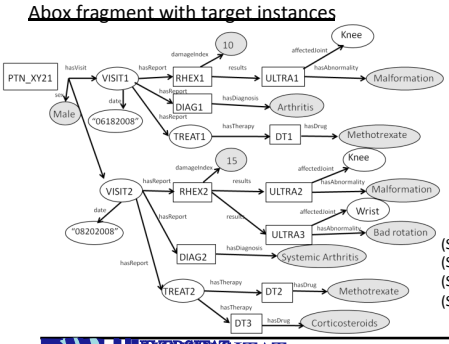
$$score(n) = ratio\_share(n) * entropy(n)$$

- Reconstruct hierachy with only good nodes (global vs. local)

Tbox fragment and Contexts Graph



Abox fragment with target instances

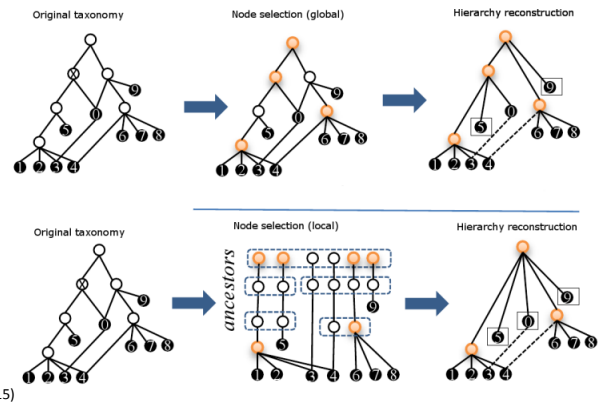


FACTS

(Disease,Drug,Gender,damageIndex)

(Arthritis,Malformation,Methotrexate,Male,10)

(SystemicArthritis,Malformation,Methotrexate,Male,15)
(SystemicArthritis,Malformation,Corticosteroids,Male,15)
(SystemicArthritis,BadRotation,Methotrexate,Male,15)
(SystemicArthritis,BadRotation,Corticosteroids,Male,15)



Nebot, V., Berlanga, R. (2009). Efficient retrieval of ontology fragments using an interval labeling scheme. *Information Sciences* 179(24):4151-4173.
Nebot, V., Berlanga, R. (2011). Building Data Warehouses with Semantic Web Data. *Decision Support Systems (Accepted)*.