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 (C1, C2, Csub) are least common reachable concepts/instances. E.g. Contexts(Diagnosis, Rheumatology, Patient)=Visit, C contexts(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 recursively 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_{Sig} : upper module of ontology for Sig(D_i)
  - Select “good” nodes from M_{Sig} to favor dense regions of the taxonomy (share) and good classification nodes (entropy).

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