Identifying Basic Level Entities in a Data Graph

Marwan Al-Tawil[1], Vania Dimitrova[1], Dhaval Thakker[2], Brandon Bennett[1]

[1] School of Computing, University of Leeds, UK

[2] School of Electrical Engineering and Computer Science, University of Bradford, UK

Data graphs (in the form of RDF linked data) become widely available on the Web and are being used in a myriad of applications. Gradually, data graphs are being exposed to users who are not familiar with the specific domain and can be unaware of the full structure of encoded knowledge. In other words, the cognitive structures users have about a domain may not match the structure of the data graph. This can provide major obstacles to exploration, especially when the users need to learn new things to make sense of data, which can lead to confusion and frustration. Our earlier research has shown that it is critical to identify anchoring entities from a data graph that serve as knowledge bridges to learn new concepts. From such anchors, new knowledge can be introduced following subsumption strategies for meaningful learning.

Identifying anchoring entities is not a trivial task because the graph usually includes thousands of entities at different levels of abstraction. We have utilised the Cognitive Science notion of basic level objects in a domain to develop algorithms for identifying basic level entities in a data graph. These entities will refer to the most inclusive categories at which objects are easily identified, and hence should provide good anchors for knowledge exploration. Two groups of algorithms are developed:

- *distinctiveness algorithms* which identify the most differentiated basic categories whose attributes are shared amongst the category members but are not associated to members of other categories; and
- *homogeneity algorithms* which identify basic categories whose members share many attributes together.

We applied the algorithms on an existing data set in the Music domain to identify basic level entities. The performance of the algorithms was examined using a benchmarking set with basic level entities in the data graph identified by humans. It corresponds to the cognitive structures humans form on the part of the world represented in the ontology, and is used as a ‘gold standard’ to evaluate the algorithms. Quantitative and qualitative data will be presented showing the strengths and limitations of each algorithm, leading to the development of a hybridisation approach based on the taxonomy level of each entity.