Towards Mining Generalized Patterns from RDF Data and a Domain Ontology

Tomas Martin, Victor Fuentes, Petko Valtchew, Abdoulaye Baniré Diallo, René Lacroix, Maxime Leduc, Mounir Boukadoum
Université du Québec à Montreal (CRIA, LACIM), LactaNet

1. Complex heterogeneous data for Dairy production

- 1.4M Cows
- 6500 Herds

- Part of a larger project on a comprehensive decision-support system (DSS) to optimize cow profitability in dairy farming
- Dairy dataset: phenotypic data and milk yield for all the heifers (past 20 years) as well as management conditions for herds
- Increasingly detailed genetic profile for each individual animal

Combining both genotypic and phenotypic data sources into an analytical model of dairy production should help dairy producers optimize their management decisions.

2. An Ontology to encode Domain Knowledge

- A rich unified schema, i.e. domain ontology (DO), to federate heterogeneous sources of dairy data
- A flexible data format to build a compatible dataset on top of it.

Ontologies = machine-readable structured representations of domain concepts and their relationships. Serve as unified data schemas, standardized and structured vocabularies, conceptual schemas, knowledge repositories, etc.

3. OGPS: Ontologically-generalized Graph Patterns

- Graph patterns = recurring fragments, i.e. sub-graphs, within the data
- Represent symbolic summaries of the commonalities in the data records.
- In OGP, vertex/edge labels = entities from the DO

Albeit more challenging to mine, they provide context to any shared element and a varying degree of abstraction.

4. OGP help share the hidden shared conceptual structure

Ontological entities in the patterns make explicit the shared conceptual structure that remains otherwise invisible in the raw data. While raw numbers and labels may mismatch, higher-order abstractions from the DO describing them may well coincide. The higher abstraction level in OGP = better generalization and increased expert readability.

Figure 3: Interesting patterns, too far down the pattern space.

However, graph mining comes at a relatively high computational cost and the DO amplifies the problem. Figure 3 shows an example of a pattern deemed interesting by our experts but hardly reachable with available tools. Thus, the design of computationally efficient DO-aware graph miner of sufficiently compact output is our current research target.

5. Two methods: gSpan-OF and Tax-ON

Two workaround solutions to dissociate topology and labels:
- gSpan-OF: a flat set of ontological labels in a pure graph mining task
- Tax-ON: (1) pure graph mining on dataset rewritten with only node labeled properties as labels; (2) successive label specialization on graph patterns from (1).

Both methods forsake part of the available structure:
- gSpan-OF: ignores the hierarchies in the DO (flattening).
- Tax-ON: ignores the constraints of the graph structure - graphs brought down to vertex sets (disconnecting).

Neither (flattening / disconnecting) ensure deep enough exploration of the pattern space.

6. A direct ontology-aware graph miner

OGPs are still beyond the reach of existing methods. Both approaches presented here suffer high computational costs due to the combinational explosion of the ontology induced pattern space.

Figure 5 clarifies the number of candidates. Taxogram examines while testing all possible specializations of a specific pattern (left 1168, 1767). Specializations are required to reach a most specific pattern while the peak number of candidates is generated at depths four and five values. The mapping aspect & scoring the tax 10k specializations tested, some 50% were duplicates (ratio increases with the pattern size).

Conclusion: A more direct approach is needed to deal with both topology enrichment and label specialization. Major challenges ahead are non-redundant candidate pattern generation (i.e. canonical representations) and efficient support computation.

Corresponding author contact information: E-mail address: valtchev.petko@uqam.ca. Tel: +1 514 987 1000. Fax: +1 514 987 8477.