Implicit Representation of Bigranular Rules for Multigranular Data

Stephen J. Hegner
DBMS Research of New Hampshire, USA

M. Andrea Rodríguez
University of Concepción, Chile

DEXA 2018
29th International Conference on
Database and Expert Systems Applications
Regensburg, Germany
04 September 2018

The Idea of Multigranular Attributes

Spatio-temporal attributes

Thematic attributes

cmn = comuna/county
prv = provincia/province
rgn = región/region

Place	Time	Births
Puerto_Montt_cmn	*Y2017Q1*	b_1
Llanquihue_prv	*Y2017Q1*	b_2
Los_Lagos_rgn	*Y2017Q1*	b_3
Puerto_Montt_cmn	*Y2017*	b_4

Granules: The domain values are called *granules*.

Granular order: The granules of spatial and temporal attributes have inherent order structure.

Spatial containment: *Puerto_Montt_cmn* $\sqsubseteq$ *Llanquihue_prv* $\sqsubseteq$ *Los_Lagos_rgn*

Temporal interval containment: *Y2017Q1* $\sqsubseteq$ *Y2017*

Typical constraints: Functional dependency (FD) $\{\text{Place}, \text{Time}\} \rightarrow \text{Births}$.

- The number of births is monotonic w.r.t. space and time, so $b_1 \leq b_2 \leq b_3$, $b_2 \leq b_4$.

Lattice-Like Operations on Granules

Place	Time	Births
Osorno_prv	*Y2017Q1*	b_1
Llanquihue_prv	*Y2017Q1*	b_2
Chiloé_prv	*Y2017Q1*	b_3
Palena_prv	*Y2017Q1*	b_4
Los_Lagos_rgn	*Y2017Q1*	b_5

Join: The four provinces join to the region.

$$Los\_Lagos\_rgn = \bigsqcup\{Osorno\_prv, Llanquihue\_prv, Chiloé\_prv, Palena\_prv\} \quad .$$

Meet: Distinct provinces are disjoint (six possibilities in all).

$$\bigsqcap\{Osorno\_prv, Llanquihue\_prv\} = \bot$$

Disjoint Join: The four provinces join *disjointly* to the region.

$$Los\_Lagos\_rgn = \bigsqcup^{\perp}\{Osorno\_prv, Llanquihue\_prv, Chiloé\_prv, Palena\_prv\} \quad .$$

Consequence: $\sum_{i=1}^{4} b_i = b_5$.

Observation: These lattice-like operations are *partial*.

Approaches to Modelling Multigranular Domains

Question: How can multigranular attributes be modelled for effective implementation? Two possibilities:

Single-structure model: Widely used in *Geographic Information Systems*.

Fix a domain $\mathcal{D}$; semantics of granule g defined by a subset $\mathrm{Sem}(g) \subseteq \mathcal{D}$.

- For a spatial attribute, $\mathcal{D} =$ suitable subset of $\mathbb{R} \times \mathbb{R}$.
- $\mathrm{Sem}(g)$ the geographic region represented by g.

Advantages: Extensive model; well-developed theory and practice.

Disadvantages: Extreme resource demands, both space and time.

Constraint-based model: Work directly with constraints of the form $g_1 \sqsubseteq g_2$, $g \sqsubseteq \biguplus\{g_i \mid 1 \leq i \leq k\}$, and $g = \biguplus\{g_i \mid 1 \leq i \leq k\}$, among others.

Advantages: Only represent as much information as needed.

Challenges: Constraint inference,

Constraint retrieval (based upon features), Constraint consistency.

A Logic for Multigranular Domains

- A logic for representing knowledge within multigranular attributes has been developed [Hegner & Rodríguez, 2016, 2017].

Granule Expressions (terms): $g, \quad \bigsqcup\{g \mid g \in S\}, \quad \bigsqcap\{g \mid g \in S\}, \quad \bot, \quad \top$

Granule Rules (sentences): Combine granule expressions using $\sqsubseteq$ (and equality). Examples: $g_1 \sqsubseteq g_2, \quad \bigsqcap\{g_1, g_2\} = \bot$.

Semantics/Models: Set semantics. Fix a domain $\mathcal{D}$.

- A model assigns to each granule a semantics $\mathsf{Sem}(g) \subseteq \mathcal{D}$ in a manner which respects the operations.

Example: $g_1 \sqsubseteq g_2$ iff $\mathsf{Sem}(g_1) \subseteq \mathsf{Sem}(g_2)$.

Satisfiability (of a set of rules): Related to distributivity of order operators.

- Very complex problem in theory.
- Not a problem in practice — axioms are models of "real" things.

Needs for an inference/lookup mechanism: The theory does not provide any such mechanism, beyond raw testing.

Goal of this work: Provide such a mechanism for the *common case*.

Requirements for Join-Rule Lookup

Join rule: One of the following forms: $g = \bigsqcup S \quad g = \biguplus S$

Head $g \sqsubseteq \bigsqcup S \quad g \sqsubseteq \biguplus S$ Body

Primary Lookup requirements:

Head lookup: Given a granule g', find all rules with head g'.

Body lookup: Given a set T of granules, find all rules with T contained in the body ($T \subseteq S$).

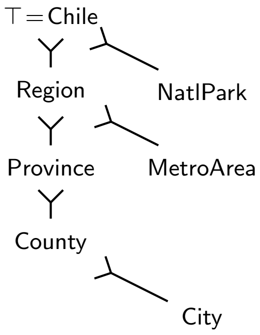
Complication: In the general case, the rules to be found may need to be derived first.

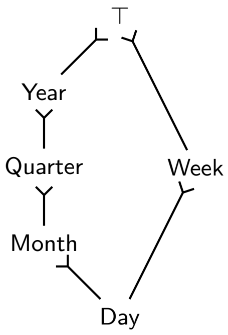
- A *knowledge base* of rules, not just a *database*.

Goal: For rule retrieval, *make the common case fast*.

Question: What is the common case?

Granularities — Organizing Granules





- The granules of each attribute are partitioned into a hierarchy of *granularities*.

Order: $G_1 \leq G_2 \Leftrightarrow ((\forall g_1 \in \text{Granules}\langle G_1 \rangle)(\exists g_2 \in \text{Granules}\langle G_2 \rangle)(g_1 \sqsubseteq g_2))$.

Examples: Every county is contained in a (unique) province.
Every day is contained in a (unique) week.

Disjointness: Distinct granules of the same granularity are disjoint.

Common Cases of Bigranular Rules

A common case: A family of *bigranular* rules between a pair of granules.

Bigranular rule: All granules in the body are of the same granularity.

- The head is necessarily of a different granularity.

Common case 1: Equality-join order property:

$G_1 \trianglelefteq G_2 \stackrel{\text{def}}{=}$ every granule of G_2 is the (disjoint) join of granules of G_1.

Example of common case 1: Province $\trianglelefteq$ Region.

$Los\_Lagos\_rgn = \bigsqcup^{\perp} \{Osorno\_prv, Llanquihue\_prv, Chiloé\_prv, Palena\_prv\}$.

$BíoBío\_rgn = \bigsqcup^{\perp} \{Arauco\_prv, BíoBío\_prv, Concepción\_prv, \cancel{Ñuble\_prv}\}$.

Common case 2: Subsumption-join order property:

$G_1 \lessdot G_2 \stackrel{\text{def}}{=}$ every granule of G_2 is contained in the (disjoint) join of granules of G_1.

Example of common case 2: MetroArea $\lessdot$ Province

$Gran\_Puerto\_Montt\_urb \sqsubseteq \bigsqcup^{\perp} \{Puerto\_Montt\_cmn, Puerto\_Varas\_cmn\}$

Resolvability and the Nondisjointness Relation

Context: Let $\mathcal{C}$ denote the set of constraints which hold on the multigranular attribute under consideration.

- Given a rule φ, there are three possible cases.

  (a) $\mathcal{C} \models \varphi$ (b) $\mathcal{C} \models \neg\varphi$ (c) Neither of these

Resolvability: The rule φ is *resolvable* from $\mathcal{C}$ if one of (a) or (b) holds.

- Written $\mathcal{C} \models^{\pm} \varphi$.

Context: $\langle G_1, G_2 \rangle$ a pair of granularities.

Full disjointness resolvability: $\langle G_1, G_2 \rangle$ is *fully disjointness resolvable* if $(\forall g_1 \in \mathrm{Granules}\langle G_1 \rangle)(\forall g_2 \in \mathrm{Granules}\langle G_2 \rangle)(\mathcal{C} \models^{\pm} (\prod\{g_1, g_2\} = \bot))$.

A compact relational representation under full disjointness resolvability:

$$\mathrm{NRel}_{\langle G_1, G_2 \rangle} = \{\langle g_1, g_2 \rangle \mid \mathcal{C} \models (\prod\{g_1, g_2\} \neq \bot)\}$$

- $\mathcal{C} \models ((\prod\{g_1, g_2\} = \bot))$ iff $(g_1, g_2) \notin \mathrm{NRel}_{\langle G_1, G_2 \rangle}$.

Symmetry: Note that $\mathrm{NRel}_{\langle G_1, G_2 \rangle} = \mathrm{NRel}_{\langle G_2, G_1 \rangle}$ always holds.

- Use $\mathrm{NRel}_{\langle G_1, G_2 \rangle}$ rather than $\mathrm{DRel}_{\langle G_1, G_2 \rangle}$ (the corresponding relation for disjointness) because $\mathrm{NRel}_{\langle G_1, G_2 \rangle}$ is usually much smaller.

Main Representation Theorem for $\trianglelefteq$

Recall: $G_1 \trianglelefteq G_2$ iff

$$(\forall g_2 \in \mathsf{Granules}\langle G_2 \rangle)(\exists S \subseteq_f \mathsf{Granules}\langle G_1 \rangle)(\mathcal{C} \models (g_2 = \bigsqcup S)).$$

Main Theorem: If $G_1 \trianglelefteq G_2$ holds, then so do the following.

(a) $\langle G_1, G_2 \rangle$ is fully disjointness resolvable. In other words, for $G_1 \trianglelefteq G_2$ to hold, there must be complete information about disjointness of the collective granules of G_1 and G_2.

(b) In the above "recall" formula, for each $g_2 \in \mathsf{Granules}\langle G_2 \rangle$,

$$S = \{g_1 \in \mathsf{Granules}\langle G_1 \rangle \mid \langle g_1, g_2 \rangle \in \mathsf{NRel}_{\langle G_1, G_2 \rangle}\}.$$

In words, each $g_1 \in \mathsf{Granules}\langle G_2 \rangle$ is the join of those granules in $\mathsf{Granules}\langle G_1 \rangle$ with which it is not disjoint. This is the only possibility.

Fast head-driven lookup: To identify the body of the rule with head $g_2 \in \mathsf{Granules}\langle G_2 \rangle$, it suffices to find all matches to g_2 in $\mathsf{NRel}_{\langle G_1, G_2 \rangle}$.

Fast body-driven lookup: To identify the head of the rule with $S' \subseteq \mathsf{Granules}\langle G_1 \rangle$ as part of its body, it suffices to find the unique g_1 which matches every member of S' in $\mathsf{NRel}_{\langle G_1, G_2 \rangle}$.

Main Representation Theorem for $\oslash$

Recall: $G_1 \oslash G_2$ iff

$$(\forall g_2 \in \mathsf{Granules}\langle G_2\rangle)(\exists S \subseteq_f \mathsf{Granules}\langle G_1\rangle)(\mathcal{C} \models (g_2 \sqsubseteq \bigsqcup\!\!\!\!\perp S)).$$

- The result is similar to that of $\trianglelefteq$, with one additional condition necessary.
- With subsumption, a *resolved minimality* condition is necessary.

Motivating example: Consider MetroArea $\trianglelefteq$ Province.

Trivial "solution": *Gran_Puerto_Montt_urb* $\sqsubseteq$ Granules$\langle$Province$\rangle$.

Resolved minimality: In the "recall" formula, for any proper subset $S' \subsetneq S$,

$$(\mathcal{C} \models \neg(g_2 \sqsubseteq \bigsqcup\!\!\!\!\perp S')).$$

- In other words, if any element is removed from S, the assertion becomes false (not just fails to be true).
- Under those conditions, a theorem analogous to that for $\trianglelefteq$ holds.

Conclusions and Current Directions

Conclusions:

Representation of common-case join rules:

- Handles both equality- and subsumption-join order.
- Applies in a constraint-based framework with incomplete information.
- Uses only (non)disjointness information about granules.

Current Directions:

Implementation in MGDB: *MGDB* is a PostgreSQL-based multigranular DBMS under development at the University of Concepción.

- Test data of administrative and political subdivisions of Chile provide many instances of equality- and subsumption-join order.
- The ideas of this paper are being applied to the efficient implementation of the associated rules.