NOTES ON RELATIONAL ALGEBRA

IORDAN GANEV

1. Introduction

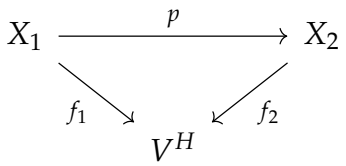
The organization of data is fundamental to all empirical endeavors. A simple way to organize data is through the use of a rectangular table, where the columns specify different attributes and each row records the attribute values of a single sample. This organizational approach is known as the relational model, and rows are often called 'tuples'. In what follows, we formulate this basic definition in somewhat more abstract terms; this formulation will give us a general perspective on basic table manipulations.

2. Basic definitions

Let V be the set of atomic values; for our purposes, we take $V = \mathbb{R}$ to be the set of all real numbers[1], and consider tables whose entries are in V.

A *header* is a finite set of (distinct) strings. Each element of the header corresponds to a column of our table; we refer to an element of the header as a *column heading*. An *H-tuple* is a function $H \to V$ from a header H to the set V. Thus, an H-tuple is simply a row in our table, and can represent a single sample. We denote the set of all H-tuples as V^H or $[H \to V]$.

A *relation* with header H is a function $f : X \to V^H$ where X is a finite set, called the *index set*. Thus, a relation is another name for a table: the elements of the index set X specify the rows, and we have an H-tuple for every row. Given two relations $f_1 : X_1 \to V^H$ and $f_2 : X_2 \to V^H$ with the same header H, a *map of relations* from f_1 to f_2 is a function $p : X_1 \to X_2$ between the index sets such that $f_2 \circ p = f_1$. In other words, we require the following diagram to commute:



In this way, we obtain a category $\mathcal{C}_H$ whose objects are relations with header H and whose morphisms are maps of relations. Extracting the index set defines a forgetful

[1]One can generalize some of the constructions we present below to the case were V is the set of all strings, Booleans, etc., or to distinguish between integers and floats. Some operations presented below rely on the total ordering of the real numbers.

functor For : $\mathcal{C}_H \to \mathsf{FSet}$ to the category of finite sets. Let $\mathsf{Rel}(H)$ be the set of equivalence classes of relations with header H. We also write $\mathsf{Rel}_X(H)$ for the set of equivalence classes of relations with header H and index set X.

3. Union

Recall the disjoint union functor on the category of finite sets:

$$\coprod : \mathsf{FSet} \times \mathsf{FSet} \to \mathsf{FSet}$$

For any headers H_1 and H_2, we describe a functor $\mathcal{C}_{H_1} \times \mathcal{C}_{H_2} \to \mathcal{C}_{H_1 \cap H_2}$ that extends the disjoint union operation in the sense that the following diagram commutes:

$$\begin{array}{ccc} \mathcal{C}_{H_1} \times \mathcal{C}_{H_2} & \longrightarrow & \mathcal{C}_{H_1 \cap H_2} \\ {\scriptstyle \text{For}}\downarrow & & \downarrow{\scriptstyle \text{For}} \\ \mathsf{FSet} \times \mathsf{FSet} & \xrightarrow{\coprod} & \mathsf{FSet} \end{array} \tag{3.1}$$

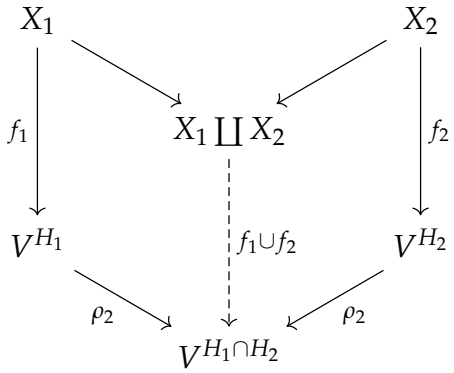
To this end, first observe that any function $H_1 \to V$ restricts to a function on the intersection $H_1 \cap H_2$, and similarly for H_2. Thus, we have restriction maps:

$$\rho_1 : V^{H_1} \to V^{H_1 \cap H_2} \qquad \text{and} \qquad \rho_2 : V^{H_2} \to V^{H_1 \cap H_2}.$$

Now let $f_1 : X_1 \to V^{H_1}$ and $f_2 : X_2 \to V^{H_2}$ be two relations. Then there is a natural[2] induced map from the disjoint union of X_1 and X_2 to $V^{H_1 \cap H_2}$:

$$f_1 \cup f_2 : X_1 \coprod X_2 \to V^{H_1 \cap H_2}$$

Specifically, this map takes $x \in X_i$ to $\rho_i \circ f_i(x)$, for $i = 1, 2$, and fits into the following commutative diagram:



The relation $f_1 \cup f_2$ defines an object of $\mathcal{C}_{H_1 \cap H_2}$. It is straightforward to check that the assignment $(f_1, f_2) \mapsto f_1 \cup f_2$ is functorial. Hence, we obtain the (generalized) union functor

$$\cup : \mathcal{C}_{H_1} \times \mathcal{C}_{H_2} \to \mathcal{C}_{H_1 \cap H_2}$$

making Diagram 3.1 commute.

[2]It is 'natural' in the sense of category theory, as it is induced from the universal property of the coproduct $X_1 \coprod X_2$.

In terms of tables, the union operation makes a new table whose columns are the common columns of the two tables and whose rows are all the (now truncated) rows of the original two tables. Special cases:

- If $H_1 = H = H_2$, we obtain a functor $\cup : \mathcal{C}_H \times \mathcal{C}_H \to \mathcal{C}_H$ matching the usual union operation from database theory.
- If $H_1 \cap H_2 = \emptyset$, we obtain the empty table for any pair of inputs:

$$\cup : \mathcal{C}_{H_1} \times \mathcal{C}_{H_2} \to \mathcal{C}_\emptyset = \{\emptyset\}$$

4. Product

Recall the product functor on the category of finite sets:

$$\times : \mathsf{FSet} \times \mathsf{FSet} \to \mathsf{FSet}$$

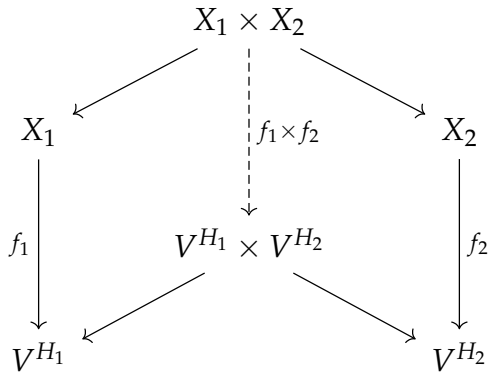
For any headers H_1 and H_2, we describe a functor $\mathcal{C}_{H_1} \times \mathcal{C}_{H_2} \to \mathcal{C}_{H_1 \coprod H_2}$ that extends the product operation in the sense that the following diagram commutes:

$$\begin{array}{ccc} \mathcal{C}_{H_1} \times \mathcal{C}_{H_2} & \longrightarrow & \mathcal{C}_{H_1 \cap H_2} \\ {\scriptstyle \text{For}}\downarrow & & \downarrow{\scriptstyle \text{For}} \\ \mathsf{FSet} \times \mathsf{FSet} & \xrightarrow{\times} & \mathsf{FSet} \end{array} \tag{4.1}$$

To this end, let $f_1 : X_1 \to V^{H_1}$ and $f_2 : X_2 \to V^{H_2}$ be two relations. Then there is a natural induced map from the product of X_1 and X_2 to $V^{H_1} \times V^{H_2}$:

$$f_1 \times f_2 : X_1 \times X_2 \to V^{H_1} \times V^{H_2}$$

Specifically, this map takes the pair (x_1, x_2) to $(f_1(x_1), f_2(x_2))$ and fits into the following commutative diagram:



where the slanted maps are the natural projections. In other words, we concatenate column headings and fill in the table in all possible ways from the original table. It is straightforward to check that the assignment $(f_1, f_2) \mapsto f_1 \times f_2$ is functorial. Observing that $V^{H_1 \coprod H_2} \simeq V^{H_1} \times V^{H_2}$, we obtain the product functor

$$\times : \mathcal{C}_{H_1} \times \mathcal{C}_{H_2} \to \mathcal{C}_{H_1 \coprod H_2}$$

making Diagram 4.1 commute.

5. Linked join (fiber products)

Let H_1 and H_2 be headers, and recall the restriction maps $\rho_i : V^{H_i} \to V^{H_1 \cap H_2}$ from above. Observe that the fiber product of the maps ρ_i can be identified with the space $V^{H_1 \cup H_2}$, that is, the following diagram is a pullback square:

$$\begin{array}{ccc} V^{H_1 \cup H_2} & \longrightarrow & V^{H_2} \\ \downarrow & & \downarrow{\scriptstyle \rho_2} \\ V^{H_1} & \xrightarrow[\rho_1]{} & V^{H_1 \cap H_2} \end{array}$$

Furthermore, given two relations $f_1 : X_1 \to V^{H_1}$ and $f_2 : X_2 \to V^{H_2}$, we can consider the fiber product of the maps $\rho_i \circ f_i$, namely:

$$X_1 \times_{V^{H_1 \cap H_2}} X_2 = \{(x_1, x_2) \in X_1 \times X_2 \mid \rho_1 \circ f_1(x_1) = \rho_2 \circ f_2(x_2)\}$$

In other words, this is the set of all pair of rows (one from the first table, one from the second) that have the same entries on the common columns. This space fits into the following pullback square:

$$\begin{array}{ccc} X_1 \times_{V^{H_1 \cap H_2}} X_2 & \longrightarrow & X_2 \\ \downarrow & & \downarrow{\scriptstyle \rho_2 \circ f_2} \\ X_1 & \xrightarrow[\rho_1 \circ f_1]{} & V^{H_1 \cap H_2} \end{array}$$

By universal properties of pullbacks, there is a natural map connecting these two fiber products:

$$f_1 \bowtie f_2 : X_1 \times_{V^{H_1 \cap H_2}} X_2 \to V^{H_1 \cup H_2}$$

The resulting table is known as the linked join of the original two tables. It contains all the columns of the original two tables, but only those rows that are compatible on the common columns. It is straightforward to check that the assignment $(f_1, f_2) \mapsto f_1 \bowtie f_2$ is functorial. We obtain the linked join functor:

$$\bowtie : \mathcal{C}_{H_1} \times \mathcal{C}_{H_2} \to \mathcal{C}_{H_1 \cup H_2}$$

Note that it does not extend and operation on finite sets, since the fiber product depends on the maps f_1 and f_2. Special cases:

- If $H_1 = H_2 = H$, we recover the usual intersection operation $\cap : \mathcal{C}_H \times \mathcal{C}_H \to \mathcal{C}_H$. In this case, the restriction maps ρ_i are identity maps, and the fiber product becomes:

  $$X_1 \times_{V^H} X_2 = \{(x_1, x_2) \in X_1 \times X_2 \mid f_1(x_1) = f_2(x_2)\}.$$

  and the linked join map $X_1 \times_{V^H} X_2 \to V^H$ takes (x_1, x_2) to the common value $f_1(x_1) = f_2(x_2)$.

- If $H_1 \cap H_2 = \emptyset$, then the union $H_1 \cup H_2$ is a disjoint union $H_1 \coprod H_2$, and we recover the product operation from above.

Remark 5.1. In practice, the linked join of two tables with no common column headings is conventionally deemed to be the empty table, that is, their union. The product is implemented separately.

6. Other operators

6.1. **Projection.** Let H_1 and H_2 be headers. Given a function $\phi : V^{H_1} \to V^{H_2}$, we can transform relations with header H_1 into relations with header H_2 by postcomposition. Explicitly, if $f : X \to V^{H_1}$ is a relation with header H_1, then $\phi \circ f : X \to V^{H_2}$ is a relation with header H_2. It is straightforward to check that the assignment $f \mapsto \phi \circ f$ is functorial between the categories $\mathcal{C}_{H_1}$ and $\mathcal{C}_{H_2}$. The resulting functor is known as the projection corresponding to ϕ:

$$\mathrm{Proj}_\phi : \mathcal{C}_{H_1} \to \mathcal{C}_{H_2}$$

Thus, we see that functions $V^{H_1} \to V^{H_2}$ give rise to functors $\mathcal{C}_{H_1} \to \mathcal{C}_{H_2}$.

We remark on a connection between the projection operator and the union operator. Suppose $H' \subseteq H$ is a subheader of H. The inclusion $i : H' \hookrightarrow H$ induces a pullback map $i^* : V^H \to V^{H'}$, and we have the corresponding projection functor:

$$\mathrm{Proj}_{i^*} : \mathcal{C}_H \to \mathcal{C}_{H'}.$$

Let $\emptyset_{H'} \in \mathsf{Rel}(H')$ be the empty relation. Taking the union of any relation in $\mathcal{C}_H$ with $\emptyset_{H'}$ gives a functor: $- \cup \emptyset_{H'} : \mathcal{C}_H \to \mathcal{C}_{H'}$. One can check that this functor coincides with Proj_{i^*}.

6.2. **Select.** Let $f : X \to V^H$ be a relation in $\mathcal{C}_H$. For any subset $W \subseteq V^H$, we can consider the relation f restricted to the preimage $f^{-1}(W)$ of W. Thus, we obtain a map from the power set of V^H to subobjects of f:

$$\mathcal{P}(V^H) \to \mathrm{SubObj}(f), \qquad W \mapsto [f^{-1}(W) \to V^H]$$

6.3. **Difference.** For any set Y, there is a difference operator:

$$\mathrm{diff}_Y : \mathsf{Rel}(H) \to \mathsf{Rel}(H)$$

taking a relation $f : X \to V^H$ to the restriction of f to $X \setminus (X \cap Y)$. This operator is not functorial.

7. Operations with in the category $\mathcal{C}_H$

Throughout this section, H is a fixed header and all relations will have header H.

7.1. **Pullback.** We first explain how relations pull back under maps between the rows. Let $\alpha : X \to Y$ be a function between the finite sets X and Y. There is a pullback operation:

$$\alpha^* : \mathsf{Rel}_Y(H) \to \mathsf{Rel}_X(H), \qquad [\alpha^*(g)](x) = g(\alpha(x))$$

which takes a relation $g : Y \to V^H$ with index set Y to the relation $g \circ \alpha : X \to V^H$ with index set X. The push forward operation requires aggregation functions, which we now discuss.

7.2. **Aggregating functions.**

7.2.1. *Aggregating a single column.* We denote the one-point set by $\{\mathrm{pt}\}$ and consider relations $\mathsf{Rel}(\{\mathrm{pt}\})$ with header equal to the one-point set. Such a relation is the same as a function $X \to V$; in other words, it is a table with a single column. The *dual space* of $\mathsf{Rel}(\{\mathrm{pt}\})^*$ is the set of functions from $\mathsf{Rel}(\mathrm{pt})$ to V. In other words, an element of the dual space gives a function from X^V to V for every finite set X. An element of the dual space $\mathsf{Rel}(\{\mathrm{pt}\})^*$ is known as an *aggregating function*. Thus, an aggregating function FUN is a map:

$$\mathrm{FUN} : \bigcup_{X \text{ finite set}} V^X \to V$$

When $V = \mathbb{R}$, examples include AVG, MIN, MAX, COUNT, SUM, etc.

7.2.2. *Aggregating multiple columns.* Let H be a header. The dual space to $\mathsf{Rel}(H)$ is defined as $\mathsf{Rel}(H)^* = (\mathsf{Rel}(\mathrm{pt})^*)^H$. Thus, an element of the dual space is a choice of aggregating function for every attribute in H. Note that there is no notion of an 'index set' for an element of the dual space. We will use the same notation FUN for elements of $\mathsf{Rel}(H)^*$ as for elements of $\mathsf{Rel}(\{\mathrm{pt}\})^*$; it will be clear from context which meaning is intended. There is an *evaluation map*:

$$\mathrm{ev} = \langle -, - \rangle : \mathsf{Rel}(H)^* \times \mathsf{Rel}(H) \to V^H$$

which applies an aggregating function to each column of a relation. In symbols, let $f : X \to V^H$ be a relation and let $\mathrm{FUN} = (\mathrm{FUN}^{(h)})_{h \in H}$ be a tuple of aggregating functions, one for every attribute in H. Then

$$\langle\ \mathrm{FUN}\ ,\ f\ \rangle(h) = \mathrm{FUN}(\mathrm{ev}_h \circ f)$$

where $\mathrm{ev}_h : [H \to V] \to V$ is the 'evaluation at h' map.

7.3. **Aggregate push forward.** We are now ready to explain the aggregate push forward operation. For any aggregating function $\mathrm{FUN} \in \mathsf{Rel}(H)^*$, we have the FUN-pushforward[3] along α, given by:

$$\alpha_*^{\mathrm{FUN}} : \mathsf{Rel}_X(H) \to \mathsf{Rel}_{\alpha(X)}(H), \qquad [\alpha_{\mathrm{FUN}}(f)](y) = \left\langle\ \mathrm{FUN}\ ,\ f\big|_{\alpha^{-1}(y)}\ \right\rangle.$$

[3]The usual pushforward function operation on vector-valued functions on finite sets involves on summing over fibers, that is, $\alpha_* : \mathsf{Rel}_{\alpha(X)}(H)$ with $[\alpha_*(f)](y)(h) = \sum_{x \in \alpha^{-1}(y)} f(x)(h)$. The general version replaces summation by any aggregating function.

The resulting relation has as its index set the image $\alpha(X) \subseteq Y$ of X in Y under α.

7.4. **Group by.** Suppose we have a relation f with attributes H. Suppose H_1 and H_2 are disjoint subsets of H, so that we have the relations[4] $f_1 \in \mathsf{Rel}(H_1)$ and $f_2 \in \mathsf{Rel}(H_2)$. Then grouping f by H_1 and aggregating H_2 is a function:

$$\mathsf{Rel}(H_2)^* \times \mathsf{Rel}_X(H_1 \coprod H_2) \to \mathsf{Rel}_{f_1(X)}(H_1)$$

defined in terms of pushing forward f_2 along f_1 with a specified aggregating function $\mathrm{FUN} \in \mathsf{Rel}(H_2)^*$:

$$(f_1)_*^{\mathrm{FUN}}(f_2)(\mathbf{v}) = \left\langle\ \mathrm{FUN}\ ,\ f_2\big|_{f_1^{-1}(\mathbf{v})}\ \right\rangle$$

for $\mathbf{v}$ in the image $f_1(X) \subseteq V^{H_1}$.

7.5. **Window functions.** Recall that $\mathcal{P}(X)$ is the power set of X. Given a relation $f \in \mathsf{Rel}_X(H)$ and a subset $X' \subseteq X$, we can restrict f to X' to obtain a relation $f\big|_{X'} \in \mathsf{Rel}_{X'}(H)$. A *window assignment* is a choice of subset of X for every element of X, i.e., a map

$$\Phi : X \to \mathcal{P}(X).$$

Given a relation, aggregating function, and window assignment, we obtain a new relation which aggregates according to the windows. More precisely, we have a *window function*:

$$\mathsf{Rel}(H)^* \times \mathsf{Rel}_X(H) \times [X \to \mathcal{P}(X)] \to \mathsf{Rel}_X(H)$$

taking (f, FUN, Φ) to the relation $\overline{\Phi}^{\mathrm{FUN}}(f) : X \to V^H$ sending $x \in X$ to $\left\langle\ \mathrm{FUN}\ ,\ f\big|_{\Phi(x)}\ \right\rangle \in V^H$. Note that the index set X remains unchanged.

7.6. **Pullback and window functions.** Recall that relations pull back under maps between the rows: $\alpha : X \to Y$ induces $\alpha^* : \mathsf{Rel}(Y) \to \mathsf{Rel}(X)$. Such a map induces a window assignment taking x to the fiber to which it belongs:

$$\Phi_\alpha : X \to \mathcal{P}(X); \qquad x \mapsto \alpha^{-1}(\alpha(x)).$$

One can show that $\overline{\Phi}_\alpha^{\mathrm{FUN}} = \alpha^* \alpha_*^{\mathrm{FUN}}$. We also have a map:

$$\mathsf{Rel}(H)^* \times \mathsf{Rel}_X(H) \times [X \to Y] \to \mathsf{Rel}_{\alpha(X)}(H)$$

taking f, FUN, α to the relation taking $y \in \alpha(X)$ to $\left\langle\ \mathrm{FUN}\ ,\ f\big|_{\alpha^{-1}(y)}\ \right\rangle$.

[4]More explicitly, for each $x \in X$, the function $f_i(x) : H_i \to V$ is defined as the restriction of $f(x) : H \to V$ to the subset $H_i \subseteq H$.

8. Dependencies

Let H be a header and let $f : X \to V^H$ be a relation. For $h_1, h_2 \in H$, we say that f exhibits a dependency $h_1 \to h_2$ if there is a function $g : V \to V$ such that

$$f(x)(h_2) = g(f(x)(h_1))$$

for all $x \in X$. One can write down the corresponding commutative diagram. The dependency is only in the relation f in question; other relations with the header H may not exhibit the dependency.

Another way to think about dependencies is as follows. Consider the subset $\{h_1, h_2\}$ of H. This is a two-element set, so relations on it are functions $f : X \to V \times V$, where the first coordinate corresponds to h_1 and the second to h_2. We say that a relation on $\{h_1, h_2\}$ exhibits a dependency $h_1 \to h_2$ if the image $f(X) \subseteq V \times V$ is contained in the graph of a function. Now, we say that $f : X \to V^H$ exhibits a dependency $h_1 \to h_2$ if its projection $\mathrm{Proj}_{\{h_1,h_2\}}(f)$ does. This is equivalent to the above definition (indeed, consider the graph of g).

Heath's theorem asserts that, if f exhibits a dependency $h_1 \to h_2$, then there is an isomorphism:

$$\mathrm{Proj}_{\{h_1,h_2\}}(f) \bowtie \mathrm{Proj}_{H\setminus\{h_2\}}(f) = f$$