CSC343 Assignment 1: Relational Algebra

February 7, 2019

1 Constraints

For each of the following constraints give a one sentence explanation of what the constraint implies, and why it is required.

- $\pi_{species}(Artifact) - \pi_{species}(Species) = \emptyset$.

  Every artifact is assigned to a known species in the Catalogue of Life database upon arrival at the institute, for purposes of proper taxonomic practices.

- $\pi_{rank}(Staff) \subseteq \{\text{'technician', 'student', 'pre-tenure', 'tenure'}\}$.

  All staff ranks must fall under technician, student, pre-tenure or tenure, otherwise a staff member could have an unknown rank.

- $\pi_{family}(Genus) - \pi_{family}(COL) = \emptyset$.

  Every genus belongs to a family name that is known in the Catalogue of Life, as per proper taxonomic practice.

- $\pi_{genus}(Species) \subseteq \pi_{genus}(Genus)$.

  All species belong to a known genus in the Catalogue of Life, as per proper taxonomic practice.

- $\pi_{CID}(Collected) = \pi_{CID}(Collection)$.

  The collections that artifacts belong to are known entire collections from field trips.

- $\pi_{AN}(Artifact) = \pi_{AN}(Collected)$.

  All artifacts belong to collections from field trips.

- $\pi_{SID}(Collection) \subseteq \pi_{SID}(Staff)$.

  Collections from field trips may only be done by the institute's scientific staff, however, not all staff members have found collections.

- $\pi_{SID}(Artifact) \subseteq \pi_{SID}(Staff)$.

  Personnel who maintain an artifact must be a member of the institute scientific staff, ensuring that the artifact is safely stored.

- $\pi_{type}(Artifact) \subseteq \{'tissue','image','model','live'\}$

  All artifacts types must fall under tissue, image, model or live.

- $\pi_{AN}(Published) \subseteq \pi_{AN}(Artifact)$

  Artifacts mentioned in scholarly publications must be known single artifacts collected in the field.

2 Queries

Write relational algebra expressions for each of the queries below.

1. Rationale: Performance reviews include seeing how current the work is of staff who have held their current rank for a long time.

   Query: Find the most recent collection date of any artifact collected by a staff member who has held their current rank the longest. Keep ties.

   We make a key assumption: Date is a numeric value with smaller values denoting older dates.

   - We begin by determining the SIDs of staff who have not held the same rank for the longest time:

   $$\text{notLongestRankStaff} := \pi_{\text{S1.SID}}\sigma_{\text{S1.date}>\text{S2.date}}(\rho_{\text{S1}}\ \text{Staff} \times \rho_{\text{S2}}\ \text{Staff})$$

   - We now find the staff who have held their rank the longest:

   $$\text{longestRankStaff} := \pi_{\text{SID}}(\text{Staff}) - \text{notLongestRankStaff}$$

   - Determine the collections corresponding to the longest-rank staff from the above line:

   $$\text{tempCollection} := \sigma_{\text{Collection.SID = longestRankStaff.SID}}(\text{Collection})$$

   - We find the collections (gathered by our longest-rank staff) that do not have the newest date:

   $$\text{notNewestCollection} := \pi_{\text{T1.date}}\sigma_{\text{T1.date}<\text{T2.date}}(\rho_{\text{T1}}\ \text{tempCollection} \times \rho_{\text{T2}}\ \text{tempCollection})$$

   - Finally, find the most recent collection dates from the longest-rank staff:

   $$\text{newestCollectionDate} := \pi_{\text{date}}(\text{Collection}) - \text{notNewestCollection}$$

2. Rationale: Staff who maintain every artifact in some collection should be considered favourably in performance reviews.

   Query: Find all staff who maintain all artifacts in at least one collection.

   - We first determine the CIDs of collections and link them with the staff members who perform maintenance:

   $$\text{CollectionStaff} := \pi_{\text{CID, SID}}(\text{Collected} \bowtie \text{Artifact})$$

   - We then find the collections that are maintained by more than one staff member:

   $$\text{MultiStaffCollection} := \pi_{\text{C1.CID}}\left[\sigma_{\substack{\text{C1.CID=C2.CID}\wedge \\ \text{C1.SID}\neq\text{C2.SID}}}(\rho_{\text{C1}} \mathit{CollectionStaff} \times \rho_{\text{C2}} \mathit{CollectionStaff})\right]$$

   - We now find collections maintained by only one staff member:

   $$\text{SingleStaffCollection} := (\pi_{\text{CID}}\text{Collection}) - \text{MultiStaffCollection}$$

   - Finally we get the SIDs of the staff who maintain all artifacts in at least one collection:

   $$\text{AllArtifactStaff} := \pi_{\text{SID}}(\text{SingleStaffCollection} \bowtie \text{Collection})$$

3. Rationale: An artifact collected and maintained by the same staff may have some special requirements that should be investigated.

   Query: Find all artifacts that were collected by the same staff who maintains them.

- We find the artifact numbers where the maintenance staff is the same as the collection staff.

$$\pi_{\text{AN}}(\text{Collection} \bowtie \text{Collected} \bowtie \text{Artifact})$$

4. Rationale: Identify multi-talented field workers.

   Query: Find all staff who have collected at least 3 artifacts from every species in some family.

   - We first construct a relation that ties together ANs, CIDs and SIDs. This will allow us to isolate for staff who have collected at least three unique artifacts under the same species:

   $$\text{Triples} := (\pi_{\text{AN, species}}\text{Artifact}) \bowtie \text{Collected} \bowtie (\pi_{\text{CID, SID}}\text{Collection})$$

   - We will now generate all possible combinations of three artifacts, and project the relevant columns of AN, species and SID:

   $$\text{AllCombinations} := \pi_{\substack{\text{A1.AN, A1.species, A1.SID,}\\ \text{A2.AN, A2.species, A2.SID,}\\ \text{A3.AN, A3.species, A3.SID}}}(\rho_{\text{A1}}\text{Triples} \times \rho_{\text{A2}}\text{Triples} \times \rho_{\text{A3}}\text{Triples})$$

   - We construct a relation that links all species with their families:

   $$\text{SpecFam} := \pi_{\text{species, family}}(\text{Species} \bowtie \text{Genus})$$

   - We now select only the rows where the same staff member has collected at least three different artifacts under the same species, and we add the appropriate family to each row via a theta join:

   $$\text{TripleSpecies} := \pi_{\substack{\text{A1.SID,}\\ \text{A1.species,}\\ \text{SpecFam.family}}}\left(\sigma_{\substack{\text{A1.AN}\neq\text{A2.AN}\wedge\\ \text{A1.AN}\neq\text{A3.AN}\wedge\\ \text{A2.AN}\neq\text{A3.AN}\wedge\\ \text{A1.species=A2.species}\wedge\\ \text{A1.species=A3.species}\wedge\\ \text{A2.species=A3.species}\wedge\\ \text{A1.SID=A2.SID}\wedge\\ \text{A2.SID=A3.SID}}}\text{AllCombinations}\right) \bowtie_{\text{A1.species=SpecFam.species}} \text{SpecFam}$$

   - Each SID that has at least 3 artifacts from one species in the associated family:

   $$\text{StaffFam} := \pi_{\text{SID, family}}\text{TripleSpecies}$$

   - Find all the possible species a staff member would need to collect three artifacts from in order to complete a family. This only considers the families from which they have already collected at least three artifacts from at least one species, not all possible families.

   $$\text{AllPossibleStaffSpeciesFamilies} := \text{StaffFam} \bowtie \text{SpecFam}$$

   - We form a difference to remove all confirmed tuples where the staff member has collected at least three artifacts of the corresponding species. The remaining tuples represent missing species that a staff member has failed to collect from the associated family.

   $$\text{MissingSpeciesFromFamily} := \text{AllPossibleStaffSpeciesFamilies} - \text{TripleSpecies}$$

   - Get answer

   $$\text{multiTalentedStaff} := StaffFam$$

5. Rationale: Which publications might have some specialized niche focus?

   Query: Find all publications that have used exactly 2 of our artifacts.

- We first determine the publications where at least two of the artifacts have been used:

  $$\text{atLeastTwo} := \sigma_{\substack{\text{P1.journal} = \text{P2.journal}\wedge \\ \text{P1.date} = \text{P2.date}\wedge \\ \text{P1.AN}\neq\text{P2.AN}}}(\rho_{\text{P1}}\ \text{Published} \times \rho_{\text{P2}}\ \text{Published})$$

- We now determine the publications where at least three of the artifacts have been used:

  $$\text{atLeastThree} := \sigma_{\substack{\text{P1.journal} = \text{P2.journal}\wedge \\ \text{P1.date} = \text{P2.date}\wedge \\ \text{P1.AN}\neq\text{P2.AN}\wedge \\ \text{P2.journal} = \text{P3.journal}\wedge \\ \text{P2.date} = \text{P3.date}\wedge \\ \text{P2.AN}\neq\text{P3.AN}}}(\rho_{\text{P1}}\ \text{Published} \times \rho_{\text{P2}}\ \text{Published} \times \rho_{\text{P3}}\ \text{Published})$$

- Now, we isolate for the publications which have used exactly two artifacts:

  $$\text{exactlyTwo} := \text{atLeastTwo} - \text{atLeastThree}$$

6. Rationale: Identify motherlode locations.

   Query: Find all locations where at least one artifact from every family has been collected.

   - We begin by finding a list of families that have been found at every location:

     $$\text{familiesAtLocations} := \pi_{\text{location, family}}(\text{Artifact} \bowtie \text{Species} \bowtie \text{Genus})$$

   - Next, we determine all possible combinations of locations and families:

     $$\text{allFamilyLocationCombinations} := \pi_{\text{location}}(\text{familiesAtLocations}) \times \text{COL}$$

   - We then isolate all the locations that do *not* have at least one artifact from every family collected there:

     $$\text{notMotherlode} := \pi_{\text{location}}(\text{allFamilyLocationCombinations} - \text{familiesAtLocations})$$

   - Finally, we remove these non-motherlode locations from all possible locations to find the ones which *are* motherlode locations:

     $$\text{motherlodeLocations} := \pi_{\text{location}}(\text{Artifact}) - \text{notMotherlode}$$

7. Rationale: Exclusively tissue sample collectors may need extra support for special reagents and shipping costs.

   Query: Find all staff who have collected only tissue samples.

   - We first determine any staff who have collected non-tissue samples:

     $$\text{notTissueStaff} := \pi_{\text{SID}}\sigma_{\substack{\text{type} = \text{`image'}\vee \\ \text{type} = \text{`model'}\vee \\ \text{type} = \text{`live'}}}\text{Artifact}$$

   - We then remove these staff to isolate for the tissue-only collectors:

     $$\text{tissueStaff} := \pi_{\text{SID}}\text{Artifact} - \text{notTissueStaff}$$

8. Rationale: Collection staff who should be encouraged to diversify their network.

   Query: Find all staff pairs who have worked only with each other on collections.

- Create a relation that contains the SID of each staff member and which artifacts they have collected:

$$\text{artifactsCollected} := \text{Collection} \bowtie \text{Collected}$$

- Next, determine all collectors who have worked with others:

$$\text{collaborativeCollectors} := \pi_{\text{artifactsCollected.SID}}(\sigma_{\substack{\text{artifactsCollected.AN=Artifact.AN}\wedge \\ \text{artifactsCollected.SID}\neq\text{Artifact.SID}}}(\text{artifactsCollected} \times \text{Artifact}))$$

- Next, determine all maintainers who have worked with others:

$$\text{collaborativeMaintainers} := \pi_{\text{Artifact.SID}}(\sigma_{\substack{\text{artifactsCollected.AN=Artifact.AN}\wedge \\ \text{artifactsCollected.SID}\neq\text{Artifact.SID}}}(\text{artifactsCollected} \times \text{Artifact}))$$

- We then determine staff that have only worked alone:

$$\text{aloneStaff} := \pi_{\text{SID}}\text{Staff} - \text{collaborativeCollectors} - \text{collaborativeMaintainers}$$

- Now we find all the staff who have worked with at least two other people on a collection. We begin by doing a large combination of two Artifact relations (A1, A2) and two artifactsCollected (C1, C2) relations. This gives us four SID columns along with the artifact numbers they worked on. We first ensure that the SIDs in C1 and C2 are the same so that we can check the remaining two SIDs (A1 and A2) are all unique. If C1 or C2, A1, and A2 are not unique, then we eliminate the tuple. We also check that A1, C1, and A2, C2 are working on the same artifact. The remaining tuples contain SIDs of staff who have worked with at least two other people. We get all these unique SIDs.

$$\text{atLeastTwoOthers1} := \pi_{\text{C1.SID}}(\sigma_{\substack{\text{C1.SID = C2.SID}\wedge \\ \text{A1.AN = C1.AN}\wedge \\ \text{A2.AN = C2.AN}\wedge \\ \text{A1.SID}\neq\text{C1.SID}\wedge \\ \text{A2.SID}\neq\text{C2.SID}\wedge \\ \text{A1.SID}\neq\text{A2.SID}}}(\rho_{\text{A1}}\text{Artifact} \times \rho_{\text{C1}}\text{artifactsCollected} \times \rho_{\text{A2}}\text{Artifact} \times \rho_{\text{C2}}\text{artifactsCollected}))$$

$$\text{atLeastTwoOthers2} := \pi_{\text{A1.SID}}(\sigma_{\substack{\text{C1.SID = C2.SID}\wedge \\ \text{A1.AN = C1.AN}\wedge \\ \text{A2.AN = C2.AN}\wedge \\ \text{A1.SID}\neq\text{C1.SID}\wedge \\ \text{A2.SID}\neq\text{C2.SID}\wedge \\ \text{A1.SID}\neq\text{A2.SID}}}(\rho_{\text{A1}}\text{Artifact} \times \rho_{\text{C1}}\text{artifactsCollected} \times \rho_{\text{A2}}\text{Artifact} \times \rho_{\text{C2}}\text{artifactsCollected}))$$

$$\text{atLeastTwoOthers3} := \pi_{\text{A2.SID}}(\sigma_{\substack{\text{C1.SID = C2.SID}\wedge \\ \text{A1.AN = C1.AN}\wedge \\ \text{A2.AN = C2.AN}\wedge \\ \text{A1.SID}\neq\text{C1.SID}\wedge \\ \text{A2.SID}\neq\text{C2.SID}\wedge \\ \text{A1.SID}\neq\text{A2.SID}}}(\rho_{\text{A1}}\text{Artifact} \times \rho_{\text{C1}}\text{artifactsCollected} \times \rho_{\text{A2}}\text{Artifact} \times \rho_{\text{C2}}\text{artifactsCollected}))$$

- We now have all staff who have only worked with themselves and all staff who have worked with at least two other people. We can subtract these from all staff to get the SIDs of those who have worked exclusively with each other.

$$\begin{aligned}\text{exclusiveStaff} := {} & (\pi_{\text{SID}}\text{Staff}) - \text{aloneStaff} \\ & - \text{atLeastTwoOthers1} - \text{atLeastTwoOthers2} - \text{atLeastTwoOthers3}\end{aligned}$$

- Now that we have all exclusive staff SIDs, we need to put them in pairs. We create two copies of a relation to get the artifacts that exclusive staff have worked on.

$$\text{exclusiveStaffArtifacts1} := \text{Artifact} \bowtie \text{exclusiveStaff}$$
$$\text{exclusiveStaffArtifacts2} := \text{Artifact} \bowtie \text{exclusiveStaff}$$

- Finally, we need to find pairs of staff who have worked on the same artifact. Since we have only selected artifacts that exclusive staff have worked on, if two unique people have worked on the same artifact, then they must be an exclusive pair.

$$\begin{aligned}\text{exclusiveStaffPairs} := {} & \pi_{\text{E1.SID, E2.SID}} \\ & (\rho_{\text{E1}}\text{exclusiveStaffArtifacts1} \bowtie_{\substack{\text{E1.SID} \neq \text{E2.SID} \wedge \\ \text{E1.AN = E2.AN}}} \rho_{\text{E2}}\text{exclusiveStaffArtifacts2})\end{aligned}$$

9. Rationale: Track the influence of a given staff member.

   Query: Staff member SID_1 is influenced by staff member SID_2 if (a) they have ever worked together on a collection or (b) if SID_1 has ever worked with a staff member who is influenced by SID_2. Find SIDs of staff members influenced by SID 42.

   - Cannot be expressed

3 Constraints

1. No species is also a genus.

$$\sigma_{\text{species} = \text{genus}}(\text{Species}) = \emptyset$$

2. No genus belongs to more than one family.

$$\rho_{\text{G1}}\text{Genus} \bowtie_{\text{G1.genus} = \text{G2.genus} \wedge \text{G1.family} \neq \text{G2.family}} \rho_{\text{G2}}\text{Genus} = \emptyset$$

3. All publications must be published after all artifacts they use have been collected.

   We make a key assumption: Date is a numeric value with smaller values denoting older dates.

$$\begin{aligned} &\text{jointCID} := \text{Collection} \bowtie \text{Collected} \\ &\text{Published} \bowtie_{\text{jointCID.AN} = \text{Published.AN} \wedge \text{Collection.date} > \text{Published.date}} \text{jointCID} = \emptyset \end{aligned}$$

4. Students may not catalogue live artifacts.

$$\sigma_{\text{type} = \text{‘live’} \wedge \text{rank} = \text{‘student’}}(\text{Artifact} \bowtie \text{Staff}) = \emptyset$$