Conjunctive Queries

Jef Wijsen

April 4, 2025

1 Preliminaries

We assume two disjoint, infinite sets: the set $\mathbf{var} = \{x, y, z, \ldots\}$ of *variables* and the set $\mathbf{dom} = \{a, b, c, \ldots\}$ of *constants*. We define $\mathbf{sym} = \mathbf{var} \cup \mathbf{dom}$, the set of *symbols*. A *substitution* is a mapping $\theta : \mathbf{sym} \to \mathbf{sym}$ such that for every constant a, $\theta(a) = a$. A *valuation* is a substitution θ such that for every variable x, $\theta(x)$ is a constant.

We assume denumerably many *relation names* $R, S, T, \ldots$, each of which has a fixed *arity* (a nonnegative integer). If R is a relation name of arity n, and $s_1, \ldots, s_n$ are symbols, then $R(s_1, \ldots, s_n)$ is an *atom*. If each s_i ($1 \leq i \leq n$) is a constant, then the atom is said to be *ground*. The letters L, H will be used to denote atoms.

A *database schema* $\mathbf{S}$ is a finite set of *relation names*. A *database* I over $\mathbf{S}$ is a finite set of atoms using only the relation names of $\mathbf{S}$. A database is *ground* if it contains only ground atoms.

Valuations and substitutions extend to atoms and databases in the natural way:

$$\begin{aligned} \theta(R(s_1, \ldots, s_n)) &= R(\theta(s_1), \ldots, \theta(s_n)) \ ; \\ \theta(I) &= \{\theta(L) \mid L \in I\} \ . \end{aligned}$$

2 Conjunctive queries

A *rule-based conjunctive query* (or simply *rule*) q over database schema $\mathbf{S}$ is an expression:

$$H \leftarrow B$$

where B is a (usually nonground) database over $\mathbf{S}$, H is a single atom with a fresh relation name not in $\mathbf{S}$, and such that each variable that occurs in H, also occurs in B. The atom H is called the *head* of the rule; B is called the *body*.

Let I be a database over $\mathbf{S}$. The *answer* to q on I, denoted $q(I)$, is defined by:

$$q(I) = \{\theta(H) \mid \theta \text{ is a substitution such that } \theta(B) \subseteq I\} \ .$$

Example 1 Assume a relation name Emp of arity 3, where $\mathsf{Emp}(\mathrm{Ed}, 10\mathrm{K}, \mathrm{UMH})$, for example, means that Ed is employed by UMH, earning 10K. The relation name Loc of arity 2 is used to store company locations, for example, $\mathsf{Loc}(\mathrm{UMH}, \mathrm{Mons})$. The query "Get name and salary of each employee working for a company with a location in Mons," can be formulated as follows:

$$\mathsf{Answer2}(x, y) \leftarrow \{\mathsf{Emp}(x, y, z), \mathsf{Loc}(z, \text{“Mons”})\}$$

The question "Get names of companies located in both Mons and Charleroi," can be formulated as follows:

$$\mathsf{Answer1}(x) \leftarrow \{\mathsf{Loc}(x, \text{“Mons”}), \mathsf{Loc}(x, \text{“Charleroi”})\}$$

3 Query Containment

The following definitions are relative to a fixed database schema. Let q_1 and q_2 be two queries with the same relation name in the head. We say that q_1 is *contained in* q_2, denoted $q_1 \sqsubseteq q_2$, if for every **ground** database I, $q_1(I) \subseteq q_2(I)$.[1] We say that q_1 and q_2 are *equivalent*, denoted $q_1 \equiv q_2$ if $q_1 \sqsubseteq q_2$ and $q_2 \sqsubseteq q_1$.

Let $q_1 : H_1 \leftarrow B_1$ and $q_2 : H_2 \leftarrow B_2$ be two queries with the same relation name in the head. A *homomorphism* from q_2 to q_1 is a substitution θ such that $\theta(B_2) \subseteq B_1$ and $\theta(H_2) = H_1$.

Example 2

$$\begin{array}{lcl} q_1 : \mathsf{Answer1}(x) & \leftarrow & \mathsf{Emp}(x,y,z), \mathsf{Loc}(z, \text{“Mons”}), \mathsf{Loc}(z, \text{“Charleroi”}) \\ q_2 : \mathsf{Answer1}(u) & \leftarrow & \mathsf{Emp}(u,v_1,w_1), \mathsf{Loc}(w_1, \text{“Mons”}), \mathsf{Emp}(u,v_2,w_2), \mathsf{Loc}(w_2, \text{“Charleroi”}) \end{array}$$

Arguably, for every ground database I, $q_1(I) \subseteq q_2(I)$. The mapping $\theta = \{u/x, v_1/y, w_1/z, v_2/y, w_2/z\}$ is a homomorphism from q_2 to q_1.

Theorem 1 (Homomorphism Theorem) *Let q_1 and q_2 be two rules with the same relation name in the head. Then, $q_1 \sqsubseteq q_2$ if and only if there exists a homomorphism from q_2 to q_1.*

Proof. Let $q_1 : H_1 \leftarrow B_1$ and $q_2 : H_2 \leftarrow B_2$.

$\boxed{\Leftarrow}$ Assume a substitution θ such that $\theta(B_2) \subseteq B_1$ and $\theta(H_2) = H_1$. Let I be an arbitrary ground database and $L \in q_1(I)$. Then, there exists a valuation ν such that $\nu(B_1) \subseteq I$ and $\nu(H_1) = L$. Then, $\nu \circ \theta(B_2) \subseteq \nu(B_1) \subseteq I$ and $\nu \circ \theta(H_2) = \nu(H_1) = L$.[2] It follows $L \in q_2(I)$.

$\boxed{\Rightarrow}$ Assume $q_1 \sqsubseteq q_2$. Let ν be a valuation mapping each variable in B_1 to a new fresh constant not occurring elsewhere. Since ν is injective, the inverse mapping ν^{-1} is well-defined. Let $I = \nu(B_1)$, and $L = \nu(H_1)$. Obviously, $L \in q_1(I)$. Since $q_1 \sqsubseteq q_2$, $L \in q_2(I)$. Then, there exists a valuation θ such that $\theta(B_2) \subseteq I$ and $\theta(H_2) = L$. Then, $\nu^{-1} \circ \theta(B_2) \subseteq B_1$ and $\nu^{-1} \circ \theta(H_2) = H_1$. Hence, $\nu^{-1} \circ \theta$ is a homomorphism from q_2 to q_1. $\square$

Corollary 1 *Let $q_1 : H_1 \leftarrow B_1$ and $q_2 : H_2 \leftarrow B_2$ be two rules with the same relation name in the head. Then, $q_1 \sqsubseteq q_2$ if and only if $H_1 \in q_2(B_1)$.*

Corollary 2 *Two rules q_1 and q_2 with the same relation name in the head are equivalent if and only if there are homomorphisms from q_1 to q_2 and from q_2 to q_1.*

4 Query Optimization by Rule Minimization

We say that a rule $q_1 : H_1 \leftarrow B_1$ is *minimal* if there is no equivalent rule $q_2 : H_2 \leftarrow B_2$ such that $|B_2| < |B_1|$ (it is understood that H_1 and H_2 have the same relation name). Note that minimality is with respect to cardinality.

Theorem 2 *Let $q_1 : H \leftarrow B_1$ be a rule. Then, there exists a subset $B_2 \subseteq B_1$ such that $q_2 : H \leftarrow B_2$ is a minimal rule and $q_2 \equiv q_1$.*

[1] Some textbooks write $q_1 \subseteq q_2$ instead of $q_1 \sqsubseteq q_2$.

[2] $\nu \circ \theta$ is the substitution satisfying for each symbol s, $\nu \circ \theta(s) = \nu(\theta(s))$.

Proof. Let $q_3 : H_3 \leftarrow B_3$ be a minimal rule such that $q_3 \equiv q_1$. By Corollary 1, we can assume a homomorphism θ from q_1 to q_3 and a homomorphism μ from q_3 to q_1. Let $B_2 = \mu(B_3)$ and $q_2 : H \leftarrow B_2$.

We show that $\mu \circ \theta$ is a homomorphism from q_1 to q_2: first, from $\theta(B_1) \subseteq B_3$ and $\mu(B_3) = B_2$, it follows $\mu \circ \theta(B_1) \subseteq B_2$; second, from $\theta(H) = H_3$ and $\mu(H_3) = H$, it follows $\mu \circ \theta(H) = H$. Conversely, the identity substitution is a homomorphism from q_2 to q_1. By Corollary 1, $q_1 \equiv q_2$.

Clearly, $|B_2| \leq |B_3|$. Since q_3 is minimal, $|B_2| = |B_3|$. $\square$

Example 3 From [1]. Let R be a relation name with arity 3. Every satisfiable SPJR query can be translated into an equivalent rule, for example, by an inductive algorithm.

$$\underbrace{\underbrace{\pi_{AB}(\sigma_{B=5}(R))}_{W(x,5)} \bowtie \underbrace{\pi_{BC}(\underbrace{\underbrace{\pi_{AB}(R)}_{T(x,y)} \bowtie \underbrace{\pi_{AC}(\sigma_{B=5}(R))}_{S(x,y)}}_{U(x,y,z)})}_{V(x,y)}}_{\mathsf{Answer3}(x,y,z)}$$

We obtain:

$$\begin{array}{rcl} W(x,5) & \leftarrow & R(x,5,z) \\ T(x,y) & \leftarrow & R(x,y,z) \\ S(x,y) & \leftarrow & R(x,5,y) \\ U(x,y,z) & \leftarrow & T(x,y), S(x,z) \\ V(x,y) & \leftarrow & U(z,x,y) \\ \mathsf{Answer3}(x,y,z) & \leftarrow & W(x,y), V(y,z) \end{array}$$

Hence,

$$\begin{array}{rcl} W(x,5) & \leftarrow & R(x,5,z_1) \\ T(x_1,5) & \leftarrow & R(x_1,5,z_2) \\ S(x_1,z) & \leftarrow & R(x_1,5,z) \\ U(x_1,5,z) & \leftarrow & T(x_1,5), S(x_1,z) \\ V(5,z) & \leftarrow & U(x_1,5,z) \\ \mathsf{Answer3}(x,5,z) & \leftarrow & W(x,5), V(5,z) \end{array}$$

Hence, the SPJR query is equivalent to:

$$\mathsf{Answer3}(x,5,z) \leftarrow R(x,5,z_1), R(x_1,5,z_2), R(x_1,5,z) \ .$$

An equivalent minimal rule is obtained by deleting the second body atom (use the substitution that maps z_2 to z and that is the identity otherwise):

$$\mathsf{Answer3}(x,5,z) \leftarrow R(x,5,z_1), R(x_1,5,z) \ .$$

So the original query is equivalent to:

$$\pi_{AB}(\sigma_{B=5}(R)) \bowtie \pi_{BC}(\sigma_{B=5}(R)) \ .$$

A *variable renaming* μ is a substitution such that whenever x and y are distinct variables, then $\mu(x)$ and $\mu(y)$ are distinct variables. Two rules q_1 and q_2 with the same relation name in the head are *isomorphic* if there exists a variable renaming μ such that $\mu(q_1) = q_2$.

Corollary 3 *Let q_1 and q_2 be* **minimal** *rules with the same relation name in the head such that $q_1 \equiv q_2$. Then, q_1 and q_2 are isomorphic.*

Proof. Left as an exercise. $\square$

5 Unions of Conjunctive Queries

A *union-of-rules* Q is a finite, nonempty set of rules, all with the same relation name in the head. Given a database I, the *answer* $Q(I)$ is defined by $Q(I) = \bigcup_{q \in Q} q(I)$. Query containment and equivalence are defined as before.

Theorem 3 *Let $P = \{p_1, \ldots, p_m\}$ and $Q = \{q_1, \ldots, q_n\}$ be two unions-of-rules, where all rules have the same relation name in the head. Then, $P \sqsubseteq Q$ if and only if for each $i \in \{1, 2, \ldots, m\}$, there exists $j \in \{1, 2, \ldots, n\}$ such that $p_i \sqsubseteq q_j$.*

Proof. $\boxed{\Leftarrow}$ Trivial. $\boxed{\Rightarrow}$ Assume $P \sqsubseteq Q$. We show that $p_1 \sqsubseteq q_j$ for some $j \in \{1, 2, \ldots, n\}$ (the proof for p_i with $i \neq 1$ is analogous). Let $p_1 : H_1 \leftarrow B_1$. Let ν be a valuation mapping distinct variables to new distinct constants not occurring elsewhere. Let $I = \nu(B_1)$ and $L = \nu(H_1)$. Clearly, $L \in P(I)$. Since $P \sqsubseteq Q$, $L \in Q(I)$. Then, we can assume the existence of $j \in \{1, 2, \ldots, n\}$ such that $L \in q_j(I)$. Assume $q_j : G_j \leftarrow A_j$. It follows that there exists a substitution θ such that $\theta(A_j) \subseteq I$ and $\theta(G_j) = L$. Then, $\nu^{-1} \circ \theta$ is a homomorphism from q_j to p_1. By Theorem 1, $p_1 \sqsubseteq q_j$. $\square$

6 Exercises

1. [2] Find all equivalences and containments among the following rules:

$$
\begin{aligned}
q_1 : R(x,y) &\leftarrow S(x,u), S(u,v), S(v,y) \\
q_2 : R(x,y) &\leftarrow S(x,u), S(u,v), S(v,w), S(w,y) \\
q_3 : R(x,y) &\leftarrow S(x,u), S(v,w), S(z,y), S(x,v), S(u,w), S(w,y) \\
q_4 : R(x,y) &\leftarrow S(x,u), S(u,5), S(5,v), S(v,y)
\end{aligned}
$$

   Minimize each rule.

2. Prove Corollary 3.

3. Generalize Corollary 3 for unions-of-rules.

4. Let R be a relation name of arity 3. Minimize the number of joins in

$$\pi_A(\pi_{AB}(R) \bowtie \sigma_{A=B}(\pi_A(R) \bowtie \pi_B(R))) \ .$$

References

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases.* Addison-Wesley, 1995.

[2] J. D. Ullman. *Principles of Database and Knowledge-Base Systems, Volume II.* Computer Science Press, 1989.