

Homework #4,
Information Integration, CSCI 6967-01
Due March 6, 2008 at 2pm

Question 1. You are given the following views for source definitions.

$$v_1(A, B, D) : -p(A, B, C), s(B, C, D), C > 100$$

$$v_2(E, F) : -s(F, F, E), r(E, F), F > 20$$

$$v_3(H, G) : -p(H, I, G), H > I$$

Given the following query:

$$Q1 : ans(X, Y) : -p(X, Y, Z), s(Y, Z, K), r(K, Y), X > Y, Z > 10$$

Does this query have a rewriting using the above views? Find all valid rewritings. Show your work.

Question 2. Suppose you are given the following inclusion dependencies. The primary keys for each relation are underlined.

$$p(\underline{X}, \underline{Y}, Z, W), s(\underline{W}, M, N), r(\underline{X}, \underline{Y}, W, K)$$

and

$$r(X, Y) \subseteq p(X, Y)$$

$$r(W) \subseteq s(W)$$

$$p(W) \subseteq s(W)$$

Suppose we are given the following views from three different sources:

$$v_1(X, Y, Z, W) : -p(X, Y, Z, W)$$

$$v_2(W, M, N) : -s(W, M, N)$$

$$v_3(X, Y, W, K) : -r(X, Y, W, K)$$

Note that the given views may not contain all the tuples in the corresponding relation, i.e. the views may be incomplete. In this case, find maximal way to answer the following query:

$$Q : ans(X, Y, W) := p(X, Y, Z, W), Y > W$$

Question 3. Suppose you are given that the views from Question 2 above have the following completeness descriptions:

$$LC(v_1, p, Y > W)$$

$$LC(v_2, s, p(X, Y, Z, W))$$

$LC(v_3, r, K < 10)$

Note that the second definition says that v_2 is complete with respect to s for all the W values in p .

Which of the following queries are complete with respect to these definitions? Explain why, why not.

$Q1 : ans(X, Y, M) : -p(X, Y, Z, W), s(W, M, N), Y = W$

$Q2 : ans(X, Y, M) : -p(X, Y, Z, W), s(W, M, N), Y > 5, W < 3$

$Q3 : ans(X, Y, M) : -p(X, Y, Z, W), s(W, M, N), r(X, Y, W, K), K = 5$

Question 4. You are given the following entities and relationships between them. Assume this is the database after initial matches for the entities are completed. The relationship $paper(x, y, z)$ denotes that x, y, z wrote a paper together. We assume $paper$ is a relationship with arbitrary arity.

$paper(e1, e3, e7)$

$paper(e2, e4, e8)$

$paper(e1, e5, e7)$

$paper(e1, e9, e10)$

$paper(e2, e10)$

$paper(e4, e7, e8)$

$paper(e5, e7, e8)$

$paper(e7, e8)$

$paper(e12, e13)$

$paper(e12, e14)$

$paper(e13, e9)$

$paper(e13, e10)$

$paper(e14, e3)$

$paper(e14, e4)$

$paper(e14, e10)$

You are going to use the clustering approach to find the two most similar entities and merge them. You also know that entities appearing in the same paper cannot be the same. Continue this process until you found the top 3 matches. You can use one of the following three heuristics in your merge operation:

- $|Nbr(c_1) \cap Nbr(c_2)|$,
- $|Nbr(c_1) \cap Nbr(c_2)| / |Nbr(c_1) \cup Nbr(c_2)|$
- $\sum_{e \in c_1 \cap c_2} u(e) / \sum_{e \in c_1 \cup c_2} u(e)$ where $u(e)$ is $\log(N/E)$ where N is the total number of entities and E is the number of entities attached to E .

Evaluate the precision (total number of correct decisions at top 3) of each heuristic for the following correct associations: $e1 = e2 = e14$, $e3 = e4$, $e6 = e7$. (Clearly, you may need to write a small program to compute this. As I have not tested this yet, I do not know if there will be a difference in the different heuristics.)