INFSCI 2711: Advanced Topics in Database Management

Spring 2024

Homework 3: Distributed Query Processing and Data Linkage

I.

1. **[10pt]** Consider the relations.

   Book (b_id, b_name, b_authorId, b_year)
   Author (a_id, a_name, a_country)
   located at DC and NYC correspondingly. Specify two semi-join strategies to execute the following query depending on the order of relations in the semi-join operation (the result can
   be generated either in NYC or in DC):

   SELECT * FROM Book, Author
   WHERE Book.b_authorId= Author. a_id
   AND Author. a_country ='USA'

   Solution:
   Two semi-join strategies for the given query, which joins `Book` and `Author` tables based on the author ID and filters authors by the country 'USA'.

   Strategy 1: NYC to DC Semi-Join

   - In NYC, select the 'a_id` from the `Author` table where `a_country` is 'USA'.
   - Send these `a_id` values from NYC to DC.
   - In DC, use the received list of `a_id` values to select all entries from the `Book` table where `b_authorId` matches any of the `a_id` in the list.
   - Join the filtered `Book` entries with the `Author` entries based on `b_authorId = a_id` to produce the result in DC.

   Strategy 2: DC to NYC Semi-Join

   - In DC, select the `b_authorId` from the `Book` table.
   - Send these `b_authorId` values from DC to NYC.
   - In NYC, use the received list of `b_authorId` values to select all entries from the `Author` table where `a_id` matches any of the `b_authorId` in the list and `a_country` is 'USA'.
   - Join the filtered `Author` entries with the `Book` entries based on `a_id = b_authorId` to produce the result in NYC.

   The semi-join reduces the amount of data that must be transferred between locations by only sending the necessary keys (`a_id` or `b_authorId`) to perform the join operation on the other side. The direction of the semi-join (NYC to DC or DC to NYC) would typically be chosen based on which operation results in the least amount of data being transferred and hence the least network traffic.

2. **[20pt]** Illustrate the benefit and disadvantages from using each of the semi-join strategies considered in the previous problem with actual *Book* and *Author* tuples. Explain how that benefit depends on **actual values of the *Book* and *Author* tuples (with real examples)**. Provide no more than two sentences of explanation per strategy.

Solution:

Strategy 1 (NYC to DC Semi-Join) Benefits and Disadvantages

Advantage: If the `Author` table has fewer authors from the 'USA', this strategy efficiently filters the data before transferring, thus reducing the volume of data sent from NYC to DC. For example, if there are 1,000 authors but only 10 from the 'USA', only 10 `a_id` values are sent to DC.

Disadvantage: However, if `Author` has many authors from the 'USA', the strategy becomes less efficient as many `a_id` values are sent to DC. If, say, 800 out of 1,000 authors are from the 'USA', this doesn't significantly reduce the data sent to DC.

Strategy 2 (DC to NYC Semi-Join) Benefits and Disadvantages

Advantage: This strategy is beneficial if the `Book` table has many books but a smaller subset of unique author IDs, thus limiting the amount of data sent from DC to NYC. For instance, if the `Book` table has 10,000 books but only includes works from 100 unique authors, only 100 `b_authorId` values are sent to NYC.

Disadvantage: Conversely, if most books have different authors, this approach can result in almost the entire `Book` table's author IDs being sent to NYC. If there are 10,000 books with 9,500 unique authors, then 9,500 `b_authorId` values would be sent, negating the benefit of the semi-join.

II.

1. **[20pt]** Consider a set of candidate pairs C, which is a subset of cross production of X and Y
X = {1, 5, 8, 11} Y = {2, 4, 7, 9}
Compute the cross product of X and Y.

Solution:

The cross product of X and Y ($X \times Y$) = {(1,2), (1,4), (1,7), (1,9), (5,2), (5,4), (5,7), (5,9), (8,2), (8,4), (8,7), (8,9), (11,2), (11,4), (11,7), (11,9)}

2. For the following distance function $(x + y)^2 < 36$, we use following fillers to simplify the computation, try to identify whether it's indicating upper bound or lower bound, and find the corresponding true negative or true positive tuples.
   a) $(2 \times \max(x, y))^2$
   b) 4xy
   c) $2(x^2 + y^2)$

Solution:

a) $(2 \times \max(x, y))^2$

Pairs	$(x + y)^2$	$(x + y)^2 < 36$	$(2 \times \max(x, y))^2$	$(2 \times \max(x, y))^2 > = 36$
(1,2)	9	Yes	16	No
(1,4)	25	Yes	64	No
(1,7)	64	No	196	Yes
(1,9)	100	No	484	Yes
(5,2)	49	No	100	Yes
(5,4)	81	No	100	Yes
(5,7)	144	No	196	Yes
(5,9)	196	No	484	Yes
(8,2)	100	No	256	Yes
(8,4)	144	No	256	Yes
(8,7)	225	No	256	Yes
(8,9)	289	No	484	Yes
(11,2)	169	No	484	Yes
(11,4)	225	No	484	Yes
(11,7)	324	No	484	Yes
(11,9)	400	No	484	Yes

- **Upper Bound**: because $(2 \times \max(x, y))^2 > = (x + y)^2$
- **True Negative Tuples:** Any tuples (x, y) such that $(2 \times \max(x, y))^2 > = 36$ will not satisfy $(x + y)^2 < 36$, so they can be considered true negatives.

b) 4xy

Pairs	$(x + y)^2$	$(x + y)^2 < 36$	4xy	4xy $< = 36$
(1,2)	9	Yes	8	Yes
(1,4)	25	Yes	16	Yes
(1,7)	64	No	28	Yes
(1,9)	100	No	36	Yes
(5,2)	49	No	40	No
(5,4)	81	No	80	No
(5,7)	144	No	140	No
(5,9)	196	No	180	No
(8,2)	100	No	64	No
(8,4)	144	No	128	No
(8,7)	225	No	224	No
(8,9)	289	No	288	No
(11,2)	169	No	88	No
(11,4)	225	No	176	No
(11,7)	324	No	308	No
(11,9)	400	No	396	No

- **Lower Bound:** $4xy < = (x + y)^2$
- **True Positive Tuples**: Any tuples (x, y) such that ($4xy < 36$) and ($x, y > 0$) might satisfy $(x + y)^2 < 36$, so they can be considered potential true positives.

c) . $2(x^2 + y^2)$

Pairs	$(x + y)^2$	$(x + y)^2 < 36$	$2(x^2 + y^2)$	$2(x^2 + y^2) >= 36$
(1,2)	9	Yes	10	No
(1,4)	25	Yes	34	No
(1,7)	64	No	100	Yes
(1,9)	100	No	164	Yes
(5,2)	49	No	58	Yes
(5,4)	81	No	82	Yes
(5,7)	144	No	148	Yes
(5,9)	196	No	212	Yes
(8,2)	100	No	136	Yes
(8,4)	144	No	160	Yes
(8,7)	225	No	226	Yes
(8,9)	289	No	290	Yes
(11,2)	169	No	250	Yes
(11,4)	225	No	274	Yes
(11,7)	324	No	340	Yes
(11,9)	400	No	404	Yes

- **Upper Bound:** $2(x^2 + y^2) > = (x + y)^2$
- **True Negative Tuples:** Any tuples (x, y) such that $2(x^2 + y^2) > = 36$ will not satisfy $(x + y)^2 < 36$, so they can be considered true negatives.

III.

Consider the following relations.

Student:

s_num	s_name	grade
S1	Johns Smith	1
S2	Bill Evasn	2
S3	Daniel Brown	1
S4	Mel Gibson	5

Library:

person_name	Borrow_book
J Smith	Pride and Prejudice
Bill J Evasn	Harry Potter
Daniel Brown	Vampire
Mal Gibson	Little Mermaid

1. Also consider the following query:
   Find the names, grade and borrowed book of all students in the student table.
   For this query:
   [10pt] Specify an SQL expression using equijoin and show the resulting table.

   <u>Solution:</u>

   ```
   SELECT s.s_name, s.grade, l.Borrow_book
   FROM Student s
   JOIN Library l ON s.s_name = l.person_name;
   ```

2. **[20pt]** Consider the similarity join based on edit distance between any pair of c_name in Student and Library table. Please specify string edit distance between any pair of c_name in Student and Library table.

   <u>Solution:</u>

   ```
   SELECT s.s_name, s.grade, l.Borrow_book
   FROM Student s, Library l
   WHERE sed(s.s_name, l.person_name) <= k;
   ```

   a. J Smith to Johns Smith
      - Insert o – 1.
      - Insert h – 2.
      - Insert n – 3.
      - Insert s - 4.

   b. Bill J Evasn to Bill Evasn
      - Delete J – 1.

   c. Daniel Brown to Daniel Brown -0

   d. Mal Gibson to Mel Gibson
      - Delete a – 1.
      - Insert e – 2.

   Result of edit distance join for **k = 4**

3. **[10pt]** Specify an SQL expression and show the resulting table using an approximate join with string edit distance (sed) for the threshold k = 5

Solution:

```
SELECT s.s_name, s.grade, l.Borrow_book
FROM Student s, Library l
WHERE sed(s.s_name, l.person_name) <= 5;
```

s_name	grade	Borrow_book
Johns Smith	1	Pride and Prejudice
Bill Evasn	2	Harry Potter
Daniel Brown	1	Vampire
Mel Gibson	5	Little Mermaid

4. **[10pt]** What is the best threshold value for edit distance so that each name in the student? table have only one match in the library table (show the resulting table for it)?

Solution:

Number of all record pairs, for which the similarity is computed:

Student X Library = 16

Best Threshold Value for edit distance join is **k = 4**

s_name	grade	Borrow_book
Johns Smith	1	Pride and Prejudice
Bill Evasn	2	Harry Potter
Daniel Brown	1	Vampire
Mel Gibson	5	Little Mermaid