# Big Data Analytics

**Homework 5 (Queries)**

Name: - Sreenath Reddy Kurukunda

ID: - 811211580

In this homework, there are 5 questions + 1 bonus question, covering the topic of various queries. If you can answer the bonus question correctly, you can obtain 20 extra points. The maximum mark for this homework is **120 points**, which will be later scaled.

1.Please give the formal definitions of the range query, nearest neighbor query, and top-*k* query. [12 points]  
**ANSWER**:

Range query:

Range Query is a common or standard database operation that used to retrieves all the data from the database where it can be upper and lower boundary within a given range R.

nearest neighbor query

Nearest neighbor query is known as a form of proximity search, is the optimization problem of finding the point in each set that is closest to a given point. In a spatial database, the nearest neighbor query returns all data items that are close to the query point q.

top-*k* query

The query is used to obtain the number of top data objects that we require. Consider a database with n data objects and a scoring function to rank the data objects in the database. The top-k query returns the top k data objects as ranked by f.

2. **(The Range Query)** Given a query range centered at *q* with radius *r* and an MBR node *e*,

2(a). Please provide the pruning condition (i.e., the condition to prune MBR node *e*) for the range query. [10 points]

**ANSWER**:

The distance between the pruning object and the query point, a range query centred on q and with radius r can be reduced. d can be used to find the distance between two objects (x, y).

The condition [d (Q, q) > r(Q) + r(q)] must be satisfied for the objects to prune without any errors.

2(b). Please formally prove the pruning condition in 2(a). [10 points]

Answer:

MBR e = ([e1min, e1max], [e2min, e2max ],…, [edmin, edmax ])

range query Q = ([q1min, q1max], [q2min, q2max ],…, [qdmin, qdmax ])

rectangle R

Bool isOverlapping = true;

For each dimension R

{

If([ejmin, ejmax]) and [qjmin, qimax] do not overlap with each other)

{

isOverlapping = false;

break;

}

}

Return isOverlapping;

3. **(Reverse Nearest Neighbor Query)** Please read Sections 1-3 of the following paper.

Yufei Tao, Dimitris Papadias, and Xiang Lian. Reverse *k*NN Search in Arbitrary Dimensionality. In *Proceedings of the Very Large Data Bases Conference* (VLDB'04), pages 744-755, Toronto, Canada, Aug. 30-Sept. 3, 2004. *Located in the Library Course Reserves on the left-hand course menu.*

3(a). Please give the formal definition of the *reverse nearest neighbor* (RNN) query [6 points]

Answer:

For a dataset D and a point q, the reverse closest neighbor query returns all data objects with q as their nearest neighbor.

For every point p in the database D,

RNN(q) = {p d | such that dsit (p, p`) < dist (p, q)}

Where, dist = distance metric

3(b). Please provide the pruning condition for the TPL approach, and prove its correctness. [12 points]

Answer:

The TLP method is most used in reverse nearest neighbor situations to eliminate false alarms (RNN).

Consider the following example: a spatial dataset with a query point q and an RNN candidate data object c. Join the data objects now and cut the line perpendicularly in half. This divides the dataspace into two parts. The pruning region is the half that contains the object c. Any object in the pruning zones with a greater distance than c can be removed.

When several RNN objects are considered, there will be more pruning regions, and all items that fall within the pruning zone can be easily pruned.

4. Please describe the definitions of the dynamic dominance between two points, dynamic skyline query, and reverse skyline query. [12 points]

Answer:

Dynamic dominance between two points: It refers to the fact that the superiority between two points is not fixed and changes constantly. Consider a graph G with edge weights and a set of vertices V, where v1 and v2 are two of the vertices. It is claimed that v1 outperforms v2 if dist(v1, Q) > dist(v2, Q) (v2, Q).

Dynamic skyline search: A dynamic skyline query returns all data objects in the dataspace that are not dynamically dominated by other points in the dataspace. The dynamic skyline query receives the dynamic arguments.

The reverse skyline query retrieves all data objects from the dataspace for which the dynamic skyline query has query point q.

5. **(Keyword Search Query)**

G. Li, B. C. Ooi, J. Feng, J. Wang, and L. Zhou. EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In *SIGMOD*, pages 903-914, 2008. *Located in the Library Course Reserves on the left-hand course menu.*

5(a). Please read Sections 1-2 of the following paper, and describe the definition of the *r*-radius Steiner graph. [8 points]

Answer:

R-radius steiner graph:

Consider the following r-radius steiner graph G and keyword query q: If a node n's keywords appear in q, it is referred to as a content node. A node s is said to be steiner if it is in the path of two content nodes. The r-radius steiner graph is a subgraph of G that contains steiner nodes and edges. The radius of the steiner graph is always smaller than the radius of the graph G.

5(b). Please read Section 4 of the following paper, and describe the ranking function of the *r*-radius Steiner graph. [30 points]

Answer:

The essential components of the r-radius steiner graph's ranking function are the structural relevance and compactness between the content nodes. IP style ranking using IDF:

The relevance of the text in the IR literature is taken into account by this ranking function, which also takes into account the tern frequency (TF), inverse document frequency (IDF), and normalized document length (NDL).

The following function is used to determine the keyword k's and the steiner graph G's document relevancy:

ScoreIR(k, G) =

Structural compactness based db ranking:

The structural compactness score should be higher if the steiner graph is compact, which indicates that it is full. The steiner graph compactness should be followed by the structural compactness and structural relevancy between the content nodes and the input keywords.

You can determine the structural compactness using,

When the r-radius graph G and its corresponding steiner graph SG are given, then

**Bonus Question [20 extra points]**

6. Read Section 7 of the paper in Question 5, and write a summary (short survey) of existing keyword search techniques.

## Submitting Your Assignment

*All work must be your own. Copying other people’s work or from the Internet is a form of plagiarism and will be prosecuted as such.*

You may submit a Microsoft Word (.docx) document as an attachment. If you attach a document for your assignment, be sure to include your name in the text of the document and in the name of the document.

You can submit multiple times and only the last submission attempt will be considered for grading.

* Submissions sent by email will NOT be accepted.