# Big Data Analytics

**Homework 5 (Queries)**

**Submitted to: Ruoming Jin**

**KSU ID: Shiva Kumar Peddapuram (**[**speddapu@kent.edu**](mailto:speddapu@kent.edu)**)**

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:***

* Let us consider a query point as q.
* Basically, range query retrieves all the objects Oi in the database falling into the query range of Q.

***Nearest Neighbor Query:***

* Given a query point q, which has objects O in the spatial database D.
* The Nearest neighbor retrieves all the objects O and a nearest neighbor (NN) query retrieves an object oi, in D that is closed to q. For other objects o’, dist(q,oi)dist(q,o’)

***Top k query:***

* Let us consider a Spatial database D.
* Each of them is categorized by “n” attributes.
* Now with a scoring function f, As per the objects which are ranked in the patial database, and the number of expected results “k”.
* Now a top-k query Q then returns the object K with the highest rank in 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:**

**Pruning Condition:**

* For pruning an object, with a center Q and radius r.
* We need to evaluate the distance between the objects and the query point.
* Distance for them can be determined as d(x, y).
* For d(Q, q)>r(q)+r(Q) now the object (node) in Q can be pruned safely.
* The pruning condition for for processing a range query centered at Q with radius r, and an MBR node “e” is if an MBR node e in the R-tree can be safely pruned. Then we cannot access all the objects inside the node e. this can help us save a lot of computational cost.

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

**Answer:**

From the above question given that:

**Input is as follows:**

MBR Node E= (x1min,x1max,x2min,x2max,… xdmin,xdmax)

**And the rectangle query range is as follows:**

Q= (q1min, q1max, q2min, q2max, …qdmin,qdmax).

Bool overlap=true;

For each dimension j

{

If([ximin,xjmax]) and (qimin,qjmax) do not overlap each other

}

For each dimension j

{

If([xjmin,xjmax]) and [qjmin,qimax] do not overlap with each other)

{

overlap=false;

Break;

}

Return overlap;

**Below is also an image format pseudo-code:**

**Text, letter

Description automatically generated**

**Conclusion and final explanation:**

The above is the example of a pseudo-code that determines whether the two hyper-rectangles overlap or not.

Two rectangles cannot overlap if a dimension exists that divide their expected intervals from one another.

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

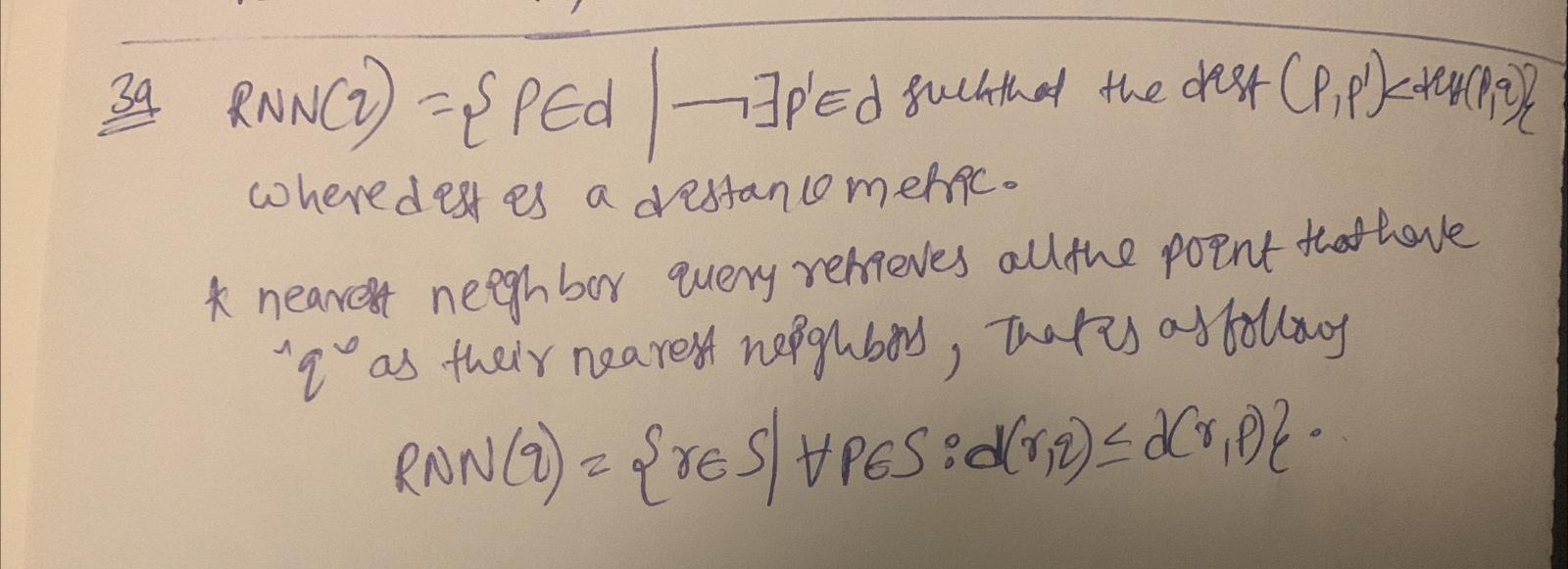
Yufei Tao, Dimitris Papadias, and Xiang Lian. Reverse *k*NN Search in Arbitrary Dimensionality. In *Proceedings of the Very Large Data Bases Conference* (VLDB'04), pages 744-755, Toronto, Canada, Aug. 30-Sept. 3, 2004. *Located in the Library Course Reserves on the left-hand course menu.*

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

**Answer:**

RNN is defined as Reverse closest neighbor (RNN) it obtains all the points that have a point q as their nearest neighbor given as a point q and a multidimensional dataset d.

Every point pd has RNN(q) =



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

**Answer:**

Reverse nearest neighbor (RNN) false alarms are filtered out via TLP pruning, uses a straightforward TPL methods as example.

Chart, scatter chart

Description automatically generated

* Given an RNN candidate 01 and a query point q.
* Between the point Q and o1 we can draw a perpendicular bisector.
* Net the data space has now been split into two half planes. like q(above bisect) and o.
* The RNN candidate containing half plane is now regarded as the pruning region.
* Any object within this pruning zone can be trimmed and it can be demonstrated by that and it cannot be an RNN.
* From the above from the point o2 to o1 is separated by a greater distance than by is o2 and q.
* Though there we have a numerous RNN candidates and pruning regions.

**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:**  if the dominance between two points is not fixed, then it is called as dynamic dominance.

**Definition:**

Given a sizable undirected and edge-weighted graph G and a collection of query vertices Q. U dominated V if and only if dist(u,Q) > dist(v,Q) where dist(V,q) is the shortest distance between v and Q where U and V are two points in Q.

**Dynamic Skyline Query:**

A dynamic skyline query is given as a question for the point q and dataset D, returns all the points that are dynamically not dominated by other points.

**Reverse Skyline query:**

Given that a query point Q, then reverse skyline query retrieves all the point p from P where Q is the dynamic skyline query of p.

***Below is the condition for the Reverse skyline query.***

A piece of paper with writing on it

Description automatically generated

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:in order to define the r-radius steiner graph centric distance and radius must be defined earlier only.

**Centric radius:** in a graph G with any of the node V, the centric radius is denoted as the maximal value among the distance between v and any node u in G.

**Radius:**  in a given graph G, radius is defined as the minimal value among the centric distances of every node in G.

r-Radius steiner Graph: given r-radius graph G and a keyword query k.

If node w directly contains some input keyword in k, then it is called as content node. The radius of an r-radius Steiner graph may be smaller than r but cannot be larger than r

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

**Answer:**

Ranking function for the r-radius Steiner graph is based on the structural relevancy of the input keywords and the compactness between the content nodes.

* The ranking function is based on the structural compactness of content nodes and the structural relevance of input keywords with respect to a Steiner graph with r-radius.
* **TF.IDF-based IR Ranking:**

TF·IDF-based IR-style ranking function weighs an r-radius Steiner graph consider textual relevancy in IR literature, which contain term frequency (tf), inverse document frequency (idf) and normalized document length (ndl). We will combine the three parameters to evaluate the document relevancy between an input keyword ki and a given Steiner graph SG, denoted as ScoreIR(ki, SG)

Text, letter

Description automatically generated

* TF·IDF-based ranking methods are efficient for textual documents, they are inefficient for semi-structured and structured data. From the IR perspective, traditional textual relevancy is important.

***Structural Compactness-based DB Ranking:***

when an r-radius Steiner graph SG is more compact, SG is more likely to be meaningful and relevant. Accordingly, the structural compactness score should be larger.

Text, letter

Description automatically generated

The compactness of SG should include the following parameters:

1. the structural compactness between content nodes in SG, and
2. the structural relevancy between input keywords w.r.t. SG

structural compactness between any two content nodes is defined as

Lemma 4. Given an r-radius graph G and its corresponding Steiner graph SG with respect to a given keyword query, the following equation holds: SimG(ni, nj ) = SimSG(ni, nj ) where ni and nj are any two content nodes in SG.

Text, letter

Description automatically generated

**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.**

Answer:

* Keyword checking for relational databases using Steiner trees was the focus of the paper. To improve online keyword query processing, EASE develops and materializes rich systemic relations.
* DBXplorer, DISCOVER-I, DISCOVER-II, BANKSI, and BANKS-II are relational database management systems. DISCOVER and DBXplorer generate tuple trees that contain all of the input's keywords and are connected by primary-foreign key connections.
* In a labeled graph, BANKS categorize linked trees using an approximation of the Steiner tree problem. DISCOVER-II takes the keyword similarity analysis problem into account in terms of disjunctive semantics as opposed to DISCOV ER-I, which considers only conjunctive semics. Kacholia et al. proposed a bidirectional strategy to increase keyword search performance over graph info.
* However, their approach only works by distinguishing Steiner trees from the whole list, which is inefficient as the recognition of relational relationships by inverted indices is very difficult. IR methods were developed to improve quest efficacy. The sub tree rooted at the lowest common ancestors (LCAs) of information nodes have been suggested as answers in terms of keyword search for XML records.
* XRANK and XSEarch are programs that promote the search for XML documents through keywords and return related sub trees as answers to keyword queries.
* XRANK provides a rating system, where a score is awarded to T with a Page Rank adaptation for XML documents for a specified tree T containing all the keywords.
* XSEarch focuses on semance and performance ranking; it uses an all-pair interconnection index to test the synchronization between nodes during execution.
* XKeyword is a framework providing proximity search for keywords over XML documents that correspond to an XML schema.
* **Reference: https://dbgroup.cs.tsinghua.edu.cn/ligl/papers/SIGMOD2008-EASE.pdf**