# Big Data Analytics

# Homework 5 (Queries)

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]

**Range Query: retrieves all objects falling into a given query range Q. Very useful in applications like location-based services, map services.**

**Nearest Neighbor Query: A nearest neighbor query retrieves an object that is the closest to query point q. An example is you want to go to a specific restaurant closest to your location so you query over all restaurants in the city, and find the closest one to your current location.**

**Top-k query: requires a system able to “rank” objects to find the best answer(s). Searching a database to find the best results based off scoring attributes in the data that best represent end solution.**

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]

**Mindist(q,E) > r**

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

**Mindist(q,E) > r : q=2, e=4, r=6 Mindist(2,4) > 6. This can be safely pruned because MBR node e is outside the query range.**

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]

**RNN Query: Given a point q, a reverse k nearest neighbor query retrieves all the data points that have q as one of their k nearest neighbors. Given a multi-dimensional dataset P and a point q, a reverse nearest neighbor (RNN) query retrieves all the points p∈P that have q as their nearest neighbor.**

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

**The pruning condition used is the concept of half-planes or half-spaces in high dimensions.**

**Mindist(p,N)< dist(p,q): p=2, N=4, q=6: midist(2,4)<dist(2,6): prune successful**

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

**Dynamic dominance between two points: to points, two attributes that are dynamicity relatable between each other that once figured out and keyed together, can provide strong results based off one another, i.e. strong as dominant, or can show they are not dynamically related.**

**Dynamic Skyline Query: given a query object u, a dynamic skyline query obtains all the objects in the database that are not dynamically dominated by other objects *with respect to* u.**

**Reverse Skyline Query: The inverse of dynamic skyline. This is given a query point q, a reverse skyline query retrieves all objects u such that the dynamic skylines of u *include* the 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]

**Given a Graph G and an input keyword query K, the r-radius Steiner Graph finds all the r-radius Steiner Graphs in G, which contain all or a portion of the input keywords in K, ranked by relevancy with K. The radius of r-radius steiner graphs may be smaller than r but cannot be longer than r. Determing the subgraphs that are r-radius graphs by using lemma enables effective retrival of r-radius graphs. A key to effieicency and cost ist o retrieve only the coorespondinng revelent r-radius graph partiotions, Bigger overlap will lead to higher similarity and mroe precise results.**

5(b). Please read Section 4 of the following paper and describe the ranking function of the *r*-radius Stein

er graph. [30 points]

**One rankinng is the TF IDF\_based ranking system. The first stem is to assign each r-radius graph a score using a standard IR-ranking formula and then combine the indivdual scores using a score aggrigation function to obtain a final score of the graphs. Another ranking system is Structural Compactness-based DB ranking. In this ranking, when r-radius Steiner graph is more compact, the graph is more likely to be meaningful and relevant. The compactness of the the structure should return a high score to show a high ranking. Structured DPBFs allow for a higher effiencency and quality of results.**

**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.

**Keyword search over relational databases:**

**Discover and DBXplorer : generate tuples connected through primary-foreeign key relationships that contain all the input keywords.**

**Banks: identifies connected trees in a labeled graph by using a approximation of the Steiner tree problem.**

**Discover I: considers the problem of the keyword proximity search in terms of conjunctive semantics.**

**Discover II: considers the problem of the keyword proximity search in terms of disjunctive semantics.**

**Keyword searches over XML documents**

**XKeyword: system that finds keyword proximity search over XML documents that are in the XML schema.**

**XRank and XSEarch: facilitate keyword search for XML documents and return connected subtrees as answers.**

**As one can see, there are more existing keyword search techniques than one person could even manage to learn themselves. With so many ways to keyword search, it can be done realitcally on almost any platform.**

## 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.