**CS 63016 & CS 73016 Big Data Analytics**

**Homework 3**

**Instructor:** Xiang Lian

**Due Date:** Please refer to the course website

1. Please list two major differences (or improvement) of X-tree from R\*-tree. [10 points]

R\*-tree is an optimized version of R-tree, where coverage and overlap is minimized. R\*-tree allows overlap among MBR index nodes. On the other hand, X-tree prevent overlap among MBRs. In X-tree, if the splitting of MBRs cannot prevent overlap, a super that contains them node will be created.

While increasing dimensionality, the performance of R\*-tree degrades dramatically to the point that it can be even worse than linear scan. That is, in high dimensional space many MBRs are most likely to overlap. X-tree performs better than R\*-tree in the high dimensional space since the overlap among MBRs is minimized.

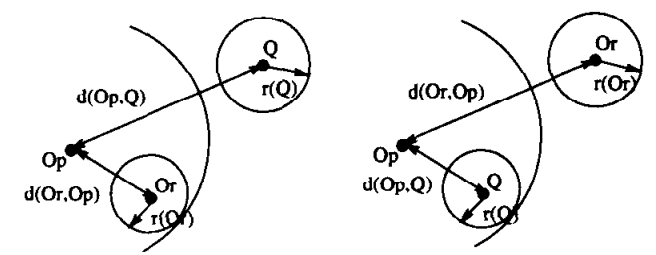
2. Please list at least 2 distance functions that are ***metric distances***, and at least 2 distance functions (or similarity measures) that are ***non-metric distances***. For each distance function, give the reason why it is (or is not) a metric distance, and give references or URL links. (***Hint:*** *please search on the Web or Wikipedia to find the answers*) [20 points].

Distance functions that are metric distance:

Distance functions that are metric distance:

3. Please read the lecture slide of Chapter 3, "Range Queries Over M-Tree", and prove the pruning strategy for the range query below (***Hint****: use the triangle inequality*) [30 points]:

If |*d*(*Op*, *Q*)-*d*(*Or*, *Op*)|>*r*(*Q*)+*r*(*Or*), then *d*(*Or*, *Q*) > *r*(*Q*) + *r*(*Or*) holds and node centered at *Or* with radius *r*(*Or*)can be safely pruned.



4. **(The Curse of Dimensionality)** [40 points]

4a. What is the curse of the dimensionality? Please provide the reason for the dimensionality curse. [10 points]

Overlap in directory increases rapidly with growing dimensionality.

4b. Read Section 2 of the following paper, and write a short survey about existing dimensionality reduction techniques and high dimensional data structures mentioned in this section (*Please cite reference papers in your survey and provide a list of reference papers after the survey. You may need to read abstract or introduction of some reference papers, if they are unclear in the section.* ***Note:*** *please use your own words to describe the techniques;* ***DO NOT*** *copy any sentences from the paper*). [30 points]

H. T. Shen, X. Zhou, and A. Zhou. An adaptive and dynamic dimensionality reduction method for high-dimensional indexing. In *VLDBJ*, 2006. <http://staff.itee.uq.edu.au/zxf/_papers/VLDBJ06.pdf>

**Bonus Question [20 extra points]**

6. Read Section 2.2 of the following paper, as well as the cited papers in this subsection, and write a short survey about the *intrinsic dimensionality*. (***Note:*** *please use your own words to describe the problem definition and solutions;* ***DO NOT*** *copy any sentences from the paper*).

R.F.S. Filho, A. Traina, C. Traina, and C. Faloutsos. Similarity search without tears: the OMNI-family of all-purpose access methods. In *ICDE*, 2001. <http://repository.cmu.edu/cgi/viewcontent.cgi?article=1565&context=compsci>

**Submission**

Submit an electronic copy of your homework solution to the [Blackboard](https://learn.kent.edu/).