# Big Data Analytics

**Homework 3 (X-Tree, M-Tree, and OMNI-Family)  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
Name:- Sreenath Reddy Kurukunda   
ID:- 811211580**

In this homework, there are 4 questions + 1 bonus question, covering the X-tree, M-tree, and OMNI-family. 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 list two major differences (or improvement) of X-tree from R\*-tree. [10 points]  
**ANSWER:**

1) The R\*-tree can perform very good with small data, whereas X-tree can perform efficiently on high dimensional data.

2)  The Insertion of data in the X-tree is faster. The insertion of the data in the R\*-tree is not so fast.

3)   In X-tree, searching the tree with a query is faster but in R\*-tree it is not faster.

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].  
**ANSWER:**

The metric distances are as follows,

**Euclidean distance**:

It is used to measurement the distance between two locations in Euclidean space. It is provided by,

D (p, q) = D(q, p) =

=

*Reference:* [*https://en.wikipedia.org/*](https://en.wikipedia.org/)

**Graph edit distance:** It is used to compare the similarity or dissimilarity of two graphs. The distance between two graphs g1 and g2 can be calculated using the following formula:

GED (g1, g2) =

*Reference:* [*https://en.wikipedia.org/*](https://en.wikipedia.org/)

**Non-metric distances:** Distances that does not follow the triangle inequality are referred to non-metric distances. The following are non-metric distances.

**Quasinorm spaces:** Except for triangular inequality, which employs a new inequality in place of the norms that satisfy the norm axioms. Quasinorm's space is defined as follows:

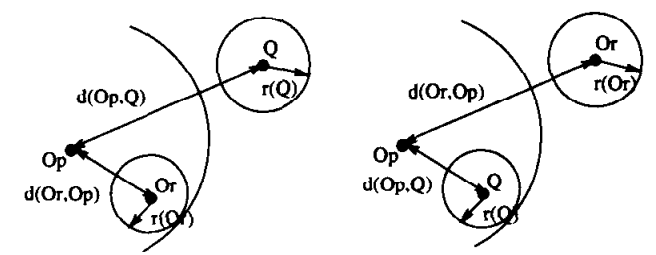
*Reference:* [*https://en.wikipidea.org/*](https://en.wikipidea.org/)

**Kullback-leibler divergence:** It is a metric for determining how two probabilities differ from one another. It does not satisfy the triangle inequality because it is a distribution-wise asymmetric measure. This is how the kullback-leibler divergence is calculated:

*Reference:* [*https://en.wikipidea.org/*](https://en.wikipidea.org/)

3. Please read the lecture slide of Chapter 5, "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.



**ANSWER:**

The triangle inequality as given,

D(x, y) <= D(x, z) + D(z, y)

The given equation to be proved is

If |*d*(*Op*, *Q*)-*d*(*Or*, *Op*)| > *r*(*Q*)+*r*(*Or*) then *d*(*Or*, *Q*) > *r*(*Q*) + *r*(*Or*)

According to the triangle inequality,

d(Or, Op) <= d(Or, Q) + d(Q, Op)

which can be written as,

d(Or, Q) >= d(Or, Op) – d(Q, Op)

d(Or, Op) – d(Q, Op) can be written as,

= Or – Op – Q + Op

= Qr – Q

If d(Or, Op) – d(Q, Op) > r(Qr), then, d(Or, Q) >= d(Or, Op) - d(Q, Op) > r(Qr) + r(Q)

Therefore, using the triangle inequality,

d(Or, Q) > r(Qr) + r(Q) if d(Or, Op) - d(Q, Op) > r(Qr) + r(Q)

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]  
**ANSWER:**

When working with huge or high dimensional data, the curse of dimensionality is a problem. These issues do not arise when working with low-dimensional data, such as in databases and data mining.

The fundamental problem in this situation is that as the data's dimensions increase, so does its volume, making the data less dense or sparse. As the complexity of the data increases.

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. *Located in the Library Course Reserves on the left-hand course menu.***ANSWER:**Both the global dimensionality reduction and the local dimensionality reduction methods are used to reduce dimensionality. In n-dimensional data which is initially reduced to one-dimensional data with the global dimensionality reduction, which is used to optimise the searching time and data accessing cost.

the correlation between the data is utilised to shape the data into multiple clusters which is using local dimensionality reduction, then the multiple clusters are generated and made into in independent clusters. The detection of clusters is difficult since the dimensions are not dependant.

All the clusters are on distinct axes after the dimensionality reduction. To index the different sub-spaces, a global indexing structure with high dimensionality is used.

Data approximation and data transformation are two ways that can be used to index high-dimensional data.

The original points of the data are represented using the reduced representation in the data approximation approach. It's difficult to create vector forms for high-dimensional data.

High-dimensional data is separated into two pyramids using the pyramid technique, and each pyramid is then sliced to a distinct page. This makes it easier to refer to each dimension on its own pyramid page. The iDistance approach, on the other hand, assists in the conversion of high-dimensional data into a single-dimensional distance vector. However, searching for points in all dimensions using the iDistance technique is tough.

**Bonus Question [20 extra points]**

5. 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. *Located in the Library Course Reserves on the left-hand course menu.*

**ANSWER:**

The performance of a smaller database that is a portion of a larger data set is referred to as intrinsic dimensionality.

Hausdorff asserts that datasets with a single standard deviations dimension works better when analysing embedded data with high dimensionality because of embedded data spaces are split into cubic grid cells. The performance of data was enhanced by utilising cubic grid cells. Maximised approximation is a brand-new dimension that Elizaveta and Peter developed to gauge intrinsic dimensionality. The data can be displayed in one dimension if the value of the maximised approximation is low; however, if the value of the maximised approximation is large, the data

## 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.