Metric Spaces

- Elements:
  - ***U***: universe of objects.
  - ***d***: distance function.
- Definition:

Metric Space **(*U, d*).**

- Data set $\boldsymbol{X} \subseteq \boldsymbol{U}$, $|\boldsymbol{X}| = N$.
  - Objective: given $\boldsymbol{q} \in \boldsymbol{U}$, retrieval all similar objects objects to ***q*** in ***X***.

Properties

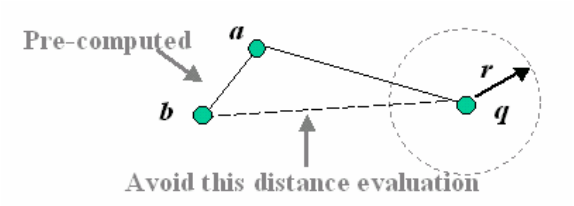
d: $\boldsymbol{U \times U \rightarrow \mathbf{R}^{+}}$ distance function:

- $\forall\ x, y \in \boldsymbol{U},\quad d(a, b) \geq 0$ positiveness
- $\forall\ x \in \boldsymbol{U},\quad d(a, a) = 0$ strictly positiveness
- $\forall\ x, y \in \boldsymbol{U},\quad d(a, b) = d(b, a)$ symmetry
- . $\forall\ x, y, z \in \boldsymbol{U},\ d(a, b) \leq d(a, c) + d(c, b)$

  Triangle inequality

Triangle inequality

- Reduce the number of distance evaluations performed during a range search operation:



$$\text{abs}(\ d(q,a) - d(b,a)\ ) > r$$

HYPOTESIS: DISTANCE EVALUATION ARE EXPENSIVE TO COMPUTE
OBJETIVE: REDUCE THE NUMBER OF DISTANCE EVALUATIONS

Applications

- Information retrieval
- Recognition of images
- Recognition of faces, voice or fingerprint
- Genetic databases
- Medical databases
- etc.

Example

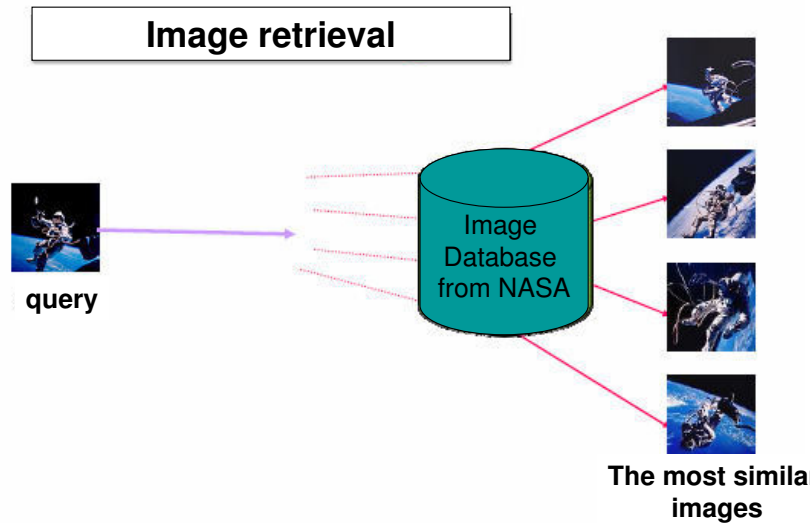
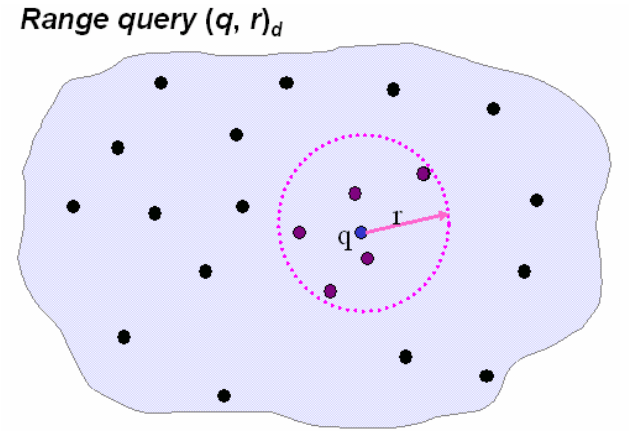


Image retrieval
query
Image Database from NASA
The most similar images

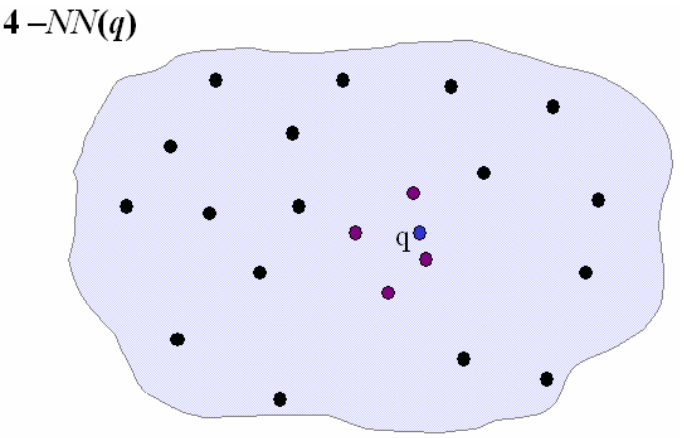
Similarity search

- There are basically two kinds of searches in a Metric Space:
  - Range Search (***q***, ***r***) : retrieval all objects in ***X*** at distance at most r from the query ***q***.
  - ***k-NN*(*q*)** : retrieval the ***k*** most similar objects to a query ***q***.

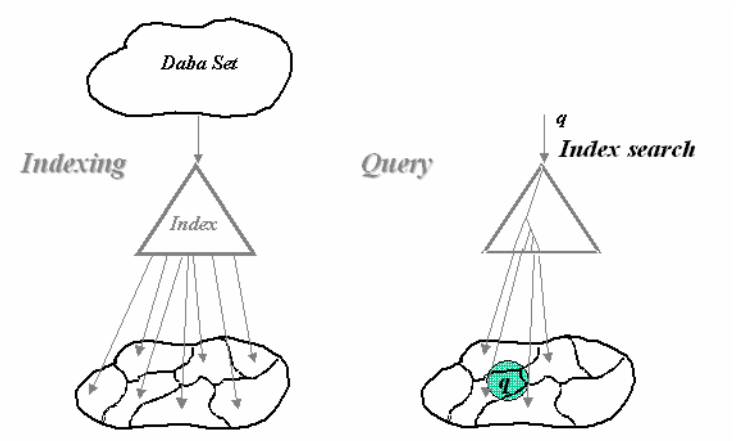
Range Search



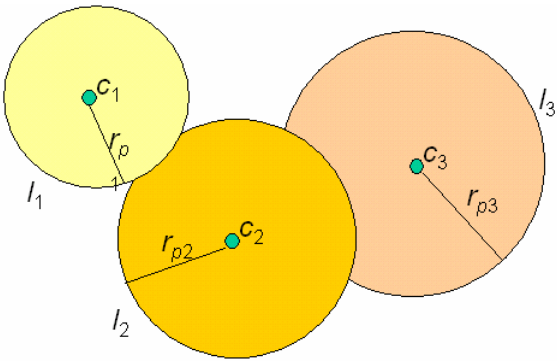
K-NN search

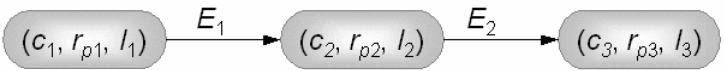


Indexing



List of Clusters (LC)

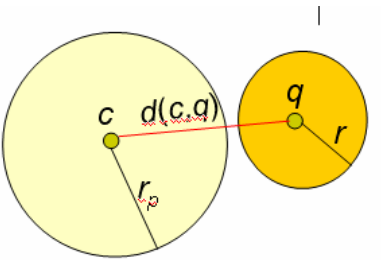


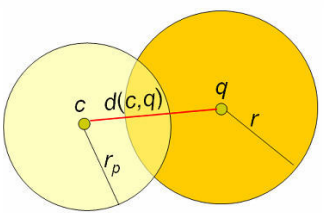


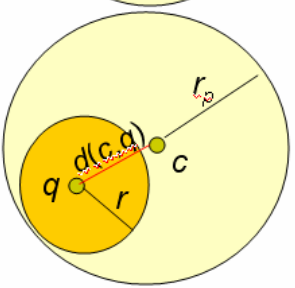
Build(U)
1. If $U = \emptyset$ Then Return an empty list
2. Select a center $c \in U$
3. $I \leftarrow kNN(c)$ in $U - \{c\}$
4. Let be $r_c = \max_{x \in I} d(x, c)$
5. $E \leftarrow U - I$
6. Return (c, r_c, I):**Build**(E)

Searching on the LC

Search(L, q, r)
1. If L is empty Then Return
2. Let $L = (c, r_c, I) : E$
3. Compute the distance $d(c, q)$
4. If $d(c, q) \leq r$ Add c to the set of results
5. If $d(c, q) \leq r_c + r$ Then Search I exhaustively
6. If $d(c, q) > r_c - r$ Then **Search**(E, q, r)







Sparse Spatial Selection - SSS

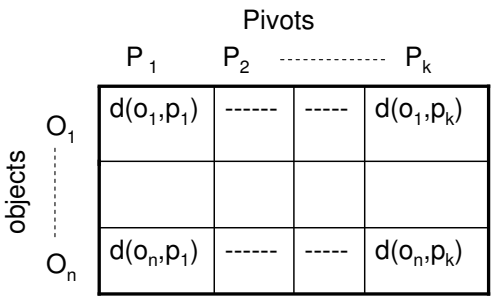
•Maximum distance between two objects in the database

$$M = \max \{d(x, y) / x, y \in X\}$$

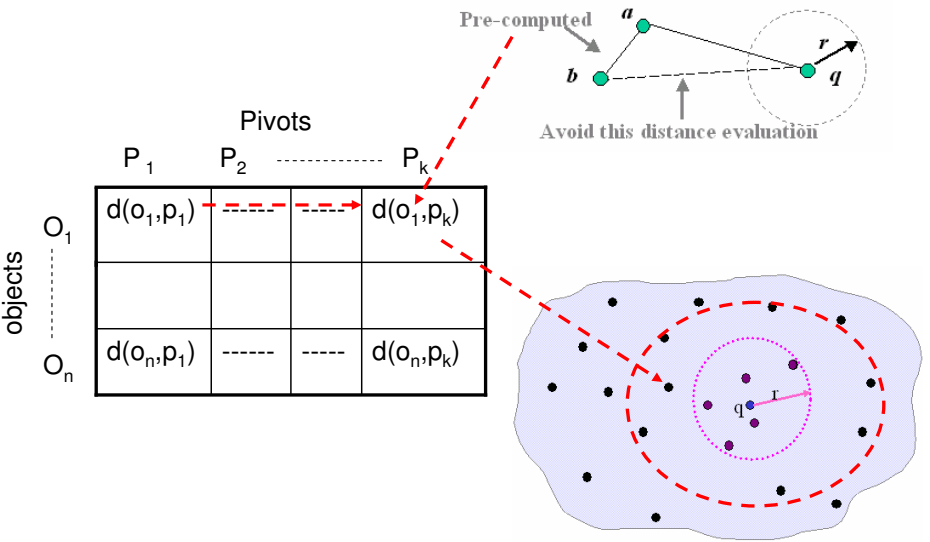
Pivots Selection Stage

PIVOTS ← {x1}
for all xi ∈ U **do**
  if ∀ p ∈ PIVOTS, d(xi, p) >= 0.5**M* **then**
    PIVOTS ← PIVOTS ∪ {xi}

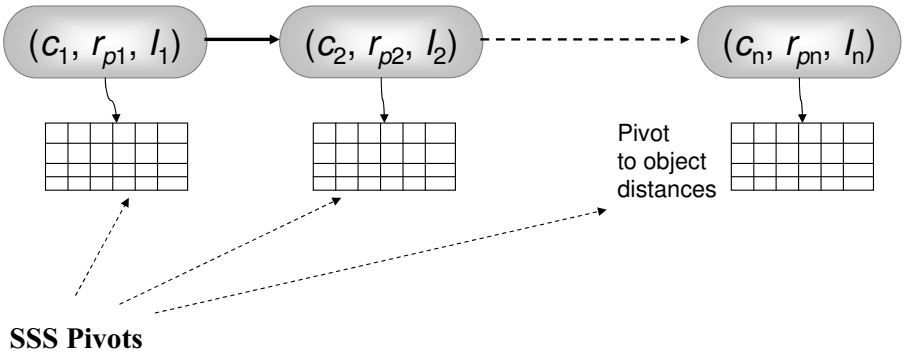
Searching on the SSS



Searching on the SSS



LC-SSS Combination



The same SSS pivots in each table (bucket).

Also the same ordering of pivots in the table columns, the first two are the most distant ones, and so on.