

Subspace Similarity Search

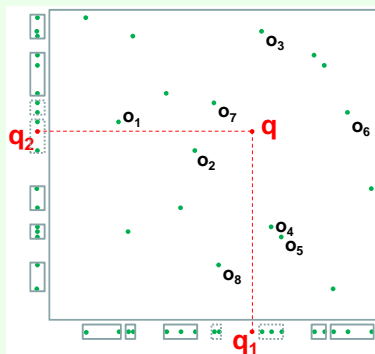
- Application: Database of images represented by color, shape, texture \Rightarrow similarity can be related to the shape only, but not to the color description
- Subspace: $S = (S_1, \dots, S_D) \in \{0,1\}^D$
- Subspace Dimensionality: $d = \sum_{i=1}^D S_i \leq D$
- Subspace Distance: $dist_S(x, y) = \sqrt[p]{\sum_{i=1}^D S_i |x_i - y_i|^p}$
- Subspace k-NN query: $\forall o \in NN(k, S, q), \forall o' \in DB \setminus NN(k, S, q): dist_S(o, q) \leq dist_S(o', q)$
- Ad-hoc definition of an arbitrary subset of attributes

Bottom-Up: Dimension-Merge Index [1]

- Each dimension is organized separately in a one-dimensional modified R*-tree, bounds for each index I_i

$$ub_S(q, o) = \sqrt[p]{\sum_{i=1}^D S_i \cdot \begin{cases} |o_i - q_i|^p, & \text{if } o_i \text{ has been reported} \\ \max(|I_i^{min} - q_i|, |I_i^{max} - q_i|)^p, & \text{if not} \end{cases}}$$

$$lb_S(q, o) = \sqrt[p]{\sum_{i=1}^D S_i \cdot \begin{cases} |o_i - q_i|^p, & \text{if } o_i \text{ has been reported} \\ |I_i^{next} - q_i|^p, & \text{if not} \end{cases}}$$



obj	d1	d2	lb	ub
o ₁	0.5	2.94	9.01	
o ₂	1.0	3.06	9.06	
o ₃	0.4	1.94	8.51	
o ₄	1.0	2.15	8.56	
o ₅	1.5	2.42	8.63	
o ₆	1.0	3.07	9.06	
o ₇	1.9	1.5	2.42	2.42
o ₈	1.7	2.55	8.67	

index bounds	dim	l ^{min}	l ^{max}
1	2.9	9.0	
2	1.9	8.5	
L ₂	3.47	12.4	

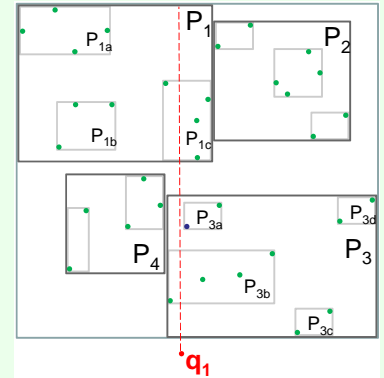
- Refinement: min. distance of a new object is greater than the upper bound of the k-NN distance, where

$$minObjectDist_S(q, I) = \sqrt[p]{\sum_{i=1}^D S_i \cdot mindist(I_i^{next}, q)^p}$$

- Heuristics to choose an appropriate index in each step: Round Robin, *GlobalMinDist* (closest page to q), *MinScore* (distance normalization)

Top-Down: Projected R-tree

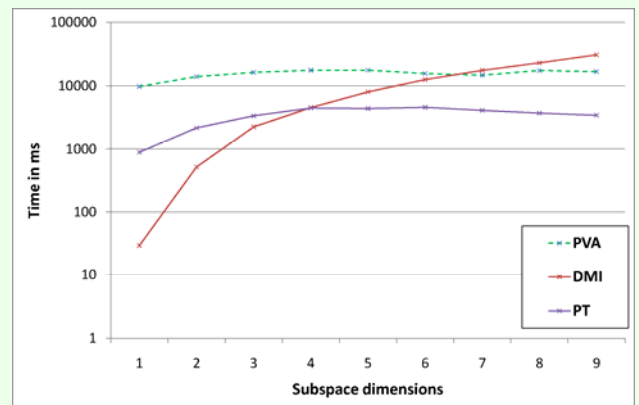
- Single index on the full-dimensional space
- Query processing in a best-first manner
- Based on the minimum distance between an index page P and the query q in a subspace S



$$mindist_S(q, P) = \sqrt[p]{\sum_{i=1}^D S_i \cdot \begin{cases} (P_i^{min} - q_i)^p, & \text{if } P_i^{min} > q_i \\ (q_i - P_i^{max})^p, & \text{if } P_i^{max} < q_i \\ 0, & \text{else} \end{cases}}$$

Evaluation

- The bottom-up approach performs better with a lower-dimensional subspace
- The top-down approach is more appropriate for d approaching D
- Comparison partners, e.g. [2], are outperformed in these regions (CLOUD dataset, D = 9, k = 10)



[1] T. Bernecker, T. Emrich, F. Graf, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, A. Zimek: Subspace Similarity Search Using the Ideas of Ranking and Top-k Retrieval. In Proc. ICDE, DBRank Workshop 2010.

[2] H.-P. Kriegel, R. Kröger, M. Schubert, Z. Zhu: Efficient Query Processing in Arbitrary Subspaces Using Vector Approximations. In Proc. SSDBM 2006.