Motivation

● Similarity search is useful in many areas, such as multimedia retrieval, pattern recognition, and computational biology.
● A generic model is desirable that is capable of accommodating not just a single data type and similarity metric, but a wide spectrum.
● Pivot-based methods outperform compact partitioning methods in terms of CPU costs, but their I/O costs are high because the data is not clustered well.

Contributions

● Develop the SPB-tree, which integrates the compact partitioning with a pivot-based approach, and proposes an efficient pivot selection algorithm to identify a few but effective pivots.
● Present efficient similarity search algorithms and corresponding cost models.
● Conduct extensive experiments with both real and synthetic data sets to demonstrate the performance of the SPB-tree and our proposed algorithms.

Metric Spaces

● A metric space is a tuple \((M, d)\), in which \(M\) is the domain of objects and \(d\) is a distance function which defines the similarity between the objects in \(M\).
● Symmetry: \(d(q, o) = d(o, q)\); non-negativity: \(d(q, o) \geq 0\); identity: \(d(q, o) = 0\) iff \(q = o\); and triangle inequality: \(d(q, o) \leq d(q, p) + d(p, o)\).

SPB-tree Construction

● Map the objects in a metric space to data points in a vector space using well-chosen pivots.
● Map the data points in the vector space into integers in a one-dimensional space using the SFC.
● Employ a B^*-tree with MBB information to index the resulting integers.

SPB-tree Structure

● A pivot table stores selected objects (e.g., \(o_1\) and \(o_2\)) to map a metric space into a vector space.
● A B^*-tree is employed to index the SFC values of objects
  - The minimum SFC value key in its subtree
  - The pointer ptr to the root node of its subtree
  - The SFC values min and max to represent the MBB \(M\)
● A RAF to keep objects separately
  - An object identifier id
  - The length len of the object
  - The real object obj

Pivot Selection Algorithm

● Precision. Given a set \(OP\) of object pairs in a metric space, the quality of a pivot set \(P\) is evaluated as

\[
\text{precision}(P) = \frac{1}{|OP|} \sum_{o1,o2\in OP} \frac{D(\phi(o1), \phi(o2))}{d(o1, o2)}
\]

● HF based incremental pivot selection algorithm (HFI)
  - Employ the HF algorithm [2] to obtain outliers as candidate pivots CP
  - Incrementally select effective pivots \(P\) from CP to maximize precision

Metric Range Query Algorithm

● Metric Range Query: Given an object set \(O\), a query object \(q\), and a search radius \(r\) in \(M\), \(RQ(q, r) = \{o \in O \mid d(q, o) \leq r\}\).

Pruning Rule: Given a pivot set \(P\), if an object \(o\) is enclosed in \(RQ(q, r)\), then \(\phi(o)\) is certainly contained in the mapped range region \(RR(r)\).

Metric kNN Search Algorithm

● Metric kNN Search. Given an object set \(O\), a query object \(q\), and an integer \(k\) in a generic metric space \(M\), \(kNN(q, k) = \{R \subseteq O \mid \forall r \in R, \exists o \in O - R, d(q, r) < d(q, o)\}\).

Pruning Rule. Given a query object \(q\) and a B^*-tree entry \(E\), \(E\) can be safely pruned if \(\text{MIN}(E, q) \geq \text{curND}\).

- \(\text{MIN}(E, q)\) denotes the mapped minimum distance between \(q\) and \(E\)
- \(\text{curND}\) represents the distance from \(q\) to the current \(k\)-th NN

Metric kNN Search Algorithm

- Traverse SPB-tree in a depth-first paradigm to verify objects contained in \(RR(r)\)

Performance

● Comparisons among Pivot Selection Algorithms (Words)

- Number of I/O accesses vs. number of pivots
- CPU cost vs. number of pivots

● Comparisons among Metric Similarity Algorithms (Color)

- Number of I/O accesses vs. number of pivots
- CPU cost vs. number of pivots

References