Technical Report

iDistance: An Adaptive B^+-tree Based Indexing Method for Nearest Neighbor Search

Authors:

H.V. Jagadish
University of Michigan
Beng Chin Ooi
National University of Singapore
Kian-Lee Tan
National University of Singapore
Cui Yu
Monmouth University
Rui Zhang
National University of Singapore

iDistance: An Adaptive B^+-tree Based Indexing Method for Nearest Neighbor Search

H.V. Jagadish
University of Michigan
Beng Chin Ooi
National University of Singapore
Kian-Lee Tan
National University of Singapore
Cui Yu
Monmouth University
Rui Zhang
National University of Singapore

Abstract

In this paper, we present an efficient B^+-tree based indexing method, called iDistance, for K-nearest neighbor (KNN) search in a high-dimensional metric space. iDistance partitions the data based on a space- or data-partitioning strategy, and selects a reference point for each partition. The data points in each partition are transformed into a single dimensional value based on their similarity with respect to the reference point. This allows the points to be indexed using a B^+-tree structure and KNN search to be performed using one-dimensional range search. The choice of partition and reference point adapt the index structure to the data distribution.

We conducted extensive experiments to evaluate the iDistance technique, and report results demonstrating its effectiveness. We also present a cost model for iDistance KNN search, which can be exploited in query optimization.

1 Introduction

Many emerging database applications such as image, time series and scientific databases, manipulate high-dimensional data. In these applications, one of the most frequently used and yet expensive operations is to find objects in the high-dimensional database that are similar to a given query object. Nearest neighbor search is a central requirement in such cases.

There is a long stream of research on solving the nearest neighbor search problem, and a large number of multidimensional indexes have been developed for this purpose. Existing multi-dimensional indexes such as R-trees [Gut84] have been shown to be inefficient even for supporting range queries in high-dimensional databases; however, they form the basis for indexes designed for high-dimensional databases [KS97, WJ96]. To reduce the effect of high dimensionality, use of larger fanouts [BKK96, SYU00], dimensionality reduction techniques [CM00, CM99], and filter-and-refine methods [BEK$^+$98, WSB98] have been proposed. Indexes have also been specifically designed to facilitate metric based query processing [BO97, CPZ97, TSF00, FTF01]. However, linear scan remains an efficient search strategy for similarity search [BGRS99]. This is because there is a tendency for data points to be nearly equidistant to query points in a high-dimensional space. While linear scan is effective in terms of sequential read, every point incurs expensive distance computation, when used for the nearest neighbor problem. For quick response to queries, with some tolerance for errors (i.e., answers may not necessarily be the nearest neighbors), *approximate* nearest neighbor (NN) search indexes such as the P-Sphere tree [GR00] have been proposed. The P-Sphere tree works well on static databases and provides answers with assigned accuracy. It achieves its efficiency by duplicating data points in data clusters based on sample query set. Generally, most of these structures are not *adaptive* to data distributions. Consequently, they tend to perform well for some datasets and poorly for others.

In this paper, we present iDistance, a new technique for KNN search that can be adapted to different data distributions. In our technique, we first partition the data and define a reference point for each partition. Then we index the distance of each data point to the reference point of its partition. Since this distance is a simple scalar, with a small mapping effort to keep partitions distinct, it is possible to use a classical B$^+$-tree to index this distance. As such, it is easy to graft our technique on top of an existing commercial relational database. This is important as most commercial DBMSs today do not support indexes beyond the B$^+$-tree and the R-tree (or one of its variants). The effectiveness of iDistance depends on how the data are partitioned, and how reference points are selected.

For a KNN query centered at q, a range query with radius r is issued. The iDistance KNN search algorithm searches the index from the query point outwards, and for each partition that intersects with the query sphere, a range query is resulted. If the algorithm finds K elements that are closer than r from q at the end of the search, the algorithm terminates. Otherwise, it extends the search radius by Δr, and the search continues to examine the unexplored region in the partitions that intersects with the query sphere. The process is repeated till the stopping condition is satisfied. To facilitate efficient KNN search, we propose partitioning and reference point selection strategies as well as a cost model to estimate the page access cost of iDistance KNN searching.

This paper is an extended version of our earlier paper [YOTJ01]. There, we present the basic iDistance method. Here, we have extended it substantially to include a more detailed discussion of the technique and algorithms, a cost model, and comprehensive experimental studies. In this paper, we conducted a whole new set of experiments using different indexes for comparison such as the Omni-sequential [FTF01] and the M-tree [CPZ97] on both synthetic and real datasets.

The rest of this paper is organized as follows. In the next section, we present the background for metric-based KNN processing, and review some related work. In Section 3, we present the iDistance indexing method and KNN search algorithm, and in Section 4, its space- and data-based partitioning strategies. In Section 5, we present the cost model for estimating the page access cost of iDistance KNN search. We present the performance studies in Section 6, and finally, we conclude in Section 7.

2 Background and Related Work

In this section, we provide the background for metric-based KNN processing, and review related work.

2.1 KNN Query Processing

In our discussion, we assume that DB is a set of points in a d-dimensional data space. A *K-nearest neighbor query* finds the K objects in the database closest in distance to a given query object. More formally, the KNN problem can be defined as follows:

Given a set of points DB in a d-dimensional space DS, and a query point $q \in DS$, find a set S which contains K points in DB such that, for any $p \in S$ and for any $p' \in DB - S$, $dist(q, p) < dist(q, p')$.

Table 1 describes the notation used in this paper.

To search for the K nearest neighbors of a query point q, the distance of the K^{th} nearest neighbor to q defines the minimum radius required for retrieving the complete answer set. Unfortunately, such a distance cannot be predetermined with 100% accuracy. Hence, an iterative

Table 1: Notation

Notation	Meaning
C_{eff}	Average number of points stored in a page
d	Dimensionality of the data space
DB	The dataset
DS	The data space
m	Number of reference points
K	Number of nearest neighbor points required by the query
p	A data point
q	A query point
S	The set containing K NNs
r	Radius of a sphere
$dist\_max_i$	Maximum radius of partition P_i
O_i	The i^{th} reference point
P_i	The i^{th} partition
$dist(p_1, p_2)$	Metric function returns the distance between points p_1 and p_2
$querydist(q)$	Query radius of q
$sphere(q, r)$	Sphere of radius r and center q
$furthest(S, q)$	Function returns the object in S furthest in distance from q

KNN Basic Search Algorithm

```
1.    start with a small search sphere centered at query point
2.    search and check all partitions intersecting the current query space
3.    if $K$ nearest neighbors are found
4.        exit;
5.    else
6.        enlarge search sphere;
7.        goto 2;
end KNN;
```

Figure 1: Basic KNN Algorithm.

approach can be employed (see Figure 1). The search starts with a query sphere about q with a small initial radius, which can be set according to historical records. We maintain a candidate answer set which contains points that could be the K nearest neighbors of q. Then the query sphere is enlarged step by step and the candidate answer set is updated accordingly until we can make sure that the K candidate answers are the true K nearest neighbors of q.

2.2 Related Work

Many multi-dimensional structures have been proposed in the literature, including various KNN algorithms [BBK01]. Here, we briefly describe a few relevant methods.

In [WSB98], the authors describe a simple vector approximation scheme, called VA-file. The VA-file divides the data space into 2^b rectangular cells where b denotes a user specified number of bits. The scheme allocates a unique bit-string of length b for each cell, and approximates data points that fall into a cell by that bit-string. The VA-file itself is simply an array of these compact, geometric approximations. Nearest neighbor searches are performed by scanning the entire approximation file, and by excluding the vast majority of vectors from the search (filtering step) based only on these approximations. After the filtering step, a small set of candidates remain. These candidates are then visited and the actual distances to the query point q are determined. VA-file reduces the number of disk accesses, but it incurs higher computational cost to decode the bit-string, compute all the lower and some upper bounds on the distance to the query point, and determine the actual distances of candidate points. Another problem with the VA-file is that it works well for uniform data, but for skewed data, the pruning effect of the approximation vectors turns worse greatly. The IQ-tree [BBJ$^+$00] extends the notion of the VA-file to use a tree structure where appropriate, and the bit-encoded file structure where appropriate. It inherits many of the benefits and drawbacks of the VA-file discussed above and the M-tree discussed next.

In [CPZ97], the authors proposed the height-balanced M-tree to organize and search large datasets from a generic metric space, i.e., where object proximity is only defined by a distance function satisfying the positivity, symmetry, and triangle inequality postulates. In an M-tree, leaf nodes store all indexed (database) objects, represented by their keys or features, whereas internal nodes store the routing objects. For each routing object O_r, there is an associated pointer, denoted $\mathrm{ptr}(\mathrm{T}(O_r))$, that references the root of a sub-tree, $\mathrm{T}(O_r)$, called the covering tree of O_r. All objects in the covering tree of O_r, are within the distance $\mathrm{r}(O_r)$ from O_r, $\mathrm{r}(O_r) > 0$, which is called the covering radius of O_r. Finally, a routing object O_r, is associated with a distance to $\mathrm{P}(O_r)$, its parent object, that is, the routing object that references the node where the O_r entry is stored. Obviously, this distance is not defined for entries in the root of the M-tree. An entry for a database object O_j in a leaf node is quite similar to that of a routing object, but no covering radius is needed. The strength of M-tree lies in maintaining the pre-computed distance in the index structure. However, the node utilization of the M-tree tends to be low due to its splitting strategy.

Omni-concept was proposed in [FTF01]. The scheme chooses a number of objects from a database as global 'foci' and gauges all other objects based on their distances to each focus. If there are l foci, each object will have l distances to all the foci. These distances are the Omni-coordinates of the object. The Omni-concept is applied in the case that the correlation behaviors of database are known beforehand and the intrinsic dimensionality (d_2) is smaller than the embedded dimensionality d of database. A good number of foci is $\lceil d_2 \rceil + 1$ or $\lceil d_2 \rceil \times 2 + 1$, and they can either be selected or efficiently generated. Omni-trees can be built on top of different indexes such as the B$^+$-tree and the R-tree. Omni B-trees used l B$^+$-trees to index the Omni-coordinates of the objects. When a similarity range query is conducted, on each B$^+$-tree, a set of candidate objects are obtained and intersection of all the l candidate sets will be checked for the final answer. For the KNN query, the query radius is estimated by some selectivity estimation formulas. The Omni-concept improves the performance of similarity search by reducing the number of distance calculations during search operation. However, multiple sets of ordinates for each point increases the page access cost, and searching multiple B-trees (or R-trees) also increases CPU time. Finally, the intersection of the l candidate sets incurs additional cost. In iDistance, only one set of ordinates are used and also only one B$^+$-tree is used to index them, therefore iDistance has less page accesses while still reducing the distance computation. Besides, the choice of reference points in iDistance is quite different from the choice of foci bases in Omni-family techniques.

The P-Sphere tree [GR00] is a two level structure, the root level and leaf level. The root level contains a series of <*sphere descriptor, leaf page pointer*> pairs, while each leaf of the index corresponds to a sphere (we call it the *leaf sphere* in the following) and contains all data points that lie within the sphere described in the corresponding sphere descriptor. The leaf sphere centers are chosen by sampling of the dataset. The NN search algorithm only searches the leaf with the sphere center closest to the query point q. It searches the NN (we denote it as p) of q among the points in this leaf. When finding p, if the query sphere is totally contained in the leaf sphere, then we can confirm that p is the nearest neighbor of q; otherwise, a second best strategy is used (such as sequential scan). A data point can be within multiple leaf spheres, so the points are stored multiple times in the P-Sphere tree. This is how it trades space for time. A variant of the P-Sphere tree is the non-deterministic (ND) P-Sphere tree, which returns answers with some probability of being correct. The ND P-Sphere tree NN search algorithm searches k leaf spheres whose centers are closest to the query point, where k is a given constant (note that this k is different from the K in KNN). A problem arises in high-dimensional space for the deterministic P-Sphere tree search, because the nearest neighbor distance tends to be very large. It is hard for the nearest leaf sphere of q to contain the whole query sphere when finding the NN of q within this sphere. If the leaf sphere does contains the whole query sphere, the radius of the leaf sphere must be very large, typically close to the side length of the data space. In such case, major portion of the whole data set is within this leaf, scanning a leaf is not much different from scanning the whole dataset. Therefore, the authors also hinted that using deterministic P-Sphere trees for medium to high dimensionality is impractical. In [GR00], only the experimental results of ND P-Sphere are reported, which is shown to be better than sequential scan at the cost of space. Again, iDistance only use one set of ordinates and hence has no duplicates. iDistance is meant for high dimensional KNN search where P-Sphere tree can not address efficiently. The ND P-Sphere tree has better performance in high-dimensional space, but our technique, iDistance is looking for exact nearest neighbors.

Another metric based index is the Slim-tree [TSF00], which is a height balanced and dynamic tree structure that grows from the leaves to the root. The structure is fairly similar to that of the M-tree, and the objective of the design is to reduce the overlap between the covering regions in each level of the metric tree. The split algorithm of the Slim-tree is based on the concept of minimal spanning tree [Kru56], and it distributes the objects by cutting the longest line among all the closest connecting lines between objects. If none exits, an uneven split is accepted as a compromise. The slim-down algorithm is a post-processing step applied on an existing Slim-tree to reduce the overlaps between the regions in the tree.

While more indexes have been proposed for high-dimensional databases, other performance speedup methods such as dimensionality reduction have also been performed. The idea of dimensionality reduction is to pick the most important features to represent the data, and an index is built on the reduced space [CM00, FL95, LJF95, Jol86, PKF00]. To answer a query, it is mapped to the reduced space and the index is searched based on the dimensions indexed. The answer set returned contains all the answers and some false positives. In general, dimensionality reduction can be performed on the datasets before they are indexed as a means to reduce the effect of dimensionality curse on the index structure. Dimensionality reduction is lossy in nature, and hence the query accuracy is affected as a result. How much information is lost depends on the specific technique used and on the specific dataset at hand. For instance, Principal Component Analysis (PCA) [Jol86] is a widely used method for transforming points in the original (high-dimensional) space into another (usually lower dimensional) space. Using PCA, most of the information in the original space is condensed into a few dimensions along which the variances in the data distribution are the largest. When the dataset is globally correlated, principal component analysis is an effective method in reducing the number of dimensions with little or no loss of information. However, in practice, the data points tend not to be globally correlated, and the use of global dimensionality reduction may cause a significant loss of information. As an attempt to reduce such loss of information and also to reduce query processing due to false positives, a local dimensionality reduction (LDR) technique was proposed in [CM00]. It exploits local correlations in data points for the purpose of indexing.

3 The iDistance

In this section, we describe a new KNN processing scheme, called iDistance, to facilitate efficient distance-based KNN search. The design of iDistance is motivated by the following observations. First, the (dis)similarity between data points can be derived with reference to a chosen reference or representative point. Second, data points can be ordered based on their distances to a reference point. Third, distance is essentially a single dimensional value. This allows us to represent high-dimensional data in single dimensional space, thereby enabling reuse of existing single dimensional indexes such as the B$^+$-tree. Moreover, false drops can be efficiently filtered without incurring expensive distance computation.

3.1 An Overview

Consider a set of data points DB in a unit d-dimensional metric space DS, which is a set of points with an associated distance function $dist$. Let $p_1 : (x_0, x_1, \ldots, x_{d-1})$, $p_2 : (y_0, y_1, \ldots, y_{d-1})$ and $p_3 : (z_0, z_1, \ldots, z_{d-1})$ be three data points in DS. The distance function $dist$ has the following properties:

$$dist(p_1, p_2) = dist(p_2, p_1) \qquad \forall p_1, p_2 \in DB \tag{1}$$

$$dist(p_1, p_1) = 0 \qquad \forall p_1 \in DB \tag{2}$$

$$0 < dist(p_1, p_2) \qquad \forall p_1, p_2 \in DB; p_1 \neq p_2 \tag{3}$$

$$dist(p_1, p_3) \leq dist(p_1, p_2) + dist(p_2, p_3) \qquad \forall p_1, p_2, p_3 \in DB \tag{4}$$

The last formula defines the triangular inequality, and provides a condition for selecting candidates based on metric relationship. Without loss of generality, we use the Euclidean distance as the distance function in our paper, although other distance functions also apply for iDistance. For Euclidean distance, the distance between p_1 and p_2 is defined as

$$dist(p_1, p_2) = \sqrt{(x_0 - y_0)^2 + (x_1 - y_1)^2 + \cdots + (x_{d-1} - y_{d-1})^2}$$

As in other databases, a high-dimensional database can be split into partitions. Suppose a point, denoted as O_i, is picked as the reference point for a data partition P_i. As we shall see shortly, O_i need not be a data point. A data point, p, in the partition can be referenced via O_i in terms of its distance (or proximity) to it, $dist(O_i, p)$. Using the triangle inequality, it is straightforward to see that
$dist(O_i, q) - dist(p, q) \leq dist(O_i, p) \leq dist(O_i, q) + dist(p, q)$.
When we are working with a search radius of $querydist(q)$, we are interested in finding all points p such that $dist(p, q) \leq querydist(q)$. For every such point p, by adding this inequality to the above one, we must have:
$dist(O_i, q) - querydist(q) \leq dist(O_i, p) \leq dist(O_i, q) + querydist(q)$.
In other words, in partition P_i, we need only examine candidate points p whose distance from the reference point, $dist(O_i, p)$, is bounded by this inequality, which in general specifies an annulus around the reference point.

Let $dist\_max_i$ be the distance between O_i and the point furthest from it in partition P_i. That is, let P_i have a radius of $dist\_max_i$. If $dist(O_i, q) - querydist(q) \leq dist\_max_i$, then P_i has to be searched for NN points, else we can eliminate this partition from consideration altogether. The range to be searched within an affected partition in the single dimensional space is $[dist(0_i, q)$-$querydist(q)$, $min(dist\_max_i$, $dist(O_i, q)$+ $querydist(q))]$. Figure 2 shows an example where the partitions are formed based on data clusters (the data partitioning strategy will be discussed in detail in Section 4.2). Here, for query point q and query radius r, partition P_1 and P_2 need to be searched, while partition P_3 needs not. From the figure, it is clear that all points along a fixed radius have the same value after transformation due to the lossy transformation of data points into distance with respect to the reference points. As such, the shaded regions are the areas that need to be checked.

To facilitate efficient metric-based KNN search, we have identified two important issues that have to be addressed:

1. What index structure can be used to support metric-based similarity search?
2. How should the data space be partitioned, and which point should be picked as the reference point for a partition?

We focus on the first issue here, and will turn to the second issue in the next section. In other words, for this section, we assume that the data space has been partitioned, and the reference point in each partition has been determined.

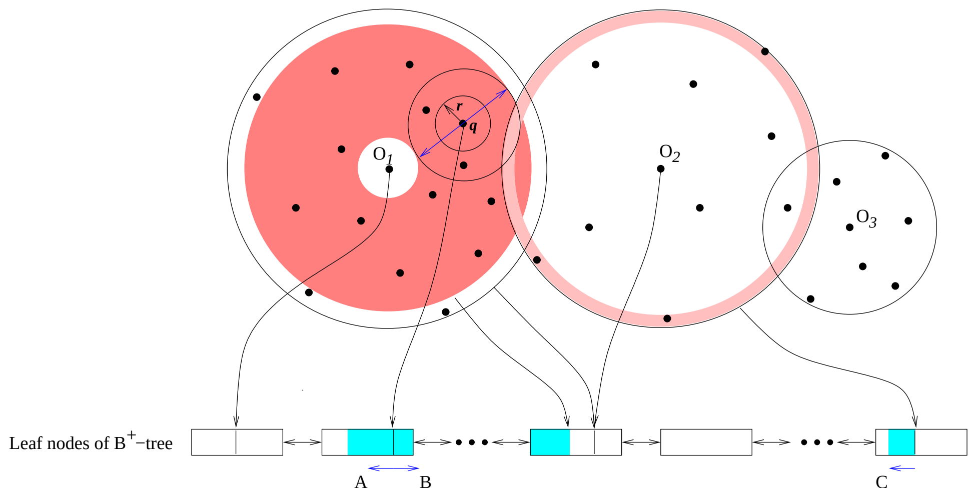


Figure 2: Search regions for NN query q

3.2 The Data Structure

In iDistance, high-dimensional points are transformed into points in a single dimensional space. This is done using a three-step algorithm.

In the first step, the high-dimensional data space is split into a set of partitions. In the second step, a reference point is identified for each partition. Suppose that we have m partitions, $P_0, P_1, \ldots, P_{m-1}$ and their corresponding reference points, $O_0, O_1, \ldots, O_{m-1}$.

Finally, in the third step, all data points are represented in a single dimensional space as follows. A data point $p : (x_0, x_1, \ldots, x_{d-1})$, $0 \leq x_j \leq 1$, $0 \leq j < d$, has an index key, y, based on the distance from the nearest reference point O_i as follows:

$$y = i \times c + dist(p, O_i) \tag{5}$$

where c is a constant used to stretch the data ranges. Essentially, c serves to partition the single dimension space into regions so that all points in partition P_i will be mapped to the range $[i \times c, (i+1) \times c)$. c must be set sufficiently large in order to avoid the overlap between the index key ranges of different partitions. Typically, it should be larger than the length of diagonal in the hypercube data space.

Figure 3 shows a mapping in a 2-dimensional space. Here, O_0, O_1, O_2 and O_3 are the reference points, points A, B, C and D are data points in partitions associated with the reference points and, c_0, c_1, c_2, c_3 and c_4 are range partitioning values which represent the reference points as well. For example O_0 is associated with c_0, and all data points falling in its partition (the shaded region) have their distances relative to c_0. Clearly, iDistance is lossy in the sense that multiple data points in the high-dimensional space may be mapped to the same value in the single dimensional space. That is, different points within a partition that are equidistant from the reference point have the same transformed value. For example, data point C and D have the same mapping value, and as a result, false positives may exist during search.

In iDistance, we employ two data structures:

- A B$^+$-tree is used to index the transformed points to facilitate speedy retrieval. We choose the B$^+$-tree because it is an efficient indexing structure for one-dimensional data and also it is available in most commercial DBMSs. In our implementation of the B$^+$-tree, leaf nodes are linked to both the left and right siblings [RG00]. This is to facilitate searching the neighboring nodes when the search region is gradually enlarged.

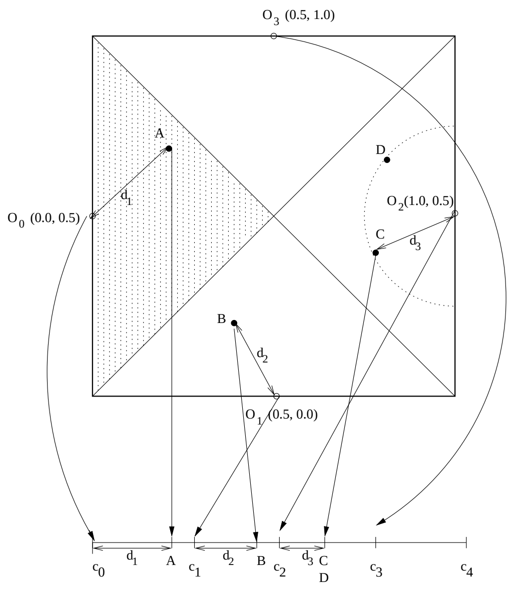


Figure 3: Mapping of data points

- An array is used to store the m data space partitions and their respective reference points. The array is used to determine the data partitions that need to be searched during query processing.

3.3 KNN Search in iDistance

Figures 4–6 summarize the algorithm for KNN search with iDistance method. The essence of the algorithm is similar to the generalized search strategy outlined in Figure 1. It begins by searching a small 'sphere', and incrementally enlarges the search space till all K nearest neighbors are found. The search stops when the distance of the furthest object in S (answer set) from the query point q is less than or equal to the current search radius r.

```
**KNN Search Algorithm iDistanceKNN**($q$, $\Delta r$, $max\_r$)

1.      $r = 0$;
2.      Stopflag = FALSE;
3.      **initialize** $lp[\ ]$, $rp[\ ]$, $oflag[\ ]$;
4.      while Stopflag == FALSE
5.          $r = r + \Delta r$;
6.          **SearchO**($q$, $r$);
end iDistanceKNN;
```

Figure 4: iDistance KNN main search algorithm

Before we explain the main concept of the algorithm *iDistanceKNN*, let us discuss three important routines. Note that both routines SearchInward and SearchOutward are similar, and we shall only explain routine SearchInward. Given a leaf node, routine SearchInward examines the entries of the node towards to the left to determine if they are among the K nearest neighbors, and updates the answers accordingly. We note that because iDistance is lossy, it is possible that points with the same values are actually not close to one another — some may be closer to q,

```
**SearchO**($q$, $r$)

1.   $p_{furthest}$ = furthest(S,q)
2.   if **dist**($p_{furthest}$, $q$) $< r$ **and** $|S| == K$
3.       Stopflag = TRUE;
4.           /* need to continue searching for correctness sake before stop*/
5.   for $i = 0$ to $m - 1$
6.       $dis$ = **dist**($O_i$, $q$);
7.       if not $oflag[i]$ /* if $O_i$ has not been searched before */
8.           if $sphere(O_i, dist\_max_i)$ **contains** $q$
9.               $oflag[i]$ = TRUE;
11.              $lnode$ = **LocateLeaf**($btree$, $i * c + dis$);
12.              $lp[i]$ = **SearchInward**($lnode$, $i * c + dis - r$);
13.              $rp[i]$ = **SearchOutward**($lnode$, $i * c + dis + r$);
14.          else if $sphere(O_i, dist\_max_i)$ **intersects** $sphere(q, r)$
15.              $oflag[i]$ = TRUE;
16.              $lnode$ = **LocateLeaf**($btree$, $dist\_max_i$);
17.              $lp[i]$ = **SearchInward**($lnode$, $i * c + dis - r$);
18.      else
19.          if $lp[i]$ not **nil**
20.              $lp[i]$ = **SearchInward**($lp[i] \rightarrow leftnode$, $i * c + dis - r$);
21.          if $rp[i]$ not **nil**
22.              $rp[i]$ = **SearchOutward**($rp[i] \rightarrow rightnode$, $i * c + dis + r$);
end SearchO;
```

Figure 5: iDistance KNN search algorithm: SearchO

while others are far from it. If the first element (or last element for SearchOutward) of the node is contained in the query sphere, then it is likely that its predecessor with respect to distance from the reference point (or successor for SearchOutward) may also be close to q. As such, the left (or right for SearchOutward) sibling is examined. In other words, SearchInward (SearchOutward) searches the space towards (away from) the reference point of the partition. Lets consider again the example shown in Figure 2. For query point q, the SearchInward search on the partition P_1 will search towards left sibling as shown by the direction of arrow A, while the SearchOutward will search towards right sibling as shown by the direction of arrow B. For partition P_2, we only search towards left sibling by SearchInward as shown by the direction of arrow C. The routine LocateLeaf is a typical B$^+$-tree traversal algorithm which locates a leaf node given a search value, and hence the detailed description of the algorithm is omitted. It locates the leaf node either based on the respective value of q or maximum radius of the partition being searched.

We now explain the search algorithm. Searching in iDistance begins by scanning the auxiliary structure to identify the reference points, O_i whose data spaces intersect the query region. For a partition that needs to be searched, the starting search point must be located. If q is contained inside the data sphere, the iDistance value of q (obtained based on equation 5) is used directly, else $dist\_max_i$ is used. The search starts with a small radius. In our implementation, we just use Δr as the initial search radius. Then the search radius is increased by Δr, step by step, to form a larger query sphere. For each enlargement, there are three cases to consider.

1. The partition contains the query point, q. In this case, we want to traverse the partition sufficiently to determine the K nearest neighbors. This can be done by first locating the leaf node whereby q may be stored (Recall that this node does not necessarily contain points whose distance is closest to q compared to its sibling nodes), and search inward or outward

```
SearchInward(node, ivalue)

1.    for each entry $e$ in $node$ ($e = e_j, j = 1, 2, \ldots, Number\_of\_entries$)
2.        if $|S| == K$
3.            $p_{furthest}$ = furthest(S,q);
4.            if dist($e$, $q$) < dist($p_{furthest}$, $q$)
5.                $S = S - p_{furthest}$;
6.                $S = S \cup e$;
7.        else
8.            $S = S \cup e$;
9.    if $e_1$.key > $ivalue$
10.       $node$ = SearchInward($node \rightarrow leftnode$, $i * c + dis - r$);
11.   if end of partition is reached
12.       $node$ = nil;
13.   return($node$);
end SearchInward;
```

Figure 6: iDistance KNN search algorithm: SearchInward

of the reference point accordingly. For the example shown in Figure 2, only P_1 is examined in the first iteration and q is used to traverse down the B$^+$-tree.

2. The query point is outside the partition but the query sphere intersects the partition. In this case, we only need to search inward. Partition P_2 (with reference point O_2) in Figure 2 is searched inward when the search sphere enlarged by Δr intersects P_2.

3. The partition does not intersect the query sphere. Then, we do not need to examine this partition. An example in point is P_3 of Figure 2.

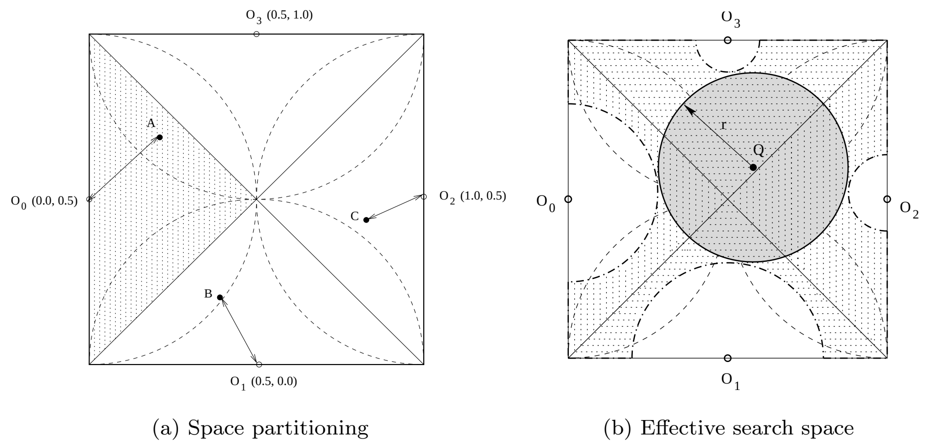
The search stops when the K nearest neighbors have been identified from the data partitions that intersect with the current query sphere and when further enlargement of the query sphere does not change the K nearest list. In other words, all points outside the partitions intersecting with the query sphere will definitely be at a distance D from the query point such that D is greater than *querydist*. This occurs at the end of some iterations and when the distance of the furthest object in the answer set, S, from query point q is less than or equal to the current search radius r. At this time, all the points outside the query sphere has a distance larger than *querydist*, while all candidate points in the answer set have distance smaller than *querydist*. In other words, further enlargement of the query sphere would not change the answer set. Therefore, the answers returned by iDistance are of 100% accuracy.

An interesting by-product of iDistance is that it can provide initial approximate KNN answers quickly. In fact, at each iteration of algorithm iDistanceKNN, we have a set of K candidate NN points. These candidate NN points can be partitioned into two categories: those that are in the final KNN answer set (the true KNN) and those that are not. The points in the latter categories will be replaced by the true KNN in subsequent iterations of the algorithm. As we shall see in our study, it turns out that iDistance can generate a large percentage of real NNs within the first few iterations. This property is attractive for online processing as users may choose to terminate the processing once they find these initial results adequate for decision-making.

4 Selection of Reference Points and Data Space Partitioning

To support distance-based similarity search, we need to split the data space into partitions and for each partition, we need a reference point. In this section we look at some choices. For ease of exposition, we use 2-dimensional diagrams for illustration. However, we note that the complexity of indexing problems in a high-dimensional space is much higher; for instance, the distance between points larger than one (the full normalized range in a single dimension) could still be considered close since points are relatively sparse.

4.1 Space-based Partitioning



(a) Space partitioning (b) Effective search space

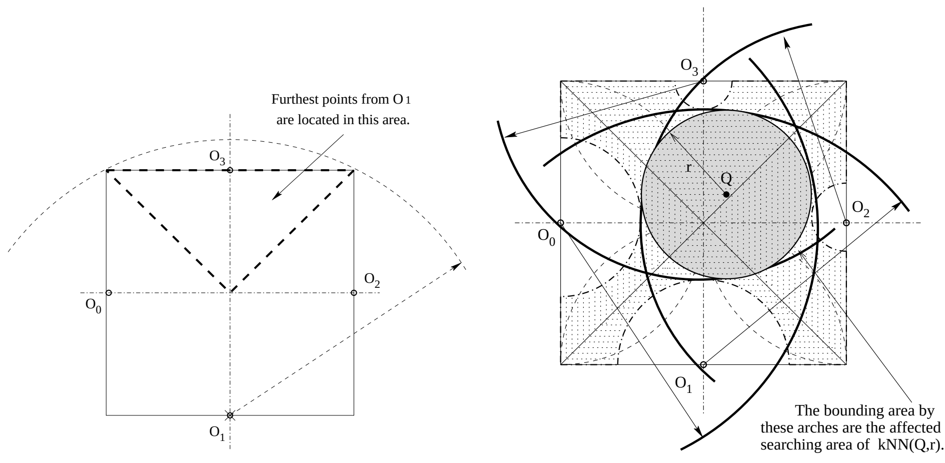
Figure 7: Using (centers of hyperplanes, closest distance) as reference point.

A straightforward approach to data space partitioning is to sub-divide the space into equal partitions. In a d-dimensional space, we have $2d$ hyperplanes. The method we adopted is to partition the space into $2d$ pyramids with the center of the unit cube space as their top, and each hyperplane forming the base of each pyramid.[1] We study the following possible reference points selection and partition strategies.

1. **Center of Hyperplane, Closest Distance.** The center of each hyperplane can be used as a reference point, and the partition associated with the point contains all points that are nearest to it. Figure 7(a) shows an example in a 2-dimensional space. Here, O_0, O_1, O_2 and O_3 are the reference points, and point A is closest to O_0 and so belongs to the partition associated with it (the shaded region). Moreover, as shown, the actual data space is disjoint though the hyperspheres overlap. Figure 7(b) shows an example of a query region, which is the dark shaded area, and the affected space of each pyramid, which is the shaded area bounded by the pyramid boundary and the dashed curve. For each partition, the area not contained by the query sphere does not contain any answers for the query. However, since the mapping is lossy, the corner area outside the query region has to be checked since the data points have the same mapping values as those in the area intersecting with the query region.

   For reference points along the central axis, the partitions look similar to those of the Pyramid tree. When dealing with query and data points, the sets of points are however not exactly identical, due to the curvature of the hypersphere as compared to the partitioning along axial hyperplanes in the case of the Pyramid tree.

[1]We note that the space is similar to that of the Pyramid technique [BBK98]. However, the rationale behind the design and the mapping function are different; in the Pyramid method, a d-dimensional data point is associated with a pyramid based on an attribute value, and is represented as a value away from the center of the space.


Furthest points from O1
are located in this area.
O3
O2
O0
O1
r
Q
O3
O2
O0
O1
The bounding area by
these arches are the affected
searching area of kNN(Q,r).

(a) Space partitioning (b) Effect of reduction on query space

Figure 8: Using (center of hyperplane, furthest distance) as reference point.

2. **Center of Hyperplane, Furthest Distance.** The center of each hyperplane can be used as a reference point, and the partition associated with the point contains all points that are furthest from it. Figure 8(a) shows an example in a 2-dimensional space. Figure 8(b) shows the affected search area for the given query point. The shaded search area is that required by the previous scheme, while the search area caused by the current scheme is bounded by the bold arches. As can be seen in Figure 8(b), the affected search area bounded by the bold arches is now greatly reduced as compared to the closest distance counterpart. We must however note that the query search space is dependent on the choice of reference points, partition strategy and the query point itself.

3. **External Point.** Any point along the line formed by the center of a hyperplane and the center of the corresponding data space can also be used as a reference point.[2] By *external point*, we refer to a reference point that falls outside the data space. This heuristic is expected to perform well when the affected area is quite large, especially when the data are uniformly distributed. We note that both closest and furthest distance can be supported. Figure 9 shows an example of the closest distance scheme for a 2-dimensional space when using external points as reference points. Again, we observe that the affected search space for the same query point is reduced under an external point scheme (compared to using the center of the hyperplane).

4.2 Data-based Partitioning

Equi-partitioning may seem attractive for uniformly distributed data. However, data in real life are often clustered or correlated. Even when no correlation exists in all dimensions, there are usually subsets of data that are locally correlated [CM00, PKF00]. In these cases, a more appropriate partitioning strategy would be used to identify clusters from the data space. There are several existing clustering schemes in the literature such as K-Means [Mac67], BIRCH [ZRL96], CURE [GRS98] and PROCLUS [APW$^+$99]. While our metric based indexing is not dependent on the underlying clustering method, we expect the clustering strategy to have an influence on retrieval

[2]We note that the other two reference points are actually special cases of this.

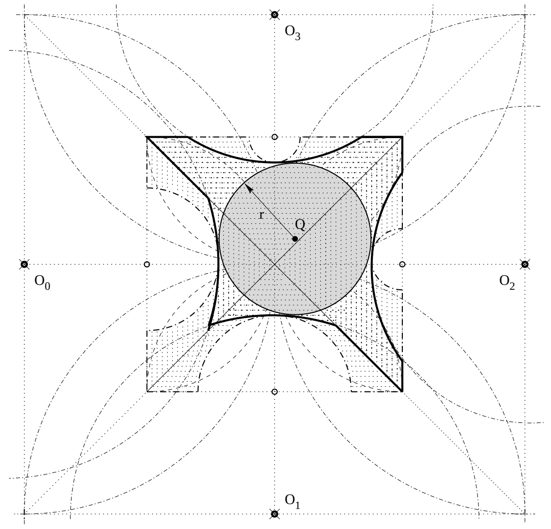


Figure 9: Space partitioning under (external point, closest distance)-based reference point.

performance. In our implementation, we adopted the K-means clustering algorithm [Mac67]. The number of clusters affects the search area and the number of traversals from the root to the leaf nodes. We expect the number of clusters to be a tuning parameter, and may vary for different applications and domains.

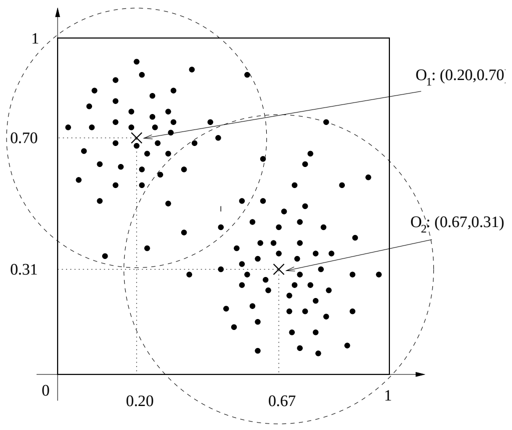


Figure 10: Cluster centers and reference points

Once the clusters are obtained, we need to select the reference points. Again, we have two possible options when selecting reference points:

1. *Center of cluster.* The center of a cluster is a natural candidate as a reference point. Figure 10 shows a 2-dimensional example. Here, we have 2 clusters, one cluster has center O_1 and another has center O_2.

2. *Edge of cluster.* As shown in Figure 10, when the cluster center is used, the sphere area of both clusters have to be enlarged to include outlier points, leading to significant overlap in the data space. To minimize the overlap, we can select points on the edge of the partition as reference points, such as points on hyperplanes, data space corners, data points at one side of a cluster and away from other clusters, and so on. Figure 11 is an example of selecting the

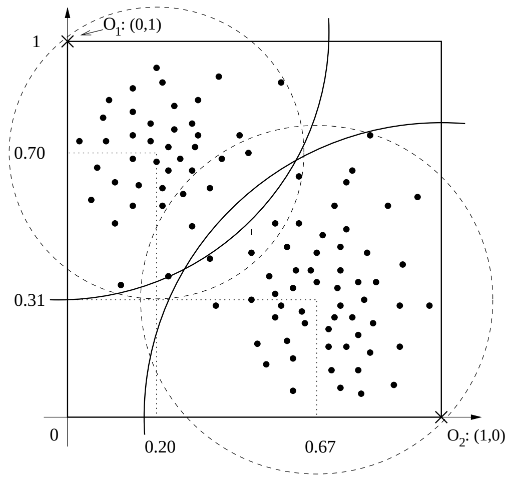


Figure 11: Cluster edge points as reference points

edge points as the reference points in a 2-dimensional data space. There are two clusters and the edge points are $O_1 : (0, 1)$ and $O_2 : (1, 0)$. As shown, the overlap of the two partitions is smaller than that using cluster centers as reference points.

In short, overlap of partitioning spheres can lead to more intersections by the query sphere, and more points having the same similarity (distance) value will cause more data points to be examined if a query region covers that area. Therefore, when we choose a partitioning strategy, it is important to avoid or reduce such partitioning sphere overlap and large number of points with close similarity, as much as possible.

5 Cost Models for iDistance

iDistance is designed to handle KNN search efficiently. However, due to the complexity of very high-dimensionality or the very large K used in the query, iDistance is expected to be superior for certain (but not all) scenarios. We therefore develop cost models to estimate the page access cost of iDistance, which can be used in query optimization (for example, if the iDistance has the number of page accesses less than a certain percentage of that of sequential scan, we would use iDistance instead of sequential scan). In this section, we present two cost models: a cost-effective cost model for uniform datasets, and a cost model for any data distribution but requires maintenance of histograms.

Based on the assumption of uniform data distribution, we derived some formulas to estimate the average page accesses of both the spaced-based and the data-based partitioning schemes. Experiments show that the formula for the space-based partitioning has good accuracy on uniform datasets but not so good for nonuniform datasets. The formula for the data-based partitioning is only satisfactory for limited workloads on uniform datasets. However, these cost models still add values: for the space-based partitioning approaches, the model provides fast estimation when the data are not so skewed; for the data partitioning approaches, the model provides theoretical insight to the schemes and facilitate the determination of an effective number of reference points.

To cope with data skewness, we also estimate the cost of each query based on both the Power-method [TFP03] and a histogram of the key distribution. This histogram-based cost model applies to all partitioning strategies and data distributions, and it predicts individual query processing cost in terms of page accesses instead of average cost. However, the overhead is higher since we need to maintain and retrieve the information used in the Power-method and the histogram. Fortunately, the experiments show that these overhead are much less than the query processing cost.

Table 2: Notation for Cost Model

Notation	**Meaning**
C_{eff}	Average number of points stored in a page
r	Radius of query sphere
N	Number of data points of the data space
N_{pt}	Total number of partitions
N_p	Number of partitions accessed by one query point
O	Reference point of a partition
d	Dimensionality of the data space
R_{max}	Maximum radius of partition
v	for some volume
v_a	average volume
V	total volume
V_s(r)	Function returns the Euclidean volume of the sphere with radius r, i.e. $V_s(r) = \frac{\sqrt{\pi^d}}{\Gamma(\frac{d}{2}+1)} r^d$, where $\Gamma(x+1) = x\Gamma(x), \Gamma(1) = 1, \Gamma(\frac{1}{2}) = \pi^{1/2}$
$dist(p_1, p_2)$	Function returns the distance between points p_1 and p_2

The notation used in the derivation is summarized in Table 2. We assume that the data space is normalized to a unit hypercube. All the data values for each dimensions are normalized to the range [0,1].

5.1 Cost Model for Space-based Partitioning

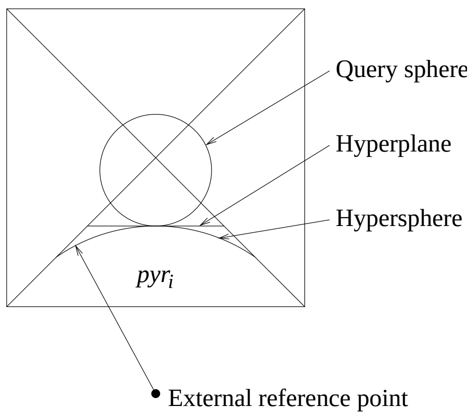


Figure 12: Effect of reference point on the number of data accessed

We assume a uniform distribution of both data points and query points. Figure 12 shows the effect of reference points on the region accessed by a query sphere. When the reference point is outside of the data space, it is called an *external reference point*. In the space-based partitioning strategy, each partition is a pyramid. For the partition pyr_i, the part which is above the surface of the hypersphere as shown in the figure will be searched. The farther the external reference point, the flatter the surface of the hypersphere, and as a result a smaller region is accessed. When the reference point is infinitely far, the surface of the hypersphere becomes a hyperplane, and this is the best case for external reference point. In practice, we can use a point that is far enough, say 10 times the side length of the data space. We estimate the page access cost based on an infinitely far reference point for each partition.

The answer set of a KNN query is contained in a hypersphere. The light grey region in Figure 13

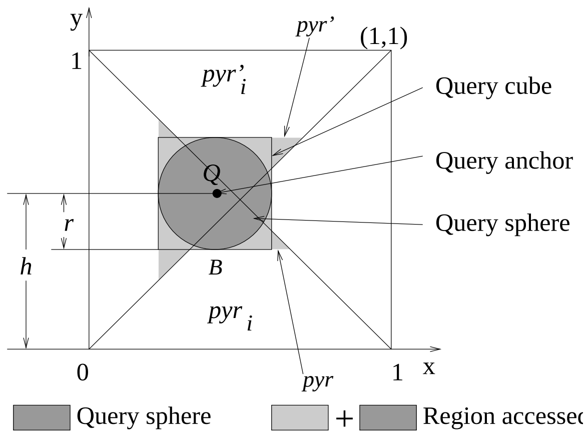


Figure 13: Modeling the page access cost for spaced-based partitioning

is accessed by the query sphere of query radius r. The query point Q is the anchor point of the query sphere. Observe that as long as the bottom of the query sphere, B, is within pyr_i, the region accessed in pyr_i is the same as if the query sphere were a query cube identical to the minimum bounding rectangle of the query sphere. Even if B is outside of pyr_i, as long as it is not very far from pyr_i, the region accessed is still similar. In fact, the query radius of KNN search is typically larger than 0.5, which satisfies the above condition. Therefore we calculate the region accessed by the query cube as an estimation for the region accessed by the query sphere.

The region accessed in pyr_i is also a pyramid pyr with base parallel to the base of pyr_i and similar to pyr_i. Their volume is proportional to h_p^d, where h_p is the height of the pyramid. The volume of the whole data space is 1. Let h be the coordinate of Q in dimension y. Then the height of pyr is $r-(h-0.5)$. The volume of pyr_i is $v_{pyr_i}=\frac{1}{2d}$ and the height of pyr_i is 0.5. Therefore the volume of pyr is

$$v_{pyr}=\left(\frac{r-(h-0.5)}{0.5}\right)^d\cdot v_{pyr_i}=\left(\frac{r-(h-0.5)}{0.5}\right)^d\cdot\frac{1}{2d}$$

Similarly, we can calculate the volume of pyr'

$$v_{pyr'}=\left(\frac{h+r-0.5}{0.5}\right)^d\cdot\frac{1}{2d}$$

The total volume accessed in the pyramid pair pyr_i and pyr'_i is

$$v=v_{pyr}+v_{pyr'}=\left(\frac{r-(h-0.5)}{0.5}\right)^d\cdot\frac{1}{2d}+\left(\frac{h+r-0.5}{0.5}\right)^d\cdot\frac{1}{2d}$$

Q is uniformly distributed in the data space, so h is uniformly distributed in [0,1]. When h is different, the expression of v is different, but the derivation is similar as above. As such, we only list v for different scenarios as follows:

If $0.25\le r\le 0.5$

1. when $0\le h\le 0.5-r$, $v_1=\frac{1}{2d}-\left(\frac{0.5-h-r}{0.5}\right)^d\frac{1}{2d}$
2. when $0.5-r\le h<r$, $v_2=\frac{1}{2d}+\left(\frac{h+r-0.5}{0.5}\right)^d\frac{1}{2d}$
3. when $r\le h<1-r$, $v_3=\left(\frac{h+r-0.5}{0.5}\right)^d\frac{1}{2d}+\left(\frac{0.5-(h-r)}{0.5}\right)^d\frac{1}{2d}$

4. when $1 - r \le h < 0.5 + r$, $v_4 = \frac{1}{2d} + \left(\frac{0.5-(h-r)}{0.5}\right)^d \frac{1}{2d}$

5. when $0.5 + r \le h \le 1$, $v_5 = \frac{1}{2d} + \left(\frac{h-r-0.5}{0.5}\right)^d \frac{1}{2d}$

We can obtain an average of v by integrating over h and then dividing the result by the size of the interval of h, 1. Consequently, the average volume accessed in an opposite pyramid pair is

$$v_a = \int_0^{0.5-r} v_1 dh + \int_{0.5-r}^{r} v_2 dh + \int_r^{1-r} v_3 dh + \int_{1-r}^{0.5+r} v_4 dh + \int_{0.5+r}^{1} v_5 dh$$
$$= (r + \frac{1-(1-\frac{r}{0.5})^{d+1}}{2(d+1)})\frac{1}{d}$$

There are d pyramid pairs in total, and therefore the total volume accessed by the KNN query sphere is

$$v_t = r + \frac{1-(1-\frac{r}{0.5})^{d+1}}{2(d+1)}$$

We can derive v_t for r within other ranges similarly.

If $0 \le r \le 0.25$, we obtain

$$v_t = r + \frac{1-(1-\frac{r}{0.5})^{d+1}}{2(d+1)}$$

If $0.5 \le r \le 1$, we obtain

$$v_t = r + \frac{1-(\frac{r}{0.5}-1)^{d+1}}{2(d+1)}$$

We note that we can combine the above cases for all $0 \le r < 1$ into one equation:

$$v_t = r + \frac{1-|1-\frac{r}{0.5}|^{d+1}}{2(d+1)} \tag{6}$$

When $r > 1$, almost all the data in the data space are accessed. Then the accessed volume is 1.

We store the data points (vectors) in the leaf nodes of the B$^+$-tree, so the number of page accesses equals the number of data points accessed divided by the average number of points stored in a page, C_{eff}. The number of page accesses is given by the following equation

$$E_{spacebased} = v_t \cdot \frac{N}{C_{eff}} \tag{7}$$

where N is the total number of data points in the dataset.

To use Equations (7) to calculate page access cost, we still need to know the query radius. [Boh00] provides a method to estimate the expectation of KNN query radius and we just sketch the method as follows:

The probability that at least k points are inside the volume $V_s(r)$ is

$$P_k(r) = 1 - \sum_{0 \le i < k} C_n^i \cdot V_s^i(r) \cdot (1 - V_s(r))^{n-i} \tag{8}$$

The probability density function p(r) can be derived by differentiation

$$p_k(r) = \frac{\partial P_k(r)}{\partial r}$$

Then the expected value of the k-th NN distance is the following integration

$$R_k = \int_0^{\infty} r \cdot p_k(r) \partial r \tag{9}$$

Note that to calculate $V_s(r)$ in high-dimensional space, boundary effects should be considered. Please refer to [Boh00] for details of these formulas.

5.2 Cost Model for Data-based Partitioning

We also derived a cost model to estimate the page access cost of iDistance indexing based on data-based partitioning. We still assume the data and queries are uniformly distributed in the data space. Because the derivations are quite complex and the model is only accurate in some limited workloads, we do not present them here but show a figure which compares the results from the cost model and from the experiments. Interested readers can read **Appendix A** for the derivations. Figure 14 shows the estimated page accesses and the actual average page accesses

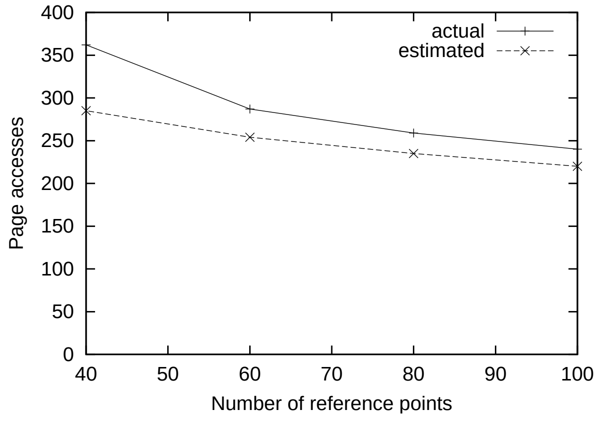


Figure 14: Data-based cost model

of 200 KNN queries on a 8-dimensional uniform dataset as a function of the number of reference points. The relative error of the cost model decrease as the number of reference points increase. This is expected since our cost model is based on probability, where the accuracy increases as the number of events increases. As the number of reference points we use is usually not too large, and results for high-dimensional data have greater errors because of the boundary effects of both the query sphere and the clusters, it is hard to use this model to predict the page access cost accurately for these workloads. However, the cost model is still meaningful in the sense that, together with the experiments, we can see both theoretically and empirically that the page access cost of iDistance decreases as the number of reference points increases; at the same time, the amount of gain decreases as well. This indicates that a reasonably large number of reference points suffices to provide good performance while keeping the clustering cost low. Results of the cost model and experiments for other datasets (different dimensionality, dataset size, etc) show similar trends.

5.3 Histogram-based Cost Model

The above two cost models are formula derivations, which are based on the uniform distribution assumption. They can be easily implemented and computed, but they only work for the uniform datasets and the cost estimation is for average cases, which may be quite far from the cost of an individual query, especially in the case of skewed data. We therefore propose a more practical but costlier approach to estimate the access cost of iDistance, which is based on the Power-method [TFP03] and a histogram of the key distribution. The basic idea of the Power-method is to pre-compute the local power law for a set of representative points and perform the estimation using the local power law of a point close to the query point. In the key distribution histogram, we divide the key values into buckets and maintain the number of points that are in each bucket.

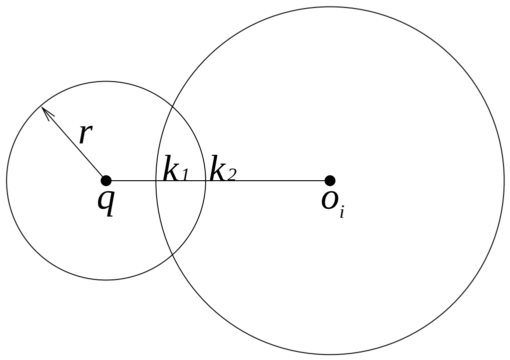


Figure 15: Histogram-based cost model

Figure 15 shows an example of how to estimate the page access cost for a partition P_i, whose reference point is O_i. q is the query point and r is the query radius. k_1 is on the line qO_i and with the largest key in the partition P_i. k_2 is the intersection of the query sphere and the line qO_i. First, we use the Power-method to estimate the K^{th} nearest neighbor distance r, which equals the query radius when the search terminates. Then we can calculate the key of k_2, $|qO_i| - r + i \cdot c$, where i is the partition number and c is the constant to stretch the key values. Since we know the boundary information of each partition and hence the key of k_1, we know the range of the key accessed in partition P_i, that is, between the keys of k_2 and k_1. By checking the key distribution histogram, we know the number of points accessed in this key range, $N_{a,i}$, and then number of pages accessed in the partition is $\lceil N_{a,i}/C_{eff} \rceil$. The summation of the number of page accesses of all the partitions provides us the number of page accesses for the query.

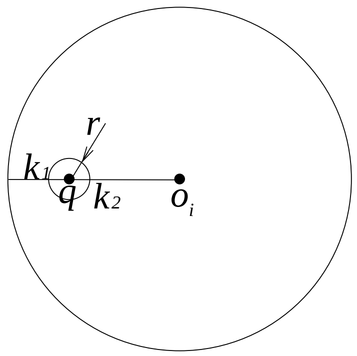


Figure 16: Histogram-based cost model, query sphere inside the partition

Note that, if the query sphere is inside a partition as shown in Figure 16, both k_1 and k_2 are intersections of the query sphere and the line qO_i. Different from the above case, the key of k_1 is $|qO_i| + r + i \cdot c$ here. The number of pages accesses is derived in the same way as above.

The histogram-based cost model takes more effort to implement and incurs higher computational cost, but it overcomes all the drawbacks of the formula-based cost models. It can estimate the page access cost of each individual query and work for any data distribution. The different methods of choosing the reference points do not affect this cost model, either.

6 A Performance Study

In this section, we present results of an experimental study performed to evaluate iDistance. First we compare the space-based partitioning strategy and the data-based partitioning strategy and find that the data-based partitioning strategy is much better. Then we focus our study on the behavior of iDistance using the data-based partitioning strategy with various parameters and under different workloads. At last we compare iDistance with other metric based indexing methods, M-tree and Omni-sequential. We have also evaluated iDistance against iMinMax [OTYB00] and A-tree [SYU00], and our results which have been reported in [YOTJ01] showed the superiority of iDistance over these schemes. As such, we shall not duplicate the latter results here.

We implemented the iDistance technique and associated search algorithms in C, and used the B^+-tree as the single dimensional index structure. Each index page is 4096 Byte. All the experiments are performed on a computer with Pentium(R) 1.6GHz CPU and 256MB RAM. We conducted many experiments using various datasets. Each result we show was obtained as the average (number of page accesses or total response time) over 200 queries which follow the same distribution of the data.

In the experiment, we generated 8, 16, 30-dimensional uniform and clustered datasets. The dataset size ranges from 100,000 to 500,000 data points. For the clustered datasets, the default number of clusters is 20. The cluster centers are randomly generated and in each cluster, the data follow the normal distribution with the default variance of 0.05. Figure 17 shows a 2-dimensional image of the data distribution.

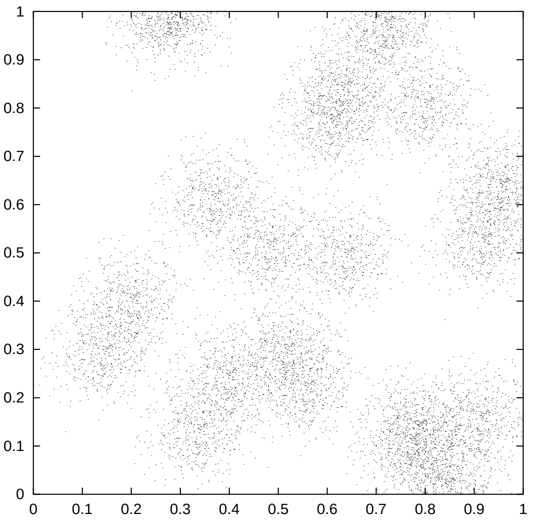
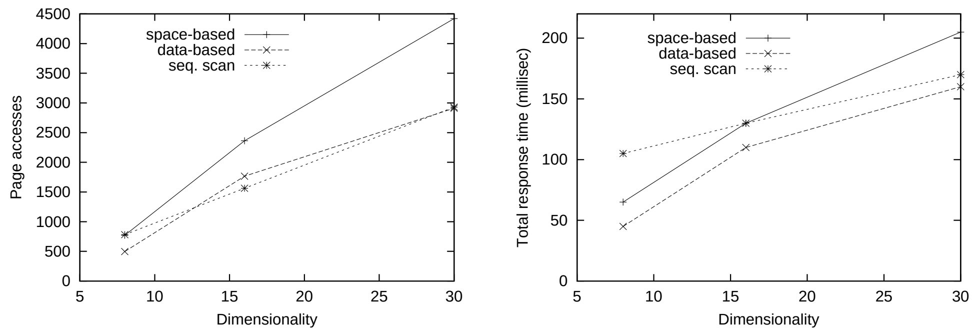


Figure 17: Distribution of the clustered data

We also used a real dataset, the Color Histogram dataset. This dataset is obtained from http://kdd.ics.uci.edu/databases/CorelFeatures/CorelFeatures.data.html. It contains image features extracted from a Corel image collection. HSV color space is divided into 32 subspaces (32 colors:



(a) Page accesses (b) Total response time

Figure 18: Space-based partitioning vs. data-based partitioning, uniform data

8 ranges of H and 4 ranges of S). And the value in each dimension in a Color Histogram of an image is the density of each color in the entire image. The number of records is 68,040. All the data values of each dimension are normalized to the range [0,1].

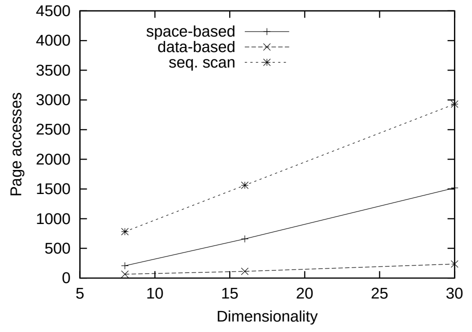
In our evaluation, we use the number of page accesses and the total response time as the performance metric. Default value of Δr is 0.01, that is, 1% of the side length of the data space. The initial search radius is just set as Δr.

6.1 Comparing Space-based and Data-based Partitioning Strategy

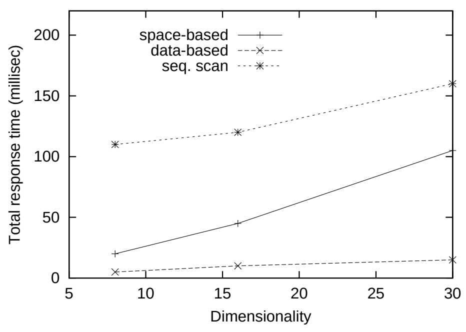
We begin by investigating the relative performance of the partitioning strategies. Note that the number of reference points is always $2d$ for the space-based partitioning approach, and for a fair comparison, we should therefore also use $2d$ reference points in the data-based partitioning approach. Figure 18 shows the result of 10NN queries on the 100,000 uniform dataset. The space-based partitioning has almost the same page accesses as sequential scan when dimensionality is 8 and more page accesses than sequential scan in high dimensionality. The data-based partitioning strategy has fewer page accesses than sequential scan when dimensionality is 8, more page accesses when dimensionality is 16 and almost the same page accesses when dimensionality is 30. This is because the pruning effect of the data-based strategy is better in low dimensionality than in high dimensionality. The relative decrease (compared to sequential scan) of page accesses when dimensionality is 30 is because of the larger number of reference points. While iDistance's page accesses performance is not attractive relative to sequential scan, the total response time performance is better because of its ability to filter data using a single dimensional key. The total response time of the space-based partitioning is about 60% that of sequential scan when dimensionality is 8, same as sequential scan when dimensionality is 16 but worse than sequential scan when dimensionality is 30. The total response time of the data-based partitioning is always less than both the others, while its difference from the sequential scan decreases as dimensionality increases.

Figure 19 shows the result of 10NN queries on the 100,000 clustered dataset. Both partitioning strategies are better than sequential scan in both page accesses and total response time. This is because for clustered data, the K^{th} nearest neighbor distance is much smaller than that in the uniform data. In this case, iDistance can prune a lot of data points in the searching. The total response time of the space-based partitioning is about 20% that of sequential scan. The total response time of data-based partitioning is less than 10% that of sequential scan. Again, the data-based partitioning is better than both the others.

In Section 4.1, we have discussed using external point as the reference points of the space-based partitioning. A comparison between using external points and the center point as the reference

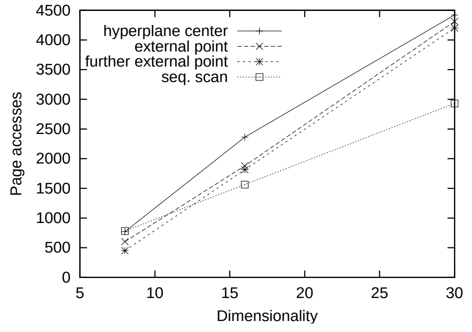


(a) Page accesses

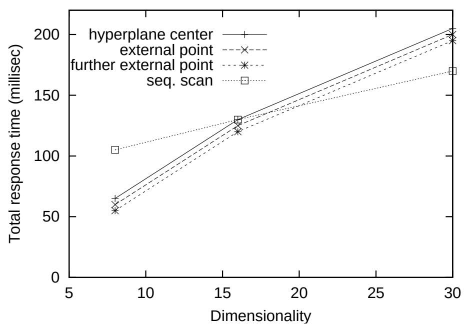


(b) Total response time

Figure 19: Space-based partitioning vs. data-based partitioning, clustered data



(a) Page accesses



(b) Total response time

Figure 20: Effect of reference points in space-based partitioning, uniform data

point on the uniform datasets is shown in Figure 20.

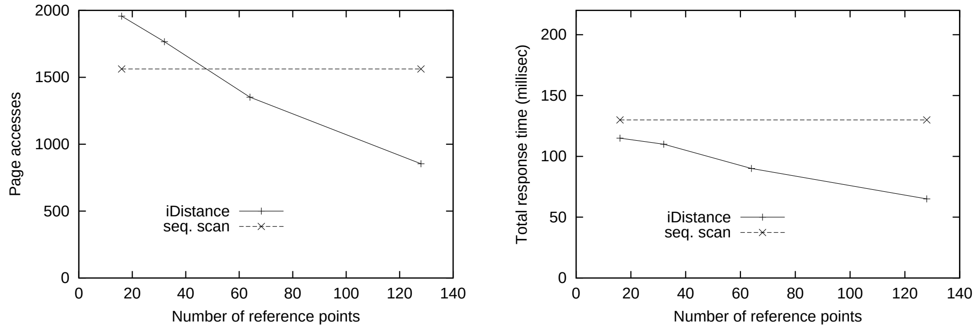
Using an external point as the reference point has slightly better performance than using the center point and using a farther external point is slightly better than using the external point in turn, but the difference between them is not big and all of them are still worse than the data-based partitioning approach (compare with Figure 18). Here, the farther external point is already very far (more than 10 times the side length of the data space) and the performance of using an even farther points almost does not change, they are not presented.

From the above results, we can see that the data-based partitioning scheme is always better than the space-based partitioning approach. Thus, for all subsequent experimental study, we will mainly focus on the data-based partitioning strategy. However, we note that the space-based partitioning is always better than sequential scan in low and medium dimensional spaces (less than 16). Thus, it is useful for these workloads. Moreover, the scheme incurs much less overhead since there is no need to cluster data to find the reference points as in the data-based partitioning.

6.2 iDistance Using Data-based Partitioning

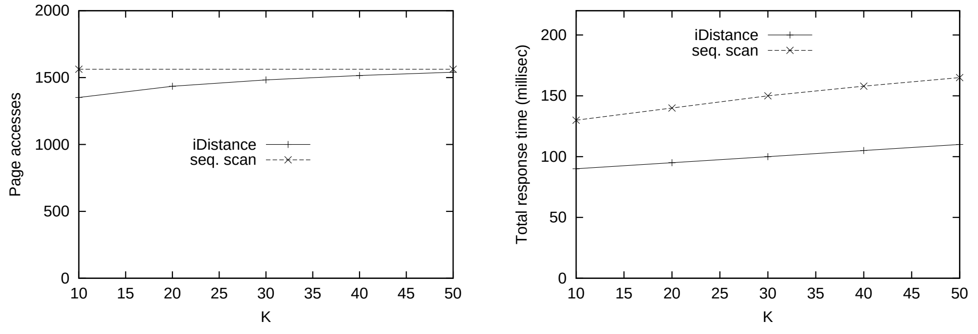
In this subsection, we further study the performance of iDistance using a data-based partitioning strategy (iDistance for short in the sequel). We study the effects of different parameters and different workloads. As reference, we compare the iDistance with sequential scan. Although iDistance (with data-based partitioning) is better than sequential scan for the 30-dimensional uniform dataset, the difference is small. To see more clearly the behavior of iDistance, we use 16-dimensional data when we test on uniform datasets. For clustered data, we use 30-dimensional datasets since iDistance is still much better than sequential scan for such high dimensionality.

Experiments on Uniform datasets



(a) Page accesses (b) Total response time

Figure 21: Effects of number of reference points, uniform data

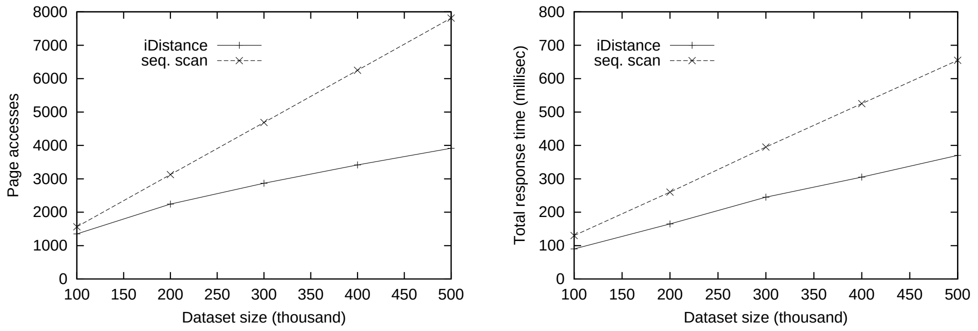


(a) Page accesses (b) Total response time

Figure 22: Effects of K, uniform data

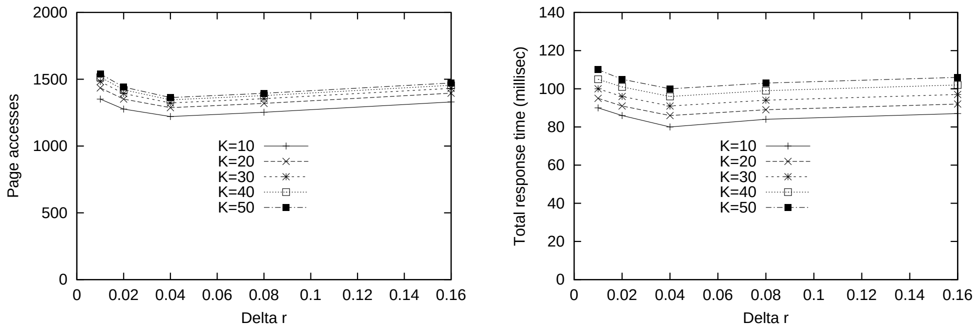
In the first experiment, we study the effect of the number of reference points on the performance of iDistance. The results of 10NN queries on the 100,000 16-dimensional uniform dataset are shown in Figure 21. We can see that as the number of reference points increases, the total response time reduces. This is expected as smaller and fewer clusters need to be examined (i.e., more data are pruned). The amount of the decrease in time also decreases as the number of reference points increases. We can choose a very large number of reference points to improve the performance, but on the other hand, CPU time increases because we need to check on more reference points and the time needed for clustering to find the reference points also increases. Moreover, there are more fragmented pages. So a moderate number of reference points is fine. In our other experiments, we have used 64 as the default number of reference points.

The second experiment studies the effect of K on the performance of iDistance. We varied



(a) Page accesses (b) Total response time

Figure 23: Effects of dataset size, uniform data



(a) Page accesses (b) Total response time

Figure 24: Effects of $\Delta\ r$, uniform data

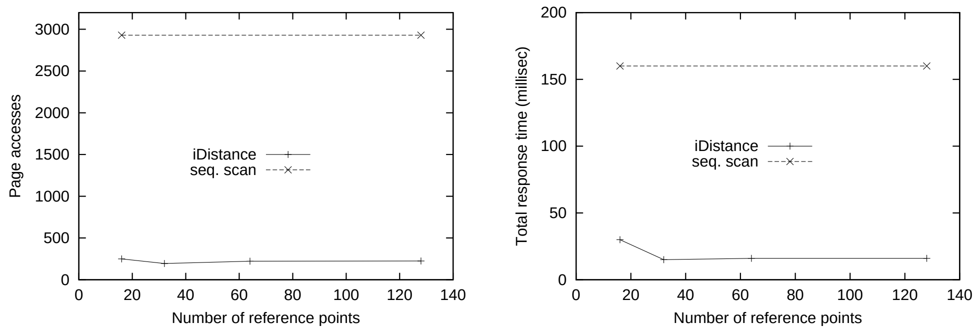
K from 10 to 50 at the step of 10. The results of queries on the 100,000 16-dimensional uniform dataset are shown in Figure 22. The total response time increases linearly as K increases, while the rate of increase for iDistance is slower than sequential scan.

This is because as K increases, the number of distance computation also increases for both iDistance and sequential scan. But, iDistance not only has fewer distance computation, the rate of increase in the distance computation is also smaller (than sequential scan).

The third experiment studies the effect of the dataset size. We varied the number of data points from 100,000 to 500,000. The results of 10NN queries on five 16-dimensional uniform datasets are shown in Figure 23. The total response time of both iDistance and sequential scan increases linearly as the dataset size increases, but the increase for sequential scan is much faster. When the dataset size is 500,000, the total response time of iDistance is about half of that of sequential scan.

The fourth experiment examines the effect of the Δr in the iDistance KNN Search Algorithm presented in Figure 4. Figure 24 shows the performance when we varied the values of Δr. We can observe that, as Δr increases, the total response time decreases at first but then increases. For a small Δr, there will be more iterations to reach the final query radius and consequently, more CPU time is incurred. On the other hand, if Δr is too large, the query radius may exceed the KNN distance at the last iteration and redundant data pages are fetched for checking. We note that it is very difficult to derive an optimal Δr since it is dependent on the data distribution and the order in which the data points are inserted into the index. Fortunately, the impact on performance is marginal (less than 10%). Considering that, in practice, small K may be used in KNN search, which implies a very small KNN distance. Therefore, in all our experiments, we have safely set $\Delta r = 0.01$, that is, 1% of the side length of the data space.

Experiments on Clustered datasets

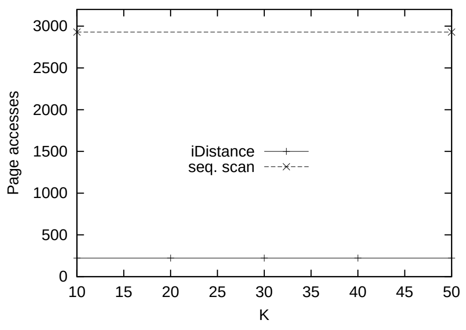


(a) Page accesses (b) Total response time

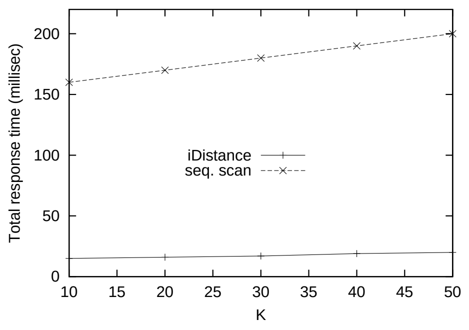
Figure 25: Effects of number of reference points, clustered data

For the clustered datasets, we also study the effect of the number of the reference points, K and dataset size. By default, the number of reference point is 64, K is 10 and dataset size is 100,000. Dimensionality of all these datasets is 30. The results are shown in Figures 25, 26 and 27 respectively. These results exhibit similar characteristics to those of the uniform datasets except that iDistance has much better performance compared to sequential scan. The speedup factor is as high as 10. The reason is that for clustered data, the K^{th} nearest neighbor distance is much smaller than that in uniform data, so many more data points can be pruned from the search. Figure 25 shows that after the number of reference points exceeds 32, the performance gain becomes marginal. We chose the number 64 in our experiments.

Each of the above clustered datasets consists of 20 clusters, each of which has the variance of 0.05. To evaluate the performance of iDistance on different distributions, we tested three other

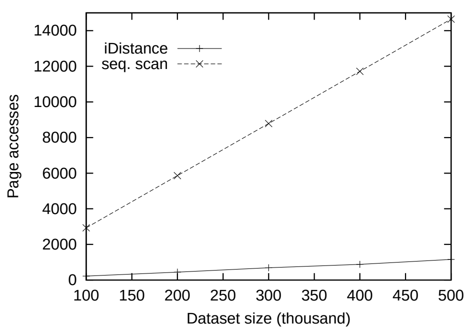


(a) Page accesses

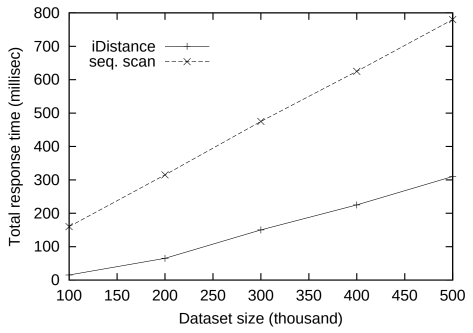


(b) Total response time

Figure 26: Effects of K, clustered data

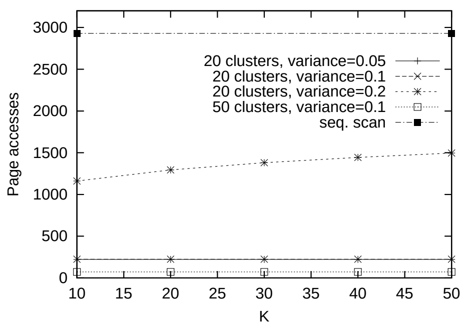


(a) Page accesses

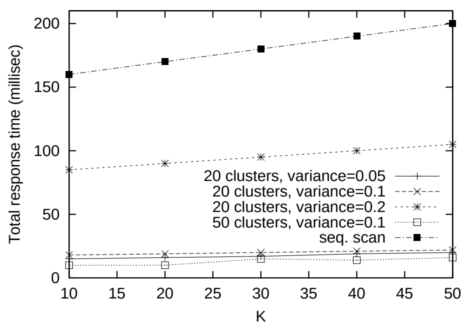


(b) Total response time

Figure 27: Effects of dataset size, clustered data

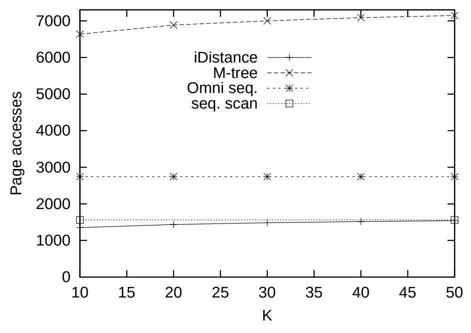


(a) Page accesses

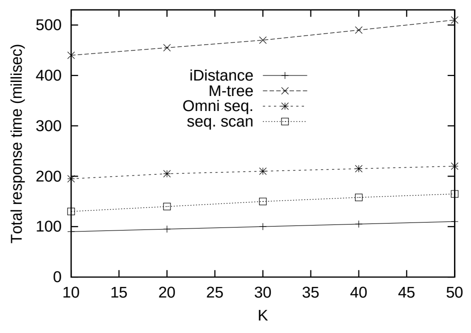


(b) Total response time

Figure 28: Effects of different data distribution, clustered data



(a) Page accesses



(b) Total response time

Figure 29: Comparative study, 16D uniform data

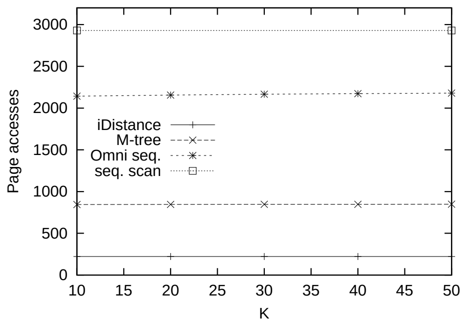
datasets with different number of clusters and different variance, while other settings are kept at the default values. The results are shown in Figure 28. Because all these datasets have the same number of data points but only differ in distribution, the performance of sequential scan is almost the same for all of them and hence we only plot one curve for sequential scan on these datasets. We observe that the total response time of iDistance remains very small for all the datasets with variance less than or equal to 0.1 but increases a lot when the variance increases to 0.2 This is because as the variance increases, the dataset becomes closer to uniform distribution, which is when iDistance becomes less efficient (but is still better than sequential scan).

We also studied the effect of different Δr on the clustered datasets. Like the results on the uniform datasets, the performance change is very small.

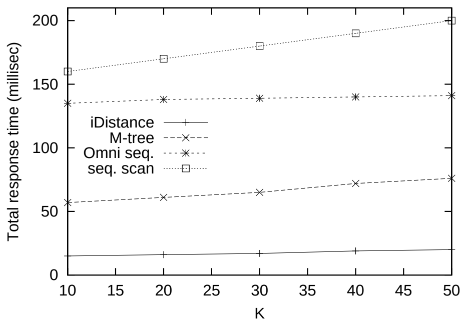
6.3 Comparative Study of iDistance and Other Techniques

In this subsection, we compare iDistance with sequential scan as well as two other metric based indexing methods, the M-tree [CPZ97] and Omni-sequential [FTF01]. In [FTF01], several indexing schemes of the Omni-family were proposed. Omni-sequential has the best average performance according to the experimental results reported in the paper. We therefore chose Omni-sequential from the family for comparison. Omni-sequential needs to select a good number of foci bases to work efficiently. In our comparative study, we tried Omni for several numbers of foci bases and only presented the one giving the best performance in the sequel. We still use 64 reference points for iDistance. Datasets used include 100,000 16-dimensional uniformly distributed points, 100,000 30-dimensional clustered points and 68040 32-dimensional real data. We varied K from 10 to 50 at the step of 10.

The results on the uniform dataset is shown in Figure 29. M-tree and Omni-sequential both have more page accesses and longer total response time than sequential scan. iDistance has similar page accesses to sequential scan, but shorter total response time than sequential scan. The results on the clustered dataset is shown in Figure 30. M-tree, Omni-sequential and iDistance are all better than sequential scan because the smaller K^{th} nearest neighbor distance enables more effective pruning of the data space for these metric based methods. iDistance performs the best. It has a speedup factor of about 3 over M-tree and 6 over Omni-sequential. The results on the real dataset is shown in Figure 31. M-tree and Omni-sequential have similar page accesses as sequential scan while the page accesses of iDistance is about 1/3 those of the other techniques. Omni-sequential and iDistance has a shorter total response time than sequential scan while M-tree has a very long total response time. Omni-sequential can reduce the number of distance computation, so it takes less time while having the same page accesses as sequential scan. The M-tree accesses the pages

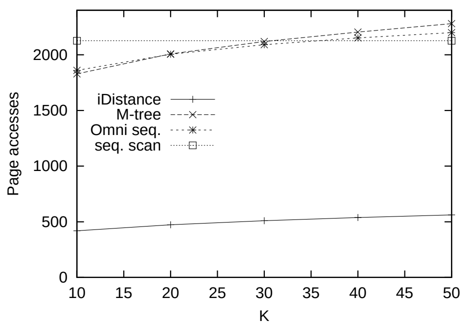


(a) Page accesses

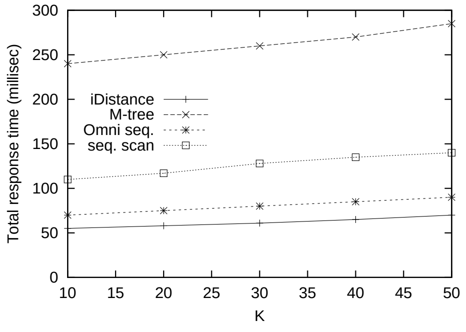


(b) Total response time

Figure 30: Comparative study, 30D clustered data



(a) Page accesses



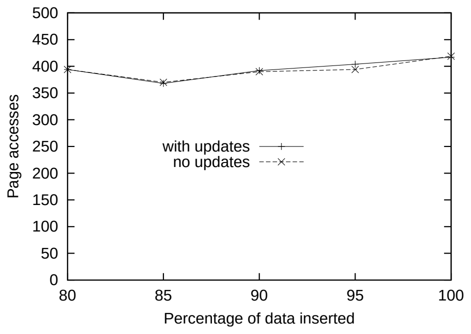
(b) Total response time

Figure 31: Comparative study, 32D real data

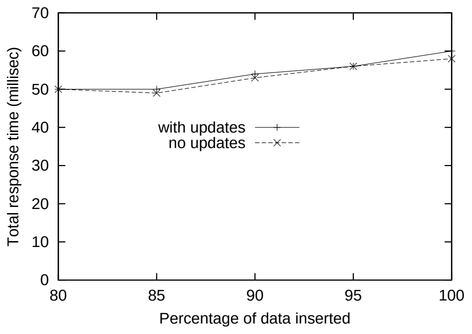
randomly, so it is much slower. iDistance has much fewer page accesses and distance computation, and hence it has the least total response time.

6.4 On Updates

We use clustering to choose the reference points from a collection of data points, and fix them from that point onwards. It is therefore important to see whether a dynamic workload would affect the performance of iDistance much. In this experiment, we first construct the index using 80% of the data points from the real dataset. We run 200 10NN queries and record the average page accesses and total response time. Then we insert 5% of the data to the database and re-run the same queries. This process is repeated until the other 20% of the data are inserted. Separately, we also run the same queries on the index built based on the reference points chosen for 85%, 90%, 95% and 100% of the dataset. We compare the average page accesses and total response time of the two as shown in Figure 32. The difference between them is very small. The reason is that real data from the same source tends to follow a similar distribution, so the reference points chosen at different times are similar. Of course, if the distribution of the data changes too much, we will need choose the reference points again and rebuild the index.



(a) Page accesses



(b) Total response time

Figure 32: iDistance performance with updates

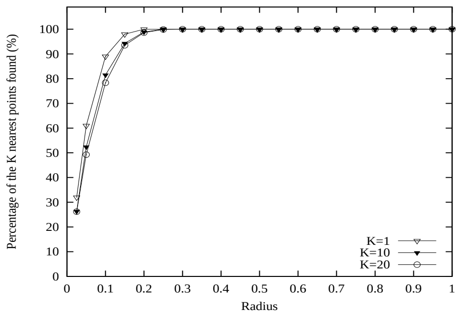
6.5 On Initial Answers

An interesting property of iDistance is that it has great probability to contain correct KNN answers in the initial KNN candidates. We did the following experiments. As the search radius of iDistance increases, we check how many points in the points that have been retrieved are in the final KNN set. In other words, we examine what percentage of the K nearest neighbors are retrieved as the search radius increases. The results on the 8, 16 and 30-dimensional uniform datasets are shown in Figure 33. Figure 34 shows the results on the increase in page accesses as the search radius increases. We observe that for a small initial search radius, a large percentage of K nearest neighbors are retrieved. Since the search radius has an almost linear relationship with the page accesses as shown in Figure 34, this means that the first small portion of the page accesses retrieve a large percentage of the K nearest neighbors. As a comparison, we plot Figure 35 for KNN search using sequential scan, which shows the percentage of K nearest neighbors found as the function of the average number page accesses when finding these answers. The percentage of K nearest neighbors found is almost linear to the number of page accesses. Since iDistance retrieves most of the initial answers in the first small portion of the page accesses, it is faster than sequential scan in returning initial answers. In fact, this is easy to understand as iDistance KNNsearch starts from the keys equal or close to the key of the query point q, those points which are nearest to q have high probability to be found in the first few iterations of enlarging the query sphere, therefore most of the true nearest neighbors are already in the candidate answer set in an early stage of the whole query processing time. While in sequential scan, the data are randomly distributed, therefore the times when finding the true nearest neighbors are randomly distributed in the whole scanning time.

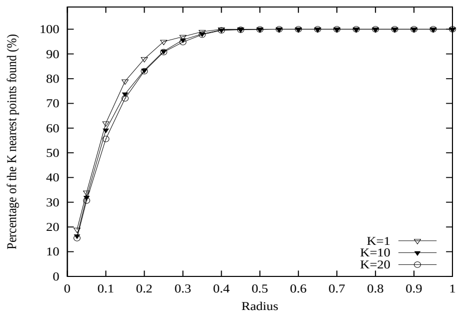
6.6 Evaluation of the Cost Models

Since our cost model estimates page accesses of each individual query, we show the actual number of page accesses and the estimated page accesses from 5 randomly chosen queries on the real dataset in Figure 36. Estimation of each of these 5 queries has the relative error below 20%. For all the tested queries, the estimations of more than 95% of them achieve the relative error below 20%. Considering that iDistance often has a speedup factor of 2 to 6 over other techniques, the 20% error will not affect the query optimization result greatly.

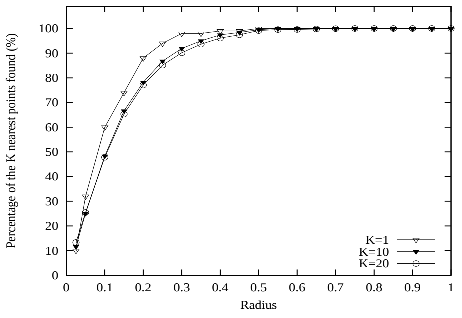
We also measured the time needed for computing the cost model. The average computation time (including the time for retrieving the number from the histogram) is less than 3% of the average KNN query processing time. So this cost model is still a practical approach for query optimization.



(a) d = 8



(b) d = 16



(c) d = 30

Figure 33: Efficiency in retrieving initial answers, iDistance

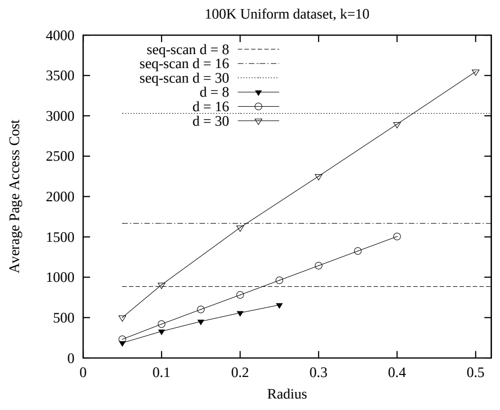


Figure 34: Page accesses vs. search radius, iDistance

6.7 Summary of the Experimental Results

The data-based partitioning approach is more efficient than the space-based partitioning approach. The iDistance using the data-based partitioning is always better than the other techniques in all our experiments on various workloads. For uniform data, it beats sequential scan in dimensionality as high as 30. Of course, due to the intrinsic characteristics of the KNN problem, we expect iDistance to lose out to sequential scan in much higher dimensionality on uniform datasets. However, for more practical data distributions, where data are skew and clustered, iDistance shows much better performance compared with sequential scan. Its speedup factor over sequential scan is as high as 10.

The number of reference points is an important tunable parameter for iDistance. Generally, the more the number of reference points, the better the performance, and at the same time, the longer time needed for clustering to determine these reference points. Too many reference points also impair performance because of higher computation overhead. Therefore, a moderate number is fine. We have used 64 as the number of reference points in most of our experiments (the others are because we need to study the effects of number of reference points) and iDistance performs better than sequential scan and other indexing techniques in these experiments. For a dataset with unknown data distribution, we suggest 60 to 80 number of reference points. Usually iDistance achieves a speedup factor of 2 to 6 over the other techniques. We can use a histogram-based cost model in query optimization to estimate the page access cost of iDistance, which usually has a relative error below 20%.

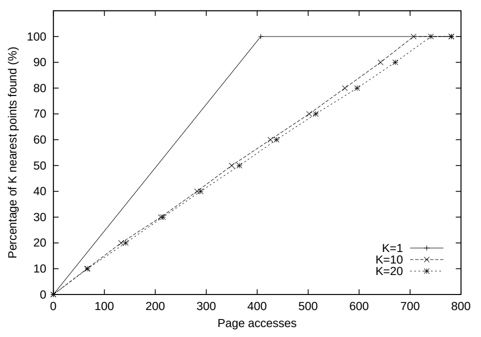
The space-based partitioning is simpler and can be used in low and medium dimensional space.

7 Conclusion

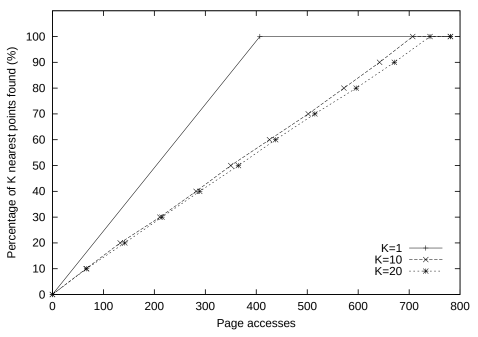
Similarity search is of growing importance, and is often most useful for objects represented in a high dimensionality attribute space. A central problem in similarity search is to find the points in the dataset nearest to a given query point. In this paper we have presented a simple and efficient method, called iDistance, for K-nearest neighbor (KNN) search in a high-dimensional metric space.

Our technique partitions the data and selects one reference point for each partition. The data in each cluster can be described based on their similarity with respect to a reference point, and hence they can be transformed into a single dimensional space based on such relative similarity. This allows us to index the data points using a B^+-tree structure and perform KNN search using a simple one-dimensional range search. As such, the method is well suited for integration into existing DBMSs.

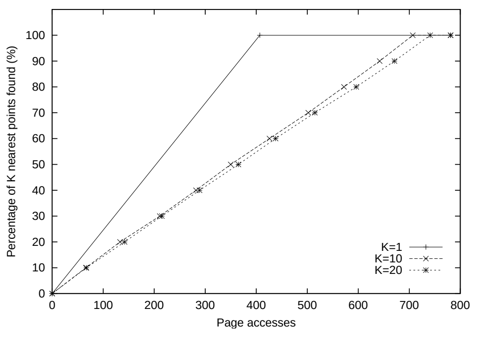
The choice of partition and reference points provides the iDistance technique with degrees of



(a) d = 8



(b) d = 16



(c) d = 30

Figure 35: Efficiency in retrieving initial answers, sequential scan

freedom that most other techniques do not have. We described how appropriate choices here can effectively adapt the index structure to the data distribution. In fact, several well-known data structures can be obtained as special cases of iDistance suitable for particular classes of data distributions. A cost model was proposed for iDistance KNN search to facilitate query optimization.

We conducted an extensive experimental study and a comparative study between iDistance, and two other metric based indexes the M-tree, and Omni-sequential. As a reference, we also compared the iDistance against sequential scan. Experimental results show that the iDistance outperforms the other techniques. Moreover, the index can be incorporated into existing DBMS cost effectively since the method is built on top of the B$^+$-tree.

APPENDIX

A Cost Model for Data-based Partitioning

We assume the data and queries are uniformly distributed in the data space. First, we estimate the number of partitions accessed by the query sphere. Second, we estimate how many data points are accessed in one accessed partition. Finally, we multiply these two. The notations listed in Table 2 still apply to the following derivations.

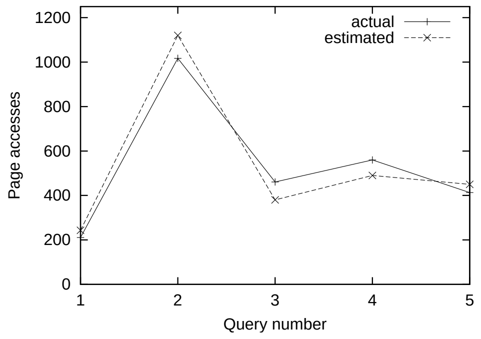


Figure 36: Evaluation of the histogram-based cost model

A.1 Number of Partitions Accessed

To estimate the number of partitions accessed by the query sphere, we view the clusters as hyperrectangles that fill up the data space. The rationale is that volume is the primary factor in the estimation. These hyperrectangles are resulted from splitting the data space in some dimensions. The number of partitions is not large, so they may not be split in all dimensions. For example, in an 8-dimensional space, we need $2^8 = 256$ partitions to make a hyperrectangle split in each dimension at least once. Since we do not have so many partitions (in the experiments, we at most used 128 partitions), we use the following equation to estimate in how many dimensions these hyperrectangles are split:

$$N_s = log_2 N_{pt}$$

where N_{pt} is the number of partitions (i.e. the number of clusters) which can be obtained after clustering. In these N_s dimensions, the extent of the cluster is 0.5 ; while in the other d-N_s dimensions, the extent of the cluster is 1.

Then we replace the query sphere with a hypercube of the same volume. The side length of the hypercube is L, where

$$\frac{\sqrt{\pi^d}}{\Gamma(\frac{d}{2}+1)} r^d = L^d$$

So

$$L = \frac{\sqrt{\pi}}{\Gamma^{\frac{1}{d}}(\frac{d}{2}+1)} r$$

If the query hyperrectangle and a cluster intersects, they have to intersect in all dimensions. Thus, they must intersect in the N_s partitions where the cluster is divided. The probability that they intersect in these N_s dimensions is

$$P_x = (L + 0.5)^{N_s}$$

Taking boundary effects into account, [TZPM] shows that we should replace L by $L - \frac{L^2}{4}$ So

$$P_x = (L - \frac{L^2}{4} + 0.5)^{N_s}$$

The probability that the query region intersects with exactly one cluster is

$$P_{inter1} = C^1_{N_{pt}} P_x (1 - P_x)^{N_{pt}-1}$$

The probability that the query region intersects with exactly two clusters is

$$P_{inter2} = C^2_{N_{pt}} P_x^2 (1 - P_x)^{N_{pt}-2}$$

and so on. Thus, the expected number of clusters intersecting the query region is

$$N_p = P_{inter1} \cdot 1 + P_{inter2} \cdot 2 + ... = \sum_{i=1}^{N_{pt}} C^i_{N_{pt}} \cdot P_x^i \cdot (1 - P_x)^{N_{pt}-i} \cdot i$$

A.2 Number of Data Points Accessed in One Accessed Partition

We assume that a data-based partition is a hypersphere and the whole hypersphere is in the data space. Both the cluster center and the cluster edge can be chosen as the reference point of a cluster. The experiments show that they have similar performance. We only consider the former case here. There are 3 cases on the relation between R_{max} and r.
Case 1: $R_{max} < r$
According to the KNN search algorithm of iDistance, Q must be outside of this partition because the query sphere will never contain a cluster center, as Figure 37 shows. The grey region represents the space that is searched. The number of data points accessed is

$$Dp_{11}(dist(O, Q)) = (V_s(R_{max}) - V_s(dist(O, Q) - r))N$$

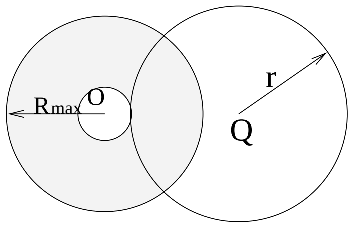


Figure 37: Case 1: $R_{max} < r$

Since this partition is accessed, $dist(O, Q)$ in the above expression ranges from R_{max} to $R_{max} + r$. We need to calculate the expectation of the number of data points accessed in this partition when $dist(O, Q)$ changes from R_{max} to $R_{max} + r$. The probability that $dist(O, Q)$ is less than a given value y, $(R_{max} < y \leq R_{max} + r)$ is

$$P_{11}(dist(O, Q) \leq y) = \frac{V_s(y) - V_s(R_{max})}{V_s(R_{max} + r) - V_s(R_{max})}$$

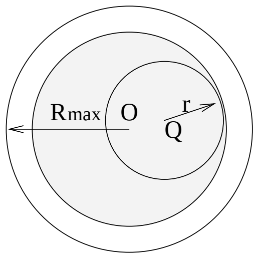
where $V_s(R_{max} + r) - V_s(R_{max})$ is the volume of the event space of the query point.
The density function $p_{11}(y)$ is

$$p_{11}(y) = \frac{dP_{11}(dist(O, Q) \leq y)}{dy} = \frac{1}{V_s(R_{max} + r) - V_s(R_{max})} \frac{dV_s(y)}{dy}$$

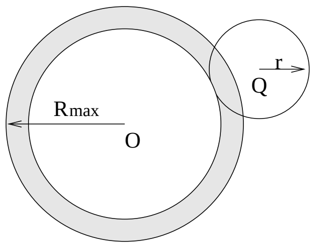
So the expected number of data points accessed in this partition is

$$\overline{Dp_{11}} = \int_{R_{max}}^{R_{max}+r} Dp_{11}(y) \cdot p_{11}(y) dy$$

Case 2: $\frac{R_{max}}{2} < r \leq R_{max}$
According to the KNN search algorithm of iDistance, there are two sub-cases:

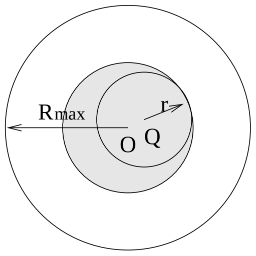


(a) Subcase 1: $\frac{R_{max}}{2} < r \leq R_{max}$ & $dist(O,Q) + r \leq R_{max}$

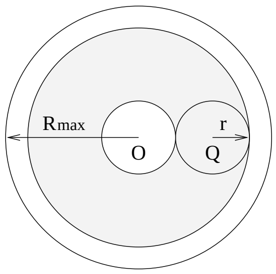


(b) Subcase 2: $\frac{R_{max}}{2} < r \leq R_{max}$ & $R_{max} - r < dist(O,Q) \leq R_{max} + r$

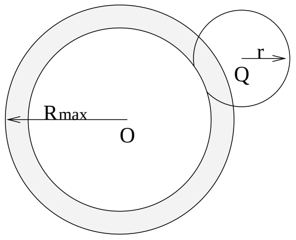
Figure 38: Case 2: $\frac{R_{max}}{2} < r \leq R_{max}$



(a) Subcase 1: $r \leq \frac{R_{max}}{2}$ & $dist(O,Q) \leq r$



(b) Subcase 2: $r \leq \frac{R_{max}}{2}$ & $r < dist(O,Q) \leq R_{max} - r$



(c) Subcase 3: $r \leq \frac{R_{max}}{2}$ & $R_{max} - r < dist(O,Q) \leq R_{max} + r$

Figure 39: Case 3: $r \leq \frac{R_{max}}{2}$

Subcase 1: $dist(O,Q) + r \leq R_{max}$
As shown in Figure 38(a), $dist(O,Q)$ ranges from 0 to $R_{max} - r$. As in case 1, we can derive that:

$$Dp_{21}(dist(O,Q)) = (V_s(dist(O,Q) + r))N$$

$$P_{21}(dist(O,Q) \leq y) = \frac{V_s(y)}{V_s(R_{max} + r)}$$

$$p_{21}(y) = \frac{dP_{21}(dist(O,Q) \leq y)}{dy} = \frac{1}{V_s(R_{max} + r)} \frac{dV_s(y)}{dy}$$

$$\overline{Dp_{21}} = \int_0^{R_{max}-r} Dp_{21}(y) \cdot p_{21}(y) dy$$

Subcase 2: $dist(O,Q) + r > R_{max}$ and $dist(O,Q) - r \leq R_{max}$
$dist(O,Q)$ ranges from $R_{max} - r$ to $R_{max} + r$, as Figure 38(b) shows

$$Dp_{22}(dist(O,Q)) = (V_s(R_{max}) - V_s(dist(O,Q) - r))N$$

$$P_{22}(dist(O,Q) \leq y) = \frac{V_s(y) - V_s(R_{max} - r)}{V_s(R_{max} + r)}$$

$$p_{22}(y) = \frac{dP_{22}(dist(O,Q) \leq y)}{dy} = \frac{1}{V_s(R_{max} + r)} \frac{dV_s(y)}{dy}$$

$$\overline{Dp_{22}} = \int_{R_{max}-r}^{R_{max}+r} Dp_{22}(y) \cdot p_{22}(y) dy$$

Case 3: $r \leq \frac{R_{max}}{2}$
There are three sub-cases:

Subcase 1: $dist(O,Q) \leq r$
$dist(O,Q)$ ranges from 0 to r, as shown in Figure 39(a).

$$Dp_{31}(dist(O,Q)) = V_s(dist(O,Q) + r)N$$

$$P_{31}(dist(O,Q) \leq y) = \frac{V_s(y)}{V_s(R_{max} + r)}$$

$$p_{31}(y) = \frac{dP_{31}(dist(O,Q) \leq y)}{dy} = \frac{1}{V_s(R_{max} + r)} \frac{dV_s(y)}{dy}$$

$$\overline{Dp_{31}} = \int_{0}^{r} Dp_{31}(y) \cdot p_{31}(y) dy$$

Subcase 2: $r < dist(O,Q)$ and $dist(O,Q) + r \leq R_{max}$
$dist(O,Q)$ ranges from r to $R_{max} - r$ as shown in Figure 39(b).

$$Dp_{32}(dist(O,Q)) = (V_s(dist(O,Q) + r) - V_s(dist(O,Q) - r))N$$

$$P_{32}(dist(O,Q) \leq y) = \frac{V_s(y) - V_s(r)}{V_s(R_{max} + r)}$$

$$p_{32}(y) = \frac{dP_{32}(dist(O,Q) \leq y)}{dy} = \frac{1}{V_s(R_{max} + r)} \frac{dV_s(y)}{dy}$$

$$\overline{Dp_{32}} = \int_{r}^{R_{max}-r} Dp_{32}(y) \cdot p_{32}(y) dy$$

Subcase 3: $R_{max} < dist(O,Q) + r$ and $dist(O,Q) - r \leq R_{max}$
$dist(O,Q)$ ranges from r to $R_{max} - r$ as shown in Figure 39(c).

$$Dp_{33}(dist(O,Q)) = (V_s(R_{max}) - V_s(dist(O,Q) - r))N$$

$$P_{33}(dist(O,Q) \leq y) = \frac{V_s(y) - V_s(R_{max} - r)}{V_s(R_{max} + r)}$$

$$p_{33}(y) = \frac{dP_{33}(dist(O,Q) \leq y)}{dy} = \frac{1}{V_s(R_{max} + r)} \frac{dV_s(y)}{dy}$$

$$\overline{Dp_{33}} = \int_{R_{max}-r}^{R_{max}+r} Dp_{33}(y) \cdot p_{33}(y) dy$$

A.3 Number of Pages Accessed

Now by multiplying the number of partitions accessed and number of data points accessed in one accessed partition, we get N_a, the number of data points accessed in the whole data space:

When $R_{max} < r$, $N_a = N_p \cdot \overline{Dp_{11}}$

When $\frac{R_{max}}{2} < r \le R_{max}$, $N_a = N_p \cdot (\overline{Dp_{21}} + \overline{Dp_{22}})$

When $r \le \frac{R_{max}}{2}$, $N_a = N_p \cdot (\overline{Dp_{31}} + \overline{Dp_{32}} + \overline{Dp_{33}})$

So we can derive the expected number of data points accessed by the integral

$$\overline{N_a} = \int_0^\infty N_a \cdot p(r)dr = \int_0^{\frac{R_{max}}{2}} N_p \cdot (\overline{Dp_{31}} + \overline{Dp_{32}} + \overline{Dp_{33}}) \cdot p(r)dr$$
$$+ \int_{\frac{R_{max}}{2}}^{R_{max}} N_p \cdot (\overline{Dp_{21}} + \overline{Dp_{22}}) \cdot p(r)dr + \int_{R_{max}}^{\infty} N_p \cdot \overline{Dp_{11}} \cdot p(r)dr \tag{10}$$

where $p(r)$ is the density function of the query radius. We can do some query simulation to obtain the density function, but this may involve a great deal of computation. As a simplification, we can only calculate page access cost at the point of the expected query radius, instead of computing the whole integral. The expected query radius is calculated using the same method as presented in Section 5.1. We shall discuss how to estimate R_{max} shortly.

We store the data points (vectors) in the leaf nodes of the B$^+$-tree, so the number of page accesses equals the number of data points accessed divided by the average number of points stored in a page, C_{eff}.

$$E_{databased} = \frac{\overline{N_a}}{C_{eff}}$$

A.4 Estimation of R_{max}

There are many clustering algorithms, and estimation of R_{max} has to be based on the underlying clustering method. For example, if BIRCH [ZRL96] is used, the threshold T − the radius of the clusters provided by the user, can be used directly to replace R_{max}. We used k-means to cluster the data points. There is no parameter that can indicate the value of R_{max} directly. Hence we estimate R_{max} by volume. Assume the data space is composed by N_{pt} partitions which are hypersphere shaped, then

$$\frac{1}{N_{pt}} = \frac{\sqrt{\pi^d}}{\Gamma(\frac{d}{2}+1)} R_{max}^d \tag{11}$$

References

[APW+99] C. Aggarwal, C. Procopiuc, J. Wolf, P. Yu, and J. Park. Fast algorithm for projected clustering. In *Proc. 1999 ACM SIGMOD International Conference on Management of Data*. 1999.

[BBJ+00] S. Berchtold, C. Böhm, H.V. Jagadish, H.P. Kriegel, and J. Sander. Independent quantization: An index compression technique for high-dimensional data spaces. In *Proc. 16th International Conference on Data Engineering*, pages 577–588. 2000.

[BBK98] S. Berchtold, C. Böhm, and H-P. Kriegel. The pyramid-technique: Towards breaking the curse of dimensionality. In *Proc. 1998 ACM SIGMOD International Conference on Management of Data*, pages 142–153. 1998.

[BBK01] C. Böhm, S. Berchtold, and D. Keim. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Computing Surveys*, 33(3):322–373, 2001.

[BEK$^+$98] S. Berchtold, B. Ertl, D.A. Keim, H.-P. Kriegel, and T. Seidl. Fast nearest neighbor search in high-dimensional space. In *Proc. 14th International Conference on Data Engineering*, pages 209–218, 1998.

[BGRS99] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbors meaningful? In *Proc. International Conference on Database Theory*, 1999.

[BKK96] S. Berchtold, D.A. Keim, and H.P. Kriegel. The X-tree: An index structure for high-dimensional data. In *Proc. 22nd International Conference on Very Large Data Bases*, pages 28–37. 1996.

[BO97] T. Bozkaya and M. Ozsoyoglu. Distance-based indexing for high-dimensional metric spaces. In *Proc. 1997 ACM SIGMOD International Conference on Management of Data*, pages 357–368. 1997.

[Boh00] C. Bohm. A cost model for query processing in high-dimensional data spaces. *ACM Transactions on Database Systems*, 25(2):129–178, 2000.

[CM99] K. Chakrabarti and S. Mehrotra. The hybrid tree: An index structure for high dimensional feature spaces. In *Proc. International Conference on Data Engineering*, pages 322–331, 1999.

[CM00] K. Chakrabarti and S. Mehrotra. Local dimensionality reduction: a new approach to indexing high dimensional spaces. In *Proc. 26th International Conference on Very Large Databases*, pages 89–100, 2000.

[CPZ97] P. Ciaccia, M. Patella, and P. Zezula. M-trees: An efficient access method for similarity search in metric space. In *Proc. 23rd International Conference on Very Large Data Bases*, pages 426–435. 1997.

[FL95] C. Faloutsos and K.-I. Lin. Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proc. 1995 ACM SIGMOD International Conference on Management of Data*, pages 163–174, 1995.

[FTF01] R. F. S. Filho, A. Traina, and C. Faloutsos. Similarity search without tears: The omni family of all-purpose access methods. In *Proc. 17th International Conference on Data Engineering*, pages 623–630, 2001.

[GR00] J. Goldstein and R. Ramakrishnan. Contrast plots and p-sphere trees: space vs. time in nearest neighbor searches. In *Proc. 26th International Conference on Very Large Databases*, pages 429–440, 2000.

[GRS98] S. Guha, R. Rastogi, and K. Shim. Cure: an efficient clustering algorithm for large databases. In *Proc. 1998 ACM SIGMOD International Conference on Management of Data*. 1998.

[Gut84] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *Proc. 1984 ACM SIGMOD International Conference on Management of Data*, pages 47–57. 1984.

[Jol86] I. T. Jolliffe. *Principle Component Analysis*. Springer-Verlag, 1986.

[Kru56] J. B. Kruskal. On the shortest spanning subtree of a graph and the travelling salesman problem. In *Proc. American Math Soc., Vol.7*, pages 48–50, 1956.

[KS97] N. Katamaya and S. Satoh. The SR-tree: An index structure for high-dimensional nearest neighbor queries. In *Proc. 1997 ACM SIGMOD International Conference on Management of Data*. 1997.

[LJF95] K. Lin, H.V. Jagadish, and C. Faloutsos. The TV-tree: An index structure for high-dimensional data. *VLDB Journal*, 3(4):517–542, 1995.

[Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Mathematical statistics and probability*, pages 281–297. University of California Press, 1967.

[OTYB00] B. C. Ooi, K. L. Tan, C. Yu, and S. Bressan. Indexing the edge: a simple and yet efficient approach to high-dimensional indexing. In *Proc. 18th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 166–174. 2000.

[PKF00] B.-U. Pagel, F. Korn, and C. Faloutsos. Deflating the dimensionality curse using multiple fractal dimensions. In *Proc. 16th International Conference on Data Engineering*, 2000.

[RG00] R. Ramakrishnan and J. Gehrke. *Database Management Systems*. McGraw-Hill, 2000.

[SYU00] Y. Sakurai, M. Yoshikawa, and S. Uemura. The a-tree: An index structure for high-dimensional spaces using relative approximation. In *Proc. 26th International Conference on Very Large Data Bases*, pages 516–526. 2000.

[TFP03] Y. Tao, C. Faloutsos, and D. Papadias. The power-method: A comprehensive estimation technique for multi-dimensional queries. In *Proc. 2003 Conference on Information and Knowledge Management*, 2003.

[TSF00] A. Traina, B. Seeger, and C. Faloutsos. Slim-trees: high performance metric trees minimizing overlap between nodes. In *Advances in Database Technology - EDBT 2000, 7th International Conference on Extending Database Technology, Konstanz, Germany, March 27-31, 2000, Proceedings*, volume 1777 of *Lecture Notes in Computer Science*, pages 51–65. Springer-Verlag, 2000.

[TZPM] Y. Tao, J. Zhang, D. Papadias, and N. Mamoulis. An efficient cost model for optimization of nearest neighbor search in low and medium dimensional spaces. *To appear in IEEE Transactions on Knowledge and Data Engineering.*

[WJ96] D.A. White and R. Jain. Similarity indexing with the SS-tree. In *Proc. 12th International Conference on Data Engineering*, pages 516–523. 1996.

[WSB98] R. Weber, H. Schek, and S. Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proc. 24th International Conference on Very Large Data Bases*, pages 194–205. 1998.

[YOTJ01] C. Yu, B. C. Ooi, K. L. Tan, and H. Jagadish. Indexing the distance: an efficient method to knn processing. In *Proc. 27th International Conference on Very Large Data Bases*, pages 421–430. 2001.

[ZRL96] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. In *Proc. 1996 ACM SIGMOD International Conference on Management of Data*. 1996.