Dimensionality Curse, Concentration Phenomenon and the KDB-tree

Nikolaos Kouiroukidis and Georgios Evangelidis

University of Macedonia, Department of Applied Informatics, 54006, Thessaloniki, Greece
{kouiruki, gevan} (at) uom.gr

Abstract: *The problem of indexing large volumes of high dimensional data is an important and popular issue in the area of database management. There are many indexing methods that behave well in low dimensional spaces, but, in high dimensionalities, the phenomenon of the curse of dimensionality renders all indexes useless. For example, when issuing range queries almost all of the index pages have to be retrieved for answering these queries. In this paper we review the state-of-the-art research regarding high dimensional spaces and we demonstrate the dimensionality curse phenomenon using the TPIE KDB-tree implementation.*

Keywords: *Dimensionality curse, KDB tree, Hypercube range queries*

I.INTRODUCTION

The term "curse of dimensionality" describes the rapid deterioration in the performance of high dimensional indexes as the number of variables (or dimensions) increases. When range or k-nearest neighbor queries are issued in high dimensional spaces, most (if not all) of the pages of the indexing structures that are employed to store the high dimensional points are visited, and the good performing in low dimensional spaces indexing methods, end up behaving as the plain sequential scan.

One of the classical indexing methods is the KDB-tree (Robinson, 1981) with TPIE (Arge et al, 2002) being one of his most efficient implementations. The KDB-tree combines some of the properties of the adaptive k-d-tree (Bentley, 1975) and the B-tree to handle multidimensional points. Each interior node corresponds to an interval-shaped region. Regions corresponding to nodes at the same tree level are mutually disjoint; their union is the complete universe. The leaf nodes store the data points that are located in the corresponding partition. Like the B-tree, the KDB-tree is a perfectly balanced tree that adapts well to the distribution of data.

In Section II, we present some observations regarding the dimensionality curse phenomenon. In Section III, we discuss the concentration phenomenon and in Section IV, we demonstrate the behavior of the KDB-tree in high dimensions. We conclude in Section V.

II.THE CURSE OF DIMENSIONALITY

The following phenomena give an insight to the notion of the dimensionality curse. See Weber et al. (1998) for further details.

1. The partitioning schemes usually split the data space in each dimension in two halves. With d dimensions there are 2^d partitions. With d<=10 and N on the order of 10^6 such a partition makes sense. However if d is larger, say d=100, there are around 10^{30} partitions for only 10^6 points. An overwhelming number of partitions are empty.

2. If we consider a hypercube range query with length s in all d dimensions the probability that a point lies within that range query is given by $P^d[s]=s^d$. This probability function is plotted in Fig. 1 below. From the formula, directly follows that even very large range queries are not likely to contain a point. At d=100 a range query with length 0.95 selects 0,59% of the data points. This hypercube range query can be placed anywhere in the data space Ω. Thus, we conclude that the data space is sparsely populated.

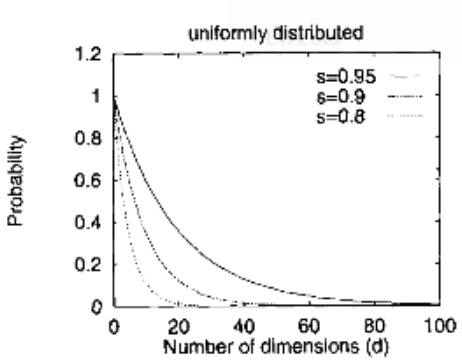

Figure 1. Plotting the probability that a hypercube query with side s contains a point.

3. The largest spherical query that fits entirely within the data space is the query $sp^d(Q,0.5)$ where Q is the centroid of the data space. The probability that an arbitrary point R lies within this sphere is given by the sphere volume

$$P[R \in sp^d(Q, \frac{1}{2})] \simeq \frac{Vol(sp^d(Q, \frac{1}{2}))}{Vol(\Omega)} = \frac{\sqrt{\pi^d} \cdot (\frac{1}{2})^d}{\Gamma(\frac{d}{2}+1)}$$

The relative volume of the sphere shrinks markedly as the dimensionality grows and it increasingly becomes improbable that any point will be found within this sphere at all. Table 1 shows this probability for various numbers of dimensions.

4. From the probability equation given above, one can determine a size a data set would have to have

such that on average at least one point falls into the sphere $sp^d(Q,0.5)$ (for even d). This is given in the following equation:

$$N(d) = \frac{(\frac{d}{2})!}{\sqrt{\pi^d} \cdot (\frac{1}{2})^d}$$

Table 1 enumerates this function for various numbers of dimensions. The number of points needed explodes exponentially. At d=20, a database must contain at least 40 million points in order to ensure that on average at least one point lies within this sphere.

D	P[R ε sp^d(Q,0.5)]	N(d)
2	0.785	1.273
4	0.308	3.242
10	0.002	401.5
20	$2.461 * 10^{-8}$	40,631,627
40	$3.278 * 10^{-21}$	$3.050 * 10^{20}$
100	$1.868 * 10^{-70}$	$5.353 * 10^{69}$

Table 1. Probability that a point is in the largest hyper-sphere

5. The expected Nearest Neighbor distance between two points in a data space Ω is given by the following formula

$$E[nn^{dist}] = \int_{Q \in \Omega} E[Q, nn^{dist}]\, dQ$$

where Q is the query point. Based on this formula, and if one estimates it with the Monte Carlo method, one finds that NN distance grows steadily with d, and except trivially small data sets, the objects are widely scattered and the probability of being able to identify a good partitioning of the data space diminishes.

6. Finally, due to the dimensionality curse phenomenon, as we will demonstrate in our experiments with the KDB-tree, when a range query is performed nearly all data pages have to be accessed in order to obtain the answer. This equals almost to a sequential scan.

III.CONCENTRATION PHENOMENON

The concentration phenomenon can be stated as follows (Ledoux, 2001): in high dimensional spaces all pairwise distances between points seem identical. Here, we'll study the concentration of the distances through the concentration of the norm. If we have n points with d dimensions each, taking values from the unit cube $[0,1]^d$ and we then consider their norms ||x||, the values of ||x|| are bounded in the interval [0,M], where M=||(1,1,....1)||.

Let us consider the euclidean norm M=sqrt(d). If we plot the minimum observed value and the maximum observed value, we observe that in low dimensions these values are close to the bounds of the domain of the norm, respectively 0 and sqrt(d). Also, the average value of the norm increases with the dimension, whereas the standard deviation seems rather constant. When the dimension is large (above 10) the minimum and maximum observed values tend to move away from the bounds. When the number of points are, for example, 100000 all the observed norms seem to concentrate in a small portion of their domain. In addition this portion gets smaller and smaller as the dimension grows when compared to the size of the total domain.

The Minkowski norms form a family of norms parametrized by their exponent p=1,2,3....

$$\|X\|_p = \left(\sum_i |X_i|^p \right)^{\frac{1}{p}}.$$

When 0<p<1, the triangle inequality does not hold so these norms are called prenorms or fractional norms. Actually, the inequality is reversed. A consequence is that the straight line is no longer the smallest path between two points. Fig. 2 depicts 2D unit balls (that is the set of x^j for which $||x^j||=1$) for various values of p. We see that for p>=1 the balls are convex and for 0<p<1 they are not.

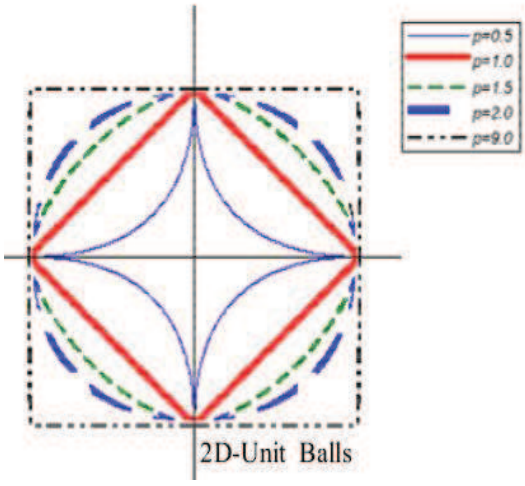

Figure 2. ***2D-Unit Balls.***

A. Concentration of the euclidean norm

If X is in R^d and is a random vector with independent and identically distributed components, and X_i follows distribution F, then

$$\mathrm{E}(\|X\|_2) = \sqrt{ad - b} + O(1/d) \text{ and}$$

$$\mathrm{Var}(\|X\|_2) = b + O(1/\sqrt{d}),$$ where a and b are constants that do not depend on the dimension (François et al., 2007; Aggarwal, 2001). This holds for

any kind of distribution. Different distributions will lead to different values for a and b but the asymptotic results remain.

This shows that the expectation of the euclidean norm of random vectors increases as the square root of the dimension, whereas its variance is constant and independent of the dimension. Therefore, when the dimension is large the variance of the norm is very small compared with its expected value. Also when the dimension is large vectors seem normalized. The relative error made while considering $E(\|X\|_2)$ instead of the real value of $\|X\|_2$ becomes negligible. As a consequence, high dimensional vectors appear to be distributed on a sphere of radius $E(\|X\|_2)$.

Since the euclidean distance is the norm of the difference between two random vectors, it's expectation and variance follow the two above laws and pairwise distances between points in high dimensional spaces seem to be all identical. Finally, if X_i are not independent the results are still valid provided that we replace d with the actual number of degrees of freedom.

In contrast to the work of Demartines (1994), where a data set X consists of n independent draws x^j from a single random vector X, Beyer (1999) considers n random vectors P^j where a dataset is made of one realization of each random vector. Beyer's theorem states that if P^j 1<=j<=n are n d-dimensional independent and identically distributed random vectors and if

$$\lim_{d\to\infty} \mathrm{Var}\left(\frac{\|\mathrm{P}^{(j)}\|}{\mathrm{E}(\|\mathrm{P}^{(j)}\|)}\right) = 0$$

then for any ε>0

$$\lim_{d\to\infty} \mathbf{P}\left[\frac{\max_j \|\mathrm{P}^{(j)}\| - \min_j \|\mathrm{P}^{(j)}\|}{\min_j \|\mathrm{P}^{(j)}\|} \le \epsilon\right] = 1.$$

This is explained as follows. Suppose there are a set of n data points randomly distributed in the d-dimensional space and some query points are supposed to be located at the origin without loss of generality. Then, if the above hypothesis is satisfied, independent of the distribution of the components of the Pj, the difference between the largest and smallest distances to the query point becomes smaller and smaller when compared with the smallest distance when the dimension increases. The ratio

$$\frac{\max_j \|\mathrm{P}^{(j)}\| - \min_j \|\mathrm{P}^{(j)}\|}{\min_j \|\mathrm{P}^{(j)}\|}$$

is called the relative contrast.

So, Beyer concluded that all points are located at approximately the same distance from the query point. Thus, the concept of NN in a high dimensional space is less intuitive than in a lower dimensional one.

B. Concentration of Minkowski norms

There is the theorem of Hinneburg (François et al., 2007; Aggarwal et al., 2001), that states the following: let P^j 1<=j<=n, n d-dimensional independent and identically distributed random vectors and $\|.\|_p$ the Minkowski norm with exponent p. If the P^j are distributed in $[0,1]^d$ then there exists a constant C_p independent of the distribution of the P^j such that

Then, there is the suprising fact that on average the

$$C_p \le \lim_{d\to\infty} \mathrm{E}\left(\frac{\max_j \|\mathrm{P}^{(j)}\|_p - \min_j \|\mathrm{P}^{(j)}\|_p}{d^{\frac{1}{p}-\frac{1}{2}}}\right) \le (n-1)\cdot C_p.$$

$$\max_j \|\mathrm{P}^{(j)}\|_p - \min_j \|\mathrm{P}^{(j)}\|_p$$

contrast grows as $d^{1/p-1/2}$. As a result, the contrast converges to a constant when the dimension increases and when the euclidean distance is used. For the L_1 norm, it increases as sqrt(d), for the euclidean norm (p=2) it remains constant and for norms with p>=3 it tends towards zero. Thus, the conclusion is that for L_p metrics with p>=3 the NN search in a high dimensional space tends to be meaningless. In other words, distance loses its discriminative power between the notions of close and far. So, on average the ratio between the contrast and $d^{1/p-1/2}$ is bounded and these bounds depend on the value of p. Furthermore, if the number of points n is large, the upper bound may be very large too. This value is much closer though to the lower bound than to the upper bound.

C. Concentration of fractional norms

Aggarwal extended Hinneburg's result to fractional p-norms (François et al., 2007; Aggarwal et al., 2001). The theorem states that if P^j 1<=j<=n are n d-dimensional independent random vectors distributed over $[0,1]^d$ then there exists a constant C independent of p and d such that

$$C\sqrt{\frac{1}{2p+1}} \le \lim_{d\to\infty} \mathrm{E}\left(\frac{\max_j \|\mathrm{P}^{(j)}\|_p - \min_j \|\mathrm{P}^{(j)}\|_p}{\min_j \|\mathrm{P}^{(j)}\|_p}\right)\cdot\sqrt{d} \le (n-1)\cdot C\cdot\sqrt{\frac{1}{2p+1}}.$$

Aggarwal notes that the constant sqrt(1/(2p+1)) may play a valuable role in affecting the relative contrast and confirmed it experimentally with synthetic data sets. It was also concluded that on average fractional norms provide better contrast than Minkowski norms in terms of relative distance. Finally, Skala (2005) showed that the ratio

$$\rho_p(d) = \frac{\mathrm{E}(\|X\|_p)^2}{2\mathrm{Var}(\|X\|_p)},$$

increases linearly with the dimension d. Here X is a random vector whose components are independent and identically distributed.

IV.EXPERIMENTS

Figures 3 and 4 demonstrate how the TPIE KDB-tree (Arge et al., 2002) behaves when the data set size is 20,000 and 1,000,000 points and we perform range queries that contain the number of points shown (of course with the relevant side length in each dimension).

As low as in 8 dimensions TPIE KDB-tree must visit all the created nodes in order to find the desired number of points. This result demonstrates the appearance of the dimensionality curse phenomenon, since a plain sequential scan is more efficient than using the KDB-tree. When the dataset is 1,000,000 points this phenomenon occurs when the dimensionality is 16.

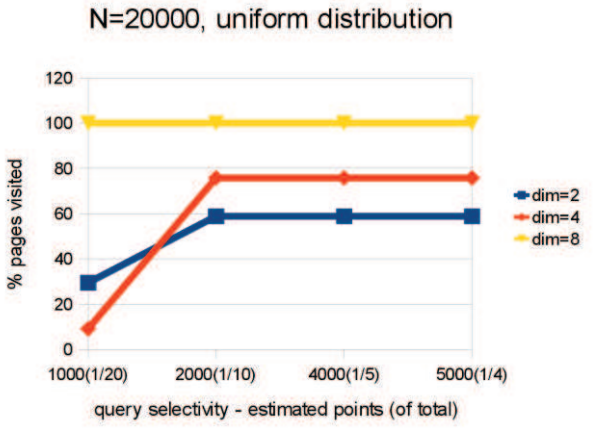

Figure 3. Percentage of visited pages for varying query selectivity and dimensionality (N=20000)

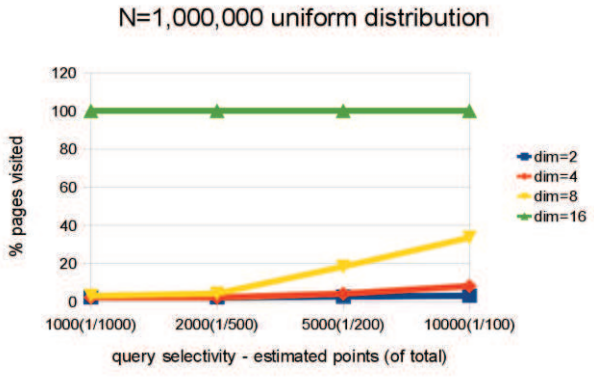

Figure 4. Percentage of visited pages for varying query selectivity and dimensionality (N=1,000,000)

V. CONCLUSIONS

In this paper we reviewed in depth the current findings on the study of high dimensional spaces. We gave many different explanations of the notion of the dimensionality curse. Finally, we demonstrated how the KDB-tree behaves in low to medium dimensions and how the dimensionality curse appears even in low dimensions and small database sizes.

REFERENCES

Charu C. Aggarwal: Re-designing Distance Functions and Distance-Based Applications for High Dimensional Data. SIGMOD Record 30(1): 13-18 (2001a)

Charu C. Aggarwal, Alexander Hinneburg, Daniel A. Keim: On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. ICDT 2001: 420-434

Lars Arge, Octavian Procopiuc, Jeffrey Scott Vitter: Implementing I/O-efficient Data Structures Using TPIE. ESA 2002:88-100

Jon Louis Bentley: Multidimensional Binary Search Trees Used for Associative Searching. Commun. ACM (CACM) 18(9):509-517 (1975)

Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, Uri Shaft: When Is "Nearest Neighbor" Meaningful? ICDT 1999: 217-235

Pierre Demartines, "Analyse de Donnees par Reseaux de Neurones Auto-Organises," PhD dissertation, Institut Nat'l Polytechnique de Grenoble, Grenoble, France, 1994 (in French)

Damien François, Vincent Wertz, Michel Verleysen: The Concentration of Fractional Distances. IEEE Trans. Knowl. Data Eng. 19(7): 873-886 (2007)

Michel Ledoux: The Concentration of Measure Phenomenon. American Mathematical Society 2001

John Robinson: The K-D-B-tree: a search structure for large multidimensional dynamic indexes, Sigmod 1981

M. Skala, "Measuring the Difficulty of Distance-Based Indexing,"Proc. 12th Int'l Conf. String Processing and Information Retrieval (SPIRE '05), M.P. Consens and G. Navarro, eds., pp. 103-114, Nov.2005

Roger Weber, Hans-Jörg Schek, Stephen Blott: A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. VLDB 1998: 194-205