THE INFINITE DISTANCE IN THE DETERMINATION OF THE NEAREST EUCLIDEAN M-NEIGHBOURS IN THE K-D-B TREE

Santana, O.; Rodríguez, G., Díaz, M., Plácido, A.

Department of Informatics and Systems
University of Las Palmas
P.O. Box 550. Las Palmas de G.C.. Spain.

ABSTRACT

In this article the search scheme of the nearest m_neighbours in the **K_D_B** tree structure is proposed. In that scheme two different strategies for the selection of alternative descent branches, that determine the order in which the criterion of the pruning tree is studied is planned. An experimental study, with the euclidean an infinite distances, in order to comparing both strategies, as soon as the influence of the distance change is realized. By last, three search schemes of the euclidean m_neighbours via the infinite distance, with the objective of improving the obtained performance with the euclidean distance are proposed for its following discussion.

INTRODUCTION

Robinson, J.T. [4] presents the **K_D_B** tree structure as a solution to the problem of retrieving points in a multidimensional space via range queries from a dynamic index. Lately, M. Díaz, O. Santana and others [1] introduced different local reorganizations to optimize the occupation in the updating of the **K_D_B** tree structure.

In this article, the search of the nearest m_neighbours in the **K_D_B** tree with the local reorganization that optimizes it, is planned. The overlapping and contention tests [3] that avoid the exploration of all the nodes of the tree are applied. Two strategies that indicate the order in which has to be verified the overlapping proof to the regions of the region page are compared. The first one according to the pre-established order in that page and the second one through a previous sort in function of the distance of each region to the query point.

In the experimental study two types of distances are used: euclidean and infinite, comparing the two selection forms of alternative descent branches with each distance and between distances. The inclusion relation that exists between hyperspheres of same radius for both distances, permit to present the search of the nearest euclidean m_neighbours through the corresponding hypersphere defined by the infinite distance [2]. Three schemes that correspond to the united or separated use of those hyperspheres in the region pages, and exclusively of united form in the point pages are studied.

All of this is separated in the detailed sections immediately. The structure of the **K_D_B** tree is described in section I. The two types of nodes that exists in the tree require to manage distances between points and from a point to a region, as is indicated in section II. The search process will have finished when the contention proof of the hypersphere showed in section III is verified. In section IV, the two strategies followed in the overlapping proof is presented. The search algorithmic of the nearest m_neighbours is developed in section V. The section VI is dedicated to describe the three search schemes of the euclidean m_neighbours through the infinite distance. Finally, the experimental study and the conclusions are presented in sections VII and VIII respectively.

1984/89/0000/0146$01.00 © 1989 IEEE

I.- K-D-B TREE

A point is defined to be an element of

$$dom_1 \times dom_2 \times \ldots \times dom_{nd},$$

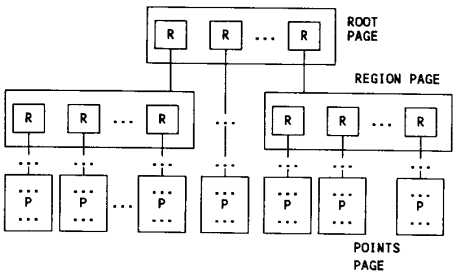
and a region to be the set of all points $X=(x_1,x_2,\ldots,x_{nd})$ satisfying:

$$min_i <= x < max_i;\ 1 <= i <= nd$$

for some collection of min_i, max_i belonging to dom_i.

There are two types of nodes in a **K_D_B** tree:

- Region pages: region pages contain a collection of (region, link) pairs, where each link points to a not empty page.
- Point pages: point pages contain a collection of points.



The following set of characteristics define the **K_D_B** structure:

- It is a multiway tree that keeps itself perfectly balanced.
- In every region page the regions in the page are disjoint, and their union is a region. If the root page is a region page, the union of its regions is the total domain, dom.
- If in a region page, the referred child page by the associated link to **"region"** is a region page, then the union of the regions in the child page is **"region"**, in other case if it is a point page, all the points of the page are contained in **"region"**.

II.- THE NEAREST M-NEIGHBOURS. DISTANCES

Given a collection of points, a distance and a query, **q**, the nearest m_neighbours to **q**, are the m_points contained in the search hypersphere of minimum radius with centre on **q**.

The search of the nearest m_neighbours in the **K_D_B** tree, given the nature of its nodes, require the calculation of distances: between points and from a point to a region.

The distances between points used in this study are euclidean and infinite.

Euclidean and infinite distances, from a point to a region are defined by the coordinate distances as follow: Given a point, X, and a region, $R=\{min_i, max_i\}$ $1\leq i\leq nd$, the **nd** functions $\mathbf{F_i(X,R)}$ are defined as follow:

$$F_i(X,R) = \begin{cases} min_i - x_i & \text{if } x_i < min_i \\ x_i - max_i & \text{if } x_i > max_i \\ 0 & \text{in other case} \end{cases}$$

are called coordinate distances from a point to a region.

The euclidean distance from X to R:

$$E(X,R) = \left(\sum_{i=1}^{nd} (F_i(X,R))^2\right)^{(1/2)}$$

and the infinite distance from X to R:

$$I(X,R) = \max_{1\leq i\leq nd} (F_i(X,R))$$

III.- CONTENTION PROOF

It is a proof that determines if the search hypersphere is contained in a region. The separation between the query and the region limits for each coordinate, are compared with the radius.The proof fails as quick as some separation $\mathbf{q_i - min_i}$ $(1\leq i\leq nd)$, were bigger than the radius; or well, some separation $\mathbf{max_i - q_i}$ $(1\leq i\leq nd)$, were bigger or equal than the radius. The proof finalizes successfully when the situation of the query respect to the region limits transgress those conditions.

IV.- CRITERIONS OF SELECTION OF ALTERNATIVE REGIONS. OVERLAPPING PROOF

The purpose of the overlapping proof is to determine if the search hypersphere overlaps with a region. Two strategies that indicate the sequence of application of this proof are studied.

The first one, **S** strategy, follows the location sequence of the regions in the region page, studying for each region its possible overlapping. For it, the distance between the query and the region is determined. There is not overlapping with the region if that distance is bigger than the radius.

The second one, O strategy, makes a previous sort of the regions, in function of the distance of the query to each one of them, testing for each region the relation between that distance and the radius of the hypersphere; so the alternatives from bigger to less proximity to the interrogation are examined and as quick as the overlapping proof fails the inspection is stopped, not being necessary to study the remaining regions. Experimentally, as it will be seen in the corresponding section, this strategy, prevails over the previous one and because of it, it is applied in the search schemes of the euclidean m_neighbours.

V.- SEARCH OF THE NEAREST M-NEIGHBOURS IN THE K-D-B TREE

The global approximation begins from the root page. When a region page is reached it is determined which is the region, **i**, that contains the query, **q**, and it is gone down by the associated link, repeating this process until reach a point page.

In the local approximation, the m_first found points define the hypersphere of initial search, which radius is given by the distance between **q** and the further point. In the return of the initial descent it is tested if the search has finalized through the contention proof in **i**, if not, the selection process of alternative regions acts through one of the proposed strategies. If there is overlapping it is gone down by the associated branch, repeating the selection and overlapping process in each new region page until reach a point page in which it is able to update the searching radius.

VI.- EUCLIDEAN M-NEIGHBOURS THROUGH THE INFINITE DISTANCE

The euclidean distance cost, from a point to a region, or between points, is bigger than the infinite distance cost. Also, for equal radius, the defined hypersphere by the infinite distance contains the hypersphere defined by the euclidean distance. It is possible to determine a search scheme of the euclidean m_neighbours making use of these relations with the infinite distance?.

Three search schemes of the nearest euclidean m_neighbours, in which is realized a filter with the infinite distance in the point pages, and that defer in the treatment of the region pages are studied: in the first one it is realized also a filter of the candidate regions to the Euclidean Overlapping through the study of the Infinite Overlapping, **SESI**, and in the second one and third one it is treated with a single distance, Euclidean Overlapping, **SE**, and Infinite Overlapping, **SI**, respectively.

a) Common features of the three schemes:
- The radius of the search hypersphere, **re**, is defined with the euclidean distance.
- When a point page is reached the infinite distance to each point is calculated, if it is less or equal than the radius the euclidean distance is calculated, updating it in the necessary case.

b) Specific features:

b.1) **SESI** search scheme:
- In the region selection process, these regions are sorted in function of the infinite distance, testing for each

region if its infinite distance to the query is less or equal than **re**, when it happens a new overlapping proof with the euclidean distance is made.

b.2) **SE** search scheme:

- The regions are sorted in function of the euclidean distance to the query and this distance is compared with the radius, **re**, to verify the overlapping proof.

b.3) **SI** search scheme:

- It is similar to the **SE** scheme establishing the order with the infinite distance.

VII.- EXPERIMENTAL RESULTS

The experimental study of the performance of the nearest m_neighbours search in the **K_D_B** tree is realized measuring the average distances, that is necessary calculate, from the query to a point or to a region, in function of: the dimension, **nd**, the maximum size of the point page, **tmpp**, the number of neighbours to retrieve, **m**, and the distance used. The study is centralized in a value of the maximum size of the region pages, **tmpr**, because it is pendent of resolution some form of local restructuration to optimize the occupation in those pages.

Nine trees of 10.000 points each one, which features are showed in the table 1, are constructed. Associated to each one of then, a file of 1.000 query points, is created. All the points are got through a function that generates pseudo-random numbers of uniform distribution in [0,1], based in the mixed congruent method.

Table 1

nd	tmpp	tmpr
2, 4, 6	5,10,15	5

VII.A.- Comparative Study of the Criterions of Alternative Regions Selection

The results got for the infinite distance, that are showed in figures 1 and

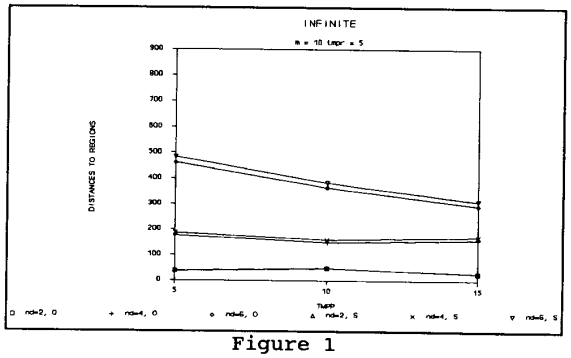


Figure 1

2, indicate that a less number of calculated distances , to regions and points, with the

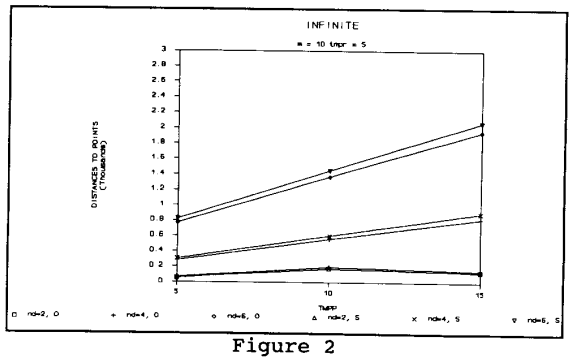


Figure 2

O strategy than the ones got with the **S** strategy, is evaluated. This difference is maintained constant when **tmpp** varies. Although, the regions sort cost must not be forgotten. The value of **tmpr** used in this study is low and because of it, the sort cost is not considered.

These results represent the exploration of a little portion of the tree. So, the maximum exploration is got for nd=6, m=10, with the **S** strategy, tmpp=15, tmpr=5, where

a 20% of the point pages and a 23% of region pages are explored. The number of region pages and the number point pages visited decrease lightly for each dimension when **tmpp** increases. For a bigger dimension, more point pages are visited and more region pages are acceded, because of exists a major dispersion of the points.

With the euclidean distance a performance similar on quality to the infinite distance behaviour is got. But in this case, quantitatively, the results are higher.

VII.A.I.- Variation of the Nearest Neighbours Number

The increase of the number of neighbours to retrieve with the infinite distance, produces an increase of the number of point pages visited and of the number of region pages acceded, and so that for a major number of distances calculated, to regions or points, as it is showed in figure 3. The

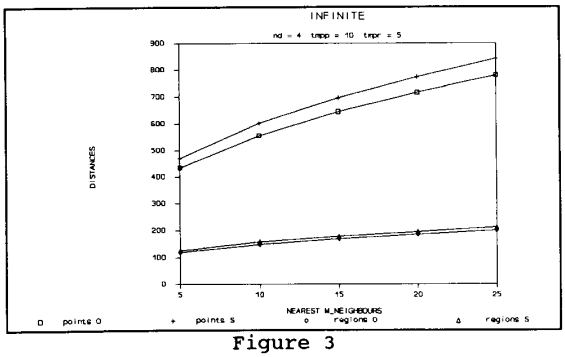


Figure 3

advantage of the second criterion of regions selection for the measured parameters is maintained. The euclidean distance doesn't improve these results.

According to the obtained results in the comparison, the second criterion of regions selection will be used for the realization of the proofs.

VII.B.- Study of the Search Schemes of the Euclidean M-Neighbours Via the Infinite Distance

The classic search scheme of the nearest m_neighbours with the euclidean distance, **E**, is compared with the three schemes that use a filter in base to the infinite distance described in the section 6: **SESI**, **SE** y **SI**.

The table 2 shows the relation got between the cost of the infinite distance from a point to a region , **cri**, and the euclidean distance, **cre**. So shows that relation for the distances between points, **cpi** and **cpe** respectively.

	nd = 2	nd = 4	nd = 6
cri/cre	0.1463	0.203651	0.246021
cpi/cpe	0.118001	0.167852	0.1969

Table 2

To compare the results, the number of equivalent euclidean distances, **deeq**, corresponding to the number of infinite distances considering the cost reasons of the table 2, is evaluated; so can be added to this value of **deeq** the number of euclidean distances involved in the adaptive filter, to get the equivalent total euclidean distances to regions and points; that constitute an acceptable measure of the

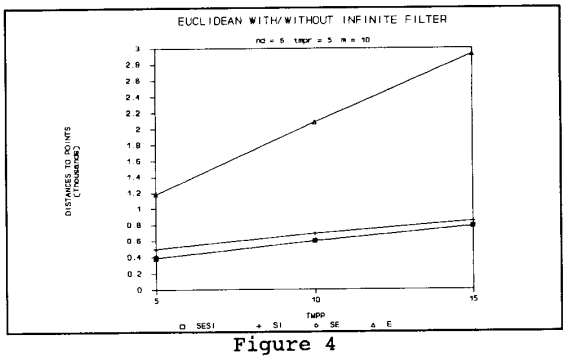


Figure 4

realization of the search scheme.

Initially, the **SESI** search scheme, that

improves the number of distances calculated in the point pages respected to the **E** scheme, is proposed, figure 4. Although, when the transformation of infinite distances to regions to their euclidean equivalents is made, it is verified that,

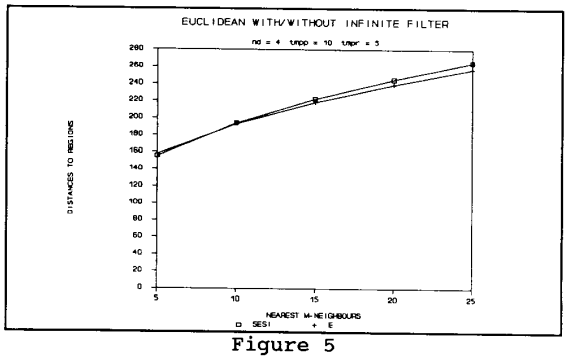


Figure 5

for a major number of neighbours to retrieve and for high dimensions, it is calculated more than with the **E** scheme, figures 5 and 6.

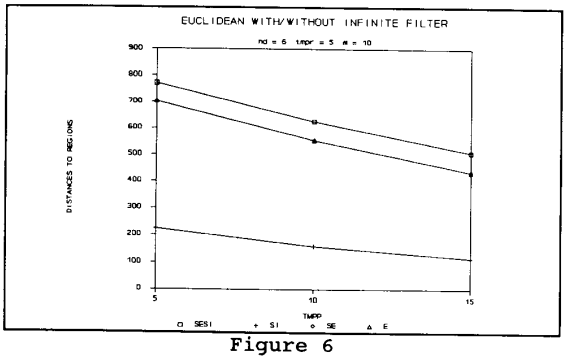


Figure 6

The number of point pages visited and the number of region pages acceded is practically the same in both schemes. The sort of the regions by the infinite distance, for a little value of **tmpr** as the one studied, doesn't cause a substantial change of the results.

The **SE** scheme maintains the efficiency in the point pages of the **SESI** and in the region pages has an equal behaviour to **E**, figures 4 and 6.

In order to get a best performance in the treatment of the region pages during the search, the **SI** scheme is introduced with the inconvenient of losing a little part of the reached efficiency in the point pages. The results showed in the figures 4 and 6, present that this lost of efficiency is perfectly compensated by the reached improvement in the distances to regions. The number of point pages visited and the number of region pages acceded in **SI** is major than in the other schemes, like was supposed. Although, the global behaviour of the search is not affected, because of the low cost of the infinite distances. So the scheme **SI** is the one that offers the best realization in the retrieval of the nearest euclidean m_neighbours.

VIII.- CONCLUSIONS

It has been planned the search scheme of the nearest m_neighbours in the **K_D_B** tree, realizing an experimental study of it with the infinite distance and the euclidean distance.

In the pruning process of the tree, during the search, with the strategy that makes a previous sort of the regions it is got to decrease the number of distances calculated to regions and points. Although, the sorting cost of the regions musn't be forgotten.

The low cost of the infinite distance respect with the euclidean and the inclusion relation that exists between the hyperspheres of equal radius, conduce to study the search problem of the euclidean m_neighbours using the infinite distance as filter. Therefore, the **SI** scheme is the one that presents a best performance of the three proposed schemes for the retrieval of the nearest euclidean m_neighbours.

REFERENCES

[1] Díaz M.; Santana O.; Rodríguez G.; Martín M.: "Reorganizaciones Locales en el Arbol K-D-B. Su Eficiencia en Situaciones Dinámicas". XIV Conferencia Latinoamericana de Informática, Vol. I, 17/32, 1988.

[2] Santana O.; Díaz M.; Rodríguez G.; Ródriguez N.: "Effect of Distance over M_Nearest Neighbours Search Performance on the Burkhard_Keller tree". Prepared to be published.

[3] Friedman J. H.; Bentley J. L.; Finkel R. A.: "An Algorithm for Finding Best Matches in Logarithmic Expected Time". ACM Transactions on Mathematical Software. Vol. 3, N° 3, 209/226, 1977.

[4] Robinson J. T.: "The K-D-B Tree: A Search Structure for Large Multidimensional Dynamic Indexes". Departament of Computer Science, Carniege-Mellon University Pittsburgh, Pennsylvania 15213, 1981.