

SOFSEM 2007

Current trends in theory and practice of computer science

Harrachov, Czech Republic. 23rd January 2007


Spatial selection of sparse pivots for similarity search in metric spaces

Oscar Pedreira, Nieves Brisaboa



University of A Coruña (Spain)

Outline

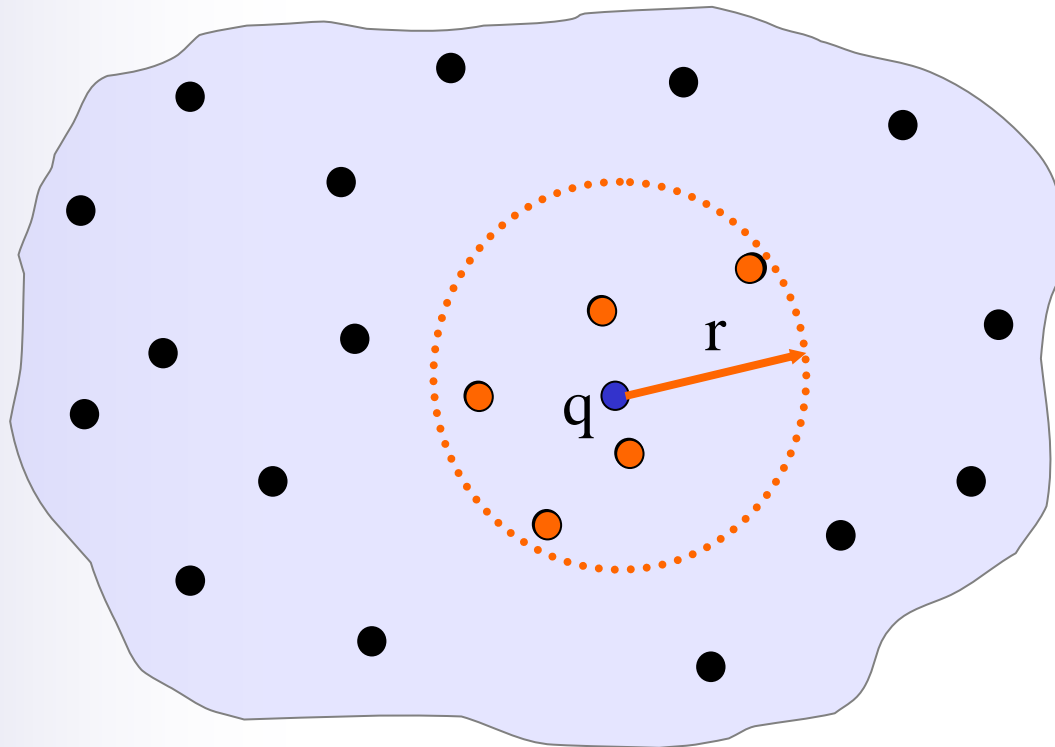
- 
1. Introduction
 2. Sparse Spatial Selection
 3. Experimental results
 4. Nested Metric Spaces
 5. Conclusions and future work

Introduction

- Traditional databases
 - Data has a well-defined structure.
 - Exact searching, using equality/inequality comparisons.
 - `SELECT name FROM Student WHERE city = 'Harrachov';`
- Non-structured databases
 - Exact searching is no possible.
 - *Similarity search* is a very common operation in several application domains.
- Some examples
 - Retrieval of texts, sound, video or fingerprints, computational biology, pattern recognition, etc.

Introduction

- Similarity search: retrieve from the database all the objects close (similar) to one given as a query.

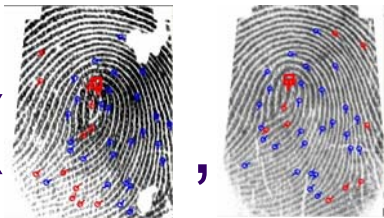


Introduction

- Distance function. Examples:

$$d(\text{latest}, \text{greatest}) = 3$$

g r e a t e s t
l a t e s t


$$d(\text{fingerprint}_1, \text{fingerprint}_2) = 1.5435$$

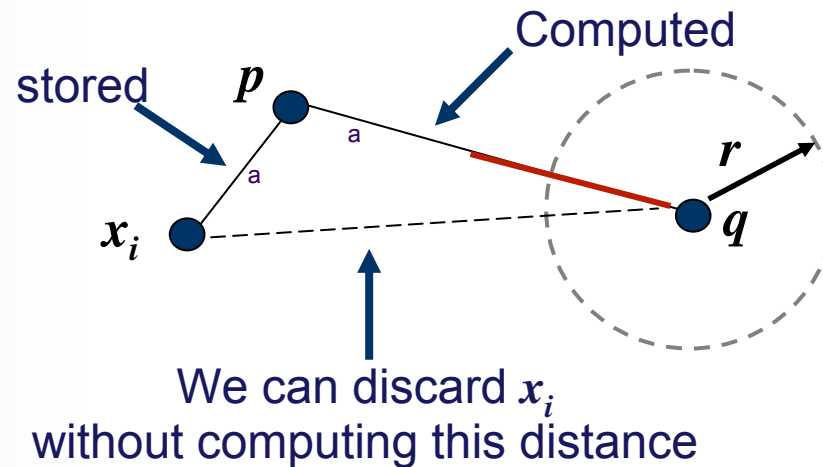
- The evaluation of d has a high computational cost. The comparison of the query with the whole database is TOO expensive



- Indexes** are built over the DB to avoid the comparison of the query with all the objects.

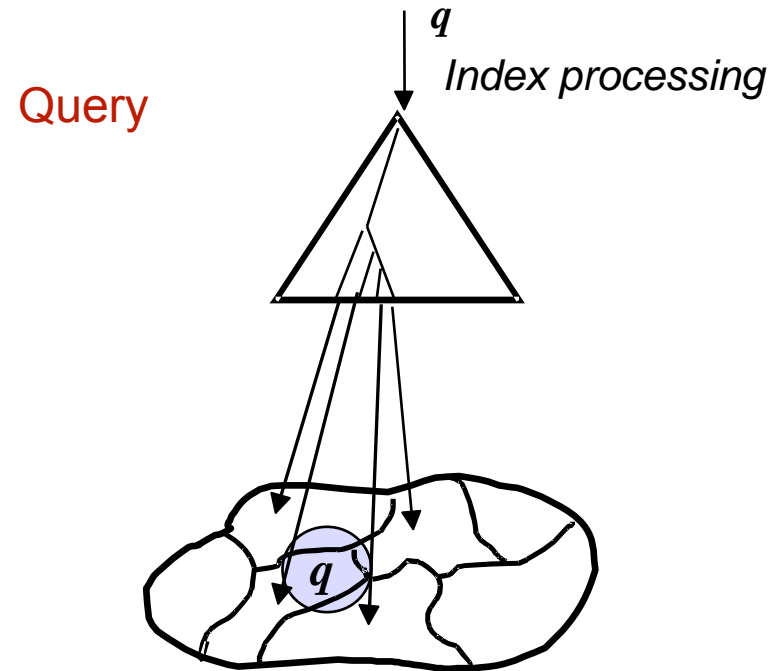
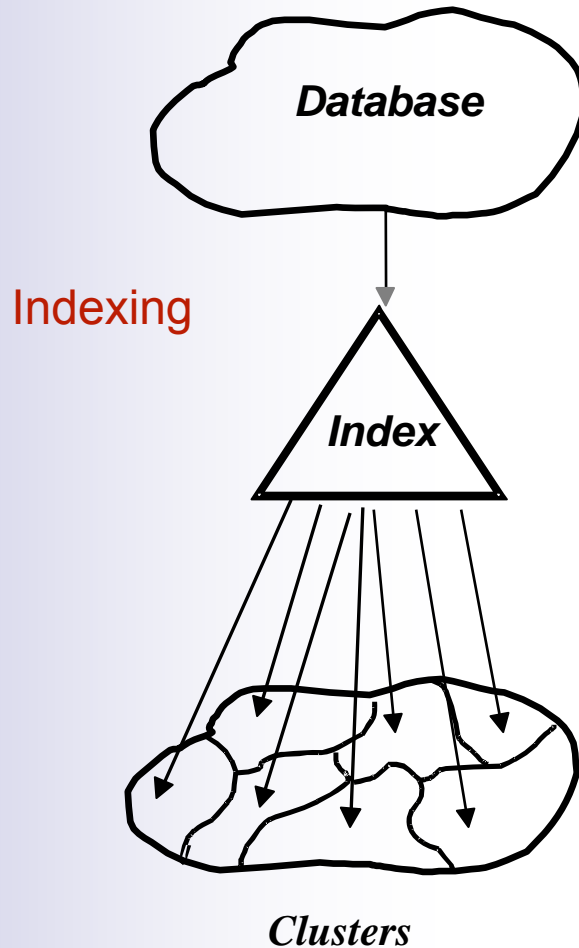
Introduction

- Metric space = Object universe + Distance function
 - E.g. Collection of words + Edit distance
- The triangle inequality, base of the indexing algorithms.
 - $\forall x, y, z \in \mathbf{U}, d(q, x) \geq d(q, p) - d(p, x)$



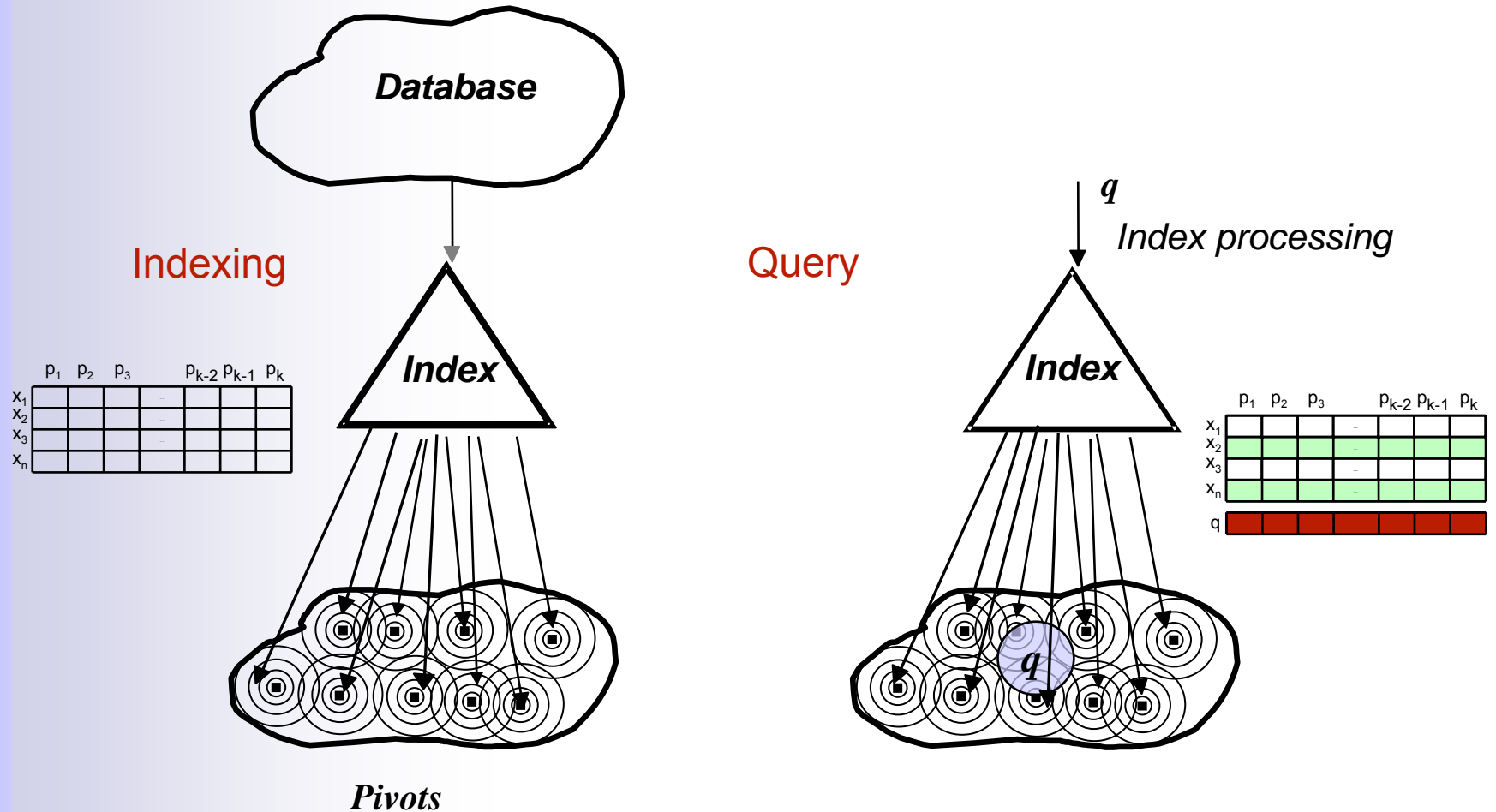
Introduction

- Clustering-based methods.



Introduction

- Pivot-based methods.



Introduction

- Clustering-based methods.
 - Bisector Tree.
 - Generalized-hyperplane Tree.
 - Geometric Near-neighbor Access Tree.
 - Voronoi Tree.
 - M-Tree.
 - Spatial Approximation Tree.
- Pivot-based methods.
 - Burkhard-Keller Tree.
 - Fixed-Queries Tree.
 - Fixed-Queries Array.
 - Vantage Point Tree.
 - Multi-Vantage Point Tree.
 - AESA.
 - Linear AESA.
 - **Sparse Spatial Selection**

Some methods were due to:

- Ricardo Baeza-Yates
- Remco Veltkamp

Introduction

- Pivot selection strategies.
 - Some have been proposed in previous work
 - ▶ *Selection* ● Bustos B., Navarro G., Chávez E. Pivot Selection Techniques for proximity search in metric spaces. In *Proceedings of SCCC 2001*. IEEE Press.
 - ▶ *Incremental*
 - ▶ *Local Optimum*
- How many pivots do they use ?
 - This number has a great influence in the search efficiency.
 - In all the existing techniques this number has to be fixed before the index construction.

Outline

- ✓ 1. Introduction
- ➔ 2. Sparse Spatial Selection
3. Experimental results
4. Nested Metric Spaces
5. Conclusions and future work

Sparse Spatial Selection

This work proposes a new and original method:

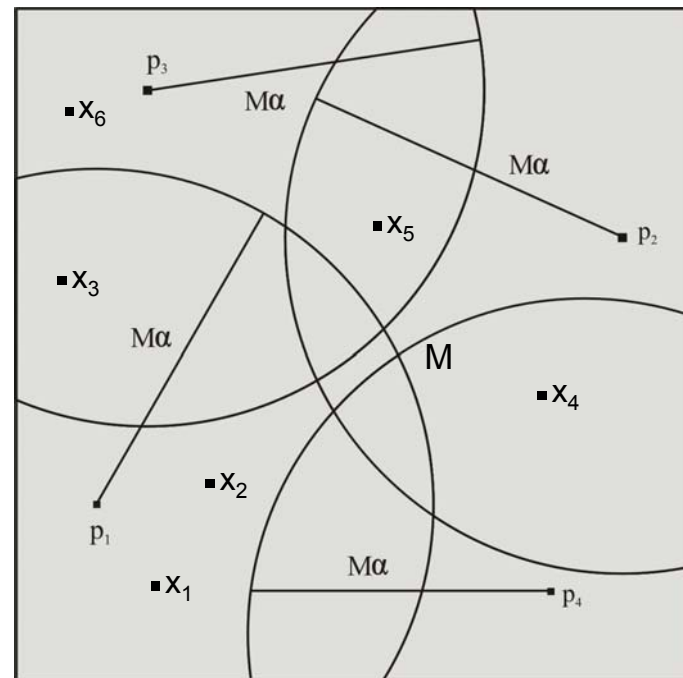
- A **pivot**-based method.
- The **number of pivots** is selected automatically.
(The algorithm sets itself the number of pivots that will be used.)
- **Dynamic** (the database can be initially empty).
- **Easier** to build than in previous techniques
- Very **efficient** in the search.
- Efficient storage in **secondary memory**.
- Useful with continuous distances

Sparse Spatial Selection

- Pivot selection strategy.
 - An object is chosen as pivot iff its distance to the all the current pivots is greater than $M\alpha$.

- M maximum distance
- $0 < \alpha < 1$

$\alpha = 0.5$



Sparse Spatial Selection

- Once the pivots are selected, the index is created.
- For example, a simple implementation ...



$\{x_1, x_2, \dots, x_n\}$



$\{p_1, p_2, \dots, p_k\}$

	p_1	p_2	p_3	...	p_{k-2}	p_{k-1}	p_k
x_1	1.3542	1.5362	2.4473	...	0.3834	3.2938	1.2532
x_2	2.3645	3.8472	2.7364	...	2.7363	3.8756	1.2837
	⋮	⋮	⋮	...	⋮	⋮	⋮
x_n	2.7463	1.2937	2.9384	...	2.8374	2.8464	1.9876

Outline

- ✓ 1. Introduction
- ✓ 2. Sparse Spatial Selection
- ➔ 3. Experimental results
4. Nested Metric Spaces
5. Conclusions and future work

Experimental results

- Test collections
 - Synthetic vector spaces
 - ▶ Collections of 1.000.000 vectors of dimensions 8,10,12, y 14, using the Euclidean distance. Uniform distribution.
 - Collections of words (edit distance).
 - ▶ 69.069 words taken from the English dictionary.
 - ▶ 51.589 words taken from the Spanish dictionary.
 - Collections of images (Euclidean distance).
 - ▶ 47.000 images from NASA's archives, represented by feature vectors of dimension 20.

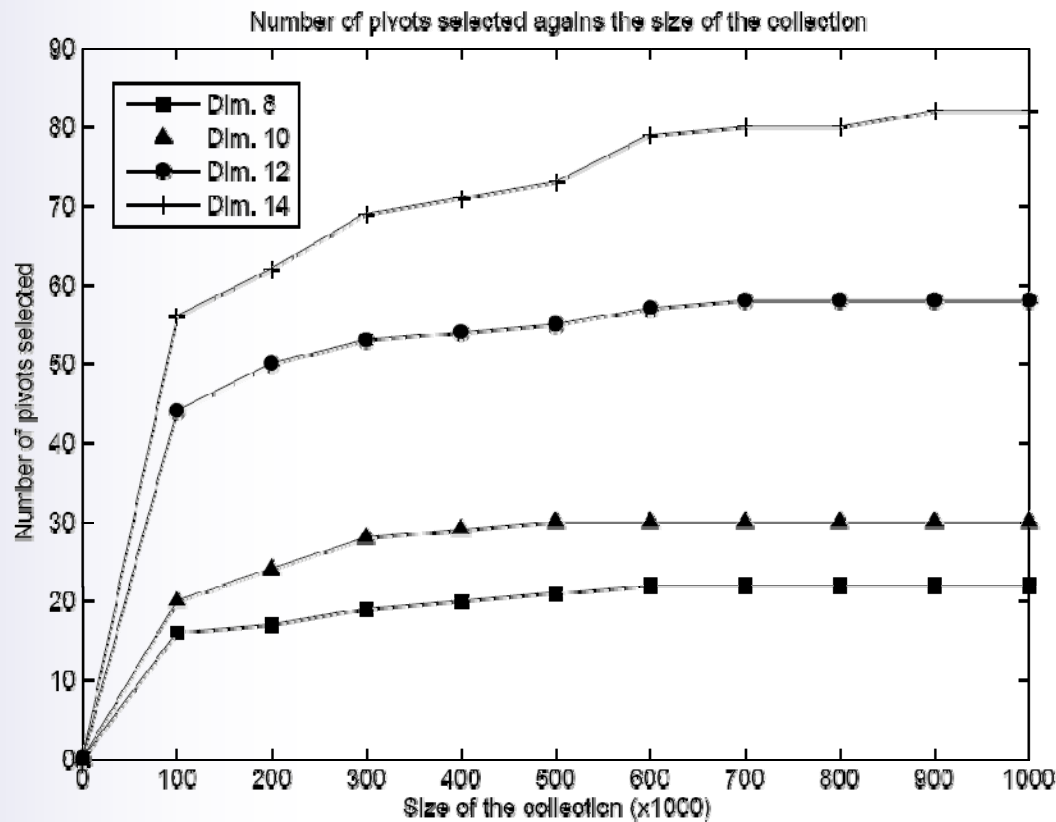
Experimental results

- Three hypothesis:

- ➔ ● The number of pivots does not depend on the collection's size, but on the space's intrinsic dimensionality. (Therefore, it should not increase indefinitely)
- The optimal values of α are in the range [0.35, 0.40].
- SSS is more efficient than other methods.

Experimental results

- Number of pivots selected by the algorithm vs. Size of the collection



Experimental results

- Three hypothesis:

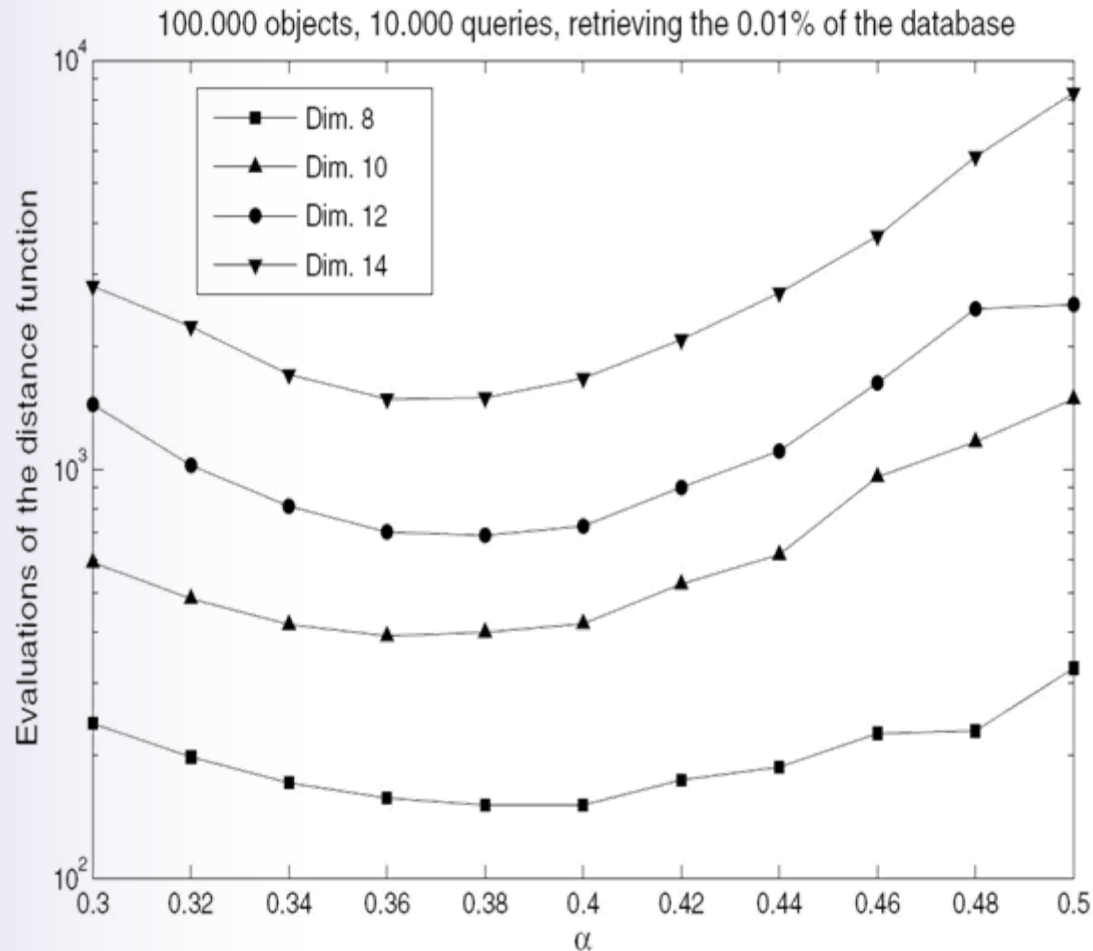
- ✓ ● The number of pivots selected should become stable in some moment because it does not depend on the collection's size, but on the space's intrinsic dimensionality.

➔ ● The optimal values of α are in the range $[0.35, 0.40]$.

- SSS is more efficient than other methods.

Experimental results

- Values of the parameter α .



Experimental results

- Three hypothesis

✓ 1. The number of pivots does not depend on the collection's size, but on the space's intrinsic dimensionality.

(Then, the number of pivots selected should become stable in some moment.)

✓ 2. The optimal values of α are in the interval [0.35, 0.40].

➔ 3. SSS is more efficient than other methods.

Experimental results

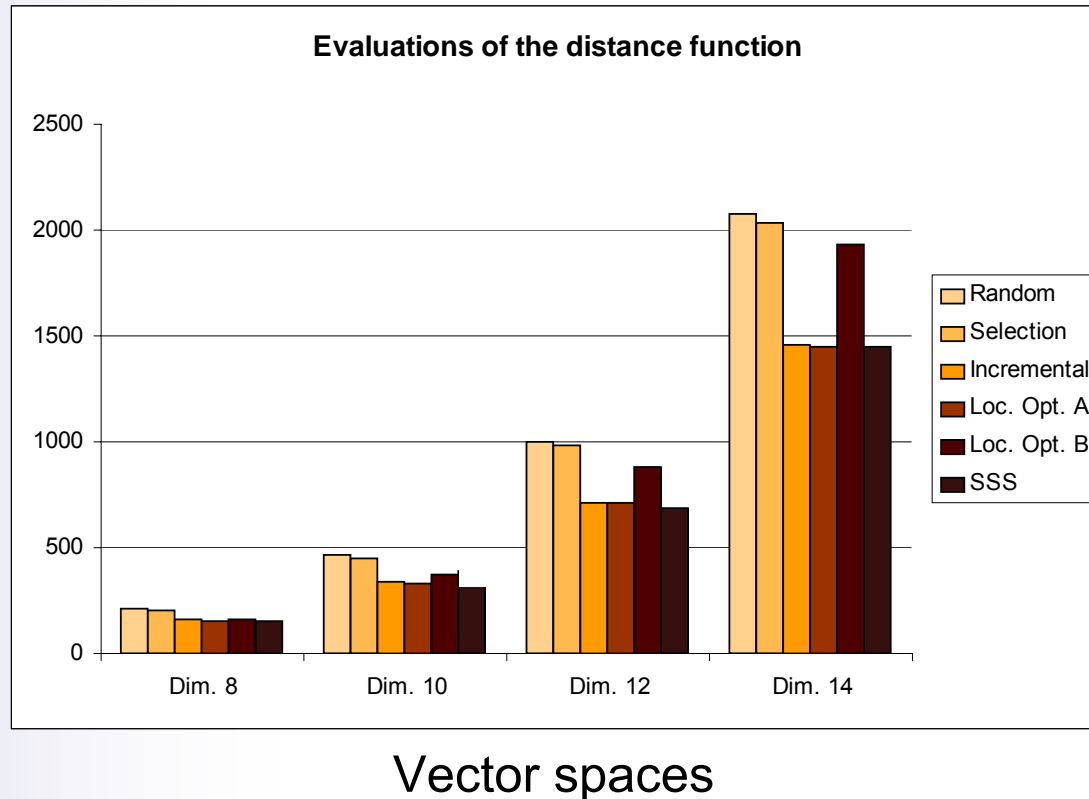
- Comparison with other techniques.
 - The proposed method has been compared with similar existing techniques.

Selection, Incremental, Local Optimum published in:

 - ▶ Bustos B., Navarro G., Chávez E. Pivot Selection Techniques for proximity search in metric spaces. In *Proceedings of SCCC 2001*. IEEE Press.
 - *Sparse Spatial Selection* is, in general, more efficient than the previous techniques, with an easier index implementation and construction.

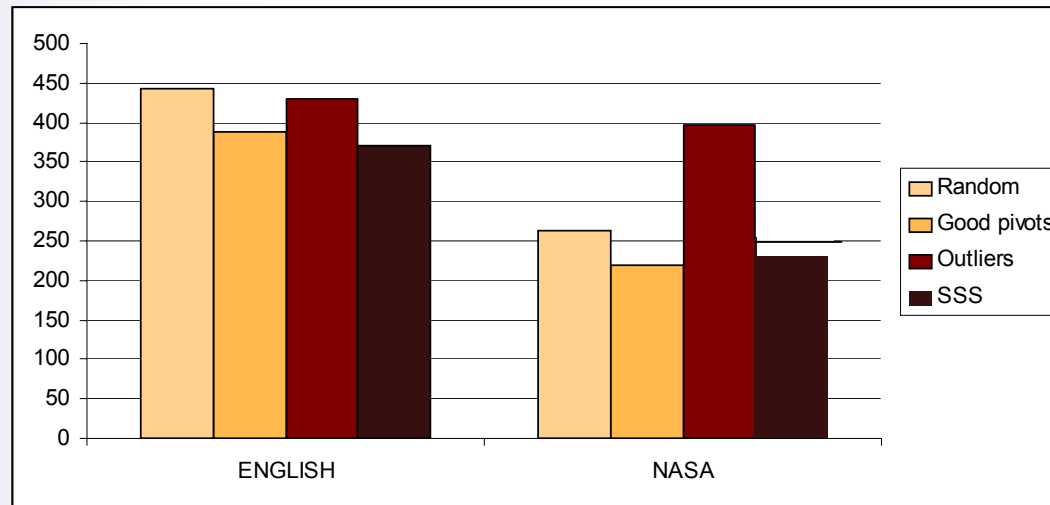
Experimental results

- Comparison with other techniques. Vector spaces:



Experimental results

- Comparison with other techniques. Collections of words and images:



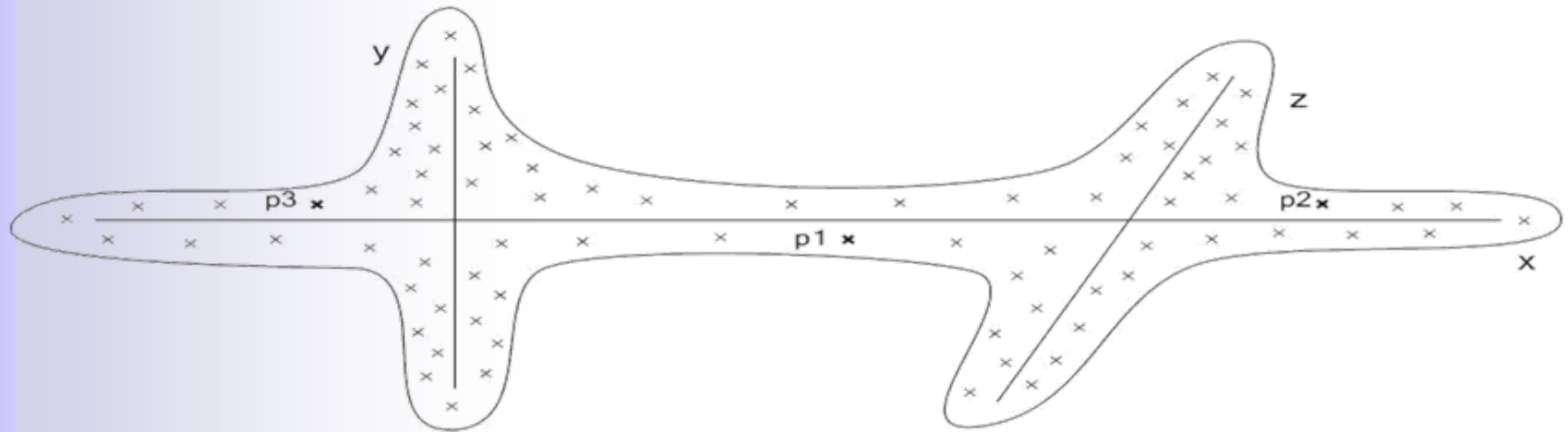
Outline

- ✓ 1. Introduction
- ✓ 2. Sparse Spatial Selection
- ✓ 3. Experimental results
- ➔ 4. Nested Metric Spaces
5. Conclusions and future work

Nested metric spaces

- Experiments with a collection of color images represented by feature vectors of dimension 112
- SSS performed **worst** than a ***random pivot selection*** !!!!!!!.
(Using the optimal number of pivot for random selection)
- *What is the reason for this (apparently) strange result ?*

Nested metric spaces



- There are small clusters very dense with specific dimensions
- Here, SSS is not able to select enough pivots in subspaces
- However, a random pivot selection can obtain some pivots from each cluster capturing their specific dimensions

Nested metric spaces

- Experiments with a collection of color images represented by feature vectors of dimension 112
- SSS performed **worst** than a ***random pivot selection !!!!!!!***.
- *What is the reason for this (apparently) strange result ?*

A new concept: Nested Metric Spaces

In some cases, the objects can be grouped in different clusters or subspaces, with different dimensions explaining the differences between objects in each of this subspaces embedded into a more general space.

Nested metric spaces

- How can we deal with this situations ?
- A possible solution consists in
 - Find the subspaces using clustering algorithms.
 - Apply SSS with a big α (0.5) to get general pivots
 - Then apply SSS in each subspaces with M adapted to that subspace to get specific pivots
- However, it can be difficult to adequately identify these subspaces, and to apply an indexing method in each of them efficiently.

Outline

1. Introduction ✓
2. Sparse Spatial Selection ✓
3. Experimental results ✓
4. Parallel processing ✓
5. Nested Metric Spaces ✓
- ➔ 6. Conclusions and future work

Conclusions

- This work presents a pivot selection strategy:
 - Dynamic
 - ▶ The database can be initially empty. Pivots are selected in an incremental way as the database grows.
 - The algorithm sets itself the number of pivots that will be used
 - ▶ Pivots are selected when they are needed to cover the space. The set of pivots adapts itself to the intrinsic dimensionality of the metric space.
 - Efficient
 - ▶ Experimental results show that this method is in most situations more efficient than previous proposals.

Future work

- New methods to deal with nested metric spaces
 - Subspace detection using clustering techniques
 - Application of adequate indexing methods in subspaces
- Dealing with efficiency.
 - Disk storage
 - Parallel indexing and searching.
- Refinements of the pivot selection strategy.



And that's all ...

Thanks for your attention