A dynamic pivot selection technique for similarity search

Benjamín Bustos
Center for Web Research, University of Chile (Chile)

Oscar Pedreira, Nieves Brisaboa
Databases Laboratory, University of A Coruña (Spain)

SISAP 2008
First International Workshop on Similarity Search and Applications
Cancún, México, 12 April 2008

Outline

1. Motivation
2. Previous work
3. Our method
   - Sparse Spatial Selection (SSS)
   - Non-Redundant Sparse Spatial Selection (NR-SSS)
4. Experimental results
5. Conclusions

Motivation

Pivot-based indexing algorithms

- Possible classification of indexing methods for similarity search:
  - Pivot-based indexes
  - Clustering-based indexes
- Pivot-based indexes:
  - Indexes are built from a set of reference points called pivots
  - The distances from the objects in the database to the pivots are computed and stored in an appropriate data structure
- Some well-known examples…
  - BKT, FQT, FQA, AESA, LAESA, etc.

Motivation

Why pivot selection techniques?

- The specific set of pivots affects the search performance
  - Which ones? Some algorithms select pivots at random, others with complex computations.
  - How can we find the optimal number of pivots? → Usually done by trial and error on the complete database, which makes the index static

Outline

1. Motivation
2. Previous work
3. Our method
   - Sparse Spatial Selection (SSS)
   - Non-Redundant Sparse Spatial Selection (NR-SSS)
4. Experimental results
5. Conclusions

Previous work

First heuristics for pivot selection (I)

- First works addressing the problem of pivot selection proposed heuristics that tried to select pivots far away from each other:
  - [Micó, Oncina, Vidal, 1994] proposes to choose pivots that maximize the sum of distances between pivots previously chosen.
  - [Yianilos, 1993] proposes a heuristic based on the second moment of the distance distribution, which selects objects far away from each other.
  - [Brin, 1995] proposes a greedy strategy that also selects objects far away from each other (though designed to select split points).

Previous work

[Bustos, Navarro & Chávez, 2003] (I)

- [Bustos, … 2003] addressed the problem of pivot selection in a formal way
- **They defined an estimator of the efficiency of a set of pivots based on a formalization of the problem**
- Using this estimator they proposed three techniques

Previous work
[Bustos, Navarro & Chávez, 2003] (II)

- Selection
  - N sets of random pivots are selected. The final set of pivots is the one maximizing the efficiency criterion.
- Incremental
  - The set of pivots is built incrementally, by adding to it the object maximizing the efficiency criterion.
- Local Optimum
  - The set of pivots is iteratively improved by replacing the worst pivot for a better one.

Previous work

Problems of the previous techniques for pivot selection

- In previous techniques the optimal number of pivots has to be obtained by trial and error using the complete database
- Insertions, updates and deletions of objects can reduce the index performance

This makes the index static

Outline

1. Motivation
2. Previous work
3. Our method
   - Sparse Spatial Selection (SSS)
   - Non-Redundant Sparse Spatial Selection (NR-SSS)
4. Experimental results
5. Conclusions

Our method

Sparse Spatial Selection [Brisaboa & Pedreira, 2007] (I)

- Sparse Spatial Selection [Brisaboa, et. al 2006] dynamically selects a set of pivots adapted to the intrinsic complexity of the space
- More efficient than previous techniques
- Dynamic and adaptive

Our method

Sparse Spatial Selection [Brisaboa & Pedreira, 2007] (II)

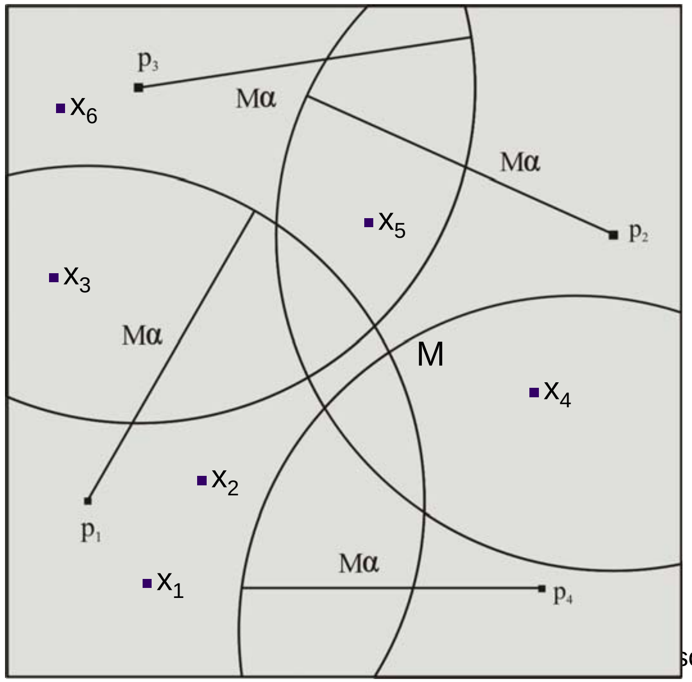
- When an object is inserted, it is selected as a new pivot if it is far away enough from the current pivots

- The object is considered "far-away" if its distance to the current pivots if greater than $M\alpha$

M maximum distance

$0 < \alpha < 1$

$\alpha = 0.5$



Our method

Sparse Spatial Selection [Brisaboa & Pedreira, 2007] (III)

Database

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

Pivots

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

	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

Our method

Sparse Spatial Selection [Brisaboa & Pedreira, 2007]

- SSS was experimentally validated, showing that

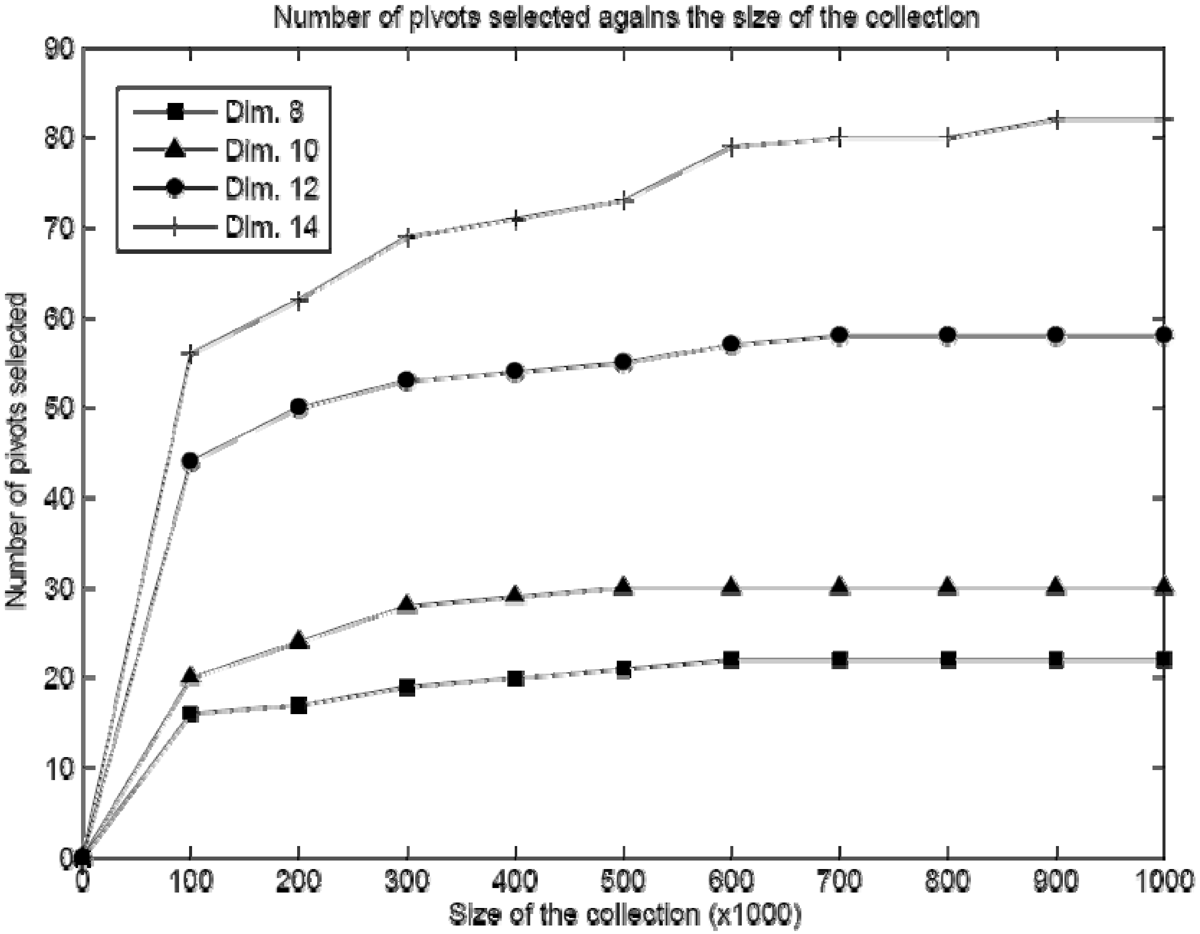
  1. The number of pivots does not depend on the collection’s size, but on the space’s intrinsic dimensionality.
     (Then, the number of pivots selected should become stable in some moment.)

  2. The optimal values of *α* are stable

  3. SSS outperforms state-of-art strategies.

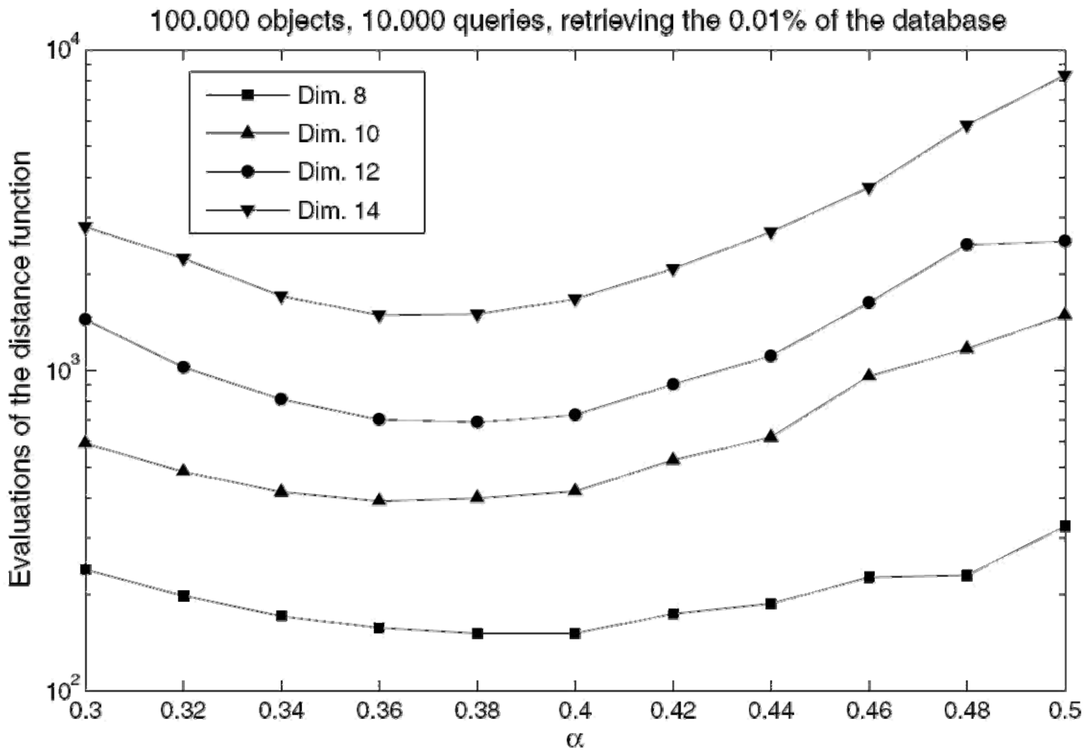
Our method

Sparse Spatial Selection [Brisaboa & Pedreira, 2007] (IV)



Our method

Sparse Spatial Selection [Brisaboa & Pedreira, 2007] (V)



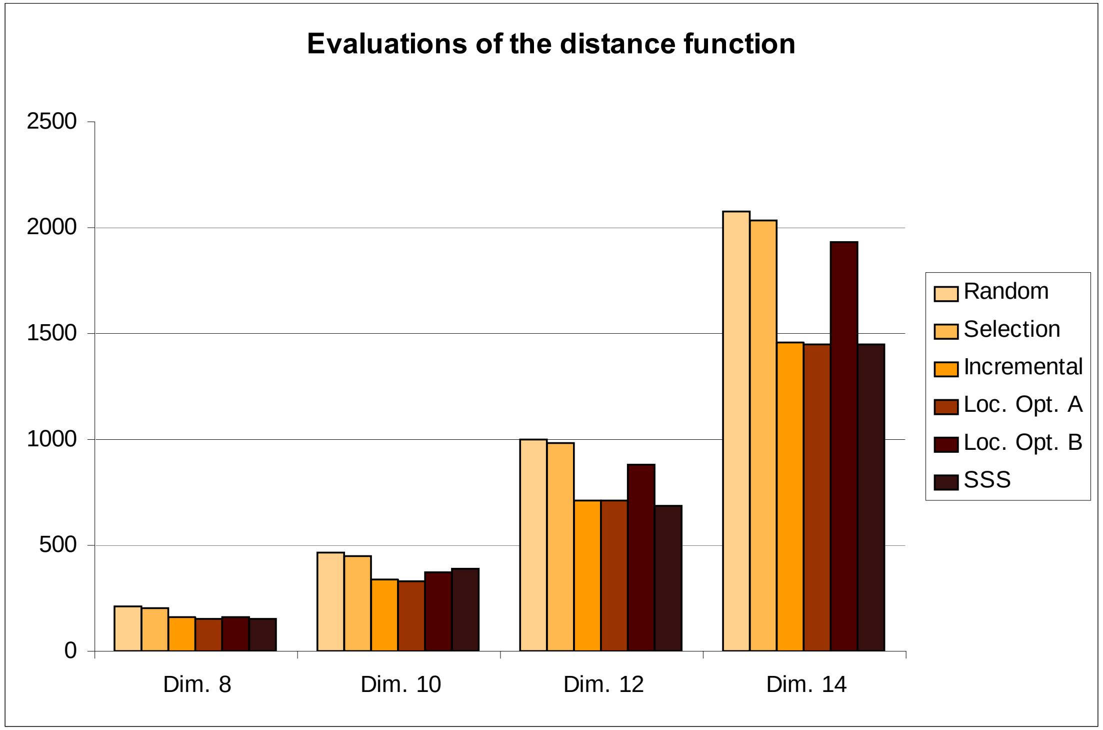
Our method

Sparse Spatial Selection [Brisaboa & Pedreira, 2007] (VI)

DB	μ	σ^2	Int. dimens.	α	pivots	α	pivots
English	8.239141	5.277638	6.085550	0.5	108	0.44	205
Spanish	8.272277	6.014831	5.688486	0.5	64	0.44	124
K = 8	1.043901	0.125227	4.351026	0.5	18	0.38	68
K = 10	1.208123	0.146074	4.995954	0.5	25	0.38	126
K = 12	1.333767	0.175158	5.078096	0.5	43	0.38	258

Our method

Sparse Spatial Selection [Brisaboa & Pedreira, 2007] (VII)



Our method

Sparse Spatial Selection [Brisaboa & Pedreira, 2007] (VIII)

- SSS presents important properties for the index…
  - **Dynamic**
    - The database can be initially empty. Pivots are selected in a incremental way as the database grows.
    - The algorithm sets itself the number of pivots that will be used.
  - **Adaptive**
    - 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.

Our method

Non-Redundant Sparse Spatial Selection (NR-SSS)

- Non-Redundant Sparse Spatial Selection (NR-SSS)

The smaller the set of pivots, the smaller the internal complexity

Our method

Non-Redundant Sparse Spatial Selection (NR-SSS)

- Non-Redundant Sparse Spatial Selection (NR-SSS)
  - When Sparse Spatial Selection (SSS) identifies a new object in the DB as a pivot, we add it to the set of pivots.
  - We also check its contribution to this set of pivots. If its contribution to the set of pivots is 0, it is **redundant**, and thus immediately discarded.
  - If the new pivot contributes more than the worst already selected pivot, we remove the worst, since it is no longer useful.

But… How can we compute the contribution of each pivot?

Our method

Contribution of a pivot

$$|d(x, p_{max}) - d(y, p_{max})| - |d(x, p_{max2}) - d(y, p_{max2})|$$

A pair of objects selected at random

	p_1	p_2	…	p_n
(x_1, y_1)	1.34	0		0
(x_2, y_2)	0	2.57		0
(x_A, y_A)	0	0		1.00
$\sum$	1.34	2.57		1.00

Contribution of each pivot for each pair of objects

Total contribution

Outline

1. Motivation

2. Previous work

3. Our method
   - Sparse Spatial Selection (SSS)
   - Non-Redundant Sparse Spatial Selection (NR-SSS)

4. Experimental results

5. Conclusions

Experimental results

Test environment

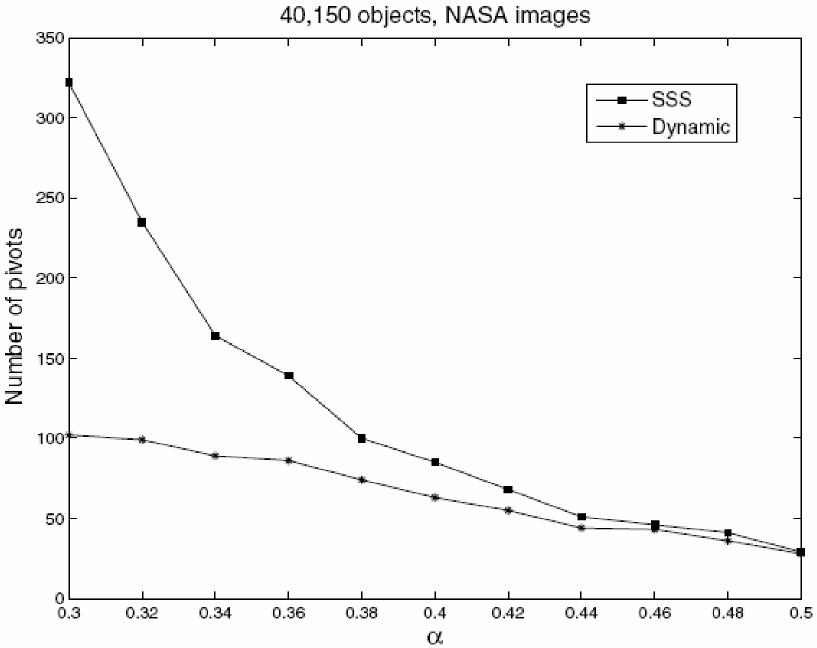
- All the collections used for experimental evaluation can be found at *SISAP Metric Spaces Library*
  - NASA: 40,150 images from NASA image and video archives, represented by feature vectors of dimension 20. Euclidean distance.
  - COLOR: 112,862 color images, each of them represented by a feature vector of 112 components. Euclidean distance.
  - SPANISH: 81,061 words taken from the Spanish dictionary. Edit distance.

Experimental results
Hypothesis

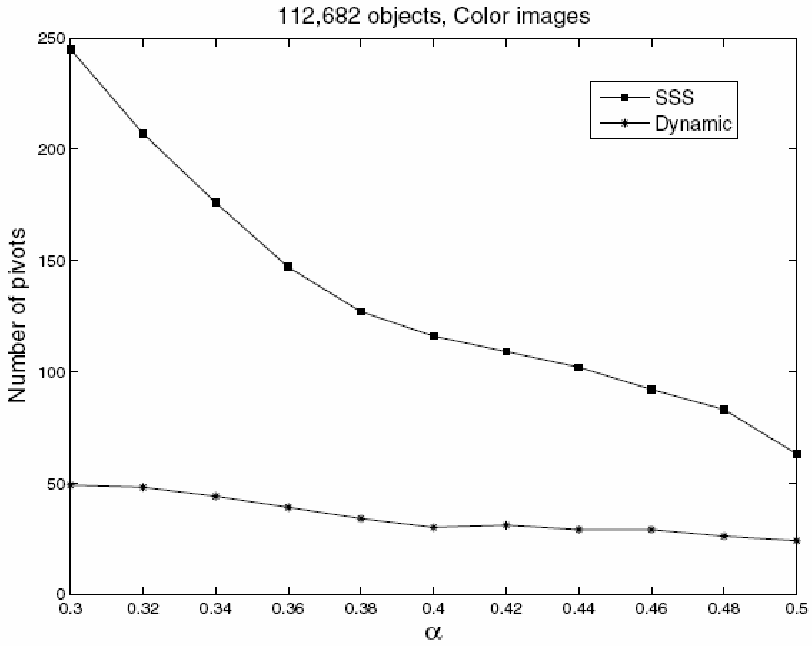
1. The set of pivots selected by Dynamic is smaller than the selected by Sparse Spatial Selection
2. The smaller the value of alpha, the higher the number of pivots replaced by Dynamic
3. The index built with Dynamic is more efficient than the one built with Sparse Spatial Selection in the search operation

Experimental results

Number of pivots selected by Dynamic and SSS



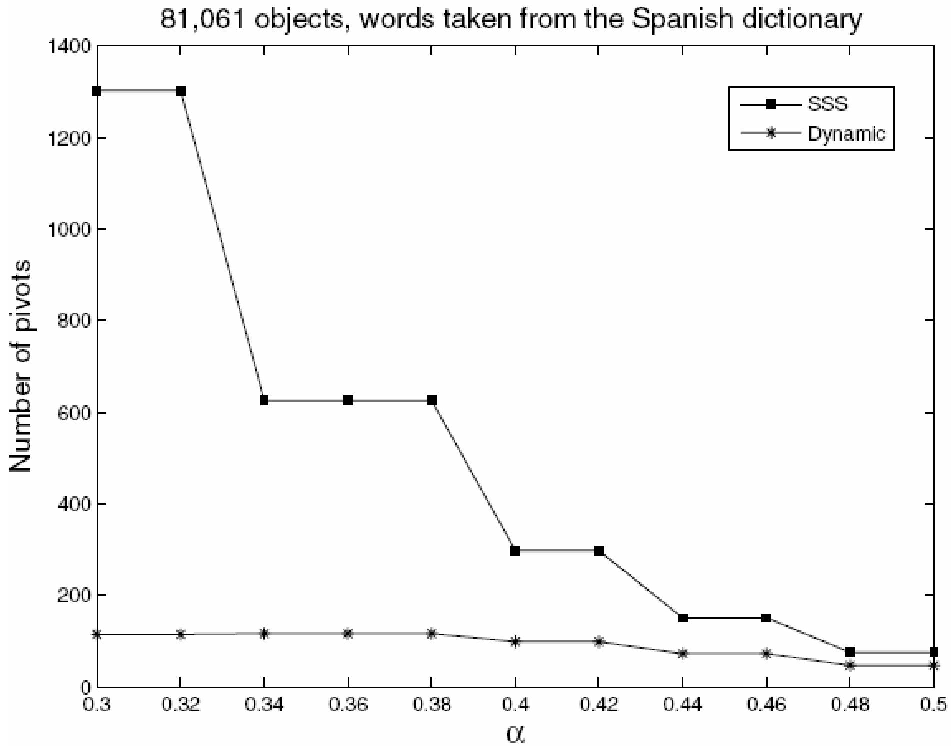
NASA Images



COLOR Images

Experimental results

Number of pivots selected by Dynamic and SSS



Words from the Spanish dictionary

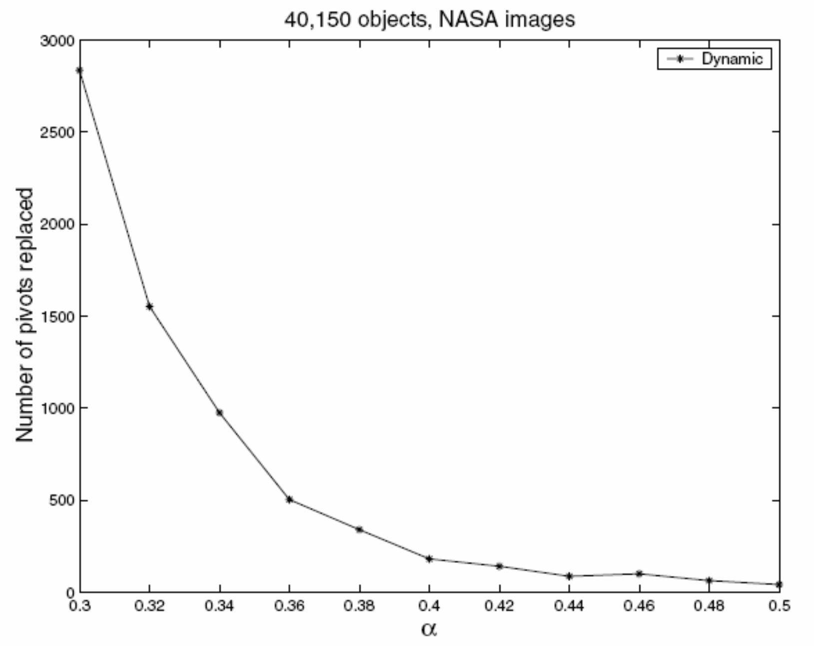
Experimental results

Hypothesis

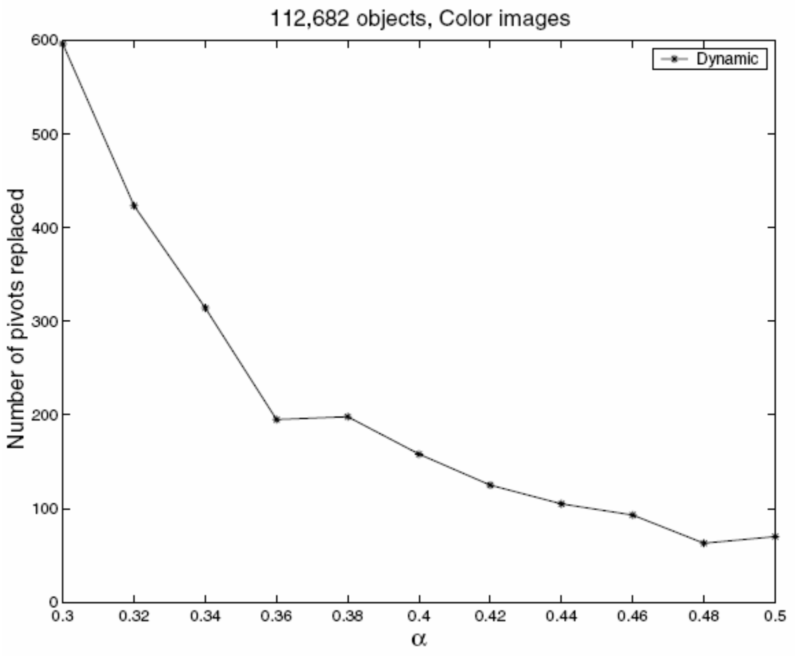
1. The set of pivots selected by Dynamic is smaller than the selected by Sparse Spatial Selection √
2. The smaller the value of alpha, the higher the number of pivots replaced by Dynamic
3. The index built with Dynamic is more efficient than the one built with Sparse Spatial Selection in the search operation

Experimental results

Pivots replaced in terms of α by Dynamic and SSS



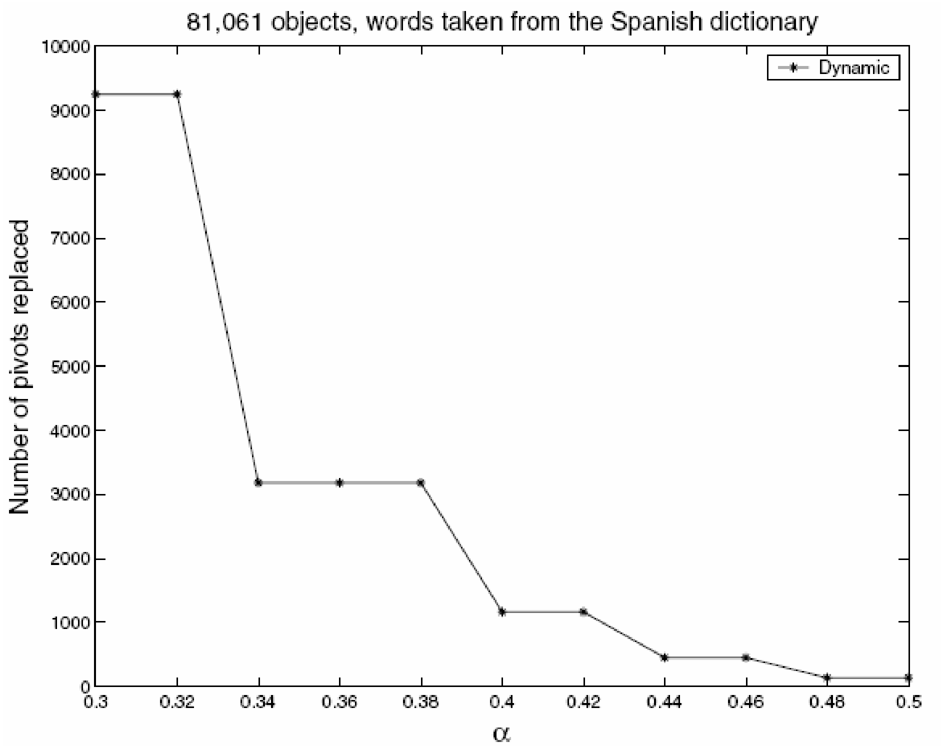
NASA Images



COLOR Images

Experimental results

Pivots replaced in terms of α by Dynamic and SSS



Words from the Spanish dictionary

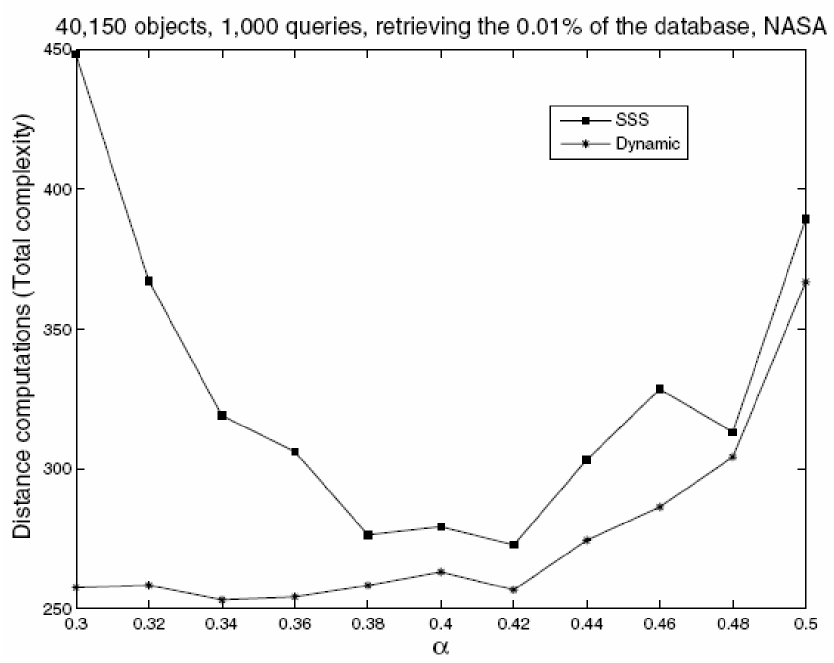
Experimental results

Hypothesis

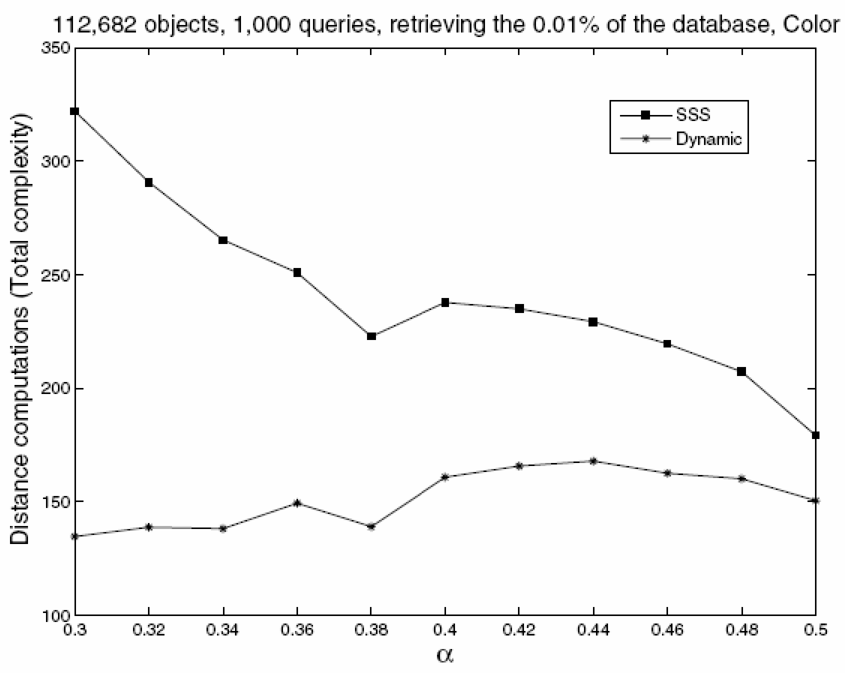
1. The set of pivots selected by Dynamic is smaller than the selected by Sparse Spatial Selection √
2. The smaller the value of alpha, the higher the number of pivots replaced by Dynamic √
3. The index built with Dynamic is more efficient than the one built with Sparse Spatial Selection in the search operation

Experimental results

Search efficiency in Dynamic and SSS



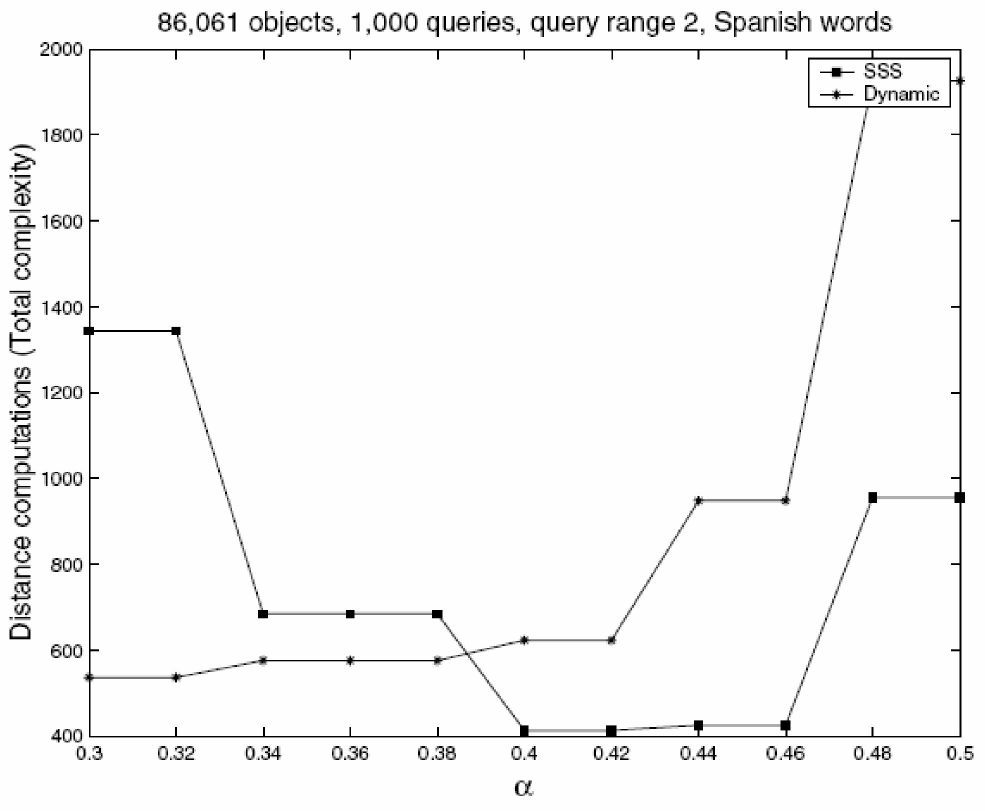
NASA Images



COLOR Images

Experimental results

Search efficiency in Dynamic and SSS



Words from the Spanish dictionary

Experimental results

Hypothesis

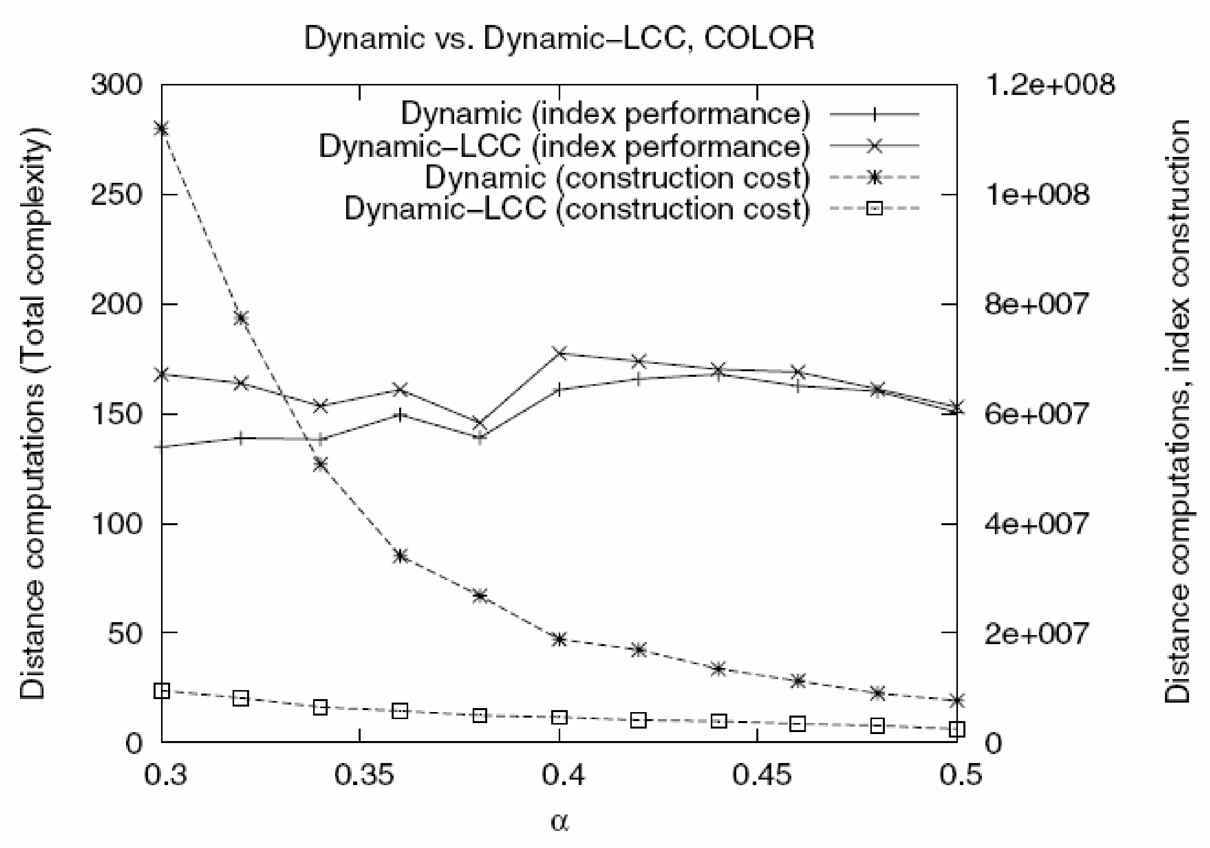
1. The set of pivots selected by Dynamic is smaller than the selected by Sparse Spatial Selection √

2. The smaller the value of alpha, the higher the number of pivots replaced by Dynamic √

3. The index built with Dynamic is more efficient than the one built with Sparse Spatial Selection in the search operation √

Experimental results

Dynamic-LCC → Low Construction Cost



Outline

1. Motivation
2. Previous work
3. Our method
   - Sparse Spatial Selection (SSS)
   - Non-Redundant Sparse Spatial Selection (NR-SSS)
4. Experimental results
5. Conclusions

Conclusions

- The paper proposes a new pivot selection technique called **Non-Redundant Sparse Spatial Selection (NR-SSS):** efficient, dynamic and that adapts itself to the space complexity.
- The pivots selected by Sparse Spatial Selection are filtered by **NR-SSS**, removing the useless ones
- The set of pivots is smaller → internal complexity is reduced
- Experimental results show the new technique outperforms state-of-art strategies

And …

Thanks for your attention!

Questions?