k-Nearest Neighbors

Tiago Royer

INE5432 — Bancos de Dados II

8 de junho de 2015

Síntese

Definição do predicado

Algoritmos e estruturas de dados

Extensões

Espaços métricos

Um espaço métrico é um par (M, d), em que $d : M \times M \to \mathbb{R}$ satisfaz

- $d(x, y) = d(y, x)$
- $d(x, y) \geq 0$
- $d(x, y) = 0 \iff x = y$
- $d(x, z) \leq d(x, y) + d(y, z)$

Definição do predicado

Seja k um inteiro positivo, a um parâmetro, e E um conjunto.

$$\sigma_{k,a}(E)$$

são os k elementos de E mais próximos de a.

Implementação diretamente em SQL ($k = 1$)

$\sigma_{1,(10.0,20.0)}(\texttt{points})$

```
SELECT *
FROM points p
WHERE NOT EXISTS (
    SELECT *
    FROM points q
    WHERE (q.x-10.0)*(q.x-10.0)+
          (q.y-20.0)*(q.y-20.0)
             <
          (p.x-10.0)*(p.x-10.0)+
          (p.y-20.0)*(p.y-20.0)
)
```

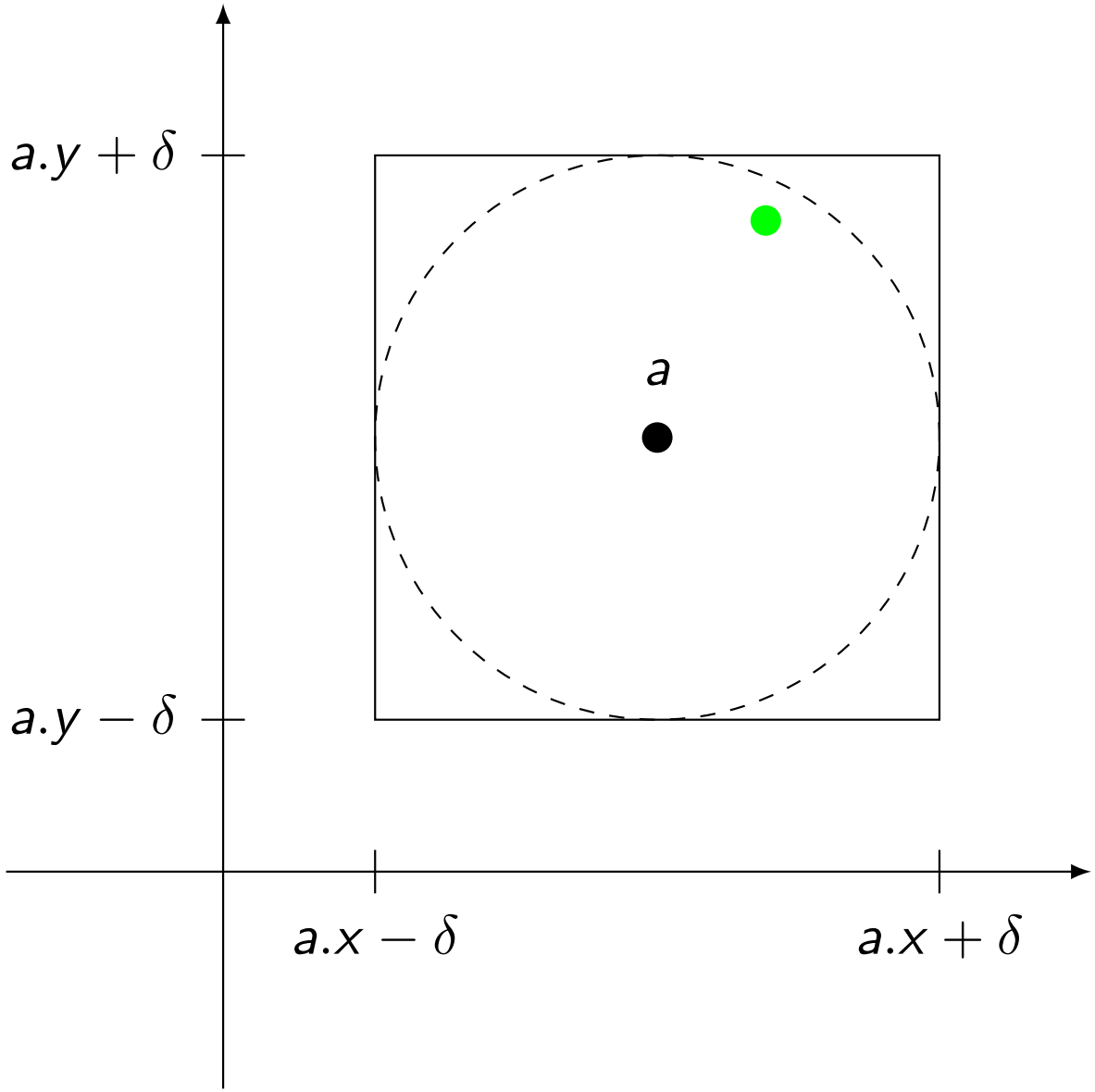
Implementação diretamente em SQL ($k = 1$)

$\sigma_{1,(10.0,20.0)}(\texttt{points})$

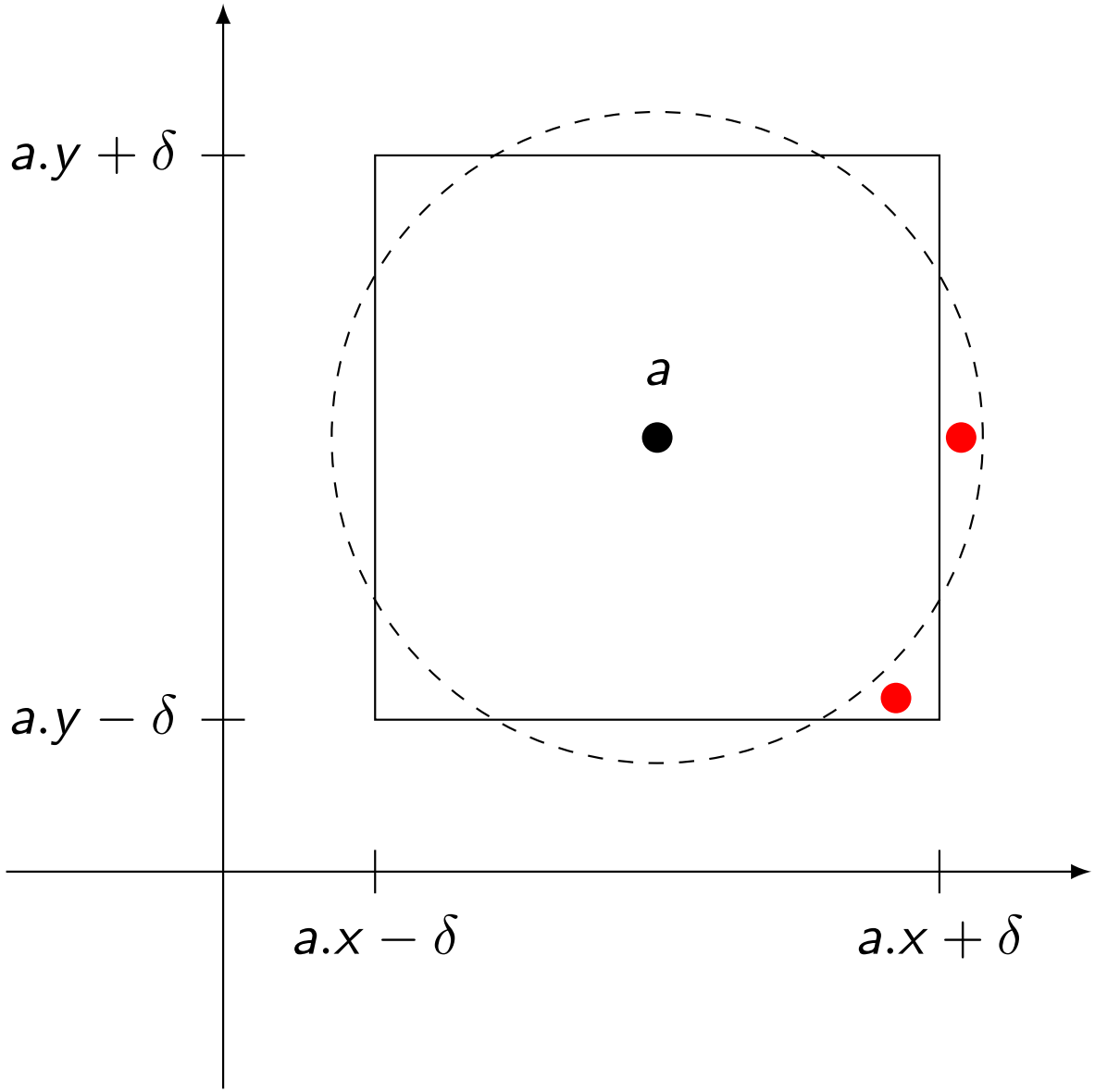
```
SELECT *
FROM points p
WHERE NOT EXISTS (
    SELECT *
    FROM points q
    WHERE (q.x-10.0)*(q.x-10.0)+
          (q.y-20.0)*(q.y-20.0)
             <
          (p.x-10.0)*(p.x-10.0)+
          (p.y-20.0)*(p.y-20.0)
)
```

Complexidade: $O(n^2)$

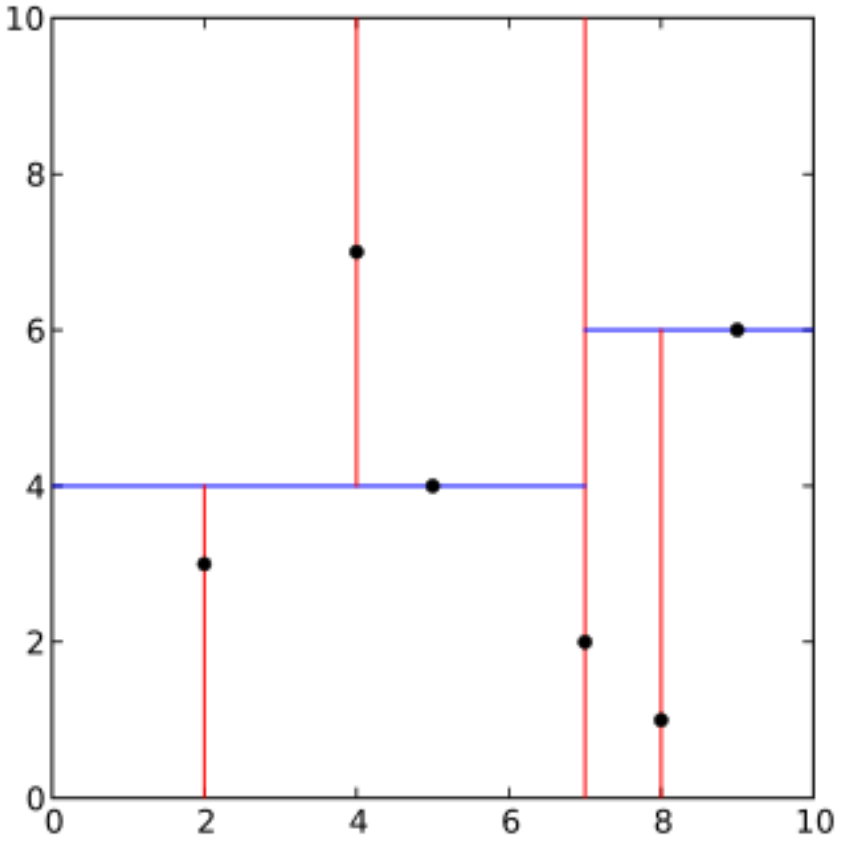
Restrições de intervalo



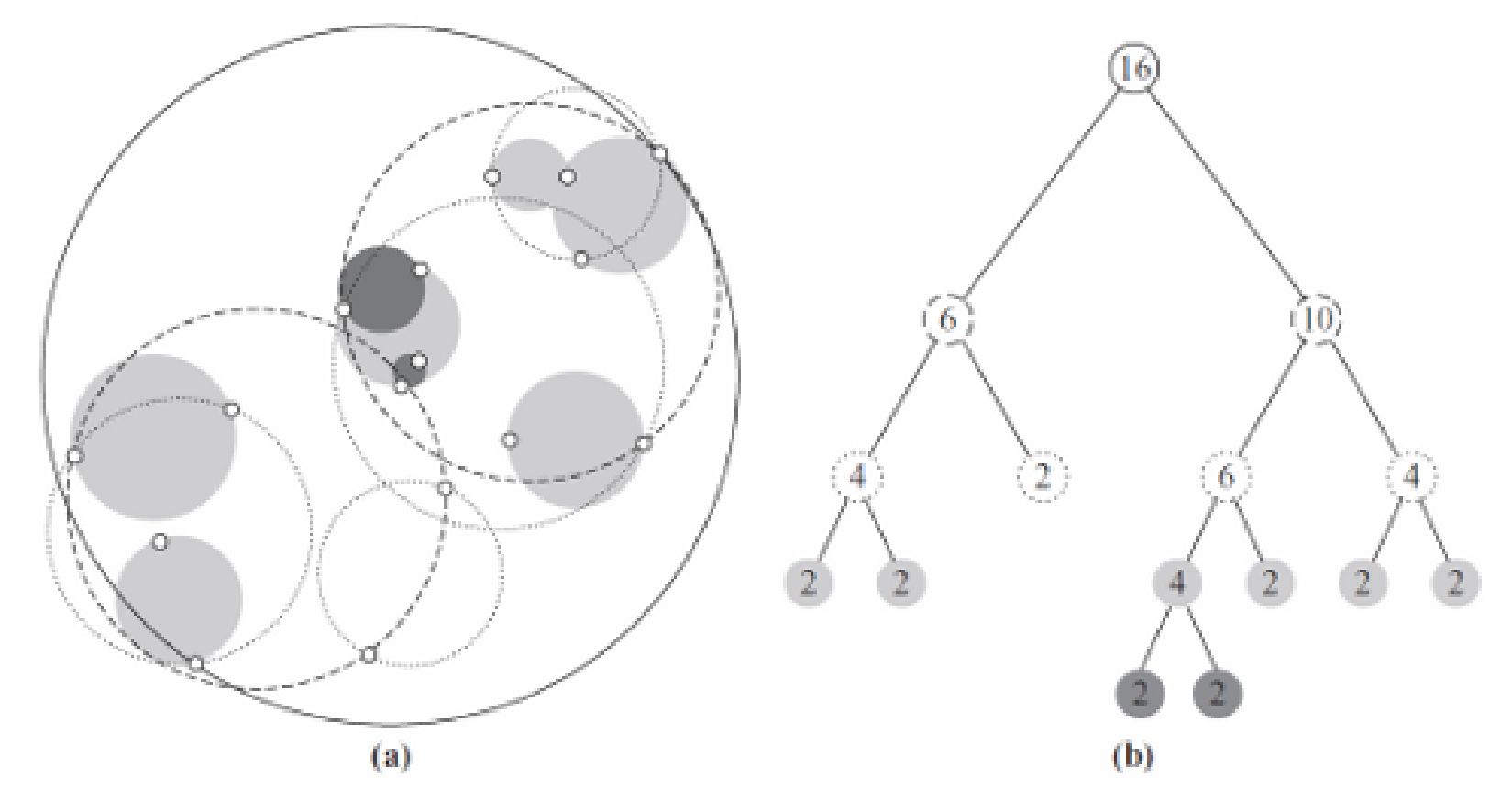
Problema



KD Tree



Ball Tree



Maldição da dimensionalidade

Para $d \leq 2$, existem algoritmos que garantem $O(\log n)$
Para $2 \leq d \leq 10$, existem algorimos que costumam ser $O(\log n)$
Para $d > 10$, os algoritmos costumam degenerar para $O(n)$

Locality-sensitive hashing

Funções de hashing que colidem objetos próximos.

Exemplo: projeções de $\mathbb{R}^n$ em $\mathbb{R}$

Extensões

- ε-NN
- Approximate Nearest Neighbor
- kNN Join

$$E_1 \bowtie_{k\mathrm{NN}} E_2$$

Pares $(e_1, e_2) \in E_1 \times E_2$ tais que e_2 é um dos k mais próximos a e_1.

- Reverse kNN
- k-Farthest Neighbor
- Reverse k-Farthest Neighbor

Referências (parcial)


Ahmed M. Aly, Walid G. Aref, and Mourad Ouzzani.
Spatial queries with two knn predicates.
CoRR, abs/1208.0074, 2012.


Ashraf Masood Kibriya.
Fast algorithms for nearest neighbour search.
Master's thesis, University of Waikato, 2007.


Flip Korn and S. Muthukrishnan.
Influence sets based on reverse nearest neighbor queries.
SIGMOD Rec., 29(2):201–212, May 2000.