

Universidade Tecnológica Federal do Paraná - Campus Pato Branco
Departamento Acadêmico de Informática
Curso de Engenharia de Computação

Consultas por similaridade em bases de dados complexas utilizando técnicas OMNI em SGBDR

Trabalho de Conclusão de Curso

Aluno: Cristiano Matsui

Orientador: Dr. Ives Renê Venturini Pola

Coorientadora: Dra. Fernanda Paula Barbosa Pola

6 de Dezembro de 2017

Sumário

- 1 Introdução
- 2 Técnica Omni
- 3 Extratores de Características
- 4 Objetivos
- 5 Materiais
- 6 Considerações Finais

Introdução

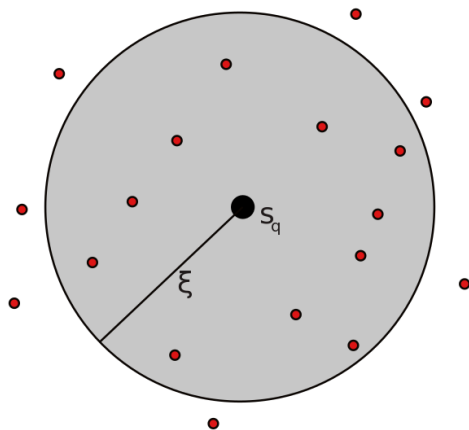
- Crescimento do uso de dados multimídia
 - Imagens, vídeos, áudio...
- “dados comuns” x “dados complexos”

Nome	Idade	Departamento	Salário	Telefone
Alberto da Silva	25	Vendas	R\$ 850,00	555-1902
Antônio dos Santos	32	Administração	R\$ 1.200,00	555-1117
Fabiana Rossi	40	Administração	R\$ 2.000,00	555-8929
Horácio Almeida	31	Recursos Humanos	R\$ 1.350,00	555-8907
João Pereira	35	Vendas	R\$ 1.500,00	555-7814
Márcia Souza	26	Vendas	R\$ 600,00	555-9800
Maria José Costa	22	Vendas	R\$ 600,00	555-6629
Mário Oliveira	54	Diretoria	R\$ 4.500,00	555-1237
Roberto Albuquerque	29	Administração	R\$ 1.200,00	555-8273
Sílvia Pires	23	Vendas	R\$ 600,00	555-8664

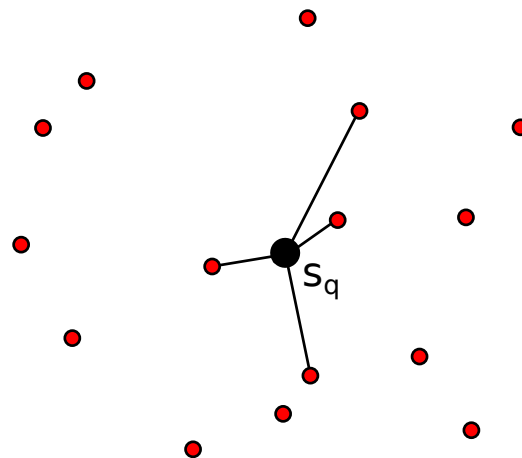
- Problema: consultas com dados complexos
- Necessidade de novos operadores de consulta

Introdução

- Consulta por abrangência (*Range query*)
- Consulta aos k-vizinhos mais próximos (*k-Nearest Neighbors query*)



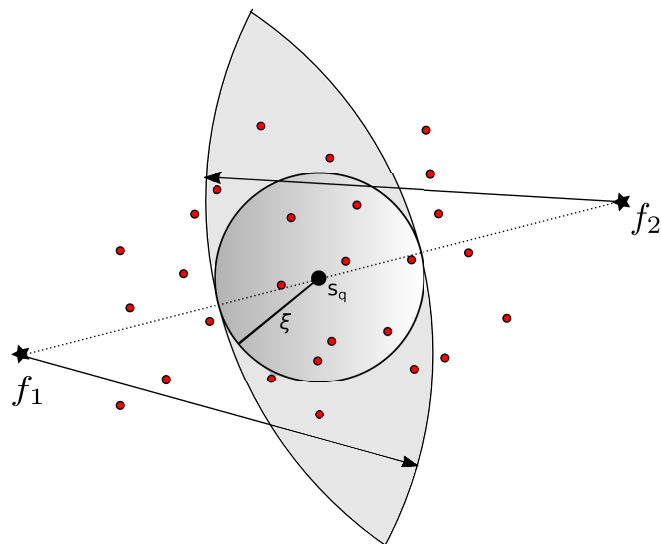
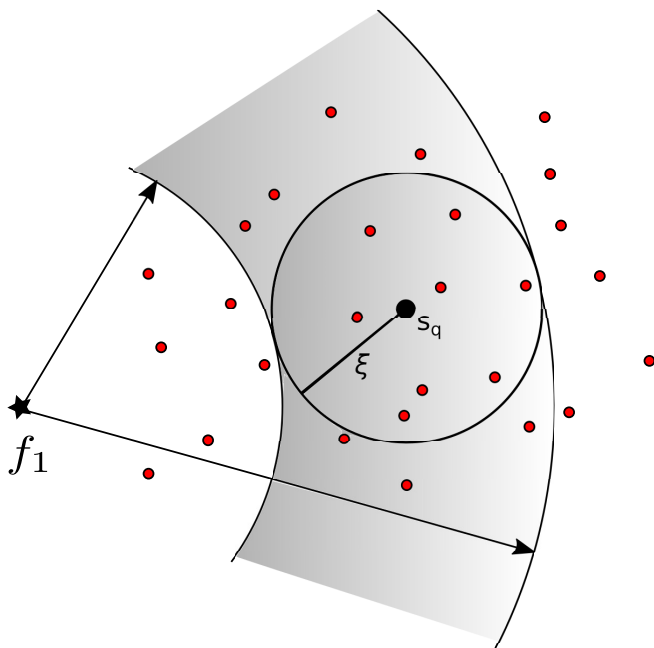
Range query



kNN query

- Consultas por similaridade são custosas
 - Complexidade dos dados
 - Tamanho da base
- Torna-se necessário otimizar estes procedimentos
- Uso da técnica Omni

- Reduz o número de cálculos de distância desnecessários
- Uso de uma base de focos
- *minimum bounding Omni region - mbOr*
- Uso da desigualdade triangular



- Coordenadas Omni
- Número de focos
 - *Box counting*
 - $(\lceil D \rceil + 1)$
- Custo do uso da base de focos
 - Espaço em disco
 - Tempo para o cálculo das coordenadas Omni

- Equação de pertinência à *mbOr*

- Desigualdade triangular

- $d_f(s_i) \leq d_f(s_q) + d(s_i, s_q)$

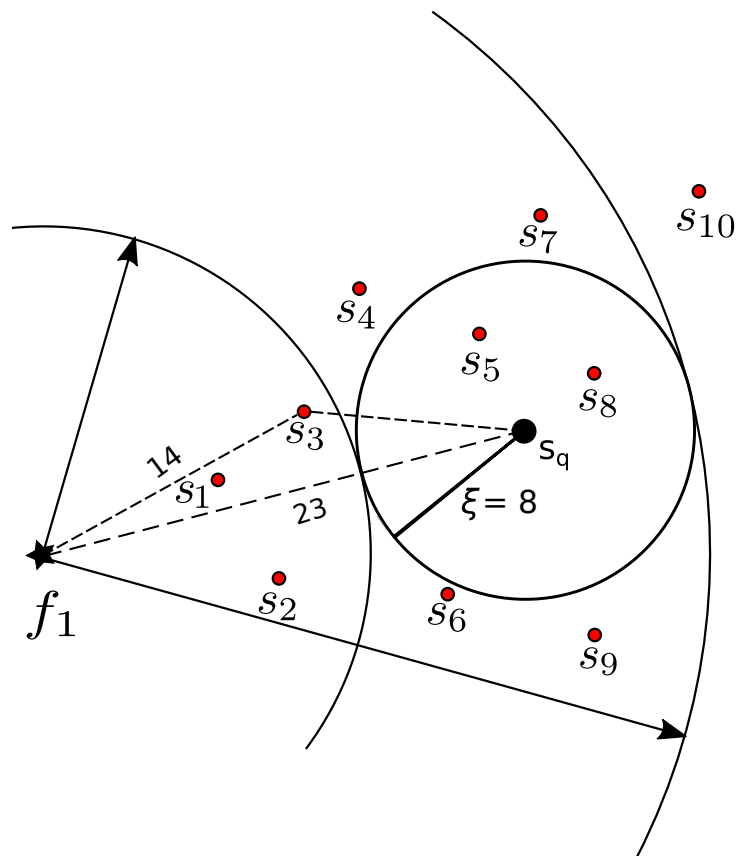
- $d(s_i, s_q) \geq |d_f(s_i) - d_f(s_q)|$

- Conceito de bola

- $d(s_i, s_q) \leq \xi$

- $\xi \geq d(s_i, s_q) \geq |d_f(s_i) - d_f(s_q)|$

- $|d_f(s_i) - d_f(s_q)| \leq \xi$



Extratores de Características

- Atributos visuais
 - Cor
 - Histograma de cores
 - Textura
 - Descritores de Haralick
 - Forma
 - Análise de contorno

Extratores de Características

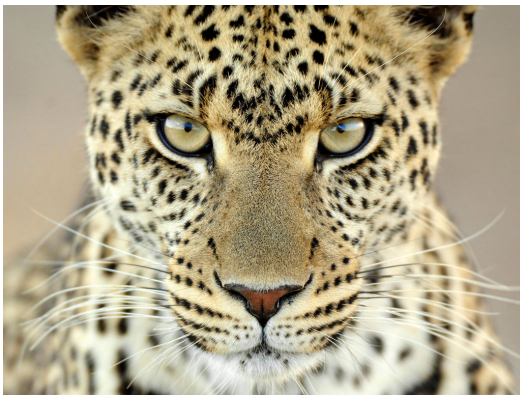
- Histograma de cores
 - Discretização e contagem
 - Computacionalmente simples
 - Pouca sensibilidade a alterações
 - Imagens diferentes podem apresentar histogramas de cor parecidos



- Descritores de Haralick
 - Matrizes de co-ocorrência
 - Análise dos níveis de cinza na imagem
 - Total de 13 características de textura
 - Contraste, correlação, entropia...
 - Especialmente útil na área médica

Extratores de Características

- Análise de forma
 - Principal análise feita pelo olho humano
 - Análise de contornos



- Objetivos Gerais

- Construção de um sistema de consultas em SGBDR por similaridade em uma base de imagens utilizando técnicas da família Omni para reduzir o custo computacional das operações de consulta

- Objetivos Específicos

- Modelar o banco de dados para atender a problemática apresentada;
- Aplicar os extratores de características das imagens utilizadas;
- Inserir no banco de dados as imagens e os valores de suas características;
- Criar a estrutura Omni necessária para a filtragem dos cálculos;
- Analisar e comparar os resultados obtidos.

- PostgreSQL
- Arboretum (GBDI)
- Software para elaboração da interface com o usuário

- Planeja-se utilizar este trabalho para a elaboração de um CBIR médico