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

Consultas por similaridade em bases de dados complexos utilizando técnica OMNI em SGBDR

Trabalho de Conclusão de Curso

Aluno: Cristiano José Mendes Matsui
Orientador: Dr. Ives Renê Venturini Pola
Coorientadora: Dra. Fernanda Paula Barbosa Pola

6 de Dezembro de 2018

Sumário

Introdução

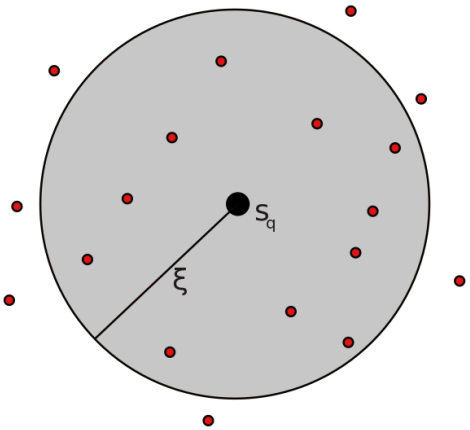
- Crescimento do uso de dados multimídia
  - Imagens, vídeos, áudio...
- "dados comuns" x "dados complexos"

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
Silvia Pires	23	Vendas	R$ 600,00	555-8664

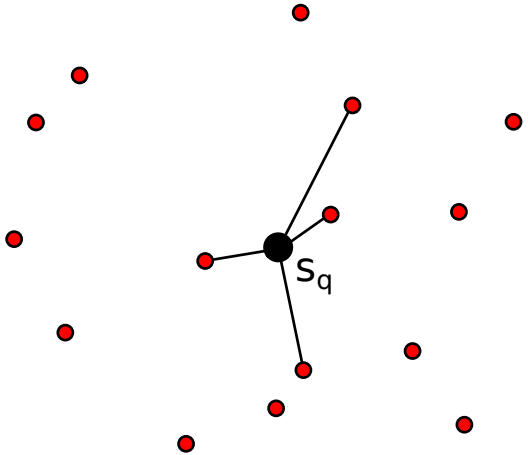
- Novos operadores de consulta

Introdução

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



Range query



kNN query

Introdução

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

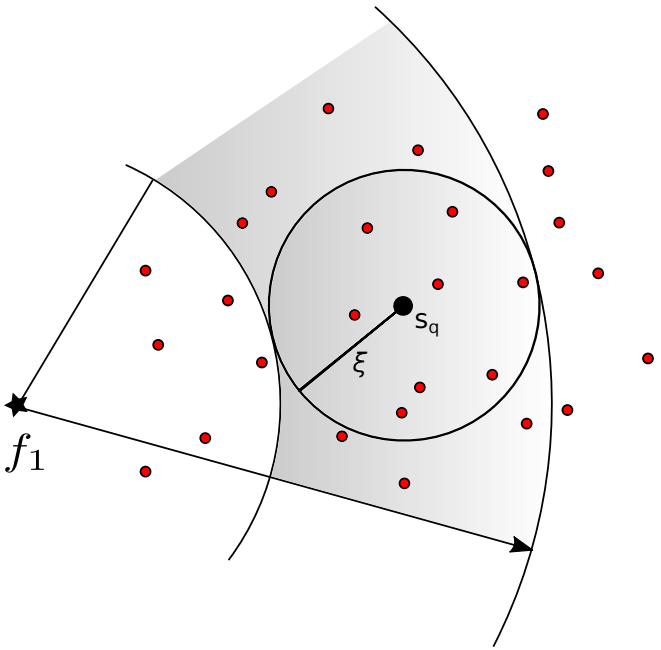
Objetivos

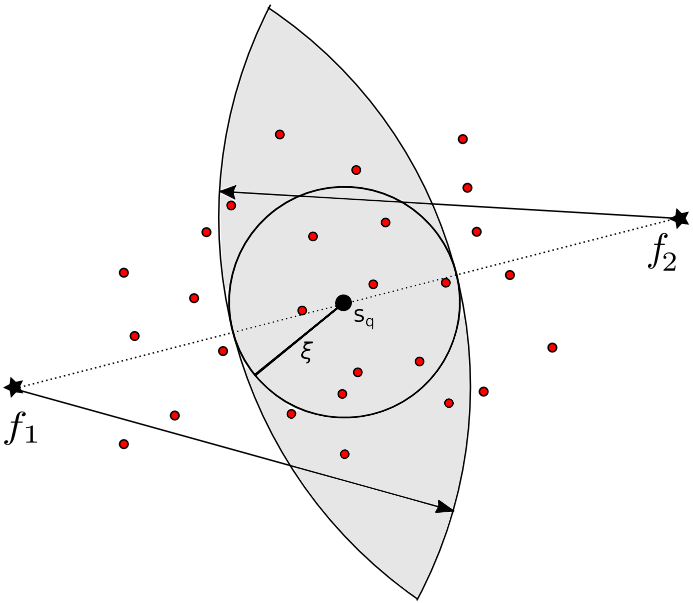
- Objetivos Gerais
  - Consultas por similaridade utilizando OMNI em SGBDR
- Objetivos Específicos
  - Modelagem;
  - Extrair características;
  - Inserção das características no banco;
  - Criação das estruturas OMNI;
  - Analisar e comparar os resultados obtidos.

Técnica OMNI

- Reduz o número de cálculos de distância desnecessários
- Uso de uma base de focos
- *minimum bounding OMNI region - mbOr*
- Desigualdade triangular e conceito de bola fechada

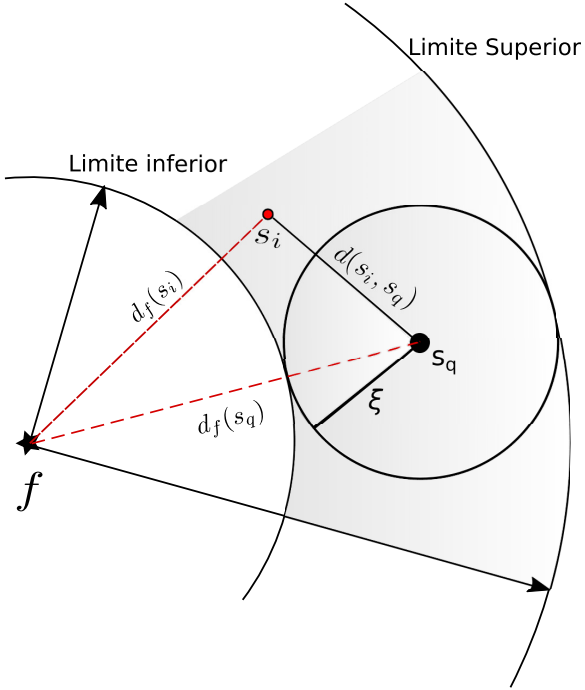
Técnica OMNI





Técnica OMNI

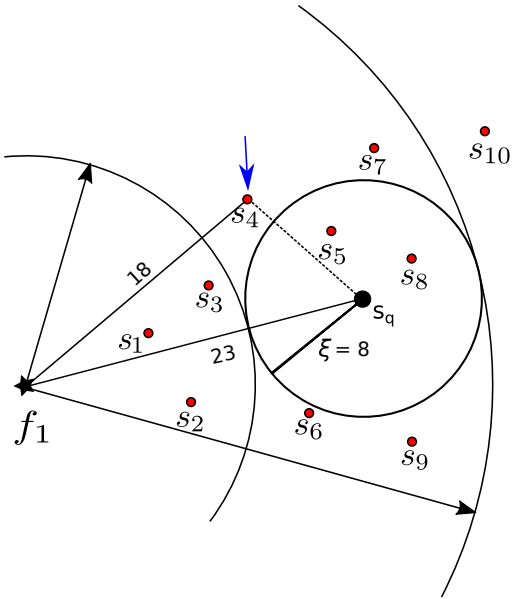
- minimum bounding OMNI region
  - Desigualdade triangular
    - $d_f(s_i) \leq d_f(s_q) + d(s_i, s_q)$
    - $d(s_i, s_q) \geq |d_f(s_i) - d_f(s_q)|$
  - Conceito de bola
    - $d(s_i, s_q) \leq \xi$



- $\xi \geq d(s_i, s_q) \geq |d_f(s_i) - d_f(s_q)|$
- $|d_f(s_i) - d_f(s_q)| \leq \xi \leftarrow$ Equação de pertinência à $mbOr$

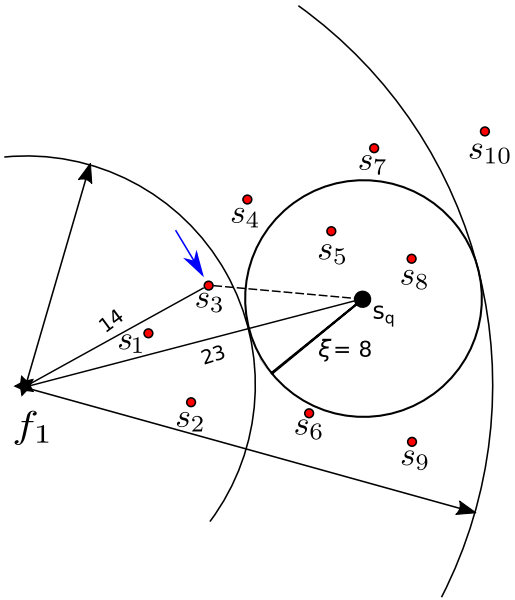
Técnica OMNI

- $|d_f(s_i) - d_f(s_q)| \leq \xi \leftarrow$ Equação de pertinência à $mbOr$
- $|d_f(s_4) - d_f(s_q)| < \xi$
- Pertence à $mbOr$!



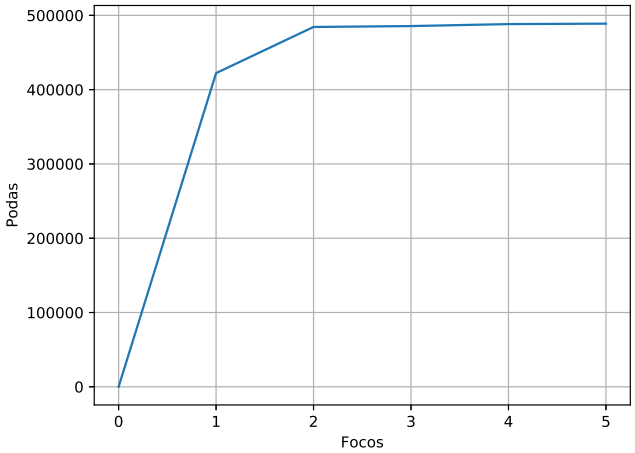
Técnica OMNI

- $|d_f(s_i) - d_f(s_q)| \leq \xi \leftarrow$ Equação de pertinência à $mbOr$
- $|d_f(s_3) - d_f(s_q)| > \xi$
- Não pertence à $mbOr$!



Técnica OMNI

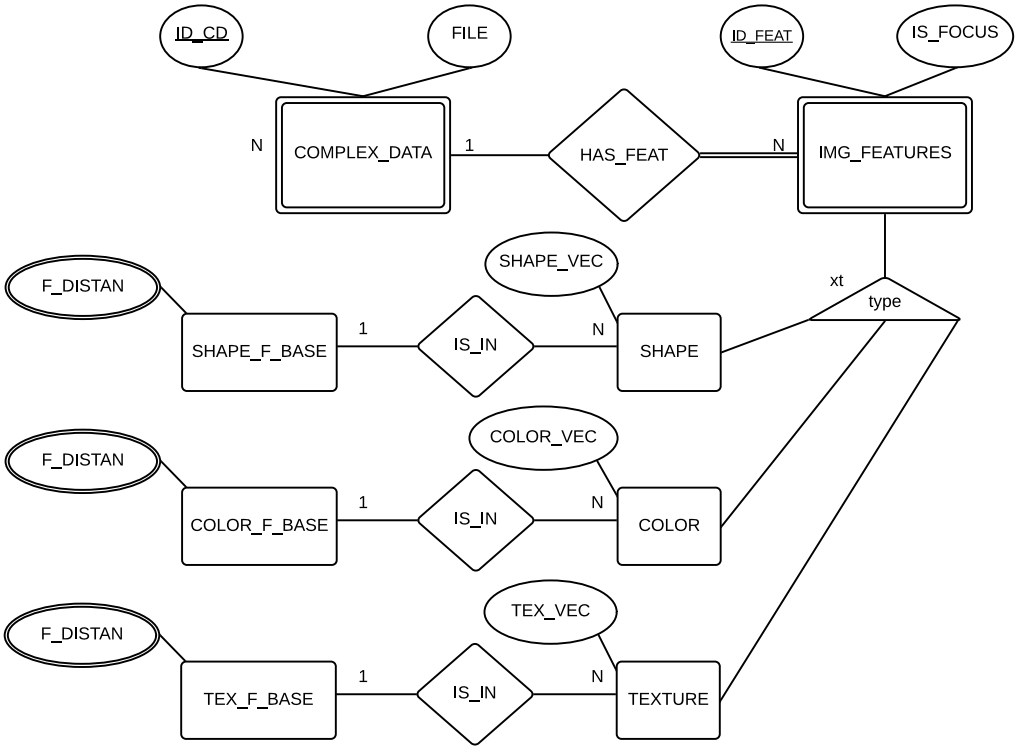
- Escolha dos focos
- Número ótimo de focos
  - *Box counting*
    - $(\lceil D \rceil + 1)$
  - Gráfico de podas $\times$ número de focos



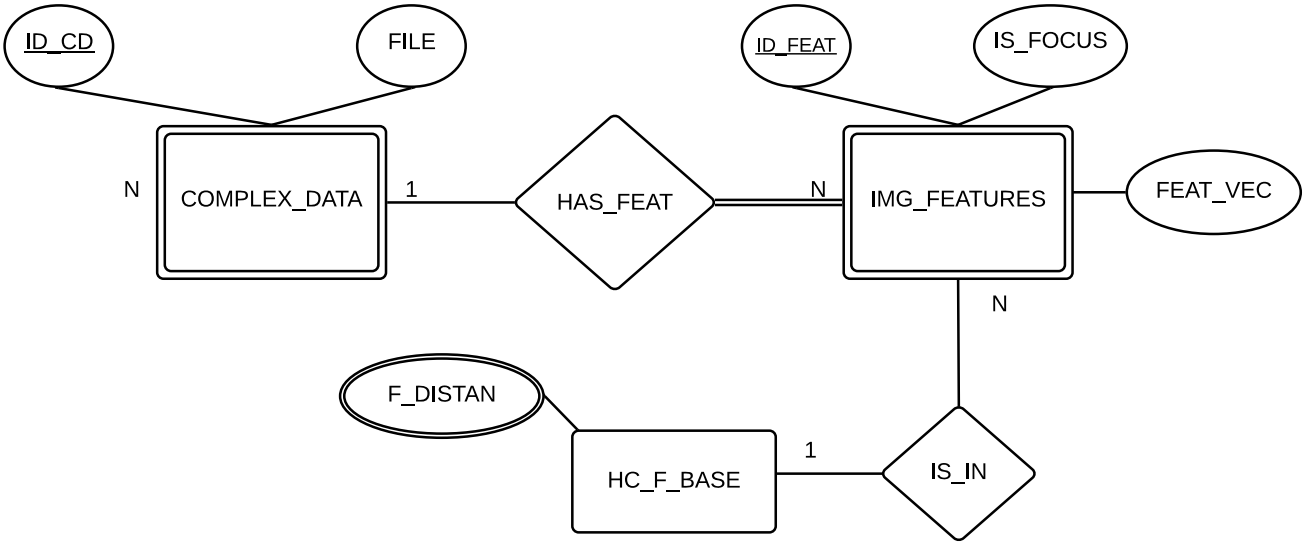
Bases de Dados

- Duas bases de dados utilizadas
- CAT_DOG - 25000 imagens
  - Imagens de cães e gatos
  - Características extraídas via Python + Scikit-Image
    - Cor - Histograma
    - Textura - dissimilaridade, contraste, correlação
    - Forma - área, excentricidade, área convexa
- HC - 500000 imagens
  - Características de imagens médicas já extraídas
  - Histograma de 256 tons de cinza

- Modelagem CAT_DOG



- Modelagem HC



Resultados do Trabalho

- Testes realizados utilizando distâncias L_1, L_2 e L_∞
- Média de 10 medições distintas para um mesmo centro
- Sequencial e OMNI com números distintos de focos
- Utilização das estruturas de indexação

Resultados do Trabalho

- Base CAT_DOG - Consulta por abrangência - L_1 - FORMA

	Tempo de Execução(ms)	Nº Podas de Cálculos
Sequencial	8,721	-
OMNI 1 Foco	**1,747**	24935
OMNI 2 Focos	5,201	24935
OMNI 3 Focos	7,606	24935
OMNI 4 Focos	10,107	24935
OMNI 5 Focos	12,271	24935

Resultados do Trabalho

- Base CAT_DOG - Consulta por abrangência - L_2 - FORMA

	Tempo de Execução(ms)	Nº Podas de Cálculos
Sequencial	8,728	-
OMNI 1 Foco	**1,825**	24935
OMNI 2 Focos	4,988	24935
OMNI 3 Focos	7,525	24935
OMNI 4 Focos	8,994	24935
OMNI 5 Focos	11,591	24935

Resultados do Trabalho

- Base CAT_DOG - Consulta por abrangência - L_∞ - FORMA

	Tempo de Execução(ms)	Nº Podas de Cálculos
Sequencial	8,805	-
OMNI 1 Foco	**2,051**	24935
OMNI 2 Focos	4,638	24935
OMNI 3 Focos	8,563	24935
OMNI 4 Focos	9,651	24935
OMNI 5 Focos	11,626	24935

Resultados do Trabalho

- Base CAT_DOG - Consulta por abrangência - L_1 - TEXTURA

	Tempo de Execução(ms)	Nº Podas de Cálculos
Sequencial	8,710	-
OMNI 1 Foco	**1,531**	24950
OMNI 2 Focos	4,512	24950
OMNI 3 Focos	6,838	24950
OMNI 4 Focos	9,447	24950
OMNI 5 Focos	11,182	24950

Resultados do Trabalho

- Base CAT_DOG - Consulta por abrangência - L_1 - COR

	Tempo de Execução(ms)	Nº Podas de Cálculos
Sequencial	8,334	-
OMNI 1 Foco	**1,675**	24923
OMNI 2 Focos	4,819	24923
OMNI 3 Focos	8,073	24923
OMNI 4 Focos	9,628	24923
OMNI 5 Focos	12,925	24923

Resultados do Trabalho

- Base HC - Consulta por abrangência - L_1

	Tempo de Execução(ms)	Nº Podas de Cálculos
Sequencial	4566,851	-
OMNI 1 Foco	448,105	484381
OMNI 2 Focos	**292,709**	490599
OMNI 3 Focos	361,712	499406
OMNI 4 Focos	462,795	499845
OMNI 5 Focos	882,159	499937

Resultados do Trabalho

- Base HC - Consulta por abrangência - L_2

	Tempo de Execução(ms)	Nº Podas de Cálculos
Sequencial	4310,409	-
OMNI 1 Foco	1215,435	422246
OMNI 2 Focos	**307,757**	484381
OMNI 3 Focos	556,683	485574
OMNI 4 Focos	598,664	488229
OMNI 5 Focos	952,008	488841

Resultados do Trabalho

- Base HC - Consulta por k-vizinhos mais próximos - L_2

	Tempo de Execução(ms)
Sequencial	**4733,579**
OMNI 1 Foco	6519,995
OMNI 2 Focos	6341,987
OMNI 3 Focos	7490,256
OMNI 4 Focos	7548,550
OMNI 5 Focos	10348,541

Resultados do Trabalho

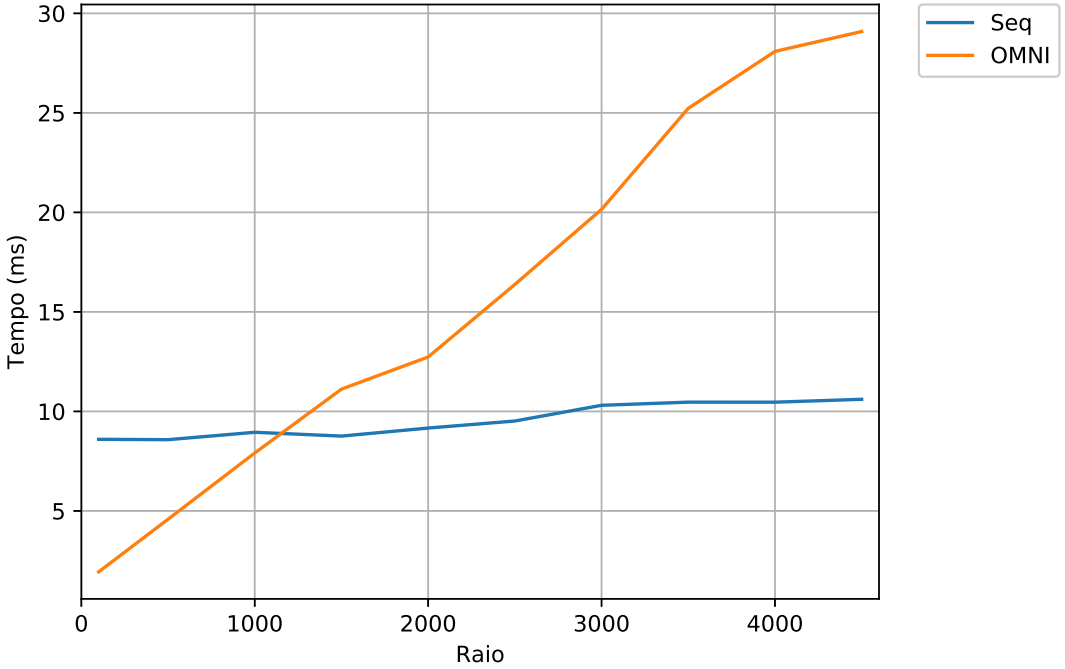
- Ganho de Performance - GP
- $GP_{\%} = (T_o/T_s - 1) * 100$

CAT_DOG	GP (%)
L_1 - FORMA	399,20
L_2 - FORMA	378,24
L_∞ - FORMA	329,30
L_1 - TEXTURA	468,91
L_1 - COR	397,55

HC	GP (%)
L_1	1460,20
L_2	1304,04
L_∞	-

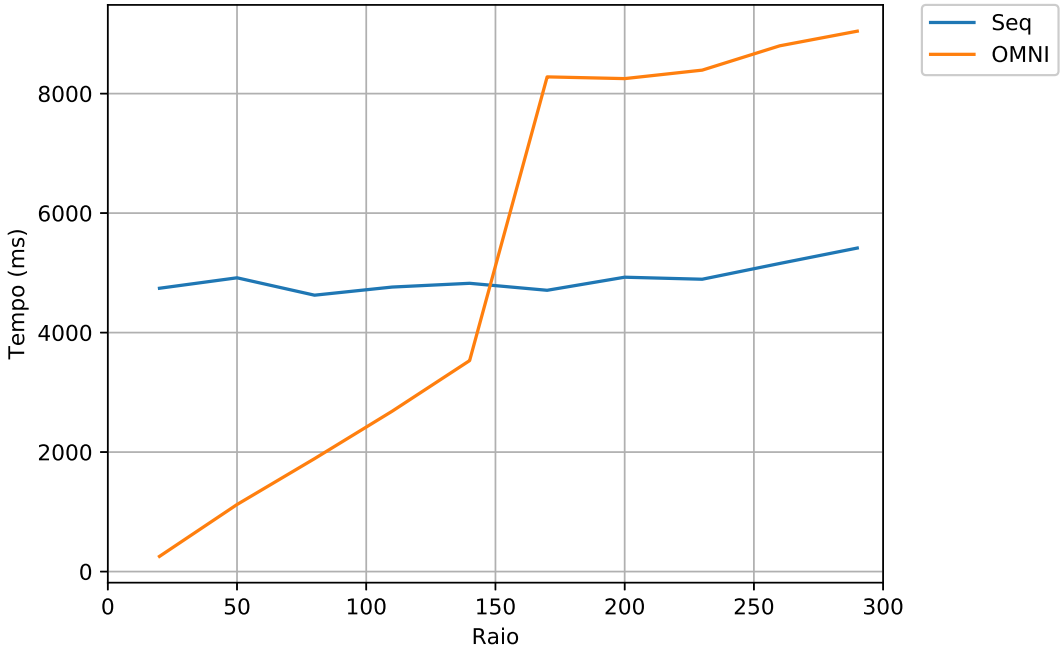
Limitações e Custos

- Aumento do raio de consulta impacta o desempenho
  - Base CAT_DOG



Limitações e Custos

- Aumento do raio de consulta impacta o desempenho
  - Base HC



Limitações e Custos

- Cálculo e preparação das estruturas OMNI
- Custo de armazenamento adicional

	Índices (MB)	Dados (MB)
SHAPE	1,62	4,86
COLOR	0,55	2,83
TEXTURE	0,55	2,83
SHAPE_F_BASE	36	1,47
COLOR_F_BASE	36	13
TEXTURE_F_BASE	36	1,47
HC_TABLE	53	1058
HC_F_BASE	263	1229

Considerações Finais

- Implementação da consulta por similaridade utilizando OMNI em SGBDR
- Ganho notável de performance para consulta por abrangência
- Trabalhos futuros:
  - Suporte a outros tipos de consulta por similaridade
  - Aprimoramento da OMNI-kNN
  - Melhoria dos extratores de características
  - Mitigação das limitações da técnica

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

Consultas por similaridade em bases de dados complexos utilizando técnica 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 2018