

CRISTIANO JOSÉ MENDES MATSUI

**CONSULTAS POR SIMILARIDADE EM BASES DE DADOS  
COMPLEXAS UTILIZANDO TÉCNICAS OMNI EM SGBDR**

Proposta de Trabalho de Conclusão de Curso de graduação, apresentado à disciplina de Trabalho de Conclusão de Curso 1, do Curso de Engenharia de Computação - Departamento Acadêmico de Informática - da Universidade Tecnológica Federal do Paraná - UTFPR - Câmpus Pato Branco, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Dr. Ives Renê Venturini Pola

Coorientador: Dra. Fernanda Paula Barbosa Pola

PATO BRANCO  
2017

## 1 INTRODUÇÃO

Nos anos recentes, foi notado um grande aumento no tráfego e armazenamento de diferentes aplicações e dados multimídias, como imagens, áudio, vídeo, impressões digitais, séries temporais, sequências de proteínas, etc. Estes tipos de dados, que apresentam muito mais atributos do que simples numerais ou pequenas cadeias de caracteres, são conhecidos como dados complexos (ZIGHED et al., 2008).

Quando tratados por um Sistema Gerenciador de Banco de Dados Relacional (SGBDR), não suportam comparações com os operadores conhecidos como "big six" da linguagem SQL:  $=$ ,  $\neq$ ,  $<$ ,  $>$ ,  $\leq$ ,  $\geq$ . Esse fato limita muito as comparações entre dados complexos inseridos em um SGBDR, ocasionando um grande problema no contexto de base de dados, uma vez que os principais sistemas de gerenciamento de base de dados são relacionais (DB-ENGINES, 2017). Com isso, tornou-se necessária a concepção de novos tipos de comparadores, como buscas por similaridade.

Estas consultas por similaridade se aplicam de maneira geral a muitos dos tipos de dados complexos (BARIONI et al., 2009). Embora equiparar duas imagens médicas (como tomografias de pacientes distintos) raramente produza um resultado diferente de falso, procurar por imagens semelhantes à original faz mais sentido e retorna resultados mais relevantes. Dentre os operadores de consulta por similaridade os mais comuns são as consultas por abrangência (*range query: Rq*) e consulta aos k-vizinhos mais próximos (*k-nearest neighbor query: kNNq*). As consultas por abrangência  $Rq(s_q, \xi)$  recebem como parâmetro um elemento  $s_q$  do domínio de dados (elemento central da consulta) e um limite máximo de dissimilaridade  $\xi$ . O resultado é o conjunto de elementos da base que diferem do elemento central da consulta por no máximo a dissimilaridade indicada.

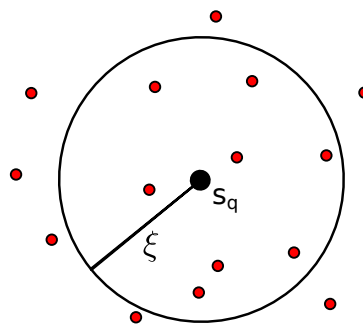


Figura 1 – Exemplo de consulta por abrangência

Uma consulta aos k-vizinhos mais próximos  $kNNq(s_q, k)$  também recebe como um de seus parâmetros um elemento central da consulta  $s_q$ , e um número inteiro  $k$  de vizinhos desejados, e retorna como resultado o conjunto dos  $k$  elementos com a menor dissimilaridade em relação ao elemento central da consulta  $s_q$  (POLA, 2010).

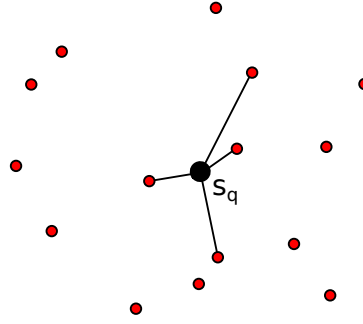


Figura 2 – Exemplo de consulta por k-vizinhos mais próximos

O SGBDR não possui suporte nativo a estes tipos de consulta, mas é possível construir estas consultas utilizando ferramentas existentes em um banco de dados relacional (como a *B-tree*). Para entender e ajustar os métodos de consultas por similaridade para diferentes tamanhos e tipos de conjuntos de dados, além de comparar diversos métodos, é importante uma análise teórica dos diferentes métodos de acesso e técnicas de estimativa do custo computacional (POLA, 2010).

A solução abordada por esta proposta é a do uso de técnicas OMNI, presentes no trabalho de (FILHO et al., 2001). Um número calculado de elementos do conjunto de dados são selecionados como "focos", e utilizados para podar cálculos desnecessários de distâncias, fazendo uso da desigualdade triangular. Para quaisquer elementos  $s_1, s_2, s_3 \in \mathbb{S}$ , sendo  $\mathbb{S}$  um domínio de elementos e uma métrica  $d : \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{R}^+$ , temos a desigualdade triangular:

$$d(s_1, s_2) \leq d(s_1, s_3) + d(s_3, s_2) \quad (1)$$

A base da técnica OMNI é calcular previamente as distâncias de todos os elementos para todos os focos selecionados, armazenando estas distâncias no banco. Quando uma consulta por similaridade (como uma consulta por abrangência) é realizada, são conhecidas as distâncias entre o elemento central da consulta  $s_q$  e o raio de abrangência  $\xi$ . Considerando um foco  $f_1$  e utilizando a desigualdade triangular, elementos que possuem uma distância entre o foco escolhido menor do que a distância de  $s_q$  até o foco menos o valor  $\xi$  serão descartados do conjunto de elementos necessários para os cálculos de distância com o elemento central. Simetricamente, elementos cuja distância até o foco seja maior do que a distância de  $s_q$  até o foco mais o valor do raio de abrangência  $\xi$  também serão descartados. Com isso, ocorre uma grande redução do número de cálculos necessários para fornecer o conjunto resposta. Essa poda também pode ser realizada por mais de um foco.

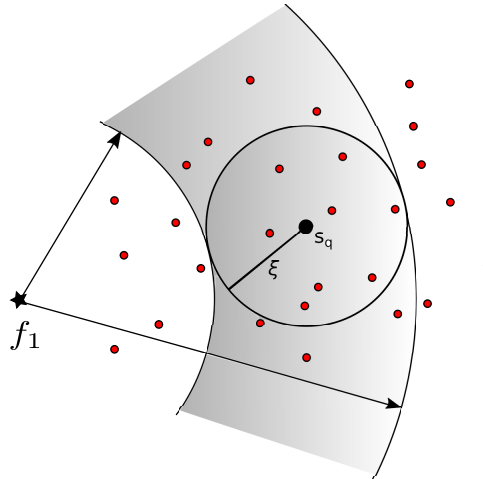


Figura 3 – Consulta por abrangência pela técnica OMNI utilizando um foco

A figura 3 ilustra a poda no número de cálculos. Apenas os elementos na área sombreada terão as suas distâncias em relação ao centro da consulta calculadas, pois estão no conjunto de elementos que não foram descartados utilizando a desigualdade triangular com as distâncias previamente calculadas em relação ao foco.

P1. Contextualização do Projeto

P2. Definição do Problema

P3. Relevância do Problema

P4. Justificativa

P5. Desafios do Projeto

P6. Contribuição

(substitua este texto pela introdução do trabalho)

## 2 OBJETIVOS

### 2.1 Objetivo Geral

O objetivo geral é tratado em seu sentido mais amplo e constitui a ação que conduzirá ao tratamento da questão abordada no problema de pesquisa, fazendo menção ao objeto de uma forma mais direta. O objetivo geral deve:

Conter descrição do que vai fazer, de forma precisa e objetiva;

Ser diretamente ligado ao título;

Ser mais detalhado que o título;

Resolver o problema proposto;

Ser Claro, Conciso e Completo (CCC) e deve ser verificável ao final do trabalho.

(substitua este texto pelo objetivo geral do trabalho)

## 2.2 Objetivos Específicos

Os objetivos específicos apresentam, de forma pormenorizada, detalhada, as ações que se pretende alcançar e estabelecem estreita relação com as particularidades relativas à temática trabalhada. Os objetivos específicos devem:

Ser Claros, Concisos e Completos (CCC) e devem ser verificáveis ao final do trabalho;

Fazem parte do detalhamento do objetivo geral ;

Devem ser iniciados com o verbo no infinitivo;

Podem ser considerados com subprodutos do objetivo geral.

(substitua este texto pelos objetivos específicos do trabalho)

## 3 METODOLOGIA

Na seção de procedimentos metodológicos ou metodologia (ver qual o nome mais adequado ao trabalho) deve ser descrito sucintamente o procedimentos metodológicos para a execução do projeto ressaltando como os objetivos serão alcançados.

Em geral, a seção descreve os procedimentos usados para resolver o problema atacado. Pode ser estruturada em tópicos, onde cada tópico representa um subproduto do objetivo geral.

No caso de desenvolvimento de sistemas deve-se descrever a metodologia a ser utilizada, por exemplo Scrum, eXtreme Programming, RUP, etc.

Também pode ser descritos técnicas de desenvolvimento de software como por exemplo TDD, BDD, SPA, etc.

(substitua este texto pelo de procedimentos metodológicos/metodologia do trabalho)

### 3.1 MATERIAIS

Coloque todos os materiais que serão utilizados. Exemplos: computadores, equipamentos de redes, licenças de software, etc. Também deverá ser colocado se os recursos estarão disponíveis. A universidade não comprará os recursos, portanto a responsabilidade de comprar algo será do aluno. (substitua este texto pelo de recursos necessários do trabalho)

### 3.2 METODOS

Aqui entram os métodos envolvidos no trabalho.

## 4 CONCLUSÃO

Na seção de Conclusão ou Considerações Finais (ver qual o nome mais adequado ao trabalho) o acadêmico deve descrever:

O planejamento do trabalho de estágio que será desenvolvido pelo aluno, ao longo do período letivo, está descrito no cronograma da Quadro 1. Neste cronograma constam todas as atividades com seus respectivos prazos para o cumprimento.

[illegible]

## Referências

BARIONI, M. C. N. et al. Seamlessly integrating similarity queries in sql. **Software: Practice and Experience**, John Wiley & Sons, Ltd., v. 39, n. 4, p. 355–384, 2009. ISSN 1097-024X. Disponível em: <<http://dx.doi.org/10.1002/spe.898>>. Citado na página 1.

DB-ENGINES. **DB-Engines Ranking**. 2017. Disponível em: <<https://db-engines.com/en/ranking>>. Acesso em: 30 de agosto de 2017. Citado na página 1.

FILHO, R. F. S. et al. Similarity search without tears: The omni family of all-purpose access methods. In: **Proceedings of the 17th International Conference on Data Engineering**. Washington, DC, USA: IEEE Computer Society, 2001. p. 623–630. ISBN 0-7695-1001-9. Disponível em: <<http://dl.acm.org/citation.cfm?id=645484.656543>>. Citado na página 2.

POLA, I. R. V. **Explorando conceitos da teoria de espaços métricos em consultas por similaridade sobre dados complexos**. Agosto 2010. Tese (Doutorado em Ciência da Computação) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2010. Citado 2 vezes nas páginas 1 e 2.

ZIGHED, D. A. et al. **Mining Complex Data**. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2008. ISBN 3540880666, 9783540880660. Citado na página 1.