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 querry: kNNq*). As consultas por abrangência *Rq*(s_q, ξ) recebem como parâmetro um elemento s_q do domínio de dados (elemento central da consulta) e um limite máximo de dissimilaridade ξ. O resultado é o conjunto de elementos da base que diferem do elemento central da consulta por no máximo a dissimilaridade indicada.

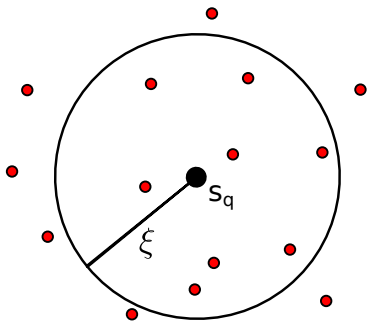

Figura 1 – **Exemplo de consulta por abrangência**

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).

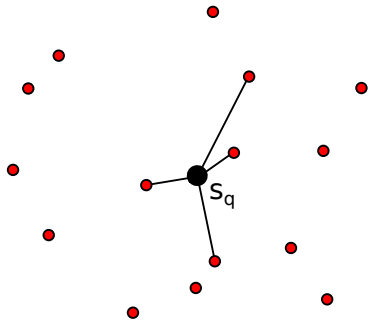

Figura 2 – **Exemplo de consulta por k-vizinhos mais próximos**

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). O cálculo do custo das operações realizadas será feito utilizando operações com B-trees, as quais o banco fornece suporte ao modelo de custo.

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 $\mathbb{S}$, sendo $\mathbb{S}$ um domínio de elementos e uma métrica $d : \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{R}^+$, temos a desigualdade triangular:

$$d(s_1,s_2) \leq d(s_1,s_3) + d(s_3,s_2) \tag{1}$$

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 ξ. 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 ξ 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 ξ 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.

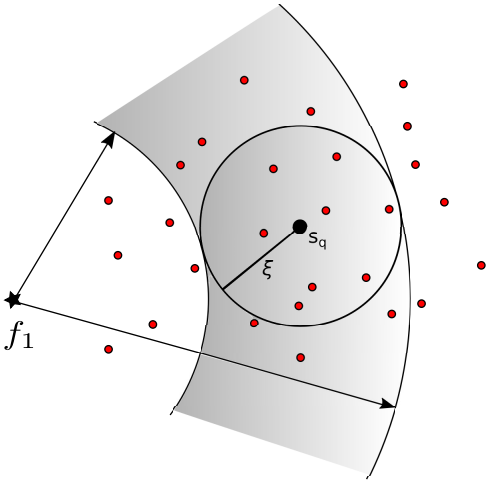

Figura 3 – **Consulta por abrangência pela técnica Omni utilizando um foco**

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. Para o armazenamento das distâncias de cada foco f_i para cada outro elemento s_k será feito utilizando uma estrutura de indexação que implementa os conceitos da técnica Omni com a estrutura da B^+-tree, originando uma nova estrutura chamada de Omni-Btree. As distâncias serão armazenadas em *l* Omni-Btrees, sendo *l* o número de focos criados para a base de dados (FILHO et al., 2001).

O principal foco deste trabalho é o emprego destas técnicas para bases de dados constituídas por imagens. Para isto, torna-se necessário o uso de uma miríade de extratores de características das imagens, para um maior refinamento do uso de consultas por similaridade. Estas características podem se referir a: atributos visuais (cor, forma, textura), atributos lógicos (identificação de elementos) e atributos semânticos (identificação de emoções humanas). As características visuais podem ser utilizadas como histogramas de cores para a análise de cor, matrizes de co-ocorrência para a análise de textura e métodos baseados em contorno para a análise de forma. Geralmente, consultas são feitas utilizando uma combinação destas características, e não apenas uma delas.

2 OBJETIVOS

O objetivo geral deste trabalho é elaborar um sistema de que seja capaz de realizar consultas por similaridade em uma base de dados complexos, mais precisamente em um conjunto de imagens. Estas imagens podem variar de fotografias de ambientes, rostos de pessoas, impressões digitais e imagens de diagnósticos médicos como radioagrafias e tomografias, por

exemplo. A ideia central é implementar todo o mecanismo de armazenamento, cálculo de distâncias e busca de dados similares no próprio banco de dados, deixando apenas a interface com o usuário fora do banco.

Após as etapas de povoamento do banco e implementação das consultas estiverem concluídas, será criada uma interface gráfica para facilitar o acesso do usuário com o sistema de consulta. Também serão estudadas maneiras de se otimizar estas consultas, utilizando diferentes métodos e formas de indexação dos elementos dentro do banco.

3 MATERIAIS E MÉTODOS

O SGBDR utilizado será o PostgreSQL, um sistema de gerenciamento de banco de dados objeto-relacional gratuito e de código-aberto. Os extratores de características das imagens utilizados são do framework Arboretum, desenvolvido pelo Grupo de Bases de Dados e Imagens (GBDI) da Universidade de São Paulo - campus São Carlos. Para a interface com o usuário, será estudado um framework que atenda as necessidades deste trabalho.

4 CONCLUSÃO

Ao término deste trabalho, espera-se que o aluno tenha obtido domínio sobre o SGBD PostgreSQL e suas ferramentas, assim como um maior entendimento e experiência no campo de estudos de dados complexos. Após o sistema ser concluído, será possível a implementação de um sistema de recuperação de imagens baseado em conteúdo (*Content-Based Image Retrieval - CBIR*). Com isto, espera-se utilizar este CBIR para o gerenciamento de imagens médicas, um campo onde há uma produção crescente de imagens utilizadas em diagnósticos e terapias. Técnicas utilizadas em diagnósticos como raciocínio baseado em casos necessitam fortemente a recuperação de imagens que podem ser valiosas para suportar certos diagnósticos (LONG et al., 2009).

5 CRONOGRAMA

Quadro 1 – Cronograma de Atividades.

Atividades	**Set**	**Out**	**Nov**	**Dez**	**Jan**	**Fev**	**Mar**	**Abr**	**Mai**	**Jun**
1. Estudo e acompanhamento da literatura	X	X								
2. Modelagem do problema	X	X								
3. Verificação das estruturas Omni		X	X							
4. Estudo dos extratores de características			X							
5. Estudo e inserção das métricas no banco		X	X							
6. Criação e povoamento da base de dados			X	X						
7. Elaboração da apresentação do TCC1			X	X						
8. Defesa do TCC1				X						
9. Implementação da busca sequencial					X	X	X			
10. Implementação da técnica Omni					X	X	X			
11. Comparação e análise de resultados							X	X		
12. Implementação da interface							X	X		
13. Elaboração da apresentação final								X	X	
14. Defesa do TCC2										X

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 2 vezes nas páginas 2 e 3.

LONG, L. et al. Content-based image retrieval in medicine: Retrospective assessment, state of the art, and future directions. v. 4, p. 1–16, 01 2009. Citado na página 4.

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.