UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANÁLISE DA EVOLUÇÃO TEMPORAL DE DADOS MÉTRICOS

ISIS CAROLINE OLIVEIRA DE SOUSA FOGAÇA

ORIENTADOR: PROF. DR. RENATO BUENO

São Carlos - SP
Outubro/2016

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANÁLISE DA EVOLUÇÃO TEMPORAL DE DADOS MÉTRICOS

ISIS CAROLINE OLIVEIRA DE SOUSA FOGAÇA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Engenharia de Software / Banco de Dados
Orientador: Dr. Renato Bueno

São Carlos - SP
Outubro/2016

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANÁLISE DA EVOLUÇÃO TEMPORAL DE DADOS MÉTRICOS

ISIS CAROLINE OLIVEIRA DE SOUSA FOGAÇA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Engenharia de Software / Banco de Dados.

Membros da Banca:

Prof. Dr. Renato Bueno

(Orientador – DC-UFSCar)

Profª. Drª. Marilde Terezinha Prado Santos

(DC-UFSCar)

Prof. Dr. Pedro Henrique Bugatti

(UTFPR/CP)

São Carlos - SP
Outubro/2016

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária UFSCar
Processamento Técnico
com os dados fornecidos pelo(a) autor(a)

F655a

Fogaça, Isis Caroline Oliveira de Sousa
Análise da evolução temporal de dados métricos / Isis Caroline Oliveira de Sousa Fogaça. -- São Carlos : UFSCar, 2017.
70 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2016.

1. Espaço métrico. 2. Evolução temporal. 3. Mapeamento. 4. Consulta por similaridade. I. Título.

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a defesa de Dissertação de Mestrado da candidata Isis Caroline Oliveira de Sousa Fogaça, realizada em 22/11/2016.

Prof. Dr. Renato Bueno
(UFSCar)

Profª. Drª. Marilde Terezinha Prado Santos
(UFSCar)

****
Prof. Dr. Pedro Henrique Bugatti
(UTFPR)

Certifico que a sessão de defesa foi realizada com a participação à distância do membro Prof. Dr. Pedro Henrique Bugatti. Depois das arguições e deliberações realizadas, o participante à distância está de acordo com o conteúdo do parecer da comissão examinadora redigido no relatório de defesa da aluna Isis Caroline Oliveira de Sousa Fogaça.

Prof. Dr. Renato Bueno
Coordenador da Comissão Examinadora
(UFSCar)

Dedico este trabalho aos meus pais, Maria Ines e José Carlos, meus verdadeiros exemplos e meu maior motivo para lutar pelos meus objetivos.

Em memória da minha amada avó Aparecida, a pessoa mais doce, forte e inspiradora que eu já conheci.

AGRADECIMENTO

Agradeço primeiramente a Deus pela vida e por ter me dado forças para chegar até aqui.

Aos meus pais, Maria Ines e José Carlos, meu esposo, Danilo, e minha irmã, Ana Helena, por me apoiarem em tudo e sempre estarem ao meu lado.

Ao meu orientador, professor doutor Renato Bueno, que com muita paciência e dedicação me conduziu na realização deste trabalho, suportando todas as minhas limitações e ansiedades.

Ao professor doutor Hermes Senger, que muito me auxiliou e incentivou no início do curso.

Aos professores do Grupo de Banco de Dados por todo apoio, incentivo e compartilhamento de conhecimento obtidos ao longo do caminho.

Aos colegas de curso, por todos os momentos compartilhados com um mesmo objetivo, pelo companheirismo, troca de aprendizados e pelas amizades. Em especial à Mirela Cazzolatto, com quem eu aprendi muito, e que muitas vezes dispôs do seu precioso tempo para me auxiliar.

Aos amigos e demais familiares por terem me apoiado. Em especial às minhas amigas Ana e Vivian, que estão sempre ao meu lado, não importa a distância.

Aos meus professores da Fatec Garça, que muito me incentivaram e acreditaram que eu conseguiria.

À CAPES, pelo período de apoio financeiro.

RESUMO

A expansão de diferentes áreas do conhecimento com os diversos tipos de informação tornou necessário o suporte a dados complexos (imagens, sons, vídeos, cadeias de DNA, entre outros), que por não possuírem uma Relação de Ordem Total (ROT), necessitam de outros mecanismos de gerenciamento, como a recuperação por conteúdo. Em geral, esses dados são representados em domínios de espaços métricos, onde apenas se tem os elementos e as distâncias entre eles. Através das características extraídas dos mesmos, realiza-se consultas por similaridade. Considerando a necessidade de associar a informação temporal a esses dados em muitas aplicações, este trabalho visa analisar a evolução temporal dos dados métricos. Para isso, uma alternativa é mapeá-los para um espaço multidimensional, a fim possibilitar a estimativa de trajetórias. Neste trabalho, foram estudados diferentes métodos de mapeamento, sendo também analisado como o mapeamento afetou a distribuição dos mesmos e, por conseguinte, a realização das estimativas. Foram propostos dois novos métodos para estimar o estado de um elemento em um tempo diferente daqueles disponíveis na base de dados, com o objetivo de reduzir no conjunto resposta a quantidade de elementos não relevantes. Os métodos propostos são baseados na redução do raio de consulta na região estimada pela delimitação do raio de consulta (*range*) e a avaliação da proximidade dos elementos retornados utilizando verificação (aproximação) do *k*-NN reverso. Foram realizados experimentos que mostraram que os métodos propostos melhoraram o resultado final das estimativas, que anteriormente eram realizadas apenas com consultas aos vizinhos mais próximos.

Palavras-chave: Espaço métrico. Evolução temporal. Mapeamento. Consulta por similaridade.

ABSTRACT

The expansion of different areas of knowledge through many types of information brought the necessity to support complex data (images, sounds, videos, strings, DNA chains, etc.), that do not have a Total Order Relationship and need other management mechanisms, like the content-based retrieval. In general, they are represented in metric space domains, where we have only the elements and the distances between them. Through the characteristics extracted from them, we perform the similarity search. Considering the necessity to associate temporal information on these data in many applications, this work aims to analyze the temporal evolve of metric data. One alternative for this is embedding them into a multidimensional space to allow trajectories estimates. We studied different methods of embedding and analyzed how this affected the data's distribution and, consequently, the estimates. Two new methods were purposed to estimate an element's status on a different time from that available in database, in order to reduce the number of non-relevant elements on search results. These methods are based on radius search reduction (range) and evaluation of retrieved element's proximity by using an approximation of reverse k-NN. We performed experiments which showed that purposed methods could improve the estimate's result, that used to be performed only using k-NN searches.

Keywords: Metric space. Temporal evolving. Embedding. Similarity search.

LISTA DE FIGURAS

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 Contexto

Conforme a complexidade das informações em diferentes áreas de conhecimento cresce, a demanda pela diversidade e complexidade de diferentes dados aumenta. Isso tem feito com que os Sistemas Gerenciadores de Banco de Dados (SGBD) necessitem suportar diferentes tipos de dados, como imagens, vídeos, sons, sequências de DNA, séries temporais, entre outros.

Em vários tipos de dados, chamados não convencionais ou complexos, não é possível realizar consultas que se baseiam na Relação de Ordem Total (ROT). Assim, não se pode ordenar imagens por seu conteúdo, como é facilmente feito com números, datas e textos curtos, considerando os operadores relacionais.

Para os dados não convencionais, portanto, restaria realizar a comparação por igualdade. Todavia, esse tipo de comparação não é de grande utilidade em muitos domínios. A possibilidade de duas imagens serem exatamente iguais, por exemplo, é mínima. Sendo assim, uma abordagem amplamente estudada e utilizada é a realização de consultas por similaridade, que recupera objetos semelhantes ao objeto de consulta baseado em suas características (HUANG, SHEN *ET AL.*, 2011), ou seja, através do conteúdo extraído dos mesmos.

Para a recuperação de um dado complexo baseado em conteúdo, são utilizados os Sistemas Baseados em Recuperação por Conteúdo (*Contend-based Retrieval* – CBR). Para o domínio de imagens, são denominados Recuperação de Imagem por Conteúdo (*Content-based Image Retrieval* – CBIR), que extraem as

informações dos dados e retornam seu vetor de características (WENGERT, DOUZE *ET AL.*, 2011), que são comparados para responder às consultas por similaridade. Assim, não se utiliza os dados em si, mas as características extraídas dos mesmos. As características de imagens representam, por exemplo: formas (KHOTANZAD E HONG, 1990), (BOSCH, ZISSERMAN *ET AL.*, 2007), textura (KRIG, 2014), (XU, YAO *ET AL.*, 2010), cores (GONZALEZ E WOODS, 2011), (WENGERT, DOUZE *ET AL.*, 2011) entre outros.

Dessa forma, extraídos os vetores de características de cada elemento do conjunto de dados, os mesmos são dispostos em um domínio métrico ou multidimensional. Para isso, uma função de distância é responsável por calcular o quão (dis)similares são os elementos (CHINO, 2004). Se essa função de distância atende às propriedades da simetria, não-negatividade e desigualdade triangular, tem-se uma métrica, ou função de distância métrica.

A necessidade de análise apresentada neste trabalho é considerar a possibilidade desses dados sofrerem modificações no decorrer do tempo, com a necessidade de uma verificação do comportamento dos mesmos, deixando de gerenciá-los como estáticos, mas assumindo-os como dinâmicos, utilizando o tempo como um parâmetro fundamental.

Dois estudos foram utilizados como embasamento para a análise proposta neste trabalho. No primeiro deles, em (BUENO, KASTER *ET AL.*, 2009b), foi proposto o espaço métrico-temporal, composto por dois espaços métricos em que um deles representa o conteúdo dos dados e o outro representa as informações do tempo. No segundo, proposto por Bueno (2009) e continuado neste trabalho, busca-se estimar a trajetória de dados métricos. Para isso, os dados métricos são mapeados para um espaço multidimensional, e o valor temporal é acrescentado a este espaço.

Com os dados dispostos em um espaço multidimensional, neste caso também chamado de espaço vetorial, é possível estimar a posição de um elemento neste espaço para um determinado instante de tempo diferente daqueles disponíveis na base, permitindo assim projetar as trajetórias desses dados no decorrer do tempo, representando seu comportamento evolutivo. Ou seja, a partir das informações reais existentes de um elemento, pode-se representar uma estimativa de como este mesmo elemento estará no futuro.

1.2 Motivação e Objetivos

Considerando os dados em domínios métricos, uma das grandes dificuldades encontradas é avaliar seu comportamento através do tempo, pois nesse domínio de dados, existem apenas os elementos e as distâncias entre eles.

Um exemplo é uma consulta a imagens de exames médicos, como uma tomografia computadorizada. Para encontrar o grau de similaridade entre duas imagens de exames realizados em datas diferentes, através da consulta por similaridade, é necessário extrair as características das imagens. Para analisar, estimar e avaliar o quanto uma terceira imagem estará diferente se o mesmo exame for realizado algum tempo depois, é preciso incluir e analisar a informação temporal das mesmas.

Considerando estes aspectos, percebe-se claramente que não apenas na medicina, mas em muitas outras áreas de aplicação, associar uma informação temporal aos dados complexos torna-se primordial.

Esse trabalho teve como objetivo, portanto, a partir do modelo proposto por Bueno (2009), mostrar a possibilidade de se gerenciar a evolução dos dados em domínios métricos considerando a informação do tempo. Mais precisamente, buscou-se avaliar e aprimorar as estimativas da trajetória desses dados no decorrer do tempo, considerando dois objetivos específicos.

No primeiro objetivo, considera-se que a influência do mapeamento sobre as estimativas pode ser analisada de modo a verificar o quanto a manutenção da distribuição dos dados, após mapeados, impacta no resultado das consultas. Buscou-se, portanto, comprovar a hipótese de que o resultado do mapeamento influencia diretamente os resultados das estimativas.

O segundo objetivo foi, com a utilização de outros dois tipos de consulta (*kAndRange* (VIEIRA, JR. ET AL., 2007) e uma aproximação do *k*-NN Reverso (KORN E MUTHUKRISHNAN, 2000)), aprimorar os resultados das estimativas, considerando a possibilidade de podar das respostas às consultas elementos considerados não relevantes por meio da utilização de mais condições de busca: delimitar a distância máxima desejada e verificar o quão próximos os elementos recuperados estão do centro de consulta.

1.3 Metodologia de Desenvolvimento do Trabalho

Após o estudo do modelo anteriormente proposto e sabendo-se que, ao utilizar dados em espaços métricos, a utilização de algoritmos de mapeamento eficientes é necessária para alcançar a melhor qualidade possível dos resultados, foram avaliados diferentes métodos de mapeamento, a saber: *Fastmap* (FALOUTSOS E LIN, 1995), *Multidimensional Scaling* (MDS), *Landmark* MDS (SILVA E TENENBAUM, 2003) e *SparseMap* (HRISTESCU E FARACH-COLTON, 1999). Nesta dissertação são apresentados os resultados para dois deles: *Fastmap* e MDS.

Esses dois métodos foram escolhidos a partir de uma revisão da literatura para encontrar os que melhor correspondiam às necessidades do trabalho. Testes comparativos foram realizados entre esses dois algoritmos para avaliar, em diversos casos, qual deles manteve mais fielmente a distribuição dos dados após a realização do mapeamento e, consequentemente, garantiu estimativas mais aproximadas aos elementos de busca, considerando também o custo.

Visto que foram utilizados dados controlados para a realização dos experimentos, as estimativas foram avaliadas através da comparação direta com os elementos estimados e os elementos reais (existentes na base de dados), tendo seus resultados exibidos através de curvas de precisão e revocação.

Além das consultas aos vizinhos mais próximos, foram realizadas consultas considerando o raio de distância e os vizinhos mais próximos reversos, com o intuito de melhorar a qualidade das estimativas, pois através do operador de busca já utilizado (consulta aos *k* vizinhos mais próximos), o número *k* de elementos sempre será retornado, independentemente das distâncias, permitindo que elementos que não sejam realmente próximos sejam recuperados.

Dessa forma, os métodos propostos possibilitam refinar os resultados das consultas ao podar do conjunto resposta elementos que não devem ser retornados, e provavelmente são irrelevantes. Esses resultados foram analisados a fim de verificar a viabilidade e efetividade do estudo realizado.

1.4 Organização do Trabalho

O restante deste trabalho está organizado da seguinte maneira:

No Capítulo 2 são apresentados os principais conceitos sobre dados complexos, espaços métricos e multidimensionais e consultas por similaridade.

No Capítulo 3 é discutido o conceito de informação temporal nos Dados Complexos, abordando a evolução temporal dos dados métricos, apresentando a abordagem proposta para a inclusão do tempo, bem como a maneira definida para estimar as trajetórias, ou seja, a evolução temporal desses dados.

A proposta do trabalho e os experimentos são apresentados no Capítulo 4, junto à análise de todas as etapas envolvidas nas consultas aos dados, desde o mapeamento até a avaliação dos conjuntos resposta dos objetos estimados.

Finalmente, no capítulo 5, é apresentada a conclusão do trabalho, elencando suas principais contribuições e as propostas para trabalhos futuros.

Capítulo 2

DADOS COMPLEXOS

2.1 Considerações Iniciais

Inicialmente, os SGBDs somente proviam suporte a tipos convencionais de dados: números, datas e cadeias curtas de caracteres, que compreendem um domínio que atende à ROT. Em outras palavras, esses tipos de dados permitem a comparação através de operadores relacionais $(<, \leq, >, \geq, =, \neq)$.

Porém, a emersão de diversos outros tipos de dados necessários para gerenciar informações em várias áreas do conhecimento tem tornado os dados convencionais insuficientes em diferentes domínios de aplicação, levando os SGBDs à necessidade de suportar outros tipos de dados: os dados não convencionais, também chamados complexos.

Dados complexos, que compreendem imagens, vídeos, sons, sequências de DNA, séries temporais, entre outros, não possuem uma ROT como os convencionais. Por isso, não podem ser recuperados da mesma forma, e a comparação por igualdade pouco contribui. Eles são, em geral, comparados por similaridade (BUENO, KASTER *ET AL.*, 2009a).

Considerando, por exemplo, duas imagens de exame de tomografia computadorizada, sabe-se que é muito difícil compará-las e constatar que elas são exatamente iguais. Mesmo que obtidas em sequência no mesmo exame, é pouco provável que elas sejam idênticas em todos os detalhes, pois fatores como a respiração do paciente, variação de luminosidade, mudança mínima de posição, entre

outros, influenciam em suas características. O mais viável, portanto, é a consulta por similaridade, que compara quão (dis)similares são dois objetos.

Nas consultas por similaridade, não se utiliza o elemento em si, mas a sua representação através do conteúdo extraído do mesmo, representado comumente por um vetor de características. Os vetores de características dos objetos são então comparados para que se possa obter o grau de similaridade entre eles: quanto menor for o valor da distância resultante, maior é a similaridade entre os mesmos (BOSCH, ZISSERMAN *ET AL.*, 2007).

Uma revisão sobre diferentes métodos e descritores de características pode ser encontrada em (KRIG, 2014).

Para calcular as distâncias entre os elementos, as funções de distância mais comumente utilizadas, em espaços multidimensionais, são as funções L_p (Minkowski), sendo as mais comuns: L_0 (*Infinity* ou *Chebychev*), L_1 (*City Block* ou *Manhatan*) e L_2 (Euclidiana) (BUENO, 2009).

Na Figura 2.1 é representado um exemplo de extração das características de uma imagem, representadas em um histograma de 256 níveis de cinza.

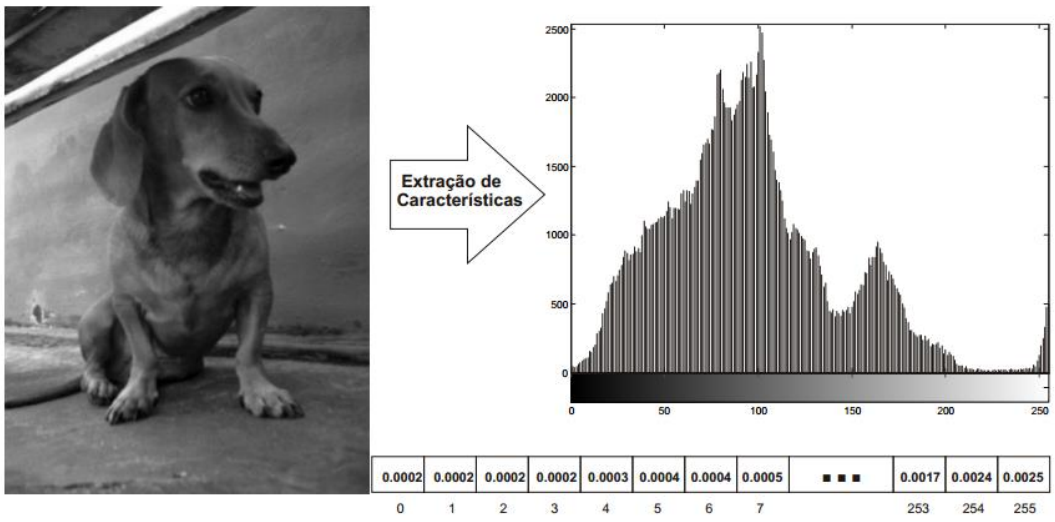

Figura 2.1: Exemplo de extração de características através de um histograma normalizado com 256 níveis de cinza (BUENO, 2009).

Em imagens, é possível realizar a recuperação do conteúdo utilizando outros tipos de características, como descritores de cores, formas ou textura (BUENO, 2009). Para aumentar a confiabilidade e alcançar melhores resultados, é possível também

utilizar combinações entre múltiplos descritores. Uma introdução a este recurso pode ser verificada mais detalhadamente em Aksoy e Haralick (2001).

Alguns descritores representam as características dos elementos através de vetores de mesmo tamanho. Outros, por sua vez, não apresentam essa igualdade na dimensão dos vetores de características, como é o caso dos Histogramas Métricos (AZEVEDO-MARQUES, TRAINA *ET AL.*, 2002).

2.2 Espaços métricos

Em espaços métricos, as informações geométricas ou dimensionais dos dados não são consideradas, isto é, as únicas informações disponíveis são os dados em si (ou seus vetores de características) e as distâncias (dissimilaridades) entre eles (AZEVEDO-MARQUES, TRAINA *ET AL.*, 2002). Por exemplo, para calcular a distância entre as palavras asa, casa e casca, aos pares, independentemente de qualquer tipo de ordem, a distância entre elas é a quantidade de símbolos que devem ser substituídos, acrescentados ou removidos para que as palavras sejam as mesmas. Uma função de distância que propicia isso é a *Levenshtein* (LEVENSHTEIN, 1996). O exemplo é ilustrado na Figura 2.2, onde pode-se notar que não há como considerar a dimensionalidade das palavras, visto que esses são diferentes, mas apenas a dissimilaridade entre as mesmas.

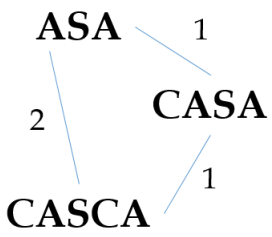

Figura 2.2: Exemplo de Espaço Métrico

Portanto, o primeiro passo é encontrar uma medida para a distância entre dois objetos, a partir de uma função de distância em que dados dois objetos x e y, a distância (ou dissimilaridade) entre eles é dada por $d(x, y)$ (FALOUTSOS, 1996).

Logo, um conjunto de dados métricos é formado por um conjunto de elementos S e uma função de distância d entre os objetos (BUENO, 2009). A função de distância deve satisfazer as três regras de um espaço métrico (TRAINA JR, TRAINA *ET AL.*, 2002):

1. Simetria: $d(x,y) = d(y,x)$;
2. Não-negatividade: $0 < d(x,y) < \infty,\ x \neq y, d(x,x) = 0$;
3. Desigualdade Triangular: $d(x,y) \leq d(x,z) + d(z,y)$.

2.2.1 Espaços Multidimensionais

Se os vetores de características dos elementos contêm o mesmo número de dimensões para todos os elementos, a dimensionalidade do conjunto pode ser considerada, tendo então os dados dispostos em um domínio multidimensional, ou espaço vetorial. Ou seja, se os objetos correspondem a vetores de valores numéricos, o espaço é denominado Espaço Vetorial, onde os objetos deste espaço são representados por *n* coordenadas de valores reais (ARANTES, 2005).

2.3 Consultas por Similaridade

A partir de um conjunto de dados S, pertencentes a um mesmo domínio $\mathbb{S}$ e uma função de distância d, são realizadas as consultas por similaridade. Os operadores de busca mais utilizados para fazer a consulta por similaridade são a consulta por abrangência e a consulta aos vizinhos mais próximos. Diversas variações desses operadores têm sido elaboradas, como a consulta aos vizinhos mais próximos reversos.

2.3.1 Consulta por abrangência (*range query*)

Na consulta por abrangência, determina-se um nível de similaridade e busca-se todos os objetos que estão dentro desse raio de abrangência. Dado um objeto de consulta q dentro de um conjunto S de objetos pertencentes a um mesmo domínio $\mathbb{S}$ e a distância máxima r_q, são retornados todos os objetos que diferem do objeto q dentro do raio estabelecido, como ilustrado na Figura 2.3.

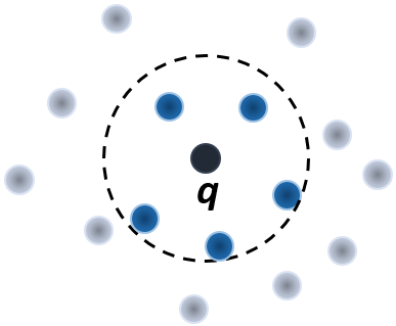

Figura 2.3: Exemplo de consulta com *Range Query*. O ponto mais escuro representa o objeto de consulta e os pontos dentro do círculo, os elementos recuperados dentro do raio de abrangência.

Através da *range query* pode-se realizar consultas como *"Quais são as cidades que estão a até 50 km de distância de São Carlos?".* Sua definição formal é dada por:

$$range(q, r_q) = \{s_i | s_i \in S, d(s_i, q) \leq r_q\}$$

É importante destacar que a consulta por abrangência depende do entendimento do usuário com relação à semântica da distância, como a unidade de medida. Por exemplo, na busca por cidades com distância de até 50 km de São Carlos, é necessário que fique claro se a unidade de medida é quilômetro, milhas, etc.

2.3.2 Consulta aos vizinhos mais próximos (*k nearest neighbors*)

Na consulta aos vizinhos mais próximos (*k*-NN), a busca consiste em encontrar um número fixo de elementos que são mais similares ao objeto de consulta. Neste tipo de consulta não se especifica um limite máximo de distância, por isso, independentemente de os elementos estarem muito próximos ou muito distantes do objeto de consulta, aqueles que obtiverem as menores distâncias dentro do limite *k* de vizinhos serão recuperados.

Com esse tipo de consulta pode-se fazer tipos de busca como *"Encontrar as 5 imagens de exames de radiografia que sejam mais parecidas com a Imagem X"*, como ilustrado na Figura 2.4.

São necessários o objeto de referência q pertencente ao domínio $\mathbb{S}$ e a quantidade $k > 0$ de elementos que deseja-se recuperar. A definição formal é dada por:

$$k - NN(q,k) = \{s_i | s_i \in A, A \subseteq S, |A| = k, \forall s_i \in A, s_j \in S - A, d(q,s_i) \leq d(q,s_j)\}$$

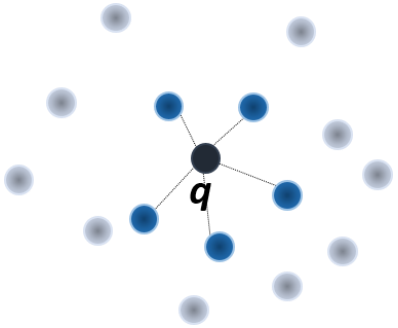

Figura 2.4: Exemplo de consulta com *k-NN*. O ponto mais escuro representa o objeto de consulta e os pontos ligados a ele, os elementos recuperados.

2.3.3 Consulta aos vizinhos mais próximos reversos (reverse *k nearest neighbors*)

A consulta aos vizinhos mais próximos reversos (KORN E MUTHUKRISHNAN, 2000) é uma variação da consulta aos vizinhos mais próximos. Dado o conjunto de dados S e um ponto q, a consulta aos *k*-NN reversos recupera todos os elementos que possuem q como um dos *k* vizinhos mais próximos (TAO, PAPADIAS *ET AL.*, 2004; TAO, PAPADIAS *ET AL.*, 2007). Uma consulta de exemplo seria *"Encontre todas as cidades que possuam São Carlos como uma de suas três vizinhas mais próximas"*, ilustrada na Figura 2.5.

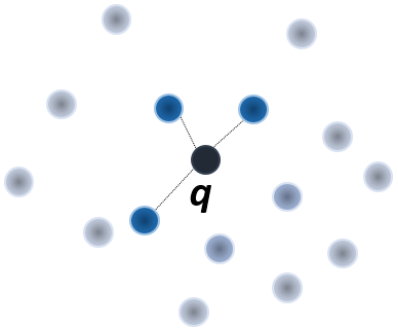

Figura 2.5: Exemplo de consulta com *k-NN* Reverso. O ponto mais escuro representa o objeto de consulta e os pontos ligados a ele, os elementos que o possuem como um dos três vizinhos mais próximos.

2.3.4 Consulta aos vizinhos mais próximos e por abrangência (*kAndRange*)

A consulta *kAndRange* (VIEIRA, JR. ET AL., 2007) corresponde à interseção das consultas *k*-NN e *range*, onde o raio inicial da consulta, que seria infinito para *k*-NN, passa a ter um valor definido. Sendo assim, a consulta *kAndRange* recupera no máximo *k* elementos cujas distâncias para o centro de consulta estejam dentro do raio estabelecido. O exemplo gráfico é ilustrado na Figura 2.6.

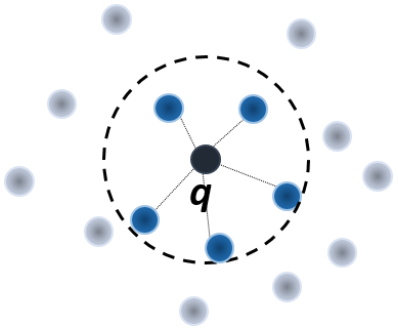

Figura 2.6: Exemplo de consulta com *kAndRange*. O ponto mais escuro representa o objeto de consulta; os pontos ligados a ele, os vizinhos mais próximos; e o círculo pontilhado, o raio de consulta.

Através dessa consulta pode-se, por exemplo, realizar a seguinte busca: *"Encontre 5 cidades mais próximas de São Carlos que estejam a até 50 km de distância"*. Se encontradas mais de 5 cidades dentro do raio de distância, apenas as 5 primeiras serão retornadas. Todavia, podem ser encontradas menos de 5, ou

nenhuma. Neste caso, a consulta limita-se a retornar apenas o que for encontrado dentro do raio de distância.

Diversas outras variações dos operadores de consulta por similaridade podem ainda ser encontrados através de uma pesquisa na literatura, sendo estes adaptáveis a diferentes necessidades de consulta, como o *K-closest pairs*, que inicia tendo infinito como o limite máximo de distância e atualiza esse limite sempre que encontrar pares de objetos cujas distâncias são menores que a atual (TAO, YI *ET AL.*, 2010; KURASAWA, TAKASU *ET AL.*, 2011).

Estruturas de indexação específicas têm sido desenvolvidas para auxiliar no gerenciamento dos dados complexos. Há diversos MAM já elaborados para a manipulação dos dados em domínios métricos, que de acordo com Chino (2004), são os métodos mais adequados para indexação e consultas por similaridade nos dados em domínios que só dispõem dos elementos e das distâncias entre eles.

2.3.5 Métodos de Acesso Métrico

Entre as diferentes estruturas de indexação utilizadas para gerenciar os diversos tipos de dados complexos, existem os Métodos de Acesso Espaciais (*Spatial Access Methods* - SAM), que auxiliam no gerenciamento de dados em espaços multidimensionais. Esse tipo método é útil quando os dados estão estruturados em vetores, onde é possível obter também informações geométricas e de coordenadas, das quais não são disponíveis em espaços puramente métricos (CHÁVEZ, NAVARRO *ET AL.*, 2001). Para esses, cujos dados são adimensionais, foram desenvolvidos os MAM, que utilizam apenas as distâncias entre os dados.

Os MAM em geral são baseados em estruturas hierárquicas. Os MAM dinâmicos permitem operações de inserções posteriores à criação da estrutura, como a *M-tree* (CIACCIA, PATELLA *ET AL.*, 1997), a *Slim-tree (TRAINA JR, TRAINA ET AL., 2000a)*, a *M*-tree* (SKOPAL E HOKSZA, 2007), a Onion-Tree (CARÉLO, POLA *ET AL.*, 2009), entre muitos outros.

2.4 Considerações Finais

Nesse capítulo foi introduzido o conceito de dados complexos, a particularidade que eles apresentam para ser gerenciados devido à ausência da ROT e algumas soluções e métodos já desenvolvidos para que a organização, recuperação e análise dos mesmos sejam possíveis.

A recuperação por conteúdo permite realizar a consulta por similaridade nesses dados, utilizando seu vetor de características. Essa recuperação pode ser aperfeiçoada através da combinação de múltiplos descritores.

Os Métodos de Acesso Métrico, que têm ganhado muitas versões, auxiliam na otimização da indexação e realização das consultas aos dados em domínios métricos.

As pesquisas e trabalhos realizados sobre recuperação por conteúdo e consultas por similaridade já é um assunto amplamente estudado. Com a associação do tempo, pode-se não só melhorar os resultados das consultas como analisar o comportamento e a evolução dos dados métricos no decorrer do tempo. No capítulo 3 uma solução já proposta é apresentada, introduzindo assim o conceito do gerenciamento do tempo nos dados métricos e a análise das trajetórias dos mesmos.

Capítulo 3

TEMPO EM DADOS COMPLEXOS

3.1 Considerações Iniciais

Sistemas de Banco de Dados convencionais em geral armazenam, em sua estrutura natural, apenas a informação no tempo presente do conjunto de dados. Se os mesmos não forem projetados para manter todos os estados dos dados, no momento em que um registro for atualizado, a informação anterior é perdida.

Porém, muitos sistemas de informação podem ter sua base de conhecimento prejudicada se em seus bancos de dados não houver a informação do tempo, isto é, um histórico da evolução dos dados. Diante disso, muitas aplicações têm sido elaboradas para que os SGBDs permitam ter a informação temporal associada ao dado, de modo que se mantenha todos os estados dos mesmos a cada atualização.

Na seção 3.2 é introduzido um breve conceito de tempo em banco de dados. Na seção 3.3 a possibilidade da informação temporal em dados métricos é discutida. Na subseção 3.3.1 é apresentado um modelo para o espaço métrico-temporal e na subseção 3.3.2 é introduzida a proposta para análise da evolução temporal em dados métricos.

3.2 Tempo em Banco de Dados

Bancos de dados temporais são projetados para permitir o armazenamento dos dados no passado, presente e futuro, registrando assim a evolução temporal das informações (EDELWEISS, 1998). Dados temporais, em geral, são dados que possuem um intervalo de tempo associado a eles, durante o qual são considerados válidos (SILBERSCHATZ, KORTH *ET AL.*, 2006). Além disso, no mundo real, mesmo que as propriedades de um objeto mudem com o passar do tempo, ele é tratado como o mesmo objeto (TANSEL, CLIFFORD *ET AL.*, 1993).

Dessa forma, diferentemente dos bancos de dados convencionais, os bancos de dados temporais possibilitam gerenciar as informações de modo que se tenha o registro histórico da transformação dos dados. Consequentemente, necessitam de um espaço maior para armazenamento, pois as informações podem aumentar muito, visto que não sobrescrevem os dados, mas mantêm armazenadas também as informações antigas.

Podem existir diferentes maneiras de interpretar os dados com a associação do tempo. Tansel, Clifford *et al.* (1993) mencionam o tempo de transação e o tempo válido, que podem ser utilizados juntos, ao que eles chamam de modelo temporal bidimensional. O tempo de transação é o tempo em que a transação acontece e o tempo válido é o tempo em que o objeto é considerado correto, o tempo que ele ocorre no mundo real. Pode ser visto como um prazo de validade inicial ou final agregado ao dado. O tempo válido pode ser usado também no futuro, quando espera-se que um fato irá ser realidade em um determinado tempo posterior (OZSOYOGLU E SNODGRASS, 1995).

O usuário também pode definir a semântica das informações e programar as aplicações quando deseja-se formular outras interpretações para o tempo. Elmasri e Navathe (2011) definem essa característica temporal como tempo definido pelo usuário.

De acordo com EDELWEISS, 1998, a definição temporal dos dados consiste em um eixo temporal, que é uma sequência de pontos consecutivos no tempo. O mais natural é que essa ordem seja linear, possibilitando total ordenação entre os instantes no tempo. Existe também a ordenação circular, para intervalos de tempo que se repetem periodicamente.

Sendo assim, entende-se o tempo como uma dimensão acrescentada ao dado, de modo que, a cada atualização, todo o conteúdo anterior continue sendo armazenado, mantendo o histórico do mesmo em uma linha temporal.

A informação temporal disponibilizada nos dados permite analisa-los de diversas outras formas. Uma delas, objeto de estudo deste trabalho, é analisar o comportamento evolutivo dos dados métricos no decorrer do tempo, analisando seus diferentes estados e verificando, por exemplo, o quanto eles evoluíram ou podem evoluir após um determinado período de tempo. Na seção 3.3 este estudo é apresentado.

3.3 Informação Temporal em Dados Métricos

De acordo com (BUENO, KASTER *ET AL.*, 2009b), dados em domínios métricos com a informação do tempo são necessários para a recuperação da informação em diferentes domínios de aplicações, como:

- Acompanhamento de grandes construções, através do monitoramento dinâmico de sensores colocados sobre as estruturas;
- Sensoriamento de equipamento, através da supervisão de dados de aparelhos industriais e científicos, como maquinários, fornos metalúrgicos e dutos de refinaria de petróleo;
- Análise de séries temporais, através do estudo climático, acompanhamento e armazenamento de mudanças, comportamento e tendências em diferentes períodos.

Além desses, pode-se citar outras áreas de conhecimento, como:

- Meteorologia: Estudos meteorológicos necessitam prover diversas estimativas com relação aos fenômenos atmosféricos. A análise da variação de indicadores climáticos durante o tempo pode auxiliar esse processo.
- Agricultura: Muitas pesquisas nessa área envolvem o acompanhamento do desenvolvimento, modificação, inovação e experimentação em diversos tipos de plantações considerando inclusive variações meteorológicas. O acompanhamento temporal, utilizando as medidas de distância entre

diferentes estágios de evolução, pode permitir identificar e interpretar esses avanços.

- Medicina: Análises de exames médicos e diagnósticos clínicos em geral, como a evolução generalizada ou individualizada de uma determinada doença.

A necessidade de gerenciar informações temporais se aplica a muitos outros domínios de aplicação em bancos de dados. Bueno, Kaster *et al.* (2009b) desenvolveram um modelo que faz a comparação dos dados complexos por similaridade considerando informações temporais, denominado espaço métrico-temporal.

3.3.1 Espaço Métrico Temporal

O modelo métrico-temporal proposto por Bueno, Kaster *et al.* (2009b) inicialmente projeta as distâncias métrica e temporal separadamente, composto por uma componente métrica e uma componente temporal.

Sendo o espaço métrico definido como $\langle S, d_s \rangle$ onde S representa o conjunto de dados pertencentes ao domínio da aplicação, e $d_s: S \times S \rightarrow \mathbb{R}^+$ uma métrica que torne possível calcular a dissimilaridade entre os elementos do domínio, define-se outro espaço métrico $\langle T, d_t \rangle$ para as medidas de tempo. Dessa forma, T representa as medidas de tempo e $d_t: T \times T \rightarrow \mathbb{R}^+$ a métrica para calcular a similaridade entre dois valores de tempo pertencentes ao domínio T.

Assim, o espaço métrico-temporal é definido como um par $\langle V, d_v \rangle$ de forma que $V = S \times T$ e $d_v: V \times V \rightarrow \mathbb{R}^+$ representa a métrica entre os objetos do espaço métrico juntamente com as informações temporais, à qual denomina-se função de distância métrico-temporal, composta pelas métricas d_s e d_t. Ou seja, um espaço métrico temporal é formado por uma componente métrica S e uma componente temporal T. É possível também fazer a representação de múltiplas informações métricas e temporais, através da agregação de outros espaços métricos, tanto na componente métrica quanto na componente temporal.

Para o cálculo da similaridade entre os elementos de um espaço métrico-temporal, é preciso que as funções de distância d_s e d_t sejam combinadas adequadamente, com o peso ideal para cada uma delas. Para isso, as métricas podem

ser agregadas em uma métrica produto, definindo assim um fator de escala que verifique a contribuição equivalente das componentes métrica e temporal no cálculo da similaridade (BUENO, 2009).

Uma forma para calcular o fator de escala é considerar os espaços métricos como mapeados em um espaço vetorial através do cálculo da dimensão fractal. Essa, por sua vez, possibilita definir a dimensão intrínseca das componentes métrica e temporal (TRAINA JR, TRAINA *ET AL.*, 2000b).

Esse modelo permite melhorar os resultados das consultas por similaridade, pois tendo a informação do tempo como parte do espaço métrico, é possível aproximar elementos com maior similaridade temporal e afastar outros com maior distância temporal.

Não possibilita, porém, analisar o comportamento evolutivo dos dados no decorrer do tempo, com a análise de sua trajetória, por exemplo, visto que os dados estão em espaços métricos e, portanto, são adimensionais, impossibilitando que suas coordenadas sejam analisadas e estimadas individualmente. Para possibilitar a estimativa de trajetória desses elementos, realizando, por exemplo, interpolações e extrapolações dos valores de suas posições no espaço, foi proposto em Bueno, 2009 o mapeamento desses dados em um espaço multidimensional, com a inclusão da relação de ordem temporal.

Portanto, não se pretende apenas saber a informação absoluta do tempo em relação aos dados, mas como a mesma pode influenciar nas características e na similaridade entre os elementos, ou seja, como ocorre a evolução dos dados métricos no decorrer do tempo.

3.3.2 Evolução temporal em dados métricos

Em Bueno (2009) foi proposta uma maneira de estimar as trajetórias de dados em domínios de espaços métricos, isto é, a partir da informação temporal associada aos elementos, verificar o comportamento evolutivo dos mesmos. Mais especificamente, a partir de um elemento existente em determinados instantes de tempo, é possível realizar uma estimativa de como estaria esse mesmo elemento em um outro instante temporal.

Para isso, é necessário que os dados em espaço métrico sejam mapeados para um espaço multidimensional, a fim de se ter o mesmo número de dimensões entre todos os elementos do conjunto e então obter o posicionamento do elemento estimado no espaço mapeado. É necessário ressaltar que não é estimado o elemento em si (ou seu vetor de características), mas sim sua posição no espaço multidimensional no qual foram mapeados os elementos métricos (seu vetor de coordenadas nesse espaço). A partir desse mapeamento, cada elemento do espaço métrico passa a ser representado no espaço multidimensional por esse novo vetor de características.

Um exemplo é ilustrado na Figura 3.1, onde os elementos *A*, *B*, *C* e *D* são mapeados para um espaço multidimensional com três dimensões.

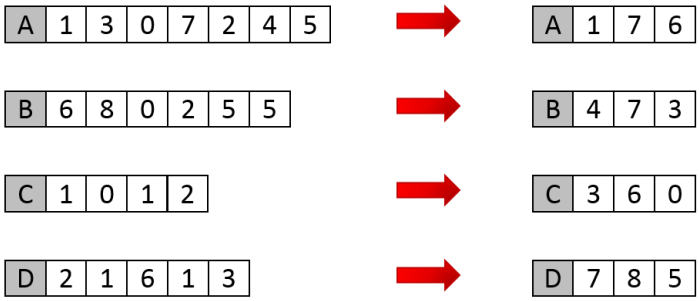

Figura 3.1: Exemplo de mapeamento do espaço métrico para o espaço multidimensional, onde os vetores de características passam a ter o mesmo número de dimensões.

Dessa forma, as consultas e estimativas são realizadas no espaço mapeado. Considerando-se que cada atributo do vetor de características representa uma dimensão, a estimativa é realizada utilizando os valores de cada atributo e o valor temporal adicionado a eles. No exemplo ilustrado na Figura 3.2, a partir do elemento *A* nos tempos *10* e *15*, estimou-se as características do mesmo no tempo *20*.

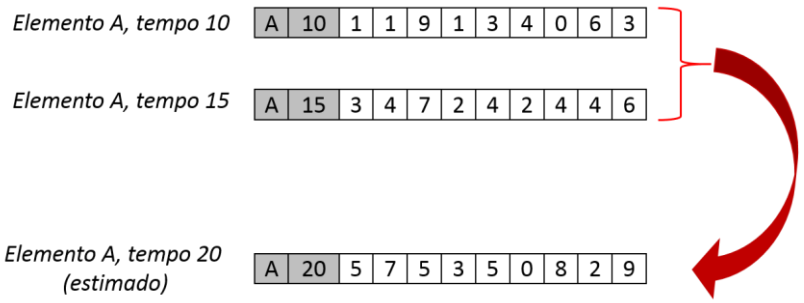

Figura 3.2: No espaço mapeado, tendo o mesmo número de atributos para todos os objetos em todos os seus instantes de tempo, estima-se seus atributos em um outro instante de tempo.

Considere-se como exemplo a necessidade de acompanhar o diagnóstico de um paciente utilizando imagens de exames médicos. Essas imagens estão dispostas em uma base de dados de imagens de exames de vários pacientes em diferentes estágios de tratamento. Na Figura 3.3 elas são representadas como pontos em um espaço bidimensional. O paciente P_A tem indexadas: uma imagem ao iniciar o tratamento $(t = 0)$ e outra com 12 meses de tratamento $(t = 12)$. A partir dessas duas imagens, deseja-se estimar qual será o estado do paciente quando chegar aos 15 meses de tratamento.

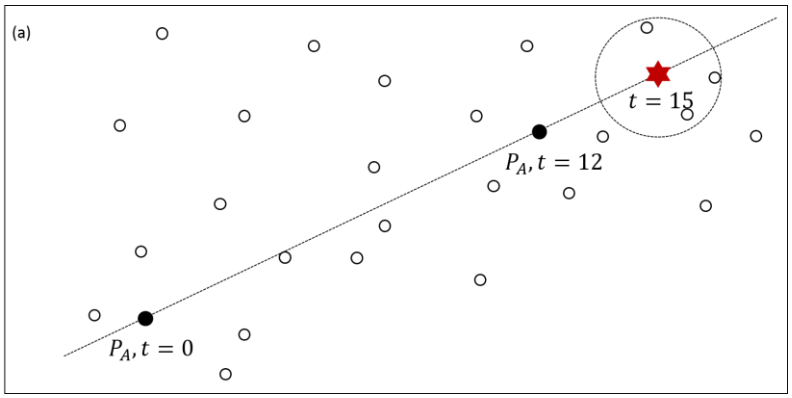

Figura 3.3: Estimativa do resultado do exame do paciente quando estiver com 15 meses de tratamento, a partir de suas imagens existentes no início do tratamento e com 12 meses de tratamento (Adaptada de Bueno, 2009).

No espaço mapeado, através das características das imagens do paciente com 0 e 12 meses de tratamento, estima-se quais seriam as características da mesma imagem com 15 meses de tratamento. Nessa estimativa é realizada uma consulta aos vizinhos mais próximos, retornando as imagens que mais se assemelham à mesma. É importante ressaltar que, a partir de um elemento estimado no espaço mapeado, não se pode "reconstruir" esse mesmo elemento no espaço métrico original (BUENO, 2009).

Para fazer a estimativa em um tempo intermediário (aos 6 meses de tratamento) realiza-se o mesmo procedimento. Suponha-se que o paciente iniciou o tratamento e apenas aos 12 meses realizou um novo exame. Deseja-se obter uma estimativa de como o mesmo estaria aos 6 meses de tratamento. Com as mesmas imagens disponíveis, estima-se suas características aos 6 meses, como ilustrado na Figura 3.4.

É possível ainda realizar a estimativa em tempo passado. Suponha-se agora que as imagens do paciente P_A sejam com 6 meses $(t = 6)$ e com 12 meses $(t = 12)$ e pretende-se estimar como a mesma estava no início do tratamento $(t = 0)$. Novamente, o mesmo procedimento é realizado, agora tendo como parâmetros $(t = 6)$ e $(t = 12)$. O exemplo é ilustrado na Figura 3.5.

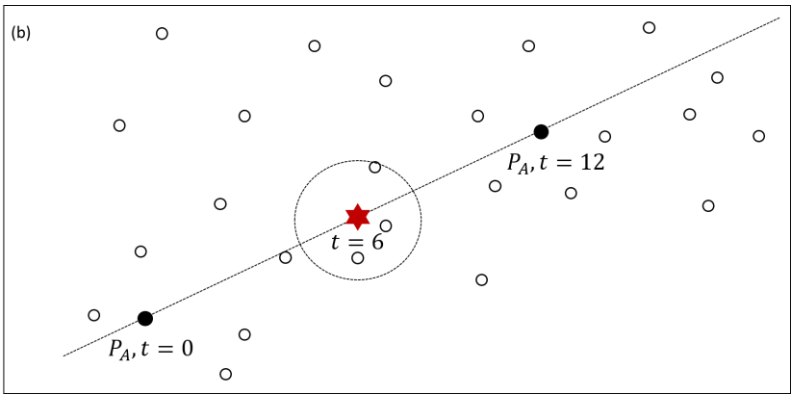

Figura 3.4: Estimativa do resultado do exame do paciente quando esteve com 6 meses de tratamento, a partir de suas imagens existentes no início do tratamento e com 12 meses de tratamento (Adaptada de Bueno, 2009).

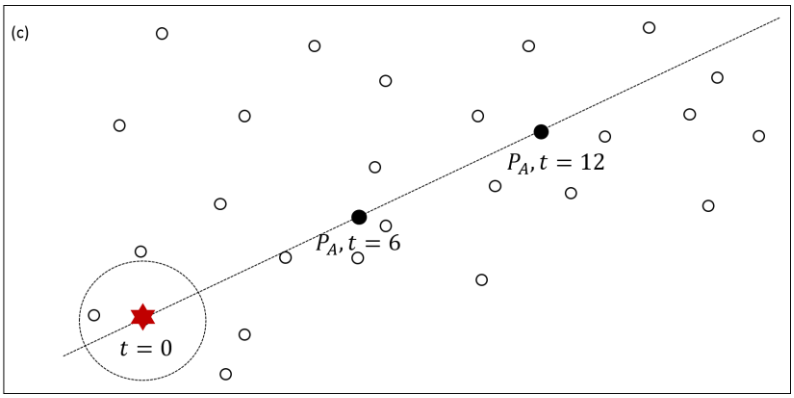

Figura 3.5: Estimativa do resultado do exame do paciente quando esteve no início do tratamento, a partir de suas imagens existentes com 6 meses de tratamento e com 12 meses de tratamento (Adaptada de Bueno, 2009).

Bueno (2009) apresentou também um experimento inicial para validar o resultado das estimativas. Foram utilizados Histogramas (níveis de cinza), *Zernike* (formas) e a informação temporal, aplicados a objetos com 36 variações representando o tempo. Foi utilizado o conjunto de imagens ALOI (*Amsterdam Library of Object Images*) (GEUSEBROEK, BURGHOUTS *ET AL.*, 2005), que consiste em um conjunto de 1000 objetos, cada um deles em vários ângulos de rotação. As variações

entre os ângulos de rotação, para o experimento, foram utilizadas para representar o tempo.

Os dados métricos temporais foram mapeados para o espaço multidimensional utilizando o algoritmo *FastMap* (FALOUTSOS E LIN, 1995), caracterizado pela sua rápida execução. Para cada elemento presente na base de dados foram utilizadas duas instâncias em tempos diferentes, a partir das quais podia-se estimar o seu estado em qualquer tempo intermediário, passado e futuro. Para cada caso foram considerados três níveis, de acordo com a distância temporal entre os elementos utilizados como referência para a consulta. Para o *intermediário* 2, por exemplo, as duas instâncias do elemento utilizado estão separadas por 20 unidades de tempo, como demonstrado na Figura 3.6. As estimativas, porém, foram limitadas a valores de tempo presentes na base de dados, para posterior verificação.

Para fazer a estimativa, por exemplo, do elemento *x* no tempo 25, a partir de duas imagens suas nos tempos 5 e 15, estimou-se suas características no tempo 25 através de interpolação/extrapolação linear. Nessa estimativa, realizou-se uma consulta aos 10 vizinhos mais próximos. No conjunto de dados original, com o elemento *x* no tempo 25 existente, realizou-se a consulta aos 10 vizinhos mais próximos, sendo este o conjunto resposta considerado correto. Os resultados das consultas ao elemento estimado e ao elemento real foram então comparados a fim de verificar se os mesmos elementos foram retornados nos dois conjuntos resposta.

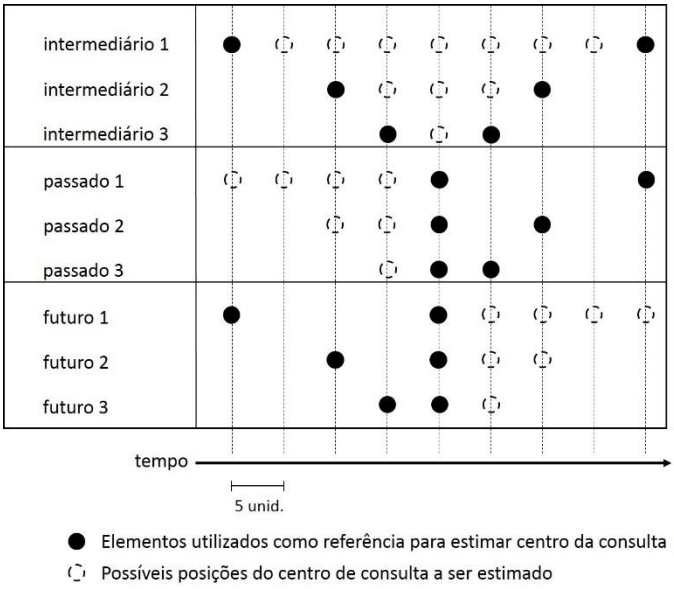

Figura 3.6: Tipos de consultas realizadas no experimento (BUENO, 2009).

Os resultados da avaliação da qualidade das estimativas (comparação das consultas ao objeto estimado e ao objeto real) são mostrados nos gráficos precisão *versus* revocação, nas Figuras 3.7 e 3.8, representando os conjuntos de Histogramas e *Zernike*, respectivamente.

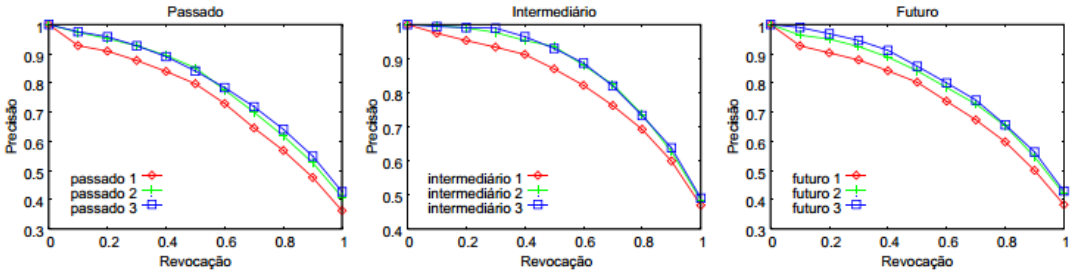

Figura 3.7: Desempenho das estimativas no conjunto de Histogramas (Bueno, 2009).

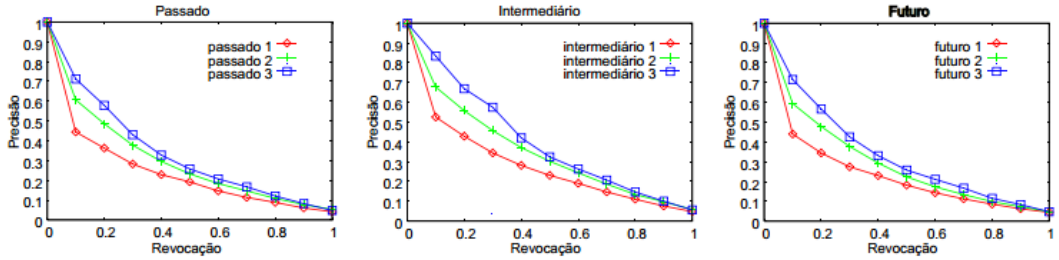

Figura 3.8: Desempenho das estimativas no conjunto *Zernike* (Bueno, 2009).

Pode-se perceber facilmente que os resultados das consultas utilizando conjuntos de histogramas e *Zernike* não atingiram os mesmos níveis de precisão. Um fator de grande impacto na qualidade das estimativas é a qualidade do mapeamento, pois quanto menor a qualidade do mapeamento, maiores as chances de os dados não manterem sua distribuição.

Portanto, a qualidade do mapeamento foi também avaliada através de uma consulta aos 10 vizinhos mais próximos diretamente em cada elemento (sem estimativas), no conjunto original e no conjunto mapeado. Como esperado, os níveis de precisão apresentaram o mesmo comportamento para os dois conjuntos de descritores (Histogramas e *Zernike*) das estimativas, como é mostrado na Figura 3.9.

É importante destacar que, para a análise proposta da evolução temporal dos dados métricos, pode-se fazer o mapeamento dos dados métrico-temporais para um espaço multidimensional, ou mapear apenas o espaço métrico e então incluir a

dimensão temporal. Nos experimentos apresentados, Bueno (2009) utilizou os dados métrico-temporais.

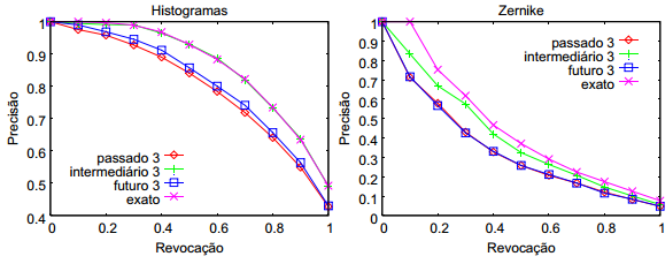

Figura 3.9: Avaliação da qualidade do mapeamento (Bueno, 2009).

A partir desse primeiro experimento, com a proposta para trabalho futuro, deu-se continuidade à análise apresentada como desenvolvimento do projeto de mestrado, através da avaliação de opções de mapeamento que melhor mantenham a distribuição dos dados e sua influência sobre as estimativas realizadas. Além disso, outras formas de consulta, através da delimitação do raio de consulta na região estimada e da avaliação da proximidade dos elementos retornados utilizando aproximação do *k*-NN reverso, foram estudadas de modo a aprimorar os resultados das estimativas.

3.4 Considerações Finais

Sabendo-se que a manipulação de dados não convencionais é de fundamental importância para o pleno gerenciamento das informações em muitos domínios de aplicação, é evidente que o gerenciamento da informação temporal agregada aos dados também é importante.

Sabe-se também, como já exemplificado, que em muitos casos, mesmo com a informação temporal, somente a consulta por similaridade não é suficiente para atender às necessidades de determinado domínio. Dessa forma, a análise das trajetórias dos espaços métricos busca justamente atender essa demanda.

Já foi possível perceber, pelos primeiros experimentos realizados por Bueno (2009) que o desenvolvimento desse modelo pode trazer resultados satisfatórios,

além de uma alternativa no que diz respeito ao gerenciamento de dados métricos, com o método proposto para a inclusão da informação do tempo.

No Capítulo 4 são mostradas as alternativas de melhorias aplicadas para refinar e aprimorar esse estudo, bem como os resultados obtidos com cada uma delas.

Capítulo 4

RESULTADOS

4.1 Considerações Iniciais

Neste capítulo são apresentados os resultados obtidos a partir do estudo proposto, que consiste na análise da evolução temporal dos dados em domínios métricos. Neste domínio de dados, as únicas informações disponíveis são os próprios dados (comumente representados por vetores de características) e as distâncias entre eles.

Considerando o trabalho realizado por Bueno (2009), onde os dados métricos são mapeados para um espaço multidimensional e, no espaço mapeado, com a informação do tempo, são realizadas estimativas aos elementos em outros instantes temporais, os estudos propostos neste trabalho foram realizados em duas frentes de trabalho.

Na primeira delas, o objetivo foi avaliar a influência do mapeamento na qualidade das estimativas, mostrando que o resultado do mapeamento interfere no resultado das estimativas.

A partir de uma revisão da literatura, foram selecionados dois algoritmos de mapeamento: o *Fastmap* (FALOUTSOS E LIN, 1995), que apresenta complexidade computacional linear de $O(kN)$; e o MDS (YOUNG E HOUSEHOLDER, 1938), (COX E COX, 2000), que se trata de um algoritmo mais custoso, $O(N^2)$, porém capaz de manter mais fielmente a distribuição das distâncias entre os elementos.

Na segunda frente de trabalho foram propostos dois métodos de busca com o objetivo de melhorar a qualidade das respostas estimadas. Para isso, aproximações de consultas por abrangência e aos vizinhos reversos foram utilizadas e analisadas a fim de proporcionar melhorarias nos resultados obtidos.

O restante deste capítulo está organizado da seguinte maneira: Na seção 4.2 são apresentadas as configurações dos experimentos. A avaliação da qualidade do mapeamento é discutida na Seção 4.3. A verificação da influência do mapeamento sobre as estimativas é apresentada na Seção 4.4. Na seção 4.5, a proposta para o aprimoramento das estimativas é introduzida. Finalmente, nas seções 4.6 e 4.7 são apresentadas as consultas alternativas utilizando os operadores *kAndRange* e a aproximação do *Reverse k*-NN.

4.2 Configuração dos Experimentos

Para a realização dos experimentos, foi utilizado o conjunto de imagens ALOI (GEUSEBROEK, BURGHOUTS *ET AL.*, 2005), que consiste em um conjunto de 1000 objetos, fotografados em 72 ângulos de visão. Cada imagem varia, portanto, em 5 graus de um ângulo de visão para o outro. Deste conjunto, foram utilizadas as 1000 imagens em 10 ângulos de rotação, variando de 0º a 45º. Cada ângulo de rotação foi utilizado para representar a variação do tempo, representando 10 instantes temporais com diferença de 5 unidades de tempo.

As imagens são classificadas de acordo com o objeto fotografado. Portanto, cada imagem é identificada individualmente pelo seu identificador de classe (ID) e seu tempo.

Na Figura 4.1 é mostrado um exemplo de uma imagem do conjunto utilizado com as dez variações, representando uma classe. Todas elas possuem o ID 15, cada uma com seu ângulo de rotação.

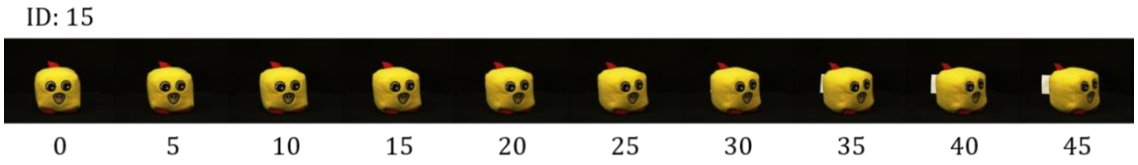

Figura 4.1: Exemplo de imagem do conjunto ALOI utilizada. Cada imagem varia seu ângulo de rotação a cada 5 graus, representando 5 unidades de tempo, de 0 a 45.

Para representar as características das imagens, foram utilizados Histogramas de 256 níveis de cinza. Dessa forma, tendo todos os elementos o mesmo número de dimensões (256), tem-se os dados em um espaço multidimensional.

A função de distância utilizada em para a realização de todos os experimentos foi a L_2 (Euclidiana).

Optou-se pela escolha de um vetor de características que pode ser representado em um espaço multidimensional devido à possibilidade de verificar a influência do mapeamento sobre os dados. Ou seja, o próprio vetor de característica pode ser utilizado na estimativa das trajetórias, possibilitando analisar a influência do mapeamento nas estimativas realizadas, a fim de evidenciar a efetividade do estudo proposto, de forma que se comprove que resultados similares são alcançados nos dados mapeados e em seu espaço original.

4.3 Análise da qualidade do Mapeamento

Na primeira etapa do trabalho, os dados do conjunto de Histogramas foram mapeados utilizando os algoritmos MDS e *Fastmap.* Para definição do número de dimensões do espaço mapeado, foi considerado o valor encontrado por Bueno (2009), calculado através da dimensão intrínseca do conjunto, definida pela dimensão fractal (TRAINA JR, TRAINA *ET AL.*, 2000b). O valor encontrado por Bueno (2009) para a dimensão intrínseca foi 9, aproximado aqui para 10 (favorecendo assim a manutenção da distribuição original dos dados).

Foram ainda realizados mapeamento para 3 dimensões a fim de verificar os resultados obtidos em um caso de dimensionalidade ainda mais baixa e facilmente representada no espaço.

4.3.1 Experimentos

Para avaliar a qualidade do mapeamento, verificou-se se a distribuição dos elementos no espaço original foi mantida no espaço mapeado, sendo que, para cada elemento de cada conjunto (original, MDS e *Fastmap*), realizou-se uma consulta aos 10 vizinhos mais próximos. As respostas das consultas foram avaliadas através de curvas de precisão e revocação, exibidos na Figura 4.2, em que as respostas do conjunto original são consideradas corretas, ou seja, relevantes.

A exibição dos resultados em curvas de precisão e revocação permite identificar o quanto os resultados de um conjunto estão de acordo com o outro. De acordo com Hjaltason e Samet (), seja R_O o conjunto de objetos resultantes de uma consulta por similaridade ao espaço métrico original e R_E o das consultas ao espaço mapeado, alguns objetos R_O podem não estar em R_E e vice-versa.

Assim, a premissa é de que o resultado correto é o de R_O. A precisão calcula a proporção dos objetos em R_E que estão em R_O, definida como $\frac{|R_E \cap R_O|}{|R_O|}$. A revocação calcula a proporção dos resultados em R_O presentes em R_E, definida como $\frac{|R_E \cap R_O|}{|R_E|}$. Logo, quando a precisão atinge 100%, todos os objetos em R_E estão corretos. Não significa porém que R_E contém todos os elementos corretos. Quando a revocação atinge 100%, todos os objetos corretos estão presentes em R_E.

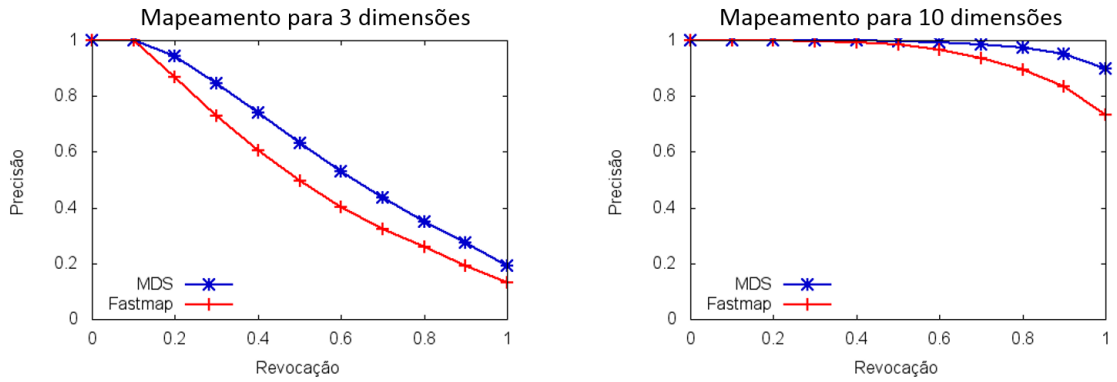

Figura 4.2: Avaliação da qualidade do mapeamento dos dados para 3 e 10 dimensões com os algoritmos MDS e *Fastmap*.

Através da visualização dos gráficos, é possível verificar que para os dois mapeamentos, o algoritmo MDS atinge melhores resultados. Percebe-se também que,

conforme o esperado, o mapeamento para 10 dimensões obteve melhores resultados do que para 3 dimensões, sendo este o selecionado para dar continuidade ao restante do trabalho.

No mapeamento para 10 dimensões, o MDS alcançou um mínimo de 90% de precisão, com precisão média de 98%, enquanto o *Fastmap* apresentou uma queda maior, chegando a 73%, com precisão média de 93%. Isso significa que o MDS mantém mais fielmente as distâncias entre os dados e, por conseguinte, a distribuição dos mesmos, bem como verificado também por Paulovich (2008) em seus experimentos. Vale ressaltar que o *Fastmap* apresenta complexidade computacional linear de $O(kN)$; e o MDS é mais custoso, com complexidade $O(N^2)$.

A partir do mapeamento dos dados, todos os experimentos das seções seguintes foram realizados utilizando os três conjuntos: original, mapeado para 10 dimensões com MDS e mapeado para 10 dimensões com *Fastmap*. No decorrer do texto, eles serão chamados Hist256, MDS10 e Fastmap10.

No restante dos experimentos, os resultados dos conjuntos mapeados também foram comparados com os do conjunto original a fim de se verificar a influência do mapeamento sobre os elementos.

4.4 Estimativas

As estimativas foram realizadas em diferentes tempos, da seguinte maneira: a partir de duas instâncias de um mesmo elemento em tempos diferentes, estima-se, através de interpolação/extrapolação linear, os valores de cada coordenada em um terceiro instante de tempo. Nessa estimativa, realiza-se uma consulta aos 10 vizinhos mais próximos, a fim de se saber como o elemento de consulta possivelmente estaria naquele instante de tempo. Vale ressaltar que a partir de um elemento estimado no espaço mapeado, não se pode “reconstruir” esse mesmo elemento no espaço métrico original (BUENO, 2009).

Realiza-se então uma consulta aos 10 vizinhos mais próximos no elemento real no conjunto original, da mesma maneira como foi feito em Bueno (2009). Os resultados das duas consultas são então comparados, sendo que, quanto maior a semelhança entre os mesmos, maior a qualidade da estimativa.

Na Figura 4.3 é ilustrado o esquema gráfico de um exemplo de estimativa e avaliação, onde, para o elemento *x* nos tempos 5 e 15, realizou-se uma estimativa do mesmo no tempo 25. Nessa estimativa, realizou-se uma consulta aos 10-NN. No elemento *x* no tempo 25, existente no conjunto de dados original, realizou-se também uma consulta aos 10-NN. Os elementos retornados para as duas consultas foram comparados a fim de saber quão próximos foram os resultados, ou seja, quanto a estimativa se aproxima do elemento real.

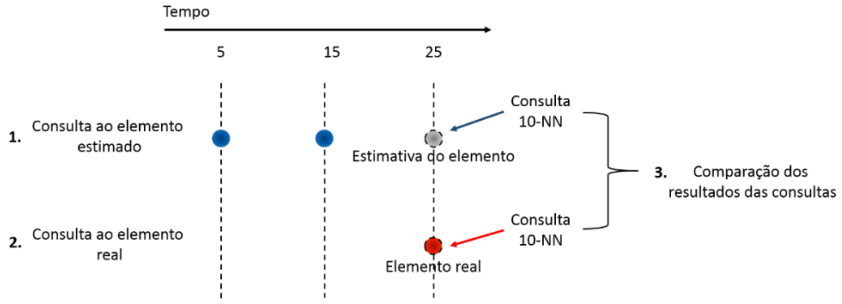

Figura 4.3: Esquema gráfico de realização e avaliação das estimativas.

4.4.1 Experimentos

As estimativas foram realizadas para os 1000 objetos do conjunto, baseando-se em duas imagens em tempos diferentes, e foram geradas as médias dos resultados das avaliações. Vale ressaltar que para realizar as estimativas dos elementos em Hist256 o procedimento é o mesmo: os dados são dimensionais, portanto é possível realizar a interpolação/extrapolação sem a realização de um mapeamento prévio para o espaço dimensional. Foram considerados 4 níveis de distância temporal entre as instâncias dos elementos, os quais foram denominados 1, 2, 3 e 4 para passado, intermediário e futuro, representando diferença de, respectivamente, 20, 15, 10 e 5 unidades de tempo, como é mostrado na Figura 4.4.

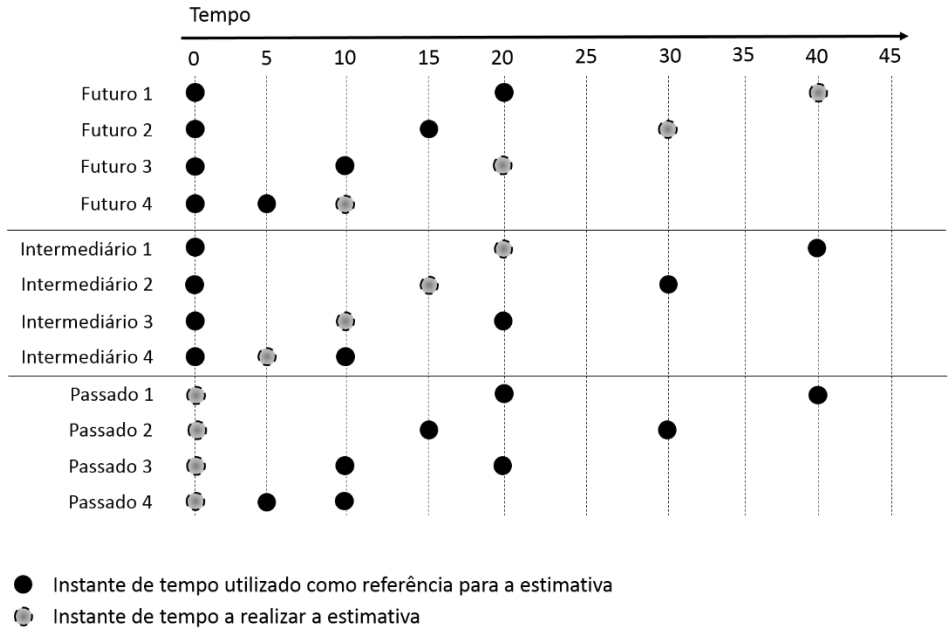

Figura 4.4: Exemplos dos intervalos de tempo para a realização das estimativas.

As avaliações das estimativas para cada conjunto são apresentadas em gráficos de precisão e revocação, exibidos nas Figuras 4.5, 4.6 e 4.7.

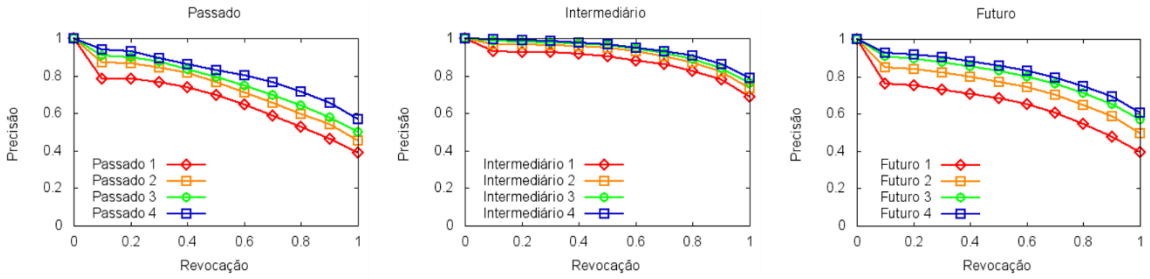

Figura 4.5: Avaliação das estimativas realizadas aos elementos no conjunto Hist256.

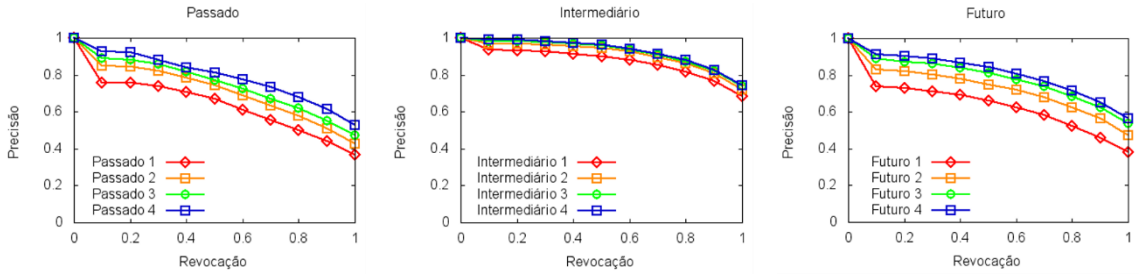

Figura 4.6: Avaliação das estimativas realizadas no conjunto MDS10.

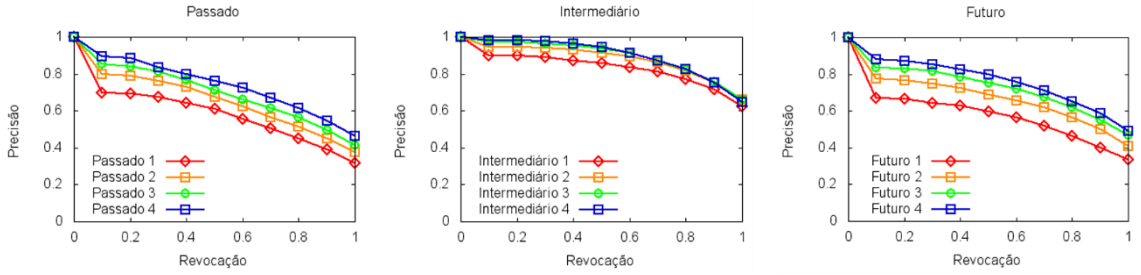

Figura 4.7: Avaliação das estimativas realizadas no conjunto Fastmap10.

Através dos gráficos, é possível perceber que o comportamento se manteve, porém com pequenas diferenças nos níveis de precisão, que ficaram acima para Hist256 e abaixo para Fastmap10, sendo a diferença um pouco mais expressiva neste último. Alcançou-se precisões médias mínima de 64% (Passado 1) e máxima de 93,6% (Intermediário 4) para Hist256. Para MDS10, as precisões médias alcançadas variaram de 61% (Passado 1) a 92% (Intermediário 4). Para Fastmap10, as precisões médias foram de 55,3% (Passado 1) a 88,6% (Intermediário 4).

Para a realização do restante dos experimentos, optou-se por utilizar as estimativas referentes ao Futuro 4. Na Figura 4.8 é possível verificar o gráfico de precisão e revocação comparativo entre os resultados das estimativas em Futuro 4 para Hist256, MDS10 e Fastmap10, onde pode-se perceber que os resultados para os conjuntos Hist256 e MDS10 são muito próximos, com precisões médias de 81,6% e 79,2%, respectivamente. Vale ainda ressaltar que se considerarmos apenas os níveis iniciais de revocação, essa diferença é ainda menor.

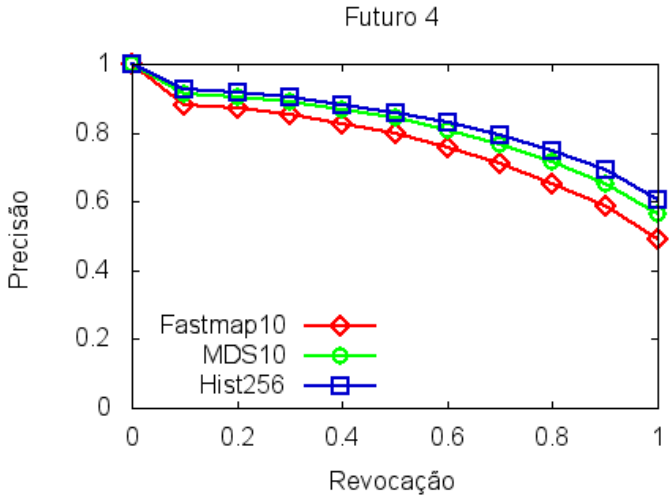

Figura 4.8: Comparação da avaliação das estimativas realizadas ao Futuro 4 nos conjuntos Hist256, MDS10 e Fastmap10.

4.5 Refinando os resultados das estimativas

Deve-se considerar, contudo, que a qualidade final dos resultados não deve ser medida apenas com a consulta *k*-NN, pois os vizinhos mais próximos (que sempre são retornados) podem não estar próximos da estimativa do estado de um elemento para serem considerados relevantes. Utilizando o exemplo do paciente descrito na seção 3.3.2, suponha que se tenha dois pacientes P_A e P_B, onde P_A tem imagens de exames nos tempos 0 e 12 e P_B nos tempos 2 e 4. Deseja-se estimar o estado de P_A no tempo 15 e de P_B no tempo 8. Realiza-se a consulta aos 5 vizinhos mais próximos nessas estimativas. Nesse exemplo, as imagens retornadas para o paciente P_B são claramente mais próximas da estimativa do estado desse paciente do que as imagens retornadas para P_A, indicando um melhor resultado, como ilustrado na Figura 4.9.

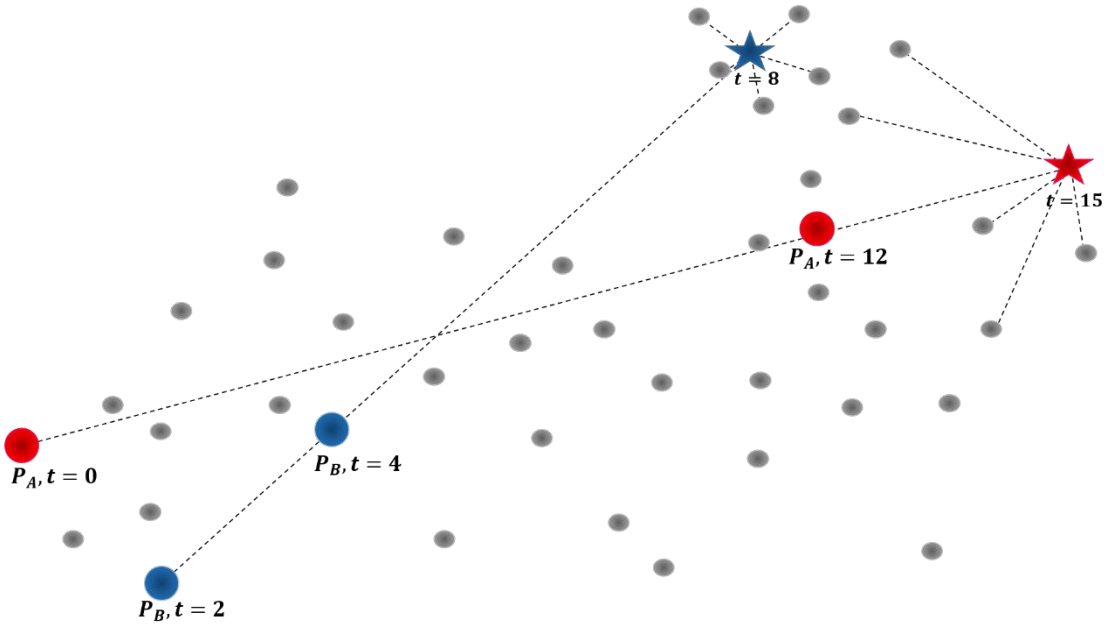

Figura 4.9: Exemplo de consultas aos vizinhos mais próximos nas estimativas referentes a dois pacientes. A proximidade entre o elemento de consulta e os elementos retornados pode variar muito entre as consultas.

Além disso, sabe-se que os 5 elementos mais próximos serão retornados de qualquer maneira, não importa qual seja a distância deles à posição estimada. Isso permite que elementos distantes sejam retornados, uma vez que podem não existir elementos realmente próximos, fazendo com que elementos não relevantes sejam

retornados. Considerando o mesmo exemplo dos pacientes P_A e P_B, possivelmente apenas dois dos elementos retornados seriam relevantes para representar a estimativa do paciente P_A no tempo 15 (os dois mais próximos). Pode ocorrer, por exemplo, de os elementos retornados para P_A serem mais próximos da estimativa P_B. Na Figura 4.10 esse exemplo é demonstrado.

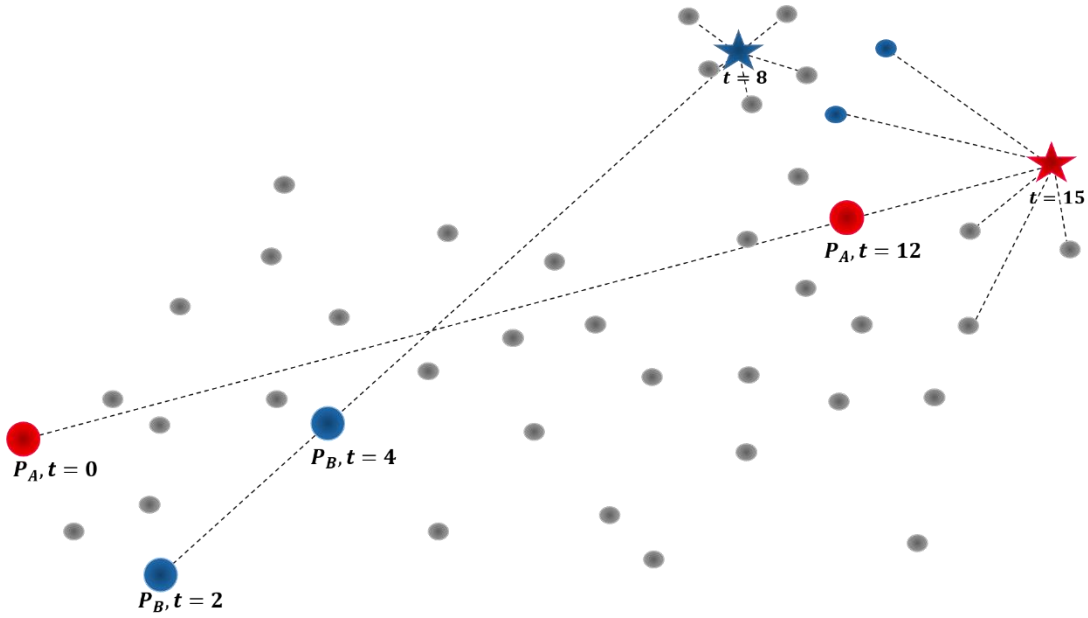

Figura 4.10: Exemplo de consultas aos vizinhos mais próximos nas estimativas referentes a dois pacientes. Os elementos retornados para P_A destacados em azul são mais próximos de P_B.

Buscando solucionar esse problema, neste trabalho de mestrado estão sendo propostas duas novas maneiras de realizar a consulta para estimar o estado dos elementos em um tempo diferente daqueles presentes na base: consultas *kAndRange* e uma aproximação do *reverse k*-NN, descritos nas próximas seções.

4.6 Refinando os resultados com *Range Query*

A primeira proposta para evitar elementos não desejados na resposta é ao utilização de operador de busca *kAndRange* (VIEIRA, JR. ET AL., 2007), ao invés de consultas aos vizinhos mais próximos. Com esse tipo de consulta, é possível delimitar a distância máxima desejada para os objetos retornados, sendo o raio dessa consulta

um indicativo da qualidade dos resultados da estimativa. Isto é, dado um determinado raio de abrangência, se não houver elementos relevantes a serem recuperados dentro desse raio, os elementos mais distantes não serão retornados, diminuindo a incidência de falsos positivos. Ao mesmo tempo, caso existam muitos elementos dentro desse raio de abrangência, apenas os *k* primeiros são retornados. Um exemplo utilizando o cenário dos pacientes P_A e P_B é ilustrado na Figura 4.11.

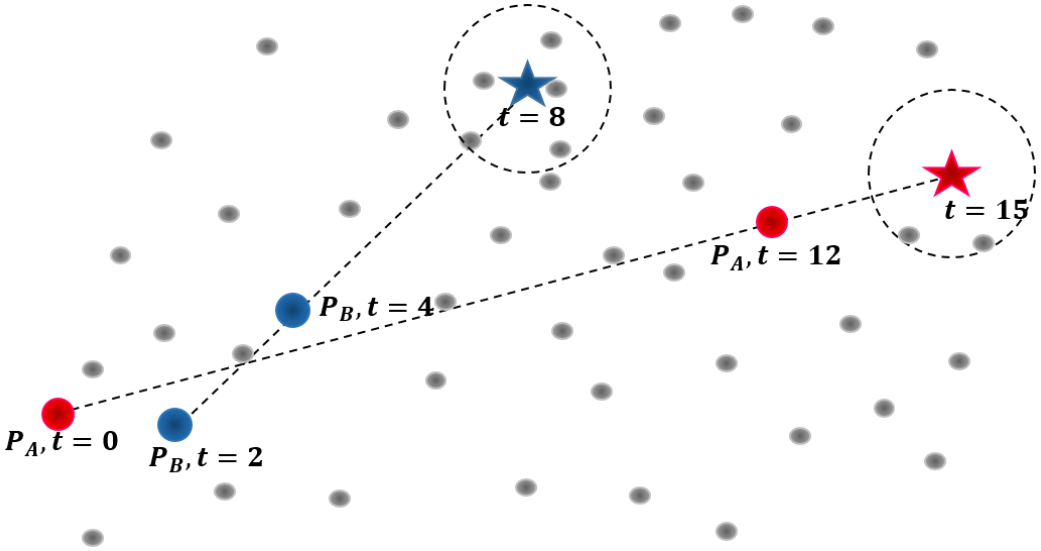

Figura 4.11: Exemplo de consulta por abrangência. Para P_B há mais elementos dentro do raio de distância estabelecido. Para P_A, apenas dois foram retornados.

4.6.1 Experimentos

Utilizando a proposta da seção anterior, a partir das estimativas de localização no espaço mapeado realizadas com os elementos da base (da mesma forma que na seção 4.4), foram realizadas consultas *kAndRange*. Portanto, a delimitação do raio de abrangência foi utilizada em conjunto com a limitação a 10-NN a fim de diminuir o número de elementos não relevantes retornados e, consequentemente, aumentar a precisão do conjunto resposta.

O raio de abrangência máximo para as consultas foi determinado através do cálculo da média dos raios de consulta 10-NN para todos os elementos do conjunto, ou seja, média das distâncias do décimo elemento retornado em um 10-NN para o elemento de consulta.

Dessa forma, obteve-se: 0,0231 como o raio de distância para o conjunto Hist256. Para o conjunto MDS10, foi encontrado o valor 0,0211. Para Fastmap10, obteve-se o raio de 0,0176.

Outras formas de calcular o raio de distância podem ser utilizadas, como o cálculo através da dimensão intrínseca do conjunto, definida pela dimensão fractal (ARANTES, VIEIRA *ET AL.*, 2003) ou considerando a distribuição homogênea (BERCHTOLD, B *ET AL.*, 1997).

Através das consultas utilizando *kAndRange* realizadas nas estimativas para cada um dos 1000 elementos, obteve-se uma média comparativa para avaliar e compreender a melhoria proporcionada pela aplicação do método proposto.

Em todas as consultas, considerou-se como relevantes os elementos retornados que pertencem à mesma classe do centro de consulta, ou seja, aqueles que correspondem à mesma categoria de imagem para a qual foi realizada a estimativa, identificada aqui pelo seu ID, conforme definido na seção 4.2.

Verificou-se que ao realizar a consulta apenas com 10-NN, dos 10000 elementos retornados no total, 72,1%, 70,7% e 65,6% eram relevantes para Hist256, MDS10 e Fastmap10. A limitação do conjunto resposta pelo raio de abrangência (*kAndRange*) permitiu diminuir o número de elementos não relevantes retornados, como é possível ver nos gráficos ilustrados na Figura 4.12.

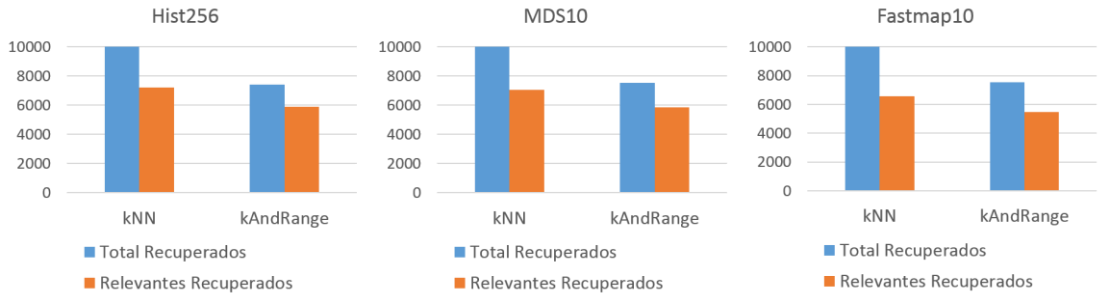

Figura 4.12: Gráficos das somatórias dos elementos retornados para as consultas *k*-NN e *kAndRange*.

Ao visualizar os gráficos, é possível perceber que, nos três casos, a diferença entre as colunas “Total Recuperados” e “Relevantes Recuperados” é mais expressiva para as consultas apenas com *k*-NN. Isso significa que nas consultas *kAndRange*, o

número de elementos retornados é mais próximo do número de relevantes, isto é, do resultado ideal.

Para o conjunto Hist256, foram podados 2570 elementos do conjunto resposta, sendo que, entre eles, 1341 eram relevantes. Para MDS10, foram podados 2456 elementos. Destes, 1230 eram relevantes. Para Fastmap10, dos 2440 podados, 1090 eram relevantes.

Em outras palavras, diminuiu-se em 25,7%, 24,5% e 24,4% para Hist256, MDS10 e Fastmap10, respectivamente, o número de elementos não relevantes retornados. Em contrapartida, aumentou-se a proporção dos relevantes recuperados em 6,9%, 6,7% e 6,7%. Portanto, a precisão das consultas aumentou de 72% para 79% em Hist256, 70,7% para 77,5% em MDS10 e 65,6% para 72,4% em Fastmap10. Para os três conjuntos, respectivamente, entre as 1000 estimativas realizadas, 568, 535 e 480 imagens retornaram em suas consultas somente imagens relevantes.

Em relação às imagens recuperadas, a limitação pelo raio na consulta permitiu podar do conjunto resposta elementos não relevantes. Observe o exemplo ilustrado na Figura 4.13, onde são mostrados resultados de consulta a 5 estimativas aleatórias, *A*, *B*, *C*, *D* e *E*, realizadas no conjunto MDS10. A primeira coluna corresponde à estimativa do elemento, utilizada como centro de consulta. Para cada um dos elementos estimados, são mostradas as imagens retornadas como os 10 vizinhos mais próximos. Ao limitar o resultado usando o raio de abrangência (*kAndRange)*, foram retornadas apenas as imagens destacadas dentro dos retângulos. Os identificadores colocados em cada imagem correspondem, respectivamente, ao ID do objeto fotografado e ao tempo, por exemplo: para a imagem 3 no tempo 10, tem-se o identificador “3 – 10”.

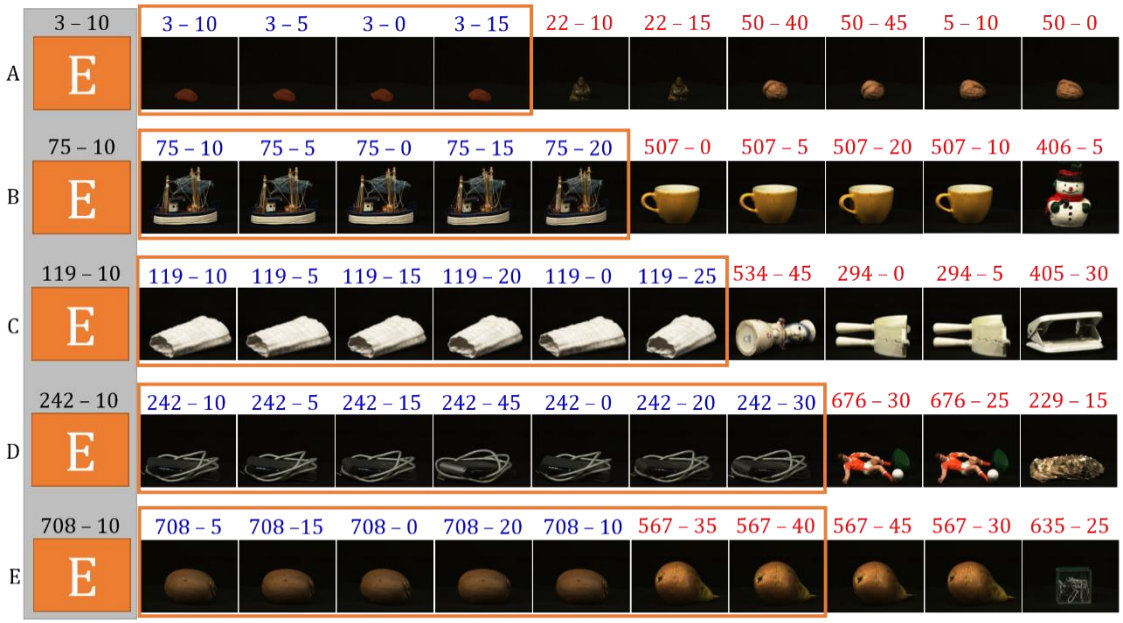

Figura 4.13: Exemplo de imagens retornadas apenas com 10-NN e podando com o raio. Somente as imagens contidas nos retângulos foram retornadas ao acrescentar o raio na consulta.

Estes são apenas alguns exemplos entre diversos outros presentes no conjunto resposta onde pode-se verificar que a delimitação do raio de distância na consulta fez com que os elementos retornados anteriormente na consulta aos 10-NN que não são relevantes não sejam recuperados. Para o elemento *A*, por exemplo, dos 10 vizinhos retornados, apenas 4 são relevantes. Ao delimitar o raio, somente esses 4 elementos foram recuperados.

Para o elemento *E*, foram recuperados outros dois elementos que não pertencem à mesma classe, considerados não relevantes. É possível observar, contudo, que se tratam de imagens visualmente semelhantes, considerando apenas a característica de cor. A partir do raio de abrangência obtido pelo cálculo da média das distâncias entre todos os elementos, os experimentos foram repetidos utilizando variações nos valores do raio. As consultas foram realizadas aumentando o valor do raio em 10%, 20%, 30%, 40%, 50% e 100%. Nas Figuras 4.14, 4.15 e 4.16 é possível verificar os gráficos dos resultados obtidos.

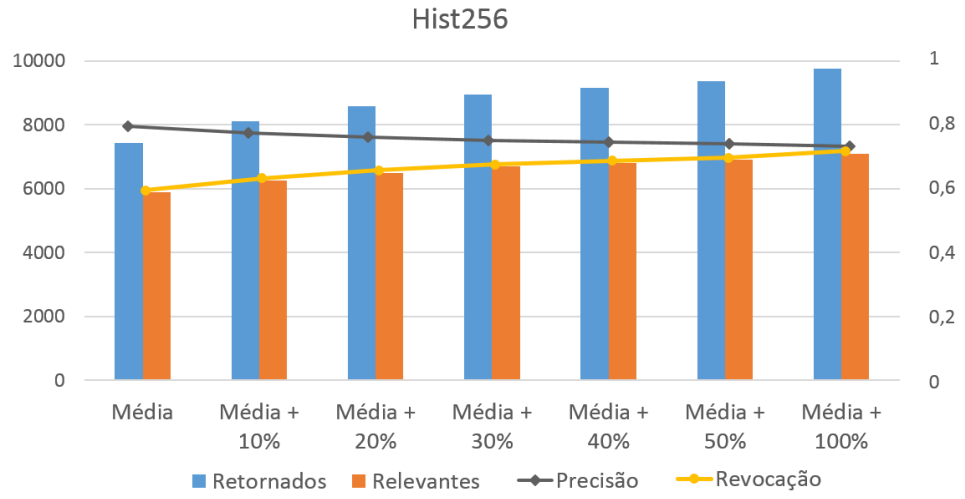

Figura 4.14: Resultados das consultas *kAndRange* realizadas variando o valor do raio de distância para o conjunto Hist256.

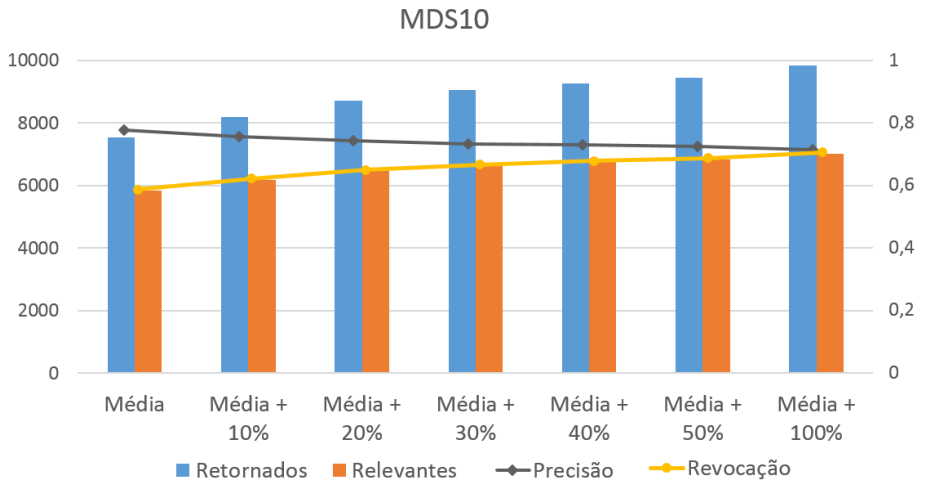

Figura 4.15: Resultados das consultas *kAndRange* realizadas variando o valor do raio de distância para o conjunto MDS10.

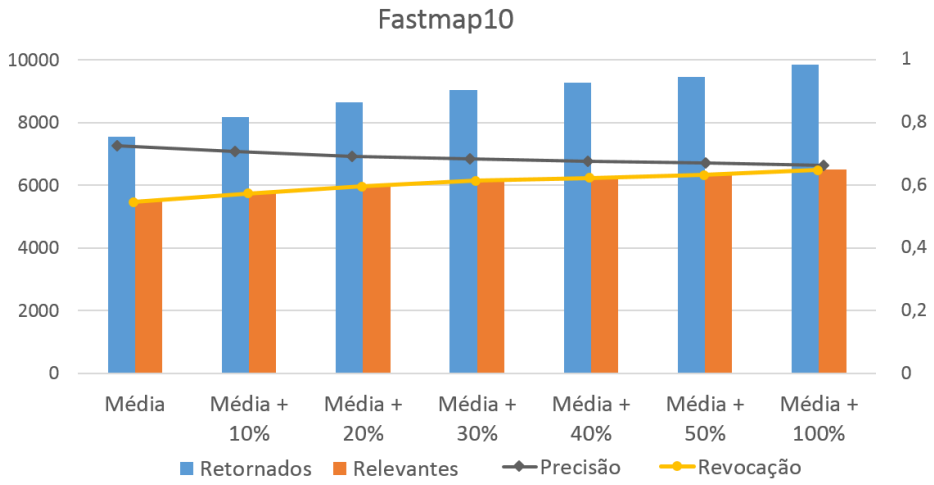

Figura 4.16: Resultados das consultas *kAndRange* realizadas variando o valor do raio de distância para o conjunto Fastmap10.

É possível verificar que, para os três conjuntos, à medida que o raio de distância aumentou, mais elementos relevantes e não relevantes foram recuperados, porém, não na mesma proporção, visto que o aumento dos não relevantes foi maior. Por conseguinte, a cada aumento no raio de distância, apesar do incremento nos níveis de revocação, a precisão dos resultados diminui, produzindo um quadro inversamente proporcional.

4.6.1.1 Experimento com elementos relevantes ausentes

Com o intuito de avaliar o refinamento das consultas utilizando o raio de abrangência e verificar a efetividade da poda dos elementos não relevantes nos resultados das consultas, nesse experimento tem-se o objetivo de verificar o comportamento do método proposto no caso onde não existem muitos elementos relevantes.

Sabe-se que para realizar a estimativa do objeto fotografado no instante de tempo desejado, utiliza-se duas instâncias temporais desse objeto existentes no conjunto. As mesmas estimativas e consultas *kAndRange* foram realizadas, porém, agora, sem utilizar as outras instâncias de tempo do objeto utilizado fotografado. Logo, para cada elemento estimado, existe na base de dados apenas as duas imagens utilizadas como referência para a estimativa.

Para Hist256, das 1000 consultas realizadas, 326 retornaram como resultado apenas os dois elementos de referência presentes na base de dados. Para MDS10, este número foi 321, e para Fastmap10, 209. Em média, foram retornados, respectivamente, 4,5, 4,5 e 6,5 elementos para cada centro de consulta.

De um total de 10000 elementos que seriam retornados numa consulta *10*-NN (consulta utilizada na abordagem anterior (BUENO, 2009)), foram retornados 4543, 4536 e 6496 para Hist256, MDS10 e Fastmap10. Desses totais, 1645, 1639 e 1663 são relevantes, equivalendo a precisões de 36,2%, 36,1% e 25,6%, sendo que na consulta *10*-NN, a precisão máxima seria 20%. Para os três conjuntos, respectivamente, 518, 493 e 294 imagens tiveram em suas respostas às consultas somente imagens relevantes.

Outros, entretanto, retornaram além dos dois elementos de referência existentes no conjunto. Isso pode ser explicado pelo fato de as imagens serem parecidas considerando apenas as características de cor, sendo inclusive um dos objetivos do trabalho. Em uma consulta onde não há mais imagens da mesma classe a serem recuperadas, uma vez que se deseja ter uma estimativa de como poderá estar esse objeto em determinado momento, imagens de outras classes visualmente similares podem ser recuperadas, como é mostrado em alguns exemplos na Figura 4.17.

Figura 4.17: Exemplos de elementos que retornaram elementos de outras classes, porém com características semelhantes.

Nestes exemplos, todas as imagens mostradas para cada estimativa foram retornadas, pois tratam-se de imagens que reproduziram histogramas similares. Porém, apenas as imagens com os identificadores destacados correspondem às duas imagens presentes no conjunto.

Para a imagem *C*, por exemplo, foram retornadas apenas as duas imagens da mesma classe utilizadas para a estimativa. As demais imagens, pertencentes a outra classe, são visualmente similares. Porém, em nossos experimentos, a precisão dessa consulta é computada como 22%.

4.7 Refinando os resultados com *k*-NN Reverso

Com o intuito de aprimorar os resultados das estimativas, foi proposta a utilização de uma aproximação do *Reverse k*-NN (KORN E MUTHUKRISHNAN, 2000), sendo que neste, são recuperados todos os que tem o elementos de consulta como um dos vizinhos mais próximos, considerando todo o conjunto. Essa aproximação ocorre da seguinte maneira: aplicando a consulta k_1-NN em um elemento estimado, aplica-se o k_2-NN reverso neste conjunto resposta. Se um dos k_1-NN retornados não possuir o elemento de consulta (elemento estimado) como um dos k_2 vizinhos mais próximos, provavelmente esse não seja um elemento relevante, e é descartado. Caso contrário, pode significar que eles são próximos o suficiente para serem pertencentes à mesma classe.

Trata-se de uma aproximação da consulta k-NN reverso, pois no operador original, todos os elementos do conjunto são verificados a fim de se encontrar quais tem o elemento de consulta como um dos *k* elementos mais próximos. Neste trabalho, essa verificação é realizada somente para os elementos retornados em k_1-NN. Neste trabalho, essa aproximação será chamada de *k_1AndRk_2-NN*.

Observe exemplo ilustrado na Figura 4.18, utilizando o mesmo exemplo dos pacientes P_A e P_B, onde 3 dos 5-NN retornados para P_A são muito mais distantes do elemento de consulta em comparação a outros elementos presentes na base.

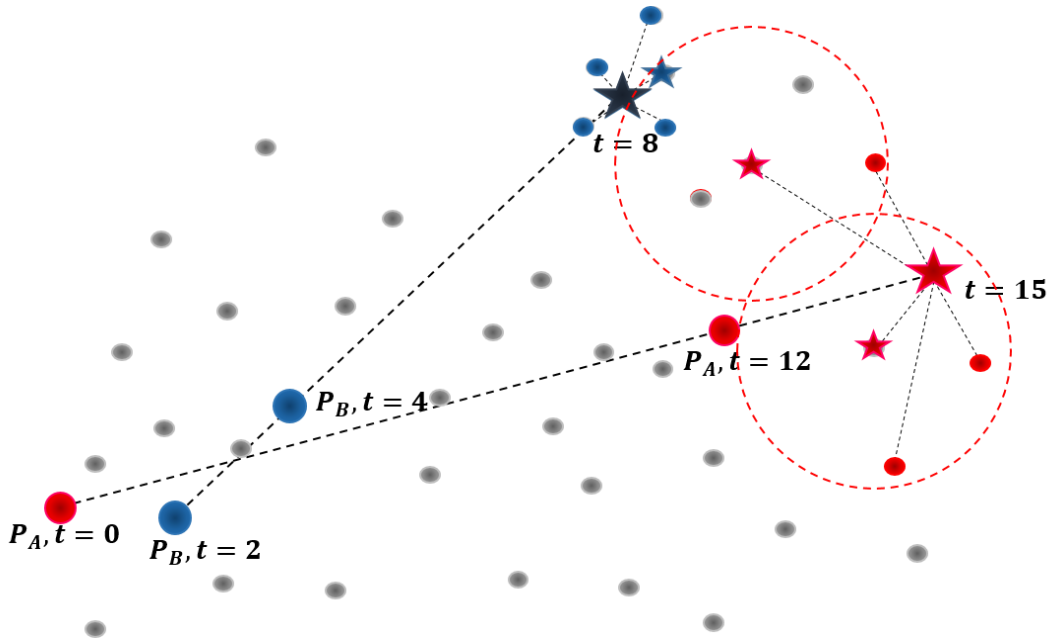

Figura 4.18: Exemplo de sequência de consultas *k*-NN e *k*-NN reverso (*k_1AndRk_2-NN)*. Alguns elementos retornados como vizinho mais próximo de P_A são mais próximos de outros elementos.

4.7.1 Experimentos

Neste experimento foram realizadas consultas *10AndR10-NN*. Verificou-se que, para Hist256, MDS10 e Fastmap10, respectivamente, 705, 607 e 539 elementos obtiveram como resultado somente imagens relevantes. A utilização da consulta *k_1AndRk_2-NN* permitiu diminuir o número de elementos não relevantes retornados e aumentar a proporção dos relevantes, como é possível ver nos gráficos ilustrados na Figura 4.21.

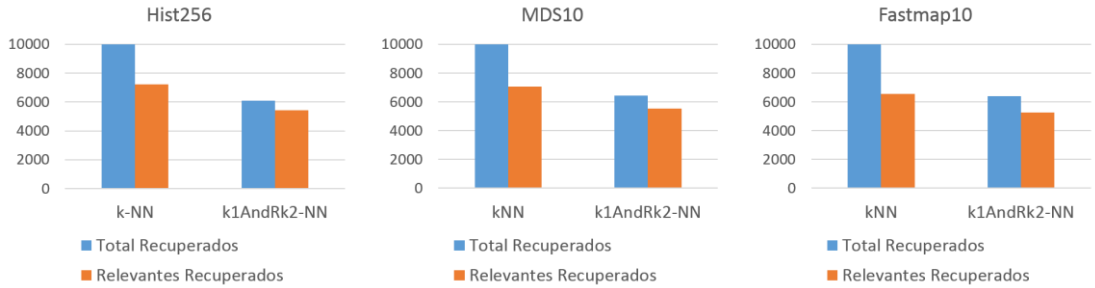

Figura 4.19: Gráficos das somatórias dos elementos retornados para as consultas *k*-NN e *k_1AndRk_2-NN*.

É possível visualizar através dos gráficos que, assim como nos experimentos realizados com *kAndRange*, diminui-se consideravelmente a diferença existente entre as colunas “Total Recuperados” e “Relevantes Recuperados” obtidas para as consultas apenas com *k*-NN.

Para o conjunto Hist256, em comparação com a consulta aos 10 vizinhos mais próximos, foram podados 3917 elementos do conjunto resposta, sendo que 1791 relevantes foram podados. Para MDS10, foram podados 3579 elementos. Entre os relevantes, 1704. Para Fastmap10, dos 3593 podados, 1972 eram relevantes.

Esses números mostram a diminuição da quantidade de elementos não relevantes retornados. Porém, aumentou-se a proporção dos relevantes recuperados em 17%, 15% e 16% para Hist256, MDS10 e Fastmap10. Portanto, a precisão das consultas aumentou de 72% para 89,2% em Hist256, 70,7% para 85,9% em MDS10 e 65,6% para 81,9% em Fastmap10.

Essa melhoria nas precisões das consultas ocorreu inclusive com relação às consultas *kAndRange*, como pode ser visualizado no gráfico comparativo entre *k*-NN, *kAndRange* e k_1AndRk_2-*NN*, ilustrado na Figura 4.20.

Na Figura 4.21 são mostrados alguns resultados de consultas com o método proposto. As imagens retornadas pelo método proposto são mostradas destacadas pelo retângulo. Também são mostradas as imagens que seriam retornadas pela consulta 10-NN. Para o elemento *B*, por exemplo, dos 10 vizinhos retornados pela consulta 10-NN, apenas 3 são relevantes. Ao aplicar o *k*-NN reverso, somente esses 3 elementos foram recuperados. Neste exemplo foi utilizado o conjunto MDS10.

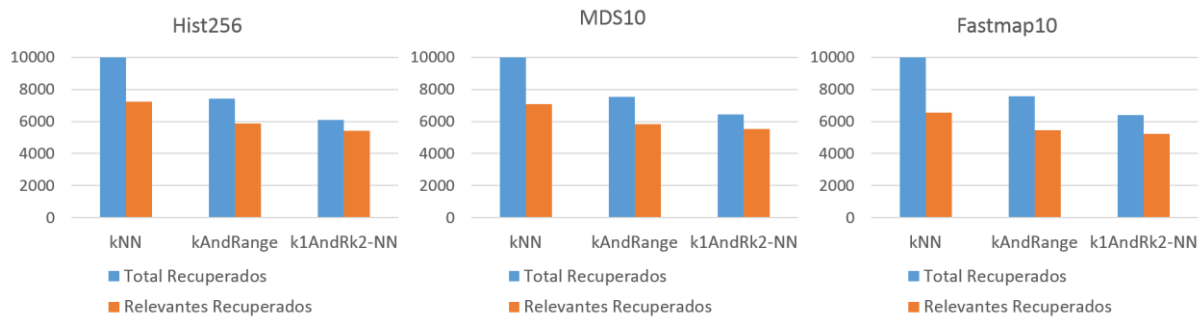

Figura 4.20: Gráfico comparativo entre os resultados das consultas *k*-NN, *kAndRange* e k_1AndRk_2-*NN* para os conjuntos Hist256, MDS10 e Fastmap10.

Para a estimativa *E* do exemplo, pode-se visualizar que as imagens retornadas não correspondem à mesma classe (a classe do objeto estimado é 764 e são retornadas várias imagens do objeto 775). Contudo, o objeto fotografado em ambas as classes é muito parecido (nas anotações que descrevem os objetos fotografados[1], ambos os objetos são descritos como "*yellow square toy*"). Este é mais um exemplo de que imagens de diferentes classes, porém semelhantes, também podem representar boa qualidade nas respostas.

Figura 4.21: Exemplo de imagens retornadas apenas com 10-NN e com *k₁AndRk₂-NN*. As imagens dentro dos retângulos são as únicas retornadas ao utilizar *k₁AndRk₂-NN*.

4.7.1.1 Experimento com elementos relevantes ausentes

Do mesmo modo como realizado com as consultas *kAndRange*, considerou-se como presente entre os elementos da mesma classe do elemento a ser estimado somente aqueles que foram utilizados como referência para a estimativa.

Dos 1000 elementos utilizados como centro de consulta, 318 retornaram apenas relevantes para Hist256, dos quais 239 recuperaram os dois elementos

[1] Disponível em http://aloi.science.uva.nl/, Object Annotation.

utilizados como referência. Para MDS10, esses números foram 237 e 211. Para Fastmap10, 205 e 167.

A proporção entre total de elementos recuperados e relevantes recuperados representou uma precisão média de 38,1% para Hist256, 33,7% para MDS10 e 31% para Fastmap10.

Neste experimento, houve também os elementos que retornaram além dos dois esperados, representando a mesma situação encontrada no experimento descrito na seção 4.6.1.1, em que as imagens são visualmente semelhantes, tendo, por conseguinte, características próximas. Na Figura 4.22 é possível visualizar alguns exemplos.

Figura 4.22: Exemplos de elementos que retornaram elementos de outras classes, porém similares.

Para a imagem *B*, por exemplo, foram retornadas as duas imagens da mesma classe existentes no conjunto. As demais imagens, pertencentes a outras duas classes, são visualmente similares.

4.7.1.2 Experimentos com variações nos valores de k_1 e k_2

O objetivo desse experimento foi verificar a influência dos valores de k_1 e k_2 nas consultas k_1AndRk_2-*NN*. Na primeira verificação, alterou-se o *k2* (valor utilizado na verificação da aproximação do *k*-NN reverso). As consultas foram realizadas aos

10AndRk$_2$-NN, onde k_2 corresponde a 5, 10 (valor considerado padrão, utilizado no experimentos anteriores), 20, 30, 40, 50. Nos gráficos exibidos nas Figuras 4.23 (Hist256), 4.24 (MDS10) e 4.25 (Fastmap10) é possível verificar visualmente as diferenças na precisão e revocação entre os resultados obtidos.

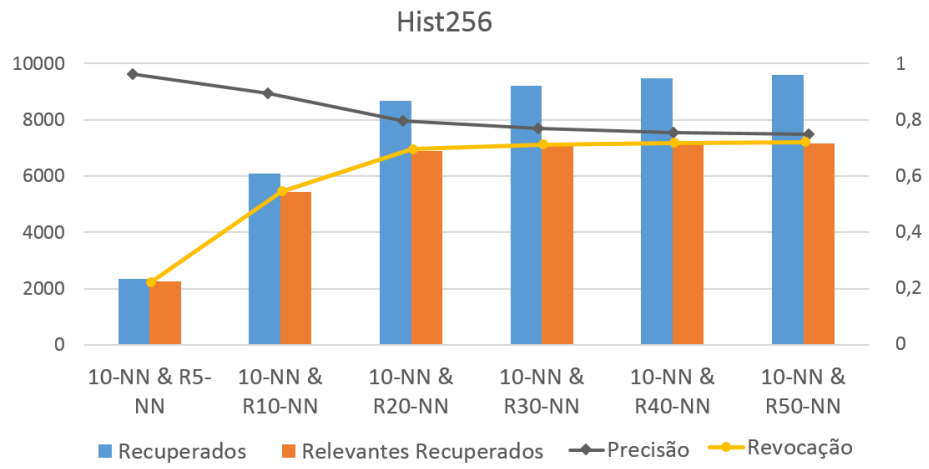

Figura 4.23: Resultados das consultas *10AndRk-NN* variando o valor de k_2 no conjunto Hist256.

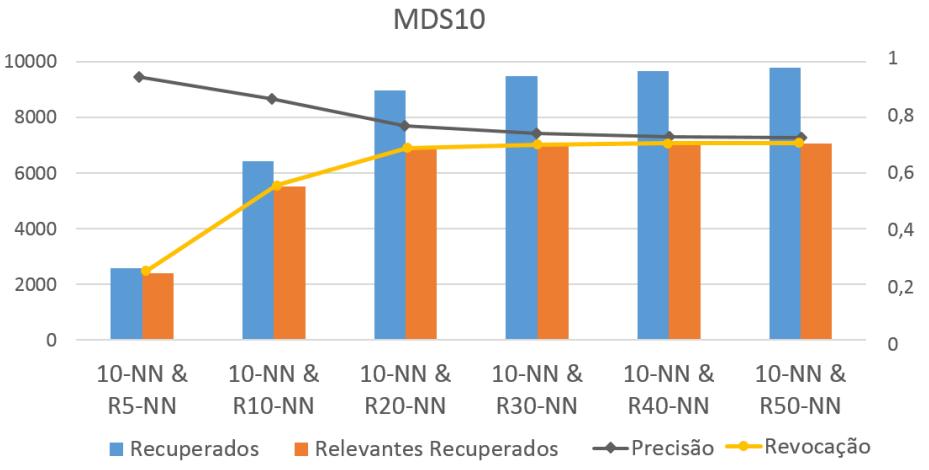

Figura 4.24: Resultados das consultas *10AndRk-NN* variando o valor de k_2 no conjunto MDS10.

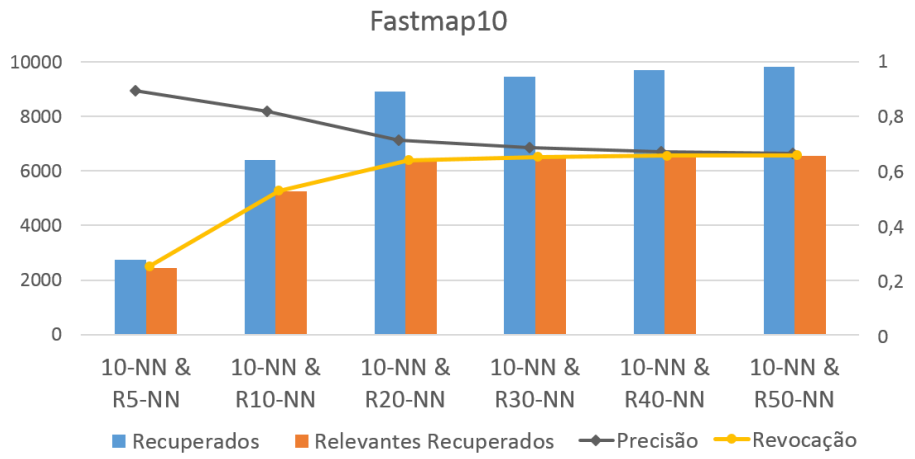

Figura 4.25: Resultados das consultas *10AndRk-NN* variando o valor de k_2 no conjunto Fastmap10.

Os resultados se mostraram similares para os três conjuntos, com variações nos níveis de precisão, que ficaram mais baixas para Fastmap10. É possível perceber que, naturalmente, à medida que k_2 aumenta, o número de elementos recuperados também aumenta. Porém, não o suficiente para aumentar significativamente os níveis de revocação e a precisão, visto que o número de elementos relevantes retornados pouco se alterou. Portanto, o total de relevantes recuperados (revocação) aumenta enquanto a precisão diminui.

Com k_2 =5, o número de elementos retornados caiu consideravelmente, onde, de 10000 elementos, foram recuperados no total 2339 para Hist256, 2587 para MDS10 e 2732 para Fastmap10. Todavia, garantiu-se que quase todos os elementos retornados são relevantes. As precisões alcançadas foram 96,1% para Hist256, 93,4% para MDS10 e 89,5% para Fastmap10, onde, respectivamente, 930, 883 e 809 de 1000 elementos retornaram apenas relevantes. Contudo, a revocação foi bastante reduzida, com totais de 2248, 2417 e 2446 no total de relevantes recuperados para Hist256, MDS10 e Fastmap10, respectivamente.

Na segunda verificação, alterou-se o valor para o k_1 (quantidade de vizinhos que serão posteriormente avaliados pelo k_2-NN reverso). As consultas foram realizadas aos *k_1AndR10-NN*, onde k_1 corresponde a 5, 10 (valor considerado padrão, utilizado nos experimentos anteriores), 20, 30, 40 e 50. Os resultados são exibidos nas Figuras 4.26 (Hist256), 4.27 (MDS10) e 4.28 (Fastmap10).

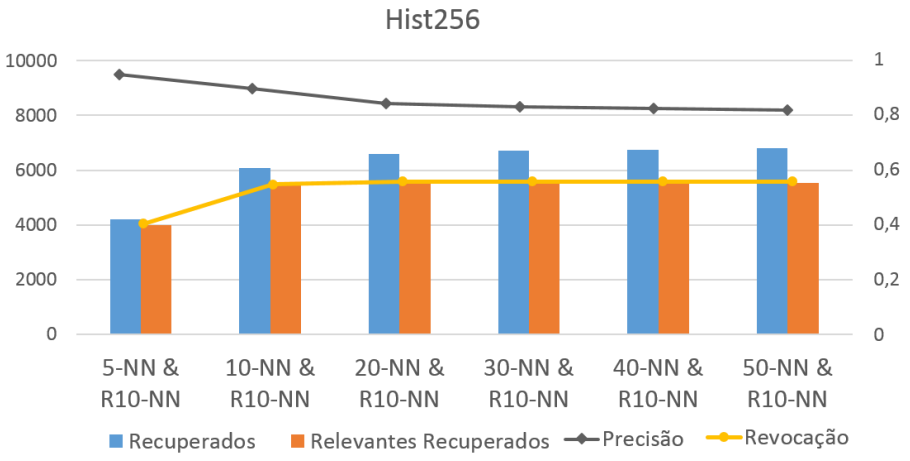

Figura 4.26: Resultados das consultas *k_1AndR10-NN* variando o valor de *k_1* no conjunto Hist256.

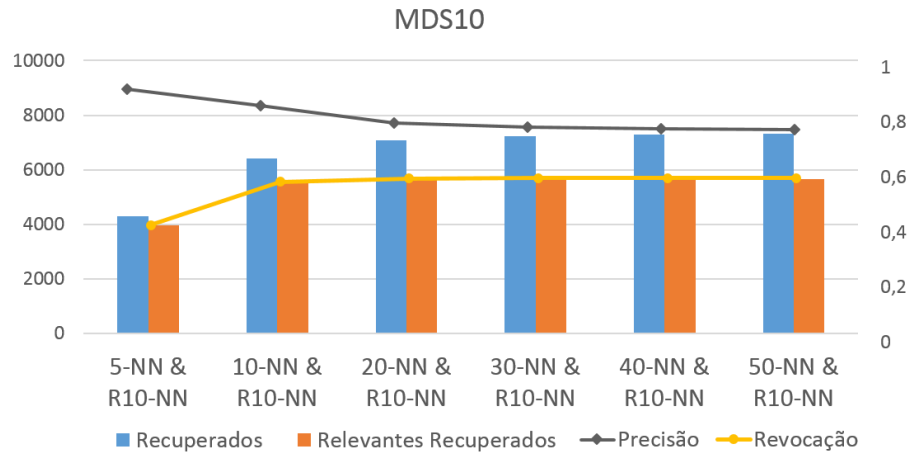

Figura 4.27: Resultados das consultas *k_1AndR10-NN* variando o valor de *k_1* no conjunto MDS10.

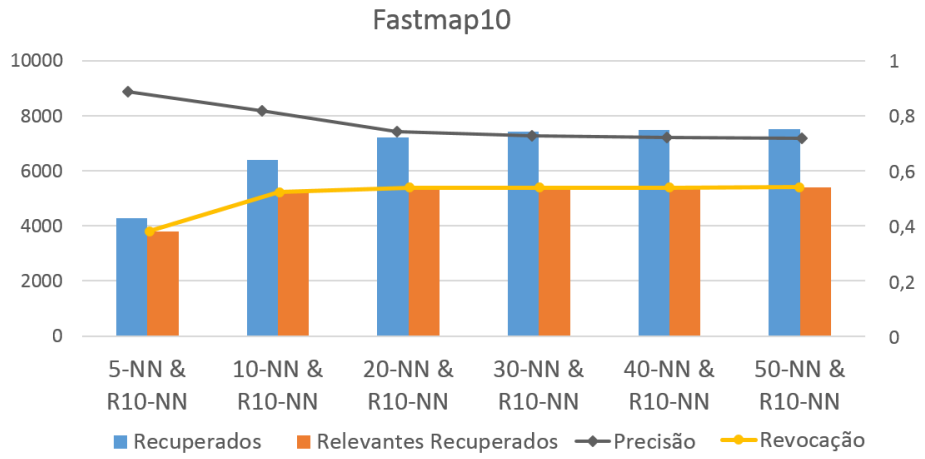

Figura 4.28: Resultados das consultas *k_1AndR10-NN* variando o valor de k_1 no conjunto Fastmap10.

É possível verificar que, para este segundo caso, o aumento no número de elementos recuperados ocorreu de forma mais contida, refletindo nos níveis de precisão das consultas para k_1 variando de 20 a 50. Novamente, à medida que o número de elementos relevantes retornados (revocação) aumentou, a precisão diminuiu.

Destaca-se, novamente, a primeira coluna, onde k_1 foi reduzido para 5. De 5000 elementos que poderiam ser retornados nas consultas, foram recuperados 4212 para Hist256, 4284 para MDS10 e 4272 para Fastmap10, com precisões de 94,4%, 92,2% e 89% para Hist256, MDS10 e Fastmap10. Das 1000 consultas realizadas, em 864, 813 e 748 (considerando respectivamente os conjuntos Hist256, MDS10 e Fastmap10) retornaram apenas imagens relevantes. Porém, os valores para revocação também foram mais baixos, com 3976 relevantes retornados para Hist256, 3949 para MDS10 e 3803 para Fastmap10.

Dessa forma, para as duas variações, entende-se que para priorizar os níveis de precisão, a redução dos valores de k_1 *e* k_2 para 5 atende à expectativa, trazendo níveis de precisão próximos de 100%. Para priorizar, porém, a revocação (quantidade de relevantes recuperados com relação ao total de elementos do conjunto), quanto maiores os valores de k_1 e k_2, maior o nível de revocação alcançado. Para o equilíbrio entre essas duas medidas, entende-se como ideal o próprio *10AndR10-NN* para os dados utilizados nos experimentos.

4.8 Considerações Finais

Neste capítulo foram apresentados os resultados obtidos a partir do estudo proposto. Foi possível verificar que os algoritmos MDS e *Fastmap* apresentaram o mesmo comportamento nos resultados das estimativas, com pequenas diferenças nos níveis de precisão, tendo o MDS um pouco mais de ganho com relação ao *Fastmap*.

Com as consultas *kAndRange* e k_1*AndRk*$_2$*-NN* utilizadas, foi possível aumentar a qualidade das estimativas, diminuindo do conjunto resposta o número de elementos não relevantes recuperados, mantendo na sua maioria apenas os elementos realmente correspondentes ao objeto de consulta.

Nas tabelas a seguir, os valores comparativos referentes às estimativas realizadas com *k*-NN, *kAndRange* e k_1*AndRk*$_2$*-NN* (com o valor 10 para k_1 e k_2) são apresentados, considerando, novamente, o tempo definido como Futuro 4. Vale ressaltar que em todos os experimentos foram realizadas 1000 consultas, correspondente aos 1000 objetos fotografados pertencentes ao conjunto de dados utilizado.

Tabela 4.1: Estimativas com *k*-NN

Estimativas com *k*-NN					
	Total Recuperados	**Relevantes Recuperados**	**Precisão Média**	**Revocação Média**	**Consultas que retornaram somente relevantes**
Hist256	10000	7219	72,19%	72,19%	255
MDS10	10000	7075	70,75%	70,75%	236
Fastmap10	10000	6561	65,61%	65,61%	182

Tabela 4.2: Estimativas com *kAndRange*

Estimativas com *kAndRange*					
	Total Recuperados	**Relevantes Recuperados**	**Precisão Média**	**Revocação Média**	**Consultas que retornaram somente relevantes**
Hist256	7430	5878	79,11%	58,78%	568
MDS10	7544	5845	77,48%	58,45%	535
Fastmap10	7560	5471	72,37%	54,71%	480

Tabela 4.3: Estimativas com *kAndRange*, sem elementos relevantes

Estimativas com *kAndRange*, sem elementos relevantes					
	Total Recuperados	Relevantes Recuperados	Precisão	Revocação	Consultas que retornaram somente relevantes
Hist256	4543	1645	36,21%	82,25%	518
MDS10	4536	1639	36,13%	81,95%	493
Fastmap10	6496	1663	25,60%	83,15%	294

Tabela 4.4: Estimativas com *10AndR10-NN*

Estimativas com *k1AndRk2-NN*					
	Total Recuperados	Relevantes Recuperados	Precisão	Revocação	Consultas que retornaram somente relevantes
Hist256	6083	5428	89,23%	54,28%	705
MDS10	6421	5515	85,89%	55,15%	607
Fastmap10	6407	5247	81,89%	52,47%	539

Tabela 4.5: Estimativas com *10AndR10-NN*, sem elementos relevantes

Estimativas com *k1AndRk2-NN*, sem elementos relevantes					
	Total Recuperados	Relevantes Recuperados	Precisão	Revocação	Consultas que retornaram somente relevantes
Hist256	4611	1758	38,13%	87,90%	318
MDS10	5184	1746	33,68%	87,30%	237
Fastmap10	5439	1682	30,92%	84,10%	205

Com relação às estimativas utilizando apenas *k*-NN (Tabela 4.1), ambas as consultas propostas (*kAndRange* e k_1AndRk_2-*NN*) obtiveram níveis de precisão mais altos, embora a revocação foi menor. Destaca-se ainda o número de elementos, entre os 1000 para os quais foram realizadas as estimativas, que retornaram apenas relevantes em suas consultas, que aumentou consideravelmente.

Em se tratando de precisão, as estimativas com k_1AndRk_2-*NN* apresentaram os melhores resultados, com 89,23%, 85,89% e 81,89% para Hist256, MDS10 e Fastmap10. Além disso, o número de consultas que retornaram apenas elementos relevantes alcançou os maiores números, sendo 705, 607 e 539, respectivamente, para os três conjuntos.

Ao realizar as consultas sem os demais elementos da mesma classe no conjunto, foi possível perceber outras características nos experimentos, onde se tem um cenário oposto: não há outras imagens da mesma classe. Os resultados, de acordo com o esperado, demonstraram as hipóteses de retornar somente os dois

elementos de referência para a estimativa, por serem da mesma classe, e/ou retornar outros elementos pela sua semelhança.

Entende-se, portanto, que os estudos realizados podem ser aprofundados e avaliados em outras bases de dados, inclusive reais, possibilitando uma maneira eficiente de gerenciar dados em espaços métricos com a informação temporal.

Capítulo 5

CONCLUSÃO

5.1 Considerações Finais

Este trabalho teve como principal objetivo estudar maneiras que possibilitem e viabilizem a realização de análises aos dados em espaços métricos considerando uma característica fundamental: a informação temporal.

Conforme abordado ao longo do texto, para muitas áreas de conhecimento, a informação temporal é mandatória para que se possa analisar os dados por completo e, consequentemente, ter um gerenciamento mais preciso dos mesmos.

A possibilidade de estimar o estado de um dado em espaço métrico requer um processo além das consultas por similaridade. Para analisar e gerenciar a informação temporal nesse domínio de dados (onde as únicas informações disponíveis são os próprios elementos, representados pelos seus vetores de características, e as distâncias entre eles), uma alternativa é mapeá-los para um espaço multidimensional.

Dessa forma, de uma consulta *"Quais são as imagens mais similares à imagem A?"*, as possibilidades se ampliam, por exemplo, para *" Quais são as imagens mais similares à imagem A quando esta estiver no tempo X?"*.

Através do estudo das opções de mapeamento realizado neste trabalho, foi possível verificar a qualidade do mapeamento realizado pelos algoritmos utilizados, em que, quanto maior a qualidade dos mesmos, maior é a qualidade das estimativas. Além disso, através da escolha de um conjunto de dados representado em espaço multidimensional (Histogramas de 256 níveis de cinza), foi possível realizar estimativas utilizando os próprios vetores de características (antes do mapeamento).

Por conseguinte, mapear esse conjunto de dados permitiu verificar o impacto do mapeamento sobre as estimativas, através da comparação das respostas das consultas, mostrando que a qualidade das mesmas pode não ser prejudicada pelo mapeamento, especialmente com MDS, para o qual obteve-se resultados similares aos obtidos realizando as estimas diretamente no espaço original.

Com a utilização dos métodos propostos – *kAndRange* e k_1AndRk_2-*NN* – os resultados das consultas tornam-se mais assertivos com relação às consultas *k*-NN convencionais aplicadas na estimativa de posicionamento no espaço dimensional, pois na consulta *k*-NN, independentemente das distâncias, *k* elementos são recuperados, permitindo que se tenha no conjunto resposta elementos distantes do centro de consulta. Dessa forma, as consultas *kAndRange* e k_1AndRk_2-*NN* permitiram diminuir o número de elementos não relevantes retornados e aumentar a precisão das consultas.

Ao limitar a proximidade dos elementos recuperados através de um raio de abrangência (*kAndRange*), garante-se, no momento da consulta, que qualquer elemento que estiver fora desse raio, ou seja, além da distância esperada, não seja retornado.

A verificação utilizando *k*-NN reverso no conjunto *k*-NN (k_1AndRk_2-*NN*), por sua vez, ao descartar aqueles elementos que não possuem o elemento de consulta como um de seus vizinhos mais próximos, poda da resposta aqueles que provavelmente estão mais próximos a elementos de outra classe de dados.

A utilização dos métodos de busca propostos pode, portanto, ser utilizada como um filtro, evitando que elementos não desejados sejam retornados como resultado, melhorando também a qualidade das estimativas feitas para tempos não disponíveis na base de dados.

5.2 Principais Contribuições

As principais contribuições desse trabalho de mestrado foram:

- Estudo da influência do mapeamento nas estimativas: através da utilização do algoritmo MDS, que apresentou melhores resultados para o mapeamento e, portanto, para as estimativas, foi possível evidenciar que

outros algoritmos, que proporcionem bons resultados no mapeamento, podem ser utilizados a fim de aprimorar as estimativas. Evidencia-se, portanto, que conforme verificado por Bueno (2009), a qualidade do mapeamento pode interferir diretamente na qualidade das estimativas.

- Proposta da realização de Estimativas usando *kAndRange* e k_1AndRk_2-*NN*: com as variações dos tipos de consultas aplicadas sobre a localização estimada, acrescentando o raio de abrangência e a busca pelos vizinhos mais próximos reversos, pode-se limitar os resultados para que sejam mais próximos do que se espera: tornar o conjunto resposta mais seleto, diminuindo as possibilidades de retorno dos elementos não desejados (considerados não relevantes), a depender da análise a ser realizada.

5.3 Trabalhos Futuros

Dentre as possibilidades de trabalhos futuros que podem ser exploradas para viabilizar a evolução deste estudo, uma delas é a utilização de mais elementos de referência para a realização das estimativas. Neste trabalho foram utilizados dois elementos e a estimativa realizada através de interpolação/extrapolação linear. Ao utilizar três ou mais elementos para as estimativas, os resultados apresentados tendem a ser melhores.

Outras maneiras de ampliar o estudo proposto incluem a utilização de outros descritores para representar as características das imagens. É possível, inclusive, ser aplicado para combinação de múltiplos descritores (BUENO, KASTER *ET AL.*, 2009a), que representa os dados em espaço métrico, e espaços métrico-temporais (BUENO, KASTER *ET AL.*, 2009b).

REFERÊNCIAS

AKSOY, S.; HARALICK, R. M. **Feature normalization and likelihood-based similarity measures for image retrieval**. Pattern Recogn. Lett.**,** v. 22, n. 5, p. 563-582, 2001. ISSN 0167-8655.

ARANTES, A. S. **Consultas por Similaridade Complexas em Gerenciadores Relacionais**. 2005. (Doutorado). Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

ARANTES, A. S. et al. **The Fractal Dimension Making Similarity Queries More Efficient**. In: PRESS, A., Second Workshop on Fractals, Power Laws and Other Next Generation Data Mining Tools (10th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining), 2003. Washington-DC.

AZEVEDO-MARQUES, P. M. et al. **The Metric Histogram: A New and Efficient Approach for Content-based Image Retrieval**. Proceedings of the IFIP TC2/WG2.6 Sixth Working Conference on Visual Database Systems: Visual and Multimedia Information Management: Kluwer, B.V.: p. 297-311, 2002.

BERCHTOLD, S. et al. **A cost model for nearest neighbor search in high-dimensional data space**. Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. Tucson, Arizona, USA: ACM: p. 78-86, 1997.

BOSCH, A.; ZISSERMAN, A.; MUNOZ, X. **Representing shape with a spatial pyramid kernel**. Proceedings of the 6th ACM international conference on Image and video retrieval. Amsterdam, The Netherlands: ACM: p. 401-408, 2007.

BUENO, R. **Tratamento do tempo e dinamicidade em dados representados em espaços métricos**. 2009. Tese (Doutorado em Ciência da Computação) Instituto de Ciências Matemáticas e de Computação, USP, São Carlos.

BUENO, R. et al. **Unsupervised scaling of multi-descriptor similarity functions for medical image datasets**. 22nd IEEE International Symposium on Computer-Based Medical Systems, 2009a. p.1-8.

BUENO, R. et al. **Time-Aware Similarity Search: A Metric-Temporal Representation for Complex Data**. Proceedings of the 11th International Symposium on Advances in Spatial and Temporal Databases. Aalborg, Denmark: Springer-Verlag: p. 302-319, 2009b.

CARÉLO, C. C. M. et al. **The Onion-Tree: Quick Indexing of Complex Data in the Main Memory**. Advances in Databases and Information Systems**,** v. 5739, p. 235-252, 2009. ISSN 0302-9743.

CHÁVEZ, E. et al. **Searching in metric spaces**. ACM Comput. Surv.**,** v. 33, n. 3, p. 273-321, 2001. ISSN 0360-0300.

CHINO, F. J. T. **Visualizando a organização e o comportamento de estruturas métricas: Aplicações em consultas por similaridade**. 2004. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação, USP, São Carlos.

CIACCIA, P.; PATELLA, M.; ZEZULA, P. **M-tree: An Efficient Access Method for Similarity Search in Metric Spaces**. Proceedings of the 23rd International Conference on Very Large Data Bases: Morgan Kaufmann Publishers Inc.: p. 426-435, 1997.

COX, T. F.; COX, M. A. A. **Multidimensional Scaling**. Second Edition. Chapman & Hall/CRC, 2000.

EDELWEISS, N. **Bancos de Dados Temporais: Teoria e Prática**. XVII Jornada de Atualização em Informática: Anais do XVIII Congresso Nacional da Sociedade Brasileira de Computação "Rumo à Sociedade do Conhecimento": Sociedade Brasileira de Computação: p. 225-282, 1998.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de Banco de Dados**. 6ª ed. São Paulo: Pearson, 2011.

FALOUTSOS, C. **Searching Multimedia Databases by Content**. Kluwer Academic Publishers, 1996. ISBN 0792397770.

FALOUTSOS, C.; LIN, K.-I. **FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets**. SIGMOD Rec.**,** v. 24, n. 2, p. 163-174, 1995. ISSN 0163-5808.

GEUSEBROEK, J.-M.; BURGHOUTS, G. J.; SMEULDERS, A. W. M. **The Amsterdam Library of Object Images**. Int. J. Comput. Vision**,** v. 61, n. 1, p. 103-112, 2005. ISSN 0920-5691.

GONZALEZ, R. C.; WOODS, R. E. **Processamento Digital De Imagens**. 3ª Ed. Pearson Education, 2011.

HJALTASON, G. R.; SAMET, H. **Contractive Embedding Methods for Similarity Searching in Metric Spaces**. Computer Science TR-4102., University of Maryland, College Park, Maryland. 2000.

HRISTESCU, G.; FARACH-COLTON, M. **Cluster-preserving Embedding of Proteins**. Center for Discrete Mathematics; Theoretical Computer Science. 1999

HUANG, Z. et al. **Effective data co-reduction for multimedia similarity search**. Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. Athens, Greece: ACM: p. 1021-1032, 2011.

KHOTANZAD, A.; HONG, Y. H. **Invariant Image Recognition by Zernike Moments**. IEEE Trans. Pattern Anal. Mach. Intell.**,** v. 12, n. 5, p. 489-497, 1990. ISSN 0162-8828.

KORN, F.; MUTHUKRISHNAN, S. **Influence sets based on reverse nearest neighbor queries**. SIGMOD Rec.**,** v. 29, n. 2, p. 201-212, 2000. ISSN 0163-5808.

KRIG, S. **Computer Vision Metrics: Survey, Taxonomy, and Analysis**. Apress, 2014.

KURASAWA, H.; TAKASU, A.; ADACHI, J. **Finding the k-closest pairs in metric spaces**. Proceedings of the 1st Workshop on New Trends in Similarity Search. Uppsala, Sweden: ACM: p. 8-13, 2011.

LEVENSHTEIN, V. **Binary codes capable of correcting delections, insertions, and reversals**. In: (Ed.). Cybernetics and Control Theory, 1996. p.707-710.

OZSOYOGLU, G.; SNODGRASS, R. T. **Temporal and Real-Time Databases: A Survey**. IEEE Trans. on Knowl. and Data Eng.**,** v. 7, n. 4, p. 513-532, 1995. ISSN 1041-4347.

PAULOVICH, F. V. **Mapeamento de dados multi-dimensionais - inegrando mineração e visualização**. 2008. Tese de Doutorado. Instituto de Ciências Matemáticas e Computacionais, USP, São Carlos.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Sistema de Banco de Dados**. 5ª edição. Rio de Janeiro: Elsevier, 2006.

SILVA, V. D.; TENENBAUM, J. B. **Global versus local methods in nonlinear dimensionality reduction**. In: (Ed.). Advances in Neural Information Processing Systems 15. Cambridge, MA: MIT Press, 2003. p.705-712.

SKOPAL, T.; HOKSZA, D. **Improving the performance of M-tree family by nearest-neighbor graphs**. Proceedings of the 11th East European conference on Advances in databases and information systems. Varna, Bulgaria: Springer-Verlag: p. 172-188, 2007.

TANSEL, A. U. et al. **Temporal Databases: theory, design and implementation**. Redwood City, California, USA: 1993.

TAO, Y.; PAPADIAS, D.; LIAN, X. **Reverse kNN search in arbitrary dimensionality**. Proceedings of the Thirtieth international conference on Very large data bases - Volume 30. Toronto, Canada: VLDB Endowment: p. 744-755, 2004.

TAO, Y. et al. **Multidimensional reverse kNN search**. The VLDB Journal**,** v. 16, n. 3, p. 293-316, 2007. ISSN 1066-8888.

TAO, Y. et al. **Efficient and accurate nearest neighbor and closest pair search in high-dimensional space**. ACM Trans. Database Syst.**,** v. 35, n. 3, p. 1-46, 2010. ISSN 0362-5915.

TRAINA JR, C. et al. **Fast Indexing and Visualization of Metric Data Sets using Slim-Trees**. IEEE Trans. on Knowl. and Data Eng.**,** v. 14, n. 2, p. 244-260, 2002. ISSN 1041-4347.

TRAINA JR, C. et al. **Slim-Trees: High Performance Metric Trees Minimizing Overlap Between Nodes**. Proceedings of the 7th International Conference on Extending Database Technology: Advances in Database Technology: Springer-Verlag: p. 51-65, 2000a.

TRAINA JR, C. et al. **Fast feature selection using fractal dimension** Brasilian Symposium on Databases (SBBD). MEDEIROS , C. M. B. E. B., K. EDITORS. João Pessoa, PB: p. 158-171, 2000b.

VIEIRA, M. R. et al. **Boosting k-Nearest Neighbor Queries Estimating Suitable Query Radii**. Proceedings of the 19th International Conference on Scientific and Statistical Database Management, IEEE Computer Society, 2007.

WENGERT, C. et al. **Bag-of-colors for improved image search**. Proceedings of the 19th ACM international conference on Multimedia. Scottsdale, Arizona, USA: ACM: p. 1437-1440, 2011.

XU, P. et al. **A robust texture descriptor using multifractal analysis with Gabor filter**. Proceedings of the Second International Conference on Internet Multimedia Computing and Service. Harbin, China: ACM: p. 147-150, 2010.

YOUNG, G.; HOUSEHOLDER, A. S. **Discussion of a set of points in terms of their mutual distances**. Psychometrika**,** v. 3, n. 1, p. 19-22, 1938.