



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Câmpus de São José do Rio Preto

Thaís Aline Bezerra Correia

**Múltiplas visões coordenadas baseadas em similaridade e
diversidade**

São José do Rio Preto
2019

Thaís Aline Bezerra Correia

**Múltiplas visões coordenadas baseadas em similaridade e
diversidade**

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Orientador: Prof^ª. Dr^ª. Danilo Medeiros Eler

São José do Rio Preto
2019

Correia, Thaisa Aline Bezerra.

Múltiplas visões coordenadas baseadas em similaridade e diversidade / Thaisa Aline Bezerra Correia. -- São José do Rio Preto, 2019

105f. : il.

Orientador: Danilo Medeiros Eler

Dissertação (mestrado) – Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto

1. Ciência da computação. 2. Visualização da informação. 3. Recuperação da informação. 4. Recuperação de dados (Computação) 5. Múltiplas visões coordenadas. I. Título.

CDU – 518.72

Thaís Aline Bezerra Correia

**Múltiplas visões coordenadas baseadas em similaridade e
diversidade**

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Comissão Examinadora

Prof. Dr. Danilo Medeiros Eler
UNESP – Câmpus de Presidente Prudente
Orientador

Prof. Dr. Ronaldo Celso Messias Correia
UNESP – Câmpus de Presidente Prudente

Prof. Dr. Ives Renê Venturini Pola
UTFPR – Universidade Tecnológica Federal do Paraná

Presidente Prudente
23 de janeiro de 2019

À minha mãe, Léia

AGRADECIMENTOS

Agradeço á Deus por seu amor e sustentação, sendo minha fortaleza em todos os momentos da minha vida.

Aos meus pais, Léia, Lenir e Valdemar, pelo amor, cuidado, dedicação e exemplos que me deram, os quais foram fundamentais na minha caminhada.

Ao meu esposo, Marcos, pela apoio, companheirismo e amor. Obrigada por inúmeras vezes me encher de ânimo e coragem para vencer meus desafios diários, sem você eu não conseguiria.

À minha família, pelo amor, incentivo, sempre acreditando e torcendo por mim, vocês são minhas riquezas.

Ao professor Danilo Medeiros Eler pela oportunidade, por ter inicialmente acreditado em mim, pelo apoio e encorajamento a continuar os estudos após a conclusão da graduação. Muito obrigada pela orientação, contribuição, sugestões e ajuda nos trabalhos desenvolvidos. E pelo exemplo de seriedade e dedicação com o ensino.

Aos meus amigos e companheiros de trabalho, Marcelo Pacheco, Thiago Tiedtke e Roberto Ruy, pelo apoio e incentivo durante a caminhada. Obrigada, pois sempre que precisei pude contar com vocês.

Aos professores e amigos que tive oportunidade de conhecer na Unesp, os quais foram fundamentais na minha trajetória como profissional e pessoa.

Pois o Senhor é bom e o seu amor leal é eterno; a sua fidelidade permanece.

(Salmos 100:5)

RESUMO

Com o avanço da tecnologia, o volume e as dimensões de dados provenientes de diferentes naturezas tem aumentado consideravelmente. Nesse cenário, diversas técnicas de visualização de informação, que abordam diferentes tipos, estruturas e dimensionalidades de dados, têm surgido e obtido destaque, apoiando diversas áreas do conhecimento, por meio de representações gráficas que oferecem uma visão ampla e geral dos dados, permitindo a sua análise e compreensão de maneira eficiente. Dessa maneira, Múltiplas Visões Coordenadas têm sido utilizada para tratar a conexão entre as diferentes técnicas de visualização, permitindo associá-las e manipulá-las mais produtivamente, ampliando a percepção do usuário em relação ao conjunto de dados, seus relacionamentos e significados. Neste trabalho, são desenvolvidas novas técnicas de coordenação, sobretudo para dados complexos, aplicando o conceito de similaridade, que permite comparar dados mediante o seu conteúdo, e de diversidade, que na recuperação de dados por similaridade, coopera para que o conjunto de resposta seja relevante e efetivo para o usuário, acrescentando um nível de diversificação entre eles. Por meio da coordenação entre as visões é possível analisar as técnicas de diversidade diretamente nas representações gráficas, verificando os conjuntos de dados como um todo e a vizinhança de seus elementos, apoiando a tarefa de definição de um parâmetro de diversificação adequado. Também são apresentados os experimentos realizados em cenários distintos, aplicando diferentes conjunto de dados, métricas de distância e parâmetros de diversidade, para a análise e comparação das técnicas, de acordo com a resposta alcançada e no contexto exploratório da coordenação entre as visões.

Palavras-chave: visualização da informação. múltiplas visões coordenadas. similaridade. diversidade.

ABSTRACT

With technology advancement, the data volume and dimensions coming from different natures has increased considerably. In this scenario, several information visualization techniques, which deal with different types, structures and data dimensionalities, have arisen and obtained prominence, supporting several knowledge areas, through graphical representations that offer a broad and general data view, allowing the their analysis and understanding efficiently. In this way, Coordinated and Multiple Views have been used to treat the connection between the different visualization techniques, allowing them to be associated and manipulated more productively, increasing the user's perception regarding the data set, their relationships and meanings. In this work, new coordination techniques are developed, especially for complex data, applying the similarity concept, which allows to compare data through its content, and diversity, which in the data recovery by similarity, cooperates so that the response set is relevant and effective for the user, adding a level of diversification among them. Through the coordination between visions it is possible to analyze the diversity techniques directly in the graphical representations, checking the datasets as a whole and the neighborhood of their elements, supporting the task of defining an appropriate diversification parameter. We also present the experiments performed in different scenarios, applying different datasets, distance metrics and diversity parameters, for the techniques analysis and comparison, according to the response achieved and in the exploratory context of the coordinated views.

Keywords: information visualization. coordinated and multiple views. similarity. diversity.

LISTA DE FIGURAS

Figura 1 – Mapa de similaridade gerado pela projeção <i>Sammon's Mapping</i>	5
Figura 2 – Exemplo de aplicação que implementa a coordenação baseada em identificador.	6
Figura 3 – Exemplo de aplicação que implementa a coordenação baseada em tópico.	9
Figura 4 – Exemplo de aplicação que implementa a coordenação baseada em tópicos no tempo.	11
Figura 5 – Exemplo de aplicação que implementa a coordenação baseada em atributos temporais.	12
Figura 6 – Novos espaços multidimensionais com atributos comuns.	13
Figura 7 – Exemplo de aplicação que implementa a coordenação baseada em distância.	14
Figura 8 – Mecanismo de coordenação baseada em semântica.	15
Figura 9 – Exemplo de aplicação que implementa a coordenação baseada em semântica.	16
Figura 10 – Exemplo de aplicação que implementa a combinação de coordenações.	18
Figura 11 – Busca por abrangência.	23
Figura 12 – Busca pelos k-vizinhos mais próximos.	24
Figura 13 – Exemplo do processo coordenação aplicando a similaridade e a diversidade, utilizando uma visão origem e duas visões destino.	39
Figura 14 – Fluxograma do processo de coordenação da Figura 13.	40
Figura 15 – Exemplo do processo coordenação aplicando a similaridade e a diversidade.	41
Figura 16 – Visualizações de diferentes conjuntos de dados no PEx.	42
Figura 17 – Tela PEx: Escolha do par de visões para efetuar o mapeamento.	42
Figura 18 – Tela PEx: Configuração da coordenação por similaridade e diversidade.	43
Figura 19 – Coordenação entre duas visões por meio da busca por similaridade aplicando a diversidade.	44
Figura 20 – Fluxograma do processo de coordenação da Figura 13.	49
Figura 21 – Consulta por similaridade, utilizando a função de distância euclidiana.	50
Figura 22 – Consulta por similaridade aplicando a técnica de diversidade Motley, utilizando a função de distância euclidiana.	52
Figura 23 – Consulta por similaridade aplicando a técnica de diversidade MMR, utilizando a função de distância euclidiana.	53
Figura 24 – Consulta por similaridade aplicando a técnica de diversidade <i>Swap</i> , utilizando a função de distância euclidiana.	54
Figura 25 – Consulta por similaridade aplicando a técnica de diversidade BRID, utilizando a função de distância euclidiana.	55
Figura 26 – Consulta por similaridade, utilizando a função de distância do cosseno.	56

Figura 27 – Consulta por similaridade aplicando a técnica de diversidade Motley, utilizando a função de distância do cosseno.	58
Figura 28 – Consulta por similaridade aplicando a técnica de diversidade MMR, utilizando a função de distância do cosseno.	59
Figura 29 – Consulta por similaridade aplicando a técnica de diversidade <i>Swap</i> , utilizando a função de distância do cosseno.	60
Figura 30 – Consulta por similaridade aplicando a técnica de diversidade BRID, utilizando a função de distância do cosseno.	61
Figura 31 – Consulta por similaridade, utilizando a função de distância mahalanobis.	62
Figura 32 – Consulta por similaridade aplicando a técnica de diversidade Motley, utilizando a função de distância de mahalanobis.	64
Figura 33 – Consulta por similaridade aplicando a técnica de diversidade MMR, utilizando a função de distância de mahalanobis.	65
Figura 34 – Consulta por similaridade aplicando a técnica de diversidade <i>Swap</i> , utilizando a função de distância de mahalanobis.	66
Figura 35 – Consulta por similaridade aplicando a técnica de diversidade BRID, utilizando a função de distância de mahalanobis.	67
Figura 36 – Consulta por similaridade, utilizando a função de distância de mahalanobis.	68
Figura 37 – Consulta por similaridade aplicando as técnicas de diversidade Motley, <i>Swap</i> e BRID, utilizando a função de distância de mahalanobis.	69
Figura 38 – Consulta por similaridade.	70
Figura 39 – Consulta por similaridade aplicando a técnica de diversidade MMR. . .	71
Figura 40 – Consulta por similaridade aplicando a técnica de diversidade MSD. . .	74
Figura 41 – Consulta por similaridade aplicando a técnica de diversidade <i>Swap</i> . . .	75
Figura 42 – Consulta por similaridade aplicando a técnica de diversidade Motley. .	76
Figura 43 – Consulta por similaridade aplicando a técnica de diversidade GMC. . .	77
Figura 44 – Consulta por similaridade aplicando a técnica de diversidade GNE, com $\alpha = 0.0$	78
Figura 45 – Consulta por similaridade aplicando a técnica de diversidade GNE, com $\alpha = 0.5$	79
Figura 46 – Consulta por similaridade aplicando a técnica de diversidade BRID. . .	80
Figura 47 – Consulta por similaridade aplicando a técnica de diversidade BRID. . .	81

LISTA DE TABELAS

Tabela 1 – Conjuntos de dados utilizados nos experimentos.	46
Tabela 2 – Especificação dos experimentos.	46

LISTA DE ALGORITMOS

1	Detecção de Tópicos Baseada em Covariância.	8
2	<i>Maximal Marginal Relevance</i> MMR	28
3	<i>Max-Sum Dispersion</i> MSD	29
4	<i>Swap.</i>	31
5	Motley.	32
6	<i>Greedy Marginal Contribution</i> GMC	32
7	<i>Greedy Randomized with Neighborhood Expansion</i> GNE	33
8	Fase Construção GNE	34
9	Fase Busca Local GNE	35
10	<i>Better Results With Influence Diversification</i> BRID	36

SUMÁRIO

	Resumo	vi
	Abstract	vii
	Lista de Figuras	viii
	Lista de Tabelas	x
	Lista de Algoritmos	xi
1	INTRODUÇÃO	1
2	TÉCNICAS DE COORDENAÇÃO APLICADAS EM PROJEÇÕES MULTIDIMENSIONAIS	4
2.1	Considerações Iniciais	4
2.2	Coordenação baseada em Identificador	5
2.3	Coordenação baseada em Tópicos	6
2.4	Coordenação baseada em Tópicos no Tempo	10
2.5	Coordenação baseada em Atributos Temporais	11
2.6	Coordenação baseada em Distância	12
2.7	Coordenação usando Semântica baseada em Ontologia	14
2.8	Combinação de Coordenações	16
2.9	Considerações Finais	17
3	SIMILARIDADE E DIVERSIDADE	20
3.1	Similaridade	20
3.1.1	Espaço métrico	21
3.1.2	Comparação e consulta por similaridade	22
3.1.3	Métodos de acesso métrico	24
3.2	Diversidade	25
3.2.1	<i>Maximal Marginal Relevance</i> - MMR	28
3.2.2	<i>Max-Sum Dispersion</i> - MSD	29
3.2.3	<i>Swap</i>	30
3.2.4	<i>Motley</i>	30
3.2.5	<i>Greedy Marginal Contribution</i> - GMC	31
3.2.6	<i>Greedy Randomized with Neighborhood Expansion</i> - GNE	33
3.2.7	<i>Better Results With Influence Diversification</i> - BRID	34

3.3	Considerações Finais	37
4	COORDENAÇÕES BASEADAS EM SIMILARIDADE E DIVERSIDADE	38
4.1	Considerações Iniciais	38
4.2	Metodologia	38
4.3	Implementação	41
4.4	Considerações Finais	43
5	EXPERIMENTOS E RESULTADOS	45
5.1	Desenvolvimento dos Experimentos	45
5.2	Resultados	47
5.2.1	Comparação entre Espaços Métricos	47
5.2.1.1	Perspectiva Geral dos Resultados	47
5.2.1.2	Função de Distância Euclidiana	50
5.2.1.3	Função de Distância do Cosseno	55
5.2.1.4	Função de Distância Mahalanobis	61
5.2.2	Análise dos Métodos de Diversidade	70
5.2.2.1	MMR	70
5.2.2.2	MSD	72
5.2.2.3	Swap	72
5.2.2.4	Motley	72
5.2.2.5	GMC	72
5.2.2.6	GNE	73
5.2.2.7	BRID	80
5.3	Considerações Finais	81
6	CONCLUSÃO E TRABALHOS FUTUROS	83
6.1	Considerações Iniciais	83
6.2	Análise dos Resultados	83
6.3	Contribuições	85
6.4	Trabalhos Futuros	85
	REFERÊNCIAS	87

1 INTRODUÇÃO

Atualmente, o volume e a complexidade dos dados provenientes de diferentes áreas e dispositivos tem aumentado significativamente. Para facilitar a organização e o entendimento dessa grande quantidade de dados, técnicas de visualização de informação podem ser empregadas para que esses dados façam sentido ao usuário, fornecendo-o significado e compreensão (WARD; GRINSTEIN; KEIM, 2015). Os sistemas de visualização de informações auxiliam o usuário a assimilar os dados e seus relacionamentos com mais clareza, permitindo analisar conjuntos de dados de uma maneira facilitada, revelando informações que não seriam percebidas sem o poder de abstração da visualização (CARD; MACKINLAY; SHNEIDERMAN, 1999).

Na literatura, são apresentadas diferentes técnicas de visualização, para diversos domínios, tipos, estruturação e dimensão de dados (OLIVEIRA; LEVKOWITZ, 2003). Essas técnicas criam uma representação gráfica a partir do conjunto de dados sob análise, para facilitar o entendimento do comportamento dos seus atributos, por exemplo, a bem conhecida coordenadas paralelas (INSELBERG, 1985). Tais técnicas possuem vantagens e desvantagens, ressaltando determinados tipos de dados, relacionamentos, hierarquias ou relação entre atributos. Em um contexto exploratório, com diferentes tipos de técnicas de visualização, as Múltiplas Visões Coordenadas (*Coordinated Multiple Views*) permitem associar diferentes tipos de representações gráficas, possibilitando explorar e mapear relacionamentos entre elas, fazendo com que a análise do usuário seja eficaz (NORTH; SHNEIDERMAN, 1999; MCEWAN; IGONIDERIGHA; BENYON, 2014).

A interação é extremamente importante em sistemas de visualização de informações, principalmente nos que aplicam Múltiplas Visões Coordenadas, pois viabiliza o contato do usuário com a representação visual. Dessa maneira, contribui para a exploração dos dados, permitindo focar em um ponto de maior importância, realizar comparações, selecionar instâncias de interesse, fazendo com que haja uma melhor compreensão sobre os dados e seus relacionamentos, cumprindo o propósito da visualização (MULLER, 2005). As técnicas de coordenação, junto às ferramentas de interação, são responsáveis por efetuar o mapeamento entre as instâncias das visões, podendo relacionar um ou mais itens entre as diferentes visualizações envolvidas no processo exploratório.

Em diversas áreas do conhecimento há muitas aplicações desenvolvidas (SEIFERT; SCHNEIDER, 2003; GODINHO et al., 2007; SABOL et al., 2009; KING et al., 2014; SADANA; STASKO, 2016) que utilizam Múltiplas Visões Coordenadas para o auxílio da resolução de vários problemas, das quais apresentam metodologias e soluções específicas para o determinado domínio. As técnicas de coordenação, em geral, se baseiam no mapea-

mento de elementos entre as visões relacionadas no processo, isto é, a partir da seleção realizada em uma visão, são mapeados elementos em outras visões de acordo com algum critério definido previamente pela técnica. Os mapeamentos mais simples são estabelecidos baseando-se em atributos convencionais, por exemplo, dados numéricos e textuais (e.g., identificadores). Porém, com o grande crescimento da produção e utilização de dados complexos – por exemplo, imagens, músicas, dados geográficos, estruturas genéticas, vídeos, textos longos, dados multidimensionais em geral, séries temporais – tem-se gerado uma demanda de desenvolvimento de técnicas de coordenação que permitam o mapeamento mediante o conteúdo desses dados.

Descrever e comparar objetos complexos por atributos textuais ou numéricos, não permite apresentar resultados fidedignos à intenção da busca, pois a marcação de objetos complexos por atributos convencionais é por muitas vezes subjetiva, e diferentes pessoas podem rotular de maneira diferente o mesmo objeto. Desta forma, isso pode não representar a intenção da busca realizada. Nesse sentido, para a comparação entre objetos complexos é importante extrair características que são relacionadas ao conteúdo do que se quer representar, as quais são utilizadas no lugar do objeto propriamente para a comparação, gerando, na maioria das abordagens, vetores de características que são comparados por meio de funções específicas, que medem o quanto os objetos são dissimilares entre si (LIU et al., 2007; RAZENTE, 2009).

A busca por similaridade efetuada em conjunto de dados homogêneos, em que há uma alta repetição de elementos semelhantes entre si, parcial ou total, pode produzir resultados com instâncias muito parecidas, sem apresentar uma relevância mínima adequada. A utilização da diversidade em buscas por similaridade tem como objetivo aumentar a relevância do resultado, fazendo com que a partir de um fator de diversidade definido pelo usuário, o resultado da busca não contenha apenas objetos similares ao centro de consulta, mas também diversos entre si, para que o usuário obtenha uma resposta mais relevante e efetiva (CARBONELL; GOLDSTEIN, 1998; ZHENG et al., 2017).

O principal objetivo deste trabalho é a evolução das técnicas de coordenação que lidam com dados complexos, em especial aquelas cujos mapeamentos são baseados em similaridade (ELER et al., 2009b; ELER et al., 2009a), para que essas técnicas também levem em consideração a diversidade nas consultas, permitindo identificar relações menos homogêneas, apresentando resultados mais significativos. Para tanto, novos mapeamentos entre os itens das diferentes visões são gerados, vinculando-os na ferramenta PEx (*Projection Explorer*) (PAULOVICH; OLIVEIRA; MINGHIM, 2007), que disponibiliza técnicas de projeções multidimensionais para o espaço bidimensional, respeitando a disposição entre os elementos dos conjuntos de dados. Além disso, o resultado deste trabalho também colabora com o entendimento das técnicas de diversidade, aplicadas em tarefas de recuperação de imagens, entre outros tipos de dados. Isso se dá, pois é possível identificar diretamente

nas visões o impacto que a alteração do fator de diversidade provoca no resultado das consultas, verificando o quanto os elementos estão distantes uns dos outros, apoiando o processo de definição de um parâmetro de diversidade adequado, que corresponda à intenção de busca.

Experimentos são definidos e executados, aplicando diferentes bases de dados existentes na literatura, tipos de funções de distância e parâmetros das metodologias abordadas, com o intuito de apresentar os resultados alcançados, tanto na perspectiva de comparação e análise dos elementos resultantes da busca baseando-se na similaridade e diversidade, quanto na contribuição no processo exploratório da coordenação entre as visualizações.

Os demais capítulos deste trabalho estão organizados da seguinte maneira. No Capítulo 2 são apresentadas técnicas de coordenação aplicáveis em visualizações geradas a partir de projeções multidimensionais (e.g., mapas de similaridades), que permitem utilizar diversos tipos de dados. Em seguida, no Capítulo 3, são abordados os conceitos de similaridade e de diversidade, como também, as principais técnicas e algoritmos existentes na literatura. No Capítulo 4 é apresentada a metodologia de coordenação baseando-se na similaridade e diversidade, a implementação da proposta do trabalho, e ferramentas utilizadas. No Capítulo 5 são relatadas as informações sobre o desenvolvimento dos experimentos, e uma elucidação dos resultados obtidos. Finalmente, no Capítulo 6, é realizada uma análise dos resultados, bem como uma descrição das contribuições geradas e futuros trabalhos.

2 TÉCNICAS DE COORDENAÇÃO APLICADAS EM PROJEÇÕES MULTIDIMENSIONAIS

2.1 CONSIDERAÇÕES INICIAIS

Em Múltiplas Visões Coordenadas, as técnicas de coordenação definem como serão realizados os mapeamentos entre as representações gráficas, de acordo com a interação aplicada em uma visão e as modificações propagadas para as demais, sendo que cada uma das técnicas são aplicáveis à determinados domínios e tipos de dados. No trabalho de [BALDONADO, WOODRUFF e KUCHINSKY \(2000\)](#), as técnicas de coordenação são divididas em duas partes: nas funções de acoplamento, que designam como serão feitas as associações entre as visões; e o modelo de propagação, o qual determina em que condições as funções são aplicadas, e as mudanças propagadas.

Existem diversas técnicas de coordenação na literatura que exploram soluções para diversos domínios de aplicações, tipo de dados e dimensionalidades, sendo que cada uma tem uma especificação e contexto de utilização. Neste trabalho, será focado na coordenação entre mapas de similaridade, que são representações visuais para diferentes tipos de dados multidimensionais, geradas por técnicas de projeções, as quais reduzem a dimensionalidade para um menor espaço resultante, conservando noções de similaridade do espaço original. Em geral, no espaço bidimensional projetado em mapas de similaridade, os elementos são representados por pontos na tela, respeitando a relação de similaridade entre eles, sendo que os elementos estão mais próximos daqueles que são similares com ele, de acordo com a métrica adotada. Como pode ser identificado na Figura 1, na qual ilustra um mapa de similaridade utilizando a projeção multidimensional *Sammon's Mapping*, em que as instâncias mais parecidas são separadas por classes, identificadas por cor. ([SAMMON, 1964](#); [PAULOVICH](#); [OLIVEIRA](#); [MINGHIM, 2007](#); [ELER et al., 2008](#)).

A seguir, são apresentadas as principais técnicas de coordenação para mapas de similaridade, sendo elas baseadas em identificador, em tópicos, em tópicos identificados durante a evolução temporal, em atributos temporais das instâncias do conjunto de dados, em distâncias, em semântica, e na combinação de diferentes técnicas de coordenação, as quais criam um mapeamento entre as visualizações utilizando abordagens específicas no processo.

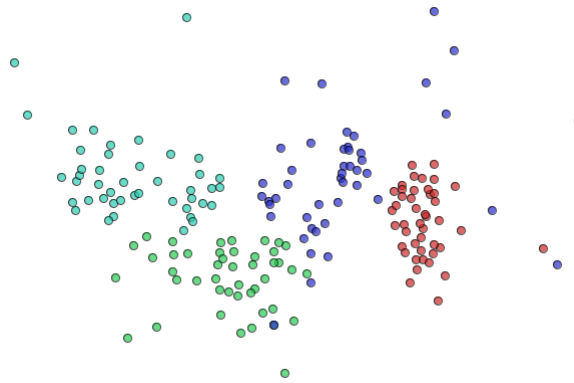


Figura 1 – Mapa de similaridade gerado pela projeção *Sammon's Mapping*.

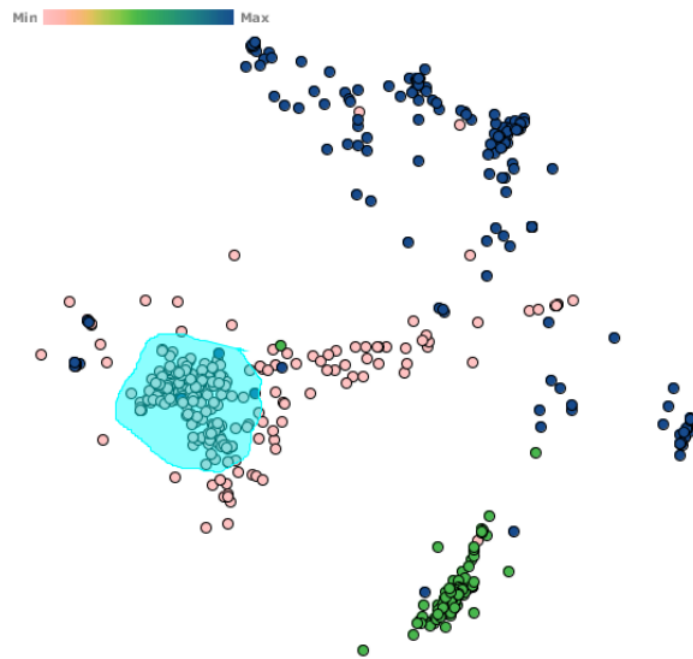
2.2 COORDENAÇÃO BASEADA EM IDENTIFICADOR

Nessa técnica, a coordenação é realizada baseando-se no identificador dos elementos entre as visões, a partir da seleção efetuada em uma visão, os mesmos elementos são encontrados e indicados em outra visão, por meio do relacionamento direto definido por um identificador, que deve ser um atributo único dos elementos que constituem o conjunto de dados.

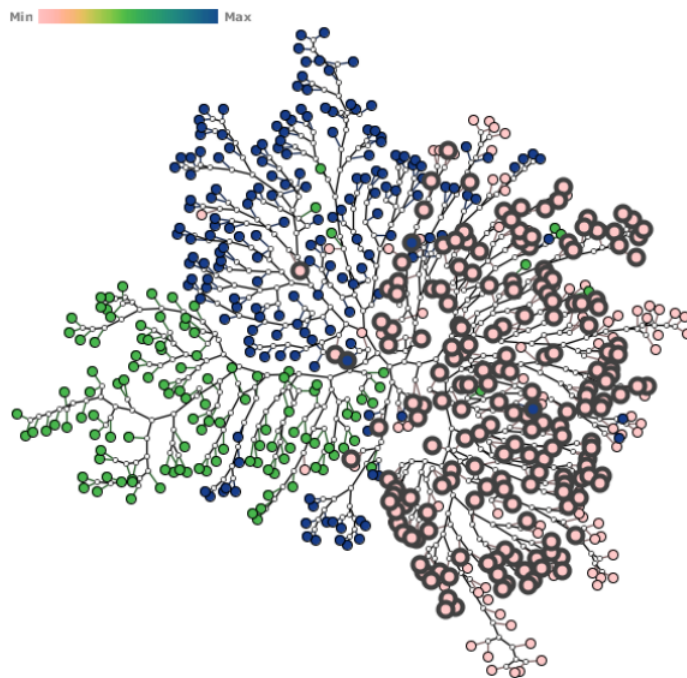
Na Figura 2, é ilustrado um exemplo de visão coordenada, baseada em identificador da aplicação *Projection Explorer* (PEX), utilizando um mesmo conjunto de dados, porém com tipos de representações distintas. Na Figura 2a a projeção é construída pela técnica *Least Squares Projection* (LSP) (PAULOVICH et al., 2008) e, na Figura 2b por *Neighbour-Joining* (NJ) (CUADROS et al., 2007). Por meio da ferramenta, na Figura 2a são selecionados um grupo de elementos que constituem a visão origem, e os mesmos elementos são destacados na visão destino, ilustrado na Figura 2b, salientando que a cor dos elementos nas visões representa a classe de cada um, definição esta, indicada pelo usuário na ferramenta (ELER, 2011).

A análise isolada realizada pelo usuário diante de diversas visões, dificulta o processo de compreensão dos relacionamentos entre os elementos, não permitindo a exploração eficiente entre as técnicas de visualização apresentadas. Porém, ao aplicar a técnica de coordenação por identificador, o usuário pode associar as mesmas instâncias em diferentes visualizações, explorando diversos espaços multidimensionais, representações gráficas e projeções, produzindo uma comparação e análise conjunta, explorando as vantagens de cada uma das visões.

A desvantagem da coordenação baseada por identificador, é que ela não é eficaz para conjuntos de dados que não possuem uma relação direta, como um identificador, restringindo assim a aplicação da solução em conjunto de dados relacionáveis.



(a) Visão gerada por LSP.



(b) Visão gerada por NJ.

Figura 2 – Exemplo de aplicação que implementa a coordenação baseada em identificador.

Fonte: (ELER, 2011)

2.3 COORDENAÇÃO BASEADA EM TÓPICOS

Como foi descrito anteriormente, na coordenação baseada em identificador os dados precisam estar diretamente ligados por um único atributo específico, fazendo com que a exploração no processo de coordenação, seja apenas aplicável à um determinado contexto.

Nesse sentido, a coordenação baseada em tópicos foi desenvolvida, para possibilitar a exploração e análise de diferentes conjuntos de dados textuais, que tenham alguma relação entre si, não necessariamente direta.

A definição de tópico, nesse contexto, consiste em uma ou mais palavras que definem um sumário representativo do assunto, ou tema relacionado a um determinado documento ou conjunto de documento, sendo essencial em situações em que a abstração e conhecimento do que se trata um particular grupo de documentos, beneficia o usuário a focar no que realmente interessa-lhe (ELER et al., 2009b).

A abordagem é realizada por meio da extração de tópicos de interesse gerados na interação com uma visão origem, onde o usuário seleciona os documentos que deseja incluir no processo. Após a seleção e extração, os tópicos gerados são propagados nas demais visões em coordenação, e apenas os elementos totalmente cobertos pelos tópicos que são indicados.

A extração dos tópicos aplicada na visão é baseada na computação da covariância dos termos dos elementos, presentes no conjunto de dados. As linhas da matriz de covariância constituem os documentos e as colunas os termos que aparecem no conteúdo do mesmo, e o valor associado na linha/coluna é o número da ocorrência do termo, ou seja, quantas vezes ele foi encontrado no documento. A covariância da matriz é calculada através da Equação 2.1, onde n é o número de documentos selecionados, $\bar{t}_i$ e $\bar{t}_j$ são as frequências médias do i -ésimo termo t_i e j -ésimo termo t_j , respectivamente, e t_{ki} e t_{kj} são as frequências dos i -ésimo termo e j -ésimo termo no documento k , a qual define o quanto os termos variam junto ao grupo dos elementos selecionados (WASSERMAN, 2004; ELER et al., 2009b).

$$cov(t_i, t_j) = \frac{1}{n-1} \sum_{k=1}^n (t_{ki} - \bar{t}_i)(t_{kj} - \bar{t}_j) \quad (2.1)$$

Após o cálculo da matriz de covariância, os dois termos com maior valor são escolhidos como tópicos iniciais para sintetizar a seleção efetuada, e então a covariância média entre eles e o restante dos termos é calculada através da Equação 2.2, onde t_i e t_j são os termos com maior covariância, e com base neles, a covariância média é calculada para cada termo t_s .

$$cov_m(t_i, t_j, t_s) = \frac{cov(t_i, t_s) + cov(t_j, t_s)}{2} \quad (2.2)$$

Após computar o valor da covariância para o termo t_s , é verificado se o valor resultante é maior que um limiar α definido, se for, então t_s é inserido na listagem de termos do tópico. Os dois termos iniciais escolhidos para compor o tópico tem um grande impacto no resultado final da abordagem, então para solucionar essa questão, se a razão entre a covariância de qualquer par de termos ainda não analisados e a covariância do par

de termos selecionados inicialmente for maior que um limiar β , então os dois termos ainda não analisados são considerados como um novo tópico e então o processo é repetido com base nos novos termos encontrados.

O Algoritmo 1 descreve o processo, sendo $\mathbf{M}$ a matriz $n \times m$, a relação de documentos por termos, α o limiar para adicionar novos termos a um tópico, β o limiar para adicionar novos tópicos à lista de tópicos e L a lista resultante de tópicos identificados.

Algoritmo 1: Detecção de Tópicos Baseada em Covariância.

```

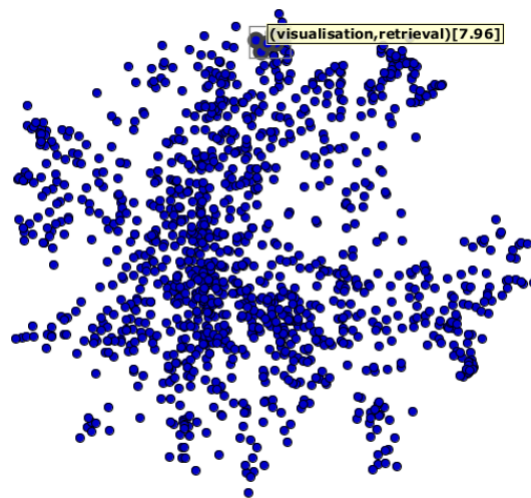
1  $t_i, t_j = \max\{cov(t_i, t_j)\} \forall t_i, t_j \in \mathbf{M}$ 
2  $covariancia = cov(t_{imax}, t_{jmax})$ 
3  $cov_{max} = covariancia$ 
4  $L = \emptyset$ 
5 enquanto  $covariancia/cov_{max} \geq \beta$  faça
6   cria um novo tópico  $T = \emptyset$ 
7   adiciona os termos  $t_i$  e  $t_j$  a  $T$ 
8   para todo  $t_k \in M$  e  $t_k \neq t_i$  e  $t_k \neq t_j$  faça
9     se  $cov_m(t_i, t_j, t_k)/covariancia \geq \alpha$  então
10       adiciona o termo  $t_k$  a  $T$ 
11     fimse
12   fim
13   adiciona o tópico  $T$  à lista de tópicos  $L$ 
14    $t_i, t_j = \max\{cov(t_i, t_j)\} \forall t_i, t_j \in \mathbf{M}$  e  $(t_i \notin l_b \text{ ou } t_j \notin l_b, \forall l_b \in L)$ 
15    $covariancia = cov(t_i, t_j)$ 
16 fim

```

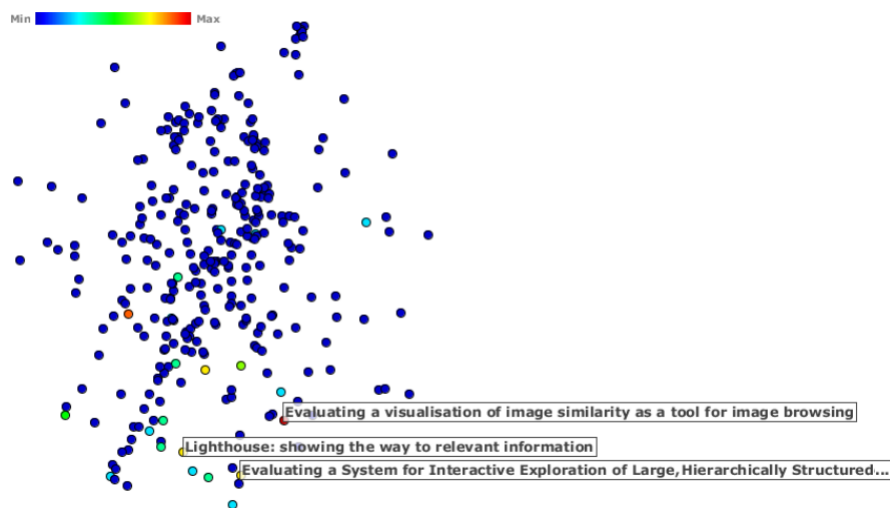
Na visão coordenada por tópicos, o usuário seleciona os documentos que deseja extrair os tópicos na visualização origem. E, então, esses tópicos são buscados nas demais visões envolvidas no processo. Se todos os termos do tópico forem encontrados em algum documento nas demais visões, então é calculado o valor da ocorrência, que é a média de quantas vezes cada termo foi detectado no documento, o qual, é normalizado para o intervalo 0 e 1. No caso de não encontrar nenhum termo do tópico no documento, então o seu valor de ocorrência será definido para 0. E, com base nesses valores resultantes, é realizado um mapeamento dos tópicos referentes aos documentos selecionados na visão origem, para as visões destinos, podendo indicar a abrangência e nível de ocorrência deles, ressaltando que na técnica apenas os termos com maior covariância são utilizados no mapeamento.

A Figura 5 mostra um exemplo de visualização coordenada baseada em tópicos, ambas visões foram formadas através da técnica de LSP. A visão ilustrada na Figura 3a

foi elaborada a partir dos artigos científicos entre 1997 e 2008, na conferência *Information Visualization*. E, a visão da Figura 3b, é construída com base nos artigos entre 1995 e 2008, da conferência *IEEE Symposium on Information Visualization*. Na visão apresentada na Figura 3a, são selecionados alguns documentos, e partir deles, um tópico é gerado, onde são indicados os dois termos com maior covariância: “*visualisation, retrieval*”. Em seguida, na visão destino, os documentos que possuem total ocorrência dos termos presentes no tópico são indicados, apoiado pela escala de cor que define a ocorrência mínima e máxima dos termos.



(a) Visão gerada por LSP.



(b) Visão gerada por LSP.

Figura 3 – Exemplo de aplicação que implementa a coordenação baseada em tópico.

Fonte: (ELER, 2011)

Nesta seção é descrita a metodologia de extração de tópicos, na qual é possível realizar a inclusão de outras técnicas de extração, detecção e análise dos tópicos. A desvantagem da técnica é que ela pode ser apenas aplicada para dados que possuem informação textual associada, porém traz uma solução robusta para a área de atuação.

2.4 COORDENAÇÃO BASEADA EM TÓPICOS NO TEMPO

O objetivo central desta técnica de coordenação é que, a partir de documentos de interesse selecionados na visão origem, seja possível realizar o mapeamento do vínculo dos mesmos com seus vizinhos, de acordo com a adição temporal de novos documentos, permitindo fazer uma análise do conteúdo relacionado, baseando-se na extração de tópicos. Como na técnica anterior, é uma abordagem para dados textuais para mapas de conjunto de documentos baseado em similaridade.

Dada a visão origem que projeta o mapa de similaridade do conjunto de dados, a cada adição de novos documentos ao conjunto, em um período de tempo, uma nova visualização é gerada, denominada de *Time-Span Maps (TS Maps)*, a qual agrega os *TS Maps* anteriores. Quando o usuário seleciona os documentos de interesse na visão origem, a extração de tópicos é aplicada sobre eles. E, conforme novos mapas são construídos, esses documentos são mapeados diretamente, como na coordenação baseada em identificadores, e uma extração de tópicos é realizada com base nos seus k vizinhos mais próximos na atual *TS Map*, permitindo compreender o assunto dos documentos que se relacionam às instâncias de origem (ELER et al., 2009b).

A Figura 4 ilustra um exemplo da coordenação baseada em tópicos no tempo, utilizando o conjunto de dados formado pela coleção de artigos da conferência *Coordinated & Multiple View (CMV)*, entre 2003 e 2008. São definidos três *TS Maps*, o primeiro é o localizado no canto superior esquerdo, no qual emprega artigos dos anos 2003 e 2004, o segundo no canto superior direito, com artigos dos anos 2003 a 2006 e o terceiro logo abaixo, definido por artigos dos anos de 2003 a 2008. No primeiro *TS Map* o usuário seleciona os documentos de interesse, e logo em seguida a extração do tópico relacionado é efetuada e indicada. O segundo *TS Map* possui os mesmos documentos do primeiro e mais de outros anos, onde é indicado em um tom mais forte de vermelho os documentos com o mesmo identificador selecionado no primeiro, e com base na vizinhança encontrada e destacada é aplicada uma nova extração de tópicos. No terceiro *TS Map*, é realizado o mesmo procedimento que no segundo, porém novos documentos são adicionados junto aos antecedentes (ELER et al., 2009b).

Como descrito na seção anterior, é possível combinar técnicas de extração de tópicos diferentes nas visões envolvidas neste processo de coordenação. No exemplo da Figura 4, nos mapas superiores é utilizada a detecção com base na covariância dos termos, e na inferior o algoritmo LWR é aplicado, baseado na detecção por regras de associação, desenvolvido por LOPES et al. (2007), sendo possível também construir o relacionamento da vizinhança com diferentes técnicas.

O resultado do processo da coordenação baseada em tópicos no tempo é a clara compreensão do relacionamento entre as instâncias de interesse e o conjunto de dados,

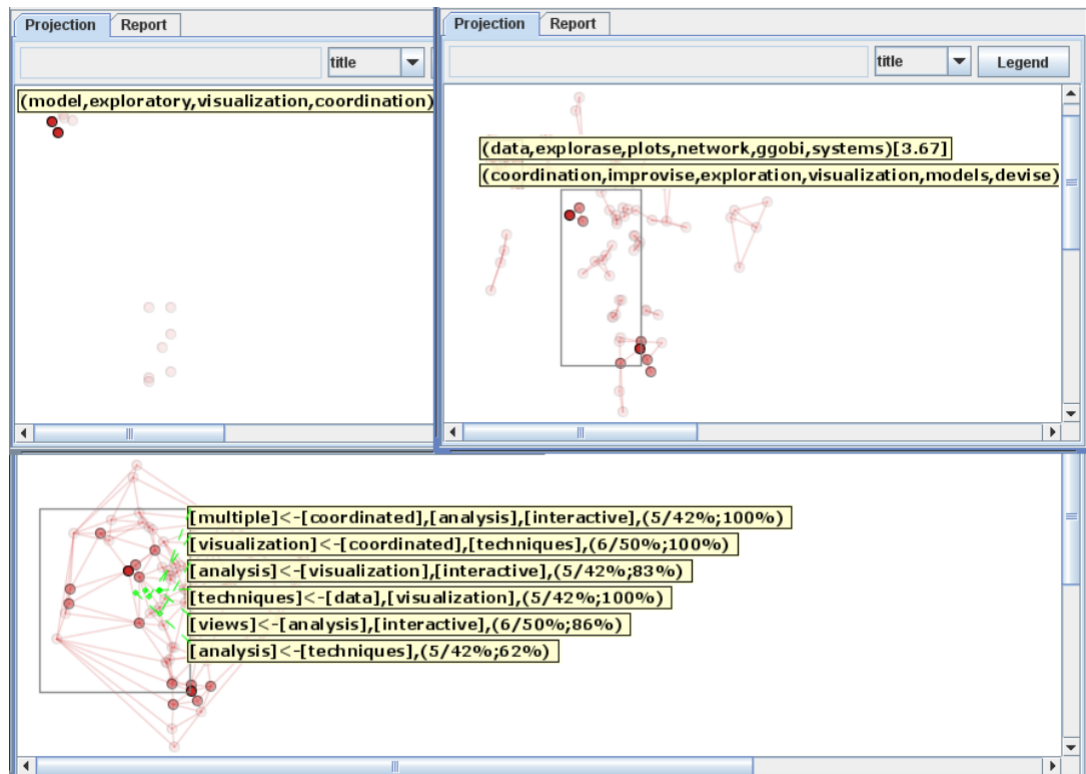


Figura 4 – Exemplo de aplicação que implementa a coordenação baseada em tópicos no tempo.

Fonte: (ELER et al., 2009b)

permitindo entender a evolução e o significado, por meio do conteúdo dessa associação.

2.5 COORDENAÇÃO BASEADA EM ATRIBUTOS TEMPORAIS

Os conjuntos de dados podem conter elementos que possuem algum atributo, que os posicione em relação ao tempo. Nesse sentido, a abordagem de coordenação baseada em atributos temporais é designada para mapear e indicar esse relacionamento temporal entre as instâncias e diferentes mapas. Nessa abordagem, a partir do conjunto de dados representado na visão origem, o usuário pode selecionar os elementos de interesse. E, a partir deles, realizar um mapeamento entre a visão origem e as demais, por meio de uma verificação do atributo de associação temporal, pela comparação do resultado booleano através de um operador de relacionamento, o qual é definido pelo usuário (ELER, 2011).

A Figura 5 apresenta um exemplo de coordenação baseada em atributos temporais, com mapas de similaridade construídos através da técnica LSP, sendo que o mapa origem, ilustrado à esquerda, é construído a partir de um conjunto de artigos da conferência *Information Visualization (IV)*, entre os anos 1997 e 2008, e o apresentado à direita, com uma coleção de artigos da conferência *IEEE Symposium of Information Visualization (IEEE InfoVis)*, entre 1995 e 2008. Na visão origem o usuário seleciona os elementos

de interesse, e define que operador relacional utilizar na comparação lógica do atributo temporal, que no caso do exemplo é o ano de publicação do documento. Em seguida, o mapeamento é realizado na visão destino, encontrando as instâncias que satisfazem a sentença lógica, ou seja, os documentos que foram publicados com data maior que os selecionados na visão origem (ELER, 2011).

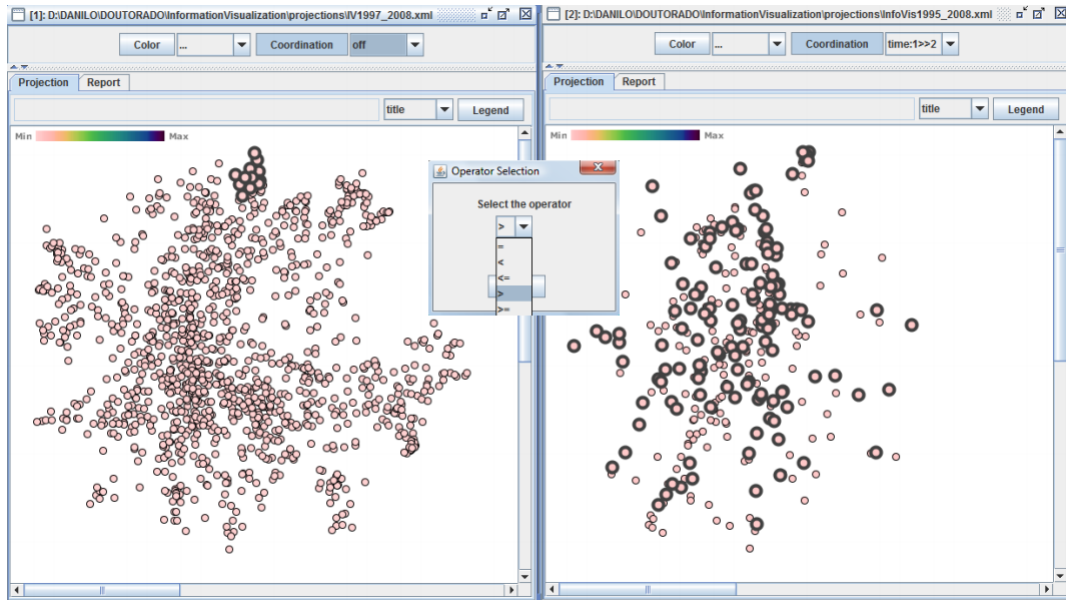


Figura 5 – Exemplo de aplicação que implementa a coordenação baseada em atributos temporais.

Fonte: (ELER, 2011)

Por meio de um mapeamento simples, a técnica permite efetuar uma análise temporal de qualquer conjunto de dados, que possua um atributo que defina tempo, produzindo uma compreensão do relacionamento das instâncias de acordo com linha de tempo e projeção empregada na coordenação.

2.6 COORDENAÇÃO BASEADA EM DISTÂNCIA

A técnica de coordenação baseada em distância permite realizar o mapeamento entre as instâncias, em relação a sua vizinhança, empregando uma abordagem arbitrária de extração de características em um espaço métrico associado. O usuário seleciona um grupo de elementos na visualização de origem, e então, é realizada uma busca no espaço original pelos k vizinhos mais próximos para cada elemento, e os mesmos são indicados na demais visões envolvidas no processo de coordenação (ELER, 2011).

Para que seja possível encontrar os vizinhos mais próximos das instâncias entre as visões, é necessário aplicar uma extração de características intrínsecas aos elementos, e definir uma função de distância para medir a similaridade entre eles. Em muitos casos,

os conjuntos de dados utilizados na coordenação não estão definidos no mesmo espaço multidimensional, como por exemplo em uma coleção de imagens. Porém é possível extrair características das mesmas com diferentes algoritmos e parâmetros, gerando um conjunto resultante de atributos específicos para cada coleção. Nesse sentido, a técnica de coordenação baseada em distância estabelece um pré-processamento que tem o objetivo de transformar cada espaço das visões envolvidas, para um espaço multidimensional comum entre elas.

A Figura 6 ilustra um exemplo de transformação do espaço multidimensional de dois conjuntos definidos distintos em um espaço multidimensional comum, aplicado pela técnica de coordenação baseada em distância. É realizada uma análise dos atributos dos espaços multidimensionais envolvidos, a fim de encontrar atributos comuns, e a partir cada espaço é transformado em um novo, formado com apenas os atributos comuns entre o conjunto de dados das visões.

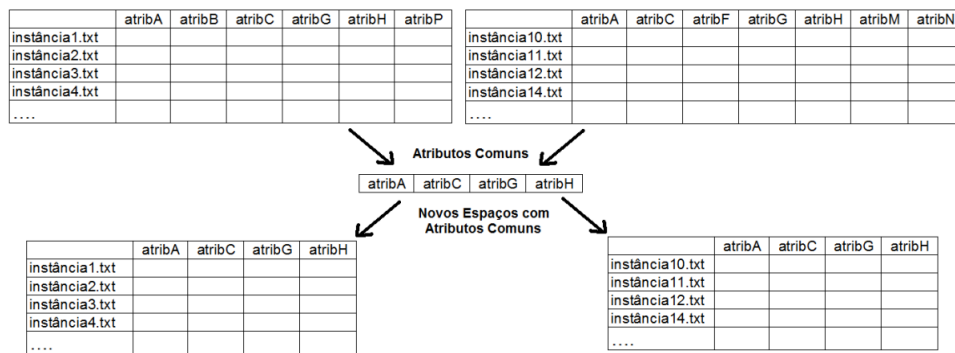


Figura 6 – Novos espaços multidimensionais com atributos comuns.

Fonte: (ELER, 2011)

Após efetuar o pré-processamento, as distâncias entre os elementos dos novos espaços multidimensionais são calculadas, por meio de métricas tradicionais, como a distância euclidiana, *city block*, cosseno, entre outras. O usuário pode então selecionar as instâncias de interesse na visão origem e definir o valor dos k vizinhos mais próximos, permitindo mapeá-los nas outras visualizações coordenadas, ou seja, cada instância selecionada na visão origem vai ter os seus k vizinhos mais próximos indicados nas visões destinos, podendo ainda ter alguma instância de interesse com vizinhos em comum nas outras visões.

É mostrado na Figura 7, um exemplo de coordenação baseada em distância. Nele a visão apresentada na Figura 7a, é gerada pela técnica LSP, composta pelo conjunto de artigos da conferência *Information Visualization (IV)*, e a visão ilustrada na Figura 7b, construída também pela técnica LSP, e formada pelos artigos da conferência *IEEE Symposium on Information Visualization (InfoVis)*. Na visão mostrada na Figura 7a, o usuário seleciona um elemento, e então na visão destino ilustrada na Figura 7b, são indicados os seus três vizinhos mais próximos. Como nas visualizações do exemplo são

externados os títulos das instâncias, nota-se que elas estão relacionadas pelo assunto “animation” (ELER, 2011).

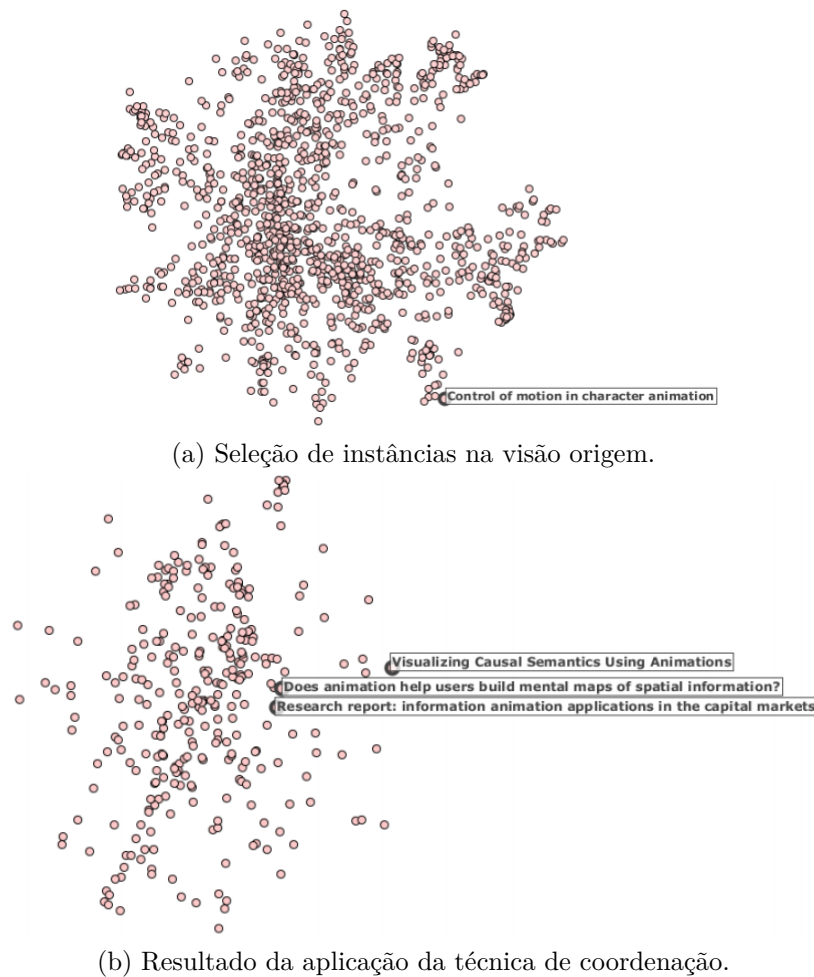


Figura 7 – Exemplo de aplicação que implementa a coordenação baseada em distância.

Fonte: (ELER, 2011)

A técnica de coordenação baseada em distância é capaz de atuar em cenários onde os conjuntos de dados não possuem exatamente o mesmo espaço multidimensional, porém é necessário ter algum atributo relacionado entre os espaços. A abordagem permite compreender as visões de uma melhor maneira, com base nas distâncias calculadas, oferecendo flexibilidade na etapa de projeção e métrica empregada.

2.7 COORDENAÇÃO USANDO SEMÂNTICA BASEADA EM ONTOLOGIA

A ontologia representa um domínio específico, dividido pelas definições de conceitos e suas relações. Os conceitos são hierarquicamente organizados, sendo que os axiomas acrescentam significado para eles, e por meio da declaração de entidades cada conceito é relacionado com o outro. Na Figura 8, é ilustrado como a ontologia é aplicada na técnica

de coordenação baseada em semântica, sendo que além de permitir a inserção de relações na representação, a ontologia possibilita que o *reasoner* efetue implicações, por meio da declaração de axiomas, e na semântica definida para construções de lógicas baseadas na linguagem que define a ontologia. Através da definição da ontologia, que é processada e fundamentada pelo *reasoner*, é possível que a coordenação estabeleça o mapeamento entre as visões (PRATES et al., 2013).

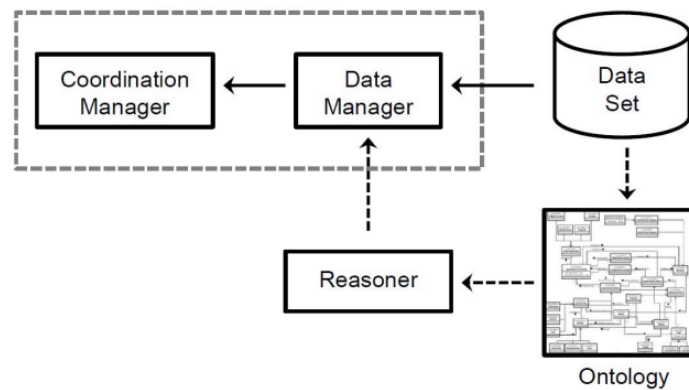


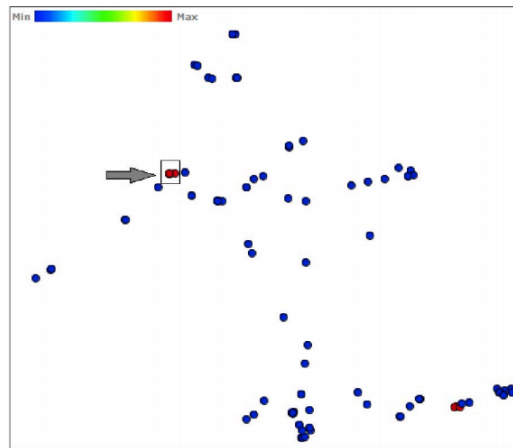
Figura 8 – Mecanismo de coordenação baseada em semântica.

Fonte: (PRATES et al., 2013)

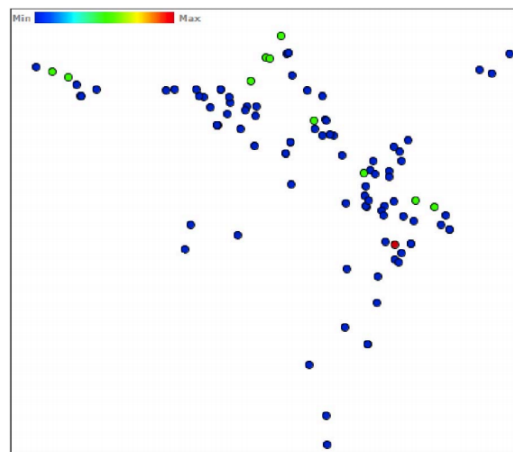
Na utilização da representação semântica, o mapeamento pode estabelecer relações pela ontologia, que ajuda a representar como os conceitos são organizados, e a definição de um conceito específico por axiomas permite classificar as instâncias sob um conceito, se eles possuem relações definidas pelos axiomas.

O processo de coordenação é dividido em quatro etapas, na primeira é feito o carregamento da ontologia no sistema, em seguida, quando um termo é selecionado pelo usuário, é realizada a correlação entre este termo, e os termos das definições de classes na ontologia. Posteriormente, é efetuada a recuperação da classe correspondente, e finalmente, a partir dela, também são recuperadas as relações pai, filhos e restrições de hierarquia.

Na abordagem de PRATES et al. (2013), a coordenação é realizada extraindo termos das instâncias selecionadas, e estes são utilizados para recuperar os tópicos relacionados nas outras visões por meio da semântica modelada pela ontologia. A Figura 9, ilustra um exemplo de coordenação semântica, composto por duas coleções de documentos diferentes, formadas por descrições de imagens médicas. Na representação gráfica, a escala de cor é utilizada para mapear as frequências dos termos selecionados nas instâncias, na qual a cor azul é o mínimo, e a cor vermelho é o máximo da frequência. Inicialmente, o usuário seleciona um ou mais termos que fazem parte da ontologia. Em seguida, são verificadas e destacadas, as instâncias que possuem a ocorrência dos mesmos, através da extração de tópicos, sendo que a coloração de cada uma é baseada na frequência desses termos, conforme é ilustrado nas Figuras 9a e 9b.



(a) Representação visual do conjunto de dados 1.



(b) Representação visual do conjunto de dados 2.

Figura 9 – Exemplo de aplicação que implementa a coordenação baseada em semântica.
Fonte: (PRATES et al., 2013)

A medida que mais termos relacionados por meio da ontologia são adicionados no processo, há um aumento considerável de instâncias relacionadas destacadas, ultrapassando a capacidade de exploração através da seleção do usuário (ELER; PRATES; GARCIA, 2014).

2.8 COMBINAÇÃO DE COORDENAÇÕES

As técnicas de coordenação tem como objetivo aumentar a compreensão dos dados, por meio das diferentes abordagens de visualização, porém em alguns casos apenas uma técnica de coordenação pode não trazer uma contribuição significativa ao processo. Nesse sentido, a *Join Coordination* propõe reunir diferentes técnicas de coordenação, para que seja possível extrair o melhor de cada uma, gerando um mapeamento resultante da combinação de todas, de modo a suprir as limitações de cada uma (ELER, 2011). A Equação 2.3 define o relacionamento das técnicas de coordenação, onde n é o número de técnicas envolvidas

no processo, $Map_{Tecnica}$ é o mapeamento da n -ésima técnica e Map_{Final} o resultado da combinação de coordenações. É possível destacar que, o mapeamento final contém apenas as instâncias que forem resultantes de todas as técnicas de coordenação envolvidas no processo, por ser empregado na definição a operação de intersecção ($\otimes$), porém é possível gerar outras maneiras de combinação, utilizando diferentes tipos de operadores.

$$Map_{Final} = Map_{Tecnica_1} \otimes Map_{Tecnica_2} \otimes \dots \otimes Map_{Tecnica_n} \quad (2.3)$$

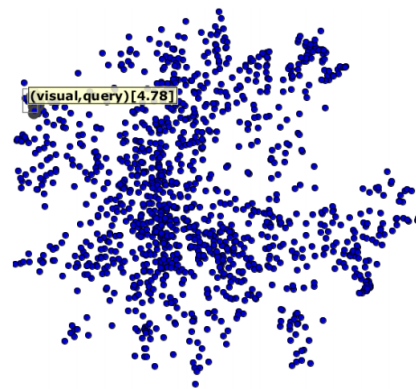
O processamento realizado pela *Join Coordination* é similar ao de outras técnicas de coordenação, descritas nos capítulos anteriores, durante o processo de exploração o usuário seleciona as instâncias de interesse em uma visão origem, e então é aplicada a coordenação nas demais visões. Porém, como há mais de uma técnica de coordenação envolvida, cada mapeamento é realizado de maneira independente. Então ao concluírem seus resultados, eles são combinados para gerar o mapeamento final.

Na Figura 10, é ilustrado um exemplo de *Join Coordination*, onde a visão apresentada na Figura 10a foi construída através da técnica LSP, com os documentos da conferência *Information Visualization (IV)*, a visão apresentada na Figura 10b foi gerada pela técnica LSP com a coleção de artigos científicos da conferência *IEEE Symposium on Information Visualization (InfoVis)*, e foram utilizadas dois tipos de técnicas de coordenação, a *Topic Coordination* e *Time Coordination*. Primeiramente as instâncias de interesse são selecionadas na visão 10a, e então a *Join Coordination* executa a *Topic Coordination*, gerando um tópico associado. Em seguida, a *Time Coordination* efetua o processo sobre o mapeamento da coordenação feita anteriormente, nesse caso emprega o operador relacional $>$, resultando na visão 10b, onde são indicadas com borda em negrito, as instâncias que são cobertas pelo tópico gerado e com atributo temporal maior que o definido pelo usuário. É importante salientar que embora tenha sido realizada a aplicação da *Join Coordination*, o mapeamento individual de cada técnica envolvida foi mantido, pois também foram indicados na visão destino.

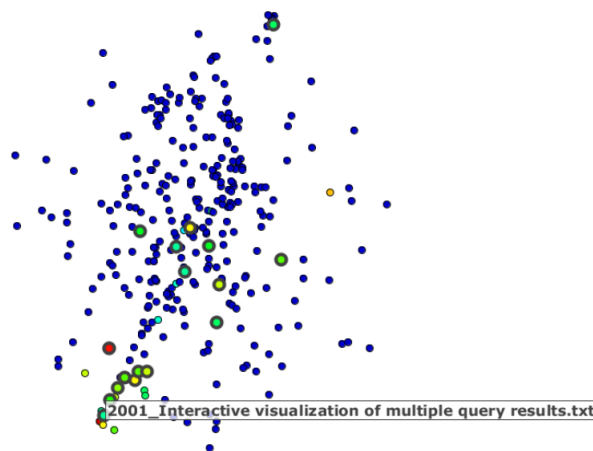
2.9 CONSIDERAÇÕES FINAIS

Na literatura a maioria das técnicas desenvolvem uma solução específica para o problema em questão, e não uma técnica genérica, ou seja, a solução é embarcada na necessidade do domínio do problema. E, muitas dessas, empregam a técnica mais simples, de associação direta por algum tipo de identificação, para realizar a coordenação entre as representações envolvidas no processo.

Muitos trabalhos abordam a coordenação sob uma perspectiva diferente, mais focada na interação e definição do *layout* da aplicação, assim como recentemente IGONIDERIGHA (2015). Nessa linha também NORTH e SHNEIDERMAN (1997), definiu uma taxonomia



(a) Seleção de instâncias na visão origem.



(b) Resultado da aplicação da técnica de coordenação.

Figura 10 – Exemplo de aplicação que implementa a combinação de coordenações.

Fonte: (ELER, 2011)

de múltiplas janelas coordenadas, focando na metodologia de interação, ação, navegação e propagação do efeito da coordenação nas visões. Não há um estudo na área que reúna as técnicas que realizam o mapeamento das instâncias através de um algoritmo de coordenação, que revele relacionamentos e permita a exploração efetiva dos dados.

As diversas técnicas existentes, podem lidar com apenas um conjunto de dados, ou vários conjuntos de dados, sejam eles do mesmo tipo ou de tipos distintos, podendo ser estruturados ou não estruturados, abordando dados unidimensionais, bidimensionais ou multidimensionais. Nesse sentido, cada uma das técnicas possuem suas vantagens e desvantagens de acordo com o contexto que é aplicada. Na seção, a ênfase foi dada em técnicas que lidam com dados multidimensionais, algumas para dados estruturados e outras para não estruturados, e aplicável para diversos domínios de dados, seja na área da pesquisa, mercado, indústria, medicina, política, entre outras.

Dentre as técnicas abordadas, é possível destacar que as baseadas em tópicos, tópicos no tempo e semântica são adequadas para dados textuais, sendo eficazes na exploração de grande conjuntos de dados. A técnica mais genérica é a baseada em identificador pois se

aplica para qualquer conjunto de dados que permita a vinculação única de um identificador para cada instância pertencente. Também a combinação de coordenadas também pode trazer benefícios para a análise pertinente ao problema enfrentado, porém a complexidade da solução aumenta consideravelmente, podendo comprometer o desempenho do resultado a ser alcançado.

3 SIMILARIDADE E DIVERSIDADE

3.1 SIMILARIDADE

Com o desenvolvimento da tecnologia, o tamanho, o volume e a dimensão das coleções de dados em geral, tem aumentado significativamente, principalmente com o avanço de técnicas para captura, processamento, armazenamento e transmissão dos dados, que buscam sempre acrescentar uma riqueza de detalhes, com o objetivo de sempre torná-los mais próximos da realidade. E, nesse contexto, recuperar esses dados eficientemente e rapidamente mediante a busca do usuário, tem se tornado uma tarefa extremamente importante em muitas áreas do cotidiano das pessoas.

As buscas convencionais realizadas em conjuntos de dados, sejam em arquivos ou sistemas de gerenciamento de banco de dados, são também efetuadas em dados convencionais, como dados textuais ou numéricos, nas quais é possível empregar a comparação de igualdade, menor ou maior. Porém para realizar buscas em dados complexos, como imagens, músicas, vídeos, dados geográficos, estruturas genéticas, textos longos, dados multidimensionais, séries temporais, entre outros, a busca convencional não se aplica, pois a descrição textual ou numérica não descreve a semântica do conteúdo do dado complexo, principalmente os que tem mais detalhes. É necessário aplicar uma busca mediante a características intrínsecas aos dados, que permita encontrar elementos mais parecidos possíveis com o dado ou conjunto que o usuário deseja comparar ([SERAPHIM, 2005](#); [LIU et al., 2007](#); [BUGATTI, 2008](#)).

A consulta por similaridade consiste na comparação do conjunto de dados com um ou mais elementos arbitrários, realizando a comparação entre eles, levando em consideração a informação inerente a cada um, realizando a consulta pelo conteúdo dos elementos, e não por algum tipo de descrição associada, como por exemplo, metadados. A comparação realizada, na maioria da vezes, é feita mediante a extração das características dos elementos, por meio de um extrator definido, formando um vetor de características associado a cada elemento, o qual é utilizado para comparação, no lugar do elemento. Com os vetores de características é possível medir a distância entre os elementos, utilizando funções de comparação, que medem o quanto os elementos são dissimilares entre si, gerando assim, um espaço de distâncias. E a busca nesse espaço, principalmente quando há uma grande quantidade de dados, é potencializada apoiando-se numa estrutura de indexação métrica ([RAZENTE, 2009](#)).

Os tipos de consultas por similaridade mais comuns existentes na literatura são por abrangência (*range query*) e pelos k-vizinhos mais próximos (*k-nearest neighbor query* -

kNN), nas quais a partir de um elemento central de consulta, a busca é realizada por meio da comparação entre os elementos no conjunto de dados alvo (RAZENTE, 2009; POLA, 2010).

A consulta por similaridade deve ser uma operação determinística, ou seja, se a base não mudar, o mesmo resultado deve ser retornado. A comparação também deve ser exata, em algumas abordagens a busca realizada é aproximada, mas a similaridade em espaços métricos é exata.

Em seguida, serão apresentados conceitos e definições pertencentes a similaridade.

3.1.1 ESPAÇO MÉTRICO

Domínios de dados que apresentam uma função de distância métrica são chamados de espaços métricos, eles englobam os espaços vetoriais e multidimensionais. Seja o espaço métrico $M = \langle \mathbb{S}, d \rangle$, onde $\mathbb{S}$ é o domínio de todos os elementos válidos, e d uma função de distância entre os elementos, também chamada de função métrica, que deve atender as seguintes propriedades, as quais permitem desenvolver consultas baseando-se na similaridade de maneira eficiente, onde $s_1, s_2, s_3 \in \mathbb{S}$ (SERAPHIM, 2005; POLA, 2010; SANTOS, 2012):

1. **Simetria:** $d(s_1, s_2) = d(s_2, s_1)$
2. **Identidade dos indiscerníveis:** $d(s_1, s_2) = 0$ então $s_1 = s_2$
3. **Não negatividade:** $0 < d(s_1, s_2) < \infty$ se $s_1 \neq s_2$
4. **Desigualdade triangular:** $d(s_1, s_2) \leq d(s_1, s_3) + d(s_3, s_2)$

Existem extratores de características genéricos e específicos, dependendo do tipo de dados, sendo eles os responsáveis em extrair valores que correspondem às características do objeto e formar o vetor de características. Para comparar dois vetores de características s_i e $s_j \in \mathbb{S}$, é necessário uma função de comparação $d(s_i, s_j)$, que meça o quanto os elementos são semelhantes entre si, mediante a distância, sendo que quanto menor a distância, mais semelhantes eles são, onde $d : \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{R}_+$. Se é possível avaliar a similaridade entre objetos a partir da função de distância métrica, então os dados podem ser considerados estar em um mesmo domínio métrico (BUENO, 2005; POLA, 2010).

As funções de distância são específicas para cada domínio de dados, e de maneira geral, é possível separá-las agrupando os domínios dos dados em cinco classes: espaços dimensionais, espaços de conjunto, espaços sequenciais, espaços de grafos e espaços de funções. Neste trabalho, o foco são as funções da classe de espaços dimensionais, como as funções euclidiana, do cosseno e mahalanobis (WILSON; MARTINEZ, 1997; SMITH, 1997; POLA, 2010).

A distância do cosseno, indicada na Equação 3.1, não considera a magnitude entre os vetores, e sim, a orientação espacial, calculando o ângulo entre eles centrados na origem do conjunto de dados. O coeficiente angular computado na função, limita-se entre os valores -1 e 1, sendo 1 quando há completa similaridade.

$$d_{\text{cosseno}}(s_i, s_j) = \frac{s_i \cdot s_j}{\|s_i\| \|s_j\|} \quad (3.1)$$

As funções de distância que pertencem a família Minkowski, definida pela Equação 3.2 para qualquer $p \geq 1$, são muito aplicadas em domínio de dados homogêneos, em que todas as dimensões possuem o mesmo peso no processo. Destacando a função euclidiana com $p = 2$, que é aplicada neste trabalho, a qual normalmente é utilizada para o cálculo de distância entre vetores, possuindo a vantagem de ser independente da origem do espaço do conjunto de dados, sendo que há completa similaridade quando o valor alcançado é 0.

$$d_{L_p}(s_i, s_j) = \sqrt[p]{\sum_{k=1}^n |s_{ik} - s_{jk}|^p} \quad (3.2)$$

Assumindo que a distância entre todos os elementos de um conjunto de dados, segue uma distribuição normal, a distância de mahalanobis, definida na Equação 3.3, realiza o cálculo utilizando a inversa da matriz de covariância dos vetores de elementos no conjunto de dados, sendo que a similaridade é considerada integral quando o seu valor é 0.

$$d_{\text{mahalanobis}}(s_i, s_j) = \sqrt{(s_i - s_j)^T S^{-1} (s_i - s_j)} \quad (3.3)$$

A função de distância mahalanobis, em geral, é utilizada para conjuntos de dados estáticos, devida complexidade do cálculo da matriz de covariância. E em contrapartida, as funções de distâncias euclidiana e cosseno são aplicáveis para conjuntos de dados dinâmicos, em que há uma constante requisição do cálculo de distância entre seus elementos.

3.1.2 COMPARAÇÃO E CONSULTA POR SIMILARIDADE

Considere um conjunto de elementos $S = s_1, s_2, s_3, \dots, s_n$ pertencente ao domínio $\mathbb{S}$ e uma função métrica d , as comparações para medir a similaridade entre os elementos podem ser realizadas pelos operadores θ de abrangência ou vizinhança. A comparação por abrangência $\theta = \text{Rng}(d, \xi_q)$, retorna verdadeiro, se o elemento s_i está à uma distância ξ_q do valor s_j , medindo a distância pela função d , isto é, $d(s_i, s_j) \leq \xi_q$. E a comparação por vizinhança $\theta = k\text{-NN}_S(d, k)$, retorna verdadeiro, se o valor s_i é um dos k mais próximos do valor s_j no espaço de busca S , com distância mensurada pela métrica d (BUENO, 2005; SERAPHIM, 2005; POLA, 2010).

Em domínios métricos, em geral as consultas por similaridade unárias são formadas por uma função de distância d e um elemento chamado centro de consulta s_q , sendo que existem basicamente dois operadores de seleção por similaridade: por abrangência (*Rng*), que considera os objetos que estejam até uma distância limite do objeto de referência, e pelos vizinhos mais próximos (*k-NN*) que limita os objetos do conjunto de resposta a um número máximo, os quais aplicam as consultas por similaridade citadas anteriormente, respectivamente (SERAPHIM, 2005; POLA, 2010).

A consulta por abrangência, definida pela Equação 3.4, recebe como parâmetros, um elemento do domínio de dados $s_q \in S$, e um grau de distância $\xi_q \geq 0$, e seleciona todos os objetos do conjunto de dados que diferem do elemento da consulta por no máximo a distância indicada, conforme é ilustrado na Figura 11. Ressaltando que ela permite, que na busca ao verificar um determinado elemento do conjunto de dados, já saiba se ele faz ou não, parte do conjunto de resposta, apresentando o resultado mais eficientemente (ZEZULA et al., 2006; RAZENTE, 2009; SANTOS, 2012).

$$range(s_q, \xi_q) = \{s_i | s_i \in S, d(s_i, s_q) \leq \xi_q\} \quad (3.4)$$

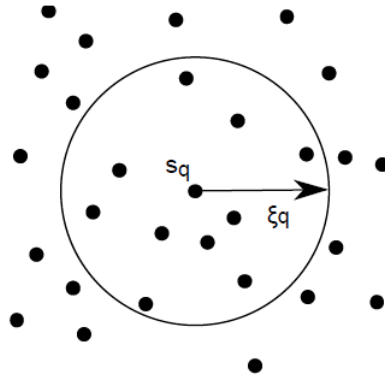


Figura 11 – Busca por abrangência.

Fonte: (SANTOS, 2012)

A consulta aos k -vizinhos mais próximos, definida pela Equação 3.5, recebe como parâmetros, um elemento do domínio de dados $s_q \in S$, e o valor $k > 0$, selecionando assim, os k -elementos da base de dados que estejam mais próximos do elemento da consulta, conforme é ilustrado na Figura 12, na qual é realizada a busca pelos seis vizinhos mais próximos. A seleção por vizinhança pode conter alguns parâmetros extras, como o de inclusividade e empate. O parâmetro de inclusividade permite definir se o elemento central da consulta pode ou não estar no conjunto de resposta, e o de empate define que, se houver empate de distância na busca, que faça exceder o valor k do conjunto de resposta, retorne todos os elementos encontrados ou apenas os k indicados pelo usuário previamente (SERAPHIM, 2005; ZEZULA et al., 2006; RAZENTE, 2009; SANTOS, 2012).

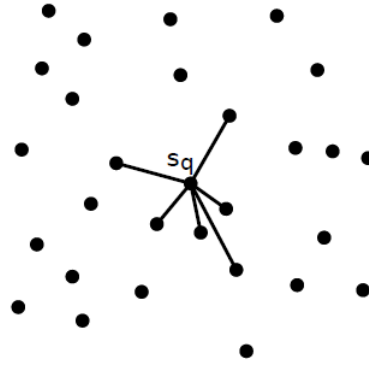


Figura 12 – Busca pelos k-vizinhos mais próximos.

Fonte: (SANTOS, 2012)

$$k\text{-}NN(s_q, k) = R \subset S, |R| = k, \forall s_j \in R, s_i \in S - R, d(s_q, s_j) \leq d(s_q, s_i) \quad (3.5)$$

Em alguns casos, quando não há conhecimento prévio do domínio de dados, a consulta por abrangência pode não ser útil para resolver o problema, pois o usuário pode não conhecer uma distância limite eficiente para utilizar na busca, então a consulta pelos k-vizinhos mais próximos vai retornar um resultado, independentemente do conhecimento das distâncias entre os elementos.

As duas consultas de seleção por similaridade, podem ser executadas a partir de uma busca sequencial entre todos os elementos do conjunto de dados alvo, por exemplo, na busca por abrangência, se o elemento estiver dentro do limite de distância ao centro de consulta, ele é adicionado ao conjunto de resposta R , sendo que o elemento centro de consulta s_q , não precisa fazer parte da coleção de documentos que serão consultados, porém ele deve permanecer no domínio métrico que define o conjunto de dados S (SERAPHIM, 2005).

Existem diversas variações desses operadores de seleção por similaridade, como por exemplo, realizar a comparação pelos k-vizinhos mais distantes (*k-farthest neighbor query k-FNQ*), a consulta pontual (*point query PQ*), equivalente à consulta por abrangência com distância nula, que retorna o próprio elemento central da consulta, se ele estiver na base. E o operador de abrangência reversa, que retorna todos os elementos que estão fora da distância indicada ξ_q , entre outras na literatura (BUENO, 2005; ZEZULA et al., 2006; RAZENTE, 2009).

3.1.3 MÉTODOS DE ACESSO MÉTRICO

A partir do domínio métrico, definido por seus dados e uma função de distância, no processo de recuperação de dados para uma dada consulta por similaridade, é necessário

mecanismos para indexar as características dos elementos, previamente extraídas, as quais irão representá-los. De acordo com a cardinalidade do conjunto de dados, ou se a função métrica for muito custosa, o tempo para obtenção da resposta pode não ser satisfatório, sendo necessário aplicar técnicas que melhorem o tempo para retorno da busca. O desempenho da similaridade é medido essencialmente por dois fatores importantes: a quantidade de acesso a disco e a quantidade de comparações entre os elementos, através dos cálculos realizados por meio da função métrica (ZEZULA et al., 2006; RAZENTE, 2009).

Em um contexto do domínio de dados onde não há relação de ordem, os Métodos de Acesso Métrico (*Metric Access Methods* MAM) são estruturas especializadas para indexar dados em espaços métricos, com o objetivo de reduzir o número de cálculos de distâncias e o número de acesso a disco, proporcionando acesso eficiente aos dados e recuperação rápida de uma dada informação armazenada, por meio da consulta por similaridade, na qual são exploradas as propriedades métricas, como não-negatividade, simetria e desigualdade triangular. Para a construção da estrutura de indexação, são calculadas as distâncias entre os elementos a partir das características extraídas dos dados complexos, e então elas são utilizadas para definir a estrutura de índices, que fornecem suporte a consultas por similaridade, oferecendo eficiência para consultas de dados de alta dimensão, complexos e de grande volume (SERAPHIM, 2005; RAZENTE, 2009).

Diversas técnicas de indexação têm sido propostas para melhorar o desempenho de consultas por similaridade, pois ter o apoio das estruturas de indexação é muito importante para podar operações de comparações, e potencializar a busca do usuário. Existem estruturas de indexação métricas baseadas em memória, denominadas dinâmicas, como gh-tree e vp-tree, baseadas em disco, designadas estáticas, como mvp-tree e vp-tree modificada, e baseadas em disco, mas também dinâmicas, como m-tree e slim-tree (BUGATTI, 2008; RAZENTE, 2009; POLA, 2010), sendo que uma revisão geral da maior parte dos métodos, pode ser encontrada em HJALTASON e SAMET (2003), ZEZULA et al. (2006).

3.2 DIVERSIDADE

Com o crescente volume de dados atualmente, é fundamental o desenvolvimento de mecanismos que auxiliam a encontrar a informação desejada. Desse modo, a busca baseada em similaridade de conteúdo tem se popularizado para a recuperação de dados complexos, sendo que a abordagem mais empregada é a consulta pelos k -vizinhos mais próximos ao centro da consulta, pois apresenta resultados efetivos ao usuário, sem a necessidade do conhecimento prévio do conjunto de dados. A consulta pelos k -vizinhos mais próximos, busca k elementos mais similares ao centro de consulta s_q , levando em consideração apenas

a relevância do elemento no conjunto de dados, em relação o centro de consulta, não avaliando a relação entre os elementos presente no conjunto de resposta (JAIN; SARDA; R., 2004).

Em conjuntos de dados que há uma alta redundância parcial ou total dos dados, o resultado da busca por vizinhança pode conter elementos muito similares entre si, não apresentando uma resposta relevante ao usuário, gerando uma frustração na pesquisa e consequente abandono da consulta. A propriedade diversidade é ferramenta muito eficaz quando o usuário realiza consultas ambíguas, principalmente em bases com alto volume de dados, pois recupera o conjunto de elementos mais próximos ao elemento central de consulta e ao mesmo tempo diverso entre si, gerando um conjunto de resposta mais heterogêneo e relevante para o usuário (ZHENG et al., 2017). A diversidade surge em um cenário onde se utiliza dados complexos ou poucos termos na busca convencional, e frequentemente a intenção do usuário está implícita, nesse sentido, o resultado apresentado ao usuário precisa apresentar elementos que são mais parecidos ao centro de consulta, mas também que carregam informação significativa e expressiva, focando realmente na relevância da informação e satisfação do usuário, não somente na acurácia da consulta em relação a similaridade (CARBONELL; GOLDSTEIN, 1998; DENG; FAN, 2014).

A diversificação em consultas por similaridade, implica em uma troca entre resultados que seguem com exatidão aquilo que o usuário introduziu como parâmetro para a busca, e resultados com elementos que possuem algum nível de diferença entre si, ou seja, conceitos que são competitivos, caracterizando o problema como uma otimização de bi-critério (VIEIRA et al., 2011a; SANTOS et al., 2015).

No trabalho de GOLLAPUDI e SHARMA (2009) é realizada uma definição axiomática para diversificação de resultados, um conjunto de propriedades que qualquer sistema de diversificação deveria satisfazer, constituído por: invariância de escala na função métrica, consistência, riqueza, estabilidade, independência dos atributos irrelevantes, monotonicidade, força da relevância e força da similaridade. É provado que não há uma função de diversificação que satisfaça todos eles, sendo que a escolha da função deve ser realizada em relação ao contexto particular da solução.

DROSOU e PITOURA (2010) apresenta uma classificação e estudo comparativo de várias definições, de algoritmos e métricas para diversificação de resultados de buscas. Há diferentes abordagens do emprego da diversidade no contexto atual, podendo ser baseada em conteúdo, que avalia a dissimilaridade entre os elementos, novidade, verificando elementos que contêm novas informações quando comparados aos encontrados anteriormente, e cobertura, analisando os elementos que pertencem a diferentes categorias. Em algumas soluções, pode haver combinação das abordagens, sendo que no trabalho a abordagem utilizada é a baseada em conteúdo, seguindo o embasamento do conceito de similaridade.

A maioria dos algoritmos de diversificação podem ser classificados em dois grupos

principais: abordagens gulosas, que sempre escolhem a melhor solução no atual momento do processamento, e a baseada em troca, que busca alcançar o melhor resultado realizando trocas de acordo com a definição. Porém, em algumas metodologias gulosas são aplicadas técnicas de heurística e randomização, para aumentar a qualidade do resultado, visto que algumas abordagens priorizam a qualidade e outras a rapidez no processamento (DRO-SOU; PITOURA, 2010). A estratégia geral amplamente utilizada consiste em utilizar algoritmos gulosos com construção incremental para encontrar respostas determinísticas em comparações aos k -vizinhos mais próximos, e o grande desafio é encontrar um *tradeoff* entre a similaridade e diversidade, para que o usuário obtenha um conjunto de resposta relevante e que atenda as suas expectativas (VIEIRA et al., 2011b).

Inicialmente os sistemas gerenciadores de banco de dados (SGBD) não suportavam diretamente consultas aproximadas, sendo necessária a implementação desse tipo de pesquisa em ferramentas externas. Como por exemplo o PgSimilar (BORGES; DORNELLES, 2006), que é um módulo que contém ferramentas para anexar funcionalidades ao PostgreSQL, que tratam vários tipos de similaridade entre *strings*. Atualmente os SGBDs implementam suas próprias extensões que dão suporte a consultas aproximadas, como o fuzzystrmatch do PostgreSQL e UTL_MATCH da Oracle, que permitem comparar textos por similaridade, e o interMedia da Oracle, que possibilita o armazenamento, gerenciamento e recuperação de imagens, áudio, vídeos, entre outros dados, mediante o conteúdo (GROUP, 2019; ORACLE, 2019).

A diversificação de resultados nas consultas pelos k -vizinhos mais próximos no âmbito geral, tem sido empregada em diversas áreas do conhecimento e tem uma grande variedade de aplicações, como em conjunto de dados na recuperação da informação (VIEIRA et al., 2011a; DIAS et al., 2017), pesquisas *web* (GRIFFITH; PFEIFER, 2008), sistemas de recomendação e propagandas patrocinadas (YU; LAKSHMANAN; AMER-YAHIA, 2009), medicina (DIAS; BUENO; RIBEIRO, 2013), pesquisas operacionais, finanças, entre outras (DENG; FAN, 2014).

De acordo com a metodologia utilizada, diversidade pode ser utilizada como uma pré-etape, para encontrar tópicos de conteúdo, e então ao encontrar o assunto desejado, o usuário possa fazer uma pesquisa mais específica por meio da similaridade (SANTOS et al., 2015).

Existem dois grupos de estratégias para a diversificação de resultados, a baseada em atributos pertinentes aos dados e a baseada apenas na computação de distância entre os elementos, de acordo com seus conteúdos. A mais conhecida é a baseada em atributos, que compara o quanto cada valor do atributo que compõe a descrição dos dados, difere de outros no conjunto de resposta. Neste trabalho, o foco é a diversidade baseada na computação de distância entre os elementos, que é modelada apenas utilizando a representação dos elementos de dados no espaço métrico e a distância entre eles, aplicando-se em dados

complexos (SANTOS, 2012). A seguir são apresentados os principais métodos na literatura, que implementam consultas por similaridade aos k -vizinhos mais próximos aplicando a diversidade.

3.2.1 MAXIMAL MARGINAL RELEVANCE - MMR

Uma proposta precursora, que consiste na combinação linear de similaridade e diversidade, denominada de *Maximal Marginal Relevance* (MMR - Relevância Marginal Máxima). Um elemento tem uma alta relevância marginal se for tão similar ao elemento de consulta quanto diverso aos elementos selecionados previamente, que fazem parte do conjunto de resposta R (CARBONELL; GOLDSTEIN, 1998).

O processo do método é descrito no Algoritmo 2, onde seja $S = s_1, s_2, \dots, s_n$, um conjunto de dados com n elementos, s_q o elemento central de consulta e k ($k \leq n$), a quantidade de elementos no conjunto de resposta. A medida de similaridade é indicada por δ_{sim} , e a medida de diversidade por δ_{div} , sendo que ambos os conceitos associados às medidas, são definidos por uma mesma função de distância ou dissimilaridade arbitrária. O método constrói incrementalmente o conjunto de resposta $R \subseteq S$, selecionando k elementos do conjunto S , que maximizem a função objetivo, definida pela Equação 3.6 (VIEIRA et al., 2011b; SANTOS, 2012).

$$mmr(s_i) = (1 - \lambda)\delta_{sim}(s_i, s_q) + 2\lambda \sum_{r_j \in R} \delta_{div}(s_i, r_j) \quad (3.6)$$

Algoritmo 2: *Maximal Marginal Relevance* MMR

```

/* Entrada: Conjunto candidato  $S$  e o tamanho  $k$  do conjunto
   resultante                                     */
/* Saída: Conjunto resultante  $R \subseteq S$ ,  $|R| = k$           */
1  $R \leftarrow \emptyset$ 
2  $s_s \leftarrow \operatorname{argmax}_{s_i \in S}(mmr(s_i))$ 
3  $S \leftarrow S \setminus s_s$ 
4  $R \leftarrow s_s$ 
5 enquanto  $|R| < k$  faça
6    $s_s \leftarrow \operatorname{argmax}_{s_i \in S}(mmr(s_i))$ 
7    $S \leftarrow S \setminus s_s$ 
8    $R \leftarrow R \cup \{s_s\}$ 
9 fim

```

O parâmetro λ que varia entre 0 e 1, representa o limiar entre a similaridade e diversidade. Quando $\lambda = 0$ a diversidade é mínima, e quando $\lambda = 1$, a máxima diversidade

é computada, quando está em um valor intermediário, uma combinação linear de ambos os critérios é otimizada, tornando o problema NP-difícil.

O critério *mmr* visa reduzir a redundância geral, mantendo a relevância da consulta dos elementos recuperados por meio da diversidade. Mesmo sendo um dos métodos com tempo de processamento melhor, é notável que o primeiro elemento tem um impacto muito grande na qualidade da construção de R , pois a função de maximização *mmr* é calculada a partir dos elementos já previamente inseridos no conjunto resultante R (SANTOS, 2012).

3.2.2 MAX-SUM DISPERSION - MSD

A técnica *Max-Sum Dispersion* (MSD - Dispersão Max-Sum) incrementalmente busca escolher pares de elementos que maximize a Equação 3.7, que atenda a similaridade de acordo com o elemento central de consulta, mas também que mantenha a diversidade do conjunto de resposta R , de acordo com o fator λ definido inicialmente. Conforme descrito no Algoritmo 3, a cada iteração são escolhidos dois elementos no conjunto de dados $s_i, s_j \in S$ que maximizem a função objetivo *msd*, e caso seja escolhido um k ímpar, na última iteração, o método escolhe aleatoriamente qualquer elemento em S como parte da resposta (GOLLAPUDI; SHARMA, 2009; VIEIRA et al., 2011b; SANTOS, 2012).

$$msd(s_i, s_j) = (1 - \lambda)(\delta_{sim}(s_q, s_i) + \delta_{sim}(s_q, s_j)) + 2\lambda\delta_{div}(s_i, s_j) \quad (3.7)$$

Algoritmo 3: *Max-Sum Dispersion* MSD

```

/* Entrada: Conjunto candidato  $S$  e o tamanho  $k$  do conjunto
   resultante                                     */
/* Saída: Conjunto resultante  $R \subseteq S$ ,  $|R| = k$           */
1  $R \leftarrow \emptyset$ 
2 enquanto  $|R| < k/2$  faça
3    $s_i, s_j \leftarrow \operatorname{argmax}_{s_i, s_j \in S} (msd(s_i, s_j))$ 
4    $R \leftarrow R \cup \{s_i, s_j\}$ 
5    $S \leftarrow S \setminus \{s_i, s_j\}$ 
6 fim
7 se  $k$  é ímpar então
8   Escolha um elemento aleatório  $s_s \in S$ 
9    $R \leftarrow R \cup \{s_s\}$ 
10 fimse

```

O método avalia a diversidade dos elementos no momento da escolha deles, não verificando o impacto que causa na inserção do conjunto de resposta R , ou seja, se são diversos entre os elementos já presentes em R .

3.2.3 SWAP

O método *Swap* (Método de Troca) é definido por um processo simples, e consiste em duas etapas, descritas no Algoritmo 4. Na primeira, são escolhidos os k elementos do conjunto S mais similares ao elemento central de consulta, e o restante dos elementos do conjunto são ordenados de acordo com a similaridade ao centro de consulta, em ordem decrescente. Então na segunda etapa, cada elemento restante no conjunto S é testado, verificando se a remoção de um dos elementos do conjunto resultante R pelo elemento em análise do conjunto S , melhora a função objetivo $\mathcal{F}$, apresentada na Equação 3.9, de acordo com a Definição de 3.8. Se houver uma melhora, então é realizada a troca e alteração do conjunto de resposta R (YU; LAKSHMANAN; AMER-YAHIA, 2009; VIEIRA et al., 2011b).

$$R = \operatorname{argmax} \mathcal{F}(s_q, S'), \quad S' \subseteq S, \quad k = |S'| \quad (3.8)$$

$$\mathcal{F}(s_q, S') = (k - 1)(1 - \lambda) \cdot \delta_{sim}(s_q, S') + 2\lambda \cdot \delta_{div}(S') \quad (3.9)$$

O valor alcançado de $\mathcal{F}$, obtido pelo conjunto de resposta R não é garantido ser ótimo, pois os elementos do conjunto S , são analisados de acordo com a ordem crescente definida pela similaridade de cada um, em relação ao centro de consulta, não levando em consideração a diversidade no processo de ordenação.

3.2.4 MOTLEY

Como o método MMR, o método Motley também constrói o conjunto de resposta R incrementalmente, sendo que inicialmente ordena todos os elementos do conjunto S em ordem crescente de acordo com o valor de similaridade de cada elemento em relação ao centro de consulta, conforme o Algoritmo 5. O primeiro elemento a ser inserido no conjunto R é o elemento mais similar ao centro de consulta, e em seguida é analisado cada elemento mais similar em S , e que mantenha a diversidade do conjunto R de acordo com um limiar θ definido previamente, se caso não atender as exigências o elemento é descartado (JAIN; SARDA; R., 2004).

O parâmetro θ define um limiar para a diversidade esperada no conjunto de resposta, uma função booleana filtra os elementos escolhidos para o conjunto de resposta, se o elemento possuir o valor de diversidade maior que o limiar, então ele é adicionado ao conjunto de resposta definitivo. Para consultas de elementos em um espaço de dados com domínio medidas de distância normalizadas entre $[0,1]$, o valor de *theta* varia também entre 0 e 1 também, sendo que sua interpretação métrica é o valor de quanto dois elementos são diversos entre si (VIEIRA et al., 2011b; SANTOS, 2012).

Algoritmo 4: *Swap*.

```

/* Entrada: Conjunto candidato  $S$  e o tamanho  $k$  do conjunto
   resultante */
/* Saída: Conjunto resultante  $R \subseteq S$ ,  $|R| = k$  */
1  $R \leftarrow \emptyset$ 
2 enquanto  $|R| < k$  faça
3    $s_s \leftarrow \operatorname{argmax}_{s_i \in S} (\delta_{sim}(s_q, s_i))$ 
4    $S \leftarrow S \setminus s_s$ 
5    $R \leftarrow R \cup \{s_s\}$ 
6 fim
7 enquanto  $|S| > 0$  faça
8    $s_s \leftarrow \operatorname{argmax}_{s_i \in S} (\delta_{sim}(s_q, s_i))$ 
9    $S \leftarrow S \setminus s_s$ 
10   $R' \leftarrow R$ 
11  para cada  $s_j \in R$  faça
12    se  $\mathcal{F}(s_q, R \setminus s_j \cup \{s_s\}) > \mathcal{F}(s_q, R')$  então
13       $R' \leftarrow R \setminus s_j \cup \{s_s\}$ 
14    fimse
15  fim
16  se  $\mathcal{F}(s_q, R') > \mathcal{F}(s_q, R)$  então
17     $R \leftarrow R'$ 
18  fimse
19 fim

```

Experimentos demonstraram que o Motley pode fornecer soluções diversas de alta qualidade a um baixo custo em termos de distância de resultados e tempo de processamento, porém, como também o método MMR, a abordagem tem seus resultados limitados a escolha do primeiro elemento do conjunto de respostas, pois depois da inserção do primeiro elemento, a escolha dos próximos é feita buscando manter o limiar θ superior ao definido previamente.

3.2.5 GREEDY MARGINAL CONTRIBUTION - GMC

O método *Greedy Marginal Contribution* (GMC - Contribuição Marginal Gananciosa), apresentado no Algoritmo 6, constrói incrementalmente o conjunto de resposta R selecionando o elemento que maximize a função, definida pela Equação 3.10. A cada iteração, o GMC faz um *ranking* de elementos candidatos em S por meio da função de avaliação de contribuição para maximização do *mmc*, que considera tanto a contribuição do elemento $s_i \in S$, utilizando a sua similaridade em relação ao centro de consulta s_q ,

Algoritmo 5: Motley.

```

/* Entrada: Conjunto candidato  $S$  e o tamanho  $k$  do conjunto
   resultante */
/* Saída: Conjunto resultante  $R \subseteq S$ ,  $|R| = k$  */
1  $s_s \leftarrow \operatorname{argmax}_{s_i \in S}(\delta_{sim}(s_q, s_i))$ 
2  $S \leftarrow S \setminus s_s$ 
3  $R \leftarrow s_s$ 
4 enquanto  $|R| < k$  faça
5    $s_s \leftarrow \operatorname{argmax}_{s_i \in S}(\delta_{sim}(s_q, s_i))$ 
6    $S \leftarrow S \setminus s_s$ 
7   se  $\delta_{div}(s_r, s_s) \geq \theta, \forall s_r \in R$  então
8      $R \leftarrow R \cup \{s_s\}$ 
9   fimse
10 fim

```

quanto os elementos já inseridos na resposta $R_{p-1}(\delta_{div})$, e os elementos que não foram inseridos no conjunto de resposta $s_j \in S - s_i(\delta_{div}^l)$ (VIEIRA et al., 2011b; SANTOS, 2012).

$$mmc(s_i) = (1 - \lambda) \cdot \delta_{sim}(s_i, s_q) + 2\lambda \sum_{s_j \in R_{p-1}} \delta_{div}(s_i, s_j) + 2\lambda \sum_{l=1}^{l \leq k-p} \delta_{div}^l(s_i, s_j) \quad (3.10)$$

Algoritmo 6: Greedy Marginal Contribution GMC

```

/* Entrada: Conjunto  $S$  e o tamanho  $k$  do conjunto resultante */
/* Saída: Conjunto resultante  $R \subseteq S$ ,  $|R| = k$  */
1  $R \leftarrow 0$ 
2 para  $p \leftarrow 1$  até  $p = k$  faça
3    $s_s \leftarrow \operatorname{argmax}_{s_i \in S}(mmc(s_i))$ 
4    $R_p \leftarrow R_{p-1} \cup \{s_s\}$ 
5 fim
6  $R \leftarrow R_p$ 

```

A abordagem do GMC é muito parecida com a do MMR, sempre escolhendo o elemento que melhor contribui para a resposta, porém o GMC aplica o *mmc* como função de avaliação para escolher o elemento que melhor compõe o conjunto de resposta. Há um benefício no desempenho ao utilizar a função de maximização *mmc*, pois o primeiro elemento do conjunto de resposta R , é escolhido mediante a sua similaridade ao centro de consulta, e também ao relacionamento de diversidade com todos os outros elementos em S , fornecendo uma alta relevância ao centro de consulta (SANTOS, 2012).

3.2.6 GREEDY RANDOMIZED WITH NEIGHBORHOOD EXPANSION - GNE

Na abordagem *Greedy Randomized with Neighborhood Expansion* (GNE), definida no Algoritmo 7, é realizada uma adaptação da meta heurística *Greedy Randomized Adaptive Search Procedure* (GRASP), sendo dividida em duas etapas: a Construção e Busca Local. Diferente da técnica GMC que sempre seleciona o elemento que maximiza a função *mmc* a cada iteração, a técnica GNE escolhe um elemento randômico na listagem com os melhores candidatos, de acordo com o parâmetro α .

Na etapa de construção, conforme o Algoritmo 8, é formada a lista de candidatos *RCL* com k elementos, baseando-se na contribuição associada a sua inclusão na solução, segundo a Equação 3.10, e ajustando o parâmetro α , que varia entre 0 e 1, para definir o quanto gananciosa ou randômica a resposta deve ser, destacando-se que quanto mais perto de 1 mais randômica a solução será (SANTOS, 2012).

A partir do conjunto de resposta provido na fase de construção, a busca local, indicada no Algoritmo 9, busca melhorar gradualmente a resposta, aplicando uma série de modificações locais na vizinhança da solução atual, realizando a troca desses elementos, e verificando se ocorre uma melhora na solução final pela Equação 3.9, com o objetivo de encontrar a solução ótima local (VIEIRA et al., 2011b).

Algoritmo 7: *Greedy Randomized with Neighborhood Expansion* GNE

```

/* Entrada: Conjunto  $S$  e o tamanho  $k$  do conjunto resultante      */
/* Saída: Conjunto resultante  $R \subseteq S$ ,  $|R| = k$                       */
1  $R \leftarrow \emptyset$ 
2 para  $i \leftarrow 0$  até  $i < i_{max}$ ,  $i \leftarrow i + 1$  faça
3    $R' \leftarrow FaseConstrução()$ 
4    $R' \leftarrow FaseBuscaLocal(R')$ 
5   se  $\mathcal{F}(s_q, R') > \mathcal{F}(s_q, R)$  então
6      $R \leftarrow R'$ 
7   fimse
8 fim

```

A técnica de diversificação GNE, dentre todas as outras, é a que possui maior complexidade computacional, por envolver diversas etapas na construção da solução, principalmente quando se aplica em um conjunto de dados com um alto número de instâncias. E além da definição do parâmetro λ , que tem como finalidade ponderar a similaridade e diversidade no cálculo da função *mmc*, é necessário definir o parâmetro α , que define o quão abrangente, mediante a randomização, será a busca dos elementos que constituem o conjunto de resposta.

Algoritmo 8: Fase Construção GNE

```

/* Saída: Conjunto resultante  $R \subseteq S$ ,  $|R| = k$  */
1  $R \leftarrow \emptyset$ 
2 para  $p \leftarrow 1$  até  $p = k$  faça
3    $s_{max} \leftarrow \operatorname{argmax}_{s_i \in S}(mmc(s_i))$ 
4    $s_{min} \leftarrow \operatorname{argmin}_{s_i \in S}(mmc(s_i))$ 
5    $RCL \leftarrow \{s_i \in S \mid mmc(s_i) \geq s_{max} - \alpha(s_{max} - s_{min})\}$ 
6    $s_i \leftarrow \operatorname{random}(RCL)$ 
7    $R_p \leftarrow R_{p-1} \cup \{s_i\}$ 
8    $S \leftarrow S \setminus s_i$ 
9 fim
10  $R \leftarrow R_p$ 

```

3.2.7 BETTER RESULTS WITH INFLUENCE DIVERSIFICATION - BRID

Todas as abordagens descritas anteriormente utilizam um parâmetro limiar para definir o quão diverso um conjunto de resposta deve ser, obtido por meio da consulta por similaridade. Definir um limiar que se ajuste à necessidade do usuário na busca, leva em consideração muitos fatores, como o conjunto de dados, o tipo de característica que se extrai dos dados, entre outros, ou seja, é preciso ter algum conhecimento prévio da base de dados, para obter uma resposta relevante da consulta, de acordo com a necessidade em questão.

O modelo de variedade, definido por SANTOS (2012), que segue as seguintes propriedades, consiste em inserir diversidade nas consultas por similaridade utilizando apenas especificações básicas do espaço métrico, sem a necessidade da definição de parâmetros prévios:

- **Independência de parâmetros** : não requer nenhum parâmetro de configuração;
- **Independência do domínio** : depende apenas que o conjunto de dados em que se aplica a busca atenda às propriedades de um espaço métrico;
- **Independência de operação de busca** : aplica-se a qualquer busca que utiliza o conceito de similaridade.

A estratégia da técnica parte do princípio de que, elementos na base de dados que são muito parecidos entre si não agregam relevância a resposta se estiverem juntos na mesma, garantindo isso com base na definição dos conceitos de elemento dominante e subconjunto dominado por ele. Nesse sentido, seja S um conjunto de dados, a partir do centro de consulta s_q e uma distância de similaridade arbitrária ξ_p , s_p é considerado o seu

Algoritmo 9: Fase Busca Local GNE

```

/* Entrada: Conjunto candidato à solução  $R$  e o tamanho  $k$  do
   conjunto resultante */
/* Saída: Conjunto resultante  $R \subseteq S$ ,  $|R| = k$  */
1  $s_{max} \leftarrow \operatorname{argmax}_{s_i \in S}(mmc(s_i))$ 
2  $s_{min} \leftarrow \operatorname{argmin}_{s_i \in S}(mmc(s_i))$ 
3  $RCL \leftarrow s_i \in S | mmc(s_i) \geq s_{max} - \alpha(s_{max} - s_{min})$ 
4 para cada  $s_i \in R$  faça
5    $R' \leftarrow R$ 
6   para cada  $s_j \in R$ ,  $s_j \neq s_i$  faça
7     para  $l \leftarrow 1$  até  $l = k - 1$  faça
8        $s_i^l \leftarrow s'_i \in S | \delta_{div}^l(s_i, s'_i)$ 
9       se  $s_i^l \notin R$  então
10         $R'' \leftarrow R' - s_j + s_i^l$ 
11        se  $\mathcal{F}(s_q, R'') > \mathcal{F}(s_q, R')$  então
12           $R' \leftarrow R''$ 
13        fimse
14      fimse
15    fim
16  fim
17  se  $\mathcal{F}(s_q, R') > \mathcal{F}(s_q, R)$  então
18     $R \leftarrow R'$ 
19  fimse
20 fim
21

```

elemento mais próximo no conjunto S . Então, todos os elementos mais distantes do que ξ_p de s_q , porém próximos uns dos outros, são considerados similares entre si, e s_p é definido por elemento dominante para a distância de similaridade ξ_p , por dominar todos os outros $s_i \in S$ com distância menor que ξ_p ao elemento s_p . E com base nisso, $\check{S}$ é estabelecido por um subconjunto de similaridade em S induzido pelo elemento dominante s_p , o qual é definido na Equação 3.11.

$$\check{S}(s_p, s_q) = s_i \in S | (\delta_{sim}(s_i, s_p) \leq \delta_{sim}(s_p, s_q)) \wedge (\delta_{sim}(s_i, s_p) \leq \delta_{sim}(s_i, s_q)) \quad (3.11)$$

Na abordagem baseada em variedade não é necessário ajustar um limiar de diversidade na busca por similaridade pelos k vizinhos. A busca é realizada inicialmente pelos

mais similares, e uma verificação é realizada por meio da Equação 3.11, a qual não permite incluir elementos tão similares entre si, a tal ponto de não agregar informação alguma para o usuário, sendo definida na Equação 3.12.

$$R = \{r_i \in S | \forall r_j \in R, r_i \neq r_j \Rightarrow r_i \notin \check{S}(r_j, s_q)\} \quad (3.12)$$

A técnica de diversidade *Better Results With Influence Diversification* (BRID), desenvolvida por SANTOS et al. (2013), realiza a consulta aos k -vizinhos mais próximos aplicando os conceitos de variedade, como é descrito no Algoritmo 10.

Algoritmo 10: *Better Results With Influence Diversification* BRID

```

/* Entrada: Conjunto candidato  $S$  ordenado por similaridade em
   relação ao  $s_q$ , e o tamanho  $k$  do conjunto resultante */
/* Saída: Conjunto resultante  $R \subseteq S$ ,  $|R| = k$  */
1  $s_t \leftarrow \operatorname{argmax}(\delta_{sim}(s_q))$ 
2  $R \leftarrow s_t$ 
3 enquanto  $|R| < k$  ou  $S$  não está vazio faça
4    $s_t \leftarrow \operatorname{argmax}(\delta_{sim}(s_q))$ 
5   // marque  $s_t$  como elemento dominante
6   para cada  $r_j \in R$  faça
7     se  $s_t \in \check{S}(r_j, s_q)$  então
8        $s_t$  é descartado, e marcado como não dominante
9       break
10    fimse
11  fim
12   $\xi_p \leftarrow \delta_{sim}(s_t, s_q)$ 
13  se  $s_t$  é dominante então
14     $R \leftarrow s_t$ 
15  fimse
16 fim

```

Mesmo em uma base de dados onde haja uma grande quantidade de dados e alta redundância, o emprego da definição de variedade nas consultas por similaridade fazem com que os k elementos presentes no conjunto de resposta sejam significativos ao usuário, evitando uma busca exaustiva, e até o abandono da pesquisa em alguns casos.

3.3 CONSIDERAÇÕES FINAIS

A consulta por similaridade pelos k -vizinhos mais próximos é a mais utilizada em diversas soluções na literatura, pois o usuário não precisa ter um conhecimento prévio do espaço métrico formado pelos elementos e função de distância aplicada. Em contrapartida, a consulta por abrangência requer a definição prévia de uma distância limite para recuperação dos elementos, não sendo intuitivo ao usuário.

O usuário nem sempre sabe o que deseja procurar. E, em algumas situações, quando ele sabe o que procurar, não consegue expressar o que anseia encontrar, principalmente quando a busca é aplicada a dados complexos. Nesse cenário, a diversidade entra como uma ferramenta na pesquisa baseada em similaridade, com o objetivo de aumentar a relevância das respostas, e reduzir o distanciamento entre a intenção do usuário, e a busca realizada por ele.

Quando há pouca incidência de dados similares no conjunto de dados onde se aplica a busca, a pura acurácia fornecida por métodos de similaridade são apropriadas, porém para um conjunto com alta redundância de dados, contendo parcialmente ou totalmente informação duplicada, é necessário utilizar algo que construa um conjunto de resposta relevante e que contenha informações substanciais para apresentar ao usuário. E nesse sentido, a diversidade é uma solução necessária que surge diante dos avanços das tecnologias e alta produção de dados, que fazem parte da vida das pessoas (CARBONELL; GOLDSTEIN, 1998).

Existem diversas técnicas de diversidade, e ZHENG et al. (2017) apresenta um estudo que explora a abordagem, desafios, desempenho e aplicabilidade das principais técnicas atualmente, sendo que a maioria das abordagens baseadas em conteúdo, aplicada à consultas por similaridade pelos k -vizinhos mais próximos são abordadas neste trabalho.

A similaridade e diversidade são propriedades concorrentes, e os algoritmos que empregam ambas tentam encontrar um *tradeoff*, para equilibrar e formar um conjunto similar e diverso, porém o usuário pode definir o que é mais importante na consulta por meio de parâmetros, gerando uma flexibilidade que permite construir resultados de consultas mais significativos para o contexto da solução (GOLLAPUDI; SHARMA, 2009).

4 COORDENAÇÕES BASEADAS EM SIMILARIDADE E DIVERSIDADE

4.1 CONSIDERAÇÕES INICIAIS

Como apresentado no Capítulo 2, foram propostas na literatura diferentes técnicas de coordenação aplicadas em mapas de similaridade, para auxiliar o usuário a compreender a relação entre as instâncias de dados, visualizados em diferentes representações gráficas. Em alguns casos, o usuário pode ter a intenção de buscar elementos que não sejam tão parecidos entre si, que tenham uma maior relevância e representatividade. Nesse sentido, há uma limitação nas consultas por similaridade pelos k -vizinhos mais próximos, principalmente em base de dados muito homogêneas, pois o conjunto de resposta pode trazer instâncias tão semelhantes, que não apresentam uma informação significativa, sem atender a finalidade da busca efetuada pelo usuário.

Atualmente, pesquisas têm focado esforços na exploração da diversidade entre as instâncias dos dados, para complementar a análise de similaridade pelos k -vizinhos mais próximos, que comumente é efetuada em sistemas de recuperação de dados complexos, principalmente em imagens. Assim, em geral, a partir de um fator de diversidade definido pelo usuário, o resultado da busca não contém apenas objetos similares ao centro de consulta, mas também diversos entre si, eliminando as redundâncias que não trazem informações consideráveis ao usuário, e dessa maneira, apresentando uma resposta mais efetiva (SKOPAL et al., 2009).

4.2 METODOLOGIA

O principal objetivo deste trabalho é o desenvolvimento de novas técnicas de coordenação baseadas na busca por similaridade pelos k -vizinhos mais próximos, aplicando as principais metodologias de diversidade na literatura, para destacar os resultados das consultas em representações visuais, geradas por projeções de conjuntos de dados, as quais mapeiam no plano de projeção resultante, as similaridades de um espaço multidimensional original (ELER, 2011).

A maioria das abordagens de recuperação de dados por similaridade aplicando a diversidade apresentam uma lista de instâncias resultantes, e exigem a definição de um parâmetro prévio de diversidade. Dessa maneira, o usuário não obtém a contextualização espacial do resultado em relação aos demais elementos no conjunto de dados, dificultando a tarefa de definição de um parâmetro de diversidade ideal, de acordo com os fatores

envolvidos na busca, como tipo de dados, extração e seleção de características. Nesse cenário, as técnicas de coordenação e representações visuais auxiliam o usuário na identificação do relacionamento entre as instâncias dos conjuntos de dados envolvidos no processo, como também, colabora na determinação de um fator de diversidade adequado, que satisfaça a sua intenção de busca.

A coordenação entre as visualizações é efetuada por meio de mapeamentos, que realizam a conexão entre os elementos dos conjuntos de dados, sendo que, para cada uma das técnicas de diversidade abordadas no Capítulo 3, é desenvolvido um mapeamento específico, implementando a busca por similaridade aplicando a diversidade no processo. A partir de um ou mais elementos selecionados na visão origem, são destacadas por cor as k instâncias recuperadas pela consulta nas demais visões destino, onde os mapeamentos são executados. Na Figura 13 é ilustrado um exemplo do processo de coordenação, em que são utilizadas uma visão origem e duas visões destino, formuladas de três conjuntos de dados diferentes, ressaltando que nos mapas de similaridade as cores associam o elemento à uma classe específica. Ao selecionar um elemento na visualização origem, que é indicado na Figura 13a, os três elementos resultantes do mapeamento baseando-se na similaridade e diversidade, são destacados nas demais visões 13b e 13c.

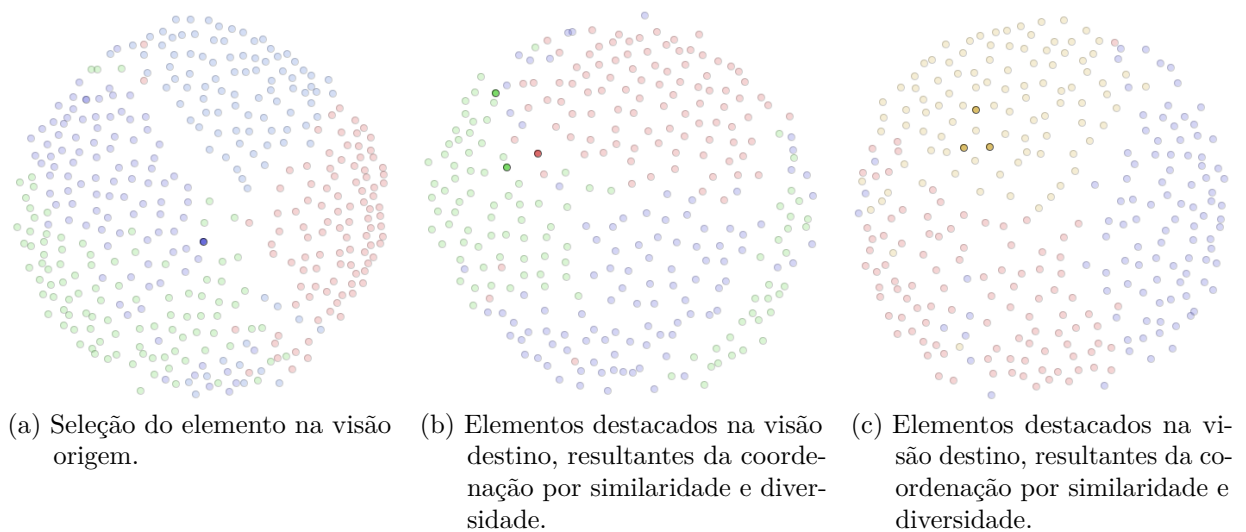


Figura 13 – Exemplo do processo coordenação aplicando a similaridade e a diversidade, utilizando uma visão origem e duas visões destino.

Na Figura 14 é apresentado um fluxograma, que ilustra o processo de coordenação realizado na Figura 13. Cada visualização é formada por um mapa de similaridade, obtidos pela projeção multidimensional dos dados arbitrários introduzidos, sendo que cada uma possui uma entidade de coordenação, que gerencia seus mapeamentos. Como pode ser observado no exemplo, com base na seleção efetuada na visão origem, por meio dos mapeamentos nas demais visões destino, é possível identificar os elementos associados nas

mesmas. O mecanismo de coordenação detecta quais elementos são selecionados, e em que visão, para então procurar nas demais a existência de mapeamentos que os relacionam.

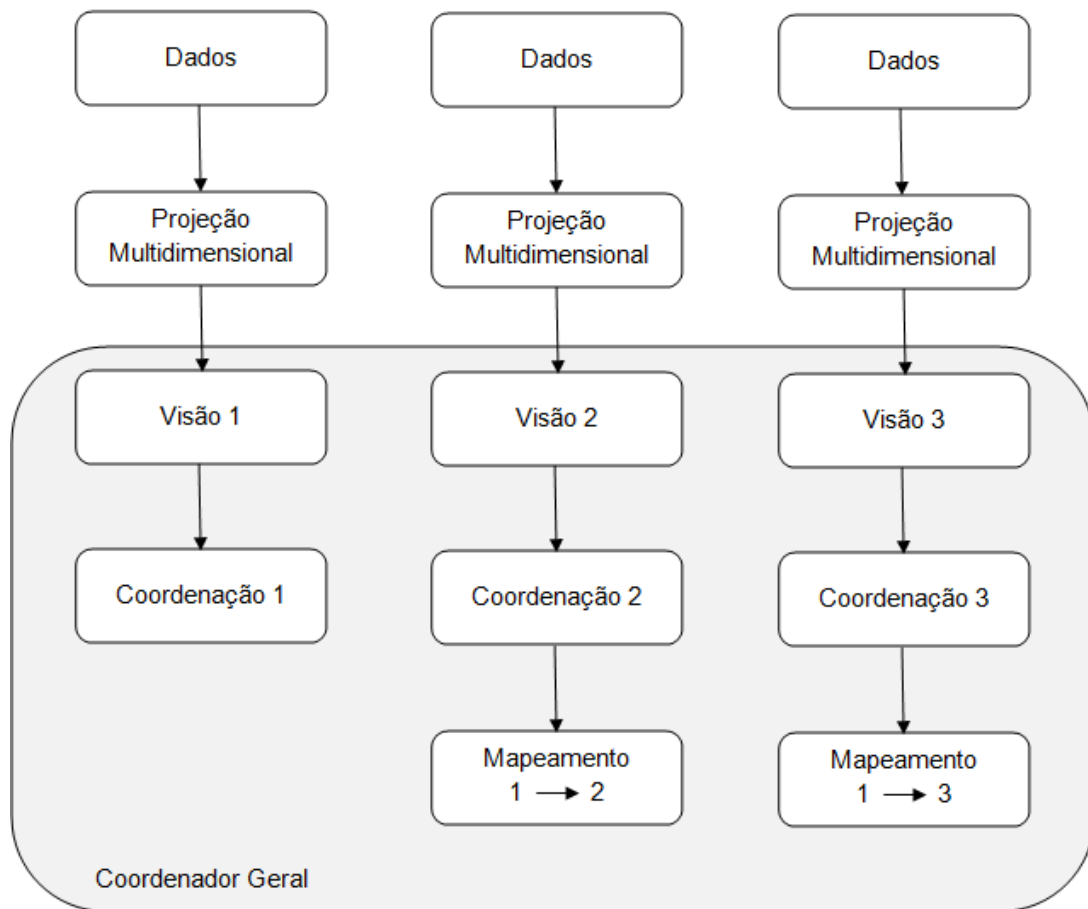


Figura 14 – Fluxograma do processo de coordenação da Figura 13.

Na Figura 15 é exibido um exemplo de coordenação entre uma visão origem e duas outras destino, sendo todas produzidas com um mesmo conjunto de dados. A partir da seleção de um elemento na visão origem, ilustrada na Figura 15a, na visão destino da Figura 15b, são destacados os cinco vizinhos mais próximos, aplicando a consulta apenas por similaridade, e na Figura 15c, são indicados os cinco vizinhos mais próximos, por meio da consulta por similaridade e diversidade, fazendo com que sejam retornados elementos não tão próximos entre si.

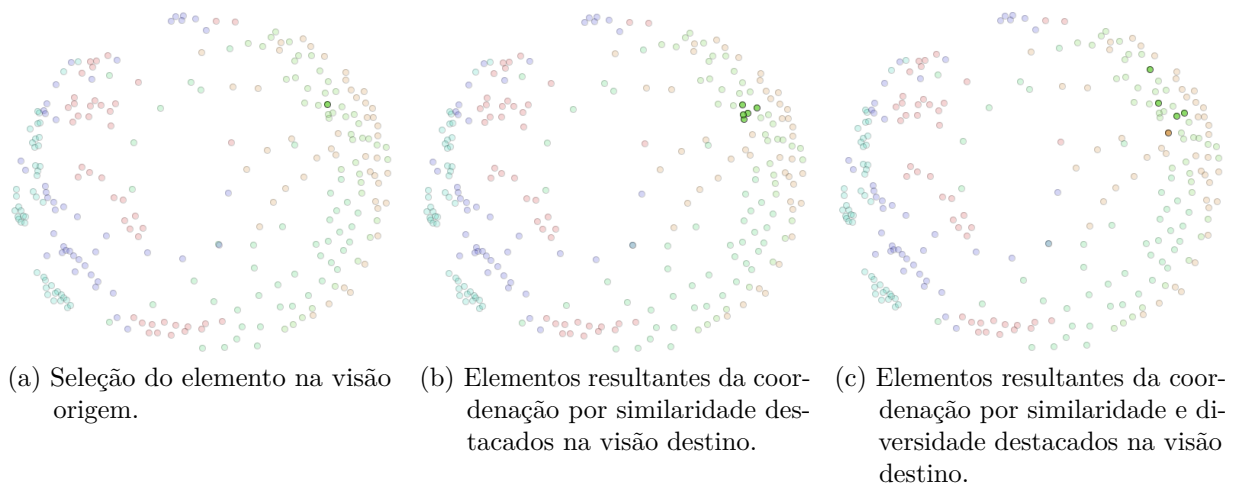


Figura 15 – Exemplo do processo coordenação aplicando a similaridade e a diversidade.

4.3 IMPLEMENTAÇÃO

A implementação das novas técnicas de coordenação que realizam o mapeamento entre representações visuais é realizada por meio da linguagem Java, e são acopladas ao processo exploratório da ferramenta *Projection Explorer* (PEX) (PAULOVICH; OLIVEIRA; MINGHIM, 2007), que possui técnicas de projeções multidimensionais para mapear, no espaço bidimensional, as relações de similaridade entre conjuntos de dados, além de possuir um arcabouço de fácil extensão para a implementação de novas técnicas de coordenação, e suporte a interação com o usuário.

De acordo com as visualizações formuladas, a partir dos dados introduzidos na PEX, o usuário pode ativar a ferramenta de coordenação, e ao selecionar algum elemento arbitrário em uma determinada visão, se houver algum mapeamento realizado previamente nas demais visões, serão destacados os k elementos associados nas mesmas. Na Figura 16, são apresentadas três visualizações de diferentes conjuntos de dados na PEX, na qual é aplicado o processo de coordenação baseando-se na similaridade e diversidade.

Os mapeamentos entre as visões precisam ser definidos previamente, sendo realizados por pares, definindo a visão origem e destino, como é mostrado na Figura 17, onde são apresentados todos os possíveis mapeamentos das três visões indicadas na Figura 16.

Ao selecionar um par de visualizações para efetuar o mapeamento entre seus respectivos elementos, por meio da similaridade e diversidade, uma tela de configuração é exibida, conforme na Figura 18. Nela, o usuário pode definir o tipo da função de dissimilaridade empregada na técnica de diversidade, e o tipo de busca realizada no processo, podendo ser pelos k -vizinhos mais próximos ou por abrangência, destacando que, as técnicas de diversidade são aplicadas apenas nas buscas pelos k -vizinhos mais próximos, por ser mais simples do usuário definir um número de elementos que serão retornados.

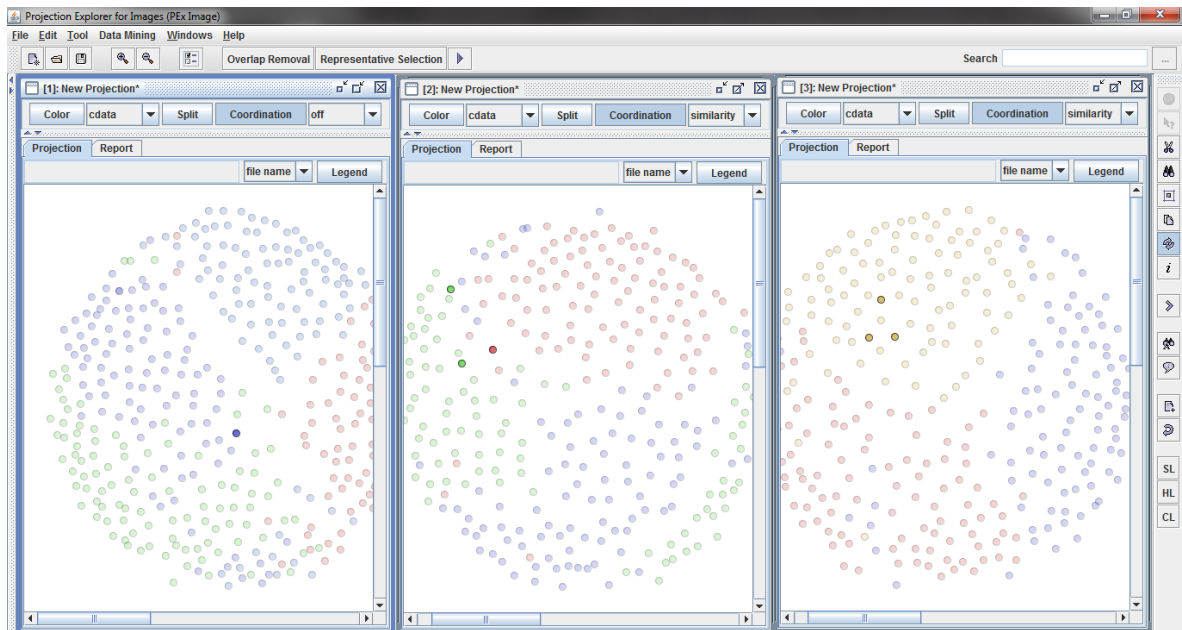


Figura 16 – Visualizações de diferentes conjuntos de dados no PEx.

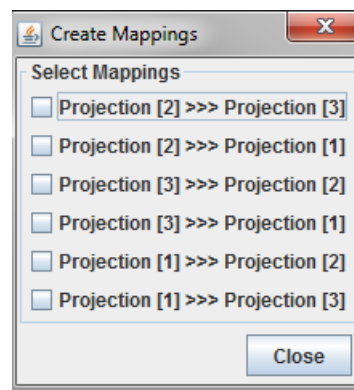


Figura 17 – Tela PEx: Escolha do par de visões para efetuar o mapeamento.

Escolhendo a busca pelos k -vizinhos mais próximos, deve-se definir a quantidade de k , ou seja, a partir de um elemento localizado na visão origem, quantos elementos na visão destino, vão ser encontrados e associados à ele pela técnica de diversidade. Finalmente, para incluir a diversidade na busca por similaridade, é preciso selecionar uma das sete técnicas disponíveis. Ressaltando que apenas na técnica BRID, que emprega o conceito de Variedade, não é necessário definir um fator de diversidade, nas demais (MMR, *Swap*, Motley, MSD, GMC e GNE), um valor entre 0 e 1 deve ser estabelecido, sendo que o valor 0 resulta em uma busca empregando apenas a similaridade, e o valor 1 apenas a diversidade.

Para cada técnica de diversidade disponível, é desenvolvido um novo mapeamento entre os elementos, levando em consideração a abordagem, estratégia, função de maximização, entre outros fatores específicos inerentes à cada uma. Salientando que, cada entidade

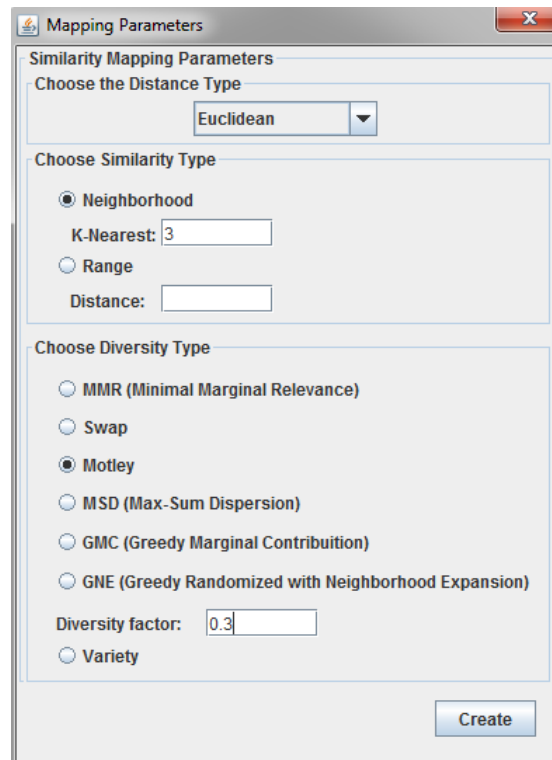


Figura 18 – Tela PEx: Configuração da coordenação por similaridade e diversidade.

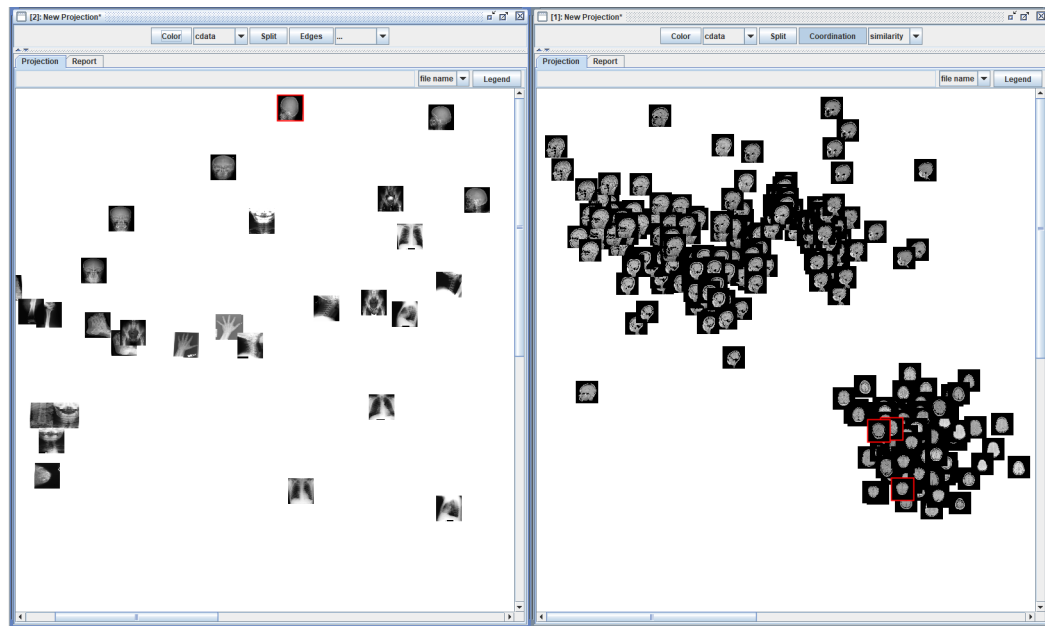
de coordenação, definida na Figura 14, pode conter diversos mapeamentos.

Na Figura 19 é apresentada uma comparação entre o mapeamento na coordenação aplicando a busca apenas por similaridade, e aplicando a diversidade, utilizando duas projeções de conjuntos de dados distintos. A partir do centro de consulta destacado na visão origem em vermelho, é possível identificar na visão destino, os três vizinhos mais próximos na Figura 19a. E na Figura 19b, são indicados os três vizinhos mais próximos, aplicando a diversidade.

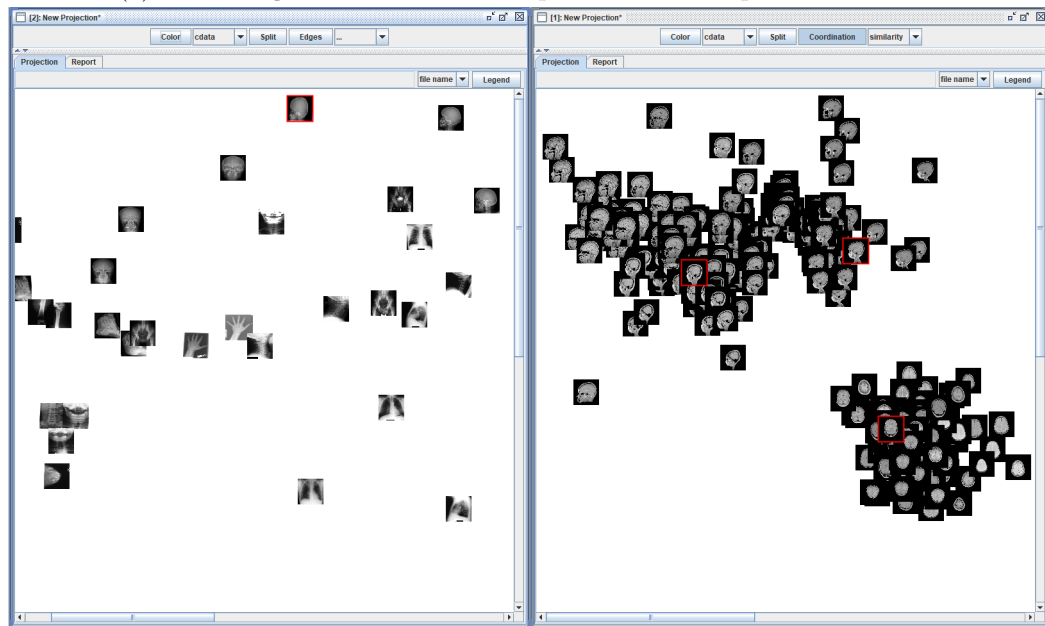
4.4 CONSIDERAÇÕES FINAIS

A visualização de diferentes conjuntos de dados, e a coordenação entre os seus respectivos elementos, contribui para o reconhecimento e compreensão de novos relacionamentos entre os mesmos. Além disso, aplicando a busca por similaridade e diversidade no processo de coordenação, auxilia o usuário na definição de um parâmetro de diversidade adequado. Pois pela visualização, por meio da indicação por destaque, é possível identificar como os elementos resultantes e selecionados estão dispostos em relação aos demais no conjunto de dados, permitindo analisar as técnicas de diversidade eficientemente, observando e compreendendo como o fator de diversidade influencia o resultado da busca.

A abordagem desenvolvida possibilita a formação de diversos tipos de mapeamentos,



(a) Coordenação entre duas visões aplicando a busca por similaridade.



(b) Listagem dos quatro vizinhos mais próximos, BRID.

Figura 19 – Coordenação entre duas visões por meio da busca por similaridade aplicando a diversidade.

entre diferentes conjuntos de dados representados graficamente por mapas de similaridade, empregando projeções multidimensionais distintas. Dessa maneira, permitindo realizar uma comparação entre diferentes contextos na aplicação das técnicas de diversidade, fornecendo uma melhor análise e compreensão de relações complexas e profundas, as quais seriam difícil identificar sem o poder da abstração da visualização dos dados.

5 EXPERIMENTOS E RESULTADOS

5.1 DESENVOLVIMENTO DOS EXPERIMENTOS

Os experimentos foram planejados com o objetivo de expressar o quão significativo é o trabalho de coordenação de Múltiplas Visões Coordenadas, baseando-se tanto na similaridade dos dados, quanto na diversidade, para agregar valor e relevância às respostas de consultas, sobretudo em base de dados com alta redundância. E também auxiliar na análise do comportamento das técnicas de diversidade, e definição de um parâmetro adequado nas consultas.

As técnicas de coordenação abordadas e desenvolvidas neste trabalho podem ser aplicadas para qualquer tipo de dado complexo, em que é possível extrair características associadas à eles. Porém são utilizadas imagens nos experimentos, por possibilitar a avaliação da efetividade das técnicas de maneira mais simples e rápida, que outros tipos de dados complexos. As seguintes bases de dados empregadas nos experimentos, são utilizadas na literatura em segmentos que trabalham tanto com recuperação de imagens por conteúdo, quanto na análise da similaridade e da diversidade:

- **ALOI (*Amsterdam Library of Object Images*)** : Coleção de imagens sintéticas coloridas, as quais capturam diferentes ângulos de 1000 objetos distintos (GEUSEBROEK; BURGHOUTS; SMEULDERS, 2005).
- **Corel (*The COREL Database for Content based Image Retrieval*)** : Coleção de 5000 imagens coloridas, de diversos conteúdos (TAO, 2018).
- **Ressonâncias magnéticas** : Coleção de 704 imagens de ressonâncias magnéticas BALAN (2007).

Foram formados subconjuntos menores, provenientes dos conjuntos de dados maiores citados anteriormente, para facilitar a visualização dos resultados, e a análise e validação do desempenho dos métodos. Na Tabela 1, são descritas as especificações de cada subconjunto formado.

Salienta-se que a técnica de extração de características aplicadas aos dados, influencia diretamente no sucesso da abordagem de busca por meio da similaridade e diversidade. Nos subconjuntos Corel₁ e Corel₂, a extração é realizada usando a última camada de classificação da CNN (*Convolutional Neural Network*) AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), sendo que a partir de uma imagem introduzida na rede, a camada final resulta em um vetor de 4096 atributos, porém são selecionadas apenas os 200 mais

Dataset	Instâncias	Dimensões	Características
Corel ₁	300	200	AlexNet
Corel ₂	400	200	AlexNet
RM ₁	6	28	<i>Fourier</i>
RM ₂	300	28	<i>Fourier</i>
ALOI ₁	288	25921	HOG
ALOI ₂	504	25921	HOG
ALOI ₃	288	130	HOG
ALOI ₄	504	130	HOG
ALOI ₅	51	257	Histograma de cor

Tabela 1 – Conjuntos de dados utilizados nos experimentos.

representativos, para reduzir a complexidade do cálculo nas abordagens de diversidade. Nos subconjuntos RM₁ e RM₂, originados do conjunto de imagens de Ressonâncias Magnéticas, as características são extraídas pelo descritor de *Fourier* (HUANG; AVIYENTE, 2006). Já nos subconjuntos ALOI₁, ALOI₂, ALOI₃ e ALOI₄, as características são extraídas por meio da técnica HOG (*Histogram Oriented Gradient*) (DALAL; TRIGGS, 2005), entretanto, nos ALOI₃ e ALOI₄ são utilizadas apenas as 130 características de maior significância, para também reduzir a complexidade da abordagem de diversidade.

Na Tabela 2, são especificadas as configurações dos experimentos realizados, sendo que os objetivos dos mesmos foram: comparar as técnicas de diversidade em representações visuais distintas, construídas a partir de diferentes funções de distâncias utilizadas na formação das projeções multidimensionais; demonstrar como a evolução da aplicação de diferentes parâmetros de diversidade, influencia no resultado da busca; e mostrar o desempenho da técnica BRID em contraste as outras técnicas de diversidade, que exigem parâmetros que são intrínsecos aos conjuntos de dados em questão.

Configuração	Dataset	Técnica de Projeção	Função de Distância
Config ₁	Corel ₁	<i>Sammon's Mapping</i>	Mahalanobis
Config ₂	Corel ₂	<i>Sammon's Mapping</i>	Mahalanobis
Config ₃	RM ₁	<i>Neighbour-Joining</i>	Euclidiana
Config ₄	RM ₂	<i>Sammon's Mapping</i>	Euclidiana
Config ₅	ALOI ₁	<i>Sammon's Mapping</i>	Euclidiana
Config ₆	ALOI ₁	<i>Sammon's Mapping</i>	Cosseno
Config ₇	ALOI ₂	<i>Sammon's Mapping</i>	Euclidiana
Config ₈	ALOI ₂	<i>Sammon's Mapping</i>	Cosseno
Config ₉	ALOI ₃	<i>Sammon's Mapping</i>	Mahalanobis
Config ₁₀	ALOI ₄	<i>Sammon's Mapping</i>	Mahalanobis
Config ₁₁	ALOI ₅	<i>Neighbour-Joining</i>	Euclidiana
Config ₁₂	ALOI ₅	<i>Sammon's Mapping</i>	Euclidiana

Tabela 2 – Especificação dos experimentos.

A codificação e experimentos são realizados em um computador que possui CPU

Intel Core i7-4770K 3.50GHz, com 16GB DDR3 RAM e Sistema Operacional Windows 10 64 bits.

5.2 RESULTADOS

Nesta seção são apresentados os experimentos efetuados, de acordo com as configurações previamente definidas, nos quais, a seleção de um elemento em uma visualização origem é mostrada, então são destacados os elementos nas demais visões destino, sendo que a seleção dos elementos nas visualizações, são evidenciadas por uma borda vermelha em torno dos elementos. Além da exibição dos resultados nas visões, os elementos pertencentes aos conjuntos de respostas alcançados, são apresentados em uma listagem.

5.2.1 COMPARAÇÃO ENTRE ESPAÇOS MÉTRICOS

Os espaços métricos formados a partir de um conjunto de dados e determinada função de distância, de acordo com as definições realizadas anteriormente no Capítulo 4, influenciam diretamente o comportamento da busca por similaridade aplicando a diversidade, pois em conjunto às técnicas de projeção multidimensional, definem a disposição e relacionamento entre as instâncias dos conjuntos de dados nas visualizações envolvidas no processo de coordenação.

A seguir, são realizadas comparações entre coordenações aplicando apenas a similaridade, e as técnicas de diversidade: Motley (JAIN; SARDA; R., 2004), MMR (CARBONELL; GOLDSTEIN, 1998), *Swap* (YU; LAKSHMANAN; AMER-YAHIA, 2009) e BRID (SANTOS et al., 2013), empregando as funções de distâncias: euclidiana, do cosseno e mahalanobis. Com o objetivo de avaliar o comportamento das técnicas em relação à diferentes espaços métricos, permitindo analisar o impacto da alteração dos parâmetros de diversidade diretamente nas visualizações.

5.2.1.1 Perspectiva Geral dos Resultados

Inicialmente na Figura 20, são mostrados previamente em forma de listagem, os resultados dos experimentos propostos neste subseção, que têm como intuito comparar diversas técnicas de diversidade e seus respectivos parâmetros, em diferentes espaços métricos, os quais aplicam as configurações Config₅ a Config₁₀, utilizando as características HOG, a partir do centro de consulta indicado na Figura 21c.

Pode-se verificar que mesmo obtendo os elementos resultantes da consulta, não é possível identificar como esses elementos estão dispostos em relação aos demais no conjunto de dados, e também como essas instâncias estão próximas umas das outras, de acordo com o espaço métrico. Nesse sentido, a projeção desses dados em representações visuais permite que o usuário compreenda como os elementos relacionam-se entre si, contribuindo para

uma melhor análise da influência da definição dos parâmetros de diversidade, facilitando a tarefa de estabelecer um fator adequado.

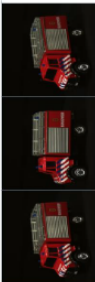
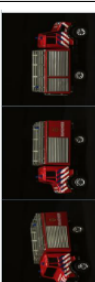
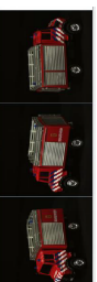
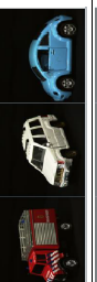
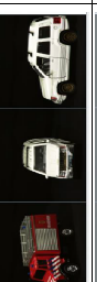
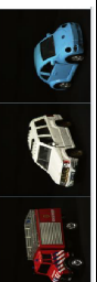
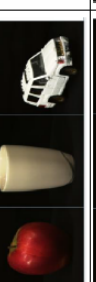
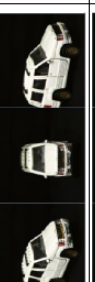
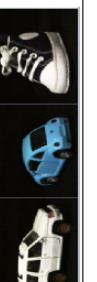
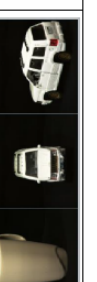
Função de Distância	Similaridade – 3NN	DIV	Motley – 3NN	DIV	MMR – 3NN	DIV	Swap – 3NN	BRID – 3NN
Euclidiana		0.7		0.1		0.1		
Euclidiana		0.8		0.3		0.2		
Euclidiana		0.9		0.5		0.5		
Cosseno		0.3		0.1		0.1		
Cosseno		0.4		0.3		0.2		
Cosseno		0.5		0.5		0.5		
Mahalanobis		0.1		0.2		0.1		
Mahalanobis		0.2		0.3		0.4		
Mahalanobis		0.3		0.5		0.8		

Figura 20 – Fluxograma do processo de coordenação da Figura 13.

5.2.1.2 Função de Distância Euclidiana

Nos seguintes experimentos as visualizações origens são formadas utilizando a configuração Config₅, e as visualizações destinos, nas quais executam-se as buscas pelos três vizinhos mais próximos, são formuladas pela Config₇, ressaltando que são subconjuntos distintos com atributos em comum, empregando a função de distância euclidiana.

Na Figura 21a a instância selecionada é destacada em vermelho, e então na visualização exibida na Figura 21b, são indicados os três vizinhos mais próximos, resultantes da busca por similaridade.

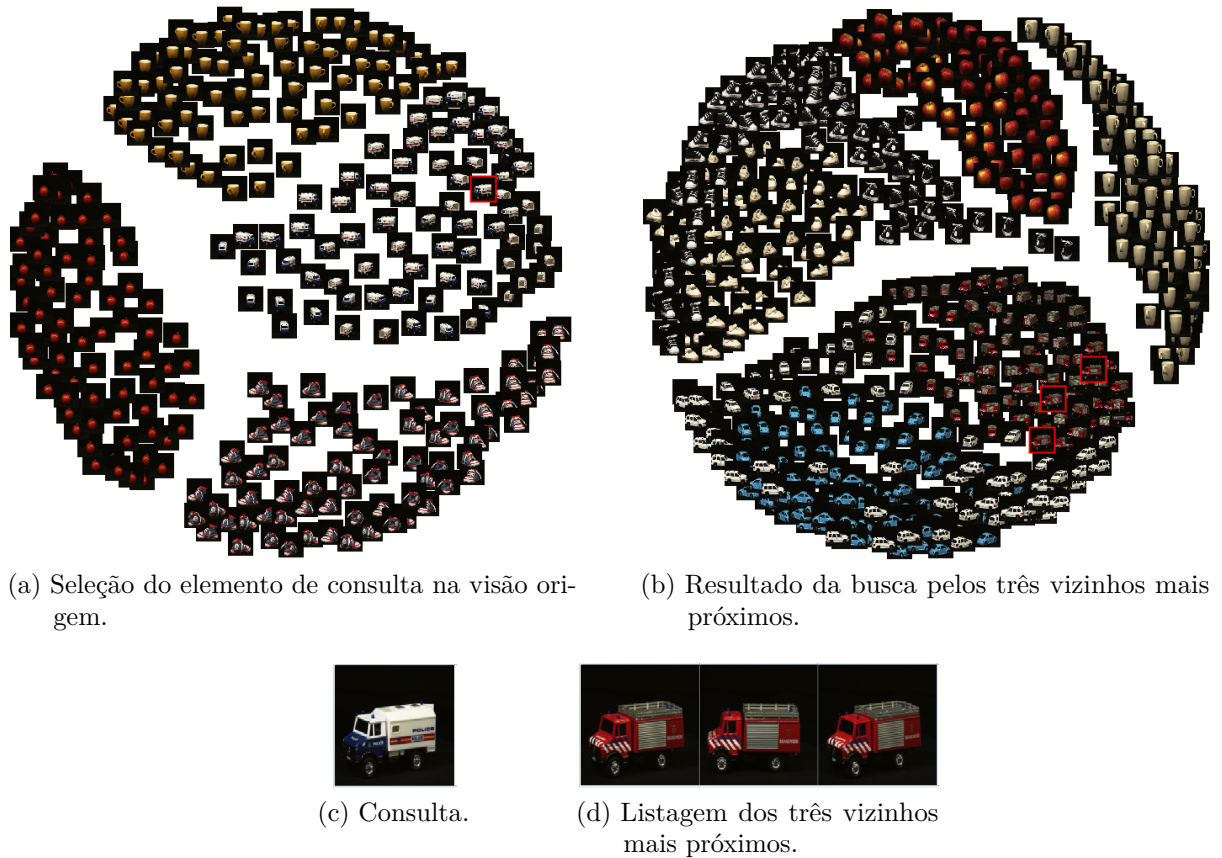


Figura 21 – Consulta por similaridade, utilizando a função de distância euclidiana.

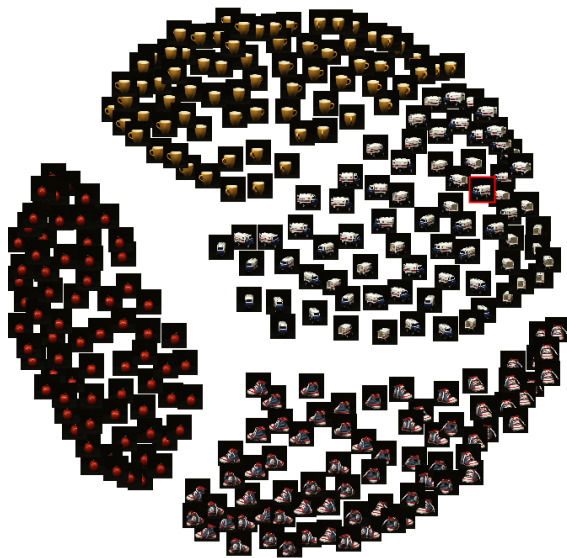
Na Figura 22a a instância selecionada é indicada em vermelho, e então nas visualizações exibidas nas Figuras 22b, 22c e 22d, são destacados os três vizinhos mais próximos por meio da busca por similaridade aplicando a técnica de diversidade Motley, com o parâmetro θ definido em 0.7, 0.8 e 0.9, respectivamente. Conforme o fator de diversidade cresce, o conjunto de resposta vai se tornando mais diverso em relação ao elemento central da consulta, salientando-se que nessa configuração do experimento, utilizando as características HOG e função de distância euclidiana, foi necessário um fator de diversidade alto para que ela tenha algum efeito na consulta por similaridade, levando em consideração que o fator máximo é 1.0.

De acordo com o resultado da técnica de diversidade MMR aplicada na busca por similaridade pelos três vizinhos mais próximos, indicados nas Figuras 23b, 23c e 23d, a partir do elemento central de consulta selecionado na visão origem exibida na Figura 23a, é possível analisar que, em comparação ao resultado aplicando a diversidade Motley, quando o fator α fica maior, os elementos do conjunto de resposta não se distanciam muito da caracterização do elemento de consulta, tendo em conta a característica extraída das imagens (HOG) que descrevem sua forma.

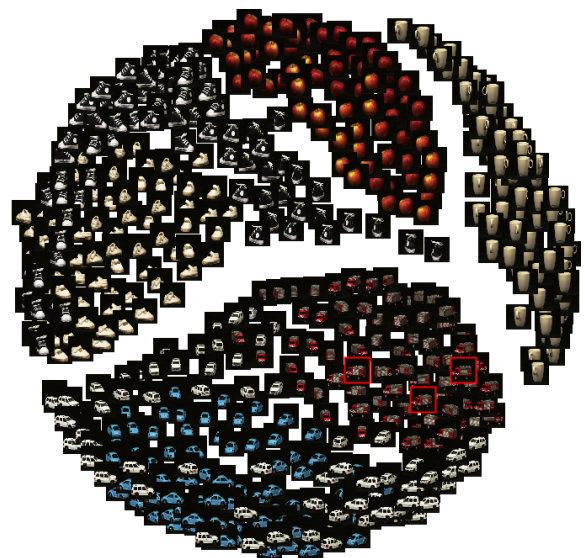
Na Figura 24a é mostrada a seleção do elemento de consulta na visão origem, e então nas Figuras 24b, 24c e 24d, são exibidos os resultados das consultas pelos três vizinhos mais próximos aplicando a técnica de diversidade *Swap*, nas quais é possível verificar que à medida que o parâmetro λ aumenta, os elementos retornados na consulta são bem diversos, distanciando-se rapidamente do centro de consulta.

A consulta pelos três vizinhos mais próximos usando a técnica de diversidade BRID, exibida na Figura 25b, apresentou elementos resultantes que são bem parecidos ao centro de consulta, mostrado na Figura 25a. É possível notar que a aplicação do conceito de Variedade no espaço métrico utilizando a distância euclidiana, apresenta elementos semelhantes ao centro de consulta, porém evita a repetição, acrescentando um nível sutil de relevância à resposta.

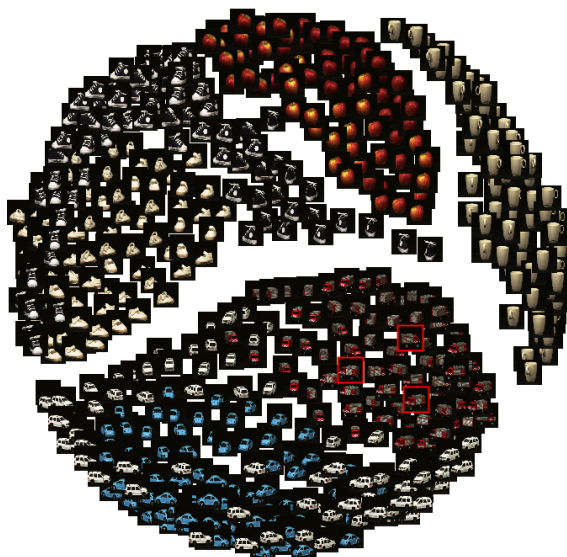
Segundo os experimentos efetuados, pode-se verificar que a projeção multidimensional dos dados baseando-se na função de distância euclidiana, realiza a disposição entre os elementos de maneira condizente ao esperado, de acordo com as características extraídas dos mesmos, sendo que os elementos mais similares estão mais próximos. A partir de um centro de consulta arbitrário em uma determinada visualização, ao efetuar a consulta por similaridade nas visões destino, acrescentando a diversidade gradativamente, é possível notar como os elementos vão se distanciando do elemento considerado mais próximo à ele, permitindo analisar o relacionamento do mesmo em relação aos demais elementos no conjunto de dados.



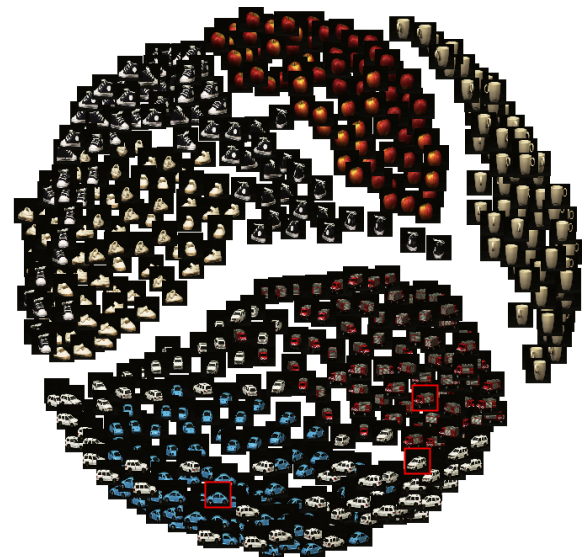
(a) Seleção do elemento de consulta na visão original.



(b) Resultado da busca pelos três vizinhos mais próximos, Motley 0.7.



(c) Resultado da busca pelos três vizinhos mais próximos, Motley 0.8.



(d) Resultado da busca pelos três vizinhos mais próximos, Motley 0.9.



(e) Consulta.



(f) Listagem dos três vizinhos mais próximos, 0.7 Motley.



(g) Listagem dos três vizinhos mais próximos, 0.8 Motley.

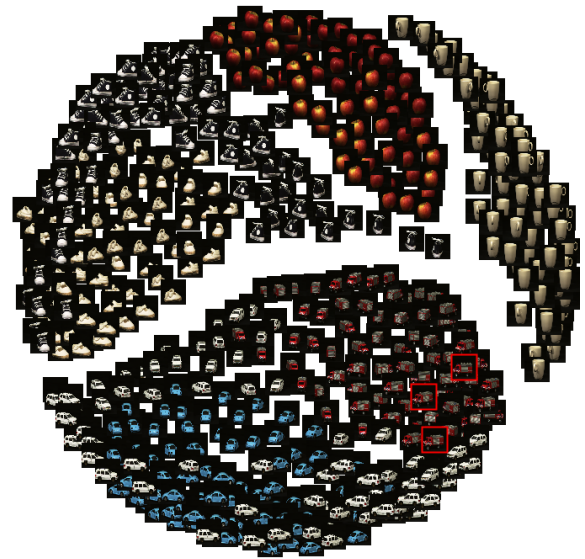


(h) Listagem dos três vizinhos mais próximos, 0.9 Motley.

Figura 22 – Consulta por similaridade aplicando a técnica de diversidade Motley, utilizando a função de distância euclidiana.



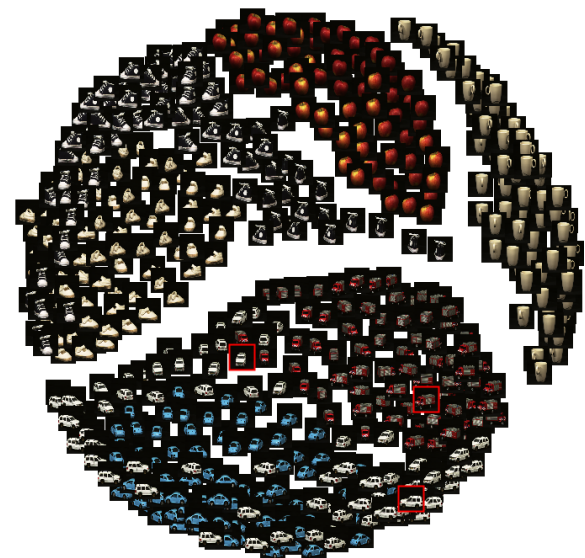
(a) Seleção do elemento de consulta na visão origem.



(b) Resultado da busca pelos três vizinhos mais próximos, MMR 0.1.



(c) Resultado da busca pelos três vizinhos mais próximos, MMR 0.3.



(d) Resultado da busca pelos três vizinhos mais próximos, MMR 0.5.



(e) Consulta.



(f) Listagem dos três vizinhos mais próximos, 0.1 MMR.



(g) Listagem dos três vizinhos mais próximos, 0.3 MMR.

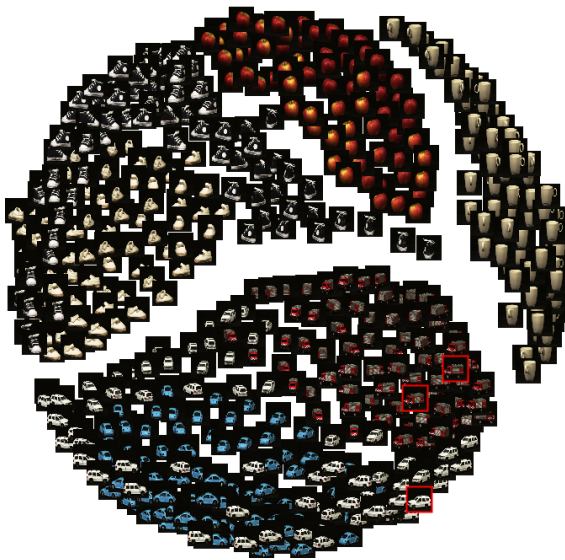
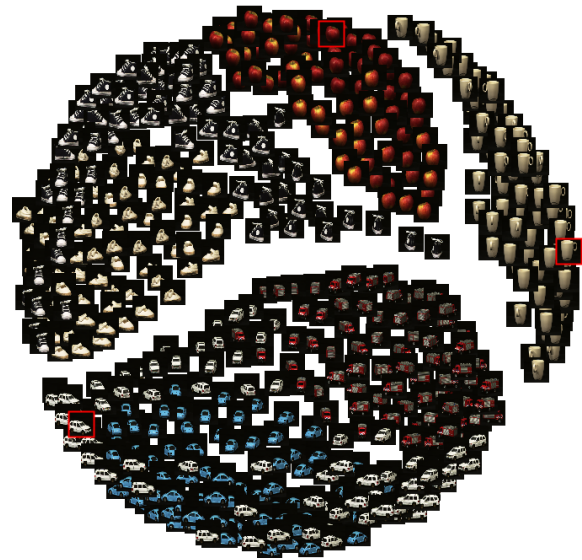


(h) Listagem dos três vizinhos mais próximos, 0.5 MMR.

Figura 23 – Consulta por similaridade aplicando a técnica de diversidade MMR, utilizando a função de distância euclidiana.



(a) Seleção do elemento de consulta na visão origem.

(b) Resultado da busca pelos três vizinhos mais próximos, *Swap* 0.1.(c) Resultado da busca pelos três vizinhos mais próximos, *Swap* 0.2.(d) Resultado da busca pelos três vizinhos mais próximos, *Swap* 0.5.

(e) Consulta.

(f) Listagem dos três vizinhos mais próximos, 0.1 *Swap*.(g) Listagem dos três vizinhos mais próximos, 0.2 *Swap*.(h) Listagem dos três vizinhos mais próximos, 0.5 *Swap*.

Figura 24 – Consulta por similaridade aplicando a técnica de diversidade *Swap*, utilizando a função de distância euclidiana.

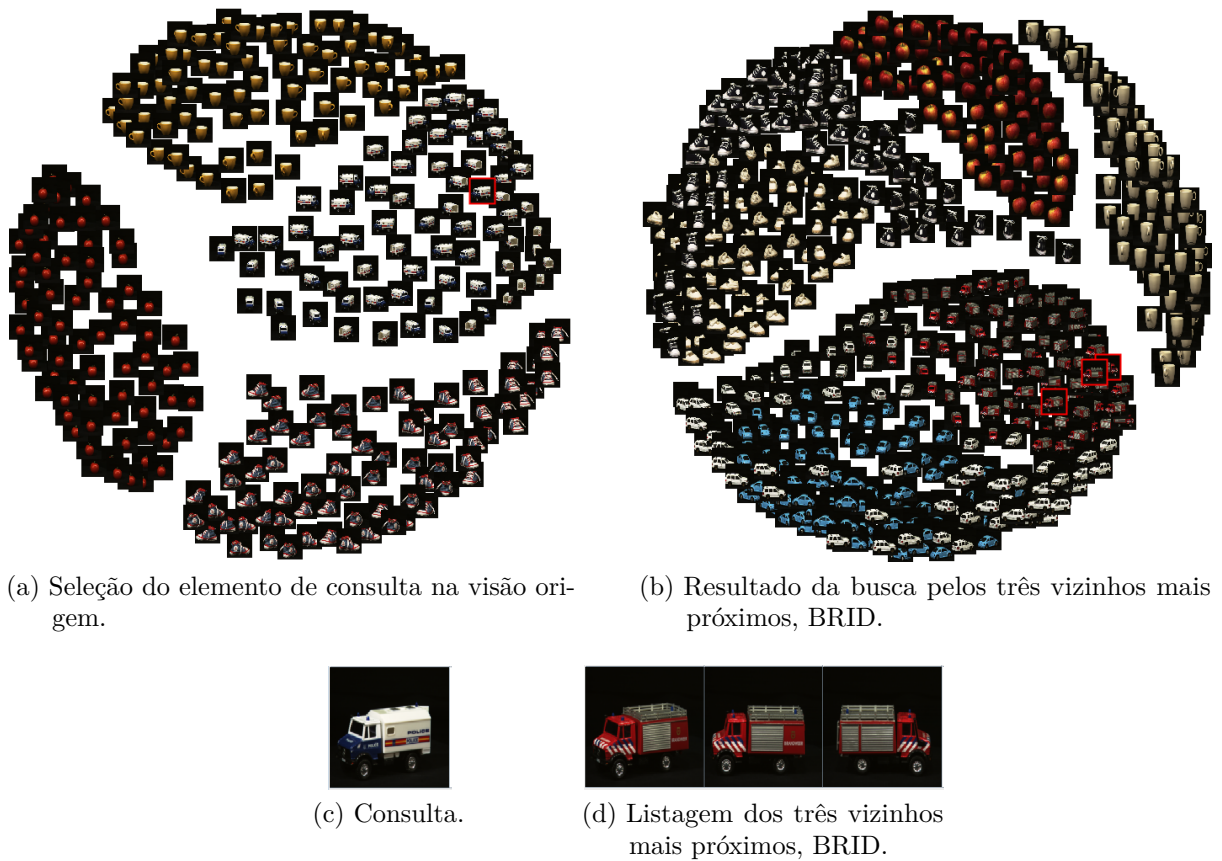


Figura 25 – Consulta por similaridade aplicando a técnica de diversidade BRID, utilizando a função de distância euclidiana.

5.2.1.3 Função de Distância do Cosseno

Nos seguintes experimentos desta subseção, as visualizações origens são formadas utilizando a configuração *Config₆*, e as visualizações destinos, nas quais executam-se as buscas pelos três vizinhos mais próximos, são formuladas pela *Config₈*, ressaltando que são subconjuntos distintos com atributos em comum, empregando a função de distância do cosseno.

Na Figura 26a, a instância selecionada é destacada em vermelho, e então na visão exibida na Figura 26b, são destacados os três vizinhos mais próximos, resultantes da busca por similaridade.

É possível notar que ao efetuar a busca por similaridade aplicando a diversidade Motley, utilizando a função de distância do cosseno, conforme é ilustrado nas Figuras 27b, 27c e 27d, obtemos resultados mais diversos com parâmetros θ menores, de acordo com a consulta mostrada na Figura 27a, em comparação ao experimento realizado anteriormente usando a função euclidiana.

Conforme o elemento central de consulta selecionado na visão origem, ilustrada na Figura 28a, os três elementos mais próximos empregando a técnica de diversidade MMR,

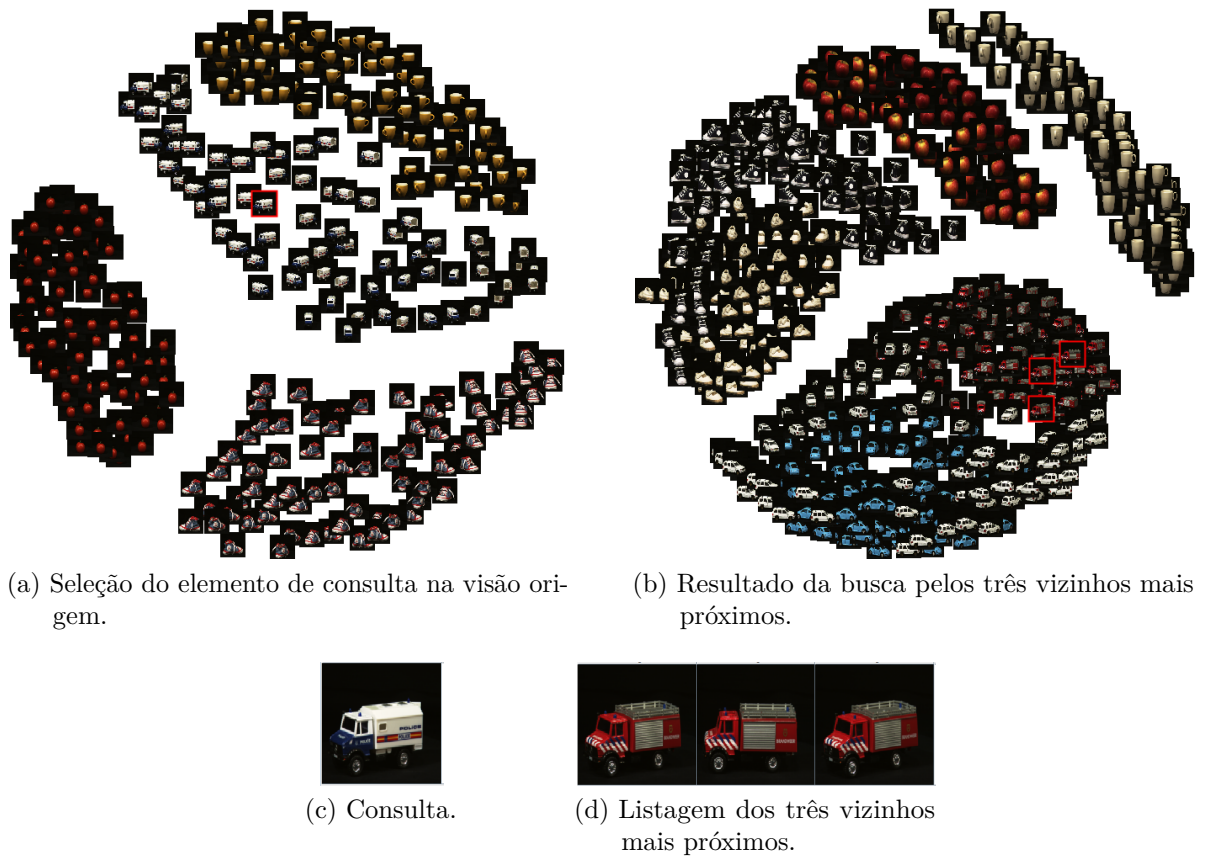


Figura 26 – Consulta por similaridade, utilizando a função de distância do cosseno.

de acordo com as Figuras 28b, 28c e 28d, são mais diversos entre si, à medida que o valor λ aumenta.

Na Figura 29a o centro de consulta é selecionado na visão origem, e então nas demais visões destino, são efetuadas as buscas pela similaridade, acrescentando a técnica de diversidade *Swap*, que apresenta um resultado análogo à diversidade *Swap* baseando-se na função de distância euclidiana, conforme é ilustrado nas Figuras 29b, 29c e 29d.

O experimento realizado com a técnica BRID, utilizando a função de distância do cosseno, tem um resultado semelhante ao experimento que emprega a distância euclidiana, como pode ser observado na Figura 30b, de acordo com o centro de consulta na Figura 30a. Os resultados das consultas apresentam os três elementos mais próximos, adicionando um pequeno nível de diferença.

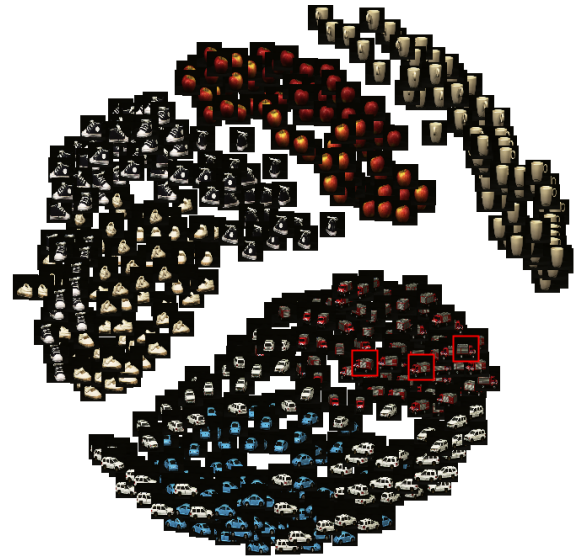
A partir de um centro de consulta arbitrário em uma determinada visualização, ao efetuar a consulta por similaridade nas visões destino, acrescentando a diversidade gradativamente, é possível notar como os elementos vão se distanciando do elemento considerado mais próximo à ele, permitindo analisar o relacionamento do mesmo em relação aos demais elementos no conjunto de dados.

Baseando-se nos experimentos apresentados, pode-se verificar que a projeção multi-

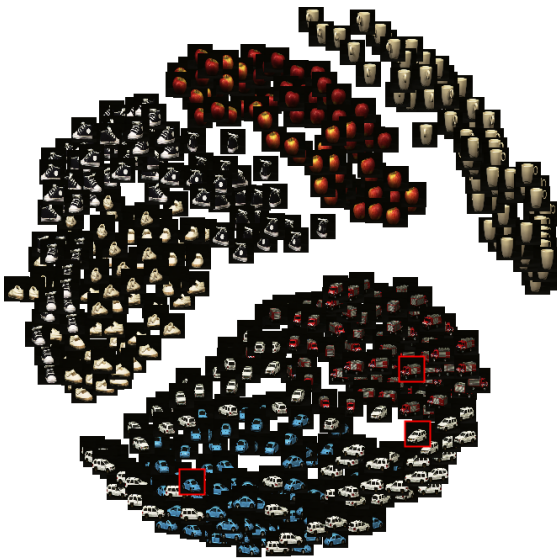
dimensional dos dados por meio da função de distância euclidiana, realiza a disposição entre os elementos de maneira condizente ao esperado, do mesmo modo que no emprego da função euclidiana, onde as instâncias mais similares estão próximas. Dessa maneira, contribuindo para o propósito do desenvolvimento da proposta, auxiliando o processo de coordenação entre as visualizações de modo adequado.



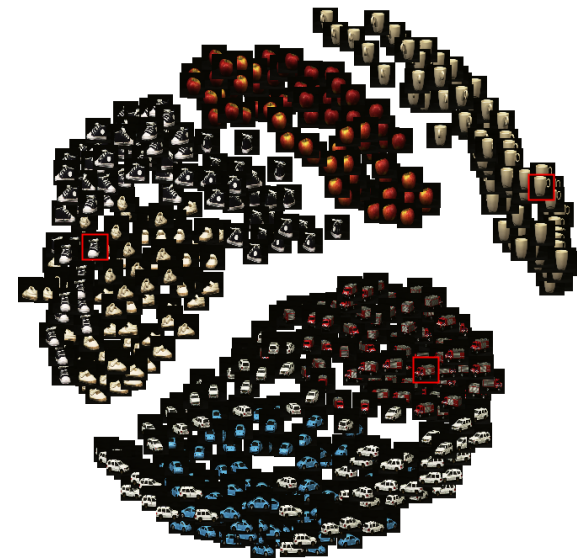
(a) Seleção do elemento de consulta na visão original.



(b) Resultado da busca pelos três vizinhos mais próximos, Motley 0.3.



(c) Resultado da busca pelos três vizinhos mais próximos, Motley 0.4.



(d) Resultado da busca pelos três vizinhos mais próximos, Motley 0.5.



(e) Consulta.



(f) Listagem dos três vizinhos mais próximos, 0.3 Motley.



(g) Listagem dos três vizinhos mais próximos, 0.4 Motley.

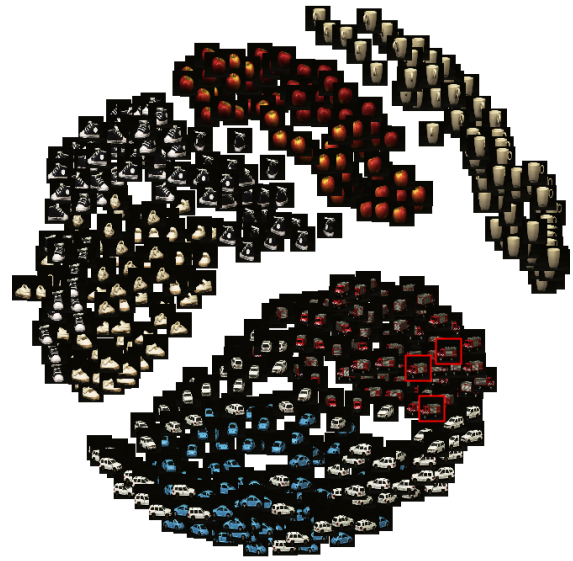


(h) Listagem dos três vizinhos mais próximos, 0.5 Motley.

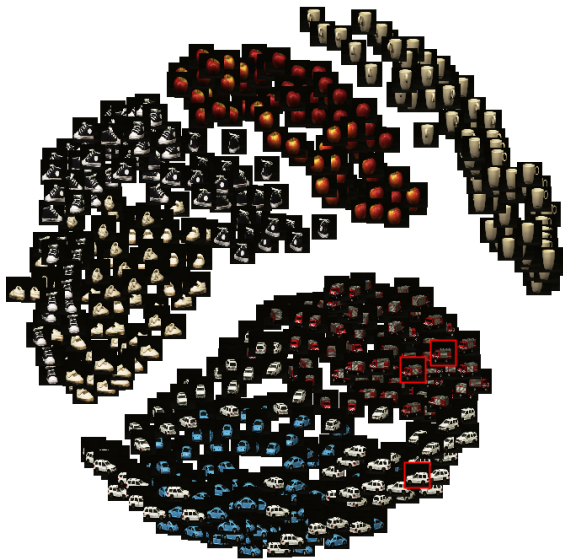
Figura 27 – Consulta por similaridade aplicando a técnica de diversidade Motley, utilizando a função de distância do cosseno.



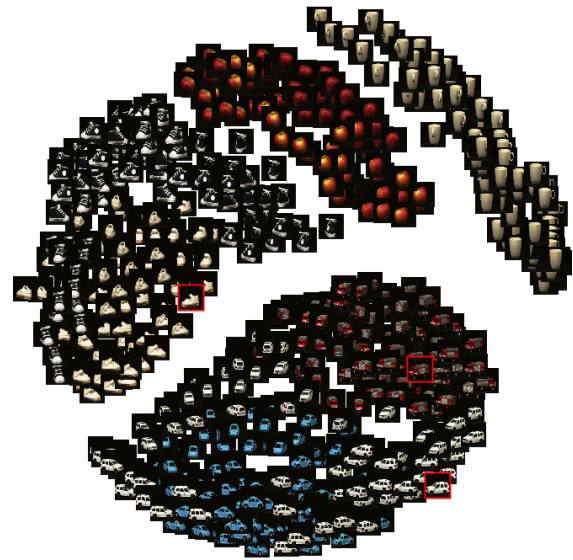
(a) Seleção do elemento de consulta na visão original.



(b) Resultado da busca pelos três vizinhos mais próximos, MMR 0.1.



(c) Resultado da busca pelos três vizinhos mais próximos, MMR 0.3.



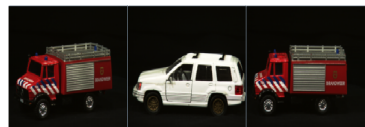
(d) Resultado da busca pelos três vizinhos mais próximos, MMR 0.5.



(e) Consulta.



(f) Listagem dos três vizinhos mais próximos, 0.1 MMR.



(g) Listagem dos três vizinhos mais próximos, 0.3 MMR.

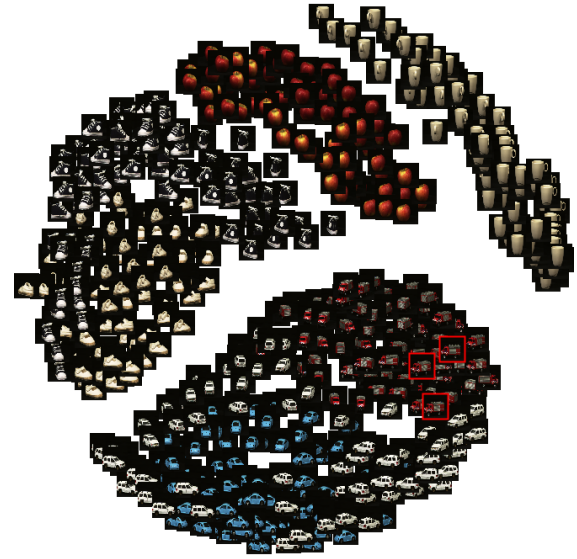


(h) Listagem dos três vizinhos mais próximos, 0.5 MMR.

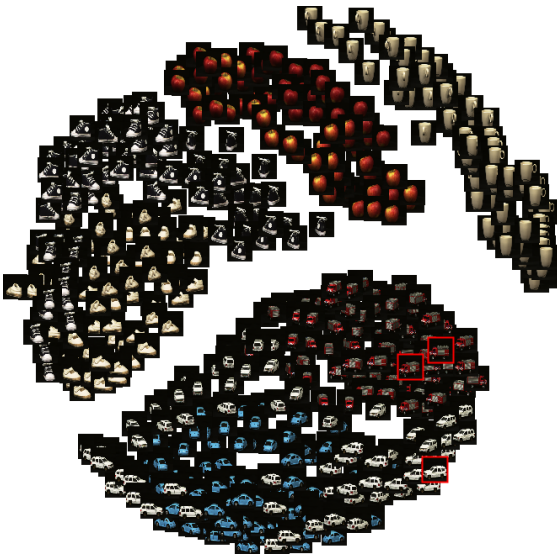
Figura 28 – Consulta por similaridade aplicando a técnica de diversidade MMR, utilizando a função de distância do cosseno.



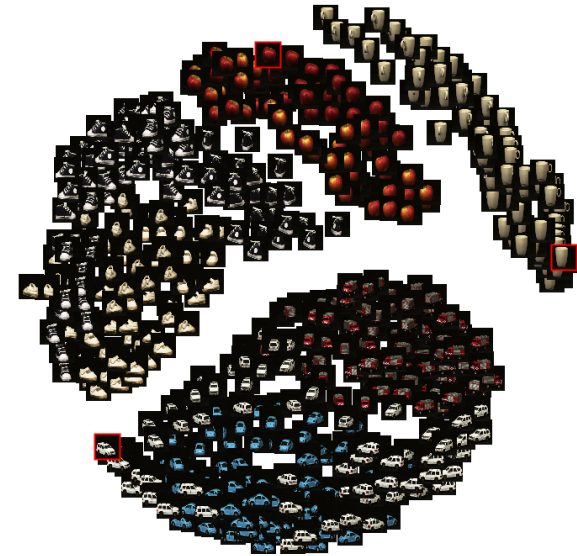
(a) Seleção do elemento de consulta na visão original.



(b) Resultado da busca pelos três vizinhos mais próximos, *Swap* 0.1.



(c) Resultado da busca pelos três vizinhos mais próximos, *Swap* 0.2.



(d) Resultado da busca pelos três vizinhos mais próximos, *Swap* 0.5.



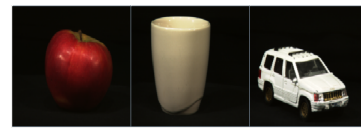
(e) Consulta.



(f) Listagem dos três vizinhos mais próximos, 0.1 *Swap*.



(g) Listagem dos três vizinhos mais próximos, 0.2 *Swap*.



(h) Listagem dos três vizinhos mais próximos, 0.5 *Swap*.

Figura 29 – Consulta por similaridade aplicando a técnica de diversidade *Swap*, utilizando a função de distância do cosseno.

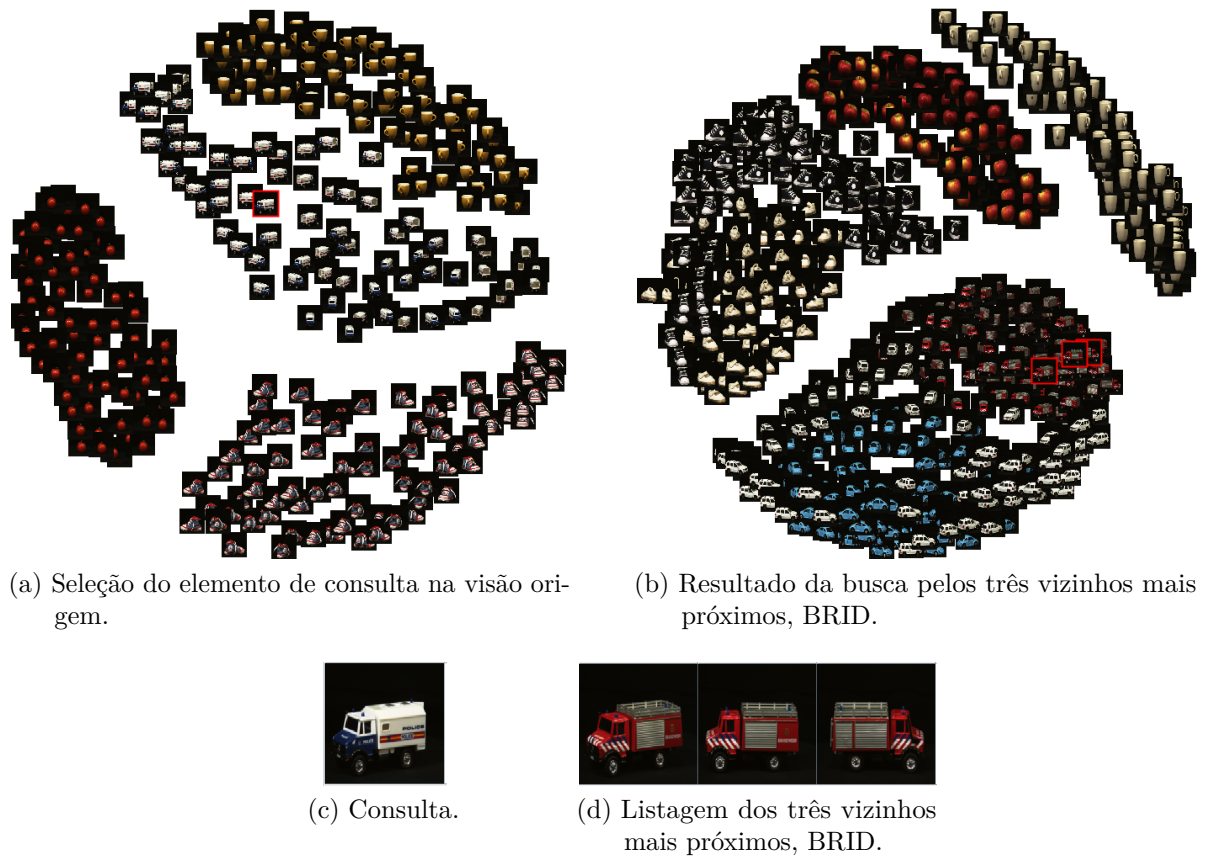


Figura 30 – Consulta por similaridade aplicando a técnica de diversidade BRID, utilizando a função de distância do cosseno.

5.2.1.4 Função de Distância Mahalanobis

A distância de mahalanobis baseia-se na correlação entre as características das instâncias. Nesse sentido, no experimento pressupõe-se que os conjuntos de dados das visualizações destino, em que são aplicadas a busca por similaridade, possuem n classes conhecidas, para as quais são calculadas matrizes distintas de covariância. Então, cada elemento do mesmo conjunto de dados, utiliza a sua respectiva matriz de covariância para gerar sua distância em relação aos demais. E seja um elemento arbitrário selecionado na visão origem, para medir a sua similaridade em relação aos outros elementos de conjuntos de dados em que se aplicam a busca, são calculadas n distâncias, utilizando cada uma das matrizes de covariância da visão destino, então a menor é escolhida, ou seja, o elemento considerado mais similar.

Os seguintes experimentos que aplicam a função de distância de mahalanobis, empregam a Config₉ na visualização origem, e nas visualizações destinos, nas quais executam-se as buscas pelos três vizinhos mais próximos, a Config₁₀, sendo que usam o mesmo subconjunto de dados que os experimentos de análise das funções euclidiana e cosseno. Porém, como o cálculo da matriz de covariância aumenta a complexidade da

abordagem, então optou-se na redução das características utilizadas na técnica, destacando que são subconjuntos distintos com atributos em comum. Na Figura 31a, a instância selecionada é destacada em vermelho, e então na visualização exibida na Figura 31b, são indicados os três vizinhos mais próximos, resultantes da busca por similaridade.

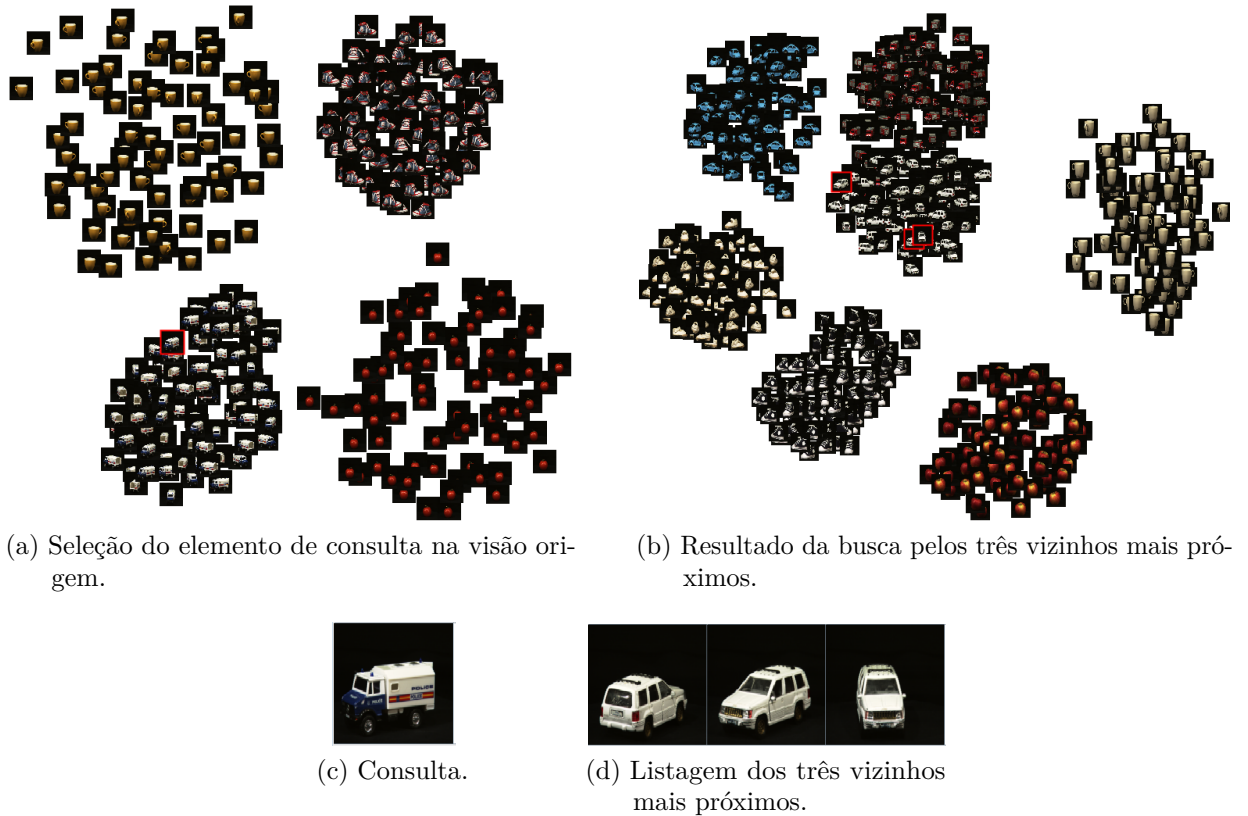


Figura 31 – Consulta por similaridade, utilizando a função de distância mahalanobis.

Na Figura 32a, o centro de consulta é selecionado na visão origem, e então a busca por similaridade associada à diversidade Motley é aplicada nas demais visões, ilustradas nas Figuras 32b, 32c e 32d. É notável, que quando o fator de diversidade θ é pequeno, a busca retorna os elementos dentro de uma mesma classe, a qual infere ser a pertencente do elemento de consulta, e então conforme esse fator aumenta, a resposta é formada por elementos que representam diferentes classes que são definidas como próximas, em relação à consulta realizada.

A técnica de diversidade MMR, aplicada às consultas por similaridade baseando-se na distância de mahalanobis, de acordo com o centro de consulta indicado na Figura 33a, não apresenta resultados satisfatórios, como pode ser identificado nas Figuras 33b, 33c e 33d. Conforme o parâmetro λ cresce, a técnica considera a diversidade entre os elementos do conjunto de resposta mais importante, e nesse cenário, com a função de distância mahalanobis, a abordagem retorna elementos muito diferentes ao centro de consulta, com parâmetros pequenos.

Da mesma maneira que a técnica MMR, lidando com a função de mahalanobis, a

técnica *Swap* também não apresenta um resultado adequado, como é mostrado nas Figuras 34b, 34c e 34d, em relação ao centro de consulta indicado na Figura 34a.

Na Figura 35a, o centro de consulta selecionado na visão origem é destacado por vermelho, e então na visão destino 35b, é indicado o resultado da busca por similaridade aplicando a técnica BRID. No âmbito da utilização da função de mahalanobis, em contraste à busca apenas por similaridade que retorna elementos de uma mesma classe, ao incluir o conceito de Variedade, são retornados elementos de outras classes que também são relacionadas ao centro de consulta, da mesma maneira que a técnica de Motley.

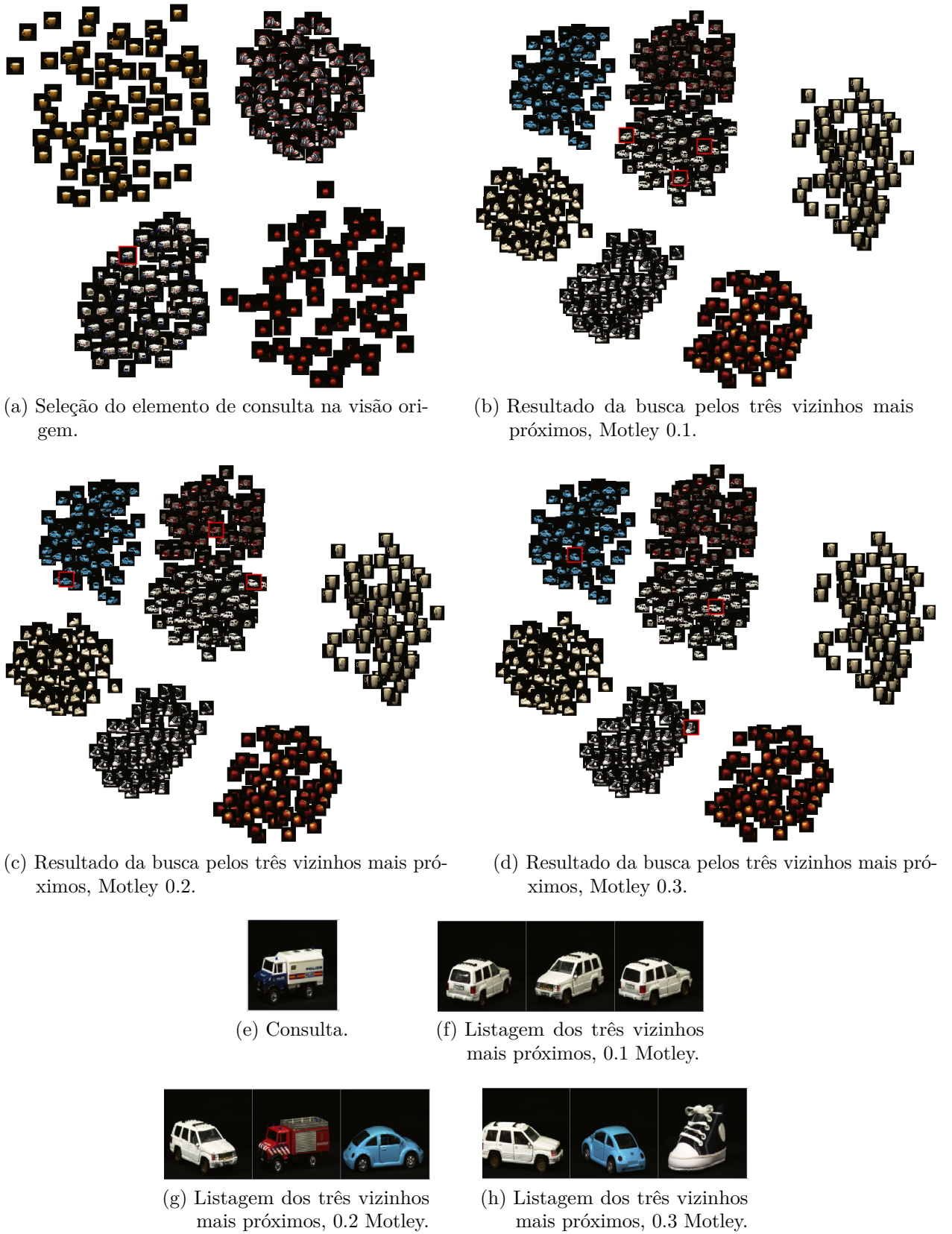
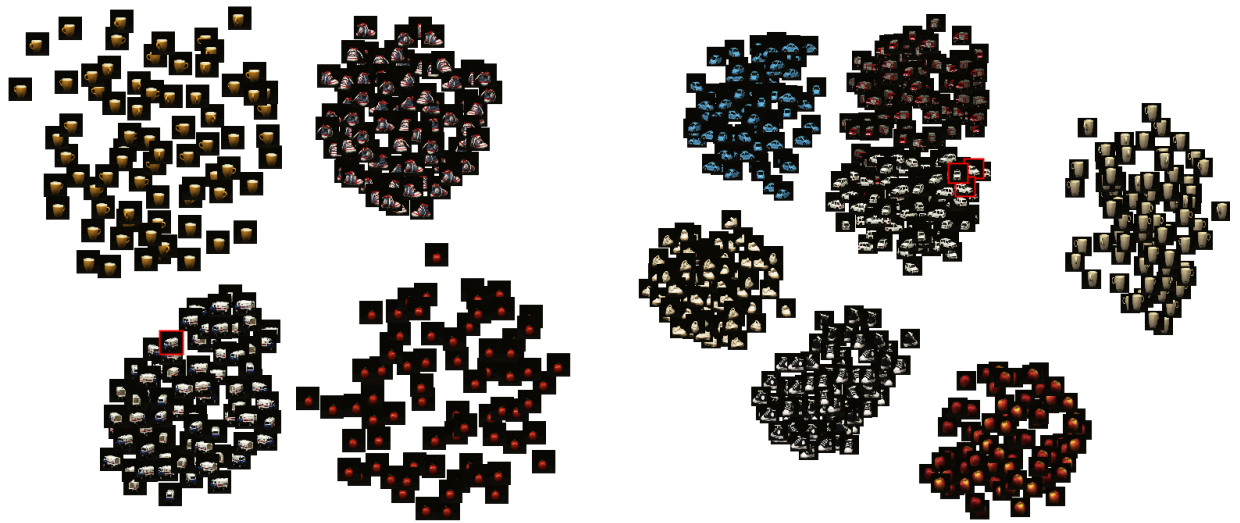
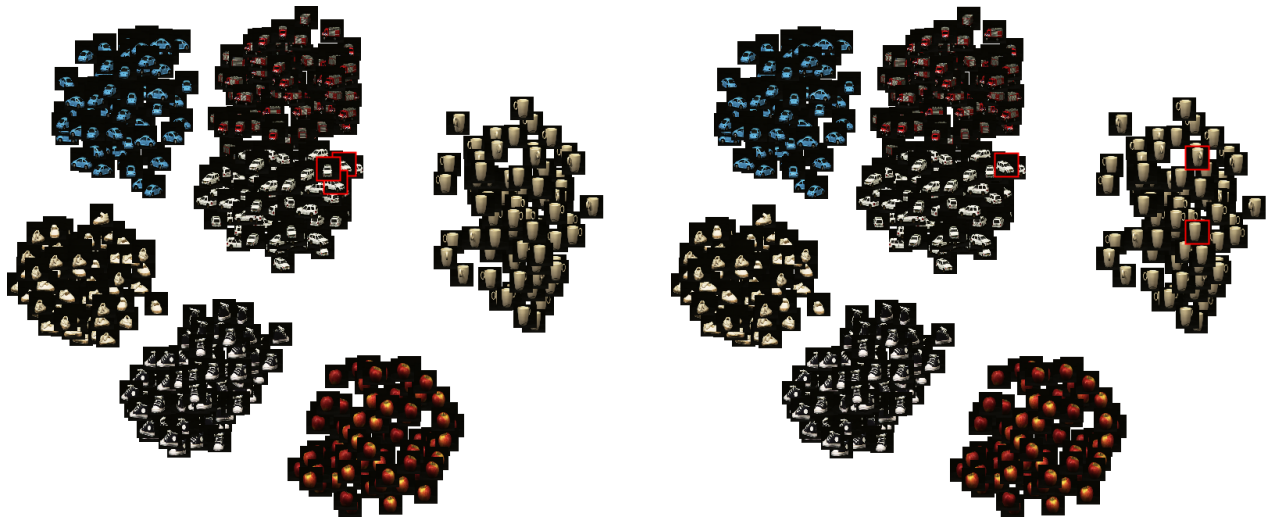


Figura 32 – Consulta por similaridade aplicando a técnica de diversidade Motley, utilizando a função de distância de mahalanobis.



(a) Seleção do elemento de consulta na visão original.

(b) Resultado da busca pelos três vizinhos mais próximos, MMR 0.2.

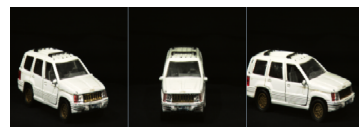


(c) Resultado da busca pelos três vizinhos mais próximos, MMR 0.3.

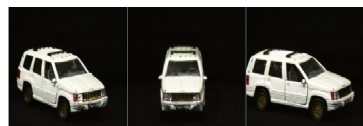
(d) Resultado da busca pelos três vizinhos mais próximos, MMR 0.5.



(e) Consulta.



(f) Listagem dos três vizinhos mais próximos, 0.2 MMR.



(g) Listagem dos três vizinhos mais próximos, 0.3 MMR.



(h) Listagem dos três vizinhos mais próximos, 0.5 MMR.

Figura 33 – Consulta por similaridade aplicando a técnica de diversidade MMR, utilizando a função de distância de mahalanobis.

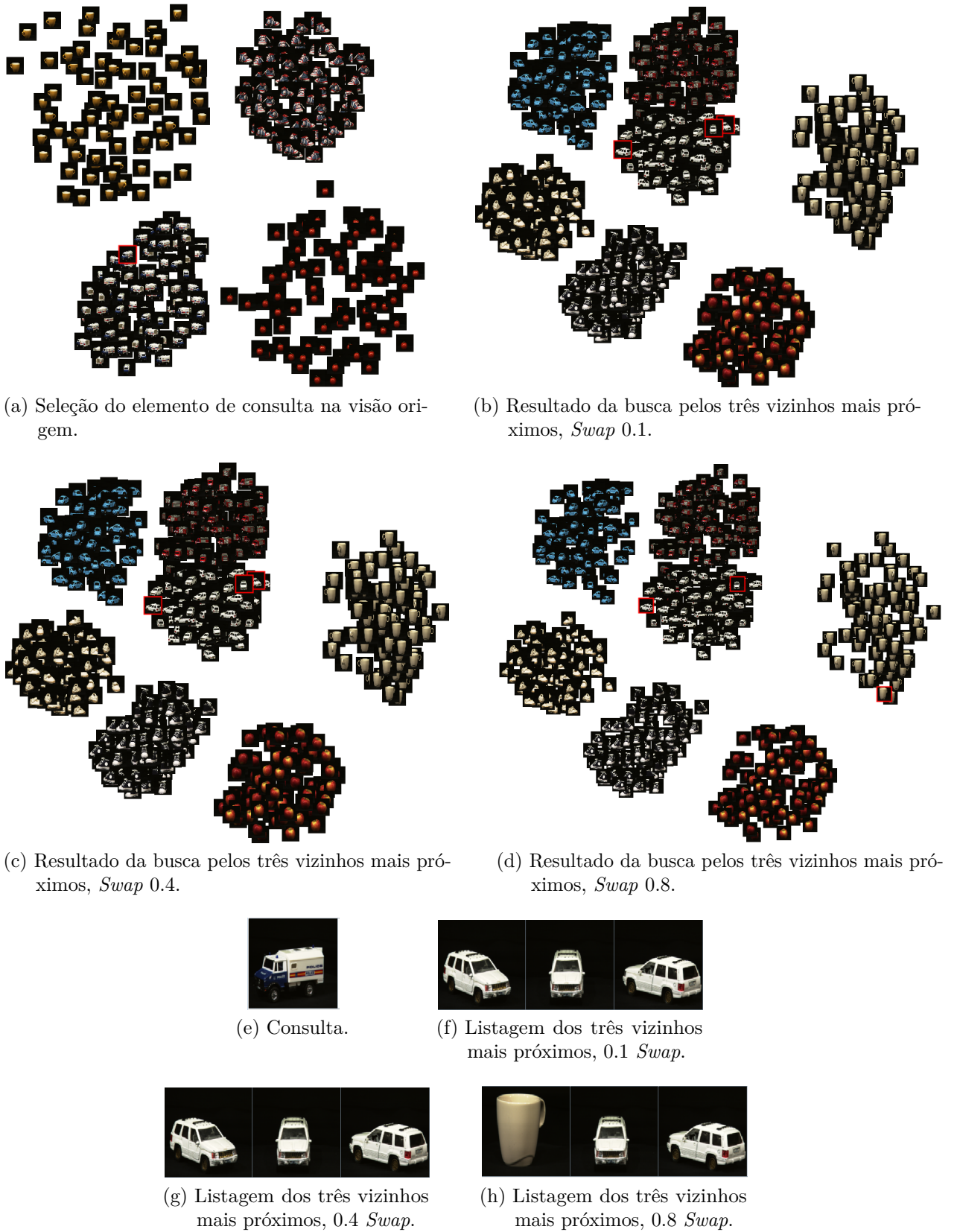


Figura 34 – Consulta por similaridade aplicando a técnica de diversidade *Swap*, utilizando a função de distância de mahalanobis.

Os experimentos apresentados nas Figuras 36 e 37, têm como objetivo introduzir

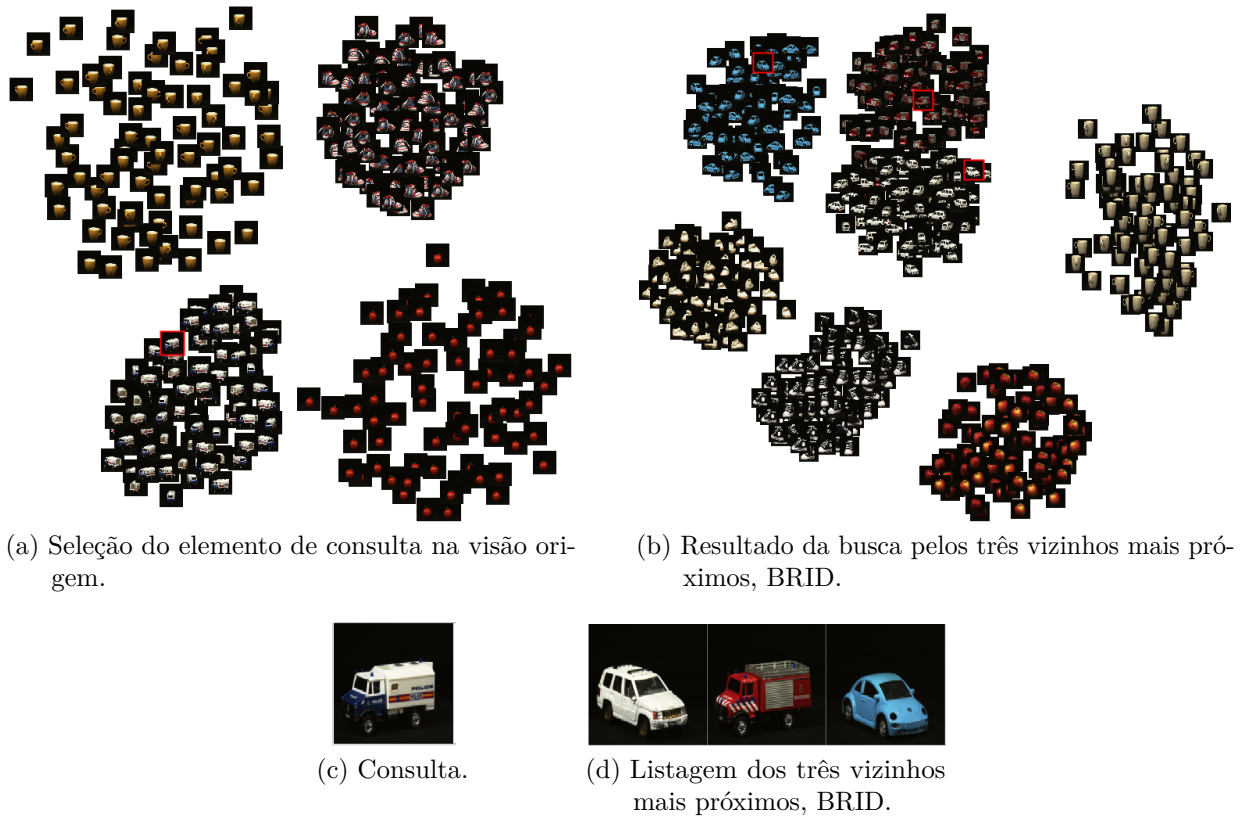


Figura 35 – Consulta por similaridade aplicando a técnica de diversidade BRID, utilizando a função de distância de mahalanobis.

mais resultados utilizando a função de mahalanobis nas consultas pela técnica de diversidade BRID, para evidenciar sua aplicabilidade em outros subconjuntos de dados. A visão origem é formada usando a $Config_1$, e nas visões destinos, em que se executa a busca pelos três vizinhos mais próximos, usa-se a $Config_2$, porém com uma quantidade reduzida de características, para facilitar o cálculo das matrizes de covariância. Na Figura 36a o centro de consulta é selecionado, e então na Figura 36b, é indicado o resultado da busca pelos três vizinhos mais próximos, calculando previamente as matrizes de covariância das classes conhecidas.

A partir da seleção do centro de consulta na Figura 37a, na Figura 37b é ilustrado o resultado da busca pela similaridade aplicando a técnica Motley com fator de diversidade θ em 0.3. Enquanto que, na Figura 37c, é mostrado o resultado aplicando a técnica de diversidade *Swap* com o valor de λ definido em 0.5. E então, na Figura 37d, é indicado o resultado utilizando a técnica BRID. É possível analisar, que a Variedade associada à função de distância mahalanobis, por meio da técnica BRID, apresenta bons resultados, pois retorna elementos de classes distintas, porém que são próximas à classe do elemento considerado mais próximo à consulta, conforme a base de dados onde se realiza a busca, eliminando a necessidade da definição prévia de um parâmetro de diversidade.

A estratégia da aplicação da função de distância mahalanobis nos experimentos,

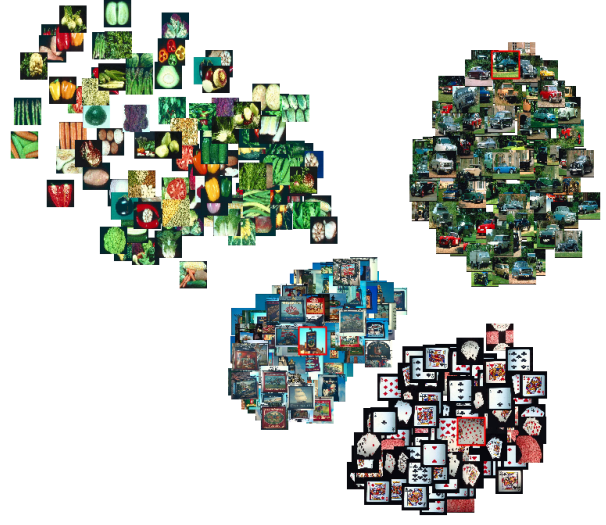


Figura 36 – Consulta por similaridade, utilizando a função de distância de mahalanobis.

carrega uma outra perspectiva de busca por similaridade aplicando a diversidade, em contraste com as funções euclidiana e cosseno, pois possibilita incluir o conceito de categorização, permitindo que o usuário encontre uma resposta diversa, que contém elementos que representam suas respectivas classes, que são consideradas próximas ao centro de consulta, além de identificar na visualização como esse relacionamento é definido no conjunto de dados.



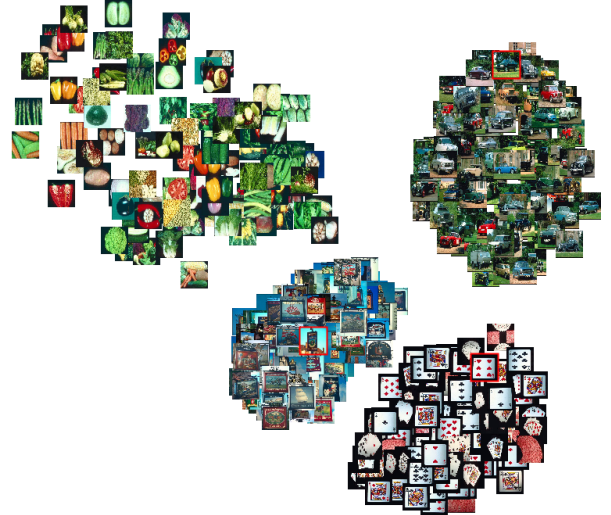
(a) Seleção do elemento de consulta na visão origem.



(b) Resultado da busca pelos três vizinhos mais próximos, Motley 0.3.



(c) Resultado da busca pelos três vizinhos mais próximos, Swap 0.5.



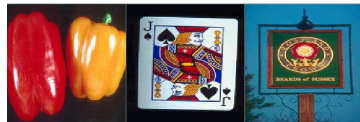
(d) Resultado da busca pelos três vizinhos mais próximos, BRID.



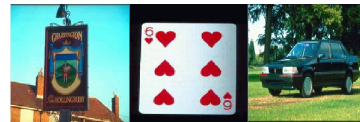
(e) Consulta.



(f) Listagem dos três vizinhos mais próximos, 0.3 Motley.



(g) Listagem dos três vizinhos mais próximos, 0.5 Swap.



(h) Listagem dos três vizinhos mais próximos, BRID.

Figura 37 – Consulta por similaridade aplicando as técnicas de diversidade Motley, Swap e BRID, utilizando a função de distância de mahalanobis.

5.2.2 ANÁLISE DOS MÉTODOS DE DIVERSIDADE

A maior parte das técnicas de diversidade requerem um fator prévio de diversidade definido pelo usuário, e nos seguintes experimentos, empregando a função de distância euclidiana, é analisado como as técnicas de diversidade se comportam com o aumento gradativo dos fatores de cada uma, verificando qual o impacto gerado na listagem dos elementos e na visualização do conjunto de dados em sua totalidade.

A partir do centro de consulta, ilustrado na Figura 38a, selecionado na visão origem, formada pela configuração Config₃, baseando-se na distância euclidiana, na visualização destino, exibida na Figura 38c, construída pela configuração Config₄, são indicados os quatro vizinhos mais próximos, resultantes da busca aplicando a similaridade.

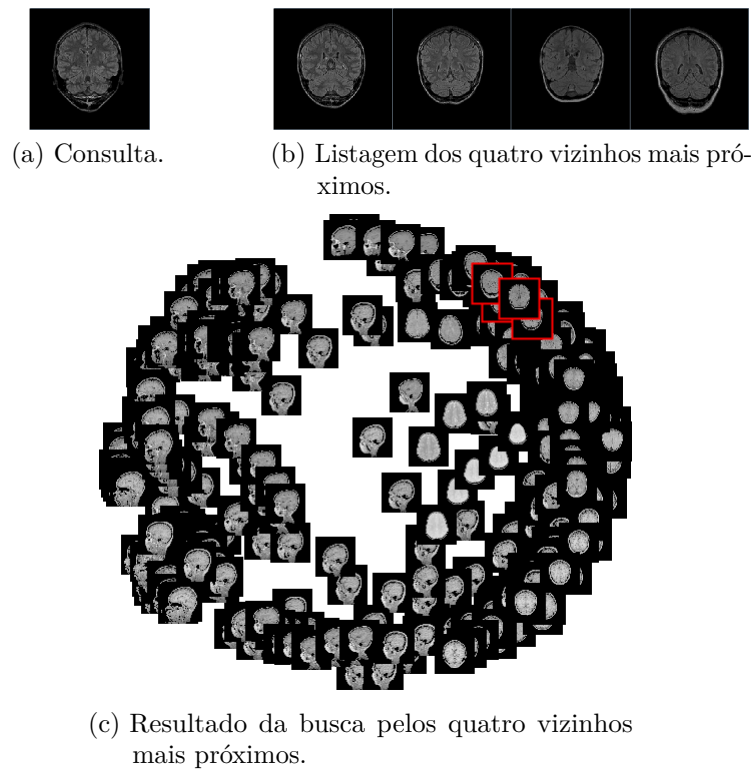


Figura 38 – Consulta por similaridade.

5.2.2.1 MMR

Com base no centro de consulta indicado na Figura 39a, nas Figuras 39e, 39f e 39g são indicados os resultados da busca aplicando a técnica de diversidade MMR. Como é verificado, conforme o parâmetro λ aumenta, a listagem dos elementos contidos nos resultados é diversificada, e também na visualização, os elementos distanciam-se entre si, devido ao fato da técnica medir incrementalmente, a diversidade entre os elementos presentes no conjunto de resposta.

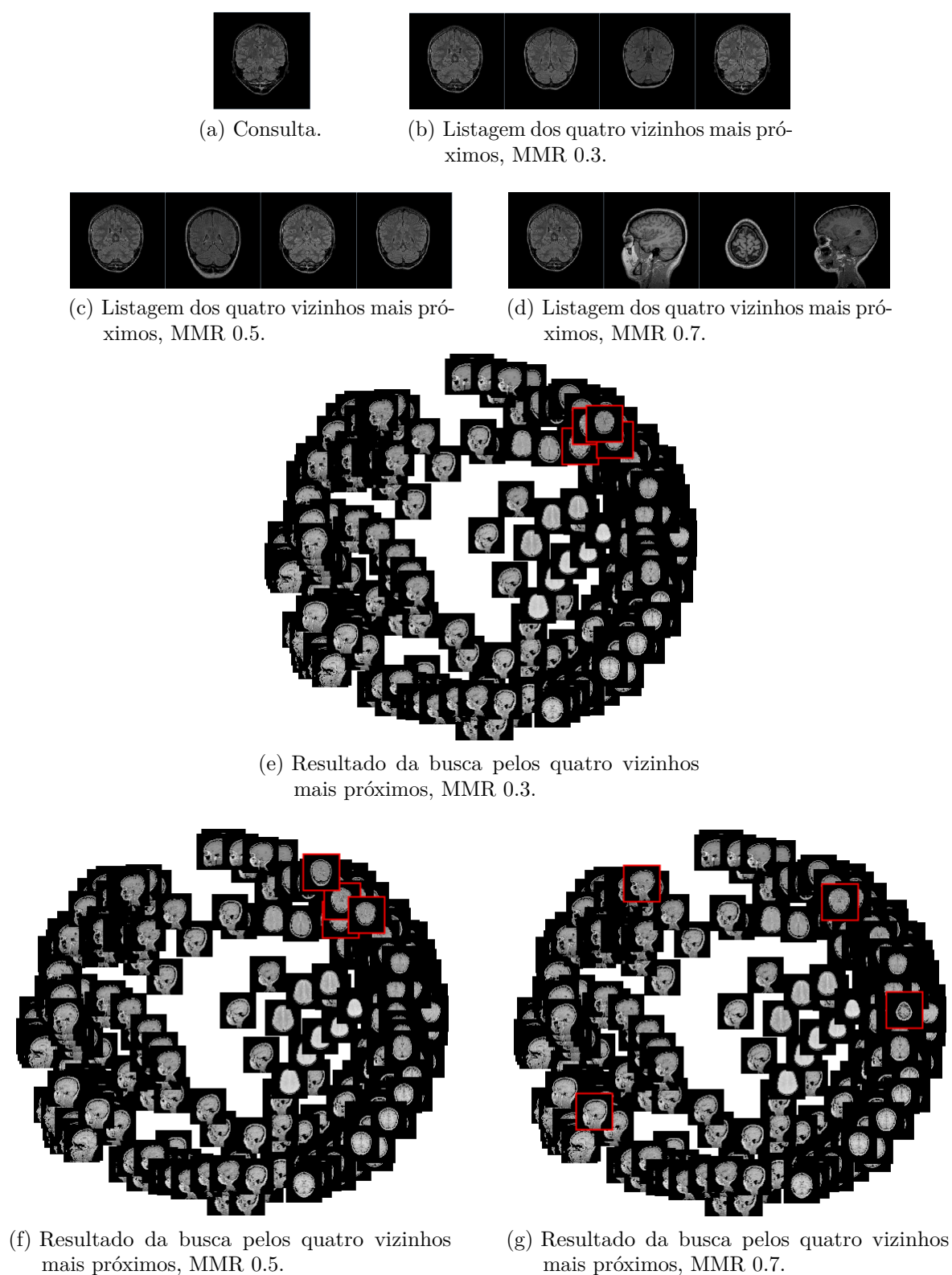


Figura 39 – Consulta por similaridade aplicando a técnica de diversidade MMR.

5.2.2.2 MSD

A Figura 40a ilustra o centro de consulta, e então nas Figuras 40e, 40f e 40g são apresentados os resultados da busca por similaridade e diversidade MSD (GOLLA-PUDI; SHARMA, 2009). A abordagem MSD seleciona pares de elementos, que sejam simultaneamente similares em relação ao elemento de consulta, mas diversos entre si, sendo que conforme o fator λ cresce, aumenta também a importância da diversidade entre os elementos do par escolhido para compor o conjunto de resposta. Nos resultados do experimento efetuado, pode-se analisar que mesmo o fator de diversidade aumentando, as soluções calculadas e alcançadas são muito parecidas.

5.2.2.3 Swap

De acordo com o centro de consulta, ilustrado na Figura 41a, nas demais Figuras 41e, 41f e 41g são indicados os quatro vizinhos mais próximos resultantes das buscas por similaridade aplicando a diversidade *Swap*, a qual define o conjunto de resposta inicial com os k elementos mais próximos, e então verifica se ao trocar um elemento da resposta com outro arbitrário no conjunto de dados, melhora o resultado final, com base na maximização de uma função que leva em consideração a diversidade da resposta e similaridade em relação ao centro de consulta. Como pode ser identificado no experimento, quando o fator λ é definido para 0.5, a resposta apresenta elementos bem diversos em relação à consulta, e conforme esse fator aumenta, como no exemplo para 0.7, a resposta continua parecida com a anterior, porém os elementos são trocados por outros, por otimizarem a função objetivo.

5.2.2.4 Motley

Na Figura 42a, é indicado o centro de consulta, que é utilizado nas buscas pela similaridade e diversidade Motley, apresentadas nas Figuras 42e, 42f e 42g. É notável que a abordagem Motley, insere um nível de diversidade ao conjunto de resposta conforme o fator θ cresce. Porém em comparação às outras abordagens, é possível identificar nas visualizações, que esse fator de diversidade impacta o conjunto de resposta de uma maneira mais sutil, retornando elementos sempre em torno do considerado mais próximo ao centro de consulta.

5.2.2.5 GMC

Na Figura 43a é indicado o elemento selecionado para consulta, e nas Figuras 43e, 43f e 43g, são exibidos os resultados da busca por similaridade aplicando a diversidade GMC (VIEIRA et al., 2011b), que compõe a resposta maximizando a função objetivo, que leva em consideração a similaridade em relação ao centro de consulta, a diversidade do conjunto de resposta, e a diversidade em relação aos elementos no conjunto de dados da visão destino. De acordo com os resultados obtidos com o experimento, apresentados nas

visualizações, analisa-se que conforme o fator de diversidade λ cresce, e são adicionados elementos ao conjunto de resposta, então são escolhidos elementos mais distantes ao considerado mais próximo do centro de consulta.

5.2.2.6 GNE

Os resultados das buscas aplicando a técnica de diversidade GNE (VIEIRA et al., 2011b), com valor de α definido em 0.0, são mostrados nas Figuras 44e, 44f e 44g, em relação ao centro de consulta destacado na Figura 44a. Como abordado anteriormente, a técnica GNE utiliza a mesma função objetivo que a técnica GMC, porém acrescenta o valor α , que quanto maior, apresenta um conjunto de resposta mais variado, devido a randomização adotada na etapa de busca local. Nesse sentido, é possível verificar que nos resultados da técnica GNE, com o valor de α em 0.5, ilustrados nas Figuras 45e, 45f e 45g, os elementos estão mais dispersos entre si nas visões destino.

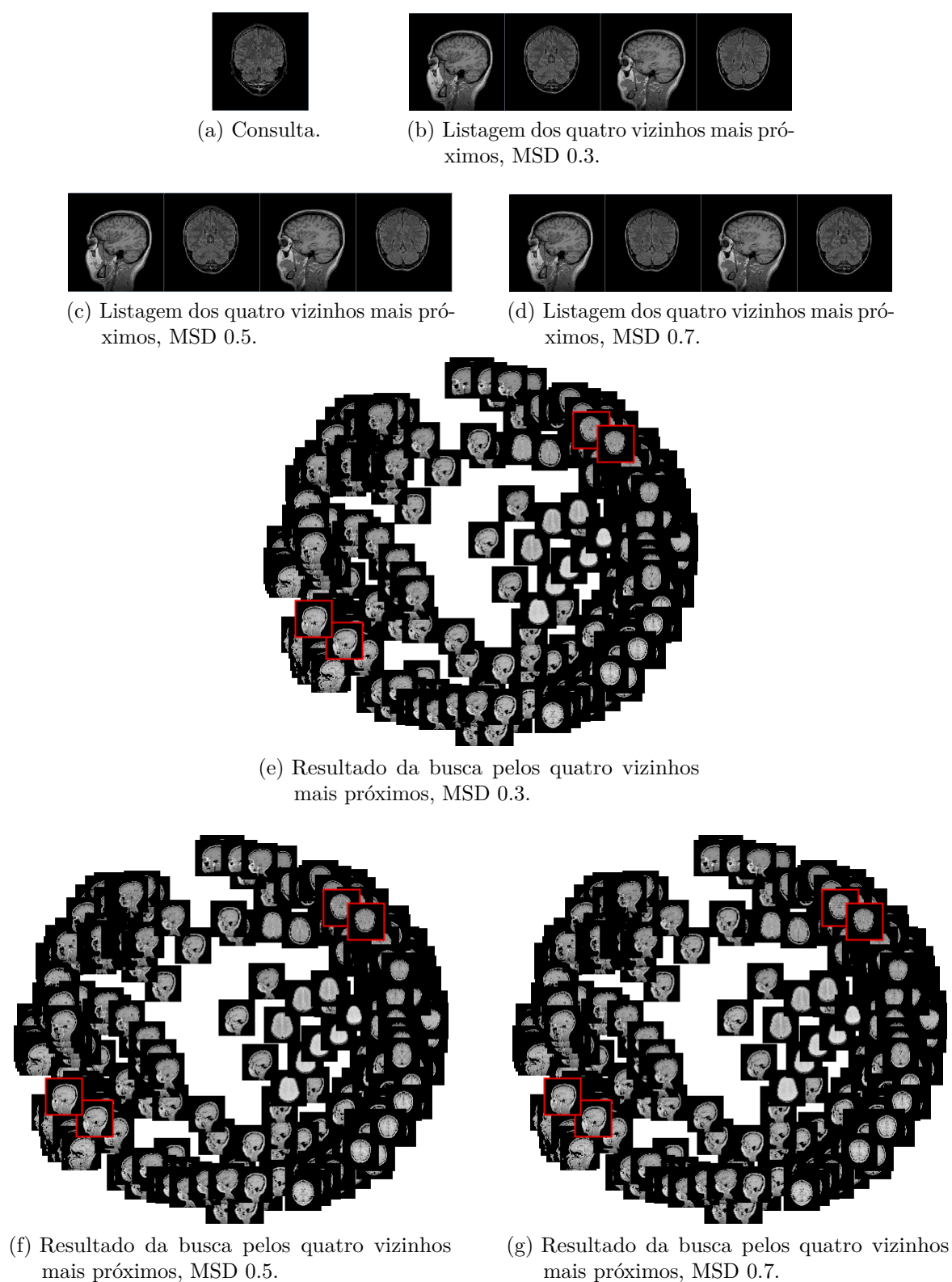


Figura 40 – Consulta por similaridade aplicando a técnica de diversidade MSD.

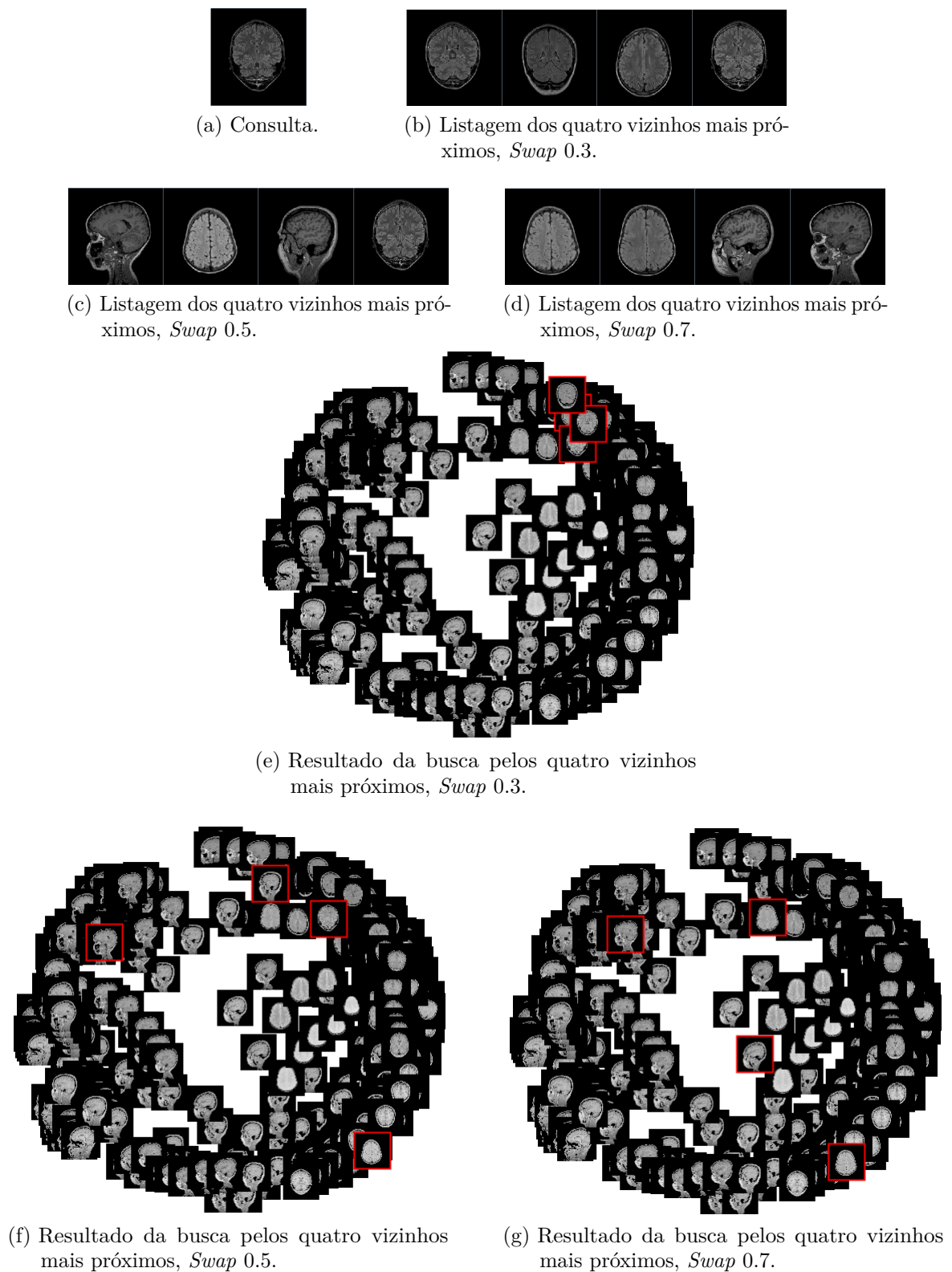


Figura 41 – Consulta por similaridade aplicando a técnica de diversidade *Swap*.

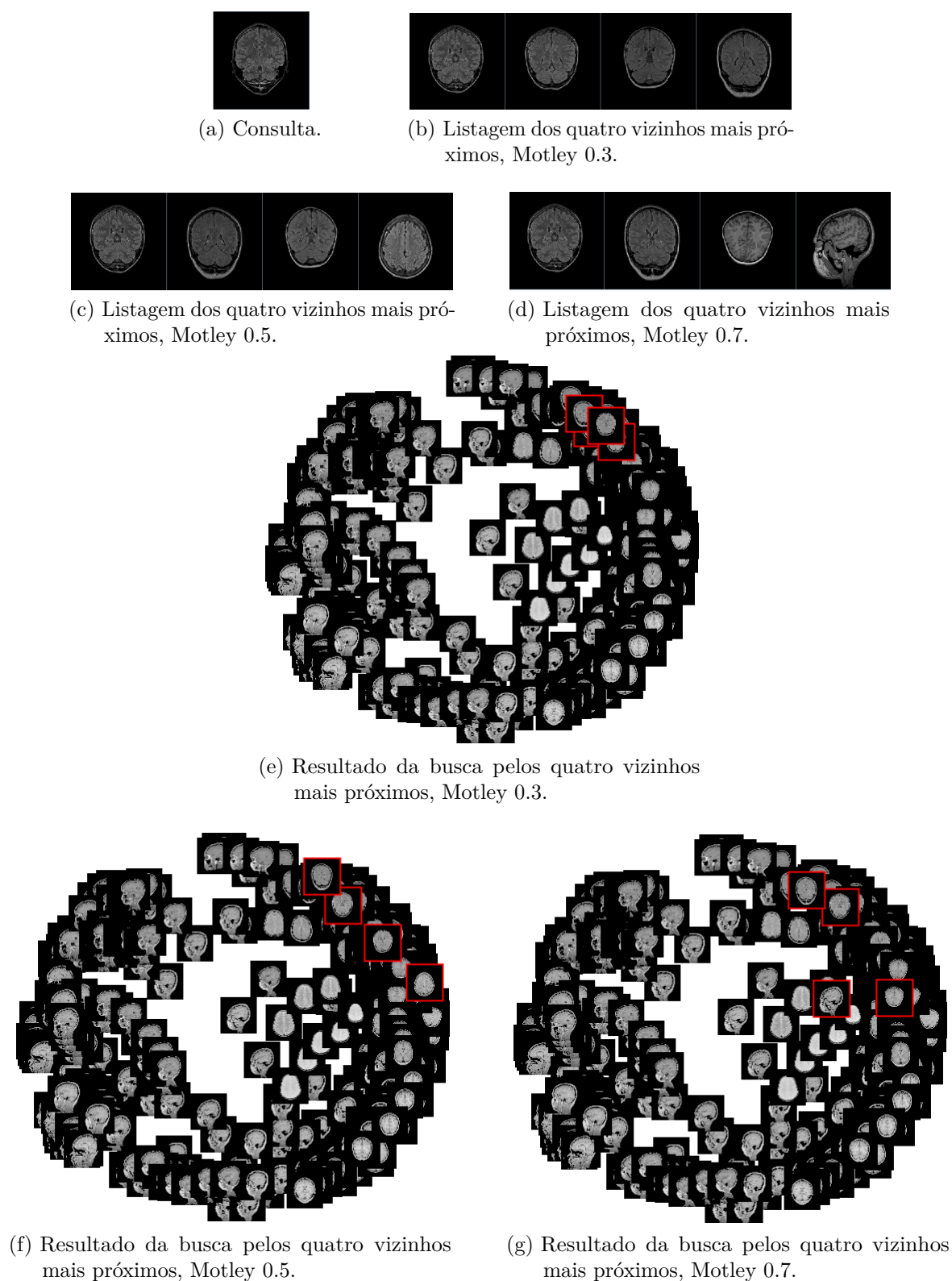


Figura 42 – Consulta por similaridade aplicando a técnica de diversidade Motley.

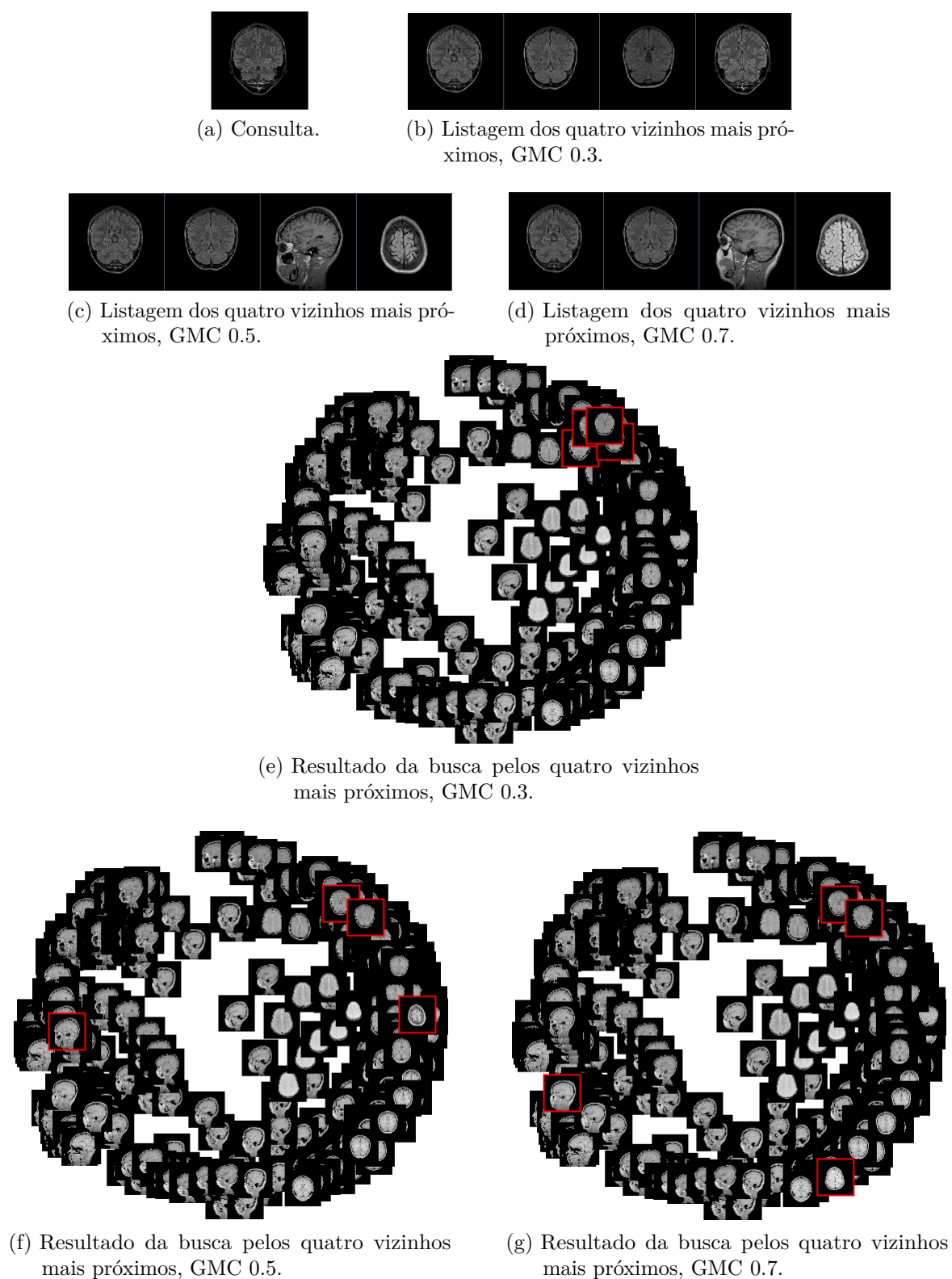


Figura 43 – Consulta por similaridade aplicando a técnica de diversidade GMC.

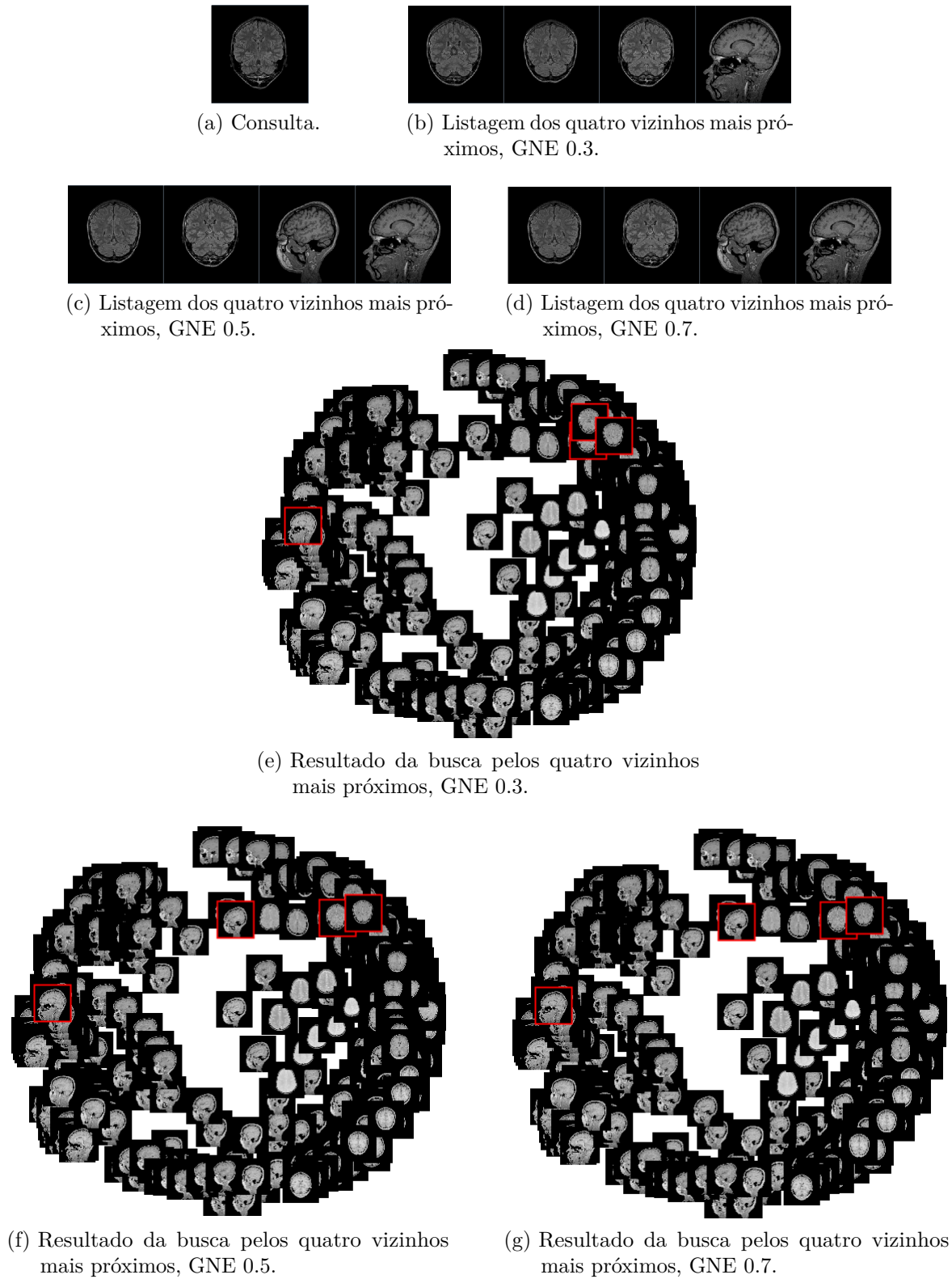


Figura 44 – Consulta por similaridade aplicando a técnica de diversidade GNE, com $\alpha = 0.0$.

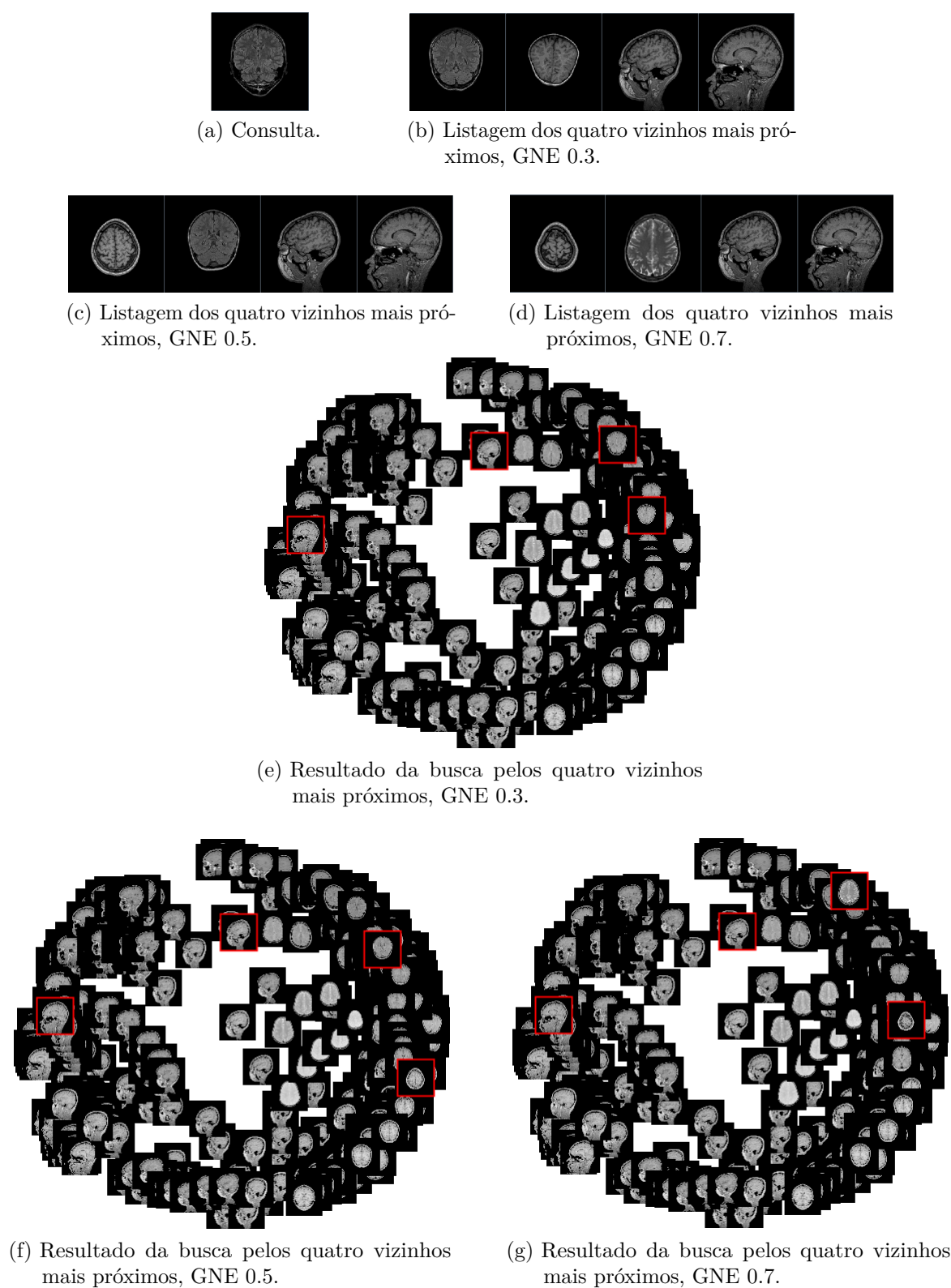


Figura 45 – Consulta por similaridade aplicando a técnica de diversidade GNE, com $\alpha = 0.5$.

5.2.2.7 BRID

Na Figura 46a, o elemento central de consulta é destacado, e então o resultado da busca por similaridade, aplicando a técnica de diversidade BRID, é exibido na Figura 46c. De acordo com as análises dos experimentos anteriores, utilizando diferentes tipos de funções de distância, pode-se identificar que a Variedade associada à função euclidiana, retorna elementos bem próximos ao centro de consulta, porém com um certo nível de relevância, evitando elementos repetidos. Como pode ser observado nos resultados apresentados nas Figuras 47b e 47c, em relação ao centro de consulta apontado na Figura 47a, os quais empregam as configurações de experimento Config₁₁ e Config₁₂, respectivamente.

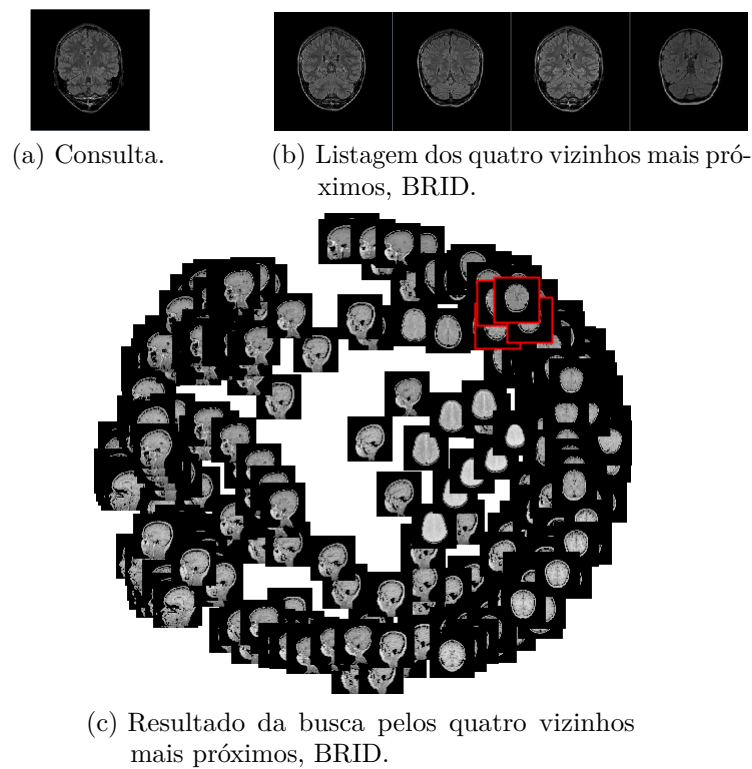


Figura 46 – Consulta por similaridade aplicando a técnica de diversidade BRID.

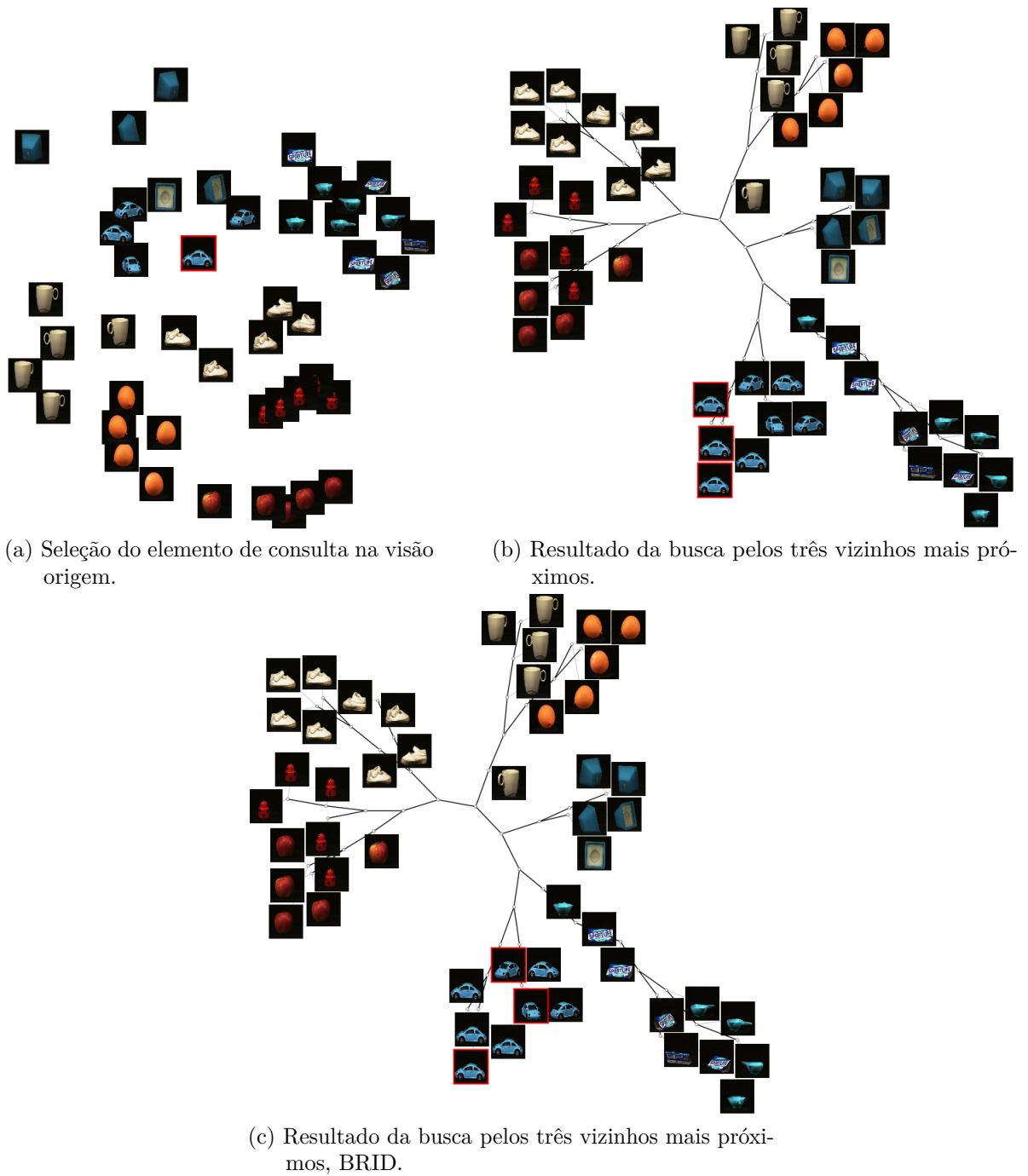


Figura 47 – Consulta por similaridade aplicando a técnica de diversidade BRID.

5.3 CONSIDERAÇÕES FINAIS

Neste Capítulo são descritos os experimentos planejados e executados, com o objetivo de demonstrar o desempenho das técnicas de diversidade implementadas e vinculadas no processo de coordenação entre as visualizações.

É possível identificar por meio dos resultados obtidos, como as técnicas de diversidade aplicadas nas consultas por similaridade se comportam utilizando as diferentes

funções de distância: euclidiana, cosseno e mahalanobis, sendo que na utilização da distância de mahalanobis, para definir as n classes é preciso ter um conhecimento prévio da base de dados em que se deseja efetuar a busca.

De acordo com as consultas efetuadas utilizando diferentes técnicas de diversidade, é possível analisar diretamente nas projeções o efeito dos parâmetros de diversidade, identificando o quanto as imagens recuperadas, pertencentes ao conjunto de resposta, estão próximas umas das outras, com base na disposição entre elas, mediante à capacidade de abstração viabilizada pela visualização coordenada. Salientando-se que a técnica BRID apresenta vantagem em relação às demais, por não ser necessário a determinação de um fator de diversidade prévio.

6 CONCLUSÃO E TRABALHOS FUTUROS

6.1 CONSIDERAÇÕES INICIAIS

Os sistemas de visualização de dados auxiliam as pessoas a assimilarem os relacionamentos entre seus elementos, revelando informações que não seriam percebidas sem o poder da abstração da visualização. Existem diferentes técnicas de visualização aplicadas em contextos distintos e, nesse sentido, Múltiplas Visões Coordenadas viabilizam a associação de diferentes representações gráficas, por meio de mapeamentos específicos entre elas (MCEWAN; IGONIDERIGHA; BENYON, 2014).

Este trabalho apresentou uma evolução das técnicas de coordenação que lidam com dados complexos, gerando novos mapeamentos baseando-se na similaridade em representações gráficas bidimensionais (ELER, 2011), elaboradas por meio de projeção multidimensional, fazendo com que essas técnicas levem em consideração a diversidade dos elementos retornados nas consultas, permitindo visualizar relações menos homogêneas e mais relevantes.

6.2 ANÁLISE DOS RESULTADOS

Os experimentos desenvolvidos neste trabalho tiveram como objetivo apresentar os resultados obtidos de acordo com o desenvolvimento da proposta do trabalho, mostrando tanto a listagem dos elementos encontrados na busca pela similaridade e diversidade, quanto a seleção dos mesmos nas visualizações envolvidas no processo de coordenação.

Uma comparação foi efetuada entre as técnicas de diversidade, Motley (JAIN; SARDA; R., 2004), MMR (CARBONELL; GOLDSTEIN, 1998), *Swap* (YU; LAKSHMANAN; AMER-YAHIA, 2009) e BRID (SANTOS et al., 2013), aplicadas na busca por similaridade, para mapear elementos na coordenação entre as visões, utilizando conjuntos de dados distintos e as funções de distância: euclidiana, cosseno e mahalanobis. As funções euclidiana e cosseno foram calculadas igualmente para todo o conjunto de dados, sendo que ambas obtiveram resultados parecidos, principalmente na disposição das instâncias nas projeções. Já a função de mahalanobis, foi calculada de maneira diferente, em que foram definidas n classes de imagens, as quais foram utilizadas na computação da matriz inversa de covariância, a partir dos vetores de elementos de cada uma, e então, posteriormente, foram aplicadas no cálculo das distâncias.

Pode-se notar que ao usar a distância de mahalanobis para compor as visões, os elementos que pertencem à mesma classe ficam mais próximos, porém no desenvolvimento

algumas dificuldades foram encontradas. Então foi necessário efetuar alguns ajustes, como a redução da dimensão do vetor de características devido à complexidade, e também o aumento expressivo da quantidade dos elementos nos conjuntos de dados manipulados nos experimentos, com o objetivo de ter amostras suficientes para o cálculo da matriz de covariância e sua inversa. Nesse sentido, embora os resultados aplicando a distância de mahalanobis sejam satisfatórios, há uma certa dificuldade no contexto prático da aplicação, agregada pela complexidade do cálculo e armazenamento da matriz de covariância.

A diversificação em consultas por similaridade envolve dois conceitos que competem entre si e, em geral, as técnicas procuram encontrar um equilíbrio entre eles, para fornecer um resultado significativo, caracterizando o problema como uma otimização de bi-critério. As técnicas percorridas anteriormente, MMR, MSD (GOLLAPUDI; SHARMA, 2009), *Swap*, GMC e GNE (VIEIRA et al., 2011b), buscam maximizar diferentes funções objetivos mediante a um *tradeoff*, que define o nível de diversificação dos resultados, por meio de uma combinação linear entre ambas definições, tornando o problema NP-difícil (VIEIRA et al., 2011a). Porém as técnicas MMR, MSD e GMC constroem a solução incrementalmente, enquanto a *Swap* baseando-se na troca, e a técnica GNE, emprega meta-heurísticas para otimizar a resposta. Com base nos experimentos, tais técnicas apresentaram resultados bons e condizentes ao esperado, de acordo com o crescimento dos seus parâmetros de diversidade. Entretanto, a complexidade proporcional à quantidade de elementos no conjunto de dados utilizado é um fator negativo. A técnica Motley também encontra a resposta incrementalmente, porém introduz uma estratégia de busca gulosa, aplicando um limiar de diversidade previamente definido, sendo que para níveis baixos de diversidade, a complexidade é bem vantajosa em relação às outras técnicas e os resultados, em geral, são próximos ao da solução ótima, conforme pode ser verificado nos experimentos.

A variedade (SANTOS, 2012) é um conceito promissor, principalmente por não ser necessário definir anteriormente um limiar de diversidade, evitando a exigência do usuário ter um conhecimento prévio do conjunto de dados, para que obtenha uma boa solução. E nos experimentos, sua aplicabilidade é verificada, mediante a técnica de diversidade BRID. Baseando-se nos resultados pode-se verificar que, quando o conjunto de dados é diversificado, a variedade não tem tanto impacto na solução, retornando elementos parecidos ao da busca apenas pela similaridade, ou seja, não se aplica para buscas em que se deseja resultados com alta diversidade. Porém em conjuntos de dados em que há uma redundância considerável, os resultados são altamente satisfatórios, resultando em um conjunto de k elementos mais similares ao centro de consulta, sem permitir apresentar elementos que sejam tão próximos entre si, que não agreguem informação relevante na resposta.

De acordo com os resultados, é possível analisar o quão enriquecedor e esclarecedor é observar o relacionamento mediante a disposição dos elementos em um mesmo conjunto

de dados projetados na visualização bidimensional, assim como, perceber relacionamentos entre conjuntos distintos e suas respectivas visualizações, mediante a coordenação entre elas. Visto que, o mecanismo de coordenação permite identificar o que a diversidade produz nas consultas por similaridade, e por meio dos mapeamentos realizados, associar tais elementos em visualizações com projeções, funções de distâncias e dados diferentes.

6.3 CONTRIBUIÇÕES

Com o atual desenvolvimento das tecnologias e consequente produção de dados em larga escala em diversos segmentos, os diferentes tipos de visualizações e a coordenação entre elas se destacam, em permitir a análise e compreensão de conjuntos de dados, de maneira facilitada e rápida. Nesse sentido, agregar o poder da similaridade e diversidade, em contextos de alta redundância de informação, com o objetivo de extrair o importante, essencial e relevante, é algo que vale muito, principalmente na atualidade, onde o usuário requer objetividade.

A principal contribuição do trabalho é o desenvolvimento de novos mapeamentos e inclusão no processo de coordenação entre as visualizações, mediante a consulta pela similaridade e diversidade, cooperando para a análise e identificação de novos relacionamentos entre elementos de diferentes conjuntos de dados, principalmente para diversos tipos de dados complexos.

A segunda contribuição e aplicação da abordagem proposta é o apoio para a análise das técnicas de diversidade, sobretudo as que necessitam da definição de um parâmetro prévio para a obtenção da resposta desejada. Em geral, as abordagens de recuperação de dados por similaridade e diversidade retornam o conjunto de resposta no formato de listagem, porém possibilitar a identificação desses elementos no conjunto de dados em sua vizinhança e totalidade, por meio da visualização e coordenação, permite compreender o efeito causado na alteração dos parâmetros envolvidos, alcançando de maneira eficiente e rápida, um fator de diversidade adequado.

6.4 TRABALHOS FUTUROS

Uma dificuldade encontrada no desenvolvimento do trabalho é a complexidade das técnicas de diversidade abordadas. Nesse sentido, é possível implementar estratégias de otimização dessas metodologias, como programação dinâmica para reduzir recálculos desnecessários, aplicando por exemplo, a abordagem de [SANTOS et al. \(2015\)](#), na qual o processo de diversificação das consultas por similaridade é dividido em duas etapas: filtragem dos melhores candidatos e então a computação efetiva da diversidade, apresentando bons resultados na melhora do desempenho de técnicas como, MMR, MSD, *Swap*, GMC e

GNE.

Acrescentaria de maneira significativa no trabalho, a realização de novos experimentos com outros tipos de dados complexos, como sons, séries temporais, dados geográficos, entre outros, bem como fazer uma análise da usabilidade do mapeamento desenvolvido no processo de coordenação na ferramenta PEx.

REFERÊNCIAS

- BALAN, A. G. R. *Métodos adaptativos de segmentação aplicados à recuperação de imagens por conteúdo*. Tese (Doutorado), 2007. Citado na página 45.
- BALDONADO, M. Q. W.; WOODRUFF, A.; KUCHINSKY, A. Guidelines for using multiple views in information visualization. In: *Working Conference on Advanced Visual Interfaces*. [S.l.: s.n.], 2000. p. 110–119. Citado na página 4.
- BORGES, E.; DORNELES, C. Pgsimilar: uma ferramenta open source para suporte a consultas por similaridade no postgresql. In: *SESSÃO DE DEMOS - SIMPÓSIO BRASILEIRO DE BANCOS DE DADOS*. [s.n.], 2006. Disponível em: <<http://repositorio.furg.br/handle/1/1696>>. Citado na página 27.
- BUENO, R. *Desenvolvimento de algoritmos genéricos para consultas por similaridade em domínios métricos*. Dissertação (Mestrado), 2005. Citado 3 vezes nas páginas 21, 22 e 24.
- BUGATTI, P. H. *Análise da influência de funções de distância para o processamento de consultas por similaridade em recuperação de imagens por conteúdo*. Dissertação (Mestrado), 2008. Citado 2 vezes nas páginas 20 e 25.
- CARBONELL, J.; GOLDSTEIN, J. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.: s.n.], 1998. p. 335–336. Citado 6 vezes nas páginas 2, 26, 28, 37, 47 e 83.
- CARD, S.; MACKINLAY, J.; SHNEIDERMAN. *Readings in Information Visualization Using Vision to Think*. First edition. [S.l.]: Morgan Kaufmann, 1999. Citado na página 1.
- CUADROS, A. M. et al. Point placement by phylogenetic trees and its application for visual analysis of document collections. In: IEEE COMPUTER SOCIETY. *IEEE Symposium on Visual Analytics Science and Technology*. [S.l.], 2007. p. 99–106. Citado na página 5.
- DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection: Ieee computer society conference. In: *Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2005. v. 1, p. 886–893. Citado na página 46.
- DENG, T.; FAN, W. On the complexity of query result diversification. *ACM Transactions on Database System*, v. 39, n. 15, p. 15:1–15:46, 2014. Citado 2 vezes nas páginas 26 e 27.
- DIAS, R. L.; BUENO, R.; RIBEIRO, M. X. Reducing the complexity of k-nearest diverse neighbor queries in medical image datasets through fractal analysis. In: *26th International Symposium on Computer-Based Medical Systems (CBMS)*. [S.l.: s.n.], 2013. p. 101–106. Citado na página 27.
- DIAS, R. L. et al. Visual-interactive k-ndn method (vik): A novel approach to visualize and interact with content-based image retrieval systems regarding similarity and diversity. In: *21st International Conference Information Visualisation*. [S.l.: s.n.], 2017. Citado na página 27.

- DROSOU, M.; PITOURA, E. Search result diversification. *ACM SIGMOD Record*, v. 39, p. 41–47, 2010. Citado 2 vezes nas páginas 26 e 27.
- ELER, D. M. *Múltiplas visões coordenadas para exploração de mapas de similaridade*. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, May 2011. Citado 11 vezes nas páginas 5, 6, 9, 11, 12, 13, 14, 16, 18, 38 e 83.
- ELER, D. M. et al. Multidimensional visualization to support analysis of image collections. In: IEEE COMPUTER SOCIETY. *XXI Brazilian Symposium on Computer Graphics and Image Processing*. [S.l.], 2008. p. 289–296. Citado na página 4.
- ELER, D. M. et al. Visual analysis of image collections. *The Visual Computer*, v. 25, n. 10, p. 923–937, June 2009. Citado na página 2.
- ELER, D. M. et al. Topic-based coordination for visual analysis of evolving document collections. In: IEEE COMPUTER SOCIETY. *13th International Conference Information Visualisation*. [S.l.], 2009. p. 149–155. Citado 4 vezes nas páginas 2, 7, 10 e 11.
- ELER, D. M.; PRATES, J. M.; GARCIA, M. B. Coordinated multiple views to support image retrieval. In: IEEE COMPUTER SOCIETY. *18th International Conference on Information Visualisation*. [S.l.], 2014. p. 139–144. Citado na página 16.
- GEUSEBROEK, J.-M.; BURGHOUTS, G. J.; SMEULDERS, A. W. M. The amsterdam library of object images. *International Journal of Computer Vision*, v. 61, p. 103–112, 2005. Disponível em: <<http://aloi.science.uva.nl/>>. Citado na página 45.
- GODINHO, P. I. A. et al. Prisma - a multidimensional information visualization tool using multiple coordinated views. In: IEEE COMPUTER SOCIETY. *11th International Conference Information Visualization*. [S.l.], 2007. p. 23–32. Citado na página 1.
- GOLLAPUDI, S.; SHARMA, A. An axiomatic approach for result diversification. In: *Proceedings of the 18th international conference on World wide web*. [S.l.: s.n.], 2009. p. 381–390. Citado 5 vezes nas páginas 26, 29, 37, 72 e 84.
- GRIFFITH, T.; PFEIFER, J. Final Project, *Applying Diversity Metrics to Improve the Selection of Web Search Term Refinements*. 2008. Citado na página 27.
- GROUP, T. P. G. D. *Documentação PostgreSQL*. 2019. Disponível em: <<https://www.postgresql.org/docs/>>. Citado na página 27.
- HJALTASON, G. R.; SAMET, H. Index-driven similarity search in metric spaces. *ACM Transactions on Database Systems*, v. 28, p. 517580, 2003. Citado na página 25.
- HUANG, K.; AVIYENTE, S. Rotation invariant texture classification with ridgelet transform and fourier transform: Ieee computer society. In: *International Conference on Image Processing*. [S.l.: s.n.], 2006. p. 2141–2144. Citado na página 46.
- IGONIDERIGHA, N. E. *Data Navigation and Visualization: Navigating Coordinated Multiple Views of Data*. Tese (Doutorado) — Napier University, March 2015. Citado na página 17.
- INSELBERG, A. The plane with parallel coordinates. *The Visual Computer*, Springer-Verlag, v. 1, n. 2, p. 69–91, 1985. ISSN 0178-2789. Citado na página 1.

- JAIN, A.; SARDA, P.; R., H. J. Providing diversity in k-nearest neighbor query results. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2004. v. 3056, p. 404–413. Citado 4 vezes nas páginas 26, 30, 47 e 83.
- KING, A. et al. A coordinated view of the temporal evolution of large-scale internet events. *Computing*, v. 96, n. 1, p. 53–65, January 2014. Citado na página 1.
- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks: Advances in neural information processing systems 25. In: *Neural Information Processing Systems Conference*. [S.l.: s.n.], 2012. p. 1097–1105. Citado na página 45.
- LIU, Y. et al. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, v. 40, p. 262–282, 2007. Citado 2 vezes nas páginas 2 e 20.
- LOPES, A. A. et al. Visual text mining using associations rules. *Computers and Graphics*, v. 31, n. 3, p. 316–326, 2007. Citado na página 10.
- MCEWAN, T.; IGONIDERIGHA, N. E.; BENYON, D. Designing coordinated multiple views of information space. In: BCS LEARNING AND DEVELOPMENT. *28th International BSC Human Computer Interaction Conference*. [S.l.], 2014. p. 82–89. Citado 2 vezes nas páginas 1 e 83.
- MULLER, F. *Granularity Based Multiple Coordinated Views to Improve the Information Seeking Process*. Tese (Doutorado) — Universität Konstanz, Mathematisch-Naturwissenschaftliche Sektion, Fachbereich Informatik, December 2005. Citado na página 1.
- NORTH, C.; SHNEIDERMAN, B. *A Taxonomy of Multiple Window Coordinations*. [S.l.], 1997. Citado na página 17.
- NORTH, C.; SHNEIDERMAN, B. *Snap-Together Visualization: Coordinating multiple views to explore information*. [S.l.], 1999. Citado na página 1.
- OLIVEIRA, M. C. F.; LEVKOWITZ, H. From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics*, v. 9, p. 378–394, 2003. Citado na página 1.
- ORACLE. *Documentação Oracle*. 2019. Disponível em: <<https://docs.oracle.com>>. Citado na página 27.
- PAULOVICH, F. V. et al. Least square projection: a fast high precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, v. 14, n. 3, p. 564–575, 2008. Citado na página 5.
- PAULOVICH, F. V.; OLIVEIRA, M. C. F.; MINGHIM, R. The projection explorer: A flexible tool for projection-based multidimensional visualization. In: IEEE COMPUTER SOCIETY. *XX Brazilian Symposium on Computer Graphics and Image Processing*. [S.l.], 2007. p. 27–36. Citado 3 vezes nas páginas 2, 4 e 41.
- POLA, I. R. V. *Explorando conceitos da teoria de espaços métricos em consultas por similaridade sobre dados complexo*. Tese (Doutorado), 2010. Citado 4 vezes nas páginas 21, 22, 23 e 25.

- PRATES, J. M. et al. Coordinating multiple views using an ontology-based semantic mapping. In: IEEE COMPUTER SOCIETY. *17th International Conference on Information Visualisation*. [S.l.], 2013. p. 192–197. Citado 2 vezes nas páginas 15 e 16.
- RAZENTE, H. L. *Adequando consultas por similaridade para reduzir a descontinuidade semântica na recuperação de imagens por conteúdo*. Tese (Doutorado), 2009. Citado 6 vezes nas páginas 2, 20, 21, 23, 24 e 25.
- SABOL, V. et al. Visual knowledge discovery in dynamic enterprise text repositories. In: IEEE COMPUTER SOCIETY. *13th International Conference Information Visualisation*. [S.l.], 2009. p. 361–368. Citado na página 1.
- SADANA, R.; STASKO, J. Designing multiple coordinated visualizations for tablets. In: MA, K.-L.; SANTUCCI, G.; WIJK, J. V. (Ed.). *Eurographics Conference on Visualization*. [S.l.]: The Eurographics Association and John Wiley & Sons Ltd., 2016. v. 35, n. 3. Citado na página 1.
- SAMMON, J. W. A nonlinear mapping for data structure analysis. In: *IEEE Transactions on Computers*. [S.l.: s.n.], 1964. p. 401–409. Citado na página 4.
- SANTOS, L. F. D. *Explorando variedade em consultas por similaridade*. Dissertação (Mestrado), 2012. Citado 10 vezes nas páginas 21, 23, 24, 28, 29, 30, 32, 33, 34 e 84.
- SANTOS, L. F. D. et al. Combine-and-conquer: improving the diversity in similarity search through influence sampling. In: *30th Symposium on Applied Computing*. [S.l.: s.n.], 2015. p. 994–999. Citado 3 vezes nas páginas 26, 27 e 85.
- SANTOS, L. F. D. et al. Parameter-free and domain-independent similarity search with diversity. In: *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*. [S.l.: s.n.], 2013. v. 5. Citado 3 vezes nas páginas 36, 47 e 83.
- SEIFERT, B.; SCHNEIDER, O. *WinDirStat: Windows Directory Statistics*. 2003. Disponível em: <<https://windirstat.net/>>. Citado na página 1.
- SERAPHIM, E. *Operadores binários para consulta de similaridade em banco de dados multimídia*. Tese (Doutorado), 2005. Citado 6 vezes nas páginas 20, 21, 22, 23, 24 e 25.
- SKOPAL, T. et al. Distinct nearest neighbors queries for similarity search in very large multimedia databases. In: *Proceedings of the Eleventh International Workshop on Web Information and Data Management*. [S.l.: s.n.], 2009. p. 11–14. Citado na página 38.
- SMITH, J. *Integrated Spatial and Feature Image Systems: Retrieval, Analysis and Compression*. Tese (Doutorado), 1997. Citado na página 21.
- TAO, D. *The COREL Database for Content based Image Retrieval*. 2018. Disponível em: <<https://sites.google.com/site/dctresearch/Home/content-based-image-retrieval>>. Citado na página 45.
- VIEIRA, M. R. et al. Divdb: A system for diversifying query results. In: *37th International Conference on Very Large Data Bases*. [S.l.: s.n.], 2011. v. 4, p. 1395–1398. Citado 3 vezes nas páginas 26, 27 e 84.

- VIEIRA, M. R. et al. On query result diversification. In: *27th International Conference on Data Engineering (ICDE)*. [S.l.: s.n.], 2011. p. 1163–1174. Citado 9 vezes nas páginas 27, 28, 29, 30, 32, 33, 72, 73 e 84.
- WARD, M.; GRINSTEIN, G.; KEIM, D. *Interactive Data Visualization - Foundations, Techniques, and Applications*. Second edition. [S.l.]: CRC Press Taylor & Francis Group, 2015. Citado na página 1.
- WASSERMAN, L. A. *All of Statistics: A Concise Course in Statistical Inference*. Second edition. [S.l.]: Springer, 2004. Citado na página 7.
- WILSON, D. R.; MARTINEZ, T. R. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, v. 6, p. 1–34, 1997. Citado na página 21.
- YU, C.; LAKSHMANAN, L.; AMER-YAHIA, S. It takes variety to make a world: Diversification in recommender systems. In: *12th International Conference on Extending Database Technology: Advances in Database Technology*. [S.l.: s.n.], 2009. p. 368–378. Citado 4 vezes nas páginas 27, 30, 47 e 83.
- ZEZULA, P. et al. *Similarity Search: The Metric Space Approach*. [S.l.]: Springer Science & Business Media, 2006. Citado 3 vezes nas páginas 23, 24 e 25.
- ZHENG, K. et al. A survey of query result diversification. *Knowledge and Information Systems*, v. 51, p. 1–36, 2017. Citado 3 vezes nas páginas 2, 26 e 37.

TERMO DE REPRODUÇÃO XEROGRÁFICA

Autorizo a reprodução xerográfica do presente Trabalho de Conclusão, na íntegra ou em partes, para fins de pesquisa.

São José do Rio Preto, 30 / 01 / 2019

Thaís Aline B. Corveia
Assinatura do autor