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

Wilson Estécio Marcílio Júnior

Uma abordagem de exploração multinível em visualizações geradas para representar projeções multidimensionais

São José do Rio Preto
2018

Wilson Estécio Marcílio Júnior

Uma abordagem de exploração multinível em visualizações geradas para representar projeções multidimensionais

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

Financiadora: FAPESP – Proc.. 16/11707-6

Orientador(a): Prof. Dr. Danilo Medeiros Eler

São José do Rio Preto
2018

Marcílio Júnior, Wilson Estécio.
Uma abordagem de exploração multinível em visualizações geradas para representar projeções multidimensionais / Wilson Estécio Marcílio Júnior. -- São José do Rio Preto, 2018
83 f. : il., tabs.

Orientador: Danilo Medeiros Eler

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

1. Computação gráfica. 2. Processamento de imagens - Técnicas digitais. 3. Visualização da informação. 4. Mineração de dados (Computação) 5. Projeções multidimensionais. 6. Algoritmos de computador. I. Título.

CDU – 518.72:76

Ficha catalográfica elaborada pela Biblioteca do IBILCE
UNESP - Câmpus de São José do Rio Preto

Wilson Estécio Marcílio Júnior

Uma abordagem de exploração multinível em visualizações geradas para representar projeções multidimensionais

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.

Financiadora: FAPESP – Proc.. 16/11707-6

Comissão Examinadora

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

Prof. Dr. Almir Olivette Artero
UNESP – Câmpus de Presidente Prudente

Prof. Dr. José Fernando Rodrigues Junior
USP – ICMC – São Carlos

Presidente Prudente
7 de dezembro de 2018

Aos meus pais, Raquel e Wilson

AGRADECIMENTOS

Primeiramente, agradeço a Deus por me dar a oportunidade.

Aos meus pais, Wilson e Raquel, pelo apoio durante o desenvolvimento desse trabalho, sem vocês eu não conseguiria. Assim como aos meus irmãos Peterson e Larissa.

À minha namorada, Karla (Mo), pelo apoio, companheirismo.

Ao professor Danilo pela oportunidade, orientação e paciência durante o desenvolvimento desse trabalho. Um amigo que a UNESP me deu desde o primeiro ano de graduação. Muito obrigado por ter me apresentado à Visualização quando eu não sabia em que especialidade seguir.

À UNESP pela minha formação acadêmica. Foram muitos aprendizados durante os últimos sete anos. Agradeço ao professor Fernando Paulovich pela grande ajuda neste trabalho. Agradeço especialmente ao professor Piteri, com o qual realizei as primeiras atividades extracurriculares no período de graduação; ao professor Milton e ao Departamento de Cartografia pela oportunidade de aprender muito com o estágio; e ao professor Rogério pelas caronas até Bauru para cursar as disciplinas do Mestrado. Por fim, agradeço a todos os professores e funcionários do Departamento de Matemática e Computação.

Às minhas professoras do ensino médio Ana Maria, Sinira e Solange. Muito obrigado por demonstrarem vontade de ensinar e por se preocuparem com a educação de seus alunos, mesmo diante de todos os problemas das escolas públicas.

Aos amigos da faculdade. Ao Rafael, Bruno e Leandro pelo companheirismo de laboratório e pelas diversas ajudas nesse trabalho.

À FAPESP pelo apoio financeiro – Processo #2016/11707-6.

A todos que contribuíram de alguma forma para que fosse possível chegar até aqui.

"Ciência da Computação está tão relacionada aos computadores quanto a Astronomia aos telescópios, Biologia aos microscópios, ou Química aos tubos de ensaio. A Ciência não estuda ferramentas. Ela estuda como nós as utilizamos, e o que descobrimos com elas." – Edsger Wybe Dijkstra

RESUMO

As projeções multidimensionais são uma ferramenta importante para análise de conjuntos de dados multidimensionais. No entanto, embora a representação gráfica de projeções multidimensionais tragam benefícios quanto à identificação de grupos e análise da similaridade entre instâncias de um conjunto de dados, tal representação apresenta dificuldades quando o número de instâncias ou a dimensionalidade do conjunto sendo analisado cresce. Neste trabalho, é apresentada uma abordagem de exploração multinível em visualizações geradas para representar projeções multidimensionais, em que o objetivo é oferecer meios para que um conjunto de dados seja explorado com uma carga cognitiva menor do que em representações comuns de projeções multidimensionais. A técnica proposta é baseada na seleção de representativos para fornecer um contexto e guiar o usuário no processo de exploração, além de utilizar diagramas de Voronoi para definição dos grupos. A abordagem pode ser empregada com qualquer técnica de projeção multidimensional, além de poderem ser utilizados os mais variados algoritmos de seleção de representativos. Nos experimentos realizados são apresentados os algoritmos mais indicados para seleção de representativos, bem como o impacto de diferentes técnicas de projeção multidimensional e do espaço de características dos conjuntos analisados. Além disso, são apresentados dois estudos de casos utilizando a técnica de exploração proposta.

Palavras-chave: projeções multidimensionais. abordagem de exploração. representativos.

ABSTRACT

Multidimensional projections are an important tool for analyzing multidimensional datasets. However, although the graphical representation of multidimensional projection brings benefits according to cluster identification and similarity analysis, such representation presents issues when the number of instances or the dimensionality of the dataset increases. In this work, a multilevel exploration approach in visualizations generated to encode multidimensional projections is presented, in which the goal is to provide subsidies for an exploration with lower cognitive load than the common approaches. The proposed technique is based on selecting representative to provide a context to guide the user in the exploration process, besides using Voronoi diagrams to define clusters. In the experiments, the best suited algorithms to select representative are presented, as well as the impact of different multidimensional projection techniques and the feature space of the analyzed dataset. Finally, two case studies are presented to show how the exploration approach works.

Keywords: multidimensional projections. exploration approach. representative.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

SUMÁRIO

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAS

Atualmente há uma grande preocupação com a quantidade de dados gerados por organizações e como retirar *insights* contidas nesses dados. Não obstante, dados pessoais também estão sendo cada vez mais difíceis de serem manipulados, tais como e-mails, histórico de pesquisas e entre outros, visto que não são dados estruturados. Sendo assim, técnicas são desenvolvidas para que o entendimento desses dados possa ser realizado de forma mais fácil. Em especial, a Visualização utiliza representações gráficas e modelos que guiam os usuários no processo de análise com o objetivo principal de fornecer entendimento acerca dos dados (CARD; MACKINLAY; SHNEIDERMAN, 1999). Por exemplo, considerando dados multidimensionais, apesar das diversas técnicas presentes na literatura para representação, como matrizes de gráficos de dispersão (ANDREWS, 1972) e Coordenadas Paralelas (INSELBERG; DIMSDALE, 1990), as técnicas de Projeção Multidimensional (TEJADA; MINGHIM; NONATO, 2003; MAATEN; HINTON, 2008; PAULOVICH et al., 2008; JOIA et al., 2011) são amplamente exploradas pela comunidade científica devido, principalmente, a sua habilidade intrínsica de construir representações que respeitam a similaridade entre as instâncias dos dados (PAULOVICH; SILVA; NONATO, 2010). Nonato e Aupetit (2018) apresentam uma ótima discussão acerca de projeções multidimensionais utilizadas para visualização de dados multidimensionais.

Formalmente, as técnicas de Projeção Multidimensional mapeiam dados de um espaço de alta dimensão ($\mathbb{R}^m$) para dados em um espaço de baixa dimensão ($\mathbb{R}^p$), de modo que esse mapeamento preserve ao máximo as relações de similaridade existentes no espaço multidimensional. As Projeções Multidimensionais são um caso especial de uma classe mais ampla chamada *Multidimensional Scaling* (MDS), que baseiam-se nas informações de distância para projetar os dados no espaço visual (PAULOVICH; SILVA; NONATO, 2010). Se uma projeção for feita com qualidade, os pontos que foram projetados distantes indicarão instâncias não similares, de acordo com a função de distância utilizada. Com isso, diferentemente da maioria das outras técnicas de visualização, as projeções multidimensionais possuem a habilidade em apresentar com facilidade os grupos presentes nos conjuntos de dados, além disso, é possível verificar o relacionamento entre diferentes grupos de instâncias que podem pertencer a diferentes agrupamentos de dados (PAULOVICH et al., 2008). Um exemplo de projeção pode ser verificado na Figura 1. Nesse caso, um conjunto de imagens de 1000 instâncias é projetado no espaço 2D, tais imagens estão divididas em 10 classes, as quais estão codificadas pelas cores.

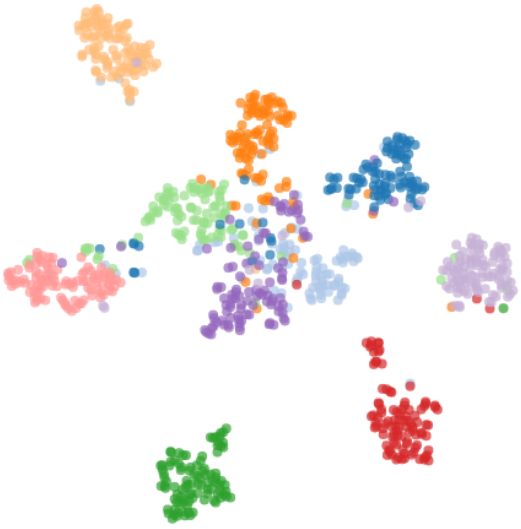
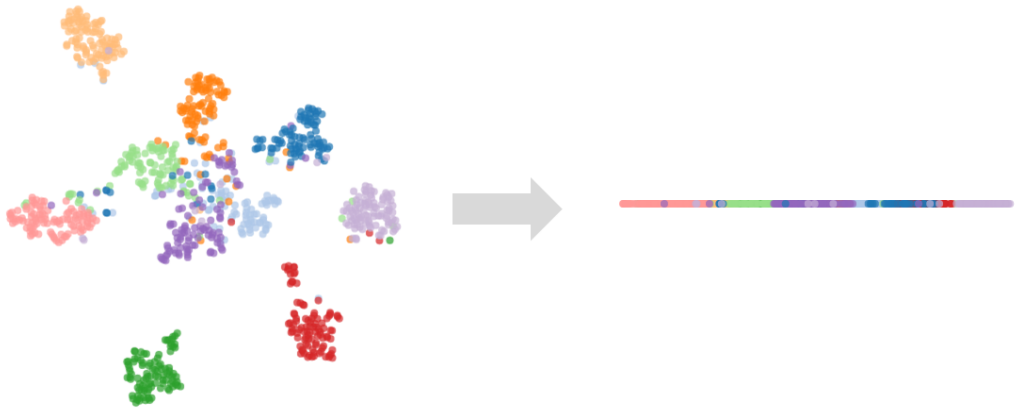


Figura 1 – Projeção do conjunto de dados *Corel* com 1000 instâncias utilizando a técnica de projeção t-SNE (MAATEN; HINTON, 2008).

Embora as metáforas geradas para representar os conjuntos de dados sejam feitas com intuito de auxiliar no processo exploratório e fornecer entendimento do conjunto, conforme o número de instâncias em um conjunto de dados é maior, as técnicas de visualização tendem a apresentar dificuldades na escalabilidade visual, exigindo uma carga cognitiva maior para análise. Consideremos, por exemplo, os métodos de projeção multidimensional. Um problema inerente ao processo de redução de dimensionalidade é a sobreposição dos marcadores utilizados para representar cada instância do conjunto de dados. Tal problema, que também está relacionado à similaridade entre as instâncias dos dados, é intensificado pelo aumento do tamanho e dimensionalidade do conjunto de dados. Na Figura 2 são exemplificados os problemas da redução de dimensionalidade.

O problema de sobreposição (ver Figura 2a) pode ser resolvido por técnicas de remoção de sobreposição, tais como *VPSC* (DWYER; MARRIOTT; STUCKEY, 2006), *PRISM* (GANSNER; HU, 2010) e *ProjSnippet* (GOMEZ-NIETO et al., 2014). No entanto, como as relações de similaridade e vizinhança precisam ser preservadas após a remoção de sobreposição, as melhores técnicas para esse propósito aumentam excessivamente o plano de projeção. Dessa maneira, a remoção de sobreposição deve ser aplicada de forma seletiva. O outro problema está relacionado ao número de instâncias do conjunto de dados, note que na Figura 2b a carga cognitiva para explorar o conjunto é proporcional ao número de instâncias. Embora algumas técnicas que estejam presentes na literatura utilizem o conceito *Overview First & Details-on-Demand* (SHNEIDERMAN, 1996) para diminuir os problemas apresentados pelas metáforas visuais das projeções multidimensionais – como *InfoSky* (ANDREWS et al., 2002), *HiPP* (PAULOVICH; MINGHIM, 2008) e *h-SNE* (PEZZOTTI et al., 2016) – é necessária maior investigação devido às suas limitações.



(a) Problema na redução da dimensionalidade. Como exemplo, o conjunto de dados projetado no espaço 2D foi projetado sob o eixo x.

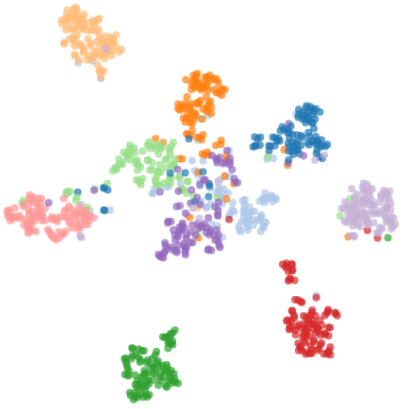
(b) Problema no aumento do número de instâncias.

Figura 2 – Problemas na representação de projeções multidimensionais.

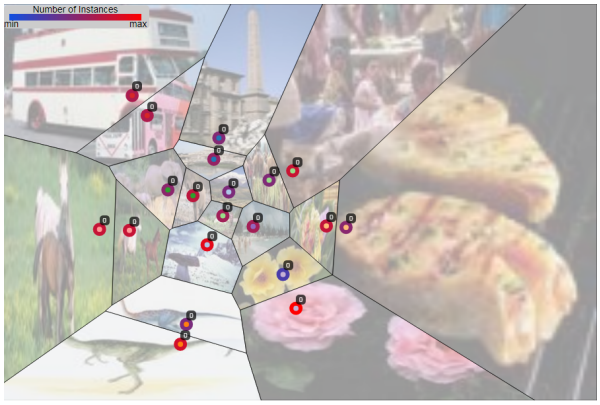
1.2 OBJETIVOS E CONTRIBUIÇÕES

Considerando as questões destacadas, uma abordagem de exploração em visualizações geradas para representar projeções multidimensionais é apresentada, cujo foco está relacionado ao problema de escalabilidade visual em projeções multidimensionais. Sendo assim, a contribuição deste trabalho é uma abordagem de exploração multinível em visualizações geradas para representar projeções multidimensionais. Primeiramente, dada uma projeção no plano, é criada uma árvore para controlar a hierarquia da abordagem com base em seleção de representativos, em que os grupos são definidos como diagramas de Voronoi rígidos. Com a seleção de representativos o usuário consegue obter uma visão geral do conjunto de dados, enquanto que pode obter detalhes por meio da interação. Dessa maneira, a carga cognitiva necessária para analisar o conjunto de dados é diminuída. Enquanto algumas técnicas de exploração multinível foram propostas na literatura – como *InfoSky* (ANDREWS et al., 2002), *HiPP* (PAULOVICH; MINGHIM, 2008) e *h-SNE* (PEZZOTTI et al., 2016) – a abordagem proposta neste trabalho apresenta algumas vantagens em relação à facilidade que pode ser aplicada para diferentes conjuntos de dados; o não aumento da sobrecarga visual ao interagir com a hierarquia; a remoção de sobreposição nos níveis mais baixos da hierarquia; além de permitir que diferentes

algoritmos de remoção de sobreposição, técnicas de seleção de representativos e técnicas de projeção possam ser utilizadas em uma forma *plug & play*. Na Figura 3 são apresentadas a projeção multidimensional por meio da representação tradicional e a representação proposta neste trabalho. Na Figura 3a é apresentada a mesma projeção da Figura 1. Na Figura 3b é apresentado o primeiro nível da abordagem proposta neste trabalho, em que imagens associadas aos representativos (círculos visíveis) são utilizados como textura – maior detalhamento acerca da metáfora visual é apresentado no Capítulo 4.



(a) Projeção do conjunto de dados *Corel* utilizando a forma tradicional de representação.



(b) Projeção do conjunto de dados *Corel* por meio da abordagem hierárquica proposta neste trabalho.

Figura 3 – Comparação da forma de representação proposta neste trabalho com a forma usual de representação.

1.3 ORGANIZAÇÃO DO TEXTO

O restante deste documento está organizado da seguinte maneira. No Capítulo 2 é apresentada uma fundamentação teórica, em que são discutidos os principais conceitos utilizados neste trabalho. No Capítulo 3 são apresentados trabalhos relacionados à esta pesquisa. No Capítulo 4 é apresentada a abordagem desenvolvida. Os resultados são apresentados no Capítulo 5. Por fim, as conclusões e trabalhos futuros são apresentados no Capítulo 6.

2 FUNDAMENTAÇÃO

2.1 PROJEÇÕES MULTIDIMENSIONAIS

As projeções multidimensionais são técnicas utilizadas para reduzir a dimensionalidade de um conjunto. Formalmente, dado um conjunto com m dimensões ($\mathbb{R}^m$), o objetivo de uma técnica de projeção multidimensional é reduzir o número de dimensões do conjunto origem para n dimensões ($\mathbb{R}^n$), de maneira que as relações de similaridade – representadas por meio da distância – no espaço multidimensional sejam preservadas ao máximo no espaço projetado. Sendo assim, dada uma função $\delta(x_i, x_j)$ que mede a similaridade entre duas instâncias x_i e x_j no espaço multidimensional e uma função de distância $d(x_i, x_j)$ que mede a distância dessas duas instâncias no espaço projetado, uma técnica de projeção multidimensional pode ser descrita como uma função f cujo objetivo é minimizar o valor de

$$|\delta(x_i, x_j) - d(f(x_i), f(x_j))| \tag{2.1}$$

tanto quanto possível.

Com o passar dos anos, diversas técnicas de projeção foram apresentadas na literatura, bem como taxonomias para caracterizá-las. Essas taxonomias organizam os métodos de projeção multidimensional de acordo com sua formulação matemática (PAULOVICH; SILVA; NONATO, 2010), habilidade de interação (CHEN et al., 2015), esparsidade do problema (LEE; VERLEYSEN, 2007), e a habilidade em preservar localmente a geometria dos dados (GRACIA et al., 2014). Recentemente, Nonato e Aupetit (2018) apresentaram uma classificação de acordo com várias características, como: Tipo de Dados, Linearidade, Flexibilidade para Supervisão e entre outras.

Assim como a organização dos métodos, também é importante destacar as características da metáfora visual utilizada para representar as projeções multidimensionais. Bertin (1983) mostrou que luminância, tamanho, orientação e posição são as melhores variáveis gráficas para codificar elementos ordinais, em que a posição é a mais efetiva, enquanto matiz e forma são as mais adequadas para codificar elementos categóricos. Gráficos de dispersão de projeções multidimensionais baseiam-se em espacialização para codificar similaridades, mapeando cada item de dados para um ponto no espaço visual de modo que as proximidades relativas de cada par reflitam as similaridades do par correspondente da melhor forma. Desse modo, a similaridade em projeções multidimensionais são codificadas por proximidade espacial, como apresentado na Figura 4.

Figura 4 – Exemplo de uma projeção multidimensional de um conjunto de dados de fotografias, realizada por meio da técnica t-SNE (MAATEN; HINTON, 2008). O conjunto possui 820 instâncias descritas por 4096 características.

Vale ressaltar que as leis de proximidade e similaridade de Ware (2012) são dois processos pré-atentivos de detecção do sistema visual humano. A *lei de proximidade* diz que os grupos de pontos espacialmente próximos uns dos outros são percebidos pré-atentivamente como compartilhando um conjunto de características. A *lei de similaridade* diz que itens cujos marcadores possuem a mesma aparência (cor ou forma) também são percebidos como compartilhando similaridade abstrata de forma pré-atentiva. Considerando a lei de proximidade, agrupamentos visuais de pontos nos mapas de dispersão de projeções multidimensionais seriam percebidos instantaneamente como grupos de instâncias similares no espaço de dados. Dessa maneira, é importante garantir que a maioria das proximidades espaciais correspondam com as similaridades dos itens para que seja possível se beneficiar da lei pré-atentiva de percepção e que seja possível incrementar a efetividade do resultado de um projeção multidimensional.

2.2 SELEÇÃO DE REPRESENTATIVOS

Quando há um aumento no número de instâncias de um conjunto, uma estratégia utilizada é a sumarização de dados, isto é, extrair um subconjunto de instâncias representativas. O subconjunto representativo é considerado mais interpretável que todo o conjunto de dados, o custo para armazenar informações dos dados pode ser significativamente reduzido e, além disso, a eficiência de métodos aplicados aos representativos é maior do que se aplicada em todo conjunto. Um conjunto representativo possui as seguintes características (PAN et al., 2005), embora durante o processo de seleção de representativos

informações acerca do conjunto de dados sejam perdidas:

- é significativamente menor que o conjunto original;
- captura a maior parte das características do conjunto original;
- possui baixa redundância, isto é, as instâncias selecionadas devem ser suficientemente dissimilares.

Na literatura, são apresentadas quatro categorias de técnicas para seleção de representativos (WANG et al., 2017), apesar de poderem se sobrepor: métodos supervisionados, métodos não-supervisionados, métodos de representação de posto pequeno e métodos de seleção de dicionário. Os métodos não supervisionados são aqueles que utilizam somente as coordenadas das instâncias, como os métodos de agrupamento (KAUFMAN; ROUSSEUW, 2005; FREY; DUECK, 2007). Os métodos supervisionados requerem um conhecimento extra sobre o conjunto, como classes, e são normalmente utilizados em tarefas de classificação. A terceira estratégia, métodos de representação de posto pequeno, tentam encontrar uma submatriz que possa se aproximar da matriz original dos dados. Por fim, os métodos de seleção de dicionário buscam representar o conjunto de dados por uma combinação linear utilizando instâncias representativas (CONG; YUAN; LUO, 2012). Neste trabalho, não foi considerada a classe de métodos supervisionados e, além disso, também foi explorado uma classe de métodos de seleção de representativos utilizada em bancos de dados complexos, chamada de métodos de acesso métrico.

2.2.1 MÉTODOS NÃO SUPERVISIONADOS

Os métodos não supervisionados são aqueles que não utilizam informações de classes das instâncias para realizar a extração de representativos. Fazem parte dessa classe de métodos, os já bem difundidos algoritmos de agrupamento (MACQUEEN, 1967; KAUFMAN; ROUSSEUW, 2005; ESTER et al., 1996) e algoritmos cujo objetivo é justamente a seleção de representativos. Por exemplo, o método de Kennard e Stone (1969) e o método *OptiSim* (CLARK, 1997) buscam encontrar representativos que respeitem uma distância média entre si. O método *CAWP* (MEI; CHEN, 2011), por sua vez, utiliza a similaridade de cada par de instâncias do conjunto de dados para definir agrupamentos por meio de objetos com diferentes pesos, em que quanto maior for o peso para um dado objeto, mais representativo esse objeto será para o agrupamento. Diferentemente, os métodos δ-medoid (LIEBMAN; CHOR; STONE, 2015) e *LampRep* (ZHANG; WEI; CHEN, 2013) utilizam um parâmetro definido pelo usuário para guiar o processo de seleção de representativos, enquanto o método δ-medoid forma agrupamentos em que os representativos tenham uma distância máxima com valor do limiar, o método *LampRep* une instâncias no mesmo agrupamento somente se possuem similaridade menor que o

parâmetro especificado. Outra estratégia é utilizar algoritmos de agrupamento, em que os representativos serão os *medoids* dos agrupamentos retornados.

2.2.2 REPRESENTAÇÃO DE POSTO PEQUENO

Técnicas de representação de posto (*rank*) pequeno são baseadas em técnicas de revelação de *rank* QR (PAN; TANG, 1999; PAN, 2000; BOUTSIDIS; MAHONEY; DRINEAS, 2009), que selecionam representativos encontrando uma permutação da matriz de dados que ofereça uma submatriz melhor condicionada. O posto de uma matriz está definido como o número de colunas ou linhas linearmente independentes. A decomposição em valores singulares pode ser utilizada para alcançar estes objetivos. A decomposição em valores singulares (*Singular Value Decomposition* - SVD) de uma matriz A é a decomposição $A = U\Sigma V^T$, onde $U = [u^1 u^2 ... u^N]$ e $V = [v^1 v^2 ... v^N]$ são matrizes ortogonais com dimensões $N \times N$ e $M \times M$, respectivamente e Σ é a matriz retangular diagonal de dimensão $N \times M$ com valores não nulos correspondentes aos valores de A. As colunas de U correspondem aos componentes principais de A, ou seja, o conjunto $\{u^i\}$ abrange o subespaço linear onde os dados originais, colunas de A, são melhor representados. Em outras palavras, a primeira coluna de U é sempre um vetor que aponta para a maior variação dos pontos, deste modo, a decomposição SVD nos permite obter uma representação $1D$ dos dados. Ainda que não seja uma boa representação dependendo da dimensionalidade do conjunto, será a melhor representação da matriz de dados em $1D$. Note que as colunas restantes representam os vetores que apontam para a maior variância dos pontos restantes, aqueles que não foram representados pela primeira coluna de U, e assim sucessivamente. Essa ideia é apresentada na Figura 5, note as duas primeiras colunas de U destacadas pelos vetores em vermelho.

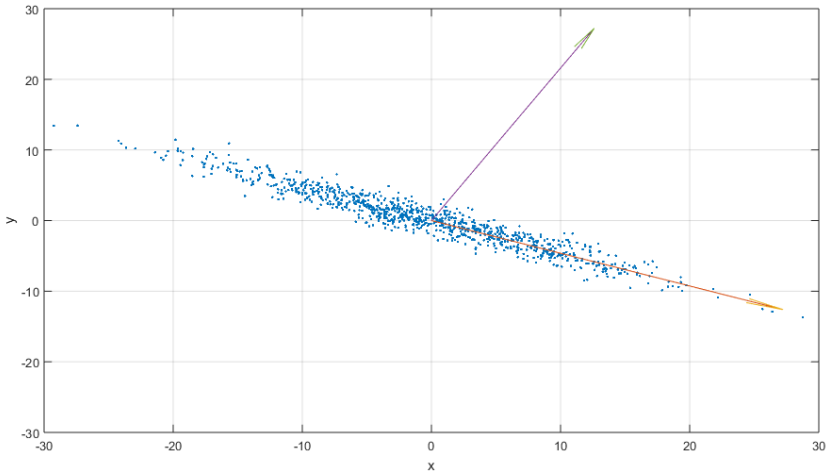


Figura 5 – As duas primeiras colunas de U representam os duas maiores variâncias para onde o conjunto utilizado aponta.

Aharon, Elad e Bruckstein (2006) apresentaram o algoritmo k-SVD que realiza

a tarefa de buscar um dicionário D que conduz a uma representação de um conjunto de instâncias $y_{i}{}_{i=1}^{N}$. O algoritmo k-SVD é um método iterativo que alterna entre codificação esparsa das instâncias baseado no dicionário corrente e um processo de atualização das instâncias do dicionário de modo que se ajustem às instâncias da melhor forma. A representação esparsa de um conjunto de dados pode ser vista como uma generalização do algoritmo k-means, onde os vetores de coeficientes podem ter mais de uma entrada com valores diferentes de zero. Joia, Petronetto e Nonato (2015) apresentaram uma metodologia para exploração de projeções multidimensionais que utiliza a extração de representativos para ajudar no processo de agrupamento. A extração de representativos é formulada por meio de um problema de decomposição em valores singulares. Seja uma matriz A e sua decomposição $A = U\Sigma V^T$, sabe-se que $U = [u^1 u^2 ... u^N]$ correspondem aos melhores vetores de base para representar todos os dados. Sendo assim, Joia, Petronetto e Nonato (2015) supõem que se um subconjunto de colunas de A se aproxima das bases $\{u^i\}$, então este subconjunto também será um bom representativo para o conjunto das colunas de U, escolhendo como representativos o subconjunto de colunas de A com alto índice de correlação. Como a técnica CSM baseia-se na correlação da matriz de dados com as colunas da matriz U para encontrar os representativos, classes com maior variação são beneficiadas na escolha de representativos (JOIA; PETRONETTO; NONATO, 2015).

2.2.3 SELEÇÃO DE DICIONÁRIO

Na seleção de dicionário, um subconjunto de instâncias é selecionado para servir como dicionário para o conjunto de dados por meio de uma combinação linear, utilizando codificação esparsa (WANG et al., 2014). Matematicamente, data uma instância $y \in \mathbb{R}^n$ e um dicionário $D \in \mathbb{R}^{n \times k}$, o problema de representação esparsa pode ser definido como

$$\min_{x} \| x \|_0 \text{ tal que } y = Dx, \tag{2.2}$$

onde $\| x \|_0$ é a norma l_0 do vetor de coeficientes $x \in \mathbb{R}^k$, n é a dimensionalidade do conjunto e k é o número de instâncias selecionadas para atuarem como um dicionário. Desse modo, dado um conjunto de instâncias $\{y_i\}_{i=1,...,N} \in \mathbb{R}^n$, o objetivo é encontrar um dicionário $D \in \mathbb{R}^{n \times k}$ tal que cada instância do conjunto possa ser representada como uma combinação linear esparsa das instâncias do dicionário. O esquema de codificação esparsa é apresentado na Figura 6, note que cada y_i representa o vetor de característica da instância i do conjunto de dados.

Para se obter uma codificação esparsa (AHARON; ELAD; BRUCKSTEIN, 2006; MAIRAL et al., 2008; RAMIREZ; SPRECHMANN; SAPIRO, 2010) é necessário que a matriz de coeficientes X seja esparsa. Visualmente, o resultado pode ser interpretado como apresentado na Figura 7. Note que nesse caso o objetivo é recuperar as instâncias

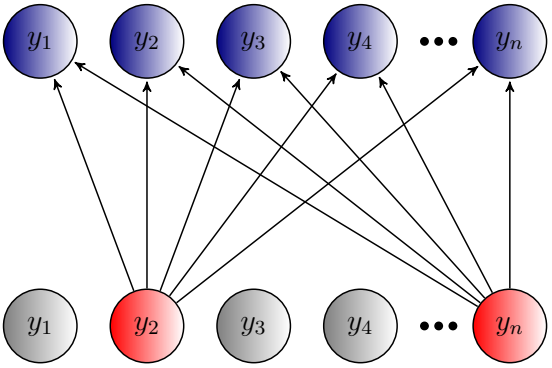


Figura 6 – Codificação esparsa do conjunto de dados Y. Note que apenas os vetores y_2 e y_n são suficientes para representar o conjunto Y.

representadas pela matriz D, que correspondem às instâncias representativas do conjunto de dados Y.

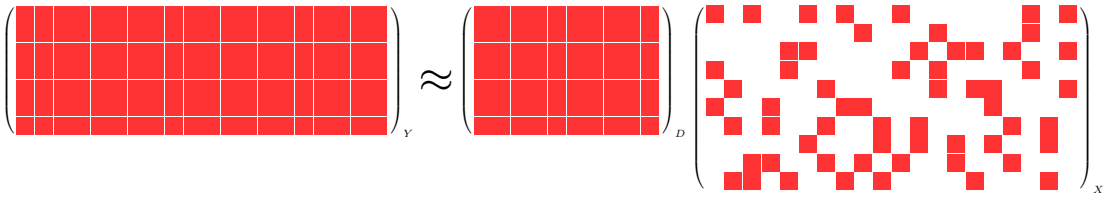


Figura 7 – Representação visual do resultado de uma codificação esparsa.

2.2.4 MÉTODOS DE ACESSO MÉTRICO

Os métodos de acesso métrico são aqueles utilizados em banco de dados complexos, que armazenam dados complexos, como imagens, áudios, mapas, sequências de DNA e entre outros. Como a comparação de dados é realizada por meio da similaridade, a seleção de representativos nesse contexto é importante já que os métodos de acesso podem se beneficiar da posição dos representativos e eliminar cálculos de distância.

Segundo Chávez et al. (2001), existem dois tipos de métodos de seleção de representativos em banco de dados complexos: baseados em agrupamentos e baseados em pivôs. Os métodos baseados em agrupamentos dividem o espaço métrico em um conjunto de regiões equivalentes, cada um representado pelo centro de um agrupamento e, sendo assim, durante as buscas regiões inteiras são descartadas dependendo da distância do centro do agrupamento para a consulta. Os algoritmos baseados em pivôs selecionam um conjunto de objetos como pivôs e um índice é construído por meio do cálculo de distância de cada objeto do conjunto para cada pivô, no qual a desigualdade triangular é utilizada para descartar objetos e eliminar cálculos de distância. Na técnica *Sparse Spatial Selection - SSS* (PEDREIRA; BRISABOA, 2007), por exemplo, um conjunto de pivôs é escolhido de forma que cubra todo conjunto de dados. Seja (X, d) um espaço métrico, $U \subseteq X$ uma coleção de objetos e M a distância máxima entre qualquer par de objetos, para cada

elemento $x_i \in U$, x_i é escolhido para fazer parte do conjunto de pivôs se sua distância para todo ponto no conjunto de pivôs é igual ou maior que $M \times \alpha$, em que $0 \leq \alpha \leq 1$ é um parâmetro que controla a proporção de representativos selecionados. Em outras palavras, um objeto na coleção se torna pivô se está localizado a mais do que uma fração de distância máxima com respeito a todos pivôs já selecionados. A restrição da distância entre os pivôs garante uma boa distribuição no espaço.

2.2.5 MÉTRICAS DE ANÁLISE DE REPRESENTATIVOS

Um dos objetivos com a seleção de representativos é reduzir o conjunto de dados enquanto mantém suas principais características. Além de podermos verificar visualmente, é possível utilizar duas métricas para avaliar a qualidade dos representativos selecionados (MA; WEI; CHEN, 2011): redundância e cobertura.

Seja o conjunto de dados com n instâncias, $Y = \{y_1, y_2, ..., y_n\}$, uma técnica de seleção de representativos irá selecionar k instâncias ($k \ll n$), resultando em Y', em que $Y' \subseteq Y$. Como a redundância e cobertura estão sendo utilizadas para avaliar a qualidade dessas técnicas, um bom método de seleção de representativos possui alta cobertura e baixa redundância. Dados dois conjuntos Y' e Y e uma instância $y \in Y$, Y' consegue representar y com grau $\max_{y' \in Y'}(F_c(y', y))$, onde $F_c(y', y)$ representa a similaridade entre y e y'. Desse modo, o grau de Y' cobre Y pode ser definido como:

$$r_C(Y', Y) = \sum_{y \in Y} \Big(\max(\underset{y' \in Y'}{F_c}(y', y)) \Big) / \mid Y \mid, \tag{2.3}$$

onde $\mid Y \mid$ é o número de instâncias em Y.

Para medir a redundância de um conjunto Y e uma instância y, $y \in Y$, o grau de que y é redundante em Y é $1 - 1/ \sum_{y' \in Y} F_c(y', y)$. Desse modo, a redundância pode ser definida da seguinte maneira:

$$r_R(Y) = \sum_{y \in Y} \Big(1 - 1/ \sum_{y' \in Y} F_c(y', y) \Big) / \mid D \mid . \tag{2.4}$$

Essas métricas são utilizadas para avaliar as técnicas de representativos no Capítulo 5.

2.3 TÉCNICAS PARA REMOÇÃO DE SOBREPOSIÇÃO DE MARCADORES

As técnicas de remoção de sobreposição utilizam diferentes abordagem para remover a sobreposição entre marcadores. Algumas técnicas criam estruturas bem definidas no

plano de projeção, enquanto outras técnicas somente consideram um grafo criado a partir dos marcadores. Durante o processo de remoção de sobreposição, essas técnicas precisam manter tanto quanto possível as relações de similaridade e vizinhança imposta pela projeção multidimensional. Neste trabalho, foram consideradas as técnicas *ProjSnippet* (GOMEZ-NIETO et al., 2014), *RWordle* (STROBELT et al., 2012), *PRISM* (GANSNER; HU, 2010) e *VPSC* (DWYER; MARRIOTT; STUCKEY, 2006)

A técnica *ProjSnippet*, que é utilizada para visualização de resultados de busca na *web*, possui uma abordagem para remover sobreposição de marcadores descrita pela soma de duas componentes, uma que quantifica a sobreposição (E_O), e a outra que quantifica as relações de vizinhança resultantes do processo de redução multidimensional (E_N):

$$E = (1 - \alpha)E_O + \alpha E_N \tag{2.5}$$

onde o parâmetro $\alpha \in [0, 1]$ é utilizado para balancear a contribuição de E_N e E_O.

Na técnica *RWordle*, a sobreposição é removida movendo o marcador em espiral até que não haja sobreposição após realizar uma ordenação ortogonal nos elementos. Dois tipos de ordenação considerando o posicionamento das instâncias podem ser utilizadas:

- *Linear Ordering* (*RWordle-L*): em que itens são ordenados de acordo com uma linha de varredura definida por um ângulo α;
- *Concentric Ordering* (*RWordle-C*): em que itens são ordenados de acordo com a distância para o centro geométrico da projeção.

Na técnica *PRISM*, a estratégia consiste em remover a sobreposição utilizando um grafo de proximidades. Primeiramente, o fator de sobreposição é calculado para cada arestas do grafo de proximidades:

$$t_{ij} = \max(\min(\frac{w_i + w_j}{|x_i^0(1) - x_j^0(1)|}, \frac{h_i + h_j}{|x_i^0(2) - x_j^0(2)|}), 1) \tag{2.6}$$

onde w_i e h_i denotam a metade da largura e altura do nó i, respectivamente, e $x_i^0(1)$ e $x_i^0(2)$ são as coordenadas x e y. Então, cada aresta do (i, j) do grafo é multiplicada por t_{ij}.

Por fim, a técnica *VPSC* tenta remover a sobreposição entre as instâncias por meio de duas etapas: primeiramente, um conjunto de restrições de não sobreposição é gerado; segundo, o objetivo é encontrar uma solução sem sobreposição baseada nesse conjunto.

2.4 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentados os conceitos utilizados para criação da abordagem de exploração proposta, assim como as métricas utilizadas para analisar a efetividade dos algoritmos na seleção de representativos.

3 TRABALHOS RELACIONADOS

Um dos fatores que devem ser considerados na criação de uma abordagem visual é a escalabilidade visual, ou seja, a capacidade de representação visual das ferramentas de visualização, da efetividade em apresentar grandes conjuntos de dados, tanto em termos do número de dimensões quanto em termos do número de instâncias de dados (EICK; KARR, 2002). As abordagens normalmente utilizam o conceito *Overview-First & Details-on-Demand* Shneiderman (1996) para atingir escalabilidade visual em técnicas de visualização tradicionais ou em novas técnicas de visualização, assim, muitas abordagens que consideram o conceito de hierarquia são propostas na literatura. Tradicionalmente, as técnicas hierárquicas de visualização são divididas em dois grupos distintos (WARD, 2002): técnicas de preenchimento do espaço visual, tais como *Treemap* (JOHNSON; SHNEIDERMAN, 1991), *Voronoi Treemap* (BALZER; DEUSSEN, 2005), *Circle Packing* (WANG et al., 2006), *NMAP* (DUARTE et al., 2014) e entre outras; e as técnicas de ligação de nós, por exemplo, *Cone Tree* (ROBERTSON; MACKINLAY; CARD, 1991) e *Hierarchical Clustering Explorer* (SEO; SHNEIDERMAN, 2002).

Neste capítulo, são apresentadas as técnicas que utilizam o conceito de hierarquia, mas possuem similaridade com o trabalho proposto.

3.1 NMAP

A técnica *Neighborhood Map-NMAP* (DUARTE et al., 2014) é uma abordagem para criar *Treemaps* retangulares. Nessa técnica, o objetivo é a criação de uma *Treemap* que preserve as relações de similaridade enquanto divide consecutivamente a área e reescala as divisões resultantes. Seja $D = \{d_1, ..., d_n\}$ um conjunto de instâncias, P : $D \rightarrow P$ uma função que atribui pesos $P \in \mathbb{R}$ a cada instância, R um retângulo que contém a área para apresentar a *Treemap* e $H = \{(x_1, y_1), ..., (x_n, y_n)\} \in \mathbb{R}^2$ as coordenadas cartesianas atribuídas aos pontos dentro de R representando as instâncias, a estratégia consiste em aplicar um processo de divisão e escala até que um retângulo seja definido por cada elemento em D.

Os retângulos podem ser verticalmente ou horizontalmente divididos. Na divisão horizontal, um segmento vertical bv é definido dividindo R em dois retângulos R_A e R_B de modo que $R_A \cup R_B = R$ e $R_A \cap R_B = \emptyset$. Pesos p_A e p_B são associados a R_A e R_B e calculados como a soma dos pesos dos elementos em cada retângulo. Sendo assim, R_A e R_B são horizontalmente reescalados de modo a apresentar áreas proporcionais a p_A e p_B, respectivamente. Na Figura 8 é apresentado o processo de partição da técnica *NMAP*. Após o primeiro posicionamento, o espaço é dividido e reescalado conforme os pesos.

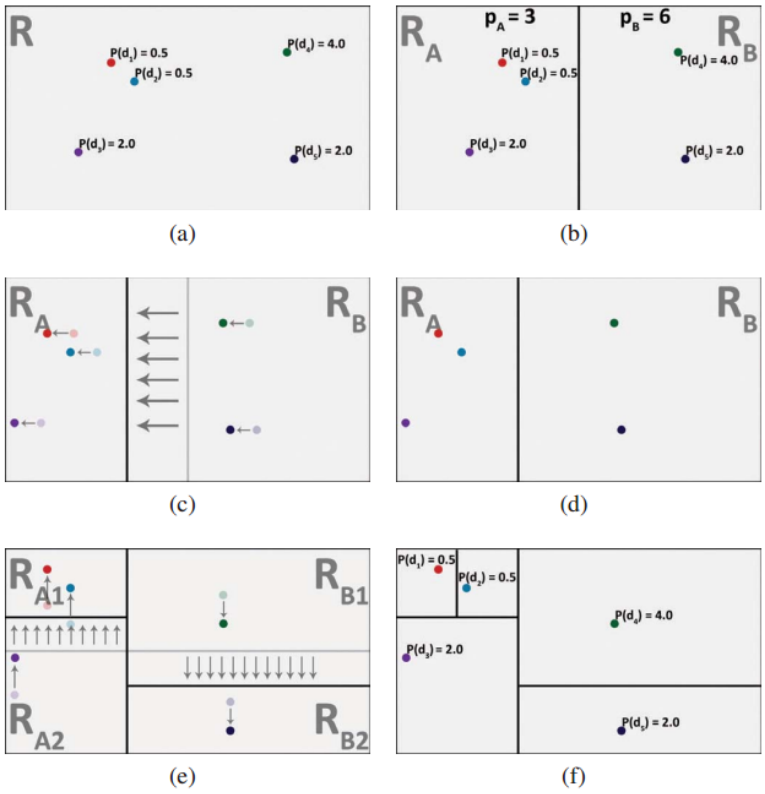


Figura 8 – Processo de particionamento da área coberta pela técnica *NMAP* (DUARTE et al., 2014).

Como o objetivo da técnica *NMAP* é preservar as relações de similaridade entre os objetos do conjunto de dados, o posicionamento inicial das instâncias é realizado por uma técnica de projeção multidimensional.

3.2 HIPP

Também baseada em técnicas de projeção multidimensional, a técnica *HiPP* (PAULOVICH; MINGHIM, 2008) define uma hierarquia em conjuntos de dados para que a exploração de documentos seja feita de acordo com a demanda de informações. Para isso, uma árvore de agrupamento hierárquica é definida, em que os elementos visuais são representados tanto por instâncias individuais, quanto por agrupamentos de instâncias relacionadas. Sua estrutura é construída de forma que o primeiro nível apresente os agrupamentos superiores, e que níveis subsequentes apresentem subgrupos mais detalhados próximos a elementos individuais. O processo exploratório é realizado de passos de refinamento, começando com uma visão geral do conjunto de dados e focando em agrupamentos de interesse, permitindo a análise dos conteúdos das instâncias de dados individuais.

O posicionamento é realizado no plano, onde grupos e subgrupos da instâncias são representados usando círculos dentro de círculos, e as posições de tais círculos reflete

similaridade entre os (sub-)grupos que eles representam. Um exemplo de exploração de documentos utilizando a técnica *HiPP* pode ser visualizado na Figura 9.

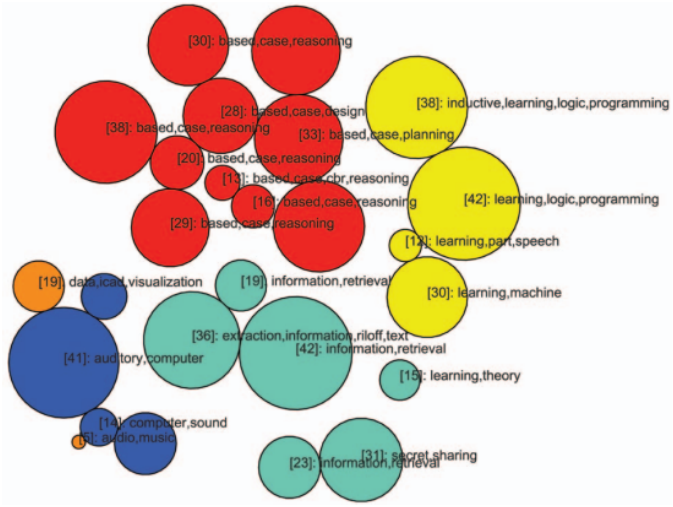


Figura 9 – Exemplo de exploração de um conjunto de documentos utilizando a técnica *HiPP* (PAULOVICH; MINGHIM, 2008).

A árvore de agrupamentos hierárquica é construída utilizando um processo de partição recursiva, em que nós internos são agrupamentos e as folhas são instâncias individuais. Esse processo inicia com um nó, $ROOT$, contendo todas as instâncias. Então, essas instâncias são divididas em $k = \sqrt{|ROOT|}$ nós, criando os nós filhos $C_1, C_2, ..., C_k$. Cada filho C_i é dividido em $\sqrt{|C_i|}$ nós, e nós resultantes são ligados aos filhos. O particionamento é recursivamente aplicado para cada novo nó até que um número especificado de instâncias seja alcançado, processo no qual os pontos são convertidos em folhas.

3.3 VISUAL SUPER TREE

Um trabalho similar ao deste projeto foi apresentado por Silva (2016), em que um método de exploração é aplicado em visualizações de árvores de similaridade. Para isso, um pré agrupamento é realizado no conjunto de dados de maneira multinível, criando um conjunto de árvores de similaridades de tamanho decrescente, cada uma representando um nível de partição do conjunto de dados. Essas árvores são agrupadas para resultar na criação de uma super árvore global, abrangendo todo conjunto de dados. A árvore resultante do agrupamento é visualizada por meio de um processo exploratório hierárquico, utilizando a metáfora *overview + detail*. Um exemplo pode ser verificado na Figura 10.

Como a *Visual Super Tree* requer que a árvore de similaridade reflita as relações de similaridade entre as observações do conjunto de dados, os super nós devem representar instâncias altamente relacionadas. Esses supernós, por sua vez, são utilizados para controlar a exploração dos dados, principalmente no sentido de reduzir a desordem visual. A partir

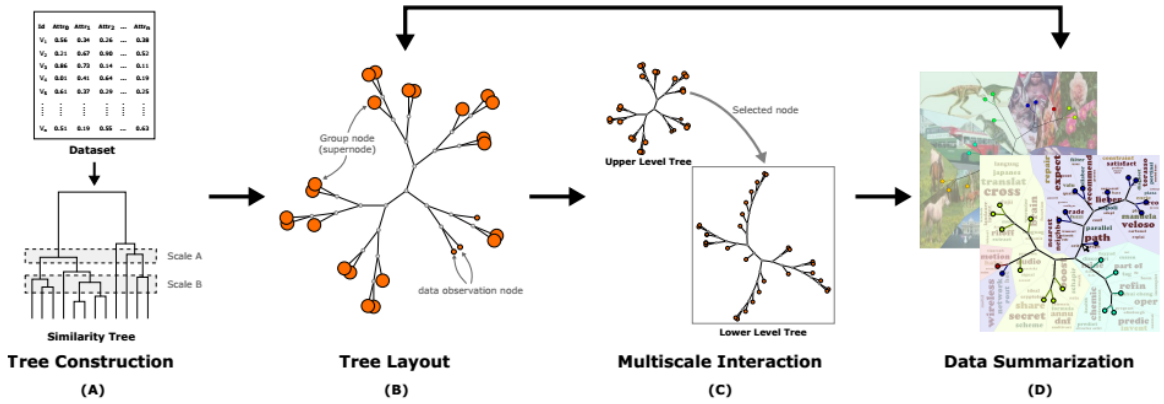
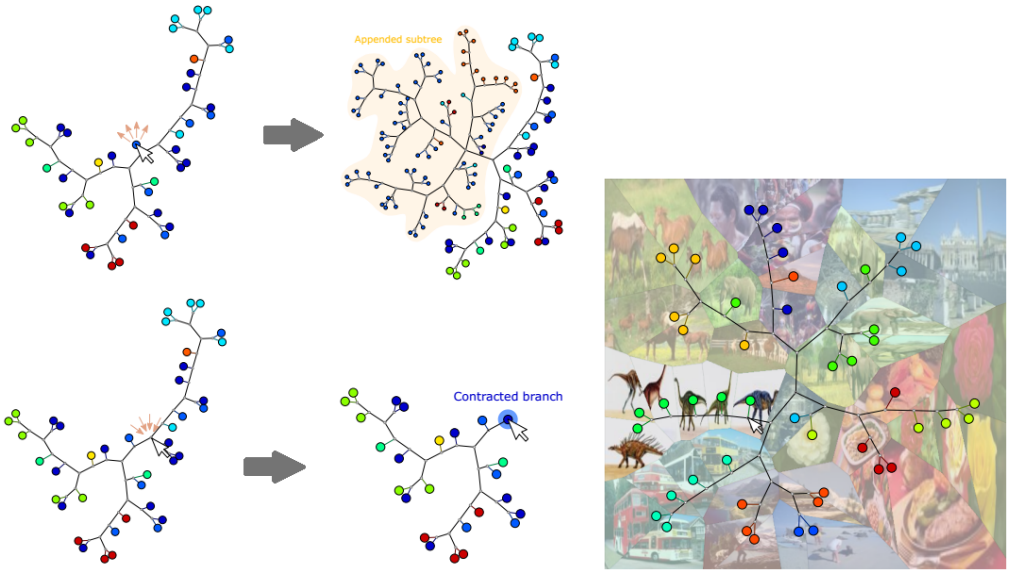


Figura 10 – Exemplo do uso da técnica *Visual Super Tree* (SILVA, 2016).

da criação da *Visual Super Tree*, são apresentados os conteúdos das instâncias projetadas, como apresentado na Figura 11b. Note que a exploração na *VST* é realizada por meio da expansão de nós e da contração de nós para formação de supernós, como apresentado na Figura 11a.



(a) Agregação e expansão de nós.

(b) Sumarização de dados.

Figura 11 – Exploração de dados utilizando a técnica *Visual Super Tree* (SILVA, 2016).

3.4 INFOSKY

A técnica *InfoSky* (ANDREWS et al., 2002) oferece uma maneira para exploração de coleções de documentos estruturados hierarquicamente. De forma similar a um telescópio real, a técnica *InfoSky* emprega uma representação gráfica planar com ampliação variável, em que documentos de conteúdo similar são posicionados próximos um dos outros, formando constelações de documentos.

Na visualização, documentos são mapeados como estrelas e documentos similares são agrupados como polígonos para representar fronteiras de constelações. Além disso, áreas menos densas ou vazias, isto é, onde não existe nenhum documento representam restrições de acesso, da mesma forma que nebulosas escuras aparecem entre galáxias. Na Figura 12 é apresentado um exemplo da técnica *InfoSky*.

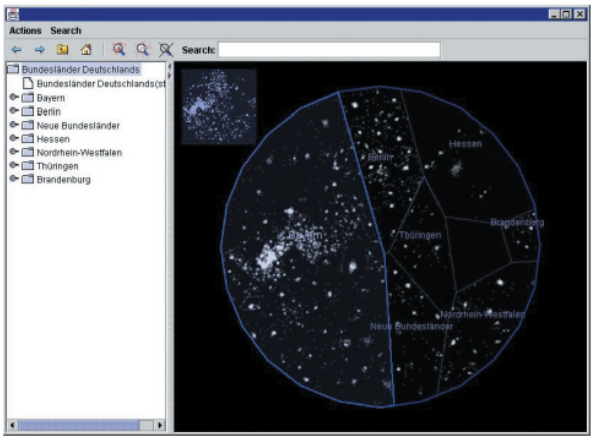


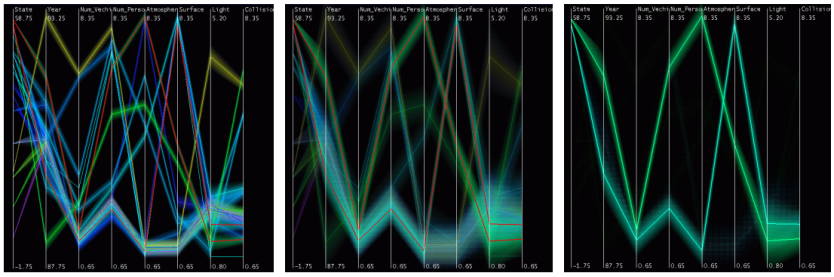
Figura 12 – Galáxia derivada de uma coleção de aproximadamente 109000 artigos de notícias alemãs. Os artigos foram classificados tematicamente em 6900 coleções e subcoleções, com 15 níveis de hierarquia (ANDREWS et al., 2002).

As fronteiras de constelações e marcações são apresentadas para o nível mais alto da hierarquia, em que todos documentos, até o nível mais baixo da hierarquia são apresentados ao usuário. Note que, como as instâncias de níveis mais baixos são abstraídas, é necessário um processo de ampliação para investigá-las.

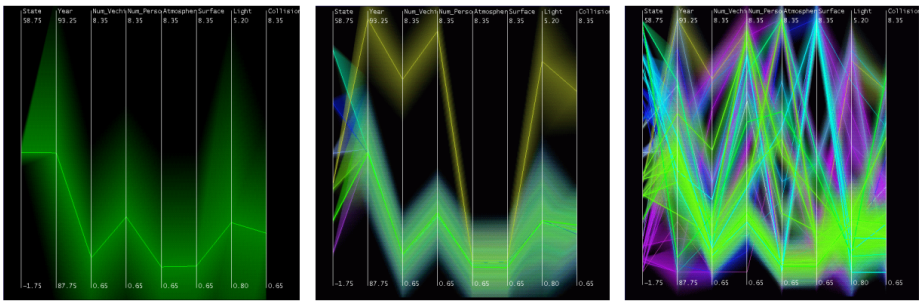
3.5 HIERARCHICAL PARALLEL COORDINATES

A técnica apresentada por Fua, Ward e Rundensteiner (1999) utiliza agrupamentos dos dados para realizar agregação e fornecer um sumário do conjunto de dados ao usuário, utilizando Coordenadas Paralelas (INSELBERG; DIMSDALE, 1990). Um exemplo é apresentado na Figura 13a, em que a média das instâncias dos agrupamentos são apresentados de forma mais opaca. Para que as instâncias possam ser apresentadas de acordo com um nível de detalhe, a técnica também realiza um agrupamento hierárquico nos dados, permitindo a variação do nível de detalhe. Um exemplo da utilização do nível de detalhe pode ser verificado na Figura 13b.

Também considerando a ideia de agrupamentos hierárquicos, Poco et al. (2011) apresentaram um *framework* para projeção de dados de alta dimensão em espaços visuais 3D, baseando na generalização da técnica *Least Square Projection* (LSP) (PAULOVICH et al., 2008). Para facilitar a exploração dos dados, uma abordagem hierárquica de agrupamento é utilizada. Primeiramente, um agrupamento é apresentado ao usuário para



(a) Agregação das polilinhas em Coordenadas Paralelas (FUA; WARD; RUNDENSTEINER, 1999).



(b) Diferentes níveis de detalhes em Coordenadas Paralelas (FUA; WARD; RUNDENSTEINER, 1999).

Figura 13 – *Hierarquical Parallel Coordinates* (FUA; WARD; RUNDENSTEINER, 1999).

que selecione os agrupamentos que deseja investigar. Os agrupamentos são divididos em outros grupos de modo que partes da projeção possam ser investigadas de acordo com a demanda de informações.

Na Figura 14 são apresentadas algumas etapas de exploração.

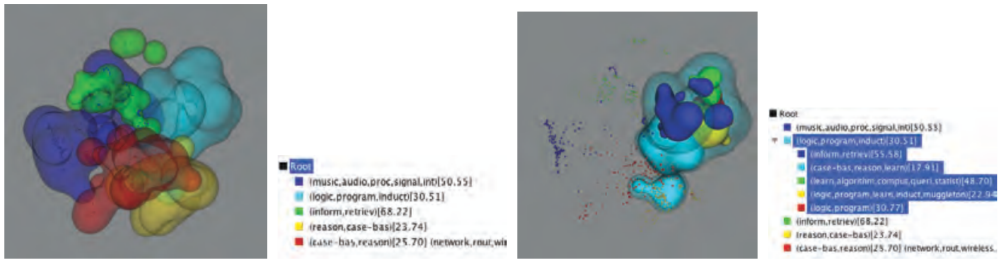


Figura 14 – Exploração de acordo com a demanda (POCO et al., 2011).

3.6 HIERARCHICAL NETWORK MAP

Mansmann e Vinnik (2006) apresentaram uma abordagem visual para detecção e avaliação de anomalias em redes, chamada de *Hierarchical Network Map*. A abordagem tem o objetivo de ajudar na análise convencional do comportamento do fluxo de redes, em que são empregados métodos estatísticos. Para visualização dos *hosts* de origem e destino dos fluxos de mensagem, uma abordagem hierárquica é empregada, fornecendo ao usuário

a possibilidade e alterar o nível de agregação apresentado ou aplicar filtros. Na Figura 15 é apresentado o esquema da visualização.

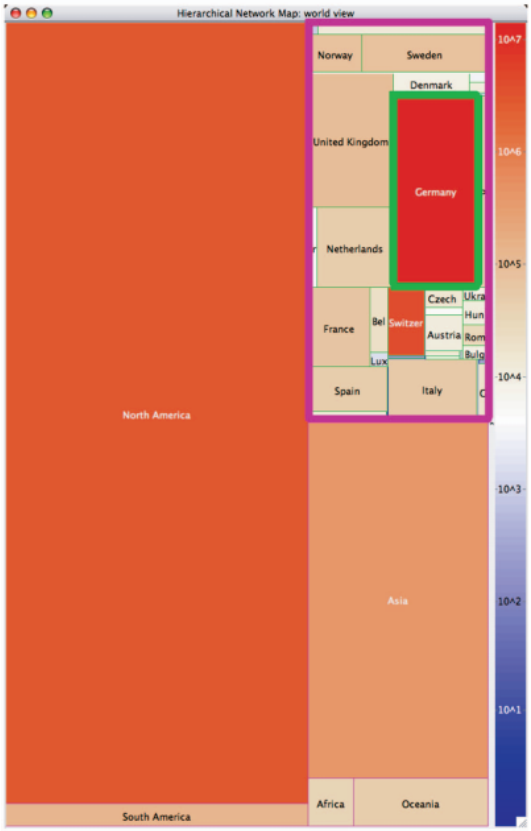


Figura 15 – Exemplo da aplicação da técnica *Hierarchical Network Map* (MANSMANN; VINNIK, 2006).

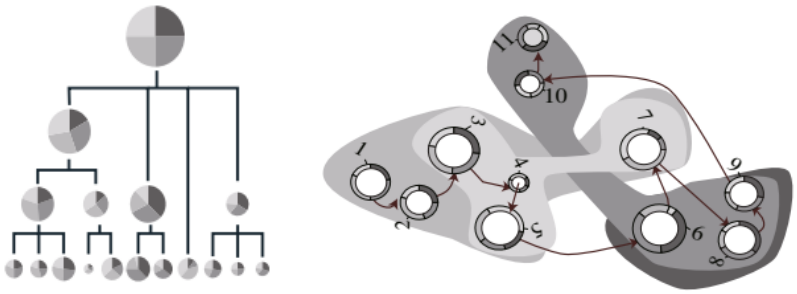
A estrutura da rede pode ser vista como uma hierarquia com endereços *IP* de *hosts* individuais como nós no nível das folhas. No topo dos níveis de rede mencionados, a hierarquia de endereços *IP* é estendida por duas classes geográficas, isto é, continentes e países. A ideia é recursivamente aninhar retângulos filhos em seu retângulo pai, desde o nível de continente até redes locais ou *hosts*. A área de cada retângulo e seu particionamento são mapeados de acordo com o total de redes e seus componentes.

3.7 TOPIC HYPERGRAPH

Outra técnica proposta para visualização de documentos foi apresentada por Wang et al. (2013), chamada de *Topic Hypergraph*. A técnica proposta caracteriza a estrutura temática de um documento longo por meio de uma representação utilizando hipergrafos, em que cada nó representa uma parte do documento e codifica seu tema por meio da composição de múltiplos tópicos. Existem dois tipos de relacionamentos entre os nós: um

nó que conecta dois temas consecutivos para representar transições e uma hiperaresta que codifica os tópicos.

A evolução temática de um longo documento pode ser sumarizada adaptativamente. Na granularidade mais grossa, todo documento pode ser sumarizado em um tema, enquanto que na granularidade mais fina, múltiplos temas caracterizam como um documento se relaciona. Na Figura 16a é apresentado um exemplo de hierarquia de temas, enquanto que na Figura 16b é apresentada a visualização do documento considerando a hierarquia.



(a) Hierarquia de tópicos. (b) Hipergrafo no segundo nível de detalhe.

Figura 16 – Hierarquia de tópicos e efeito na visualização da sequências de temas.

3.8 HIERARCHICAL-SNE

Pezzotti et al. (2016) apresentaram uma abordagem para diminuir as dificuldades importas pelos métodos de projeção, além de oferecer mecanismos para que um conjunto de dados seja explorado de forma hierárquica, seguindo o conceito *Overview-First & Details-On-Demand* (SHNEIDERMAN, 1996).

A técnica apresentada, chamada de Hierarchical-SNE (h-SNE), é derivada da técnica t-SNE e tem como principal melhoria o tempo de processamento exigido para projeção de conjuntos de dados. Para isso, a técnica h-SNE realiza a projeção em duas etapas, em que primeiramente são projetadas estruturas dominantes e, então, o restante do conjunto de dados. Essa estratégia de projeção permite que uma abordagem de exploração seja criada, ou seja, enquanto que as estruturas dominantes podem ser utilizadas para apresentar uma visão geral das projeções, o usuário pode refinar a busca e procurar por instâncias de dados individualmente. Um exemplo de projeção utilizando a técnica h-SNE pode ser visualizado na Figura 17.

3.9 CONSIDERAÇÕES SOBRE O CAPÍTULO

Neste capítulo, foram apresentadas diferentes abordagens empregadas para exploração hierárquica em diferentes conjuntos de dados. Apesar dos diferentes focos das técnicas

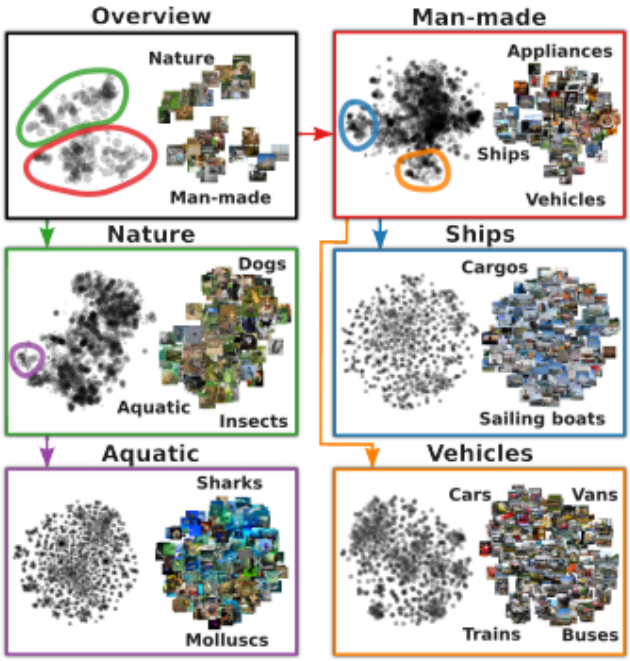


Figura 17 – Projeção de um conjunto de dados por meio da técnica h-SNE (PEZZOTTI et al., 2016).

apresentadas, todas consideram como estratégia o tão difundido mantra de Shneiderman (1996) "*Overview-First & Details-on-Demand*", em que estruturas são primeiramente apresentadas para fornecer uma visão geral do conjunto de dados analisado e guiar no processo exploratório. Conforme o usuário interage com as visualizações, as informações são apresentadas ou refinadas, fornecendo maior detalhe sobre instâncias individuais, processo correspondente à etapa "*Details-on-Demand*".

No contexto de técnicas de posicionamento de pontos, principalmente considerando técnicas de projeção multidimensional, a sobreposição entre os marcadores não é levada em consideração. Por exemplo, apesar da técnica *h*-SNE apresentar uma maneira de explorar a projeção criada, a desordem visual criada pela sobreposição ainda é mantida. Dessa forma, é necessário uma combinação de fatores, considerando tanto a exploração de acordo com a demanda de informações quanto a redução da desordem em projeções multidimensionais.

4 ABORDAGEM DE EXPLORAÇÃO MULTINÍVEL

4.1 CONSIDERAÇÕES INICIAIS

Considerando os problemas apresentados nos capítulos anteriores, foi desenvolvida uma abordagem de exploração multinível com o propósito de diminuir as seguintes dificuldades apresentadas pela forma de representação de projeções de gráficos de dispersão:

- dificuldade na exploração do conjunto de dados devido a quantidade de elementos;
- sobreposição causada pelo processo de redução de dimensionalidade.

Para criação da abordagem de exploração multinível, é necessária a redução da dimensionalidade do conjunto de entrada para o plano 2D. Dessa maneira, dada uma nuvem de pontos no plano, ocorre a primeira seleção de representativos e a definição do primeiro nível da hierarquia. Com base nos representativos, são definidos novos grupos, cuja definição é realizada por meio do domínio de Voronoi. Assim, uma nova seleção de representativos é efetuada para cada grupo, em que subgrupos também são formados. Esse processo continua até que cada grupo tenha no mínimo uma quantidade M de instâncias pré-especificada. Para visualização, diagramas de Voronoi rígidos são utilizados para codificar as fronteiras dos grupos. Neste trabalho, é apresentado uma aplicação utilizando imagens para a abordagem de exploração hierarquica e, sendo assim, imagens são utilizadas como *background* para oferecer uma visão geral do conteúdo do grupo, isto é, a imagem do representativo selecionado. Tal processo é apresentado na Figura 18.

4.2 REDUÇÃO DE DIMENSIONALIDADE E SELEÇÃO DE REPRESENTATIVOS

Dado um mapa de dispersão, o primeiro processo a ser executado para a criação da hierarquia neste trabalho é a seleção de representativos. Quando um conjunto de dados no espaço de alta de dimensão é especificado, é necessário a redução da dimensionalidade por meio de uma técnica de projeção multidimensional. Na Figura 19 é ilustrado o processo de projeção multidimensional e seleção de representativos. O primeiro passo dá-se pela especificação do arquivo de pontos – seja ele de pontos no espaço multidimensional ou não – e, a partir do mapa de dispersão gerado os representativos são selecionados, codificados na imagem pelas instâncias vermelhas.

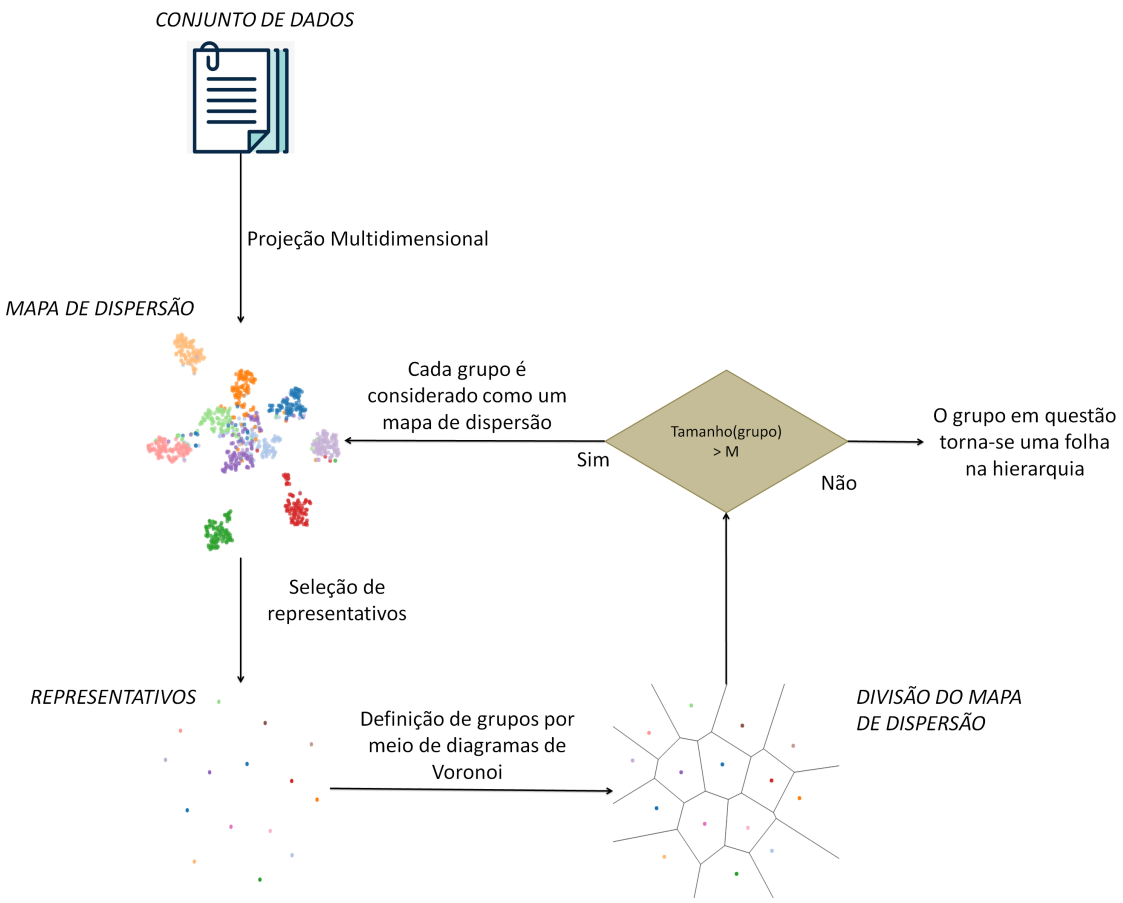


Figura 18 – Esquema geral da técnica de exploração.

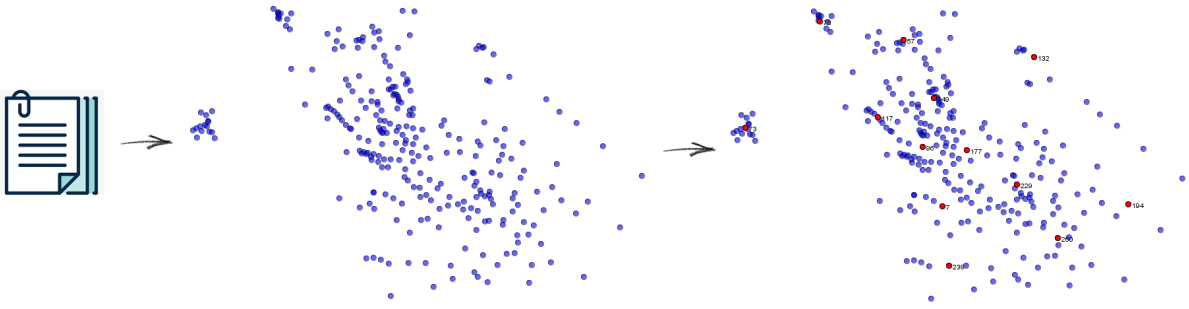


Figura 19 – Esquema geral da técnica de exploração. Da esquerda para direita, temos: o espaço de características de um conjunto de dados, a projeção desse conjunto e, por fim, os representativos desse conjunto destacados em vermelho.

Vale ressaltar que a projeção multidimensional pode ser feita com qualquer técnica, visto que a abordagem desenvolvida exige somente uma nuvem de pontos no plano. No entanto, no Capítulo 5 são apresentas algumas características de técnicas de projeção multidimensional que são ideais para utilização da abordagem de exploração proposta.

4.3 DEFINIÇÃO DA HIERARQUIA

Na seção anterior, definimos que primeiramente é realizada uma projeção multidimensional para que os representativos possam ser selecionados. No entanto, é necessário haver a criação de níveis na projeção para que a exploração dos dados possa ser realizada hierarquicamente. Sendo assim, uma nova seleção de representativos deve ser realizada nos grupos resultantes da seleção de representativos do primeiro nível. Tais grupos são formados pelo domínio de Voronoi, isto é, para toda instância x_i em um nível arbitrário da hierarquia, essa instância vai pertencer ao grupo do representativo y_j se e somente se $\forall y_k \in Y$, $d(y_j, x_i) \leq d(y_k, x_i)$. Visualmente temos o efeito como demonstrado na Figura 20.

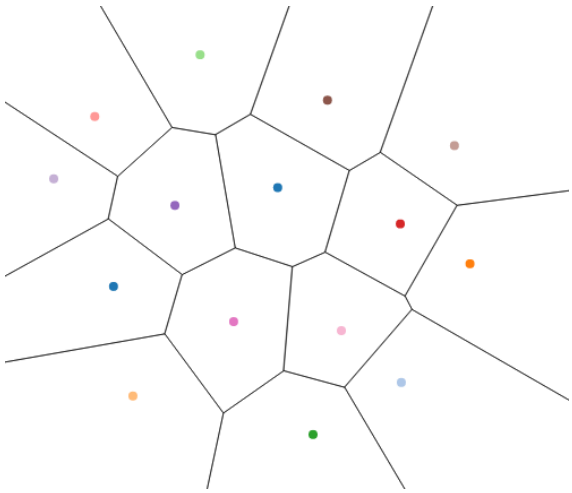


Figura 20 – Divisão do espaço imposta pela seleção de representativos.

Genericamente, para cada grupo $Y_{ji}, 2 \leq i \leq n$ definido pela seleção de representativos no nível j, é realizada uma seleção de representativos até uma condição de parada, isto é, até que $|Y_{ji}|$ seja menor que uma quantidade M pré-especificada. Esse processo cria uma árvore em que se há instâncias suficientes para a seleção de representativos em um dado grupo Y_{ji}, o qual será pai dos grupos formados pela seleção de representativos desse conjunto. O esquema da Figura 21 pode ser utilizado para facilitar o entendimento.

Conforme os representativos são selecionados e os grupos são definidos, é possível que o número de instâncias e alguns grupos sejam menores que M, de forma que precisem ser agrupados com outros grupos. Para isso, o processo reverso do algoritmo de agrupamento hierárquico é aplicado, como apresentado na Figura 22. Note que existem dois casos em

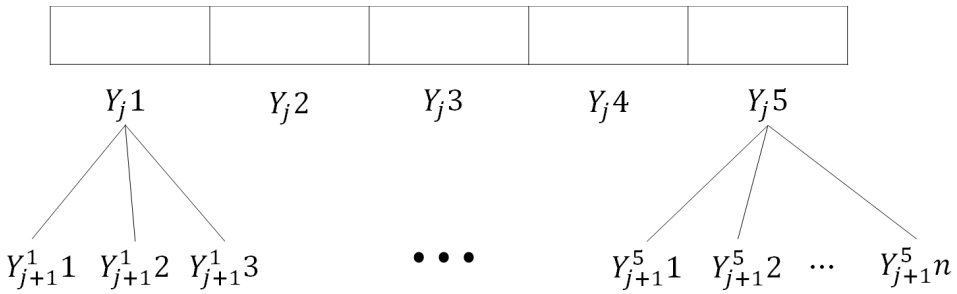
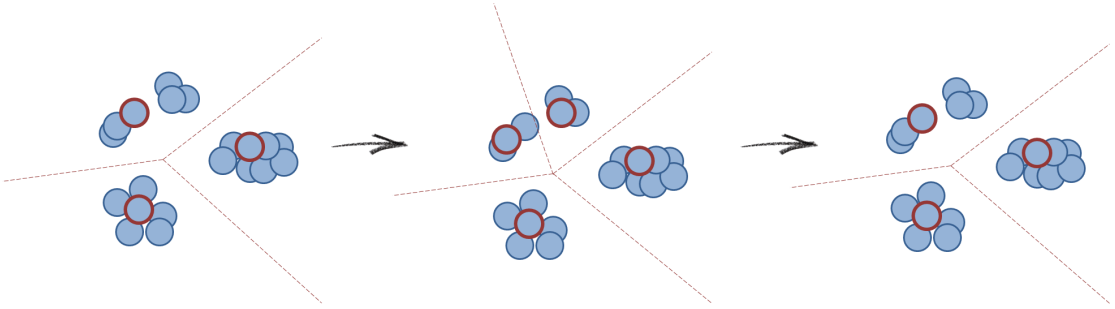
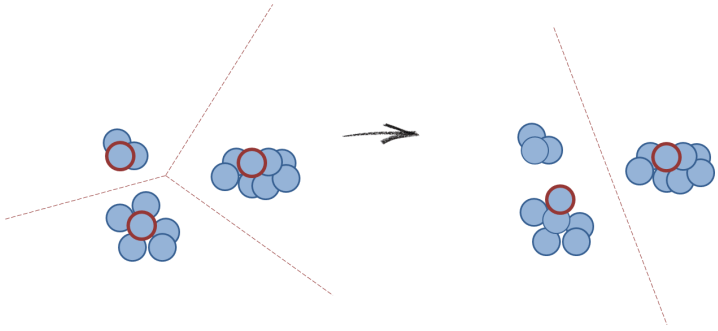


Figura 21 – Esquema da hierarquia criada pela seleção consecutiva de representativos. Note que j define um nível arbitrário da hierarquia, enquanto o índice sobrescrito é utilizado somente para mostrar qual é o pai de dado nó.

que se deve aplicar a união dos grupos. No primeiro (ver Figura 22a), um nó possui elementos suficientes para ser dividido, mas após a divisão não há uma quantidade de instâncias satisfatória. Dessa maneira, os dois nós com quantidade insuficiente são unidos pelo processo de união de nós, isto é, pelo algoritmo hierárquico aplicado de forma reversa. No segundo caso (ver Figura 22b), a partição imposta pela seleção de representativos faz com que um nó tenha um número de instâncias menor que o permitido. Assim, o nó com menor quantidade de elementos é agrupado com o elemento mais próximo, considerando os representativos dos nós.



(a) Caso 1.



(b) Caso 2.

Figura 22 – Processo de união dos nós com quantidade inferior à M instâncias.

Após o processo de definição da hierarquia, conforme a interação com o usuário, é efetuada a remoção de sobreposição. Tal remoção de sobreposição pode ser realizada

pelos algoritmos *RWordle* (STROBELT et al., 2012), *PRISM* (GANSNER; HU, 2010), *ProjSnippet* (GOMEZ-NIETO et al., 2014) ou *VPSC* (DWYER; MARRIOTT; STUCKEY, 2006), aos quais foram descritos brevemente na Seção **??**.

4.4 ABORDAGEM DE EXPLORAÇÃO

Para o desenvolvimento da abordagem de exploração multinível, foi utilizado o *framework* D3.js e, neste trabalho, foi considerada a visualização de conjunto de imagens para a elaboração das técnicas de interação e exploração do conjunto de dados. Primeiramente, o usuário deve especificar um conjunto de pontos no espaço multidimensional para ser projetado ou fornecer um conjunto de pontos no plano, de modo que seja possível a utilização do algoritmo de seleção de representativos e iniciar a construção da hierarquia.

Como o trabalho é focado para visualização de imagens, o usuário também pode fornecer um conjunto de imagens por meio da *url* dessas imagens no momento de fornecer o arquivo de projeção. Além disso, um arquivo contendo metadados do conjunto também pode ser fornecido para auxiliar na exploração de projeção. Na Figura 23 é apresentado um exemplo de projeção com o primeiro nível da hierarquia. Note que as imagens do conjunto são utilizadas como *background* de cada grupo, isto é, as imagens correspondentes aos representativos. Além disso, a cor do círculo concêntrico por fora do marcador que representa as instâncias representativas fornecem uma escala do número de instâncias pertencendo ao grupo, isto é, cores mais próximas ao vermelho codificam grupos com um maior número de instâncias. Em contrapartida, as cores dos círculos internos – dos marcadores que representam as instâncias dos dados – representam as classes das instâncias. Por fim, note que existem alguns números associados a cada marcador, tal número representa a quantidade de instâncias que foram selecionadas durante o processo exploratório. Quando o usuário atinge um nó folha da hierarquia, é possível selecionar imagens para posterior análise, sendo assim, a quantidade de imagens selecionadas é propagada no marcador que fica próximo ao círculo que codifica os representativos.

Considerando a interação, quando o cursor do mouse está sobre um representativo, uma janela com algumas informações acerca do grupo são apresentadas, como demonstrado na Figura 24. A distribuição das instâncias por classe é apresentada por um *stackedbar*, em que as classes são codificadas pelas cores e o número de instâncias pela altura do retângulo. Também são apresentadas as imagens correspondentes às instâncias similares e diversas, recuperadas utilizando os algoritmos *KNN* (COVER; HART, 2006) e *BRID* (SANTOS, 2012), respectivamente.

Para investigar um grupo é necessário clicar na instância representativa para que o processo de expansão seja realizado. Quando uma expansão é efetuada, é realizado um *zoom in* com valor proporcional ao número de elementos do grupo sendo analisado.

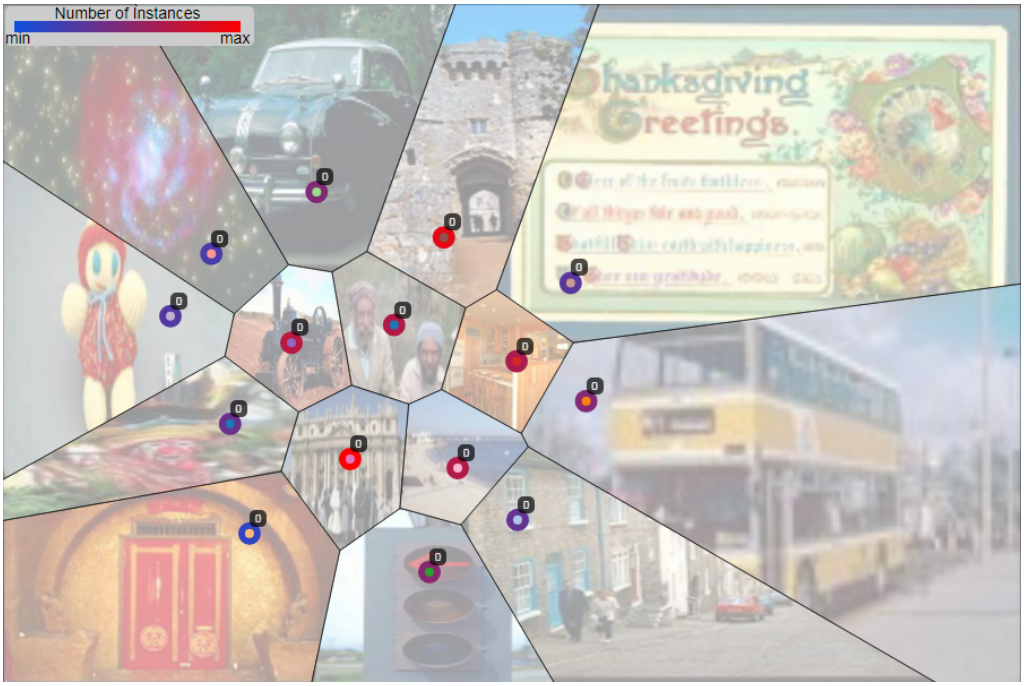


Figura 23 – Primeiro nível da abordagem de exploração.

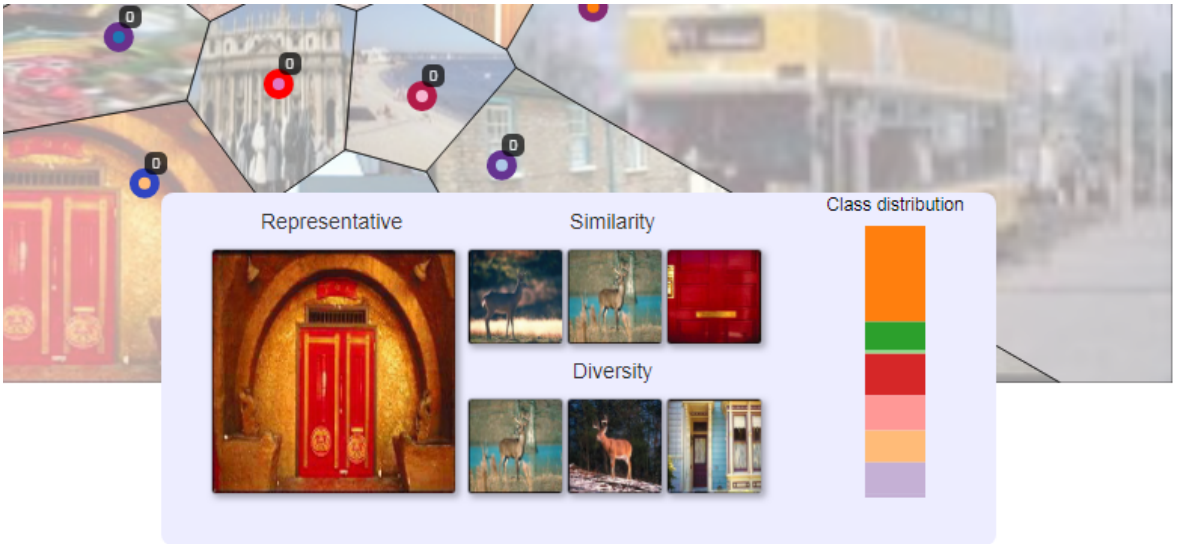


Figura 24 – *Tooltip* com informações sobre o grupo analisado.

Conforme o processo de expansão ocorre, as instâncias podem ser posicionadas para fora da projeção e, sendo assim, tais instâncias são posicionadas nas bordas do plano de projeção para que o contexto não seja perdido. Essa sequência de interação pode ser verificada na Figura 25. Note que, é possível aglomerar um grupo que foi expandido para esconder informações que não precisem ser utilizadas em dado momento.

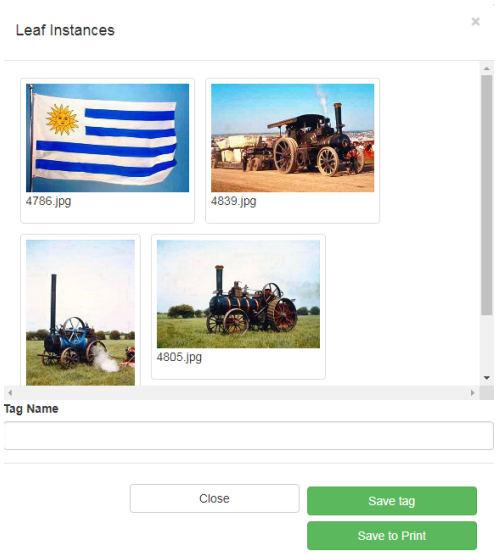
Figura 25 – Expansão de um nó. Note que como algumas instâncias iriam ser projetadas fora do plano, tais instâncias são posicionadas nas bordas da imagem.

Quando um grupo folha é alcançado, as instâncias correspondentes são apresentadas no plano de projeção quando o cursor do mouse está sobre o marcador (ver Figura 26a). Para visualizar o conteúdo das instâncias, é possível clicar sobre o representativo, como apresentado na Figura 26b.

Na Figura 26b é possível visualizar uma área para definição de *tags*, as quais são associadas às imagens selecionadas e podem servir como uma maneira de organizar o conjunto de dados sendo explorado. Não obstante, há a possibilidade de selecionar imagens para posterior análise, permitindo que o usuário possa ter maior controle das instâncias selecionadas, nesse caso, de suas imagens.

Como último detalhe acerca da abordagem de exploração multinível proposta neste trabalho, temos a codificação de metadados. Atualmente, muitos dados estão presentes para ajudar a explicar outros dados e, quando conjuntos de imagens são explorados é possível pensar em diversas características que podem ser úteis para o processo exploratório. Considerando uma rede social, por exemplo, é possível pensar na quantidade de comentários, quantidade de reações, quantidade de compartilhamentos, localização, data que a imagem

(a) Instâncias projetadas no último nível.



(b) Imagens correspondentes as intâncias projetadas.

Figura 26 – Investigação das instâncias no último nível da hierarquia.

foi capturada e entre outras. De modo a não perder tais possíveis informações, neste trabalho utilizamos mapas de calor para codificar os metadados, em um formato que se assemelhem ao da projeção. Na Figura 27 é apresentado um exemplo de tal codificação. Uma informação que sempre estará disposta é a densidade das instâncias no plano de projeção, enquanto que as outras informações dependem do fornecimento do usuário. Nesse caso, são apresentados mapas de calor para a quantidade de "Curtidas" e a quantidade de "Comentários" – gerados aleatoriamente para demonstração – para que sirvam de apoio durante a exploração.

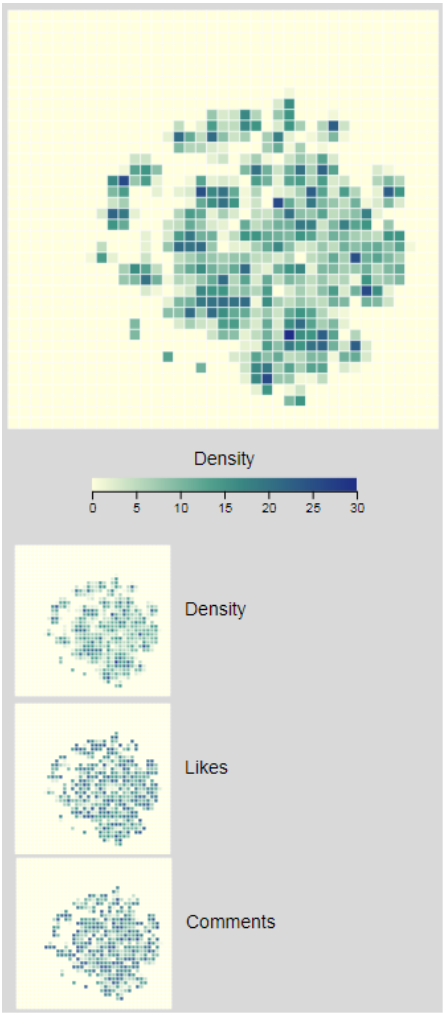


Figura 27 – Mapas de calor para codificar metadados. Cada célula do mapa de calor representa o número de instâncias presentes, para a informação de "Densidade". Enquanto, que a quantidade de "Curtidas" e de "Comentários" são utilizadas para criação dos mapas de calor das duas informações restantes. Os mapas de calor são apresentados ao lado da projeção.

4.5 CONSIDERAÇÕES FINAIS

A abordagem descrita neste capítulo tem por objetivo fornecer melhorias na exploração de conjuntos projetados utilizando técnicas de projeção multidimensional. Para isso, é empregado o modelo *Overview-First & Details-on-Demand*, em que o usuário recebe poucas informações durante o processo exploratório e detalha o conteúdo de acordo com a demanda. Além disso, uma aplicação foi apresentada como possível uso da abordagem desenvolvida.

No Capítulo 5 são apresentados alguns resultados para conjuntos de dados selecionados de forma a analisar os algoritmos necessários para criação da abordagem proposta neste trabalho, bem como estudos de casos.

5 RESULTADOS

Neste capítulo são apresentados os resultados da abordagem de acordo com alguns conjuntos de dados selecionados. Primeiramente, dois estudos de casos são apresentados para exemplificar a interação com a hierarquia, então, alguns experimentos são apresentados para verificar o desempenho das técnicas utilizadas e características da abordagem de exploração.

5.1 ESTUDOS DE CASO

5.1.1 IMAGENS DO INSTAGRAM

Nesse primeiro estudo de caso é apresentada a exploração de um conjunto de imagens retiradas do *Instagram*, cujas categorias de buscas foram *dog*, *landscape*, *flowers*, *Lofoten*, *sun*, *tree*, *sky* e *architecture*. As instâncias foram divididas em oito classes – correspondentes às chaves de busca, cada classe com dez instâncias. Na Figura 28 é apresentado um breve sumário das imagens presentes no conjunto de dados, em que duas imagens são apresentadas para cada classe.

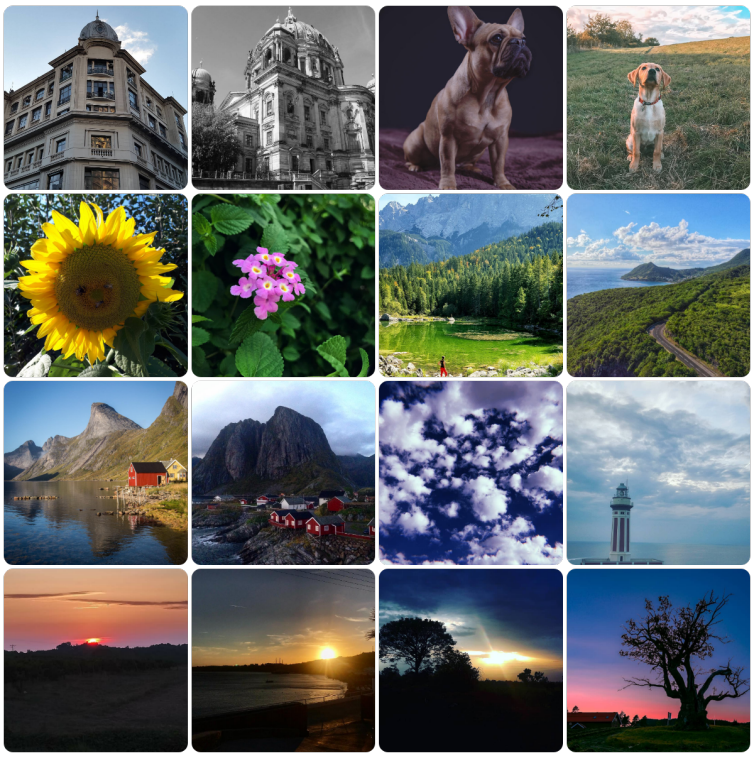


Figura 28 – Sumário das imagens presentes no conjunto de dados do primeiro estudo de caso.

A projeção inicial foi realizada com a técnica t-SNE, enquanto que a seleção de representativos foi realizada com a técnica *Affinity Propagation*. Para motivos de comparação, a projeção do conjunto de dados da forma tradicional pode ser verificada na Figura 29, as quais características foram extraídas utilizando a rede neural *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

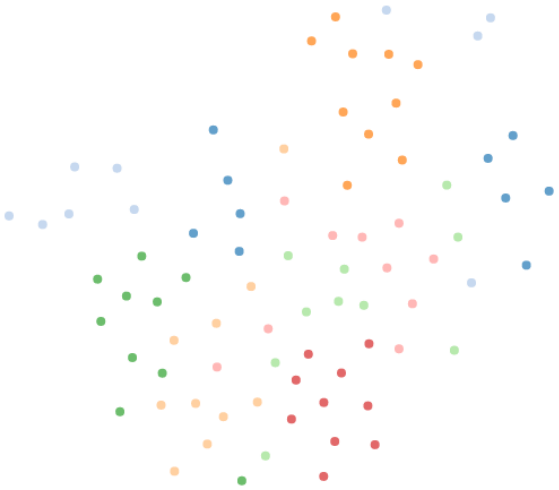
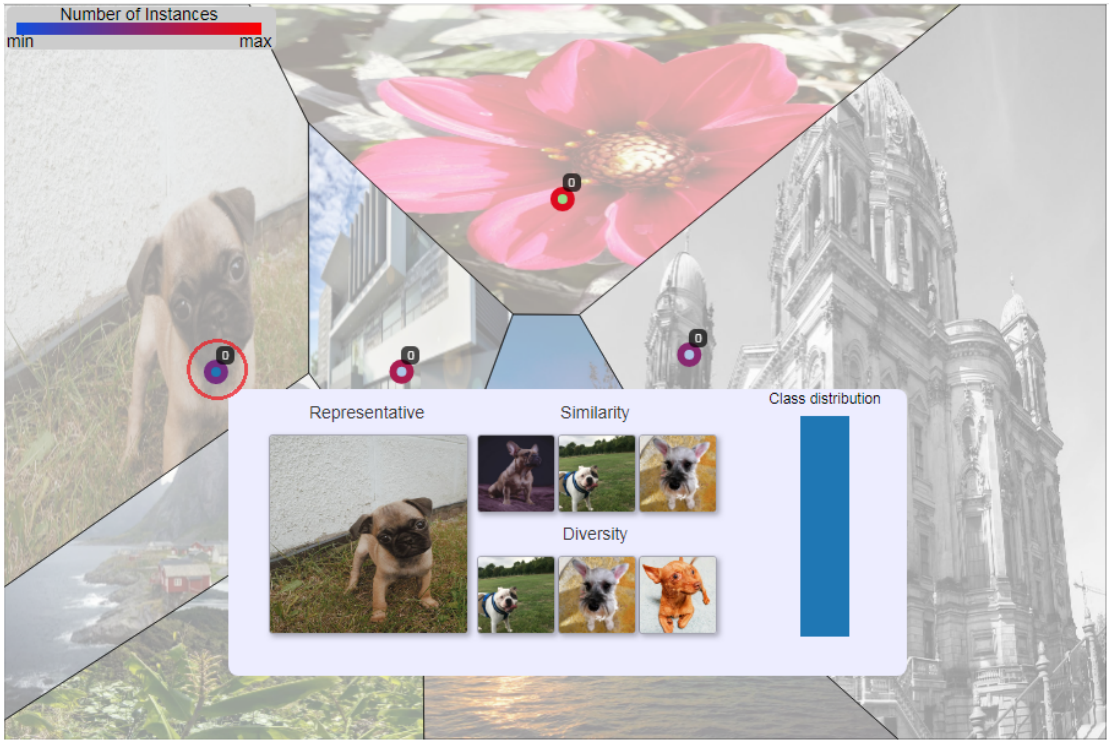


Figura 29 – Projeção do conjunto de dados de imagens retiradas do *Instagram*.

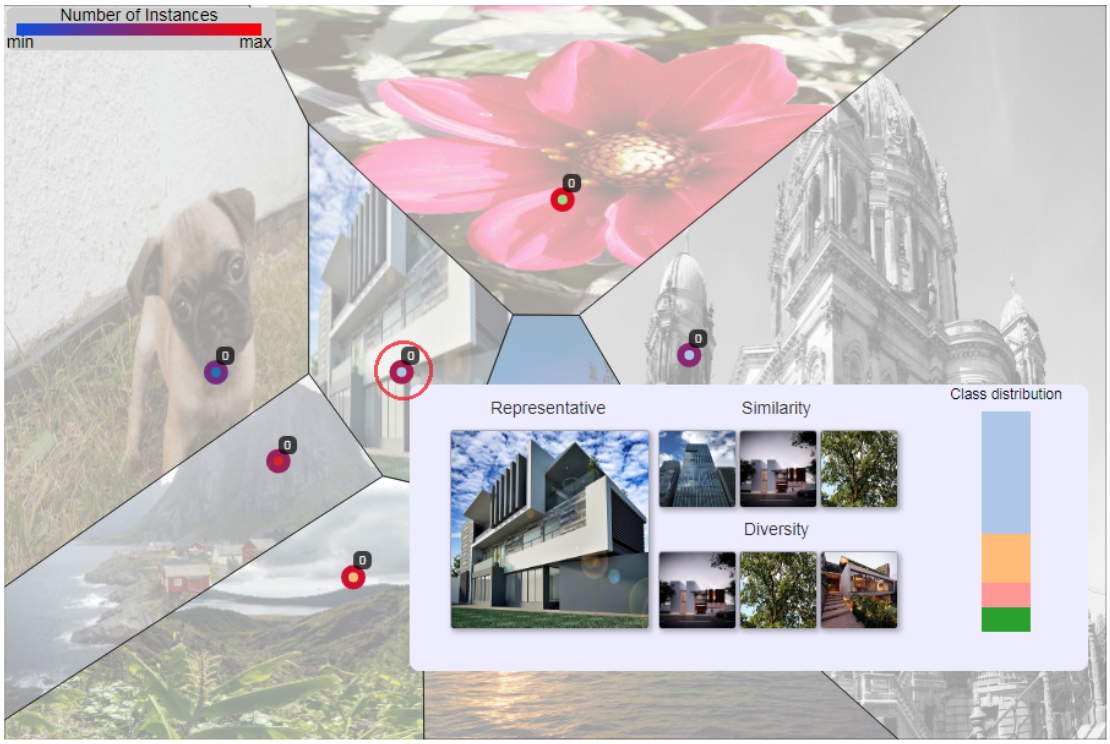
Na Figura 30 é apresentado o primeiro nível da hierarquia. Note que ao passar o mouse por cima dos representativos, *tooltips* são apresentados para melhor investigação do grupo, como apresentado nas Figuras 30a e 30b. É possível perceber que o grupo formado pelo representativo da Figura 30a é bem definido, note as instâncias similares e diversas. No entanto, existem instâncias de diferentes classes no grupo do representativo da Figura 30b, como pode ser observado pela distribuição de classes.

Na Figura 31a é apresentada a expansão do nó destacado na Figura 30a com o círculo vermelho, em que dois novos nós foram gerados. O representativo destacado em verde contém três instâncias, projetadas conforme apresentado na Figura 31b. Ao clicar nesse representativo, as imagens correspondentes aos grupos são apresentadas para possível definição de *tags* ou para selecionamento, como demonstrado na Figura 32.

Caso alguma imagem seja selecionada na Figura 32, os nós da hierarquia apresentarão o indicador dessa seleção, isto é, o marcador apresentado com o número 1 (ver Figuras 33a e 33b). Na Figura 33a existe uma diferença na quantidade de nós selecionados, no entanto, como o nó da Figura 33b é comum para ambos os nós, isto é, representa o pai dos nós da Figura 33a, tal nó contém a soma de todas imagens selecionadas.

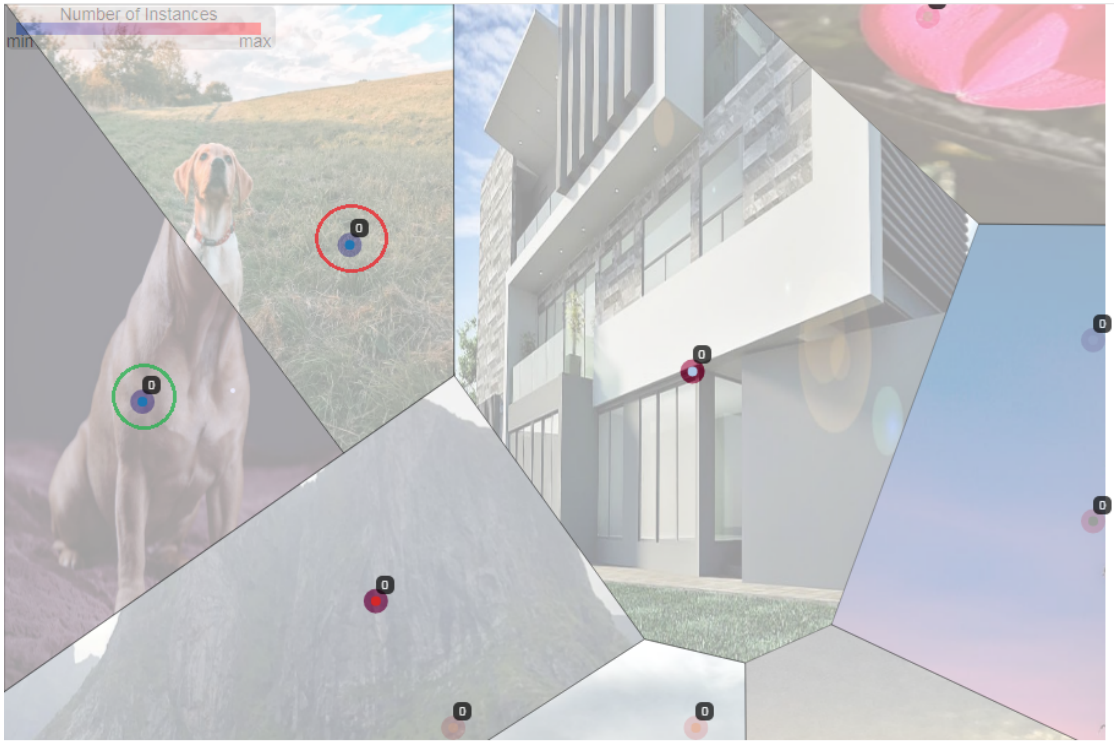


(a) Apresentação de informações dos agrupamentos.



(b) Apresentação de informações dos agrupamentos.

Figura 30 – Primeiro nível da abordagem de exploração multinível aplicada em uma projeção de imagens.



(a) Nós gerados a partir da expansão do nó destacado na Figura 30a.

(b) Instâncias pertencentes ao nó projetadas no plano.

Figura 31 – Análise da representatividade dos representativos selecionados para cada técnica.

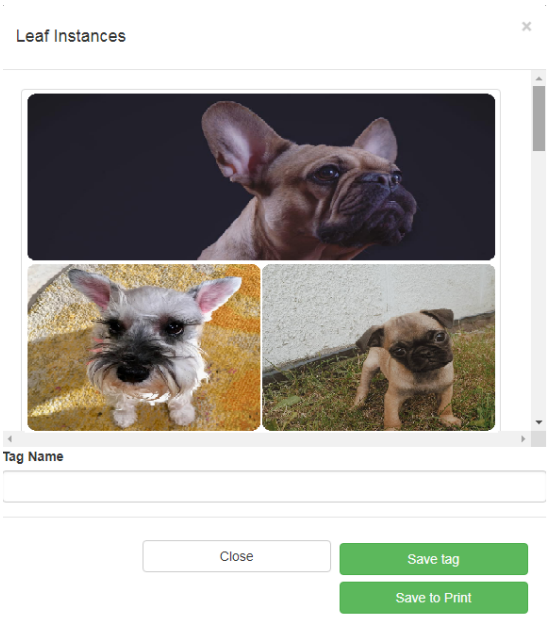


Figura 32 – Imagens correspondentes às instâncias do grupo cujas instâncias foram projetadas – grupo da Figura 31b.

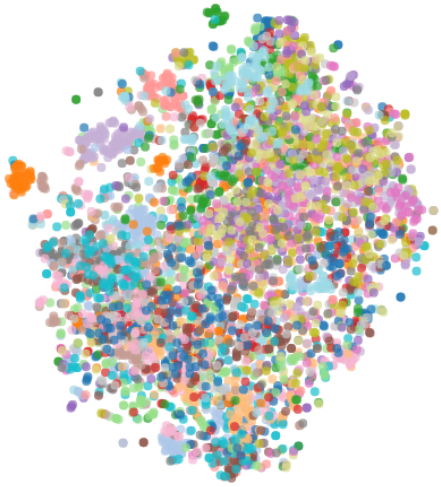
(a) Quantidade de imagens selecionadas diferentes para os dois nós filhos do nó do primeiro nível.

(b) Nó do primeiro nível com quantidade de nós igual a quantidade de imagens selecionadas nos nós filhos.

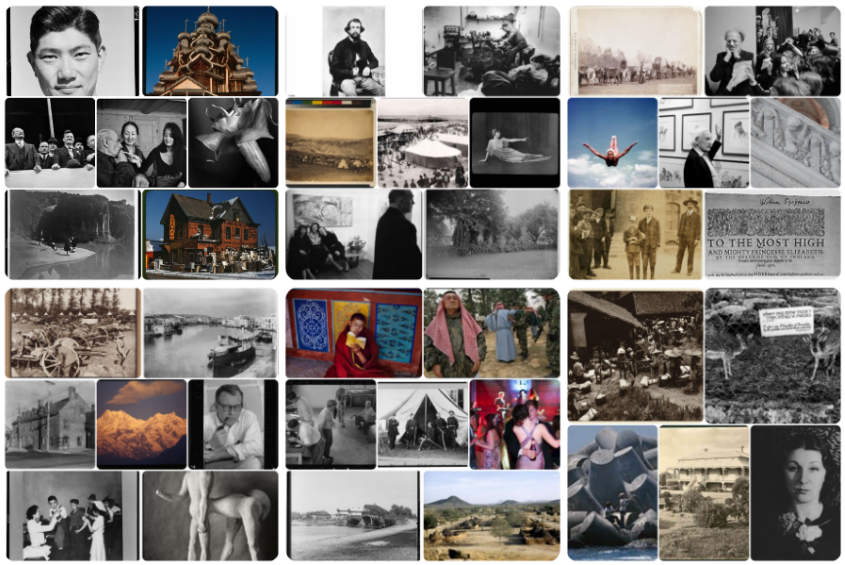
Figura 33 – Demonstração da quantidade de nós selecionadas para os nós.

5.1.2 IMAGENS DE FOTÓGRAFOS

Nesse segundo estudo de caso é explorado um conjunto de dados de imagens de fotógrafos com 4100 instâncias, cujas 41 classes representam diferentes fotógrafos. As características das imagens também foram extraídas utilizando *deep features* e a projeção inicial efetuada com a técnica t-SNE. Na Figura 34a é apresentada a projeção inicial do conjunto de dados. Note que, apesar de instâncias de diferentes classes estarem próximas umas das outras, as imagens possuem grande semelhança, o que pode ser verificado na Figura 34b – que apresenta um sumário, assim como no primeiro estudo de caso, das imagens presentes no conjunto de dados.



(a) Projeção do conjunto de dados de imagens de fotógrafos.



(b) Sumário das imagens presentes no conjunto de dados do segundo estudo de caso.

Figura 34 – Projeção e sumário do conjunto de dados de fotógrafos.

Na Figura 35 são apresentados alguns detalhes dos representativos do primeiro nível da hierarquia. Note que os grupos possuem muitas classes diferentes, no entanto, a desordem visual utilizando a abordagem proposta neste trabalho é muito menor se comparada a abordagem tradicional de visualização. Vale ressaltar que algumas texturas podem não aparecer porque como nesse estudo de caso as imagens são recuperadas da Internet – muitas imagens para serem armazenadas, algumas *url's* podem não especificar o local correto de armazenamento.

Continuando com a exploração, ao expandir o nó da Figura 35c, os nós da Figura 36 são apresentados. Note que, visualizando as distribuições de classes dos nós das Figuras 36a e 36b, é possível perceber que grupos começam a possuir instâncias com uma classe majoritária, de modo que a análise começa a ser facilitada conforme a interação com a hierarquia.

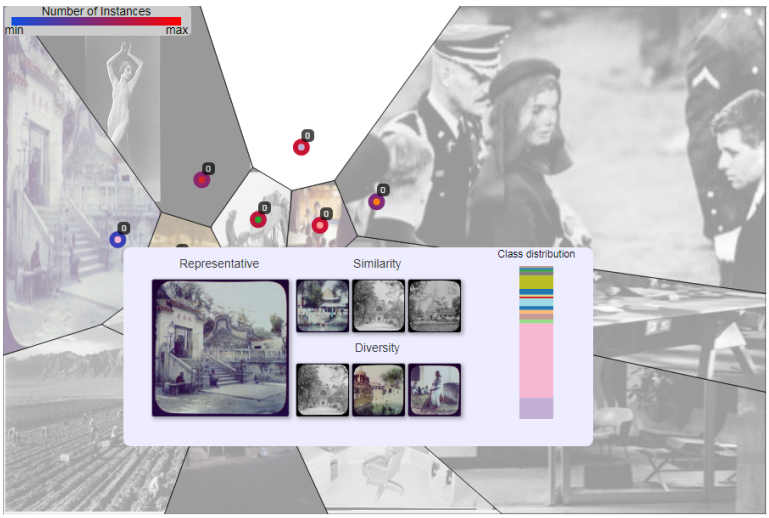
Para finalizar o estudo de caso, na Figura 37 é apresentado a seleção de instâncias para dois nós destacados na Figura 36, note que a quantidade de instâncias selecionadas são atualizadas nos nós pais.

Nesse segundo estudo de caso foi possível perceber que apesar da quantidade de instâncias do conjunto de dados, a nossa abordagem é capaz de fornecer uma exploração mais facilitada se considerarmos a carga cognitiva para análise. Por fornecer as informações aos poucos, o usuário consegue interagir com o conjunto de dados de forma seletiva.

5.2 METODOLOGIA DE EXPERIMENTAÇÃO

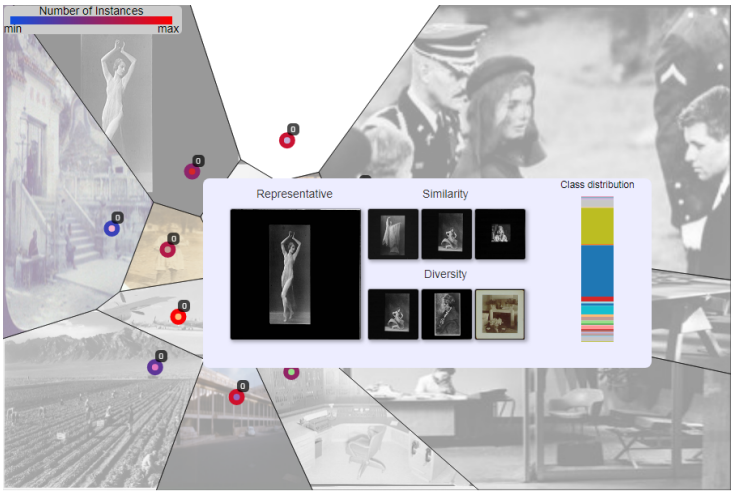
Nesta seção são apresentados experimentos para verificar o desempenho da abordagem considerando os diferentes algoritmos que podem ser utilizados para criação da abordagem hierárquica de exploração proposta neste trabalho, isto é, algoritmos de seleção de representativos, algoritmos de remoção de sobreposição e técnicas de projeção multidimensional. Além disso, também é apresentado uma comparação do algoritmo de remoção de sobreposição proposto neste trabalho com dois dos melhores algoritmos presentes na literatura.

Na Tabela 1 são apresentados os conjuntos de dados utilizados para os experimentos. Note que um mesmo conjunto de dados é utilizado com diferentes configurações para demonstração da abordagem de exploração. A coluna "Características" representa como as características foram extraídas dos conjuntos de dados, sendo assim, as características indicadas por "AlexNet" foram utilizadas as *deep features* da rede neural *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Para a projeção dos conjuntos, somente o conjunto de dados Corel_2 foi projetado por meio da técnica LSP, o restante dos algoritmos foram projetados por meio da técnica t-SNE.



(a) Sumário do grupo do representativo destacado em que há uma grande dominância de instâncias codificadas com a classe de cor rosa.

(b) Sumário do grupo destacado pelo representativo com a foto de uma criança. Note que nesse caso há grande dominânca das classes codificadas com as cores azul e verde.



(c) Sumário do grupo destacado pelo representativo com a foto em tons de cinza. As instâncias desse grupo são dominadas pelas classes de cor azul e tom amarelado.

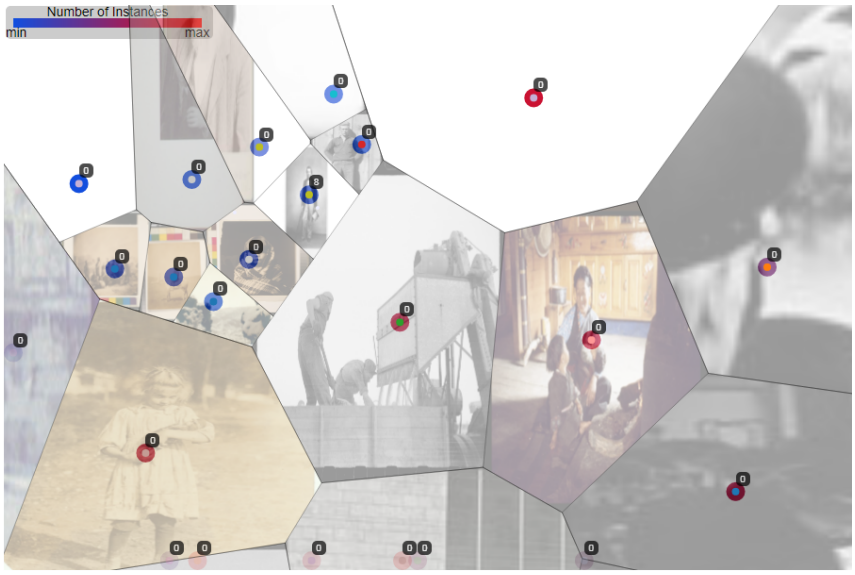
Figura 35 – Representativos e texuras do primeiro nível da hierarquia.

(a) Posicionamento das instâncias do nó folha que contém instâncias de duas classes.

(b) Posicionamento das instâncias do nó folha que contém instâncias de apenas uma classe.

Figura 36 – Expansão do nó destacado na Figura 35c.

(a) Instâncias selecionadas em ambos os nós destacados na Figura 36. Em um nó foram selecionadas três imagens e em outro nó foram selecionadas cinco imagens.



(b) Aglomeração dos nós da Figura 37a causaram a atualização do número de imagens selecionadas

Figura 37 – Demonstração da atualização do número de imagens selecionadas.

Dataset	# instâncias	# dimensões	Características
Corel_1	1000	150	Características de cores[1]
Corel_2	1000	150	Características de cores[1]
Corel_3	5000	4096	AlexNet
Corel_4	5000	2	Descritores de Fourier
Fotografos_1	4100	4096	AlexNet
Fotografos_2	4510	4096	AlexNet
Fotografos_3	9077	4096	AlexNet

[1] *Color Histogram; Color Histogram Layout; Color Moments; Co-occurence textures.*

Tabela 1 – Conjuntos de dados utilizados nos experimentos.

5.2.1 DESIGN DOS EXPERIMENTOS

Os experimentos foram realizados para avaliar as técnicas *Affinity Propagation*, k-Means, *Bisecting* k-Means, k-Medoid, SSS e k-SVD, utilizando cada métrica (descritas a seguir) para avaliar o primeiro nível da hierarquia. Essa escolha deve-se ao fato de que uma boa divisão no primeiro nível irá gerar divisões melhores nos níveis seguintes devido à separação de grupos. Além disso, também são apresentados aspectos relacionados ao impacto do espaço de características utilizado para descrever os conjuntos de dados, bem como aspectos relacionados as técnicas de projeção multidimensional.

5.2.2 IMPLEMENTAÇÃO

A abordagem foi implementada utilizando o *framework* D3.js para desenvolvimento da metáfora visual, enquanto a linguagem Java e o *framework* Spring MVC foram utilizados para implementação do *back-end*. As técnicas de remoção de sobreposição e seleção de representativos foram fornecidas por uma API Java, também desenvolvida durante este trabalho.

O computador utilizado nos experimentos possui a seguinte configuração:

- CPU Intel Core i5-4440 3.10GHz;
- 8GB DDR4 RAM;
- Windows 7 64 bits.

5.3 EXPERIMENTOS

Os experimentos para analisar os representativos foram realizados segundo quatro métricas – Tempo, Coeficiente de Silhueta (CS), *Histogram Difference Measure* (HDM) e *Nearest Neighbor Measure* (NNM) (CUI et al., 2006). Para facilitar a explicação das duas últimas métricas, é possível definir dois termos: *Data Abstraction Level* (DAL), que

se refere a razão entre o tamanho do conjunto reduzido e o conjunto original, e *Data Abstraction Quality* (DAQ), que denota o grau com qual o conjunto reduzido representa o conjunto original (CUI et al., 2006).

A métrica HDM é utilizada para atuar como DAQ. Primeiramente são calculados dois histogramas com o mesmo número de caixas do conjunto original e conjunto reduzido. O tamanho das caixas correspondem a porcentagem do número de pontos que pertencem as caixas. A diferença de histograma corresponde ao somatório das diferenças entre as caixas correspondentes aos dois histogramas. Enquanto isso, a métrica NNM é definida como a média normalizada das distâncias entre cada instância do conjunto original para seu representativo.

A métrica Coeficiente de Silhueta (KAUFMAN; ROUSSEEUW, 1990) é utilizada para interpretar a consistência dos agrupamentos e para fornecer a informação de quão bem cada agrupamento participa de seu agrupamento. Nesse caso, os agrupamentos gerados por uma técnica de projeção e seus rótulos são definidos pelas instâncias das classes. Dada uma instância i, o valor do coeficiente de silhueta pode ser calculado da seguinte maneira:

1. Calcule a média das instâncias a_i de i para todas as outras instâncias pertencendo ao mesmo agrupamento. Tal valor fornece uma medida de coesão;
2. Calcule a menor distância b_i de i para todas as outras instâncias dos outros agrupamentos, fornecendo uma medida para analisar a separação de agrupamentos;
3. A silhueta de i é definida como $s_i = \dfrac{b_i - a_i}{\max(a_i, b_i)}$.

O valor do Coeficiente de Silhueta é dado pela média da silhueta de todas as instâncias e varia de -1 à 1, em que melhor coesão e separação de agrupamentos são indicados por valores próximos de 1.

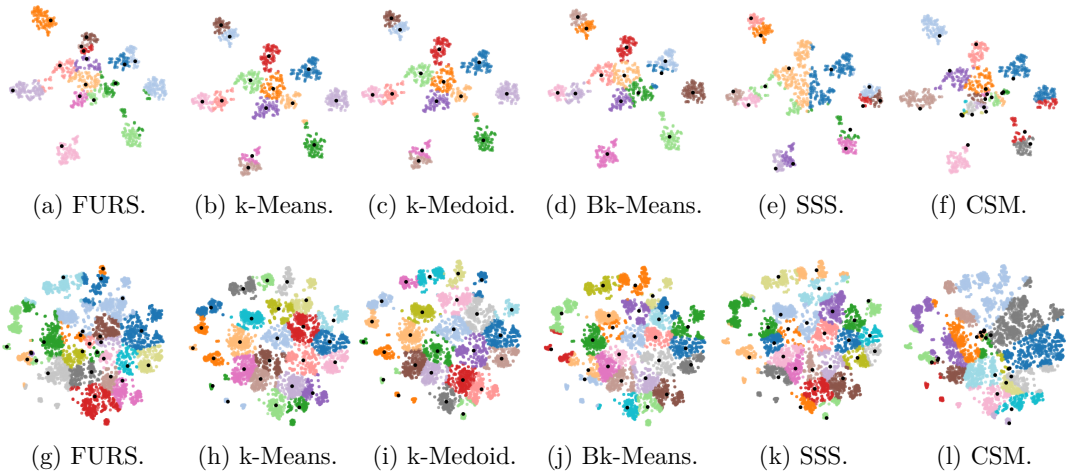
Conforme os experimentos foram realizados, alguns algoritmos foram removidos da competição pela alta complexidade e necessidade de recursos. Esses algoritmos são listados na Tabela 2.

Método	Tempo (ms)	CS	HDM	NNM
Affinity Propagation	2298	0.44534686	0.5978235	0.8727388229
DS3	1023051	0.23095012	0.790549177	0.64722630
k-SVD	2118195	0.30640015	0.6882944	0.6680884

Tabela 2 – Algoritmos removidos dos experimentos devido à alta complexidade e necessidade de recursos, isto é, tais algoritmos necessitam de quantidades muito elevadas de memória e tempo de processamento.

5.3.1 REPRESENTATIVOS

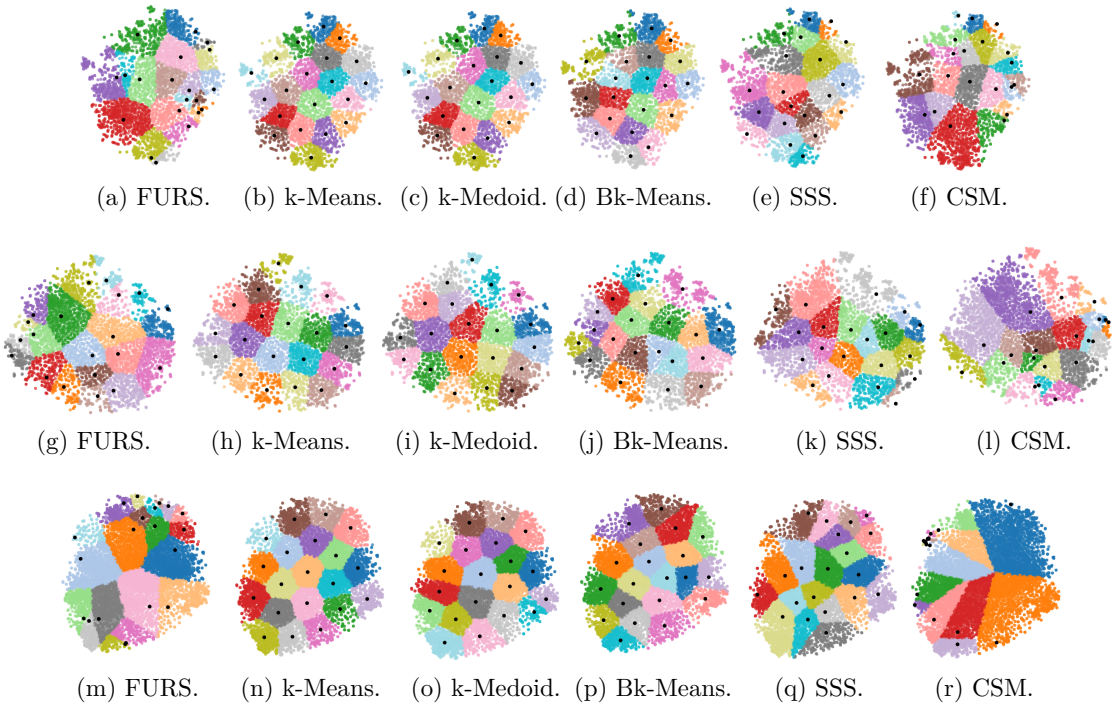
Os representativos selecionados pelas técnicas podem ser visualizados nas Figuras 38 e 39. As instâncias com a cor preta codificam os representativos selecionados, enquanto que as cores das outras instâncias indicam a divisão imposta pelos representativos.



(a) FURS. (b) k-Means. (c) k-Medoid. (d) Bk-Means. (e) SSS. (f) CSM.

(g) FURS. (h) k-Means. (i) k-Medoid. (j) Bk-Means. (k) SSS. (l) CSM.

Figura 38 – Resultado para Corel_1 apresentado pelas imagens de (a) - (f), e resultado para Corel_3 apresentado pelas imagens de (g) - (l).



(a) FURS. (b) k-Means. (c) k-Medoid. (d) Bk-Means. (e) SSS. (f) CSM.

(g) FURS. (h) k-Means. (i) k-Medoid. (j) Bk-Means. (k) SSS. (l) CSM.

(m) FURS. (n) k-Means. (o) k-Medoid. (p) Bk-Means. (q) SSS. (r) CSM.

Figura 39 – Resultado para Fotografos_1 apresentado pelas imagens de (a) - (f), resultado para Fotografos_2 apresentado pelas imagens de (g) - (l), e resultado para Fotografos_3 apresentado pelas imagens de (m) - (r).

Uma das primeiras características a serem notadas pela visualização do resultado é a distribuição dos representativos selecionados. Note que as técnicas k-Means, Bk-Means, k-Medoid e SSS sempre conseguem selecionar representativos mais distribuídos pela projeção. As implicações para essas características são discutidas com mais detalhes nas seções seguintes.

5.3.1.1 Tempo de processamento

Na Figura 40 são apresentados os resultados para o tempo de execução dos algoritmos em escala logarítmica. É possível notar a semelhança do tempo de execução entre os algoritmos k-medoid e FURS, *Bisecting* k-means e k-means, isso devido ao fato que os algoritmos k-medoid e FURS atuam sempre com uma instância para representar um agrupamento, isto é, os *medoids*. Enquanto que as técnicas k-means e sua variante *Bisecting* k-means utilizam as médias das instâncias para representar um agrupamento. Finalmente, a técnica SSS demonstrou escalabilidade com aumento do número de instâncias, enquanto que a técnica CSM apresentou tempos de processamento proibitivos – o grande tempo de processamento exigido pelo algoritmo CSM é devido a decomposição em valores singulares utilizada para selecionar os representativos.

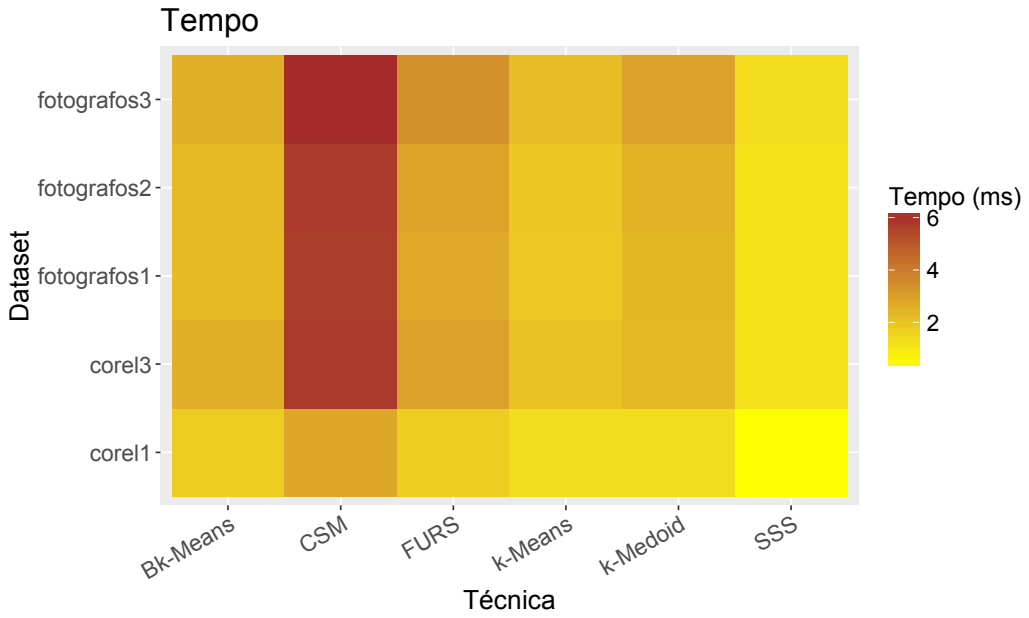


Figura 40 – Tempo de processamento dos algoritmos em escala logarítmica.

5.3.1.2 Coeficiente de Silhueta

Na Figura 41 são apresentados os resultados para a métrica Coeficiente de Silhueta, em que valores próximos de 1 indicam melhores resultados. Analisando o resultado, é possível notar que as técnicas utilizadas para agrupamento – k-Means, Bk-Means e k-Medoid – e a técnica SSS apresentaram resultados mais satisfatórios em relação as outras

justamente pelo fato de tentarem encontrar agrupamentos mais separados uns dos outros, isso faz com que, para a métrica Coeficiente de Silhueta, os agrupamentos sejam avaliados como satisfatórios. As técnicas CSM e FURS, apresentaram resultados inferiores pelo fato de selecionarem representativos mais próximos uns dos outros (ver Figuras 38 e 39). É interessante notar também que a técnica SSS apresenta resultados satisfatórios em relação as métricas de seleção de representativos propriamente dito, isso porque a técnica SSS seleciona representativos que estejam à uma distância maior ou igual a uma fração da distância máxima entre os elementos da projeção.

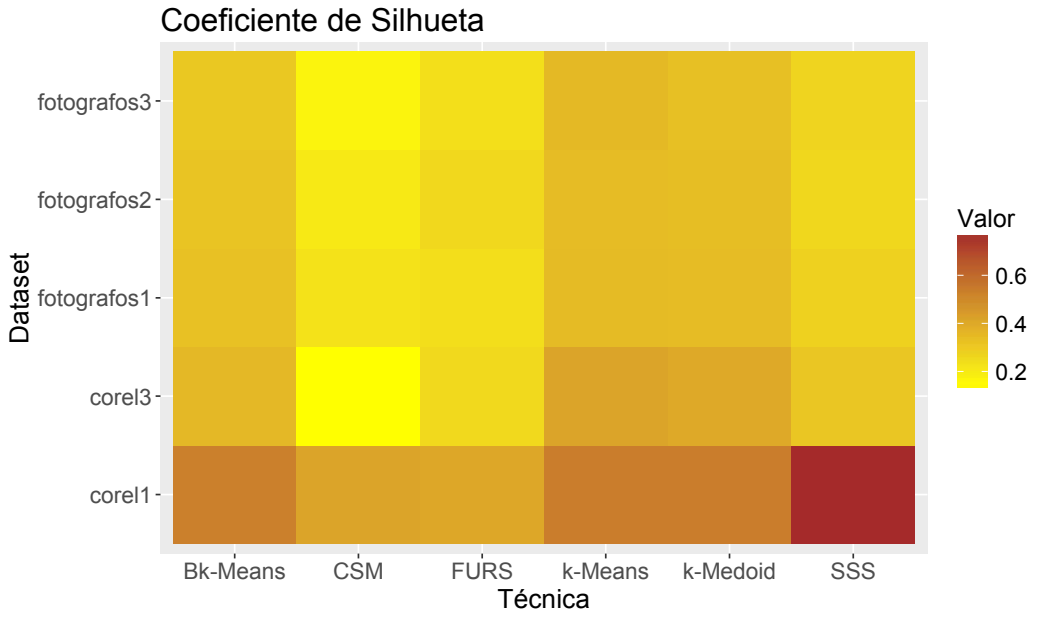
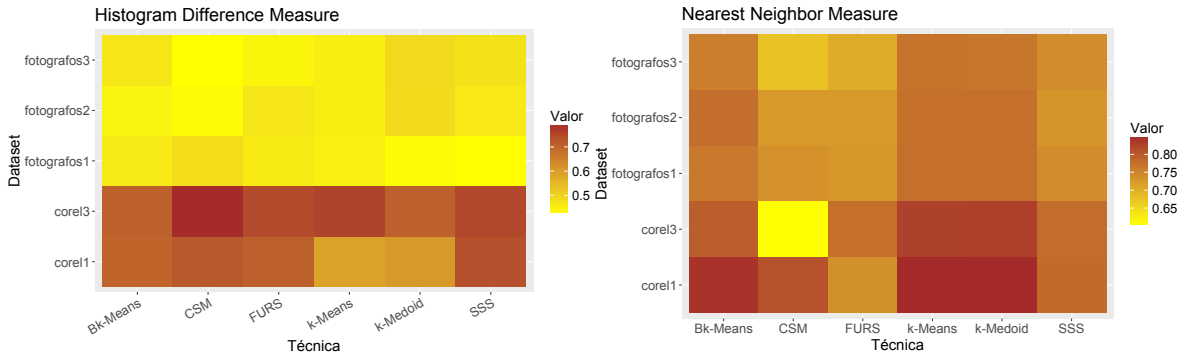


Figura 41 – Coeficiente de Silhueta.

5.3.1.3 Histogram Difference Measure and Nearest Neighbor Measure

A métrica de qualidade NNM está relacionada com a razão entre o raio da abstração e o raio do conjunto original. Conforme DAL cresce, o raio de abstração diminui, fazendo com que a qualidade aumente. Para ambas as técnicas, NNM e HDM, valores próximos de 1 indicam melhor qualidade.

Considerando o resultado para a métrica NNM, mais uma vez as técnicas de agrupamento apresentaram os melhores resultados. As técnicas FURS e CSM, com exceção dos conjuntos de dados *Fotografos1* e *Fotografos2*, selecionaram representativos próximos uns dos outros, diminuindo a qualidade segundo a métrica NNM. Para a métrica HDM, é possível notar a influência do conjunto de dados para a qualidade dos representativos segundo essa técnica. Para os conjuntos derivados de *fotografos*, em que não há uma grande separação de classes, as técnicas tendem a apresentar um resultado inferior. No entanto, é possível perceber ainda que os algoritmos apresentam resultados confusos em relação ao que foi apresentado visualmente – repare nos representativos selecionados pela técnica



(a) HDM. (b) NNM.

Figura 42 – Métricas HDM e NNM.

CSM (ver Figura 38l) para o conjunto Corel_3 e o resultado no gráfico da Figura 43e. Por isso, outra bateria de testes foi realizada considerando a representatividade do conjunto.

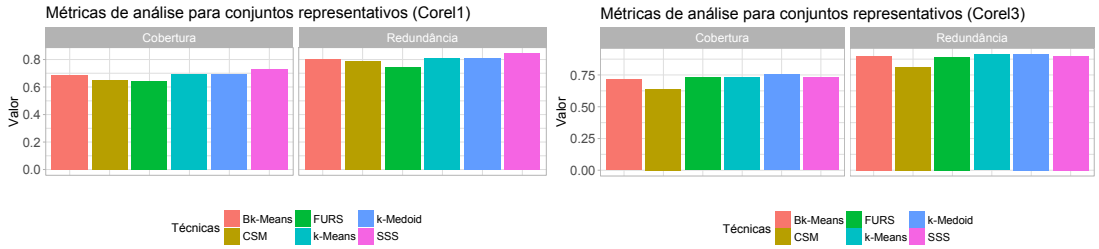
5.3.1.4 Representatividade

Para complementar a análise acerca dos representativos, os algoritmos também foram analisados segundo sua representatividade. Essa análise foi realizada por meio das métricas apresentadas na Seção 2.4. Na Figura 43 são apresentados os resultados dos algoritmos segundo a representatividade, isto é, a combinação entre *Cobertura* e *Redundância*.

As métricas de Cobertura e de Redundância conseguem mostrar algumas características que são percebíveis visualmente. Por exemplo, a técnica CSM apresenta grande redundância pelo fato de selecionar representativos que estão muito próximos um dos outros. Além disso, o desempenho para a métrica Cobertura também é prejudicado pelo fato dos representativos são conseguirem cobrir o espaço de projeção. Dessa maneira, podemos notar que segundo essas métricas, as técnicas mais suscetíveis para o uso são, em ordem: SSS, k-Medoid, k-Means, Bk-Means, FURS e CSM.

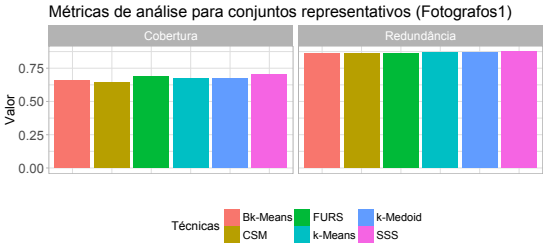
5.3.2 IMPACTO DA PROJEÇÃO E DO ESPAÇO DE CARACTERÍSTICAS

Nessa seção são discutidos o impacto do espaço de características utilizado para projeção, assim como a própria técnica de projeção utilizada para projetar as instâncias dos dados, no processo de criação da abordagem de exploração proposta neste trabalho. As considerações feitas nesta seção são importantes porque o primeiro nível da hierarquia apresenta uma visão geral do conjunto de dados, é importante que os grupos sejam formados por instâncias de classes similares. O objetivo é poder analisar as seguintes questões:

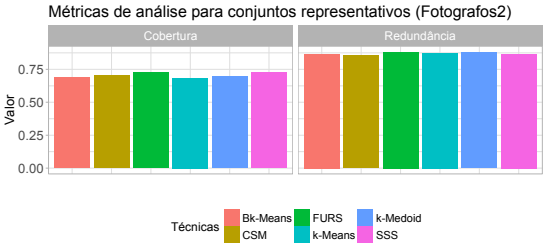


(a) Representatividade Corel_1.

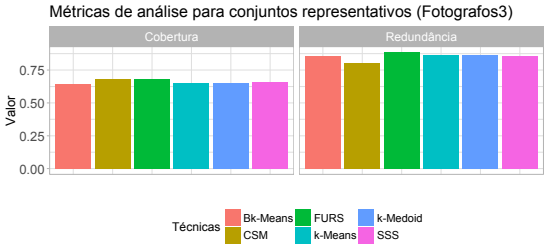
(b) Representatividade Corel_3.



(c) Representatividade Fotografos_1.



(d) Representatividade Fotografos_2.



(e) Representatividade Fotografos_3.

Figura 43 – Análise da representatividade dos representativos selecionados para cada técnica.

1. Qual é o Coeficiente de Silhueta das projeções em diferentes configurações?
2. Qual é o desempenho dessas características extraídas em um processo de classificação?
3. Qual é a distribuição de classes nos grupos gerados por essas projeções?

De modo a responder essas questões, foram realizadas projeções dos conjuntos de dados Corel_1, Corel_3 e Corel_4 por meio das técnicas LSP e t-SNE, para verificar o desempenho da abordagem de exploração considerando uma técnica que tem por característica a separação eficaz de classes (t-SNE) e uma técnica que tem por característica a preservação de vizinhança eficaz (LSP). Tais projeções podem ser verificadas na Figura 44. Um dos pontos a serem observados nas duas projeções é a proximidade das instâncias de diferentes classes, indicadas por cores diferentes. As projeções geradas com a técnica t-SNE apresentam resultados superiores visualmente que as projeções geradas com a técnica LSP, considerando a separação de classes. Na Tabela 3 é possível verificar o Coeficiente de Silhueta para as projeções. Note que o valor para técnica *LSP* considerando o conjunto

de dados Corel_4 não foi apresentado visto que o objetivo foi mostrar como um espaço de características ruim pode impactar na separação de grupos feita por uma projeção – nesse caso utilizamos a técnica t-SNE, que consegue realizar uma melhor separação das classes.

Técnica	Corel_1	Corel_3	Corel_4
LSP	0.39537978	−0.26456943	–
t-SNE	0.46870533	−0.15611145	−0.30868745

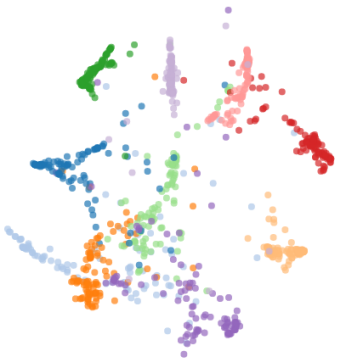
Tabela 3 – Coeficiente de Silhueta das projeções dos conjuntos de dados Corel_1, Corel_3 e Corel_4.

É possível notar como a técnica de projeção é importante para a abordagem de exploração. Ao utilizar uma técnica de projeção que é conhecida por separar as classes das instâncias de forma satisfatória (isto é, a técnica t-SNE), a desordem visual é menor. Não obstante, o impacto do espaço de características também é de grande importância para o sucesso de uma boa exploração. Note, por exemplo, a projeção que é apresentada na Figura 44e. Outro fator a ser observado é o desempenho desses espaços de características em um processo de classificação, a Tabela 4 apresenta como esses dois espaços de características influenciaram na classificação utilizando o algoritmo *Random Forest*.

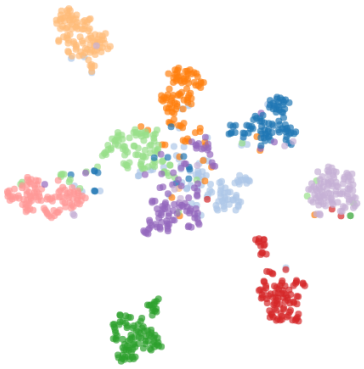
Conjunto de dados	Classificadas corretamente	Classificadas incorretamente
Corel_3	696	4304
Corel_4	2959	2041

Tabela 4 – Desempenho dos espaços de características no processo de classificação.

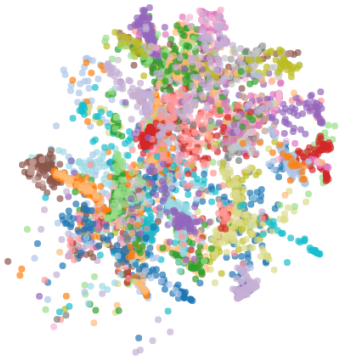
De modo a verificar a influência dos espaços de características e das técnicas de projeção para a abordagem proposta neste trabalho, é possível considerar a distribuição de classes no primeiro nível da hierarquia da abordagem de exploração, dado que esse nível influencia na qualidade dos outros níveis. Tal resultado pode ser verificado na Figura 45, em que tons mais próximos de vermelho significam a presença de um maior número de classes. Dessa maneira, conforme o que foi comentado no parágrafo anterior, note que as projeções com melhor coesão (tons mais azulados) são correspondentes às projeções realizadas pela técnica t-SNE e com espaço de características eficiente.



(a) Conjunto de dados Corel_1 projetado com LSP.



(b) Conjunto de dados Corel_1 projetado com t-SNE.

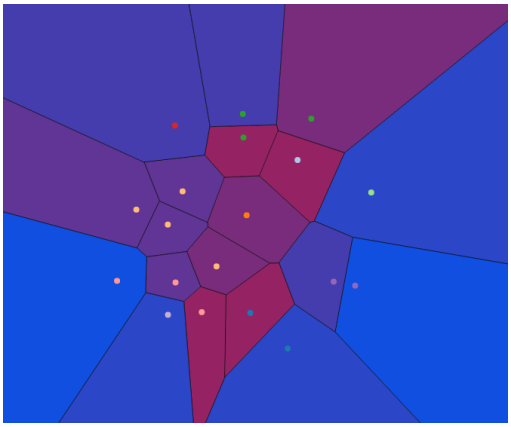


(c) Conjunto de dados Corel_3 projetado com LSP.

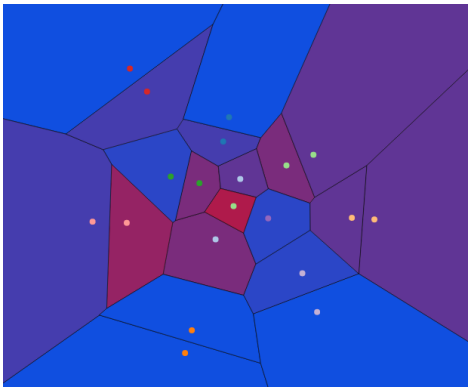
(d) Conjunto de dados Corel_3 projetado com t-SNE.

(e) Conjunto de dados Corel_4 projetado com t-SNE.

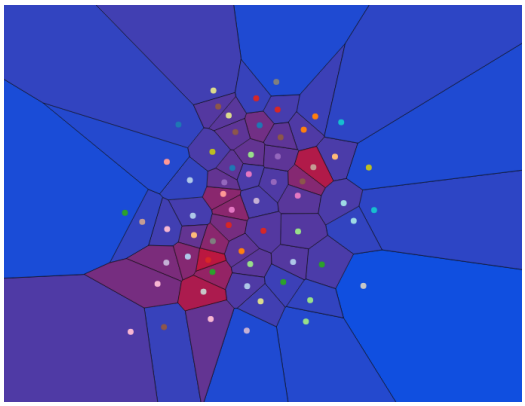
Figura 44 – Projeções dos conjuntos de dados.



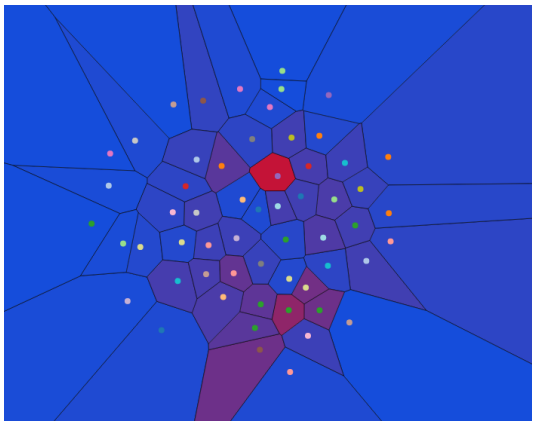
(a) Distribuição de classes do conjunto de dados Corel_1 projetado com LSP.



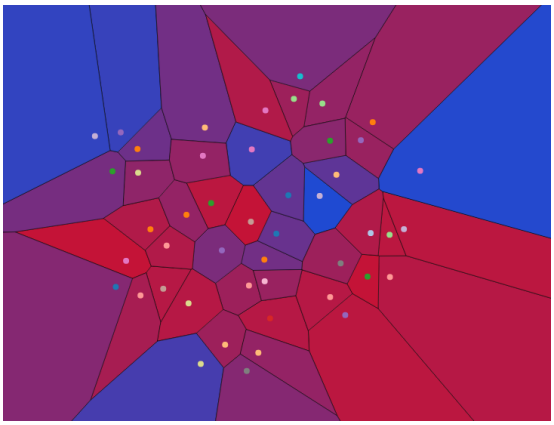
(b) Distribuição de classes do conjunto de dados Corel_1 projetado com t-SNE.



(c) Distribuição de classes do conjunto de dados Corel_3 projetado com LSP.



(d) Distribuição de classes do conjunto de dados Corel_3 projetado com t-SNE.



(e) Distribuição de classes do conjunto de dados Corel_4 projetado com t-SNE.

Figura 45 – Distribuição de classes após a criação da abordagem de exploração. Células com tons mais próximos ao azul indicam menor quantidade de classes e, portanto, maior qualidade.

5.4 ANÁLISE DO ALGORITMO DE REMOÇÃO DE SOBREPOSIÇÃO

Nesta seção é apresentada a análise do algoritmo de remoção de sobreposição proposto neste trabalho, chamado de *ExpandingNode*, comparado com as técnicas *PRISM* e *VPSC* mediante as métricas *Neighborhood Preservation* (PAULOVICH; MINGHIM, 2008) (NP), *Neighborhood Hit* (PAULOVICH et al., 2008) (NH) e tempo. Enquanto a métrica *Neighborhood Preservation* mede a preservação de vizinhança das instâncias após o processo de redução multidimensional, a métrica *Neighborhood Hit* considera a preservação de classes.

O algoritmo *ExpandingNode* se baseia no aumento das arestas formadas por pares de instâncias. O método é descrito no Algoritmo 1. Na linha 5, é verificado se existe interseção para cada par de instância (i, j), caso haja interseção, o nó i é movido na direção (j, i) usando a área de interseção. Como esse processo pode gerar mais sobreposição, tal processo é repetido para os nós de 0 até $i - 1$, como pode ser verificado nas linhas de 7 a 15.

Algoritmo 1: Função ExpandingNode.

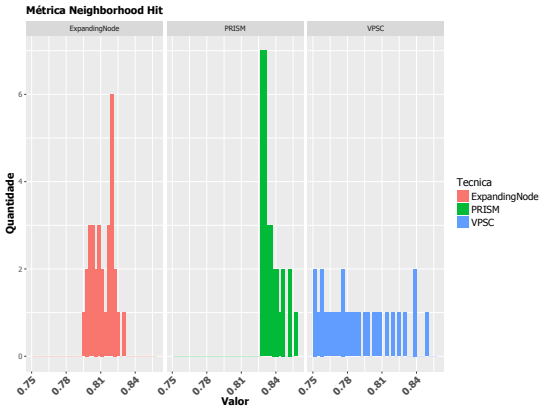
```
   /* S Conjunto de pontos                                          */
   /* e Instância base para ordenação                               */
 1 função ExpandingNode(S, e): booleano
 2     ordenar S com base na distância para e
 3     para i variando de 0 até S.length faça
 4         para j variando de i + 1 até S.length faça
 5             se S_i intersecta S_j então
 6                 mover S_i na direção (S_j, S_i) com a distância da interseção
 7                 para k variando de i até 0 faça
 8                     S_1 ← Subset(S, k − 1, 0)
 9                     ordenar S_1 com base na distância para S_k
10                     para l variando de 0 até S_1.length faça
11                         se S_k intersecta S_{1_l} então
12                             mover S_{1_l} na direção (S_k, S_{1_l}) com a distância da interseção
13                         fim
14                     fim
15                 fim
16             fim
17         fim
18     fim
19 fim
```

Para avaliação dos algoritmos foram considerados diferentes grupos das projeções, de modo a verificar o relacionamento entre a qualidade dos algoritmos utilizados considerando

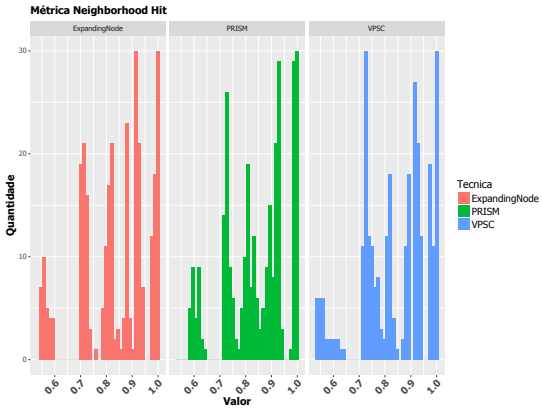
as métricas e o número de instâncias nos agrupamentos. Por exemplo, para o conjunto de dados Corel_1, a projeção foi dividida em 1 agrupamento, em 10 agrupamentos, em 20 agrupamentos, em 50 agrupamentos e em 100 agrupamentos. Para cada uma dessas configurações foram aplicados os algoritmos de remoção de sobreposição e feita a análise. Assim, para 1 agrupamento a análise foi feita sobre toda a projeção, para 10 agrupamentos a análise foi feita considerando os 10 diferentes grupos, e assim sucessivamente. Analogamente, para o conjunto de dados Fotografos_1, a análise foi realizada considerando projeção com 1, 10, 50, 200 e 300 agrupamentos.

Nas Figuras 46, 47, 48 e 49 são apresentados os resultados para os conjuntos de dados analisados segundo as métricas *Neighborhood Hit* (ver Figuras 46 e 48) e *Neighborhood Preservation* (ver Figuras 47 e 49). Note que para ambos os casos a técnica *ExpandingNode* aproxima-se do resultado das outras técnicas quando o número de agrupamentos aumenta, isto é, quando o número de instâncias é menor, a técnica proposta neste trabalho apresenta desempenho similar. A análise por meio dos histogramas fornece o entendimento de que a técnica proposta neste trabalho muito semelhante às técnicas analisadas quando o número de agrupamentos nos conjuntos de dados aumenta, isto é, já é possível notar resultados semelhantes quando o conjunto de dados é dividido em dez grupos.

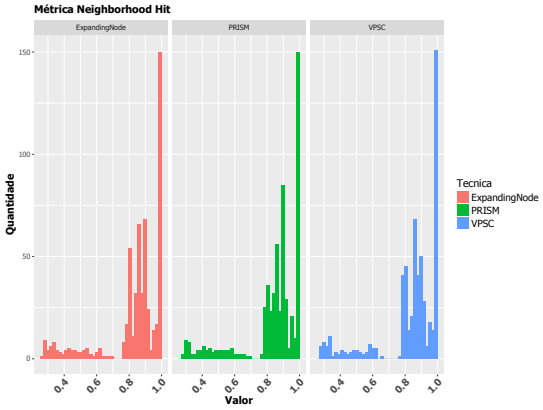
Assim como para as métricas *Neighborhood Hit* e *Neighborhood Preservation*, o resultado para o tempo começa com valores ruins e melhora conforme o número de agrupamentos aumenta. Isso fornece a ideia de que o algoritmo proposto neste trabalho é mais indicado para aplicação em pequenas partes das projeções, como acontece na abordagem de exploração proposta neste trabalho. Os resultados podem ser verificados nas Figuras 50a e 50b.



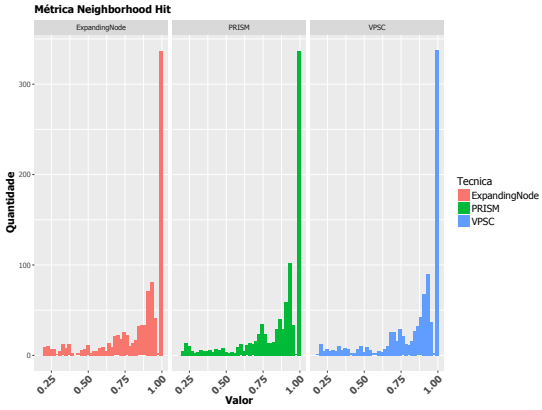
(a) NH do conjunto de dados Corel$_1$ com um agrupamento.



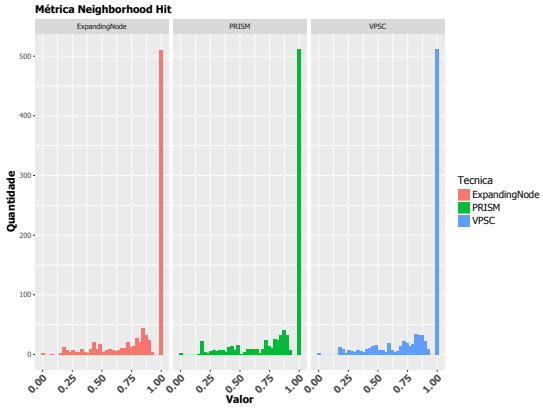
(b) NH do conjunto de dados Corel$_1$ com dez agrupamentos.



(c) NH do conjunto de dados Corel$_1$ com 20 agrupamentos.

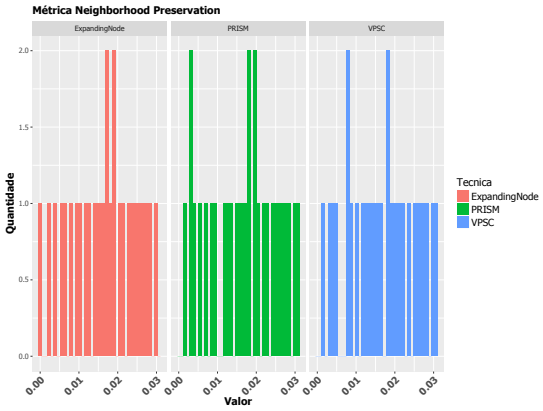


(d) NH do conjunto de dados Corel$_1$ com 50 agrupamentos.

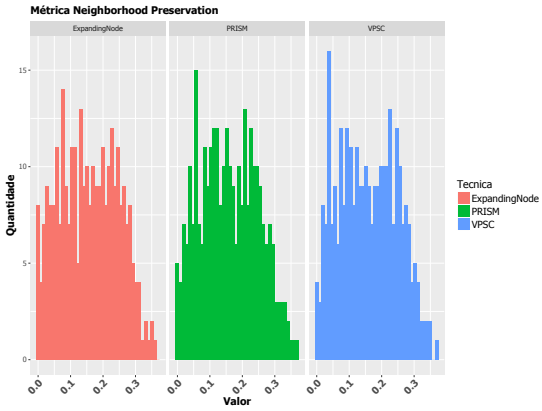


(e) NH do conjunto de dados Corel$_1$ com 100 agrupamentos.

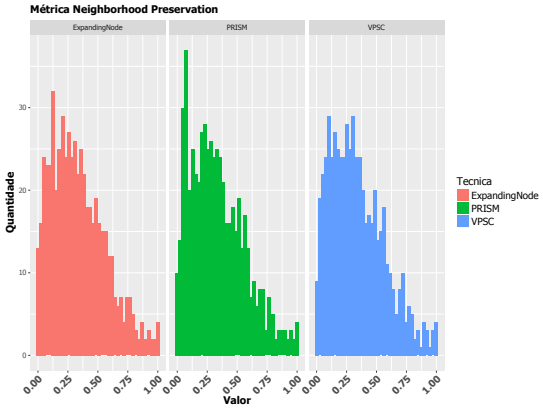
Figura 46 – *Neighborhood Hit* do conjunto de dados Corel$_1$.



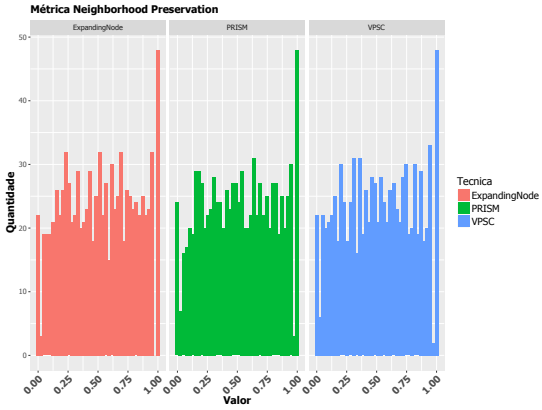
(a) NP do conjunto de dados Corel$_1$ com um agrupamento.



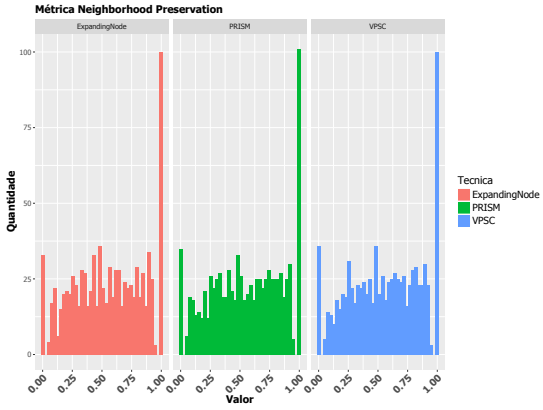
(b) NP do conjunto de dados Corel$_1$ com dez agrupamentos.



(c) NP do conjunto de dados Corel$_1$ com 20 agrupamentos.

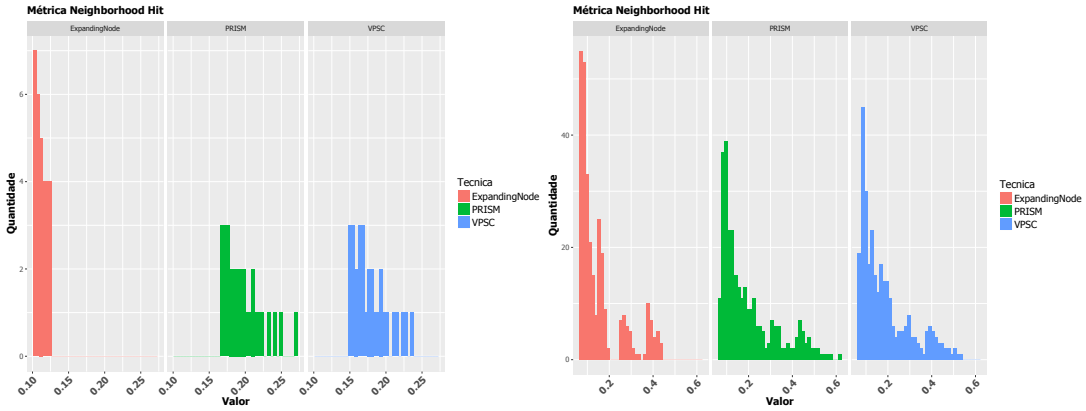


(d) NP do conjunto de dados Corel$_1$ com 50 agrupamentos.



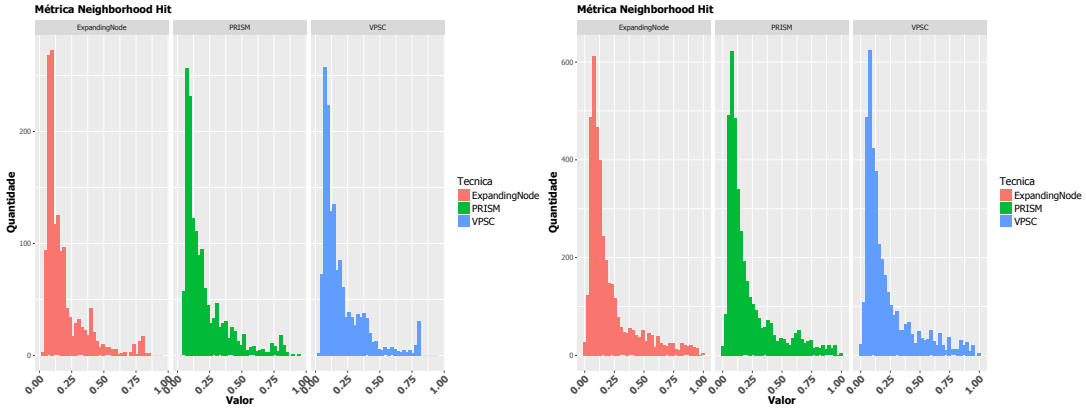
(e) NP do conjunto de dados Corel$_1$ com 100 agrupamentos.

Figura 47 – *Neighborhood Preservation* do conjunto de dados Corel$_1$.



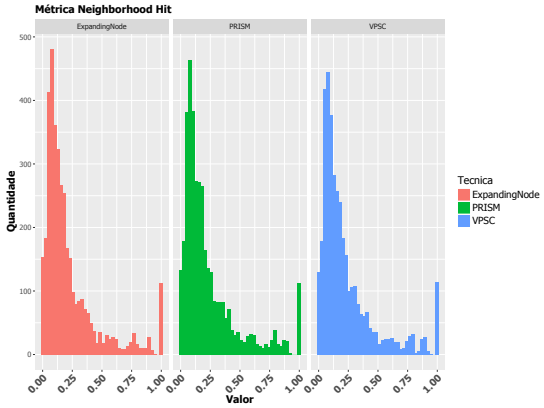
(a) NH do conjunto de dados Fotografos$_1$ com um agrupamento.

(b) NH do conjunto de dados Fotografos$_1$ com dez agrupamentos.



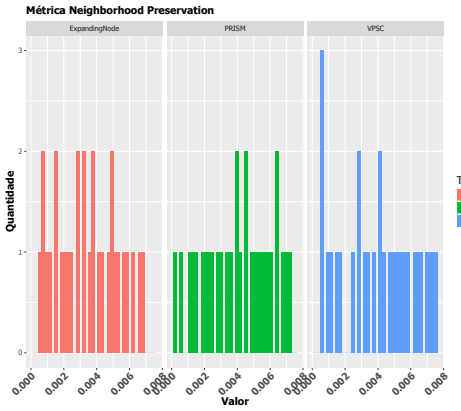
(c) NH do conjunto de dados Fotografos$_1$ com 50 agrupamentos.

(d) NH do conjunto de dados Fotografos$_1$ com 200 agrupamentos.

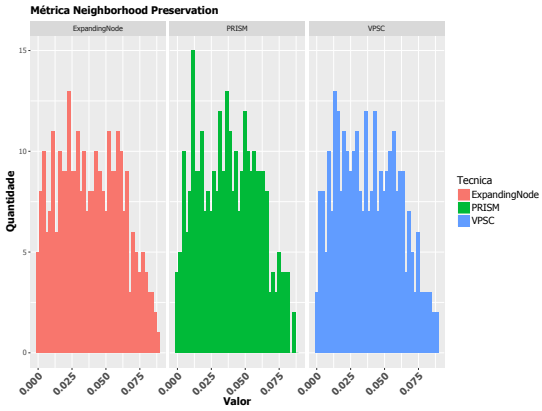


(e) NH do conjunto de dados Fotografos$_1$ com 300 agrupamentos.

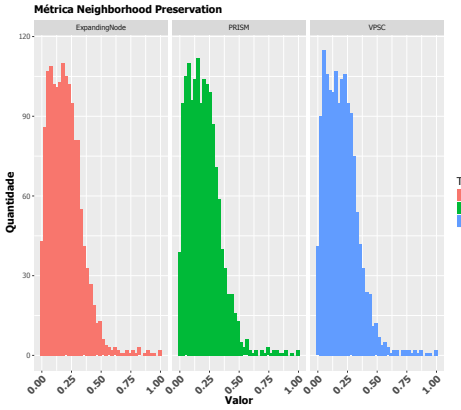
Figura 48 – *Neighborhood Hit* do conjunto de dados Fotografos$_1$.



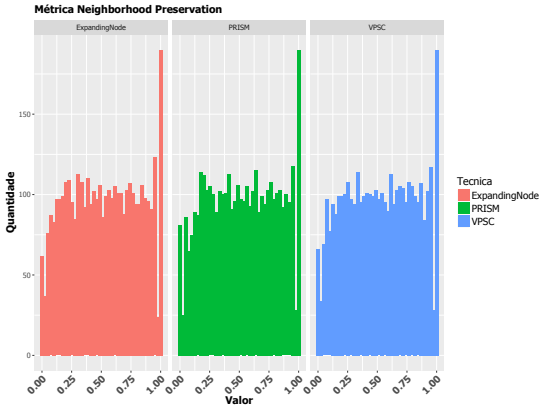
(a) NP do conjunto de dados Fotografos$_1$ com um agrupamento.



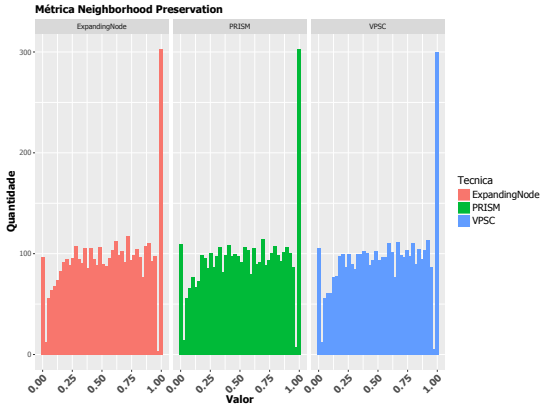
(b) NP do conjunto de dados Fotografos$_1$ com dez agrupamentos.



(c) NP do conjunto de dados Fotografos$_1$ com 50 agrupamentos.

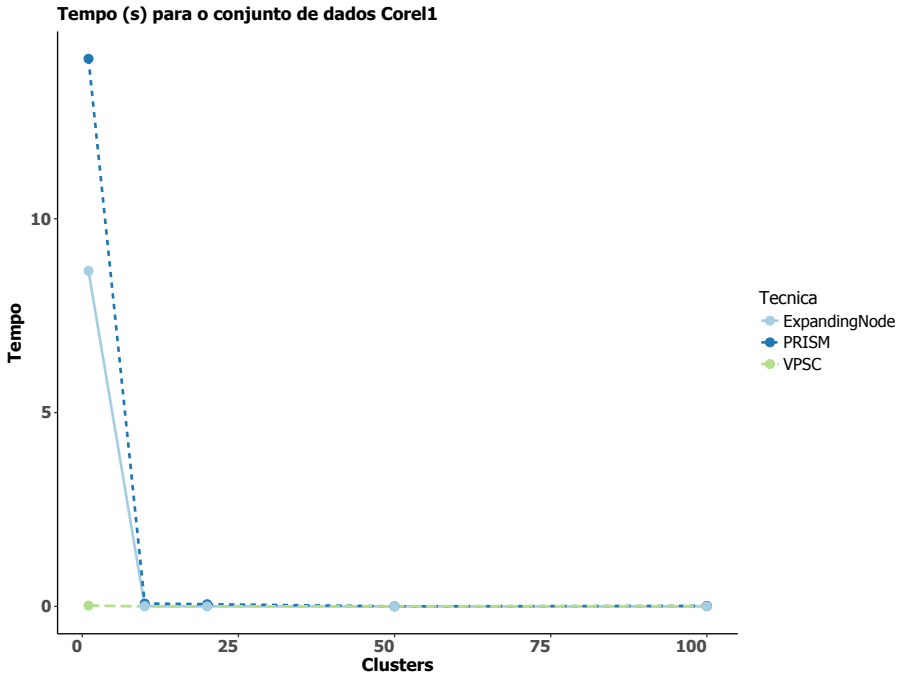


(d) NP do conjunto de dados Fotografos$_1$ com 200 agrupamentos.

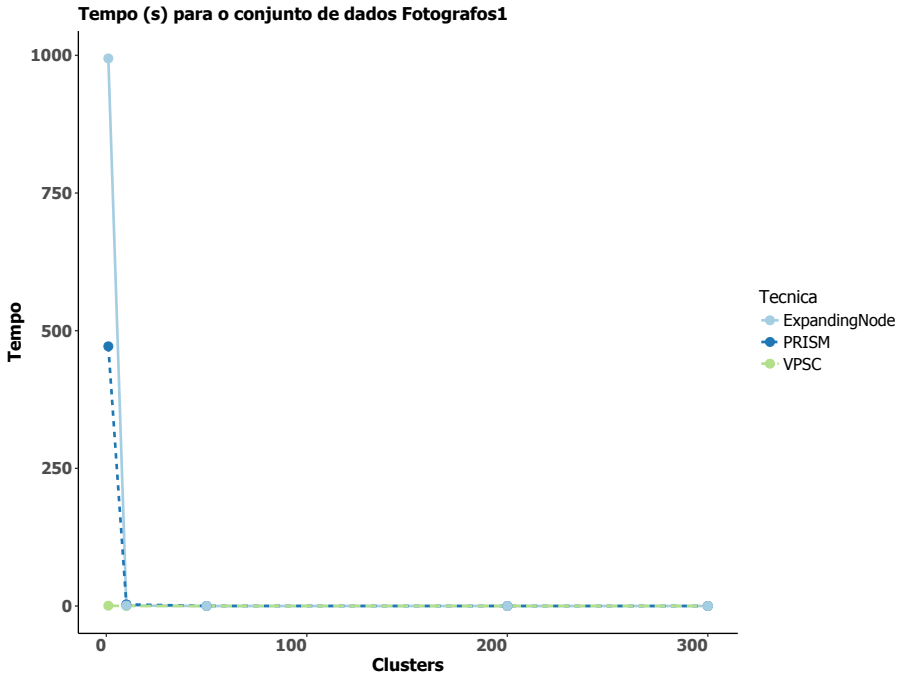


(e) NP do conjunto de dados Fotografos$_1$ com 300 agrupamentos.

Figura 49 – *Neighborhood Preservation* do conjunto de dados Fotografos$_1$.



(a) Tempo gasto pelos algoritmos para o conjunto de dados Corel_1.



(b) Tempo gasto pelos algoritmos para o conjunto de dados Fotografos_1.

Figura 50 – Tempo gasto pelos algoritmos para remover a sobreposição dos marcadores. Para cada quantidade de agrupamento, foi calculada a média do tempo gasto.

5.5 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentados experimentos acerca da abordagem de exploração desenvolvida. Foi possível verificar os aspectos relacionados a abordagem, como o desempenho dos algoritmos de seleção de representativos na divisão dos grupos de projeção, bem como a influência dos algoritmos de projeção e do espaço de características na criação da abordagem de exploração.

Como demonstração da utilização da abordagem de exploração, foram apresentados dois estudos de dados. O primeiro estudo de caso considera um conjunto de dados de imagens retirado da rede social *Instagram*, enquanto que o segundo estudo de caso utiliza um conjunto de dados de fotógrafos. No segundo estudo de caso foi possível perceber que a abordagem proposta nesse trabalho diminuir a desordem visual apresentada pela forma tradicional de apresentação dos dados de uma técnica de projeção multidimensional.

6 CONCLUSÕES E TRABALHOS FUTUROS

As projeções multidimensionais são uma ferramenta importante para análise de conjuntos de dados multidimensionais. A estratégia empregada para visualização dos dados trás benefícios porque se beneficia das características pré-atentivas para fornecer *insights* de forma rápida. No entanto, a dificuldade em analisar o conjunto é proporcional ao número de instâncias do conjunto de dados. Conforme o conjunto de dados cresce, a sobreposição dos marcadores é maior, por exemplo.

6.1 CONTRIBUIÇÕES E LIMITAÇÕES

Neste trabalho, apresentamos uma abordagem de exploração em visualizações geradas para representar projeções multidimensionais, com o objetivo de facilitar o processo exploratório e diminuir os problemas relacionados à escalabilidade visual presentes na metáfora visual tradicionalmente empregada para visualizar projeções multidimensionais, isto é, gráficos de dispersão. Para chegar nesse objetivo, dada uma projeção no plano, é criada uma árvore para controlar a hierarquia da abordagem utilizando técnicas de seleção de representativos. A seleção de representativos de forma recursiva nos fornece uma árvore para que o usuário possa interagir com a projeção de forma hierárquica e recebendo informações pertinentes acerca dos grupos, os quais são definidos por meio de diagramas rígidos de Voronoi. Por meio da análise da abordagem com diversas algoritmos de seleção de representativos, diferentes conjuntos de dados e duas técnicas de projeção multidimensional, ficou entendido que os algoritmos que apresentam melhores resultados são aqueles utilizados em processos de agrupamento devido à sua capacidade em retornar representativos com baixa redundância. Além disso, um ponto a ser ressaltado é que a utilização de técnicas de projeção multidimensional que separem as classes das instâncias eficientemente também contribuem para melhor desempenho da abordagem.

A abordagem proposta neste trabalho apresenta algumas vantagens em relações à trabalhos já propostos, como *InfoSky*, *HiPP* e h-SNE. Essas vantagens estão relacionadas à facilidade que pode ser aplicada à diferentes conjuntos de dados, como no caso da técnica *InfoSky* precisar de um conjunto de dados hierárquico; a remoção de sobreposição nos níveis mais baixos da hierarquia; ainda utilizar como base essencial a metáfora de gráficos de dispersão; permitir que diferentes algoritmos sejam aplicados de forma *plug & play*, isto é, diferentes algoritmos de remoção de sobreposição, técnicas de seleção de representativos e técnicas de projeção multidimensional; por fim, a abordagem proposta neste trabalho

permite que a interação não seja feita com incremento na sobrecarga visual. Uma limitação de nossa abordagem é a necessidade da utilização de técnicas de projeção que são eficientes na separação de classes, além disso, o espaço de características também tem influência na criação da abordagem de exploração. Esses aspectos devem ser considerados porque os grupos precisam estar bem definidos para que os usuários não sejam enganados durante o processo exploratório, isto é, instâncias não similares no mesmo grupo.

A principal contribuição deste trabalho é a abordagem de exploração multinível em visualizações geradas para representar projeções multidimensionais. A abordagem proposta permite que a exploração dos dados seja feita com o apoio de instâncias representativas, guiando o usuário no processo exploratório. Foi apresentado um *design* gráfico interativo para navegar em um conjunto de imagens como prova de conceito, baseado em divisão do espaço usando Voronoi. A abordagem foi desenvolvida de modo que seja adaptável quanto as técnicas de apoio que podem ser utilizadas – técnicas de seleção de representativos, técnicas de projeção multidimensional e técnicas de remoção de sobreposição – e, além disso, pode também ser empregada para visualização de documentos. A sobrecarga visual durante o processo exploratório é menor do que se comparada com a abordagem tradicional, visto que fornece informações aos poucos, de acordo com a demanda.

Como contribuições secundárias deste trabalho temos as análises realizadas dos algoritmos de seleção de representativos, assim como os algoritmos envolvidos no processo de criação da abordagem hierárquica. Não obstante, também temos com contribuição a técnica de remoção de sobreposição, que apesar da alta complexidade, pode ser utilizada para remoção de sobreposição em pequenos grupos de instâncias, considerando os resultados convincentes e a facilidade a qual pode ser implementada.

6.2 TRABALHOS FUTUROS

Como trabalhos futuros, pode-se incrementar a interação com o usuário. Como a definição do primeiro nível da hierarquia utilizada na abordagem de exploração é de grande importância para o restante das definições dos agrupamentos, o usuário pode realizar esse processo de identificação. Além disso, com o uso de *Active Learning* é possível melhorar o posicionamento das instâncias no plano de projeção, de modo que a separação de classes e instâncias forneça as ideias dos usuários.

Os mapas de calor podem ser utilizados para codificar as imagens que são selecionadas ou cujas *tags* foram definidas. Além disso, esse conceito pode ser utilizado para realizar anotações de conjunto de dados com maior facilidade.

Como última sugestão de trabalhos futuros, note que neste trabalho o foco foi dado a conjunto de dados de imagens. Todavia, a abordagem é facilmente generalizável para outros tipos de conjuntos de dados, como coleções de documentos. Assim, *Tag clouds*

poderiam ser utilizadas como texturas para fornecer uma visão geral do conjunto e a estrutura hierárquica poderia fornecer maneiras para extração de tópicos.

REFERÊNCIAS

AHARON, M.; ELAD, M.; BRUCKSTEIN, A. K -svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, v. 54, n. 11, p. 4311–4322, Nov 2006. ISSN 1053-587X. Citado 2 vezes nas páginas 8 e 9.

ANDREWS, D. F. Plots of high-dimensional data. *Biometrics*, v. 29, p. 125–136, 1972. Citado na página 1.

ANDREWS, K. et al. The infosky visual explorer: Exploiting hierarchical structure and document similarities. *Information Visualization*, Palgrave Macmillan, v. 1, n. 3/4, p. 166–181, dez. 2002. ISSN 1473-8716. Citado 5 vezes nas páginas ix, 2, 3, 17 e 18.

BALZER, M.; DEUSSEN, O. Voronoi treemaps. In: *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*. Washington, DC, USA: IEEE Computer Society, 2005. (INFOVIS '05), p. 7–. ISBN 0-7803-9464-x. Disponível em: <http://dx.doi.org/10.1109/INFOVIS.2005.40>. Citado na página 14.

BERTIN, J. *Semiology of Graphics*. [S.l.]: University of Wisconsin Press, 1983. ISBN 0299090604. Citado na página 5.

BOUTSIDIS, C.; MAHONEY, M. W.; DRINEAS, P. An improved approximation algorithm for the column subset selection problem. In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2009. (SODA '09), p. 968–977. Disponível em: <http://dl.acm.org/citation.cfm?id=1496770.1496875>. Citado na página 8.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. *Readings in information visualization: using vision to think*. [S.l.]: USA: Morgan Kaufmann Publishers, 1999. Citado na página 1.

CHEN, H. et al. Uncertainty-aware multidimensional ensemble data visualization and exploration. *IEEE Trans. Vis. Comput. Graph.*, v. 21, n. 9, p. 1072–1086, 2015. Citado na página 5.

CHáVEZ, E. et al. Searching in metric spaces. *ACM Computing Surveys*, v. 33, n. 3, p. 273–321, 2001. Citado na página 10.

CLARK, R. D. Optisim: An extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci.*, v. 37, p. 1181–1188, 1997. Citado na página 7.

CONG, Y.; YUAN, J.; LUO, J. Towards scalable summarization of consumer videos via sparse dictionary selection. *Trans. Multi.*, IEEE Press, Piscataway, NJ, USA, v. 14, n. 1, p. 66–75, fev. 2012. ISSN 1520-9210. Disponível em: <http://dx.doi.org/10.1109/TMM.2011.2166951>. Citado na página 7.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.*, IEEE Press, Piscataway, NJ, USA, v. 13, n. 1, p. 21–27, set. 2006. ISSN 0018-9448. Disponível em: <https://doi.org/10.1109/TIT.1967.1053964>. Citado na página 27.

CUI, Q. et al. Measuring data abstraction quality in multiresolution visualizations. *IEEE Transactions on Visualization and Computer Graphics*, v. 12, n. 5, p. 709–716, 2006. Citado 2 vezes nas páginas 43 e 44.

DUARTE, F. S. L. G. et al. Nmap: A novel neighborhood preservation space-filling algorithm. *IEEE Transactions on Visualization and Computer Graphics*, PP, n. 99, p. 1, August 2014. ISSN 1077-2626. Citado 3 vezes nas páginas ix, 14 e 15.

DWYER, T.; MARRIOTT, K.; STUCKEY, P. J. Fast node overlap removal. *Proceedings of the* 13^{th} *International Conference on Graph Drawing*, p. 153–164, 2006. Citado 3 vezes nas páginas 2, 12 e 27.

EICK, S. G.; KARR, A. F. Visual scalability. *Journal of Computational & Graphical Statistics*, v. 11, n. 1, p. 22–43, 2002. Citado na página 14.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: . [S.l.]: AAAI Press, 1996. p. 226–231. Citado na página 7.

FREY, B. J.; DUECK, D. Clustering by passing messages between data points. *Science*, v. 315, p. 2007, 2007. Citado na página 7.

FUA, Y.-H.; WARD, M. O.; RUNDENSTEINER, E. A. Hierarchical parallel coordinates for exploration of large datasets. In: *Proceedings of the Conference on Visualization '99: Celebrating Ten Years*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1999. (VIS '99), p. 43–50. ISBN 0-7803-5897-X. Disponível em: <http://dl.acm.org/citation.cfm?id=319351.319355>. Citado 3 vezes nas páginas ix, 18 e 19.

GANSNER, E. R.; HU, Y. Efficient, proximity-preserving node overlap removal. *Journal of Graph Algorithms and Applications*, v. 14, n. 1, p. 53–74, 2010. Citado 3 vezes nas páginas 2, 12 e 27.

GOMEZ-NIETO, E. et al. Similarity preserving snippet-based visualization of web search results. *TVCG*, v. 20, n. 3, p. 457–470, 2014. Citado 3 vezes nas páginas 2, 12 e 27.

GRACIA, A. et al. A methodology to compare Dimensionality Reduction algorithms in terms of quality loss. *Information Sciences*, 2014. Citado na página 5.

INSELBERG, A.; DIMSDALE, B. Parallel coordinates: A tool for visualizing multidimensional geometry. *IEEE Visualization*, v. 1, p. 361–378, 1990. Citado 2 vezes nas páginas 1 e 18.

JOHNSON, B.; SHNEIDERMAN, B. Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In: *Proceedings of the 2Nd Conference on Visualization '91*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1991. (VIS '91), p. 284–291. ISBN 0-8186-2245-8. Disponível em: <http://dl.acm.org/citation.cfm?id=949607.949654>. Citado na página 14.

JOIA, P. et al. Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics*, v. 17, n. 12, p. 2563–2571, 2011. Citado na página 1.

JOIA, P.; PETRONETTO, F.; NONATO, L. Uncovering representative groups in multidimensional projections. *CGF*, v. 34, n. 3, p. 281–290, 2015. Citado na página 9.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data : an introduction to cluster analysis*. New York: Wiley, 1990. (Wiley series in probability and mathematical statistics). A Wiley-Interscience publication. ISBN 0-471-87876-6. Disponível em: <http://opac.inria.fr/record=b1087461>. Citado na página 44.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Principles and Practice: Wiley-Interscience, 2005. Citado na página 7.

KENNARD, R. W.; STONE, L. A. Computer aided design of experiments. *Technometrics*, v. 11, p. 137–148, 1969. Citado na página 7.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F. et al. (Ed.). *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012. p. 1097–1105. Disponível em: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>. Citado 2 vezes nas páginas 33 e 39.

LEE, J. A.; VERLEYSEN, M. *Nonlinear Dimensionality Reduction*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2007. ISBN 0387393501, 9780387393506. Citado na página 5.

LIEBMAN, E.; CHOR, B.; STONE, P. Representative selection in nonmetric datasets. *Appl. Artif. Intell.*, Taylor & Francis, Inc., Bristol, PA, USA, v. 29, n. 8, p. 807–838, set. 2015. ISSN 0883-9514. Disponível em: <http://dx.doi.org/10.1080/08839514.2015.1071092>. Citado na página 7.

MA, B.; WEI, Q.; CHEN, G. A combined measure for representative information retrieval in enterprise information systems. *Journal of Enterprise Information Management*, v. 24, p. 310–321, nov 2011. Citado na página 11.

MAATEN, L. J. P. van der; HINTON, G. E. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, v. 9, p. 2579—2605, 2008. Citado 4 vezes nas páginas ix, 1, 2 e 6.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: CAM, L. M. L.; NEYMAN, J. (Ed.). *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.]: University of California Press, 1967. v. 1, p. 281–297. Citado na página 7.

MAIRAL, J. et al. Discriminative learned dictionaries for local image analysis. In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. [s.n.], 2008. Disponível em: <http://dx.doi.org/10.1109/CVPR.2008.4587652>. Citado na página 9.

MANSMANN, F.; VINNIK, S. Interactive exploration of data traffic with hierarchical network maps. *IEEE Transactions on Visualization and Computer Graphics*, v. 12, n. 6, p. 1440–1449, nov 2006. ISSN 1077-2626. Citado 3 vezes nas páginas ix, 19 e 20.

MEI, J.; CHEN, L. Document clustering around weighted-medoids. In: *8th International Conference on Information, Communications & Signal Processing, ICICS 2011, Singapore, Singapore, December 13-16, 2011*. [s.n.], 2011. p. 1–5. Disponível em: <http://dx.doi.org/10.1109/ICICS.2011.6173606>. Citado na página 7.

NONATO, L. G.; AUPETIT, M. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, p. 1, 2018. ISSN 1077-2626. Disponível em: <doi.ieeecomputersociety.org/10.1109/TVCG.2018.2846735>. Citado 2 vezes nas páginas 1 e 5.

PAN, C.-T. On the existence and computation of rank-revealing LU factorizations. v. 316, n. 1–3, p. 199–222, set. 2000. ISSN 0024-3795 (print), 1873-1856 (electronic). Disponível em: <http://www.elsevier.nl/gej-ng/10/30/19/134/24/37/abstract.html;http://www.elsevier.nl/gej-ng/10/30/19/134/24/37/article.pdf>. Citado na página 8.

PAN, C.-T.; TANG, P. T. P. Bounds on singular values revealed by qr factorizations. *BIT Numerical Mathematics*, v. 39, n. 4, p. 740–756, 1999. ISSN 1572-9125. Disponível em: <http://dx.doi.org/10.1023/A:1022395308695>. Citado na página 8.

PAN, F. et al. Finding representative set from massive data. In: *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), 27-30 November 2005, Houston, Texas, USA*. [s.n.], 2005. p. 338–345. Disponível em: <http://dx.doi.org/10.1109/ICDM.2005.69>. Citado na página 6.

PAULOVICH, F. V.; MINGHIM, R. Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE Transactions on Visulization and Computer Graphics*, v. 14, n. 6, p. 1229–1236, 2008. Citado 6 vezes nas páginas ix, 2, 3, 15, 16 e 53.

PAULOVICH, F. V. et al. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visulization and Computer Graphics*, v. 3, p. 564–575, 2008. Citado 3 vezes nas páginas 1, 18 e 53.

PAULOVICH, F. V.; SILVA, C. T.; NONATO, L. G. Two-phase mapping for projecting massive data sets. *IEEE Transactions on Visualization and Computer Graphics*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 16, n. 6, p. 1281–1290, nov. 2010. ISSN 1077-2626. Disponível em: <https://doi.org/10.1109/TVCG.2010.207>. Citado 2 vezes nas páginas 1 e 5.

PEDREIRA, O.; BRISABOA, N. R. Spatial selection of sparse pivots for similarity search in metric spaces. *Proceedings of the 33rd Conference on Current Trends in Theory and Practice of Computer Science*, p. 434–445, 2007. Citado na página 10.

PEZZOTTI, N. et al. Hierarchical stochastic neighbor embedding. In: *Proceedings of the Eurographics / IEEE VGTC Conference on Visualization*. Goslar Germany, Germany: Eurographics Association, 2016. (EuroVis '16), p. 21–30. Disponível em: <https://doi.org/10.1111/cgf.12878>. Citado 5 vezes nas páginas ix, 2, 3, 21 e 22.

POCO, J. et al. A framework for exploring multidimensional data with 3d projections. In: *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*. Chichester, UK: The Eurographs Association & John Wiley & Sons, Ltd., 2011. (EuroVis'11), p. 1111–1120. Disponível em: <http://dx.doi.org/10.1111/j.1467-8659.2011.01960.x>. Citado 3 vezes nas páginas ix, 18 e 19.

RAMIREZ, I.; SPRECHMANN, P.; SAPIRO, G. Classification and clustering via dictionary learning with structured incoherence and shared features. *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Los Alamitos CA USA, v. 00, p. 3501–3508, 2010. Citado na página 9.

ROBERTSON, G. G.; MACKINLAY, J. D.; CARD, S. K. Cone trees: Animated 3d visualizations of hierarchical information. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1991. (CHI '91), p. 189–194. ISBN 0-89791-383-3. Disponível em: <http://doi.acm.org/10.1145/108844.108883>. Citado na página 14.

SANTOS, L. F. D. *Explorando variedade em consultas por similaridade*. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2012. Citado na página 27.

SEO, J.; SHNEIDERMAN, B. Interactively exploring hierarchical clustering results. *Computer*, IEEE Computer Society Press, Los Alamitos, CA, USA, v. 35, n. 7, p. 80–86, jul. 2002. ISSN 0018-9162. Disponível em: <http://dx.doi.org/10.1109/MC.2002.1016905>. Citado na página 14.

SHNEIDERMAN, B. The eyes have it: A task by data type taxonomy for information visualizations. In: *Proceedings of the 1996 IEEE Symposium on Visual Languages*. Washington, DC, USA: IEEE Computer Society, 1996. (VL '96), p. 336–. ISBN 0-8186-7508-X. Disponível em: <http://dl.acm.org/citation.cfm?id=832277.834354>. Citado 4 vezes nas páginas 2, 14, 21 e 22.

SILVA, R. R. O. da. *Visualizing Multidimensional Data Similarities: Improviments and Applications*. Tese (Doutorado) — University of Groningen, 10 2016. Citado 3 vezes nas páginas ix, 16 e 17.

STROBELT, M. et al. Rolled-out wordles: A heuristic method for overlap removal of 2d data representatives. *Computer Graphics Forum*, p. 1135–1144, 2012. Citado 2 vezes nas páginas 12 e 27.

TEJADA, E.; MINGHIM, R.; NONATO, L. G. On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization*, v. 2, n. 4, p. 218–231, 2003. Citado na página 1.

WANG, G. et al. Topic hypergraph: hierarchical visualization of thematic structures in long documents. *Science China Information Sciences*, v. 56, n. 5, p. 1–14, May 2013. Citado na página 20.

WANG, H. et al. Representative selection with structured sparsity. *Pattern Recognition*, v. 63, p. 268–278, 2017. Disponível em: <http://dx.doi.org/10.1016/j.patcog.2016.10.014>. Citado na página 7.

WANG, W. et al. Visualization of large hierarchical data by circle packing. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2006. (CHI '06), p. 517–520. ISBN 1-59593-372-7. Disponível em: <http://doi.acm.org/10.1145/1124772.1124851>. Citado na página 14.

WANG, Y. et al. Representative selection based on sparse modeling. *Neurocomput.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 139, p. 423–431, set. 2014. ISSN 0925-2312. Disponível em: <http://dx.doi.org/10.1016/j.neucom.2014.02.013>. Citado na página 9.

WARD, M. O. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, Palgrave Macmillan, v. 1, n. 3/4, p. 194–210, dez. 2002. ISSN 1473-8716. Disponível em: <http://dx.doi.org/10.1057/palgrave.ivs.9500025>. Citado na página 14.

WARE, C. *Information Visualization: Perception for Design*. 3. ed. Amsterdam: Morgan Kaufmann, 2012. (Morgan Kaufmann Series in Interactive Technologies). ISBN 978-0-12-381464-7. Disponível em: <http://www.sciencedirect.com/science/book/9780123814647>. Citado na página 6.

ZHANG, J.; WEI, Q.; CHEN, G. Finding an λ-representative subset from massive data. In: *Joint IFSA World Congress and NAFIPS Annual Meeting, IFSA/NAFIPS, 2013, Edmonton, Alberta, Canada, June 24-28, 2013*. [s.n.], 2013. p. 585–590. Disponível em: <http://dx.doi.org/10.1109/IFSA-NAFIPS.2013.6608466>. Citado na página 7.