UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

André Luis Dias Andreotti

Abordagens interativas para exploração de coleções de documentos

Presidente Prudente
2020

André Luis Dias Andreotti

Abordagens interativas para exploração de coleções de documentos

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, da Faculdade de Ciências e Tecnologia da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de Presidente Prudente.

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

Presidente Prudente
2020

A559a

Andreotti, André Luis Dias
Abordagens interativas para exploração de coleções de documentos / André Luis Dias Andreotti. -- Presidente Prudente, 2020
77 f. : il., tabs. + 1 CD-ROM

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Danilo Medeiros Eler

1. Visualização. 2. Recuperação de textos. 3. Tag clouds. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

André Luis Dias Andreotti

Abordagens interativas para exploração de coleções de documentos

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, da Faculdade de Ciências e Tecnologia da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de Presidente Prudente.

Comissão Examinadora

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

Profa. Dra. Aretha Barbosa Alencar
UTFPR – Campo Mourão – PR

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

Presidente Prudente
7 de fevereiro de 2020

Aos meus pais, Edilene e Sérgio

AGRADECIMENTOS

Agradeço aos meus pais Edilene e Sérgio pelo incentivo que sempre me deram nos estudos, sem vocês este trabalho não existiria.

Agradeço a minha namorada Fabiana por todo empurrão dia após dia, por sempre estar do meu lado, me apoiando nas noites e finais de semana de estudos e sempre me incentivando para concluir o mestrado. Você foi fundamental meu amor, sua participação ultrapassa as fronteiras da ciência e tecnologia.

Agradeço ao meu orientador Danilo por toda a colaboração, não somente neste trabalho, mas nos demais trabalhos desenvolvidos durante o mestrado e graduação, incluindo o trabalho de conclusão de curso. Agradeço por toda a confiança e indicações durante todos esses anos.

Agradeço a todos os professores que foram a base de todo o conhecimento adquirido durante a minha formação. Vocês são e sempre serão os responsáveis por toda a formação profissional e humana.

Agradeço a todos os amigos que estiveram juntos nas aulas, presenciais e à distância, realizando trabalhos e trocando conhecimentos que são sempre essenciais para a evolução da ciência. Em especial ao amigo Fabrício Negri, aluno deste programa de mestrado, que por forças maiores já não está mais entre nós, sempre planejávamos esse momento final, agradeço aqui seu apoio neste trabalho.

Agradeço a Petrobras, pelo incentivo e financiamento do Projeto de Pesquisa e Desenvolvimento, desenvolvido em paralelo a este trabalho. Em especial ao professor Ivan, coordenador do projeto, do Centro de Geociências aplicadas ao Petróleo – UNESPetro, pela confiança e trabalho colaborativo mesmo à distância.

Agradeço pelo profissionalismo dos representantes e amigos da VCOM, FalaFreud, Unesp e Vunesp, empresas que trabalhei durante o período do mestrado, que sempre permitiram, compreenderam e incentivaram meus compromissos com os estudos.

"Stay hungry. Stay foolish" – Jobs (2005)

RESUMO

Os dados textuais têm desempenhado um papel cada vez mais importante em várias tarefas analíticas em pesquisas acadêmicas, inteligência de negócios, monitoramento de mídias sociais, jornalismo e outras áreas. A fim de explorar e dar sentido a esses dados, várias técnicas de visualização de textos surgiram nos últimos anos. Técnicas de visualizações, neste contexto, visam permitir que usuários possam explorar as relações entre documentos, descobrir documentos de interesse ou analisar padrões contidos nos documentos. Nessa forma de análise, um documento é comumente modelado por seu conteúdo subjacente e pelo conjunto de palavras que o compõe. Geralmente, técnicas de projeção multidimensional são empregadas para projetar esses documentos no espaço $2D$, porém o usuário precisa ler cada documento para entender a geração dos agrupamentos e também existem problemas quanto a sobreposição de marcadores quando o número de documentos cresce. Neste trabalho são apresentadas duas propostas de elaboração de abordagens para exploração de coleções de documentos, em que o objetivo é reduzir o esforço cognitivo necessário para explorar o conjunto de documentos comparado com representações comuns de projeções multidimensionais. A primeira é uma abordagem híbrida, que mostra o relacionamento e o conteúdo do documento em uma única visualização, utilizando "mapas de documentos" e *tag clouds*. A segunda é uma abordagem hierárquica, que utiliza *tag clouds* para preencher as texturas de agrupamentos formados por meio do domínio de Voronoi para codificar visualmente as fronteiras dos grupos de documentos. Mostramos a eficácia das abordagens propostas na exploração de coleções de documentos, fornecendo explorações em que o usuário recebe poucas informações durante o processo exploratório e detalha o conteúdo de acordo com a demanda, superando problemas de identificação de agrupamentos e sobreposição de marcadores.

Palavras-chave: visualização. recuperação de textos. tag clouds.

ABSTRACT

Textual data has played an increasingly important role in various analytical tasks in academic research, business intelligence, social media monitoring, journalism, and other fields. In order to explore and make sense of this data, various text visualization techniques have emerged in recent years. Visualization techniques, in this context, are intended to enable users to explore relationships between documents, discover documents of interest, or analyze patterns contained in documents. In this form of analysis, a document is commonly shaped by its underlying content and the set of words that compose it. Multidimensional projection techniques are often employed to project these documents into the $2D$ space, but you need to read each document to understand the generation of collations, and there are also problems with overlapping markers as the number of documents grows. This paper presents two proposals for developing approaches for exploring document collections, in which the objective is to reduce the cognitive effort required to explore the set of documents compared to common representations of multidimensional projections. The first is a hybrid approach, which shows the relationship and content of the document in a single view using "document maps" and *tag clouds*. The second is a hierarchical approach that uses *tag clouds* to fill in the textures of groupings formed through the Voronoi domain to visually encode the boundaries of document groups. We show the effectiveness of the proposed approaches in exploring document collections by providing explorations where the user receives little information during the exploratory process and details the content according to demand, overcoming cluster identification and marker overlap issues.

Keywords: visualization. text retrieval. tag clouds.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAS

Com intuito de fornecer entendimento acerca de um conjunto de dados, a visualização de informação utiliza representações gráficas e disponibiliza modelos para guiar os usuários no processo de análise e descoberta de conhecimento. Técnicas de visualização são elaboradas levando em consideração as características dos dados a serem analisados. Conjuntos de dados textuais, por exemplo, são desafiadores, pois, ao contrário de outros tipos de dados, é necessário realizar um pré-processamento visando extrair características observando as frequências de ocorrência de palavras (GRIES, 2009). E mesmo assim, ao colocar em uma ordem diferente ou em um contexto diferente, as palavras podem assumir um significado distinto (SIIRTOLA et al., 2014). Desta forma, a exploração de dados textuais é bastante explorada e diversas técnicas são desenvolvidas.

Técnicas de visualização de texto variam em como elas pré-processam e representam as informações textuais. Muitas técnicas adotam o padrão de representação *"bag-of-words"* para recuperação de informações, que modela o conteúdo do texto como um conjunto de palavras, cada uma com uma contagem de frequência associada. Para poucos documentos e tarefas simples, essa representação vetorial elementar é suficiente para criar visualizações atraentes. Esse padrão também é adotado em muitas técnicas que exibem coleções de documentos, pois permite deduzir a dissimilaridade do documento com base na comparação de frequências de palavras. Outras técnicas extraem tópicos ou outras características com significado semântico, o que normalmente requer um pré-processamento mais elaborado e computacionalmente custoso (ALENCAR; FERREIRA; PAULOVICH, 2012). Um exemplo muito utilizado para visualizar as informações textuais é exibido na Figura 1, uma técnica que utiliza *tag clouds* para representar tais frequêcias.

A partir dos métodos existentes para criar representações de similaridade entre documentos, as técnicas de projeção multidimensional, como *IDMAP* (MINGHIM; PAULOVICH; LOPES, 2006), *LSP* (PAULOVICH et al., 2008), *HiPP* (PAULOVICH et al., 2008) e *t-SNE* (MAATEN; HINTON, 2008), são possivelmente as mais comuns. As visualizações geradas por técnicas de projeção representam documentos por meio de gráficos de dispersão, quando o espaço é reduzido para o bidimensional. Geralmente, os documentos são representados por marcadores e seu posicionamento reflete a similaridade de conteúdo de seus documentos correspondentes. Um exemplo de representação gerada por uma projeção é apresentado na Figura 2, que ilustra as relações de similaridade de uma coleção de documentos de três áreas distintas, cada uma indicada por uma cor. Para gerar esse mapa,

Figura 1 – Exemplo de tag cloud gerada a partir de publicações de docentes no ano de 2016 da Faculdade de Ciências e Tecnologia, Unesp.

Fonte: Andreotti et al. (2017).

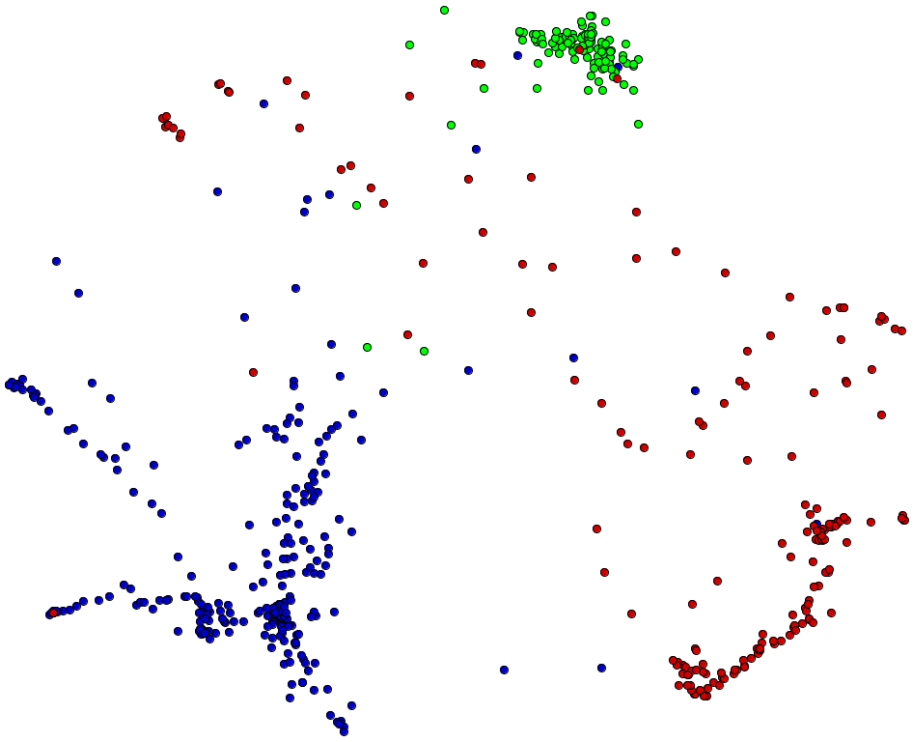
as técnicas de projeção geralmente tomam como entrada um modelo de espaço vetorial baseado em frequência da coleção, ou um vetor descrevendo tópicos ou outros recursos extraídos. Alternativamente, algumas técnicas requerem apenas uma matriz descrevendo as diferenças ou distâncias entre todos os pares de documentos (ALENCAR; FERREIRA; PAULOVICH, 2012).

1.2 OBJETIVOS E CONTRIBUIÇÕES

Uma das aplicações que tem despertado interesse na comunidade científica é a exploração de coleções de documentos. A utilização de projeções multidimensionais pode ser bastante efetiva nesse processo, resultando no que é conhecido como "mapa de documento", como aquele exemplificado na Figura 2, sendo um espaço de informação visual do conjunto de dados que permite a navegação do usuário e representa espacialmente uma ou mais propriedades dos textos. Desta forma facilitam o acesso a coleções de documentos e melhoram a efetividade de resolver problemas reais de tarefas de gerenciamento de conhecimento (BECKS; SEELING; MINKENBERG, 2002).

Por mais que existam inúmeras técnicas de projeção multidimensionais presentes na literatura, que processam ou podem ser adaptadas para processar dados textuais, ainda encontramos dificuldades na escalabilidade visual conforme o número de instâncias em um

Figura 2 – Exemplo de mapa de documentos de uma coleção de artigos científicos, obtido com técnicas de projeção multidimensional.



Fonte: Elaborado pelo autor.

conjunto de dados cresce, exigindo um maior esforço para análise dos dados. Por exemplo, 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 (MARCILIO; ELER, 2018). A literatura apresenta algumas técnicas para solucionar o problema da sobreposição de marcadores (STROBELT et al., 2012a; DWYER; MARRIOTT; STUCKEY, 2006a; GANSNER; HU, 2010a; GOMEZ-NIETO et al., 2014), mas representação gráfica sem a sobreposição necessita de uma espaço muito maior para ser visualizado, o que exige do usuário uma navegação sobre a representação gráfica e também pode gerar a perda da visão geral da representação, conforme a navegação é executada.

Este trabalho tem por objetivo apoiar a exploração de coleções de documentos por meio do aprimoramento de técnicas de projeções multidimensionais e facilitar a exploração do resultado mediante o desenvolvimento de abordagens interativas. Uma das abordagens desenvolvidas foi a de visualização híbrida para mapear as semelhanças dos documentos no espaço $2D$ e exibir *tag clouds* para cada documento, apresentando os principais termos dos dados textuais. A visualização de semelhanças e palavras chave em uma única exibição pode melhorar a exploração de dados textuais e auxiliar no entendimento da formação de

grupos em uma projeção multidimensional. A outra abordagem desenvolvida conta com uma exploração hierárquica sobre o conjunto textual analisado, utilizando *tag clouds* para auxiliar na sumarização e evitar sobreposição enquanto mantém as relações de similaridade.

Uma das principais contribuições deste trabalho é auxiliar na exploração de conjuntos de dados textuais com base na abordagem de visualização híbrida proposta, que mapeia as semelhanças e o conteúdo do texto em uma visualização exclusiva. Essa abordagem usa todos os termos do conjunto de dados para gerar a representação visual e mostrar algumas palavras chave do texto, assim, o usuário pode entender os principais tópicos de grupos distintos. Além disso, a visualização individual da *tag cloud* pode ajudar o usuário a entender a formação de grupos de instâncias semelhantes e melhorar a detecção do limite entre grupos distintos. Também introduzimos alguns mecanismos de interação para melhorar a experiência do usuário durante o processo exploratório. Outra contribuição foi o uso de uma abordagem multinível para suporte a dados textuais, acrescentada a metáfora visual com a utilização de *tag clouds* projetadas como texturas dos agrupamentos formados por meio do domínio de Voronoi.

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 hibrida desenvolvida. No Capítulo 5 é apresentada a abordagem hierarquica desenvolvida. Por fim, as conclusões e trabalhos futuros são apresentados no Capítulo 6.

2 FUNDAMENTAÇÃO

Com a diminuição do custo e a melhoria das tecnologias para armazenamento, distribuição e recuperação de dados, a quantidade de informação produzida tem crescido muito tanto em volume quanto em quantidade. Deste modo, técnicas para apoiar a interpretação de tais informações se tornam necessárias. A visualização de informação é uma área que oferece suporte a interpretação desses dados por meio da elaboração de métodos visuais de apresentação e interação com dados abstratos (PAULOVICH, 2008). A mineração de dados busca extrair padrões úteis de conjuntos de dados ou criar modelos de previsão (TAN; STEINBACH; KUMAR, 2005), usualmente como parte de um processo mais genérico de extração de conhecimento denominado *Knowledge Discovery in Databases (KDD)* (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). As duas técnicas tem objetivos similares e são complementares, a fusão das duas em uma área onde a mineração e a visualização co-existem é denominada *Mineração Visual de Dados* (WONG, 1999).

Ao trabalhar com dados textuais a etapa de pré-processamento desse conteúdo é essencial para criação de representações visuais. Uma vez que o texto livre não traga nenhum metadado acoplado ao seu conteúdo, é necessário criar modelos que possam representá-los matematicamente, uma forma de fazer isso é utilizando o padrão de representação "bag-of-words" que modela o conteúdo do texto como um conjunto de palavras, cada uma com uma contagem de frequência associada.

Por mais promissor que seja o desenvolvimento de abordagens de mineração visual de dados para apoiar tarefas textuais, com o aumento substancial de tamanho e complexidade dos conjuntos de dados atualmente disponíveis, a extração de informação relevante desses ainda permanece um desafio. Uma técnica de projeção multidimensional mapeia as instâncias de dados em um espaço uni-, bi- ou tri-dimensional, preservando alguma informação sobre as relações de distância ou similaridade entre elas de forma a revelar o máximo possível de estruturas existentes (PAULOVICH, 2008). Desta forma, aplicar essa técnica para exibir agrupamentos de coleções de documentos, sendo esses dados multi-dimensionais, mostra-se bastante efetivo, gerando "mapa de documentos".

Serão detalhados nas seções a seguir cada um dos termos apresentados acima, com intuito de utilizá-los no desenvolvimento do projeto proposto.

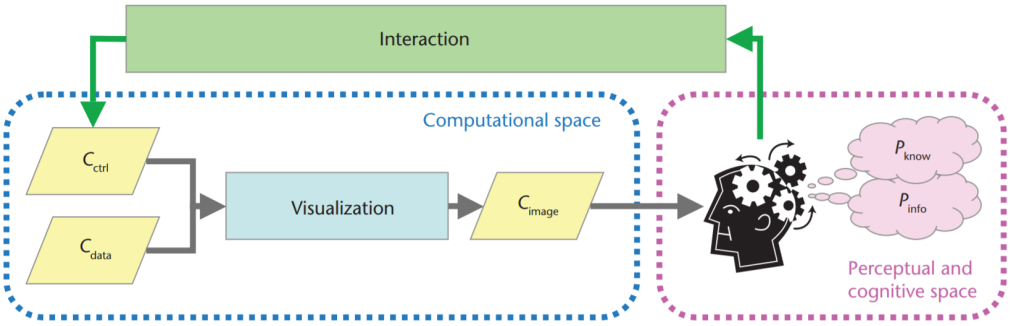
2.1 VISUALIZAÇÃO DE INFORMAÇÃO

Na visualização usamos os termos dados, informação e conhecimento, muitas vezes em um contexto inter-relacionado. O objetivo principal na visualização de dados é obter conhecimento sobre um espaço de informações (CHEN et al., 2009). A visualização de informação pretende, por meio de técnicas de mineração e visualização, apresentar um aglomerado de informações de forma simples e amigável facilitando a interpretação por parte do usuário. Os humanos tem notáveis habilidades de percepção que podem ser exploradas por técnicas de visualização, eles podem detectar alterações no tamanho, cor, forma, movimento ou textura, podem apontar para um único pixel, mesmo em uma tela com milhões de pixels. O mantra visual de busca de informações "*Overview first, zoom and filter, then details-on-demand*" de Shneiderman (1996) é muito utilizado nas visualizações modernas, visando principalmente o processamento em tempo real, uma vez que um grande conjunto de dados exige um alto poder de processamento, e realizar esse processamento apenas em uma porção de dados que o usuário deseja obter mais detalhes, no momento em que deseja, se torna mais viável.

A Figura 3 ilustra o processo definido porpor Chen et al. (2009). Dividindo o processo de visualização de informação entre o espaço computacional "*Computational space*", espaço perceptivo e cognitivo "*Perceptual and cognitive space*" e a interação "*Interation*" realizada após o usuário filtrar os dados ou mudar os parâmetros de processamento. Desta forma, a necessidade de visualização é baseada nas dificuldades que os humanos enfrentam ao adquirir uma quantidade suficiente de informações ou conhecimento diretamente de um conjunto de dados. O processo de criação de visualização transforma um conjunto de dados em uma representação visual, o que facilita um processo cognitivo mais eficiente e eficaz para aquisição de informação e conhecimento (CHEN et al., 2009).

O processo de visualização pode ser comparado ao processo típico de pesquisa, exceto que geralmente é muito mais complexo do que colocar algumas palavras-chave em um mecanismo de pesquisa. Na visualização as ferramentas para as tarefas de "pesquisa" são geralmente específicas da aplicação, o espaço de parâmetro para a "busca" é normalmente enorme, e a resposta para a interação do usuário um pouco mais lenta, especialmente no manuseio de conjunto de dados muito grandes, devido a quantidade de parâmetros que a aplicação deve processar (CHEN et al., 2009). Entretanto, o desenvolvimento da visualização de informação vem seguindo um caminho semelhante a outras tecnologias de computação, como processamento de fala, visão computacional e tecnologia *Web*, sendo desenvolvidos sistemas de visualização interativa, visualização assistida por informação e visualização assistida por conhecimento.

Figura 3 – Processo de visualização ilustrado por Chen et al. (2009). C_{data} representam os dados de entrada e C_{ctrl} os parâmetros de controle, ambos passam por uma técnica de visualização "*Visualization*" e então os resultados de visualização, C_{image}, são exibidos para o usuário e armazenados na memória do computador. P_{info} representam as informações adquiridas pelo usuário e P_{know} o conhecimento



Fonte: Chen et al. (2009).

2.2 MINERAÇÃO E PRÉ-PROCESSAMENTO DE TEXTO

Grande parte do esforço total de um projeto de visualização ou mineração é gasto coletando os dados e realizando seu pré-processamento. A qualidade do resultado final depende da qualidade do pré-processamento realizado (ALENCAR, 2013). Ao trabalhar com dados textuais um pré-processamento deve ser realizado sobre o texto livre com objetivo de filtrar as palavras relevantes e criar uma estrutura de dados para os algoritmos de processamento.

Como atividade comum na etapa de pré-processamento textual temos a tokenização do texto, que consiste em representar cada palavra em unidades distintas e significativas (os *tokens*). Levando em consideração um documento perfeitamente pontuado, a tokenização se torna uma tarefa simples, identificando espaços em brancos e pontuações como pontos de separação de *tokens*. Um exemplo é apresentado a seguir:

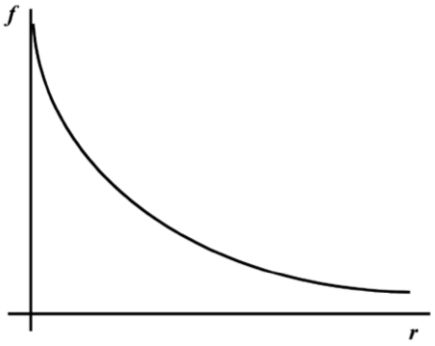
> **Entrada:** Lembre-se de olhar para o alto, para as estrelas, e não para baixo, para os seus pés.
>
> **Saída:** Lembre-se | de | olhar | para | o | alto | para | as | estrelas | e | não | para | baixo | para | os | seus | pés

Outro processamento comum é a remoção de palavras que são consideradas irrelevantes para a linguagem (*stopwords*), como: preposições, conjunções, advérbios, adjetivos e artigos (ALENCAR, 2013). Estes termos podem ser previamente estabelecidos em uma lista de *stopwords* comuns da linguagem do documento, ou utilizando bibliotecas que

conseguem identificar tais termos. Também pode ser aplicado algoritmos de *stemming*, onde as variantes de um termo são representadas por meio do radical, por exemplo reduzindo as palavras: conhecer, conhecimento, conhecendo e conhecido, para seu radical "conhe", reduzindo o tamanho do vocabulário e unificando os termos reduzidos para uma futura contagem de frequência.

Zipf (1949) realizou uma descoberta sobre um padrão existente nas frequências dos termos em uma sequência de texto, onde a frequência com que se usa uma determinada palavra é proporcional a 1 dividido pela classificação de frequência do termo entre a quantidade total. Isso significa que normalmente em uma sequência de texto o segundo item ocorre com mais ou menos a metade da frequência do primeiro, e o terceiro item com $\frac{1}{3}$ da frequência do primeiro, e assim sucessivamente. Tal constatação é agora chamada de "Lei de Zipf". Considerando um histograma de frequência destes termos em ordem decrescente obtém-se a "Curva de Zipf", a Figura 4 exibe uma representação em escala logarítmica dessa curva.

Figura 4 – Curva de Zipf. "r": termos, "f": frequência.

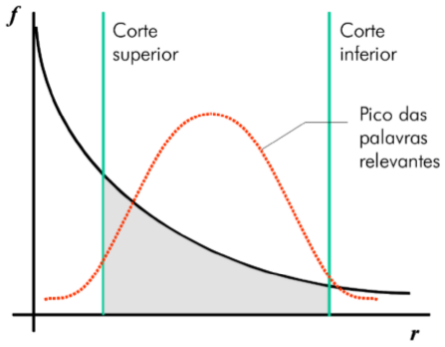


Fonte: Zipf (1949).

Sendo assim definida a distribuição dos termos, anos depois Luhn (1958) usou essa curva para especificar um limiar superior e inferior para remoção de termos poucos significativos. Os termos cuja frequência é maior que o "corte superior" são considerados poucos representativos e são removidos, também os termos cuja frequência é menor que o "corte inferior" são considerados irrelevantes para representação do texto analisado, uma vez que ocorrem raramente, e também são removidos. A Figura 5 ilustra os cortes de Luhn (1958) sobre a curva de Zipf (1949).

Após todo o processo de tokenização, seleção de palavras representativas e remoção de palavras insignificantes, é necessário a definição de uma estrutura para representar todos estes termos para que algoritmos possam realizar a mineração, recuperação e visualização desse texto. Salton, Wong e Yang (1975) definiu um modelo vetorial para representar a frequência dos termos em vários documentos distintos, denominado representação "bag-of-words". A Tabela 1 exibe o modelo definido por Salton, Wong e Yang (1975), sendo $D =$

Figura 5 – Cortes de Luhn.



Fonte: Luhn (1958).

$\{d_1, d_2, \ldots, d_N\}$ uma coleção de N documentos que inclui M termos $T = \{t_1, t_2, \ldots, t_M\}$. Cada documento d_i é um vetor $\vec{v}\,(d_i) = \{freq_{i1}, freq_{i2}, \ldots, freq_{iM}\}$, no qual o valor $freq_{ij}$ é alguma medida que determina a influência do termo t_j no documento d_i.

Tabela 1 – Representação vetorial.

	t_1	t_2	...	t_M
d_1	$freq_{11}$	$freq_{12}$	...	$freq_{1M}$
d_2	$freq_{21}$	$freq_{22}$	...	$freq_{2M}$
...	...	...	...	...
d_N	$freq_{N1}$	$freq_{N2}$	...	$freq_{NM}$

Fonte: Salton, Wong e Yang (1975).

O valor $freq_{ij}$ pode ser calculado utilizando vários padrões. Os mais utilizados na literatura são:

- **boolean:** Se $freq_{ij} = 1$ o termo t_i ocorre no documento d_j, caso contrário se $freq_{ij} = 0$ o termo não ocorre no documento d_j.
- **terms frequency (tf):** A medida considera o valor de ocorrências de t_j no documento d_i:

$$freq_{ij} = tf(t_j, d_i) = freq(t_j, d_i) \tag{2.1}$$

- **terms frequency inverse document frequency (tfidf):** Jones (1988) mostrou que um termo muito frequente na coleção, que ocorre em muitos documentos, oferece pouco poder de discriminação entre os documentos da coleção. Assim, para aumentar a representatividade de um termo foi definido um fator de ponderação chamado *inverse document frequency (idf)*, definido pela Equação 2.2, onde os termos que

aparecem em poucos documentos são favorecidos. Sendo N o número de documentos da coleção e $d(t_j)$ é o número de documentos na coleção nos quais o termo t_j ocorre.

$$idf(t_j) = log \frac{N}{d(t_j)} \tag{2.2}$$

Sendo assim, o valor de *idf* de um termo raro é alto, enquanto para um termo frequente o fator *idf* tende a ser baixo. Então, o fator de ponderação *idf* pode ser combinado com a medida *tf* resultando uma nova medida chamada *term frequency inverse document frequency (tfidf)*, exibida na Equação 2.5 (JONES, 1988):

$$tf\text{-}idf\ (t_j, d_i) = tf(t_j, d_i) \times idf(t_j) = freq(t_j, d_i) \times log \frac{N}{d(t_j)} \tag{2.3}$$

Após a contagem dos termos na matriz de representação vetorial, provavelmente será observado um cenário onde os vetores que representam os documentos possam ter normas Euclidianas muito diferentes podendo afetar o resultado final (ALENCAR, 2013). Ou seja, haverá grande diferença entre os valores dos números reais associados a cada vetor interpretados geometricamente como o "cumprimento" dos mesmos. Sendo assim, uma normalização pode ser aplicada, transformando a norma Euclidiana de cada vetor em um valor unitário conforme:

$$\|\overrightarrow{v}(d_i)\|^2 = \sum_{k=1}^{M} freq_{ik} = 1 \tag{2.4}$$

A normalização é aplicada dividindo cada coordenada do vetor $\overrightarrow{v}(d_i)$ por sua norma:

$$freq'_{ij} = \frac{freq_{ij}}{\|\overrightarrow{v}(d_i)\|} \ para\ 1 \leq j \leq M \tag{2.5}$$

2.3 PROJEÇÕES MULTIDIMENSIONAIS

Quando técnicas de projeção são aplicadas a dados textuais, como coleções de documentos, elas dão origem a mapas de documentos. Neste contexto, essas técnicas tomam como entrada representações vetoriais (ver Seção 2.2) ou outras características extraídas, por exemplo uma matriz de dissimilaridade, ou distância, entre todos os pares de documentos.

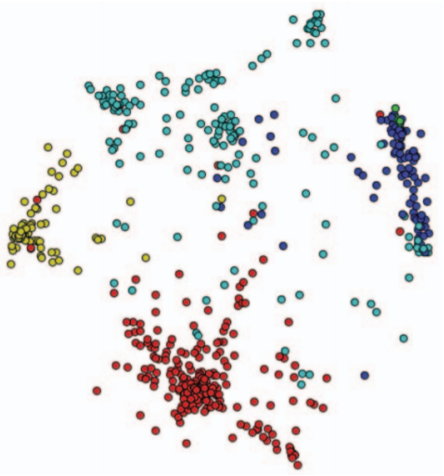
A maioria das técnicas de projeção multidimensional se baseia em informações de dissimilaridade entre instâncias de dados para incorporar dados de alta dimensão em um espaço visual. Sendo assim, a projeção multidimensional recebeu atenção significativa devido a sua capacidade de construir representações visuais que respeitam a proximidade

entre as instâncias de dados (PAULOVICH; SILVA; NONATO, 2010). Os métodos de projeção multidimensional mapeiam os dados do espaço m-dimencional cartesiano para um espaço visual p-dimensional, $p = \{2, 3\}$, preservando as distâncias o máximo possível.

A projeção multidimensional é um caso especial de uma classe mais ampla de técnicas chamada *Multidimensional Scaling* (MDS) (PAULOVICH; SILVA; NONATO, 2010). Os métodos MDS tem como objetivo colocar cada objeto no espaço m-dimensional, de modo que as distâncias entre os objetos sejam preservadas da melhor forma possível. Segundo Paulovich et al. (2008), as técnicas de projeção multidimensional são baseadas em combinações lineares de atributos de dados, definindo-os em uma nova base ortogonal de pequena dimensão, ou em um processo que tenta minimizar uma função da perda da informação ocorrida durante a projeção.

É exibido na Figura 6 um exemplo de técnica de projeção multidimensional denominada *Least Square Projection (LSP)* (PAULOVICH et al., 2008), onde foram processados documentos de diferentes áreas, e fica visível a preservação das distâncias entre os documentos processados, onde foram separados em diferentes grupos, representados por cores distintas na visualização. Algumas instâncias ainda se misturam devido a correlação entre o conteúdo textual dos documentos. A técnica engloba boas características dos métodos de projeção linear e não-linear, que mostra-se rápido no processamento em espaços espaciais de alta dimensionalidade, resultando em um posicionamento final preciso de pontos. Para o caso de documentos de texto o *LSP* pode gerar layouts de alta dimensão.

Figura 6 – Projeção de uma coleção de documentos composta por artigos científicos em quatro áreas diferentes, sendo que as cores indicam as áreas, utilizando a técnica *LSP*.



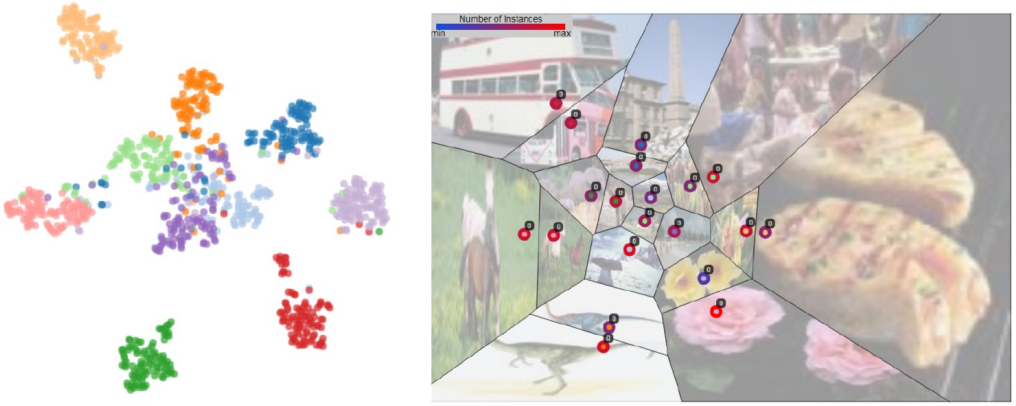
Fonte: Paulovich et al. (2008).

2.4 ABORDAGENS MULTINÍVEL

Algumas dificuldades surgem em projeções multidimensionais quando o número de instâncias ou a dimensionalidade do conjunto de dados processado aumenta, por exemplo a dificuldade na escalabilidade visual e a sobreposição dos marcadores utilizados para representar cada instância do conjunto. Na literatura existem técnicas para fazer a exploração hierárquica de projeções (MARCILIO; ELER, 2018; SILVA, 2016; PAULOVICH et al., 2008) que tem como objetivo sanar essas dificuldades encontradas pelo o usuário no momento de exploração das projeções.

Com o objetivo de tratar a dificuldade de escalabilidade visual em projeções multidimensionais, Marcilio e Eler (2018) propuseram uma abordagem de exploração multinível em visualizações geradas para representar projeções multidimensionais. Como primeira etapa, foi construída uma árvore para controlar a hierarquia da abordagem com base em seleção de instâncias representativas, onde os grupos são definidos como diagramas de Voronoi rígidos (BALZER; DEUSSEN; LEWERENTZ, 2005). O objetivo principal da técnica é fornecer meios para que o conjunto seja explorado com uma carga cognitiva menor do que em representações comuns de projeções multidimensionais (MARCILIO; ELER, 2018).

Figura 7 – Abordagem multinível proposta por Marcilio e Eler (2018). A esquerda uma projeção utilizando uma forma tradicional sobre o conjunto de dados *Corel*. A direita, sobre o mesmo conjunto de dados, uma projeção utilizando a abordagem hierárquica proposta

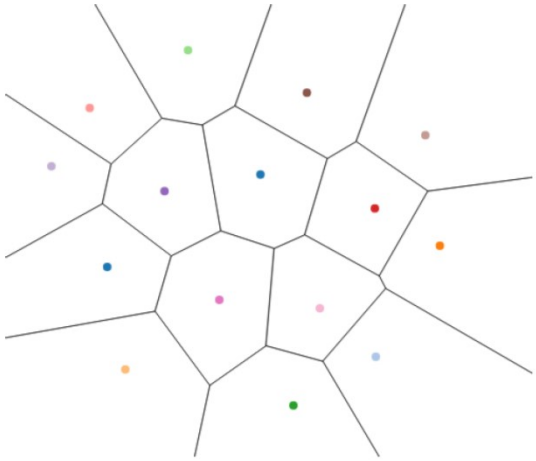


Fonte: Marcilio e Eler (2018).

O primeiro nível da hierarquia é definido com base na seleção de representativos em uma nuvem de pontos no plano. Com base nos representativos são definidos novos grupos, cuja definição é realizada por meio do domínio de Voronoi para codificar visualmente as fronteiras dos grupos, e uma nova seleção de representativos é realizada e novamente subgrupos são formados. O processamento continua até que cada grupo tenha no mínimo

uma quantidade M, pré-especificada, de instâncias. Então é criada uma árvore para que o usuário possa interagir com a projeção de forma hierárquica (MARCILIO; ELER, 2018). A Figura 8 exibe a divisão do espaço imposta pela seleção de representativos.

Figura 8 – Divisão do espaço imposta pela seleção de representativos.



Fonte: Marcilio e Eler (2018).

É possível que sejam gerados grupos em que o número de instâncias sejam menores que a quantidade mínima M pré-definida, de forma que precisem ser agrupados com outros grupos. Para isso, o processo reverso do algoritmo de agrupamento hierárquico é aplicado. Finalizado o processo de definição da hierarquia é efetuada a remoção de sobreposição (MARCILIO; ELER, 2018), que pode ser realizada pelos algoritmos *RWordle* (STROBELT et al., 2012a), *VPSC* (DWYER; MARRIOTT; STUCKEY, 2006a), *PRISM* (GANSNER; HU, 2010a) e *ProjSnippet* (GOMEZ-NIETO et al., 2014).

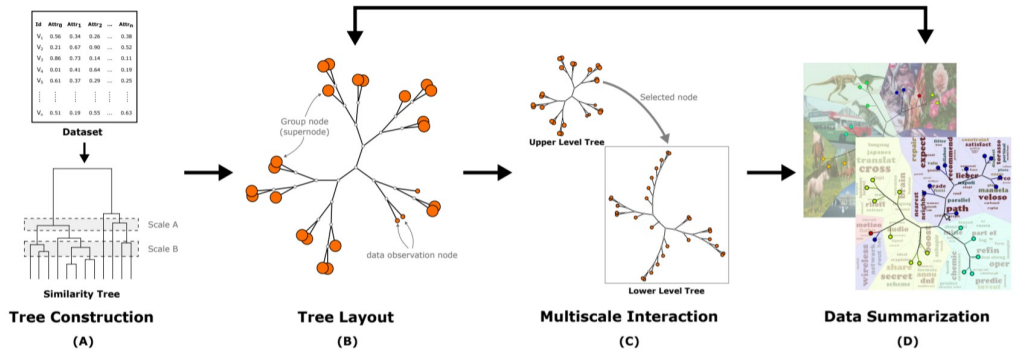
Considerando a interação, quando o cursor do mouse está sobre um representativo, uma janela com algumas informações acerca do grupo são apresentadas. Também são apresentadas as imagens correspondentes às instâncias similares utilizando o algoritmo *KNN* (COVER; HART, 1967) e instâncias diversas com o algoritmo *BRID* (SANTOS, 2012). Clicando sobre uma instância representativa é realizado o processo de expansão para que seja possível investigar um grupo. Quando uma expansão é efetuada o *zoom in* é realizado com valor proporcional ao número de elementos do grupo sendo analisado. Quando um grupo folha é alcançado, as instâncias correspondentes são apresentadas no plano de projeção quando o cursor do mouse está sobre o marcador e para visualizar o conteúdo das instâncias é possível clicar sobre o representativo. Também é possível definir *tags* para serem associadas as instâncias selecionadas que podem servir como uma maneira de organizar o conjunto de dados sendo explorado, e mapas de calor podem ser utilizados para codificar metadados do conjunto (MARCILIO; ELER, 2018).

Um trabalho correlato ao de Marcilio e Eler (2018) foi o desenvolvido por Silva

(2016) uma técnica denominada *Visual Super Tree*, na qual um método de exploração é aplicado em visualizações de árvores de similaridade. Um pré agrupamento é realizado sobre o conjunto de dados de maneira multinível, dando origem a um conjunto de árvores de similaridade de tamanho crescente, onde cada uma representa um nível do conjunto de dados particionado. A união dessas árvores forma a super árvore global, que conta com informações sobre todo o conjunto analisado.

A árvore é visualizada por meio de um processo exploratório hierárquico, garantindo que as relações de similaridade seja preservada, os super nós devem representar instâncias altamente relacionadas. Observa-se na Figura 9 uma visão geral do pipeline de construção da *Visual Super Tree*. O super nó (agrupamento) pode ser expandido (Figura 10) para exploração das instâncias relacionadas, e também as instâncias podem ser contraídas (Figura 11) formando super nós. Desta forma a técnica permite que o usuário explore todo o conjunto de dados conforme demanda.

Figura 9 – Pipeline de construção da *Visual Super Tree*. (a) Construção da árvore de similaridade sobre o conjunto de dados. (b) Layout da árvore onde nós grandes são agrupamentos, e nós pequenos são as instâncias em observação. (c) Interações de expansão e contração dos nós. (d) Sumarização dos dados.

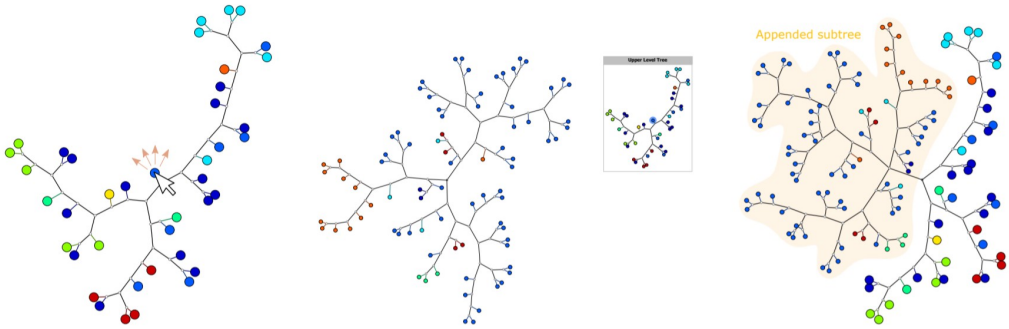


Fonte: Silva (2016).

2.5 TAG CLOUDS

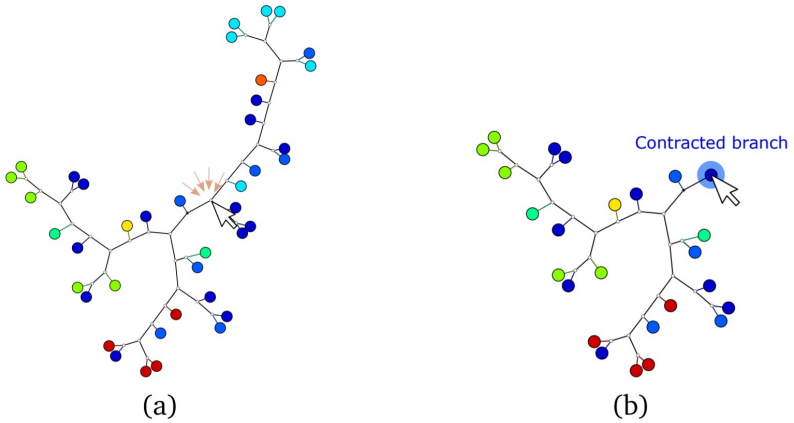
Tendo sua origem "fora do mundo dos computadores" (VIÉGAS; WATTENBERG, 2008), as nuvens de palavras, *tag clouds*, se tornaram populares no contexto de sites orientados para a comunidade, como *Flickr*, *Delicious* ou *Techorati*, que usam *tags* como um método de indexação (SMITH, 2008). Posteriormente, evoluíram como uma técnica de visualização de informação aplicada em vários contextos textuais. Alguns autores preferem chama-las de *"word clouds"* por representarem palavras de um conjunto de textos (HEIMERL et al., 2014).

Figura 10 – Expansão multinível. Quando um supernó é selecionado (a), uma nova subárvore é exibida em uma nova janela (b), ou expandida como um ramo da mesma árvore (c)



Fonte: Silva (2016).

Figura 11 – Contração multinível. Os ramos podem ser contraídos em super nós (a), economizando espaço visual para os nós restantes (b).



Fonte: Silva (2016).

Tag clouds são representações visuais baseadas em texto, por exemplo uma coleção de documentos, que utilizam vários tamanhos de fontes, cores, espaços, formas geométricas e posicionamento para representar palavras significativas. Desta forma as nuvens de palavras servem como ferramenta visuais que atraem e ajudam as pessoas a navegarem pelas informações (CHI et al., 2015). As nuvens de palavras surgiram como um método de visualização simples e visualmente atraente para o texto, tem como objetivo fornecer uma visão geral, processando o texto e dando destaque as palavras que aparecem com maior frequência.

Na Figura 12 é exibido um exemplo de *tag cloud* proposto por Heimerl et al. (2014) que tem como objetivo considerar as formas espaciais e movimentos temporais das nuvens de palavras. Neste exemplo, são processados textos sobre a evolução humana e as nuvens

são projetadas levando em consideração o formato da evolução do corpo humano.

Figura 12 – Exemplo de *tag cloud* sobre textos da evolução humana.

Fonte: Heimerl et al. (2014).

2.6 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentados os conceitos utilizados para criação da abordagem de exploração proposta neste trabalho. Esta abordagem tem como base a ferramenta de Marcilio e Eler (2018), que foi expandida para suportar conjunto de dados textuais, e adaptada para utilizar *tag clouds* como texturas para fornecer uma visão geral do conjunto textual analisado.

3 TRABALHOS RELACIONADOS

3.1 IDMAP

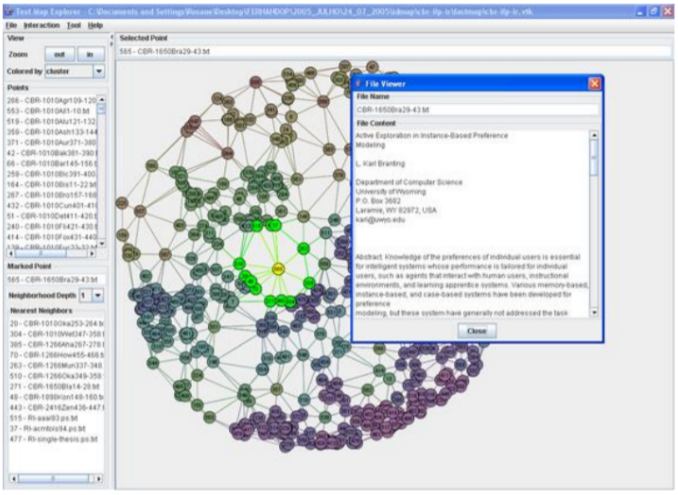
A técnica *Interactive Document Map* IDMAP, é uma abordagem para gerar mapas de documentos onde a vizinhança entre eles indicam suas similaridades. Baseado em técnicas de projeção multidimensionais e um algoritmo para melhoria de projeção, a abordagem resulta em um mapa de superfície que permite ao usuário identificar um conjunto de relações entre os documentos e subgrupos de documentos via visualização e interação. Os resultados são comparados com redução de dimensionalidade e técnicas de *cluster* para os mesmos fins (MINGHIM; PAULOVICH; LOPES, 2006).

Para criar projeções significativas cada documento é representado como um vetor de modo que seja possível calcular a distância entre os textos numericamente. Nessa técnica as etapas realizadas para construir um mapa de documentos são:

- Pré-processamento do conteúdo do texto para construir sua representação em um espaço vetorial.
- Projeção para um espaço 2D usando um algoritmo rápido, seguido por uma estratégia de melhoria denomidado *the Fource Scheme* (TEJADA; MINGHIM; NONATO, 2003).
- Agrupamento hierárquico dos dados projetados para identificação do subgrupo.

Para gerar a representação vetorial do texto a ser processado por esta técnica, os textos originais foram submetidos às etapas essenciais de pré-processamento de dados textuais, sendo elas: remoção de *stopwords*, aplicação de *stemming* para agrupar a contagem das palavras com mesmo sentido, remoção de palavras com baixa frequência aplicando o *corte de Luhn* (LUHN, 1958), representação de *bi-gramas* para palavras que ocorrem em sequência e cálculo do *term-frequency inverse document-frequency (tfidf)* para ponderar os termos de acordo com sua frequência no documento em relação a sua frequência em toda a coleção analisada (SALTON; BUCKLEY, 1988). Resultando em uma matriz $T_{n \times l}$ de documentos com n documentos e l termos. Cada linha da matriz (um documento) é um vetor, cada *bi-grama* final é uma dimensão e as *tfidfs* são as coordenadas. Por fim foi desenvolvida uma ferramenta para explorar a vizinhança de textos plotados. A Figura 13 mostra um exemplo dessa ferramenta.

Figura 13 – *TextMapExplorer*: Uma ferramenta de exploração para mapas de texto.



Fonte: Minghim, Paulovich e Lopes (2006).

3.2 VISUALIZAÇÃO ORGÂNICA DA EVOLUÇÃO DE DOCUMENTOS

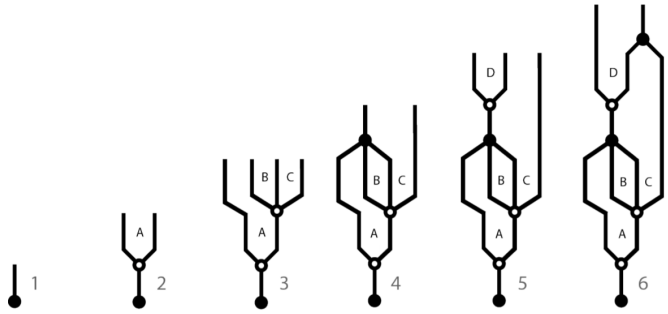
A técnica de visualização orgânica da evolução de documentos define uma estrutura de dados que captura o histórico de pressionamento de tecla na escrita de um documento e constrói uma visualização que serve como uma interface para este histórico. Seus resultados são promissores e revelam características sobre o documento como: estratégias gerais adotadas, densidade de edição local e estrutura hierárquica do texto final. Os serviços da *Web*, como o *Google Drive*, mantêm registros de alterações de documentos em nível de digitação para que os dados se tornem amplamente disponíveis, abrindo um grande leque para pesquisas de processamento e geração de linguagem natural de texto em sua foma temporal mutável (PEREZ-MESSINA; GUTIERREZ; GRAELLS-GARRIDO, 2017).

Na representação desenvolvida pela técnica, um documento vazio inicia como um nó (raiz) levando a uma única folha, ponto no qual apenas uma única inserção pode ocorrer. Quando é realizada uma inserção de caractere, ao pressionar uma tecla, a representação definida por essa técnica substitui o nó nulo na folha, e $n + 1$ arestas surgem, onde n é o número de caracteres inseridos, e cada uma dessas arestas contém um novo ponto de inserção. E quando é realizada uma exclusão de m caracteres, $m + 1$ arestas adjacentes correspondentes são agrupadas em um único nó que contém uma folha. Esse processo é exemplificado na Figura 14.

Foi utiliza uma abordagem baseados em glifos para visualizar a estrutura de dados mencionada pela técnica, que atuam de forma interdependentes e se constroem uns sobre os outros como a estrutura de uma planta. O glifo projetado para representar nós de inserção

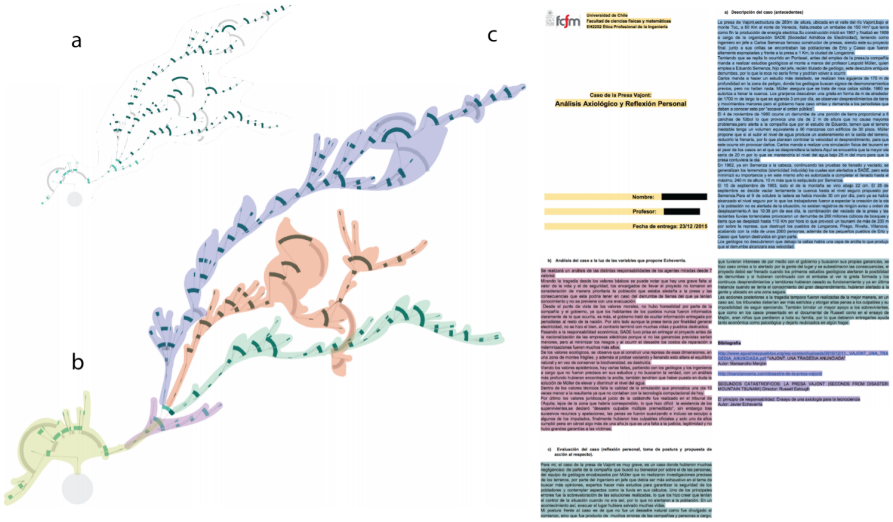
é, portanto, um multiplexador estilizado. Observamos na Figura 15 um processamento sobre um documento real desenvolvido por alunos de uma universidade pública que foram convidados a compartilhar seus documentos escritos no *Google Drive*. Observa-se pelas cores que a estrutura de ramificação da árvore é a mesma que a estrutura hierárquica do documento.

Figura 14 – Progressão da estrutura de dados de um documento. Nós deletados são circulos pretos, nós inseridos são circulos brancos e a as extremidades são pontos de inserção onde o cursor piscando pode percorrer. A direção é de baixo para cima da esquerda para a direita: (1) Documento vazio. (2) Inserção de “A”. (3) Inserção de “BC” na posição 2, resultando na string “ABC”. (4) Exclusão simultânea de “AB”. (5) Inserção de “D” antes de “C”, resultando na string “DC”. (6) Exclusão de “C”. O ducumento final contém apenas a string “D”.



Fonte: Perez-Messina, Gutierrez e Graells-Garrido (2017).

Figura 15 – Análise de um documento contendo 1567 palavras e que sofreu 7136 operações (inclusão e exclusão) sobre os caracteres. (a) Visualização bruta exibe a estrutura de ramificação. (b) Cores foram adicionadas manualmente. (c) Regiões do texto correspondentes as cores de (b).

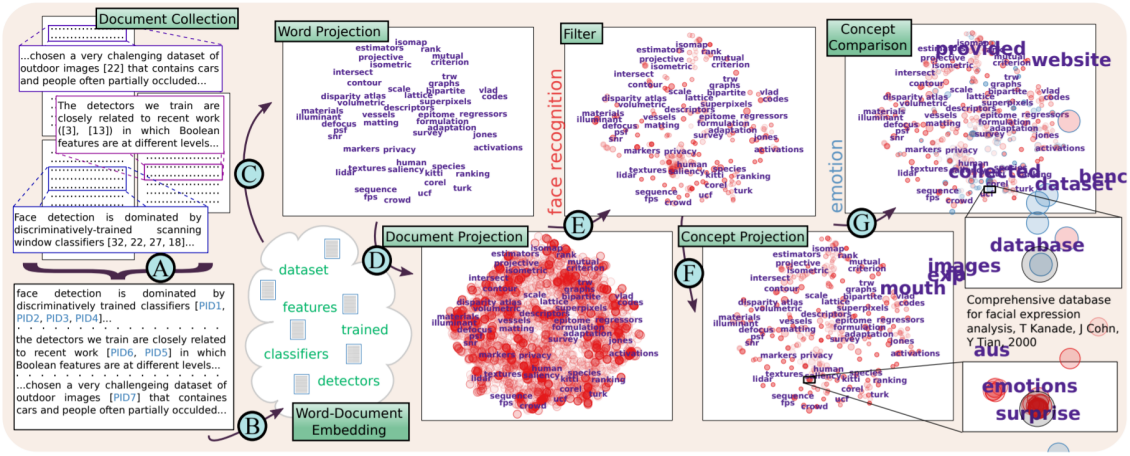


Fonte: Perez-Messina, Gutierrez e Graells-Garrido (2017).

3.3 CITE2VEC

A técnica denominada *cite2vec* define um método para explorar e descobrir coleções de documentos utilizando citações realizadas no corpo dos documentos como base de seu processamento, visto que as pessoas tendem a citar outros documentos por motivos muito precisos. A técnica projeta palavras representativas de documentos em um espaço bidimensional, e também projeta documentos nesta mesma visualização, de modo que a proximidade de um documento e uma palavra indique seu modo de uso, preservando também a similaridade de uso de documentos para documentos. O usuário pode também modificar iterativamente as projeções de documentos, especificando frases arbitrárias que alteram o significado das palavras projetadas (BERGER; MCDONOUGH; SEVERSKY, 2017). Podemos visualiza-la na Figura 16.

Figura 16 – Visão geral da técnica *cite2vec* utilizando uma coleção de documentos para seu processamento, processando as citações por meio da semântica das palavras, projetando documentos junto com palavras representativas, filtrando os documentos por meio de strings de busca inseridas pelo usuário.



Fonte: Berger, McDonough e Seversky (2017).

A primeira etapa da técnica é (A) coletar uma coleção de documentos e unificar as citações. Após, em (B) todos os documentos são agregados em uma única sequência de texto tratando cada documento como uma palavra única no processo de aprendizagem. Em (C) são projetadas as palavras relevantes para o documento. Em seguida, são projetados os documentos em (D), renderizados como círculos, mantendo os documentos próximos as palavras relevantes e as similaridade de documento para documento é preservada. Em (E), após uma frase ser fornecida, os documentos irrelevantes são filtrados e os relevantes são posicionados para sua melhor posição (F). Também é permitido a inserção de novas frases, observado em (G). É possível observar a mudança de posição de um documento específico quando o usuário fornece uma nova string à técnica (BERGER; MCDONOUGH;

SEVERSKY, 2017).

Observa-se na Figura 17 a diversidade dos agrupamentos realizados pela técnica, alguns se referem a tópicos de visão computacional, enquanto outros são conceitos semânticos mais gerais, como animais, clima e orientações.

Figura 17 – Agrupamentos de palavras associadas em *cite2vec*.

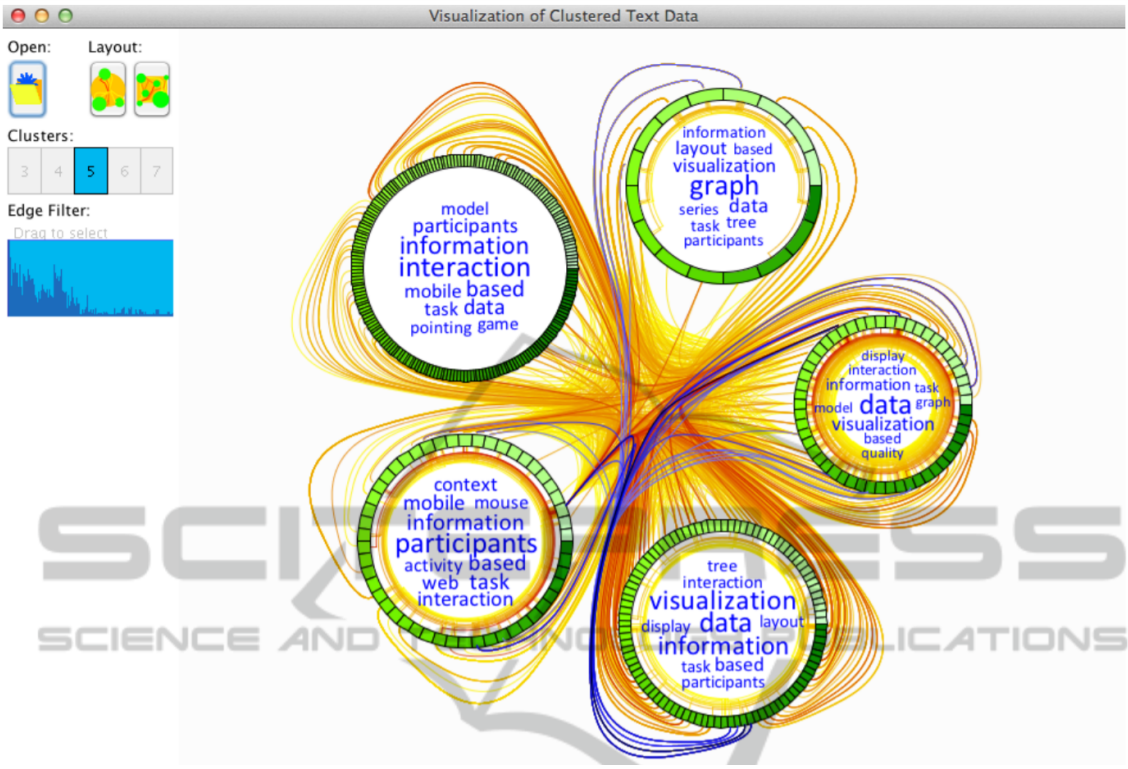
Fonte: Berger, McDonough e Seversky (2017).

3.4 EXPLORAÇÃO VISUAL ENTRE AGRUPAMENTOS DE DOCUMENTOS

Jusufi et al. (2014) propuseram uma visualização que exibe a relação entre agrupamentos de documentos baseando-se em tópicos (palavras-chaves) ou coautores comuns. A técnica exibe as relações dentro dos agrupamentos e entre os agrupamentos fornecendo informações sobre o conteúdo dos atributos dos agrupamentos. A proposta tem como objetivo facilitar a exploração interativas de redes multivariadas.

Na Figura 18 é exibido um exemplo da visualização definida por Jusufi et al. (2014), cada segmento no circulo de cada agrupamento representa um documento do conjunto de dados carregado inicialmente. A saturação da cor (verde no exemplo) de cada segmento representa a ordenação com base no valor de um atributo pré-selecionado pelo usuário ao selecionar o conjunto de dados. As linhas curvas que ligam os agrupamentos, e ligam os segmentos internos, representam a coautoria entre os documentos, o gradiente varia de amarelo a vermelho, e representa o número de autores compartilhados entre os dois documentos, sendo amarelo um menor número de coautores e vermelho o contrário, normalizado pelo total de coautores comuns (JUSUFI et al., 2014). A *tag cloud* fornece informações sobre os principais conceitos descritos em cada grupo de documentos. O usuário pode filtrar uma aresta aproximando o mouse e clicando sobre ela, o filtro de arestas também pode ser realizado manipulando o histograma.

Figura 18 – Visualização da técnica de Jusufi et al. (2014). O primeiro botão é para abrir arquivos de dados, os dois botões ao lado para alternar entre dois *layouts* de agrupamentos alternativos e os botões numerados de 3 a 7 especificam o número de agrupamentos desejados. O histograma exibe a distribuição de peso da aresta (eixo x: peso, eixo y: número de arestas). Pode ser utilizado para selecionar intervalos de pesos. Na área de visualização principal são exibidos cinco agrupamentos distintos. As arestas são destacadas em azul se forem selecionadas.



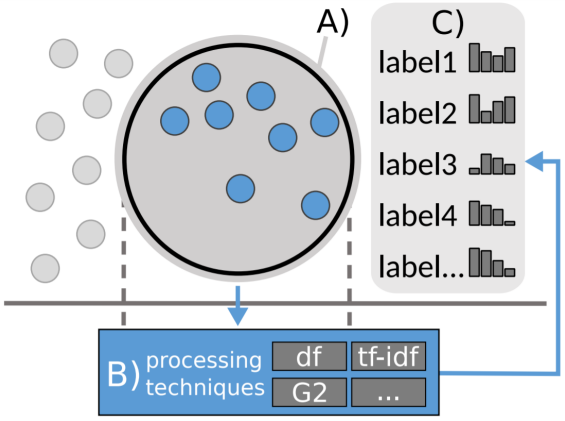
Fonte: Jusufi et al. (2014).

3.5 DOCUCOMPASS

A técnica proposta por Heimerl et al. (2016) define uma visualização composta por lentes que podem, interativamente, percorrer um espaço bidimensional onde documentos foram projetados em uma visualização principal. Ao mover essas lentes sobre uma seleção de documentos, outras técnicas são utilizadas e novas visualizações surgem dando apoio a uma investigação mais detalhada. A Figura 19 exibe os dois blocos de construção que são base para a técnica *DocuCompass*, o primeiro é a lente para focar um conjunto de documentos que o usuário pretende descobrir detalhes, o interior da lente conta com uma caracterização visual dos textos focalizados que são atualizados conforme o movimento da lente. O segundo bloco é um conjunto de técnicas de análise de documentos que alimentam a caracterização visual com seus resultados. O *DocuCompass* pode ser configurado com diferentes técnicas

de análise de documentos e uma variedade de representações visuais (HEIMERL et al., 2016).

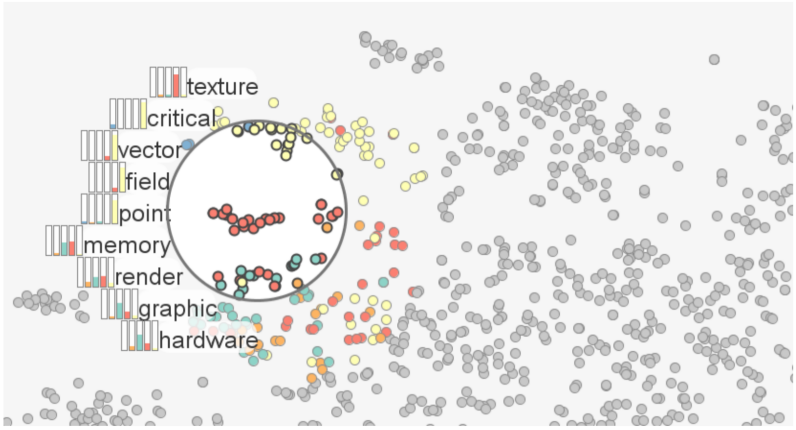
Figura 19 – Visão geral da técnica *DocuCompass*. (A) Metáfora visual flexível para que o usuário selecione um conjunto de documentos de interesse. (B) Diferentes técnicas de processamento de texto podem ser configuradas para extrair e classificar termos. (C) Um terceiro componente visual que pode ser atualizado para trazer informações para o usuário.



Fonte: Heimerl et al. (2016).

Na Figura 20 observa-se uma visualização alternativa para exibir agrupamentos sobre um conjunto de documentos focados. Se o usuário tiver interesse em "*vector field*" (campo vetorial) e "*critical point*" (ponto crítico), ele pode mover a lente para a região que contém mais documentos relacionados aos temas (região com pontos amarelos).

Figura 20 – Visualização de agrupamentos de documentos focados. Gráficos de barras com as cores dos grupos são exibidos próximos aos termos para exibir sua relevância em cada agrupamento.



Fonte: Heimerl et al. (2016).

3.6 SENTENTREE

A técnica *SentenTree* definida por Hu, Wongsuphasawat e Stasko (2017) emprega ideias de design de nuvens e árvores de palavras para explorar o conteúdo textual de mídia social, exibindo um diagrama de vértices onde os nós são palavras e os links indicam a coocorrência da palavra na mesma frase. A distribuição espacial dos nós exibe a ordenação sintática das palavras, enquanto o tamanho dos nós exibe a frequência de ocorrência de determinada palavra. A técnica tem como objetivo ajudar o usuário a obter uma compreensão rápida dos principais conceitos e opiniões em uma grande coleção de texto em mídias sociais (HU; WONGSUPHASAWAT; STASKO, 2017).

Na Figura 21 é exibido um processamento realizado utilizando dados de milhares de *tweets* extraídos da rede social *Twitter* durante o período do primeiro gol do jogo de abertura da Copa do Mundo de futebol em 2014. A técnica procura um equilíbrio entre mostrar as palavras mais frequentes e preservar a estrutura das frases.

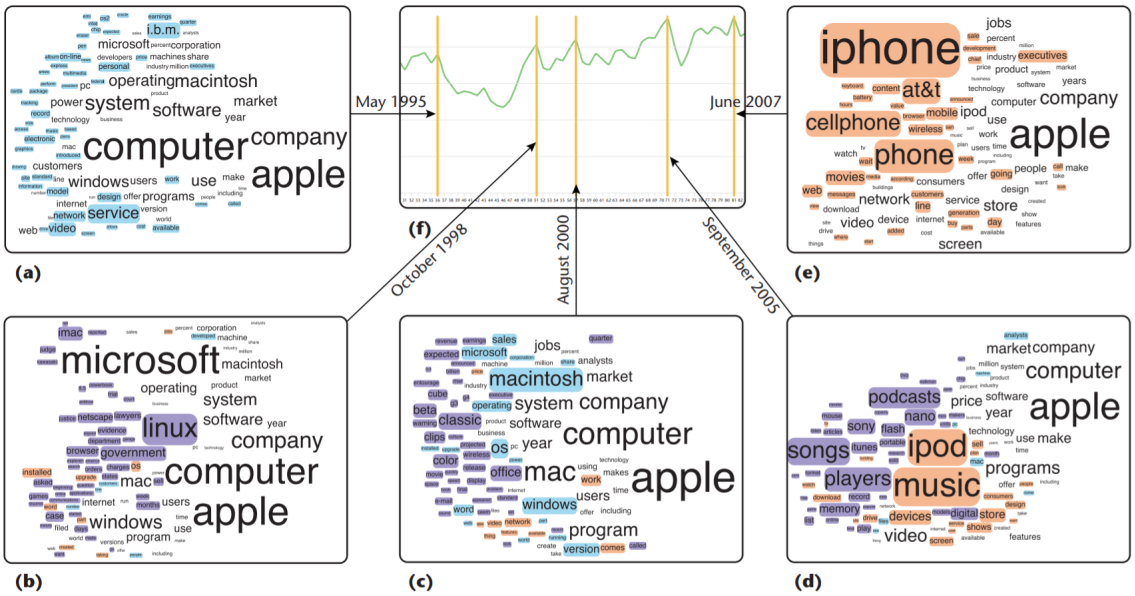
Figura 21 – Parte de uma visualização da técnica *SentenTree* sobre uma coleção de 189.450 *tweets* postados em um período de 15 minutos em torno do primeiro gol do jogo de abertura da Copa do Mundo de 2014.

Fonte: Hu, Wongsuphasawat e Stasko (2017).

3.7 PRESERVAÇÃO DE CONTEXTO EM TAG CLOUDS

Cui et al. (2010) propuseram um método dinâmico para geração de *tag clouds* que pretende garantir a coerência semântica e um bom posicionamento das palavras no espaço. O método proposto gera uma sequência de *tag clouds* agrupando as palavras relacionadas. As sequências são acopladas a um gráfico de tendências que resume as alterações de conteúdo para que os usuários possam explorar melhor as grandes coleções de documentos. É possível observar um exemplo geral da técnica na Figura 22, onde foram processadas notícias sobre a empresa *Apple* do ano de 1995 à 2007. A técnica tenta estabilizar a posição de uma palavra em diferentes nuvens facilitando o rastreamento por parte do usuário. As cores de fundo indicam se os termos estão desaparecendo (azul), surgindo (laranja), se são únicos do período (roxo) ou apenas mudam a frequência (sem cor de fundo) em comparação as outras *tag clouds* adjacentes.

Figura 22 – Uma visão geral da técnica desenvolvida por Cui et al. (2010). São criadas cinco *tag clouds* para cinco pontos de tempo. A caixa central (f) apresenta um gráfico de tendência de significância cuja curva é extraída de uma coleção de documentos com diferentes registros de data e hora.



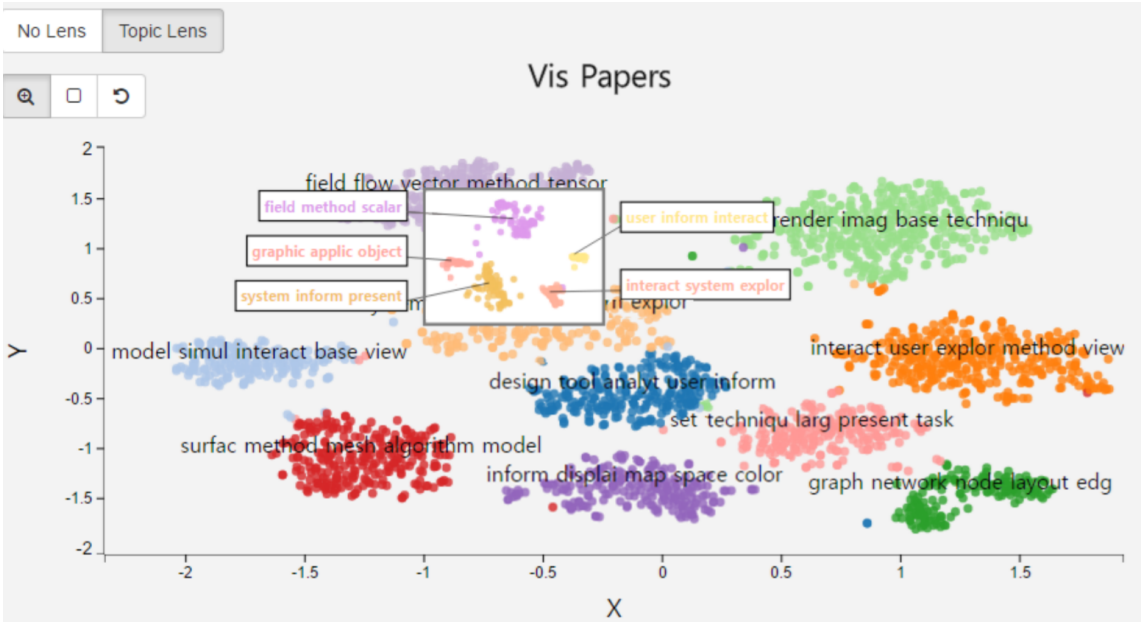
Fonte: Cui et al. (2010).

3.8 TOPICLENS

Também baseada em técnicas de projeções multidimensionais, a técnica TopicLens (KIM et al., 2017) é uma técnica de interação que permite ao usuário explorar dinamicamente os dados textuais por meio de uma interface de lente, similar a abordagem de lente citada na Seção 3.5, onde a modelagem de tópicos e a projeção bidimensional são computadas em tempo real.

A modelagem de tópicos, utilizada por Kim et al. (2017), é uma das técnicas mais amplamente utilizadas na mineração de texto, processamento de linguagem natural e aprendizado de máquina. O objetivo principal da modelagem de tópicos é derivar uma coleção dos chamados tópicos, mesmo a partir de um grande conjunto de documentos, em que cada tópico é representado por um conjunto de palavras-chave coerentes que descrevem um subconjunto dos documento. Sendo que esses tópicos fornecem aos usuários um resumo de alto nível do corpus de documento, sem a necessidade de ler documentos individuais. Kim et al. (2017) propuseram a melhoria da modelagem de tópicos tradicionais, com inserção de um processo interativo de refinamento guiado pelo usuário e melhoria do tempo de processamento para exibição de dados em tempo real.

Figura 23 – Visão geral do sistema de análise visual definido por Kim et al. (2017). Inicialmente o sistema executa uma modelagem de tópico e os documetos são projetados em um espaço bidimencional, o os agrupamentos do tópico são codificados por cores. As palavras-chave representativas são exibidas no centro de cada agrupamento. Ao mover a lente retangular o modelo é recomposto dinamicamente e revelada uma estrutura de tópicos mais detalhada com novas palavras-chave representativas.



Fonte: Kim et al. (2017).

3.9 PROJCLOUD

Segundo Paulovich et al. (2012), os métodos existentes para análise de documentos utilizando *tag clouds* são eficazes para demonstrar o conteúdo, mas não são capazes de preservar relações semânticas entre palavras chaves e seus documentos. Preservar essa relação semântica é o objetivo da técnica *ProjCloud*. Nesta técnica o usuário pode visualizar, a partir de uma projeção multidimensional, a relação de vizinhança entre documentos altamente relacionados e as suas *tag clouds* correspondentes. As nuvens de palavras são dispostas em polígonos, e a quantidade de nuvens pode ser definida pelo usuário. Na Figura 24 é possível observar a representação visual da técnica *ProjCloud*. Nesta representação foi utilizada uma coleção de 675 artigos científicos de quatro áreas distintas: raciocínio baseado em casos, programação lógica indutiva, recuperação de informação e sonificação.

Observa-se que grupos de documentos distintos são facilmente identificados e os tópicos principais que descrevem estes documentos são claramente destacados, sendo que Paulovich et al. (2012) também desenvolveram um novo algoritmo para construir

Figura 24 – *ProjCloud* de uma coleção de artigos científicos em quatro diferentes áreas do conhecimento, definindo um limite de quatro agrupamentos.

Fonte: Paulovich et al. (2012).

nuvens de palavras dentro de polígonos mantendo a relação semântica entre as palavras. Outra característica que deve ser destacada é que o algoritmo utilizado pela técnica divide os agrupamentos recursivamente, desta forma nuvens de palavras podem ser refinadas sob demanda, conforme necessidade identificada pelo usuário, para mostrar informações mais detalhadas dentro de cada agrupamento, como ilustrado na Figura 25.

Figura 25 – *ProjCloud* da mesma coleção de artigos científicos da Figura 24 definido agora em nove agrupamentos.

Fonte: Paulovich et al. (2012).

3.10 CONSIDERAÇÕES SOBRE O CAPÍTULO

Técnicas de classificações e visualização para análise textual são bastante empregadas para facilitar o processo de descoberta de conhecimento por seres humanos em tarefas que necessitam de uma análise a corpus volumosos de documentos. Existem diversas técnicas em suas mais variadas implementações disponíveis na literatura, dentre elas as citadas neste documento. Por meio deste estudo podemos perceber que técnicas que exploram a interação, e com isso, obtêm um *feedback* por parte do usuário, conseguem atingir melhores resultados pois vão de encontro com o que o usuário necessita na exploração em questão.

Utilizando agrupamento de documentos é possível criar maneiras eficientes de obter conhecimento sobre grandes coleções de documentos. Sendo que projeções multidimensionais reduzem o tempo de análise significantemente em comparação ao trabalho árduo que o usuário teria ao realizar uma exploração deste tipo manualmente. Várias aplicações que utilizam de interfaces visuais para apoiar a interpretação de algoritmos tradicionais de mineração de texto são encontradas. Mas mesmo assim, visualizações podem ser desenvolvidas para dar ao usuário um papel muito mais ativo nas tarefas de mineração de texto e atividades relacionadas, e aplicações que exploram fortemente este conceito são escassas. Muitas alternativas ricas permanecem abertas para novas ideias e uma maior exploração. Temos também que nos atentar ao tempo de processamento exigido por essas ferramentas a serem desenvolvidas para o processamento textual, sendo que uma vez que suportam a interação do usuário o tempo de resposta exigido deve ser curto, sendo que este processamento é realizado em tempo real.

4 ABORDAGEM HÍBRIDA PARA VISUALIZAÇÃO DE COLEÇÕES DE DOCUMENTOS

4.1 CONSIDERAÇÕES INICIAIS

Geralmente, técnicas de projeção multidimensional são empregadas para fazer o mapeamento do espaço m-dimensional para o espaço bidimensional. Essas técnicas podem ser usadas para representar relacionamentos entre instâncias de dados com base na distância, agrupando ou separando grupos no espaço projetado. Vários trabalhos, como *TopicLens* (KIM et al., 2017), *DocuCompass* (HEIMERL et al., 2016), *cite2vec* (BERGER; MCDONOUGH; SEVERSKY, 2017), e *IDMAP* (MINGHIM; PAULOVICH; LOPES, 2006), vêm utilizando projeções multidimensionais para auxiliar na exploração de coleções de documentos. Embora as técnicas de projeção possam organizar um conjunto de dados, o usuário precisa ler cada documento para entender os posicionamentos dos pontos baseados nas relações de similaridade e também a geração dos grupos de documentos. Como alternativa, técnicas como extração de tópicos ou *tag clouds* podem ser empregadas para apresentar um resumo do conteúdo do documento. Para minimizar o trabalho exploratório e auxiliar na análise de um grupo de documentos, este trabalho apresenta uma visualização híbrida para mostrar o relacionamento e o conteúdo do documento em uma única exibição, empregando projeções multidimensionais para relacionar documentos e *tag clouds*. Mostramos a eficácia da abordagem proposta na exploração de duas coleções de documentos compostas por notícias do mundo.

Para melhorar a exploração e análise baseadas em técnicas de projeção multidimensional, Silva e Eler (2017) propuseram uma abordagem de visualização híbrida para mapear as semelhanças das instâncias no espaço $2D$, mostrando agrupamentos de instâncias semelhantes e apresentando uma imagem para cada instância para destacar o comportamento dos atributos. Além disso, para ajudar a entender o relacionamento de similaridade entre instâncias, a abordagem anterior permite ao usuário diferenciar instâncias de classes distintas que compartilham um limite comum. O trabalho aqui apresentado fornece uma extensão da abordagem anterior, na qual propomos uma abordagem de visualização híbrida para mapear as semelhanças do documento no espaço de $2D$ e mostrar *tag clouds* para cada documento, apresentando os principais termos dos dados textuais. A visualização de similaridade e palavras chaves em uma única exibição pode melhorar a exploração de dados textuais e auxiliar no entendimento da formação de grupos.

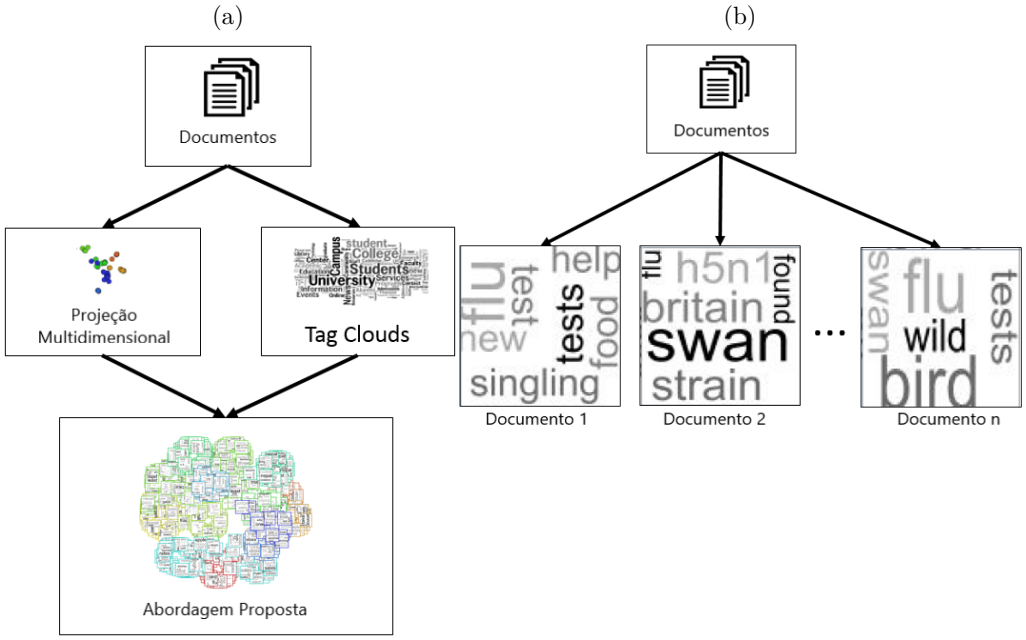
A principal contribuição deste trabalho é ajudar na exploração de conjuntos de dados textuais com base na abordagem de visualização híbrida proposta, que mapeia as semelhanças e o conteúdo do texto em uma visualização exclusiva. Essa abordagem usa todos os atributos do conjunto de dados para gerar a representação visual e mostrar algumas palavras chaves do texto, assim, o usuário pode entender os principais tópicos de agrupamentos distintos. Além disso, a visualização individual da *tag cloud* pode ajudar o usuário a entender a formação de grupos de instâncias semelhantes e melhorar a detecção do limite entre grupos distintos. Também introduzimos alguns mecanismos de interação para melhorar a experiência do usuário durante o processo exploratório.

4.2 ABORDAGEM PROPOSTA

A abordagem proposta (ANDREOTTI; SILVA; ELER, 2018) foi concebida para auxiliar na análise de coleções de documentos, criando uma representação gráfica que mistura projeção multidimensional e *tag clouds* no mesmo pipeline, como mostra a Figura 26 (a). A técnica de projeção multidimensional é empregada para agrupar documentos semelhantes no espaço de $2D$, revelando os documentos que apresentam conteúdo semelhante. A técnica de *tag clouds* é empregada para mostrar um resumo de cada documento e é usada como marcador visual na representação gráfica, como mostra a Figura 26 (b). A *tag cloud* calculada para cada documento pode revelar os principais termos dos dados textuais. A abordagem proposta é descrita aqui sem especificar as técnicas de projeção de *tag clouds*, porque qualquer técnica pode ser usada para colocar os documentos no espaço de $2D$ e qualquer técnica pode ser usada para gerar uma *tag cloud* para cada documento.

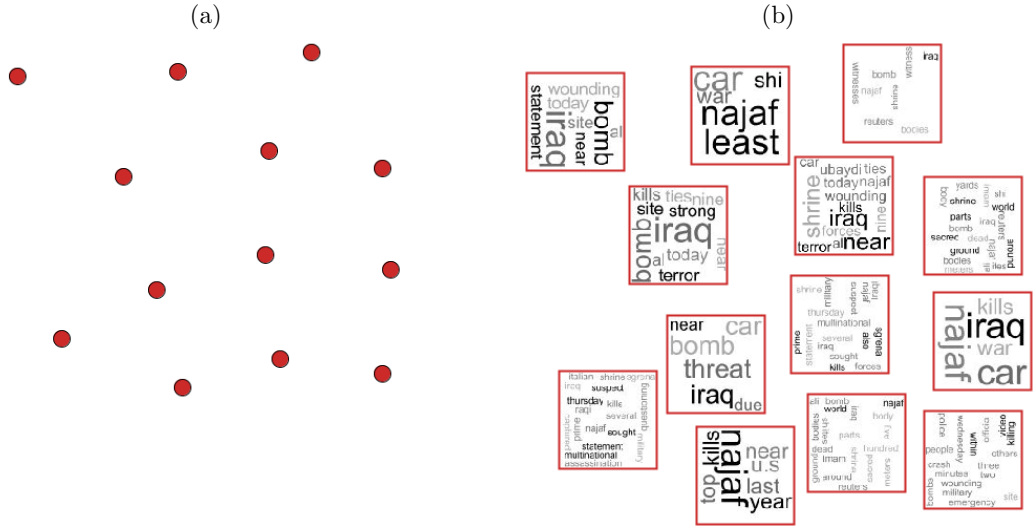
A Figura 27 apresenta um exemplo ilustrativo no qual uma coleção de documentos é projetada no espaço $2D$ e as imagens das *tag clouds* são usadas como marcadores visuais para mostrar o conteúdo dos dados textuais. O analista pode usar essa visualização para ver as semelhanças do documento no espaço de $2D$ e também analisar o conteúdo dos documentos. Essa abordagem pode ser usada para entender por que alguns grupos de documentos são gerados, pois a *tag cloud* mostra os termos mais frequentes. Geralmente, os grupos são gerados com base na ocorrência de palavras em cada documento. Neste exemplo, o grupo de documentos é composto por notícias relacionadas a um ataque a bomba no Iraque.

Figura 26 – Abordagem proposta: (a) mostra o processo completo e (b) mostra o processo detalhado de processamento das *tag clouds* para cada documento.



Fonte: Elaborado pelo autor.

Figura 27 – Exemplo de aplicação: (a) mostra uma projeção $2D$ de uma coleção de documentos e (b) mostra as imagens das *tag clouds* mapeadas como marcadores visuais.

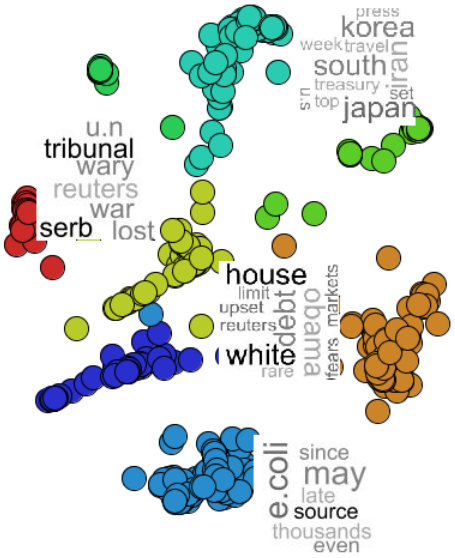


Fonte: Elaborado pelo autor.

A abordagem proposta é apoiada por ferramentas de interação para auxiliar o processo exploratório e a experiência do usuário. Por exemplo, conforme mostrado na

Figura 28, o usuário pode visualizar uma *tag cloud* de um documento passando o mouse sobre um ponto específico.

Figura 28 – Ferramenta de interação: a *tag cloud* de um documento é mostrada ao passar o mouse sobre um ponto.



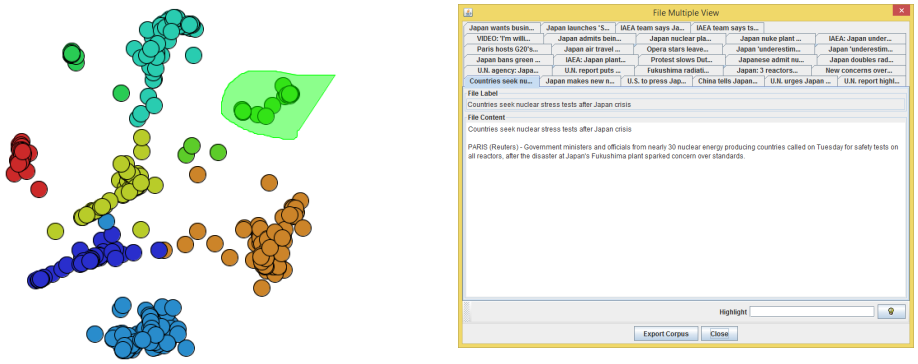
Fonte: Elaborado pelo autor.

Não obstante, o usuário ainda pode selecionar um grupo de documentos (consulte a Figura 29 (a)) para mostrar os dados de texto (consulte a Figura 29 (b)); a *tag cloud* de cada documento selecionado (consulte a Figura 29 (c)); ou uma *tag cloud* calculada a partir de dados textuais de todos os documentos selecionados (consulte a Figura 29 (d)).

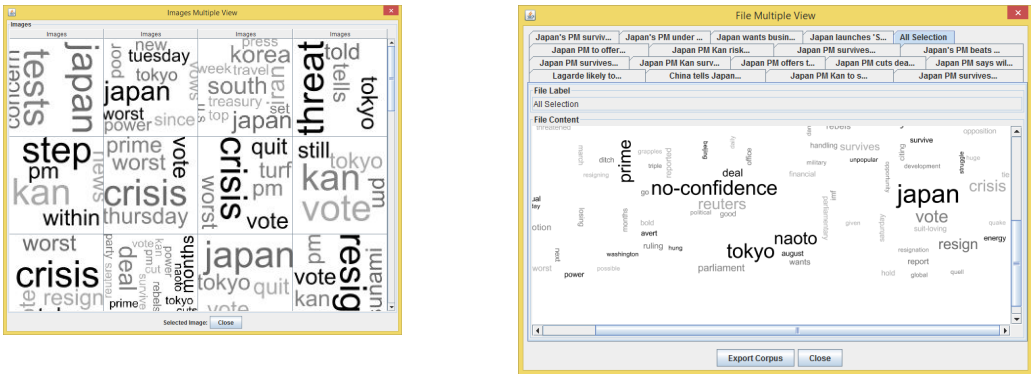
A próxima Seção mostra exemplos dessa abordagem na detecção de limites de um conjunto de dados.

Figura 29 – Ferramentas de interação: (a) seleção de um grupo de documentos; (b) o conteúdo de cada documento é mostrado em uma visão separada; (c) a *tag cloud* de cada documento é mostrada em uma visão separada; (d) uma *tag cloud* criada a partir da seleção (todos os documentos selecionados) é mostrada em uma exibição separada.

(a) (b)



(c) (d)



Fonte: Elaborado pelo autor.

4.3 APLICAÇÕES

Esta Seção apresenta aplicações para mostrar como a abordagem proposta pode ser empregada para analisar coleções de documentos. Para isso, usamos as seguintes coleções de documentos de notícias do mundo:

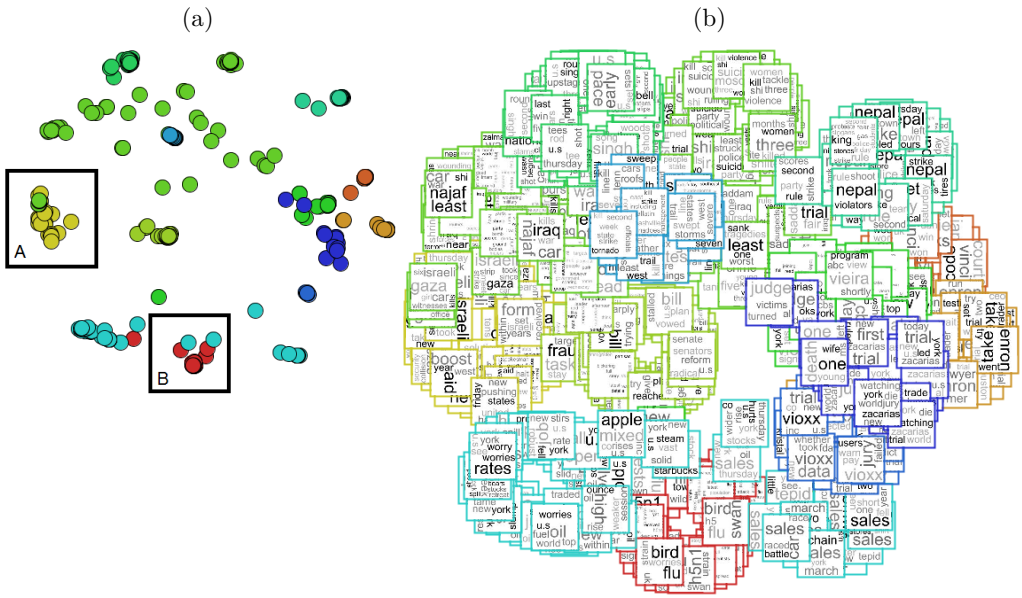
- **NEWS-8:** 495 notícias da Reuters, AP, BBC e CNN, divididas em 8 classes;
- **NEWS-13:** RSS de notícias da BBC, CNN, Reuters e Associated Press, coletadas durante dois dias em abril de 2006. Utilizamos apenas 381 documentos, divididos em 13 classes de notícias.

A abordagem proposta permite que qualquer técnica de projeção multidimensional

seja empregada para colocar os documentos no espaço $2D$. No entanto, para obter melhores resultados, é sempre preferida uma técnica de projeção capaz de preservar as relações entre documentos no espaço $2D$. Na literatura, vários trabalhos apresentaram análises detalhadas e comparações de técnicas de projeção multidimensional (PAULOVICH et al., 2008; ELER et al., 2015). Com base nesses trabalhos, empregamos a técnica de projeção dos mínimos quadrados (LSP) (PAULOVICH et al., 2008), que é uma técnica rápida proposta para projetar coleções de documentos no espaço $2D$. Além disso, usamos o método de Eler e Garcia (2013) para encontrar um bom limite para calcular os modelos de espaço vetorial (ou seja, espaços de características ou matriz de "documentos x termos").

O resultado da projeção do conjunto de dados **NEWS-13** é apresentado na Figura 30 (a). Informações sobre cores foram usadas para mapear a classe de cada notícia e dois grupos de documentos foram selecionados para uma inspeção mais aprofundada. A segunda etapa da abordagem proposta é mapear as *tag clouds* como marcador visual de cada documento – o marcador visual é uma imagem ou ícone para representar cada instância do conjunto de dados. Na Figura 30 (b) é apresentado o resultado da abordagem proposta, mostrando uma *tag cloud* para cada documento.

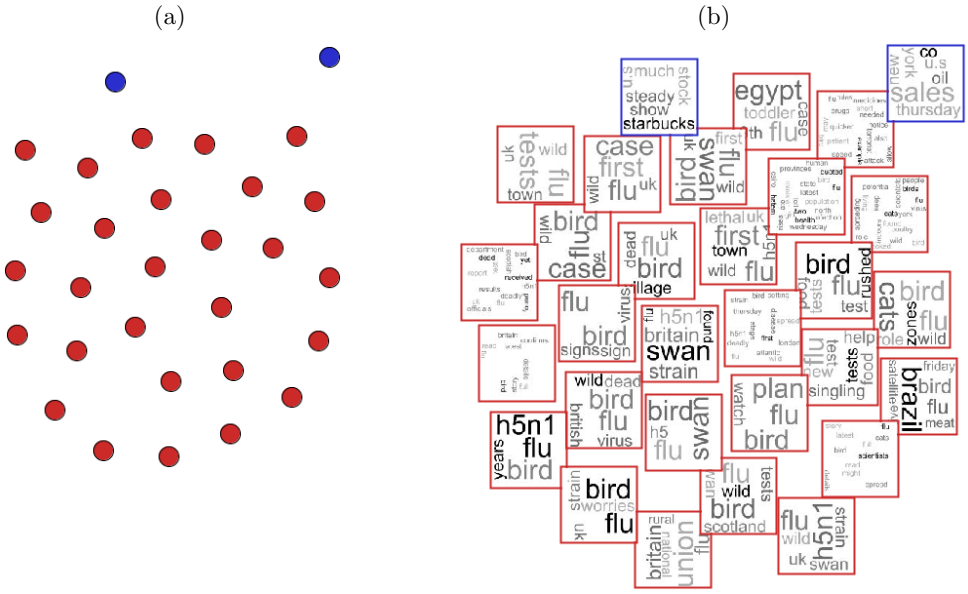
Figura 30 – A análise do conjunto de dados **NEWS-13**: (a) mostra uma projeção $2D$ da coleção de documentos e (b) mostra a *tag cloud* de cada documento como marcador visual.



Fonte: Elaborado pelo autor.

Observa-se que o usuário precisa explorar a representação gráfica com ferramentas de interação, como zoom, filtro ou seleção. A Figura 31 mostra a operação de zoom realizada na "Seleção B" da Figura 30 (a). Nesse exemplo, a abordagem híbrida é usada

Figura 31 – Análise do conjunto de dados **NEWS-13**: (a) zoom da Seleção B da Figura 30 e (b) mostra o mapeamento da *tag cloud* de (a).



Fonte: Elaborado pelo autor.

para permitir a análise da similaridade e do conteúdo do documento. A Figura 31 (a) mostra o zoom da projeção e a Figura 31 (b) mostra as *tag clouds* para cada documento, exibindo o conteúdo do grupo.

Além de usar a operação de zoom, o usuário também pode visualizar as *tag clouds* de um grupo selecionado. A Figura 32 mostra o conteúdo da visualização (*tag cloud*) para duas seleções realizadas na projeção apresentada na Figura 30 (a). O conteúdo da "Seleção A" é apresentado na Figura 32 (a), que mostra um conteúdo relacionado a notícia *"Israeli aircraft fired by missiles on a car carrying Palestinian militants in southern Gaza"*. O conteúdo da "Seleção B" é apresentado na Figura 32 (b), que mostra um conteúdo relacionado a notícia *"Bird flu and h5n1 tests confirmed on a dead swan"*.

Figura 32 – Análise do conjunto de dados **NEWS-13**: (a) mostra algumas *tag clouds* da Seleção A da Figura 30 e (b) mostra algumas *tag clouds* da Seleção B da Figura 30.

(a) (b)

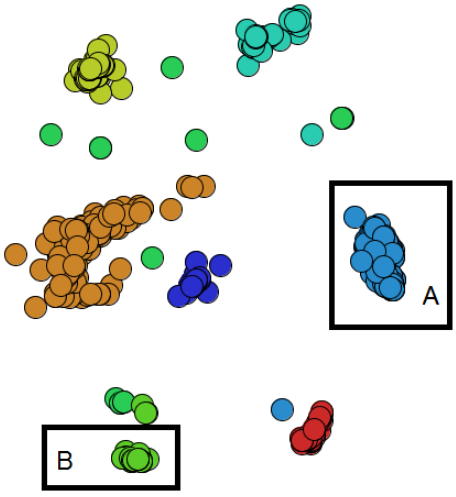
Fonte: Elaborado pelo autor.

Também aplicamos a abordagem proposta para explorar o conjunto de dados **NEWS-8**, mostrando um processo de exploração semelhante ao aplicado ao **NEWS-13**. Primeiro, a projeção do conjunto de dados coloca cada documento no espaço $2D$ para mapear as semelhanças dos documentos, como mostra a Figura 33 (a). Depois disso, as *tag clouds* são mapeadas como marcador visual de cada instância, como mostra a Figura 33 (b).

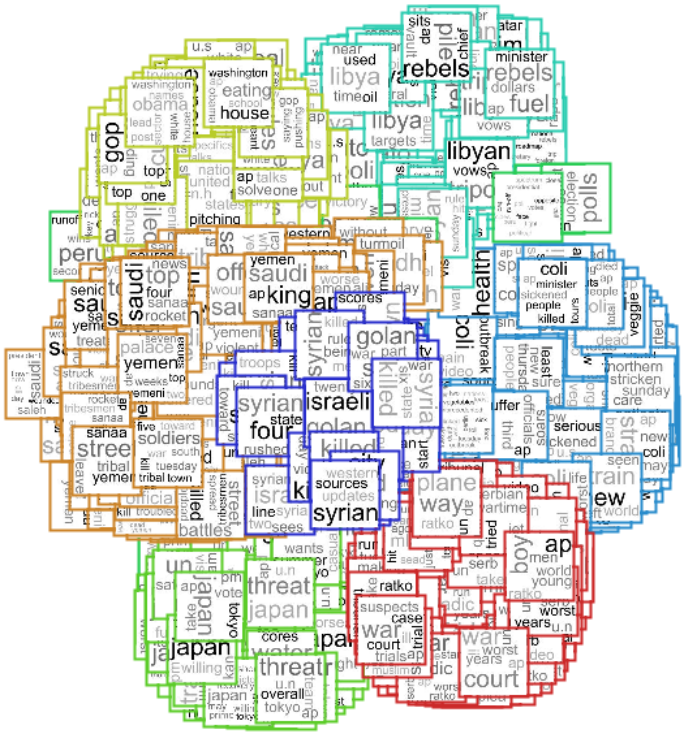
Realizamos análises detalhadas em dois grupos a partir da representação gráfica apresentada na Figura 33 (a). A abordagem proposta é empregada para mostrar as semelhanças entre os conteúdos dos documentos, conforme apresentado na Figura 34. O mapeamento da "Seleção A" é apresentado na Figura 34 (a) e a "Seleção B" na Figura 34 (b). Além disso, o conteúdo da "Seleção A" é apresentado na Figura 35 (a), relacionada a notícia *"European regions infected with e.coli"*; e o conteúdo da "Seleção B" é apresentado na Figura 35 (a), mostrando o tópico principal relacionado a notícia *"The nuclear accident in Fukushima, Japan"*.

Figura 33 – A análise do conjunto de dados **NEWS-8**: (a) mostra uma projeção $2D$ da coleção de documentos e (b) mostra a *tag cloud* de cada documento como marcador visual.

(a)



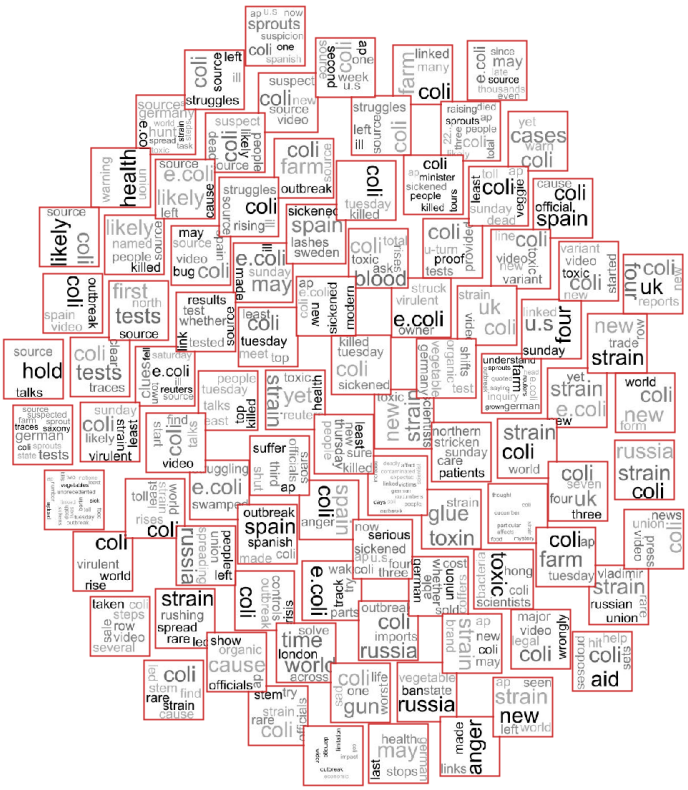
(b)



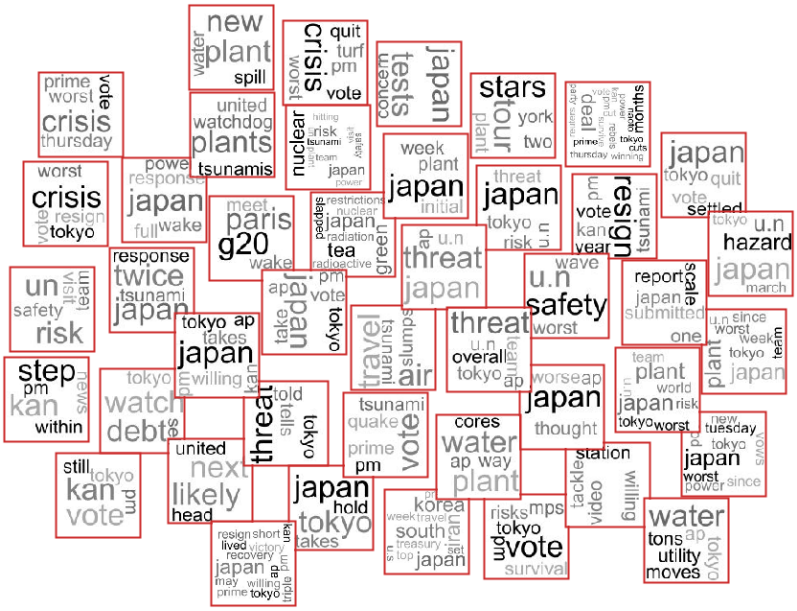
Fonte: Elaborado pelo autor.

Figura 34 – Análise do conjunto de dados **NEWS-8**: (a) mostra o zoom na Seleção A da Figura 33 e (b) mostra o zoom na Seleção B da Figura 33.

(a)



(b)



Fonte: Elaborado pelo autor.

Conforme apresentado nessas aplicações, o usuário pode usar a abordagem proposta

Figura 35 – Análise do conjunto de dados **NEWS-8**: (a) mostra algumas *tag clouds* da Seleção A da Figura 30 e (b) mostra algumas *tag clouds* da Seleção B da Figura 30.

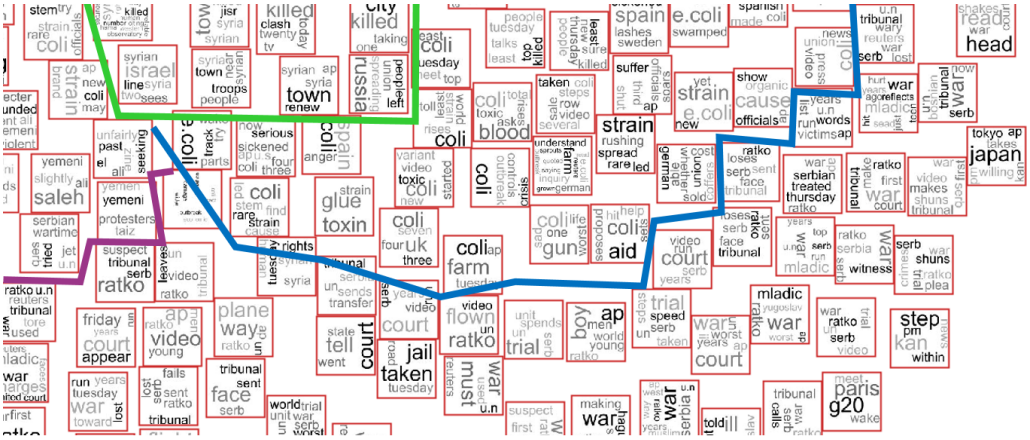
(a) (b)

Fonte: Elaborado pelo autor.

para observar as semelhanças e os conteúdos dos documentos. Além disso, nossa abordagem pode ser usada para auxiliar na compreensão da formação de agrupamentos. Por exemplo, conforme apresentado na Figura 34, os dois grupos são gerados porque os documentos têm uma ocorrência semelhante das mesmas palavras. Além disso, o relacionamento das instâncias também pode ser analisado observando a frequência das palavras de instâncias semelhantes.

A abordagem proposta também pode ser valiosa para explicar o processamento de similaridade entre os documentos apresentados no conjunto de dados analisado, ou seja, a compreensão *intra-cluster* (interna ao grupo) pode ser realizada observando as *tag clouds* geradas para cada instância. Geralmente, quando uma projeção é gerada, alguns conjuntos de documentos podem ser projetados próximos um do outro. Ao explorar um conjunto de dados rotulado, a fronteira entre cada grupo é facilmente observada; no entanto, a maioria dos conjuntos de dados não tem rótulo. Parte do conjunto de dados **NEWS-8** é apresentada na Figura 36, na qual omitimos as informações da classe para mostrar como é difícil observar a fronteira entre os grupos. Com base na análise individual, o usuário pode perceber a fronteira entre os grupos observando um resumo de cada documento por meio das *tag clouds*. Neste exemplo, desenhamos linhas para mostrar a fronteira entre cada grupo. Este exemplo mostra que uma análise individual pode oferecer vantagens na análise do grupo quando comparada com uma abordagem global, que mostra apenas a *tag cloud* computada para o grupo todo.

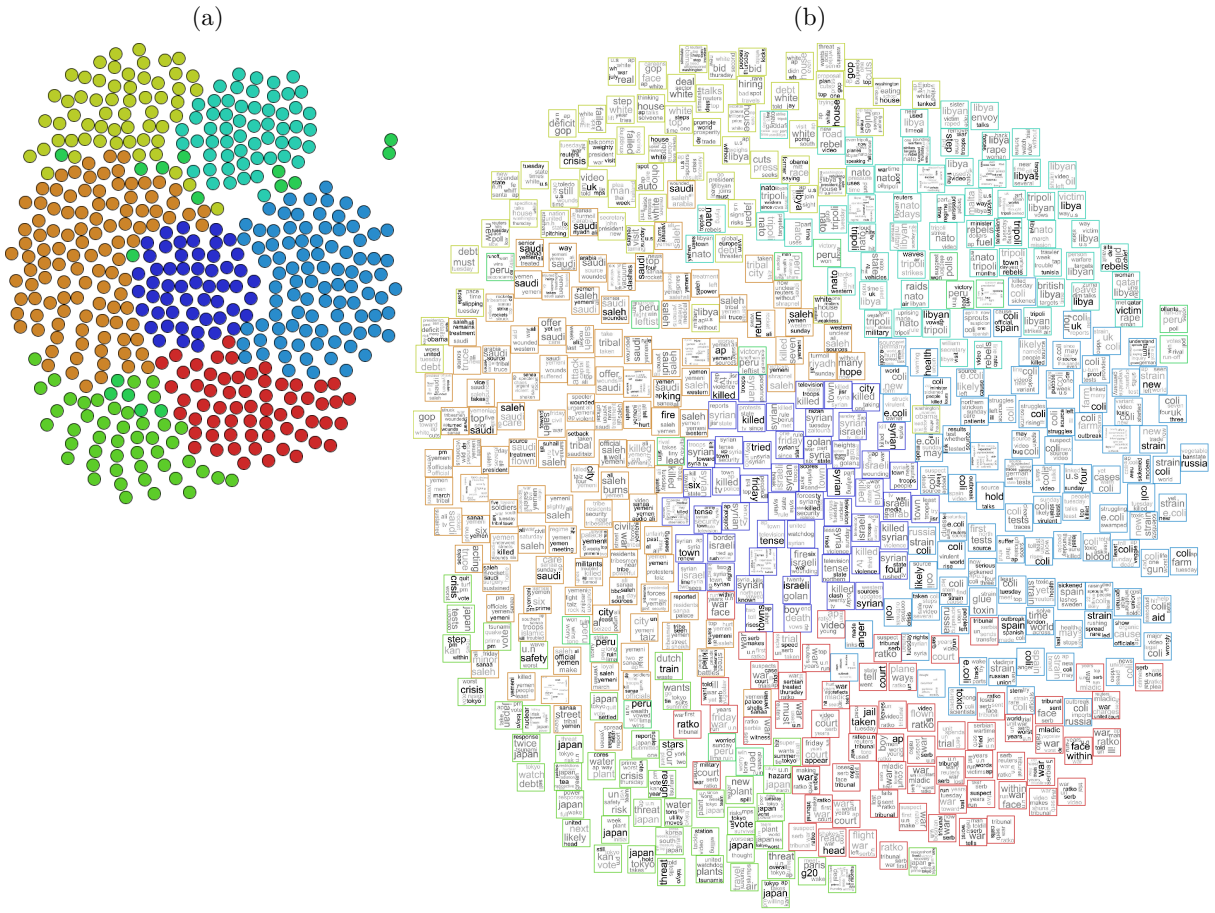
Figura 36 – Identificação de fronteira com base na inspeção individual das *tag clouds* computadas para cada instância.



Fonte: Elaborado pelo autor.

Um problema das técnicas de projeção multidimensional é a sobreposição entre marcadores quando a representação gráfica é gerada. Mesmo que o usuário possa ampliar o espaço projetado, a sobreposição ainda permanece. Para superar esse problema, nosso trabalho fornece algumas técnicas de remoção de sobreposição apresentadas na literatura; eles são *ProjSnippet* (G-NIETO et al., 2014), *Prism* (GANSNER; HU, 2010b), *VPSC* (DWYER; MARRIOTT; STUCKEY, 2006b) e *RWordle* (STROBELT et al., 2012b). Essas técnicas podem remover a sobreposição da imagem e preservar a semelhança entre os documentos. Por exemplo, como mostrado na Figura 37, empregamos a técnica RWordle para remover a sobreposição dos marcadores que se sobrepõem às projeções apresentadas na Figura 33. Assim, sem marcadores sobrepostos, o usuário pode navegar facilmente pela projeção.

Figura 37 – Técnica de remoção de sobreposição aplicada na projeção do conjunto de dados **NEWS-8** apresentado na Figura 33. A técnica *RWordle* removeu a sobreposição dos pontos (a) e imagens (b).



Fonte: Elaborado pelo autor.

4.4 CONSIDERAÇÕES FINAIS

Neste capítulo apresentamos uma técnica de visualização híbrida para auxiliar na organização, exploração e análise de coleções de documentos. A técnica proposta organiza uma coleção de documentos reunindo documentos semelhantes no espaço $2D$. O processo de exploração e análise também é aprimorado com ferramentas e representações gráficas para exibir o conteúdo dos documentos. Para isso, a abordagem proposta usa a técnica de projeção multidimensional para mostrar semelhanças entre documentos no espaço de $2D$. Além disso, uma técnica de *tag clouds* é empregada para mostrar o conteúdo do documento na mesma representação gráfica – a *tag cloud* de cada documento foi apresentada como marcador visual na representação gráfica, portanto, em uma única exibição, o usuário podia perceber os relacionamentos e conteúdos de toda a coleção.

Nas aplicações, usamos dois conjuntos de dados de notícias do mundo para mostrar a eficácia da abordagem proposta. Além disso, para auxiliar no processo exploratório, a abordagem proposta foi valiosa para auxiliar na compreensão da formação de grupos, ou seja, a abordagem poderia ser usada para mostrar por que alguns agrupamentos de documentos foram gerados e por que alguns documentos são semelhantes. Essa explicação de similaridade pode ser realizada observando a ocorrência de palavras apresentadas na *tag clouds* de cada documento.

A sobreposição é a principal limitação das técnicas de projeção multidimensional e é inevitável quando grandes conjuntos de dados são visualizados. Para superar esse problema, usamos técnicas de remoção de sobreposição e mecanismos de interação, como zoom e seleção. Entretanto, elas aumentam o tamanho da imagem necessária para criar a representação gráfica e também modificam as relações de similaridade geradas pelas técnicas de visualização. Para agregar novos mecanismos de exploração em grandes conjuntos de dados, apresentamos no próximo capítulo uma nova abordagem, na qual é considerada o uso de técnicas hierárquicas de visualização, visando minimizar o número de imagens apresentadas na representação gráfica.

5 ABORDAGEM DE EXPLORAÇÃO MULTINÍVEL

5.1 CONSIDERAÇÕES INICIAIS

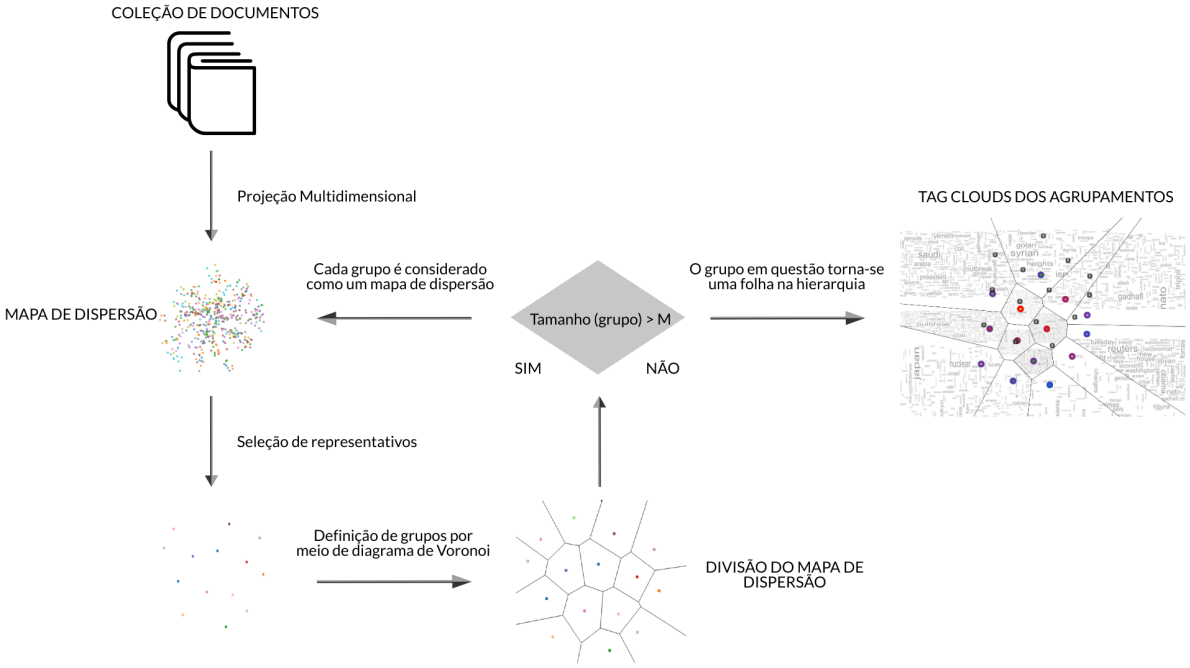
Por meio dos estudos realizados sobre as estratégias e problemas apresentados nos capítulos anteriores, visando reduzir o tamanho da imagem necessária para criar a representação gráfica e mantendo as relações de similaridade geradas pelas técnicas de visualização, com foco na abordagem de exploração multinível desenvolvida por Marcilio e Eler (2018), foi proposta uma expansão para suporte a dados textuais e acrescentada a metáfora visual para suporte a *tag clouds* projetadas como texturas, fornecendo uma visão geral dos agrupamentos de documentos analisados.

A abordagem de exploração multinível proposta neste trabalho toma como base conjuntos representados no plano $2D$, para isso qualquer técnica de projeção ou posicionamento de pontos no plano pode ser utilizada. Geralmente, a posição em que os pontos são representados no plano de projeção é baseada na similaridade do conteúdo dos documentos. 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 para codificar visualmente as fronteiras dos grupos (MARCILIO; ELER, 2018). 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. Neste trabalho, é apresentado uma aplicação utilizando documentos para a abordagem de exploração hierarquica e, sendo assim, *tag clouds* são utilizadas como *background* para oferecer uma visão geral do conteúdo do grupo, isto é, a *tag cloud* formada pelos termos de todos os documentos pertencentes ao agrupamento criado. Tal processo é apresentado na Figura 38.

5.2 REDUÇÃO DE DIMENSIONALIDADE E SELEÇÃO DE REPRESENTATIVOS

Após a etapa de projeção do mapa de dispersão o trabalho de Marcilio e Eler (2018) realiza a seleção de representativos. Na Figura 39 este processo é ilustrado, sendo que a partir de um espaço de características de um conjunto de dados, por exemplo, uma coleção de documentos, é realizada a projeção multidimensional e finalmente a seleção de representativos, onde as instâncias representativas estão destacadas na cor vermelha. A

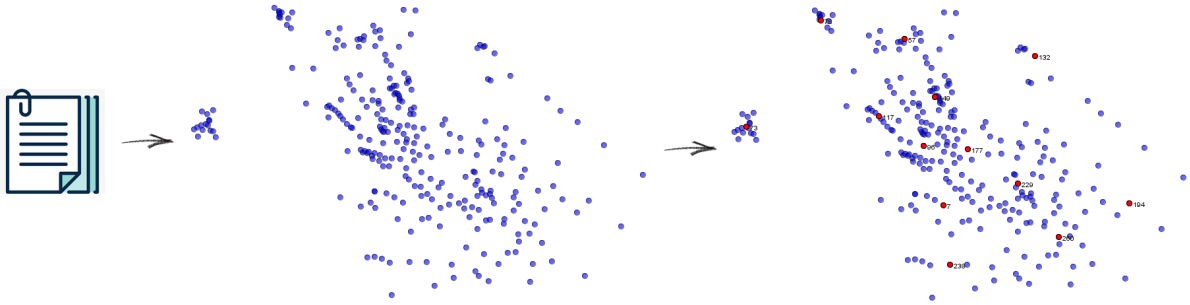
Figura 38 – Esquema geral da técnica de exploração.



Fonte: Elaborado pelo autor.

projeção multidimensional pode ser feita com qualquer técnica, visto que a abordagem desenvolvida exige somente uma nuvem de pontos no plano.

Figura 39 – 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.

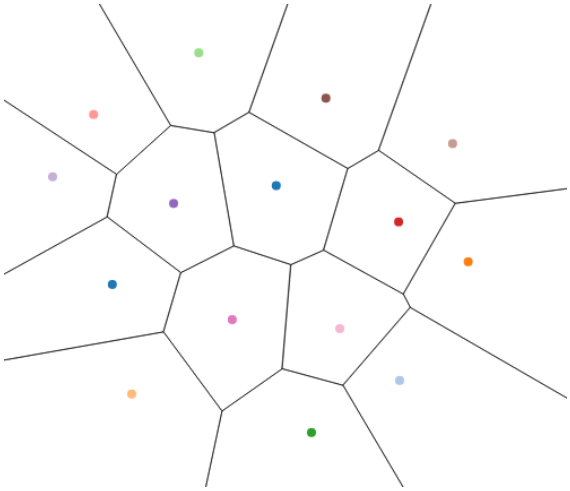


Fonte: Elaborado pelo autor.

5.3 DEFINIÇÃO DA HIERARQUIA

Após as etapas de redução de dimensionalidade e seleção de representativos sobre o conjunto de dados de entrada é necessário realizar a criação de níveis na projeção para que a exploração hierárquica dos dados possa ser realizada. Sendo assim, o processo de seleção de representativos deve ser aplicado nos grupos resultantes da seleção de representativos do primeiro nível. Posteriormente, utilizando o domínio de Voronoi, sendo que, 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 $\forall y_k \in Y$, $d(y_j, x_i) \leq d(y_k, x_i)$. Este efeito pode ser visualizado na Figura 40.

Figura 40 – Divisão do espaço imposta pela seleção de representativos.

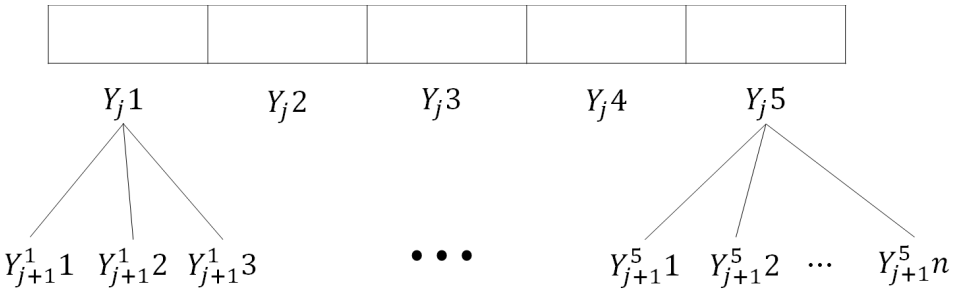


Fonte: Marcilio e Eler (2018).

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 41 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 42. Note que existem dois casos em que se deve aplicar a união dos grupos. No primeiro (ver Figura 42a), 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.

Figura 41 – 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.

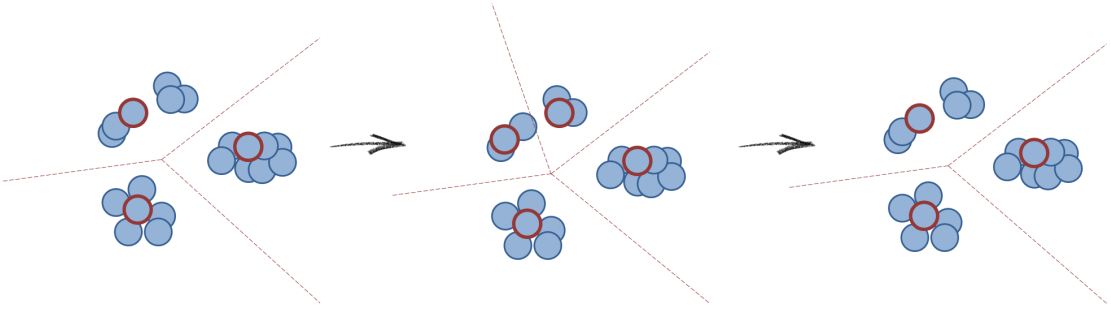


Fonte: Marcilio e Eler (2018).

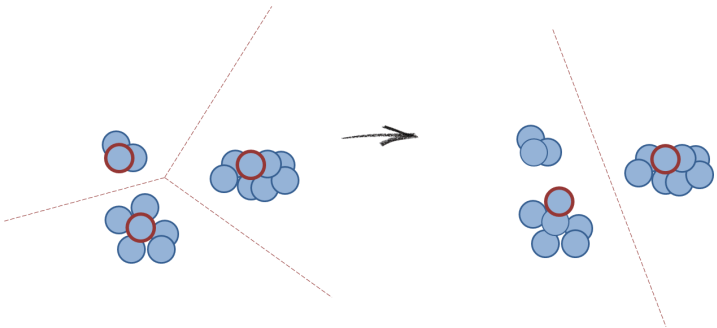
No segundo caso (ver Figura 42b), a partição imposta pela seleção de representativos faz com que um nó tenha um número de instâncias menor que o permitido. Assim, o nó com menor quantidade de elementos é agrupado com o elemento mais próximo, considerando os representativos dos nós. Após o processo de definição da hierarquia, conforme a interação com o usuário, é efetuada a remoção de sobreposição.

Figura 42 – Processo de união dos nós com quantidade inferior à M instâncias.

(a) Caso 1.



(b) Caso 2.



Fonte: Marcilio e Eler (2018).

5.4 ABORDAGEM DE EXPLORAÇÃO

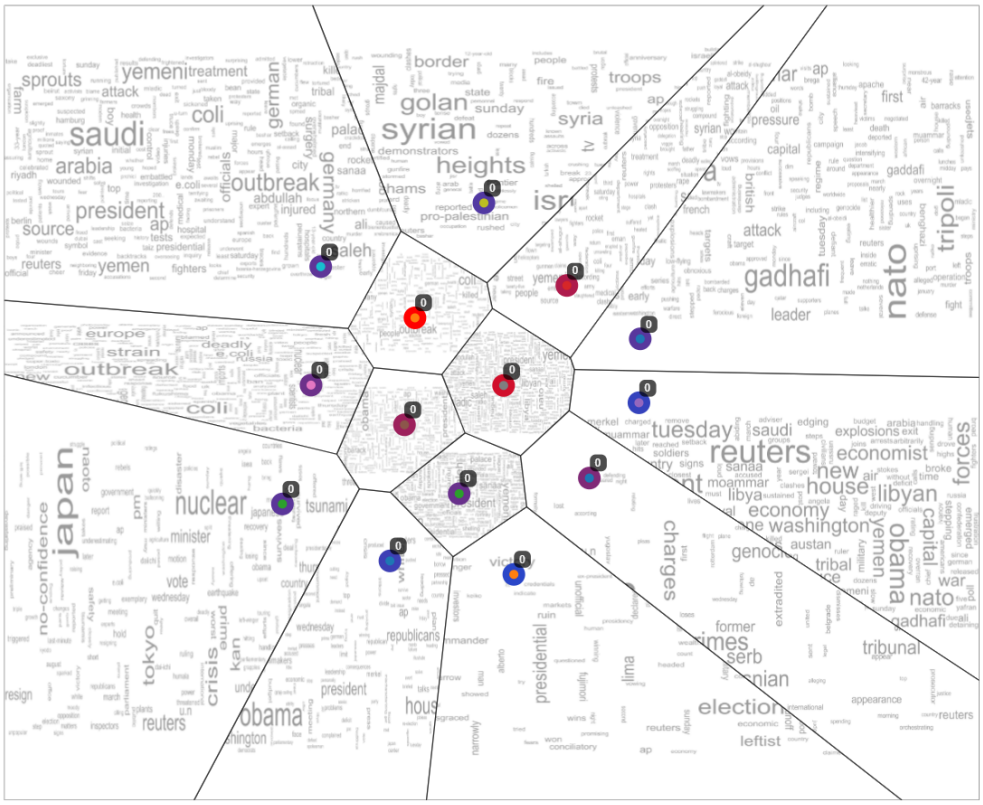
Para o desenvolvimento da aplicação, que teve sua origem por meio dos estudos descritos no presente trabalho, foi utilizado o *framework* D3.js para a projeção de imagens no plano. O usuário deve especificar um conjunto de pontos no espaço multidimensional para ser projetado ou fornecer um conjunto de pontos no plano, de modo que seja possível a utilização do algoritmo de seleção de representativos e iniciar a construção da hierarquia. O usuário também deve fornecer o conjunto de textos que originou este conjunto de pontos.

No trabalho apresentado por Marcilio e Eler (2018), para exploração de um conjunto de imagens, era necessário inserir uma lista de *urls* das mesmas, porém na extensão desenvolvida neste trabalho a própria aplicação gera estas imagens, sendo imagens que representam as instâncias dos agrupamentos por meio de *tag clouds*. Na Figura 43 é apresentado um exemplo de projeção com o primeiro nível da hierarquia. Observe que as *tag clouds* são utilizadas como textura de cada grupo, isto é, os termos ali presentes expressam a frequência de ocorrência nas instâncias de documentos de cada agrupamento. Foram utilizadas cores nos círculos concêntricos que envolvem as instâncias representativas, sendo que a cor do círculo mais externo do marcador fornece uma escala do número de instâncias pertencentes ao grupo, onde círculos que possuem cores mais próximas ao vermelho representam grupos com um maior número de instâncias, ou seja, agrupamentos onde existem bastantes documentos com uma certa similaridade de seus conteúdos textuais. E as cores dos círculos internos, representam a codificação das classes das instâncias. E finalmente, os números acima de cada marcador representam a quantidade de instâncias que foram selecionadas pelo usuário durante o processo de exploração, sendo que o mesmo pode realizar esta seleção ao atingir os nós folhas da hierarquia.

Quando o cursor do mouse é posicionado sobre um representativo, uma janela é exibida com algumas informações sobre o grupo em questão conforme Figura 44. Na janela é possível verificar a distribuição de classes das instâncias em um componente visual de *stackedbar* a direita, em que as classes são codificadas pelas cores e o número de instâncias pela área de cada retângulo. Também é possível observar as *tag clouds* geradas para as instâncias similares e diversas selecionadas utilizando os algoritmos *KNN* (COVER; HART, 2006) e *BRID* (SANTOS, 2012), respectivamente. E finalmente, com um maior destaque é exibida a *tag cloud* representativa gerada a partir de todas as instâncias do agrupamento.

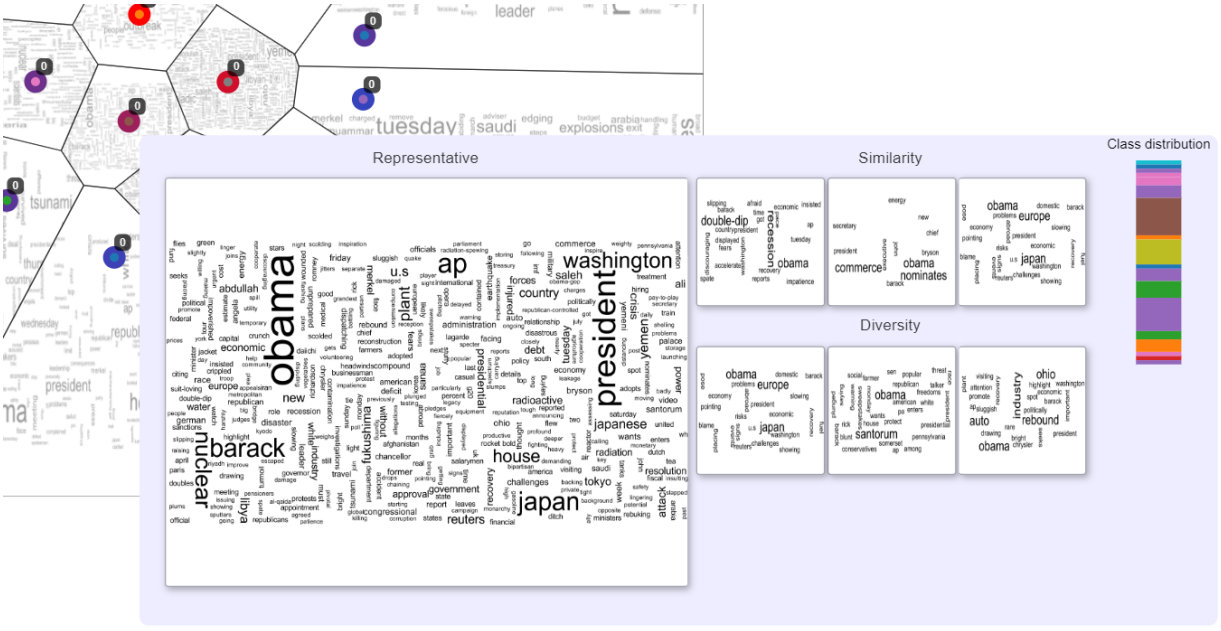
Ao clicar sobre a instância representativa, o processo de expansão é realizado, efetuando então um *zoom in* com valor proporcional ao número de elementos do grupo sendo analisado. Para que as instâncias que estavam posicionadas no nível anterior não sejam projetadas para fora da área de visualização, tais instâncias são posicionadas nas bordas do plano para que o contexto seja mantido. Tal cenário de interação pode ser verificado na Figura 45. Também é possível aglomerar um grupo que foi expandido

Figura 43 – Primeiro nível da abordagem de exploração.



Fonte: Elaborado pelo autor.

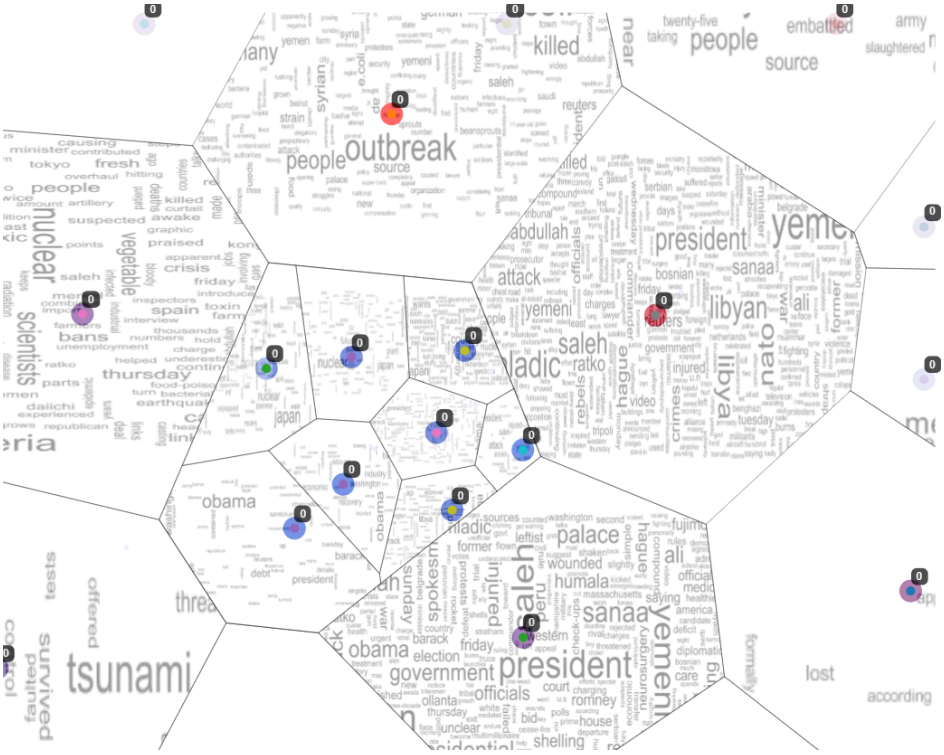
Figura 44 – Janela com informações sobre o grupo analisado.



Fonte: Elaborado pelo autor.

anteriormente, efetuando um *zoom out* e otimizando o espaço visual.

Figura 45 – Expansão de um nó. Como algumas instâncias iriam ser projetadas fora da área de visualização, tais instâncias são posicionadas nas bordas da imagem.



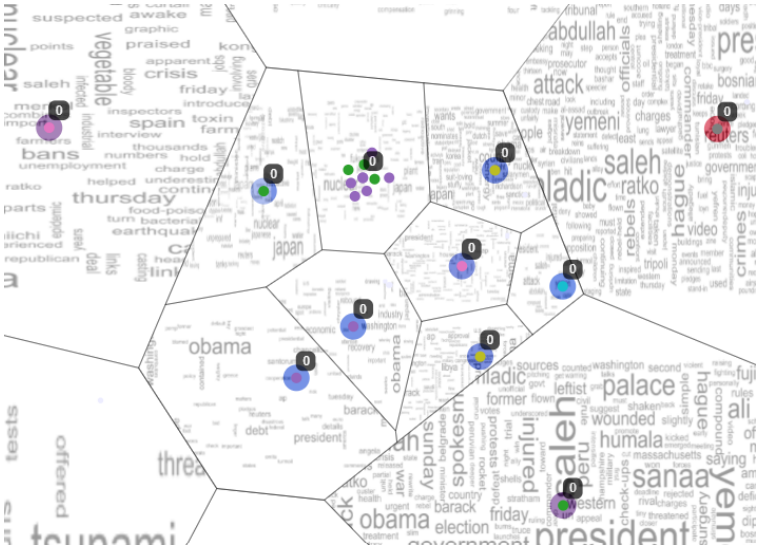
Fonte: Elaborado pelo autor.

Ao atingir um grupo folha, as instâncias são representadas individualmente no plano quando o cursor do mouse está sobre o marcador do representativo conforme Figura 46a. Ao clicar sobre o representativo, neste nível, é possível visualizar as *tag clouds* de cada instância folha em uma janela separada representada na Figura 46b.

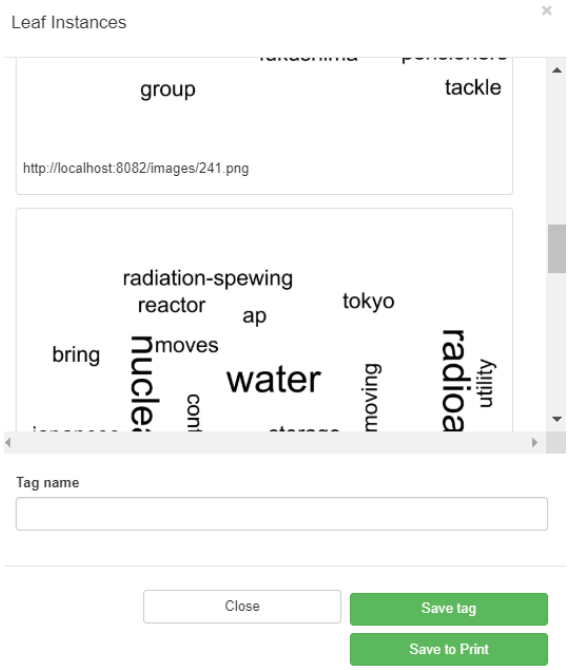
Observa-se na Figura 46b uma área para definição de *tags* que podem ser associadas às *tag clouds* selecionadas e podem servir como uma maneira de organizar o conjunto de documentos sendo explorado. Também é possível selecionar *tag clouds* para uma análise mais detalhada.

Figura 46 – Visualização das instâncias no último nível da hierarquia.

(a) Instâncias projetadas no último nível.



(b) Imagens correspondentes as intâncias projetadas.

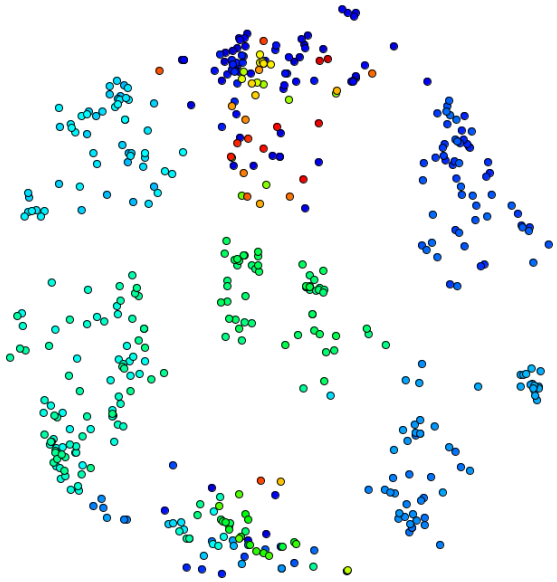


Fonte: Elaborado pelo autor.

5.5 RESULTADOS

Neste estudo de caso é apresentado a exploração de um conjunto de 495 extratos de notícias publicadas na internet pelas agências Reuters, BBC, CNN e *Associated Press*. Os arquivos foram recuperados dos *websites* das agências, e convertidos para o formato *txt*. A projeção inicial foi realizada com a técnica IDMAP (MINGHIM; PAULOVICH; LOPES, 2006) e a medida de similaridade utilizada para calcular a matriz de distância foi a distância do cosseno, enquanto que a seleção de representativos foi realizada com a técnica *Affinity Propagation*(FUJIWARA; IRIE; KITAHARA, 2011). A Figura 47 exibe um exemplo da forma clássica de projeção multidimensional para este conjunto de notícias e parâmetros descritos anteriormente.

Figura 47 – Projeção do conjunto de 495 extratos de notícias.



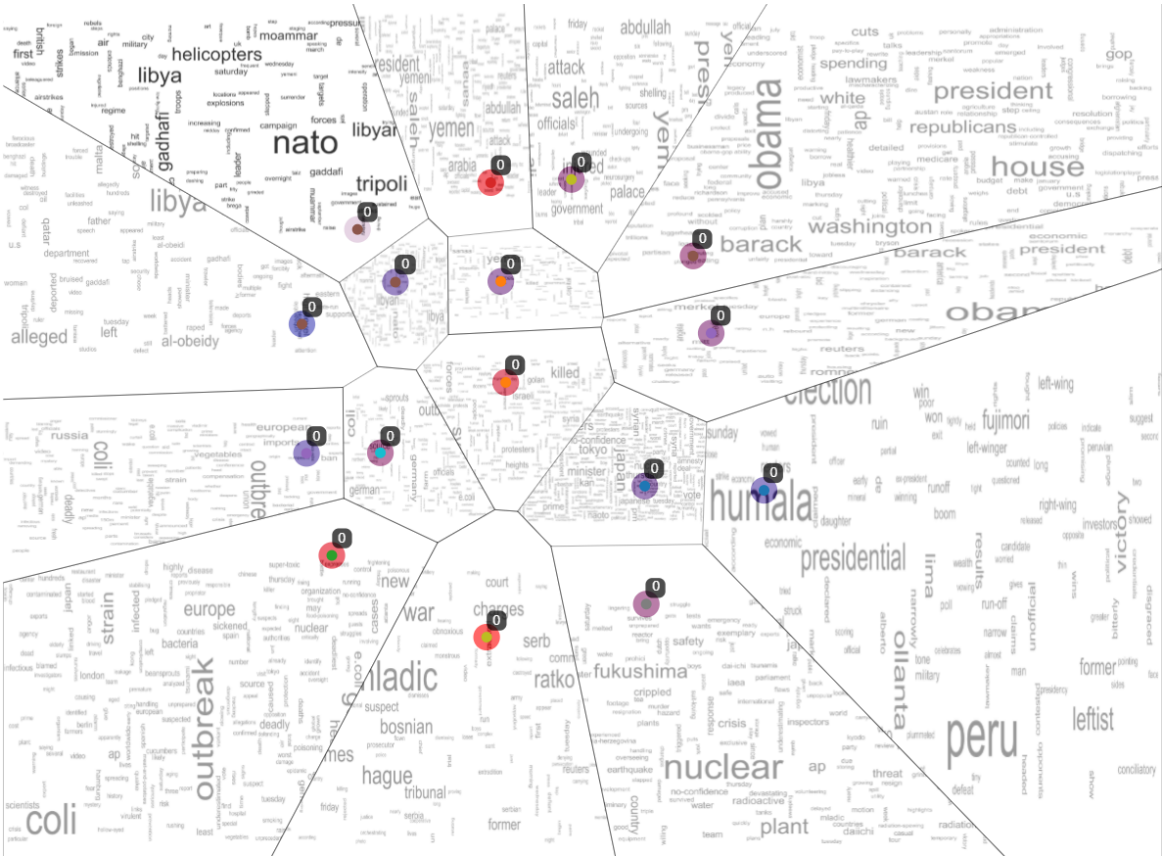
Fonte: Elaborado pelo autor.

Na Figura 48 é apresentado o primeiro nível da hierarquia proposta neste trabalho. Note que ao passar o mouse por cima dos representativos, *tooltips* são apresentados para melhor investigação do grupo, como apresentado nas Figuras 48a e 48b.

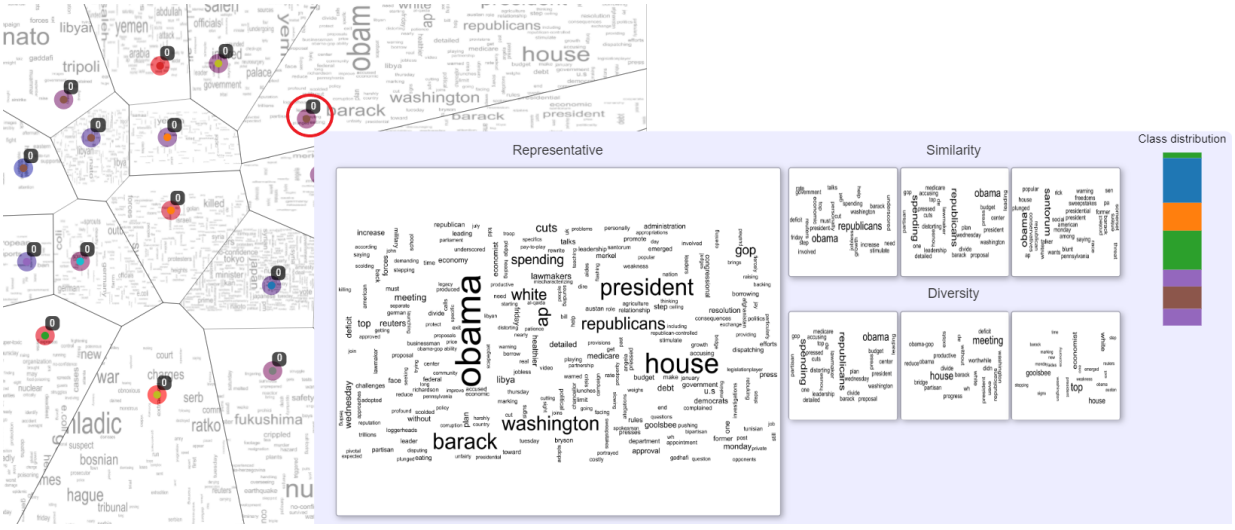
Na Figura 49 é apresentada a expansão do nó destacado com o círculo vermelho na Figura 48b, em que quatro novos nós foram gerados, também circulados em vermelho nesta Figura. O representativo destacado em verde contém seis instâncias, projetadas conforme apresentado na Figura 50. Ao clicar nesse representativo, as *tag clouds* correspondentes as instâncias (documentos) que estão na folha da hierarquia são apresentadas para possível definição de *tags* ou para que sejam selecionadas, como exibido na Figura 51, onde foram geradas *tag clouds* representativas utilizando os termos pertencentes a cada notícia individualmente.

Figura 48 – Primeiro nível da abordagem de exploração multinível aplicada em uma coleção de notícias.

(a) Visão geral da hierarquia.



(b) Apresentação de informações de um agrupamento.



Fonte: Elaborado pelo autor.

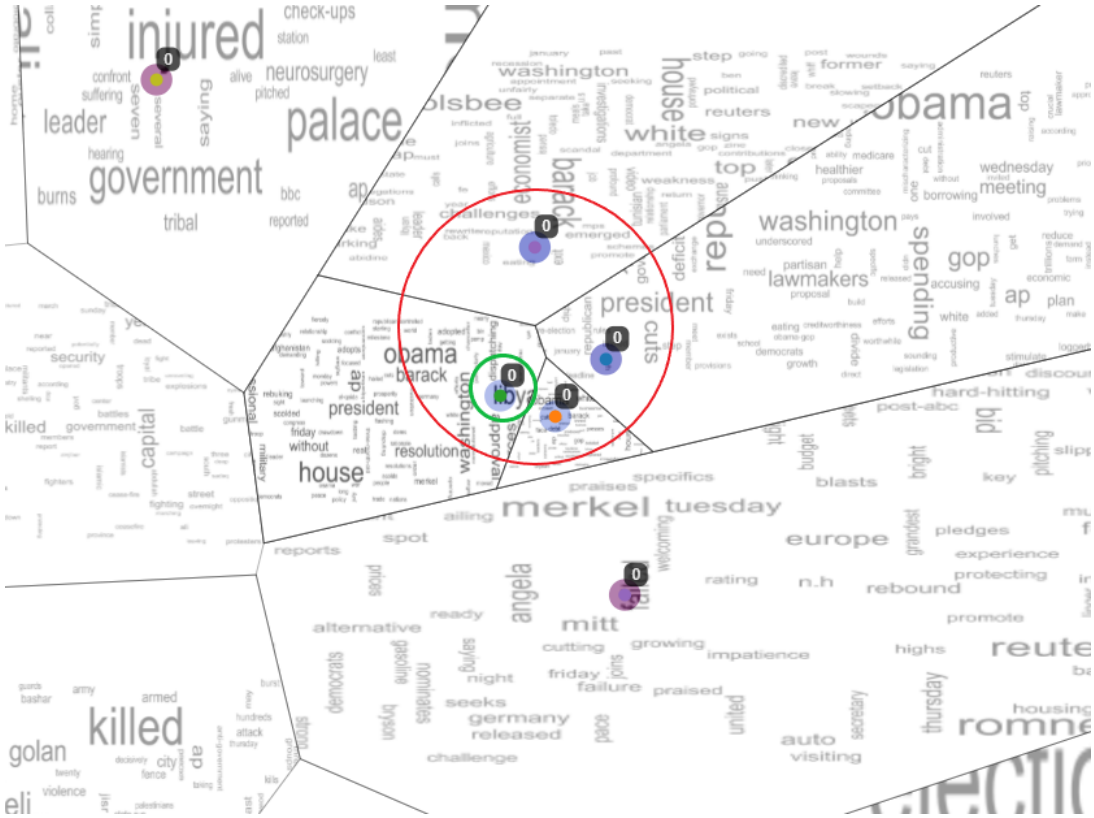
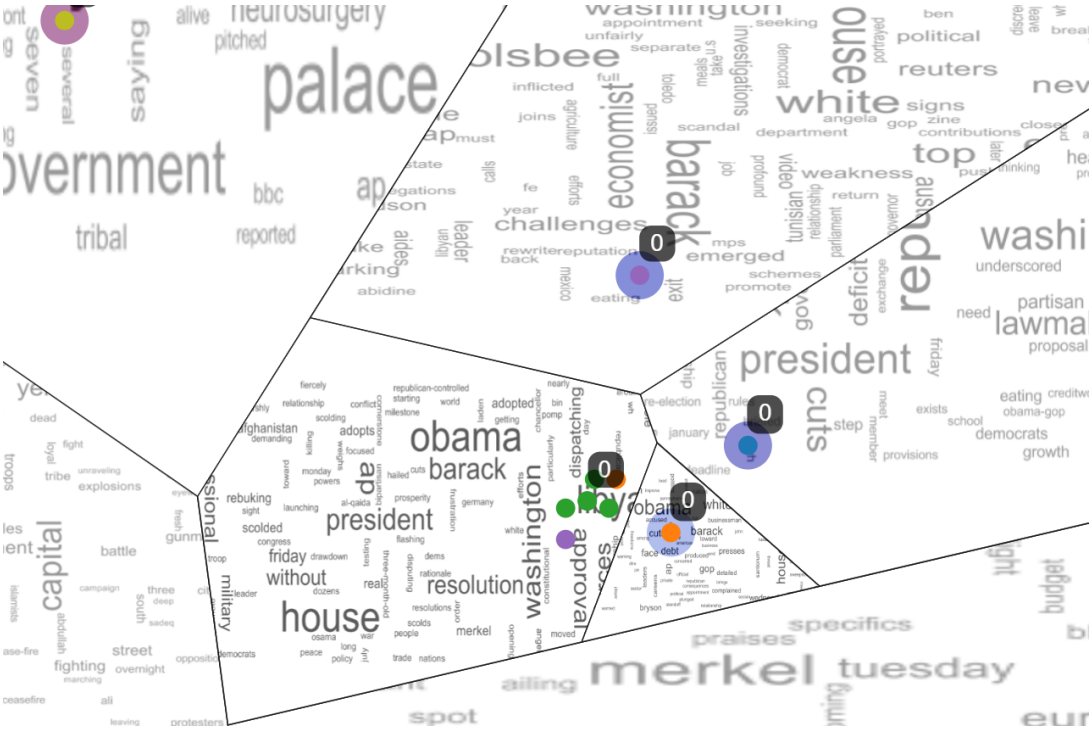


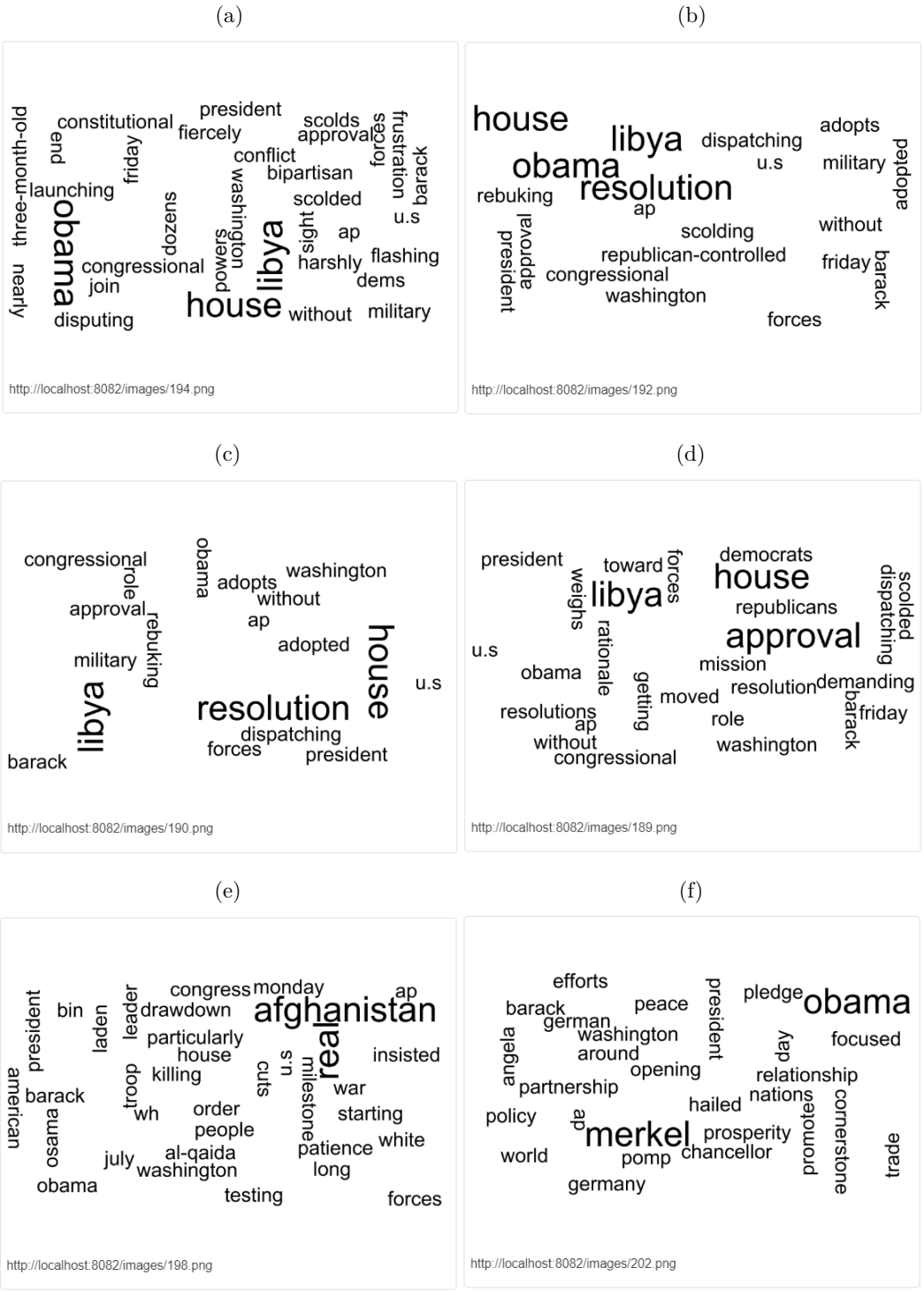
Figura 49 – Nós gerados a partir da expansão do nó destacado na Figura 48b.

Figura 50 – Instâncias pertencentes ao nó projetadas no plano.



Fonte: Elaborado pelo autor.

Figura 51 – Instâncias folhas da hierarquia representadas pelas *tag clouds* dos termos pertencentes a cada notícia isoladamente.



Fonte: Elaborado pelo autor.

5.6 CONSIDERAÇÕES FINAIS

Neste capítulo apresentamos uma abordagem de exploração multinível para suporte a dados textuais e a utilização de *tag clouds* para preencher as texturas de agrupamentos formados pelo domínio de Voronoi. Esta abordagem agregou novos mecanismos de exploração focados na resolução das limitações apresentadas na abordagem do capítulo anterior. Em seguida, foi apresentado um estudo de caso onde foi possível verificar o desempenho dos algoritmos em relação a disposição dos agrupamentos gerados, e utilizando *tag clouds* ficou claro o assunto principal em cada agrupamento e em cada instância projetada. Também foi possível perceber que a abordagem diminuiu a desordem visual da forma tradicional de uma técnica de projeção multidimensional.

A abordagem apresentada neste capítulo teve por objetivo fornecer melhorias na exploração proposta por Marcilio e Eler (2018), realizando uma expansão para suporte a exploração e processamento de dados textuais, seguindo 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.

6 CONCLUSÕES E TRABALHOS FUTUROS

Dentre os métodos existentes para criar representações de similaridade entre documentos, as técnicas de projeções multidimensionais são possivelmente as mais comuns. Quando o espaço é reduzido para o bidimensional, as visualizações geradas por essas técnicas representam os documentos por meio de "mapas de documentos", onde os documentos são representados por marcadores e seu posicionamento reflete a similaridade de conteúdo de seus documentos correspondentes. No entanto, a dificuldade em analisar o conjunto é proporcional ao número de documentos sendo explorados, sendo assim, o principal objetivo deste trabalho foi facilitar o processo exploratório e diminuir os problemas relacionados à escalabilidade visual presentes na forma tradicional de projeções multidimensionais.

6.1 CONCLUSÕES

Este trabalho apresentou duas soluções para auxiliar o processo de exploração de conjuntos de documentos. Essas soluções utilizaram projeções multidimensionais para representar esses documentos num espaço bidimensional, utilizando extração de características dos termos que os compõem. Adicionalmente, elas também se apoiaram no uso de *tag clouds* como marcador visual nas representações gráficas. Neste sentido, duas técnicas de visualização de informação foram desenvolvidas. Essas técnicas permitem a análise exploratória do conjunto de dados, aplicando o mantra visual de busca de informações "*Overview first, zoom and filter, then details-on-demand*" de Shneiderman (1996), em que o usuário recebe poucas informações durante o processo exploratório e detalha o conteúdo de acordo com a demanda.

As principais contribuições deste trabalho foram as duas técnicas exploratórias criadas para apoiar a exploração de conjuntos de dados textuais, utilizando um *design* gráfico interativo para navegação. As duas abordagens também foram desenvolvidas de modo que seja adaptável quanto ao uso de diferentes técnicas de projeções multidimensionais ou de geração de *tag clouds*.

Na literatura, há inúmeras aplicações que tratam dados textuais e criam visualizações para exploração dos mesmos. No entanto, poucas são a que focam na exploração sobre demanda de coleções de documentos e com o desenvolvimento deste trabalho foi possível notar que técnicas que permitem a interação do usuário garantem uma melhoria em relação a desordem visual presente em métodos clássicos de projeções multidimensionais, e desta forma facilita o processo de análise por parte do usuário. Além disso, o uso de

tag clouds pôde facilitar a análise de agrupamentos, sendo que a *tag cloud* calculada para cada documento pôde revelar os principais termos dos dados textuais, otimizando o entendimento a cerca dos grupos criados.

6.2 TRABALHOS FUTUROS

Após o desenvolvimento das duas técnicas de exploração textual foi possível notar que uma limitação, das duas abordagens desenvolvidas, é a necessidade da utilização de técnicas de mineração de texto e de geração de *tag clouds* que sejam eficientes, uma vez que as *tag clouds* são geradas conforme o usuário interage com a visualização, em tempo de execução. Nos estudos de casos apresentados, não utilizamos grandes coleções de documentos para que fosse possível exemplificar e facilitar a compreensão em relação as instâncias representadas pelos agrupamentos, logo um estudo de caso com um conjunto maior de documentos poderia ser analisado para identificação de técnicas eficientes.

Como trabalhos futuros, na projeção hierárquica desenvolvida, poderia ser utilizada ou adaptada uma técnica de *tag cloud* que verificasse as fronteiras dos grupos dos diagramas de Voronoi, e gerasse as *tag clouds* respeitando o formato desses grupos, eliminando as distorções e sobreposições dos termos ali gerados, aproveitando ainda mais o espaço visual disponível. Uma extensão no último nível da hierarquia poderia também ser realizada para que o usuário pudesse acessar na íntegra o conteúdo textual das instâncias de um grupo folha.

REFERÊNCIAS

ALENCAR, A. B. *Visualização da evolução temporal de coleções de artigos científicos*. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2013. Citado 2 vezes nas páginas 21 e 24.

ALENCAR, A. B.; FERREIRA, M. C. de O.; PAULOVICH, F. V. Seeing beyond reading: a survey on visual text analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 2, n. 6, p. 476–492, 2012. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1071>. Citado 2 vezes nas páginas 15 e 16.

ANDREOTTI, A. D. et al. Análise visual da evolução de coleções de documentos utilizando tag cloud. *Colloquium Exactarum. ISSN: 2178-8332*, v. 9, n. 2, p. 15–32, set. 2017. Disponível em: <http://journal.unoeste.br/index.php/ce/article/view/1607>. Citado na página 16.

ANDREOTTI, A. L. D.; SILVA, L. F.; ELER, D. M. Hybrid visualization approach to show documents similarity and content in a single view. *Information*, v. 9, n. 6, 2018. ISSN 2078-2489. Disponível em: <https://www.mdpi.com/2078-2489/9/6/129>. Citado na página 45.

BALZER, M.; DEUSSEN, O.; LEWERENTZ, C. Voronoi treemaps for the visualization of software metrics. In: *Proceedings of the 2005 ACM Symposium on Software Visualization*. New York, NY, USA: ACM, 2005. (SoftVis '05), p. 165–172. ISBN 1-59593-073-6. Disponível em: <http://doi.acm.org/10.1145/1056018.1056041>. Citado na página 26.

BECKS, A.; SEELING, C.; MINKENBERG, R. Benefits of document maps for text access in knowledge management: A comparative study. In: *Proceedings of the 2002 ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2002. (SAC '02), p. 621–626. ISBN 1-58113-445-2. Disponível em: <http://doi.acm.org/10.1145/508791.508912>. Citado na página 16.

BERGER, M.; MCDONOUGH, K.; SEVERSKY, L. M. cite2vec: Citation-driven document exploration via word embeddings. *IEEE Transactions on Visualization and Computer Graphics*, v. 23, n. 1, p. 691–700, Jan 2017. ISSN 1077-2626. Citado 3 vezes nas páginas 34, 35 e 44.

CHEN, M. et al. Data, information, and knowledge in visualization. *IEEE Computer Graphics and Applications*, v. 29, n. 1, p. 12–19, Jan 2009. ISSN 0272-1716. Citado 3 vezes nas páginas viii, 20 e 21.

CHI, M. T. et al. Morphable word clouds for time-varying text data visualization. *IEEE Transactions on Visualization and Computer Graphics*, v. 21, n. 12, p. 1415–1426, Dec 2015. ISSN 1077-2626. Citado na página 29.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27, January 1967. ISSN 0018-9448. Citado na página 27.

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 62.

CUI, W. et al. Context-preserving, dynamic word cloud visualization. *IEEE Computer Graphics and Applications*, v. 30, n. 6, p. 42–53, Nov 2010. ISSN 0272-1716. Citado 3 vezes nas páginas x, 38 e 39.

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 2 vezes nas páginas 17 e 27.

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 na página 55.

ELER, D. M.; GARCIA, R. E. Using otsu's threshold selection method for eliminating terms in vector space model computation. In: *International Conference on Information Visualization*. [S.l.]: IEEE Computer Society, 2013. p. 220–226. Citado na página 49.

ELER, D. M. et al. Simplified stress and simplified silhouette coefficient to a faster quality evaluation of multidimensional projection techniques and feature spaces. In: *International Conference on Information Visualization*. [S.l.]: IEEE Computer Society, 2015. p. 133–139. Citado na página 49.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Advances in knowledge discovery and data mining. In: FAYYAD, U. M. et al. (Ed.). Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. cap. From Data Mining to Knowledge Discovery: An Overview, p. 1–34. ISBN 0-262-56097-6. Disponível em: <http://dl.acm.org/citation.cfm?id=257938.257942>. Citado na página 19.

FUJIWARA, Y.; IRIE, G.; KITAHARA, T. Fast algorithm for affinity propagation. In: . IJCAI/AAAI, 2011. p. 2238–2243. ISBN 978-1-57735-516-8. Disponível em: <http://dblp.uni-trier.de/db/conf/ijcai/ijcai2011.html#FujiwaraIK11>. Citado na página 66.

G-NIETO, E. et al. Similarity preserving snippet-based visualization of web search results. *IEEE Transactions on Visualization and Computer Graphics*, v. 20, n. 3, p. 457–470, 2014. Citado na página 55.

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 2 vezes nas páginas 17 e 27.

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 na página 55.

GOMEZ-NIETO, E. et al. Similarity preserving snippet-based visualization of web search results. *TVCG*, v. 20, n. 3, p. 457–470, 2014. Citado 2 vezes nas páginas 17 e 27.

GRIES, S. T. *Quantitative Corpus Linguistics with R*. [S.l.]: Routledge, 2009. Citado na página 15.

HEIMERL, F. et al. Docucompass: Effective exploration of document landscapes. In: ANDRIENKO, G. L.; LIU, S.; STASKO, J. T. (Ed.). *VAST*. IEEE Computer Society, 2016. p. 11–20. ISBN 978-1-5090-5661-3. Disponível em: <http://dblp.uni-trier.de/db/conf/ieeevast/ieeevast2016.html#HeimerlJHKE16>. Citado 3 vezes nas páginas 36, 37 e 44.

HEIMERL, F. et al. Word cloud explorer: Text analytics based on word clouds. In: *2014 47th Hawaii International Conference on System Sciences*. [S.l.: s.n.], 2014. p. 1833–1842. ISSN 1530-1605. Citado 3 vezes nas páginas 28, 29 e 30.

HU, M.; WONGSUPHASAWAT, K.; STASKO, J. Visualizing social media content with sententree. *IEEE Transactions on Visualization and Computer Graphics*, v. 23, n. 1, p. 621–630, Jan 2017. ISSN 1077-2626. Citado na página 38.

JOBS, S. Stay hungry. stay foolish. 2005. Citado na página v.

JONES, K. S. Document retrieval systems. In: WILLETT, P. (Ed.). London, UK, UK: Taylor Graham Publishing, 1988. cap. A Statistical Interpretation of Term Specificity and Its Application in Retrieval, p. 132–142. ISBN 0-947568-21-2. Disponível em: <http://dl.acm.org/citation.cfm?id=106765.106782>. Citado 2 vezes nas páginas 23 e 24.

JUSUFI, I. et al. Visual exploration of relationships between document clusters. In: INSTICC. *Proceedings of the 5th International Conference on Information Visualization Theory and Applications - Volume 1: IVAPP, (VISIGRAPP 2014)*. [S.l.]: SciTePress, 2014. p. 195–203. ISBN 978-989-758-005-5. Citado 3 vezes nas páginas ix, 35 e 36.

KIM, M. et al. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE Transactions on Visualization and Computer Graphics*, v. 23, n. 1, p. 151–160, Jan 2017. ISSN 1077-2626. Citado 4 vezes nas páginas x, 39, 40 e 44.

LUHN, H. P. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, IBM Corp., Riverton, NJ, USA, v. 2, n. 2, p. 159–165, abr. 1958. ISSN 0018-8646. Disponível em: <http://dx.doi.org/10.1147/rd.22.0159>. Citado 3 vezes nas páginas 22, 23 e 31.

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 na página 15.

MARCILIO, W. E.; ELER, D. M. *Uma abordagem de exploração multinível em visualizações geradas para representar projeções multidimensionais*. Dissertação (Mestrado) — Universidade Estadual Paulista "Júlio de Mesquita Filho", 2018. Disponível em: <http://hdl.handle.net/11449/180312>. Citado 10 vezes nas páginas viii, 17, 26, 27, 30, 58, 60, 61, 62 e 70.

MINGHIM, R.; PAULOVICH, F. V.; LOPES, A. A. Content-based text mapping using multidimensional projections for exploration of document collections. *SPIE Proceedings: Visualization and Data Analysis*, v. 6060, 2006. Citado 5 vezes nas páginas 15, 31, 32, 44 e 66.

PAULOVICH, F. V. *Mapeamento de dados multi-dimensionais - integrando mineração e visualização*. Tese (Doctoral Thesis in Ciências de Computação e Matemática Computacional) — Instituto de Ciências Matemáticas e de Computação, University of São Paulo., 2008. Citado na página 19.

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 4 vezes nas páginas 15, 25, 26 e 49.

PAULOVICH, F. V.; SILVA, C. T.; NONATO, L. G. Two-phase mapping for projecting massive data sets. *IEEE Transactions on Visualization and Computer Graphics*, v. 16, n. 6, p. 1281–1290, Nov 2010. ISSN 1077-2626. Citado na página 25.

PAULOVICH, F. V. et al. Semantic wordification of document collections. *Computer Graphics Forum*, Blackwell Publishing Ltd, v. 31, n. 3pt3, p. 1145–1153, 2012. ISSN 1467-8659. Disponível em: <http://dx.doi.org/10.1111/j.1467-8659.2012.03107.x>. Citado 3 vezes nas páginas 40, 41 e 42.

PEREZ-MESSINA, I.; GUTIERREZ, C.; GRAELLS-GARRIDO, E. Organic visualization of document evolution. *CoRR*, abs/1712.06179, 2017. Disponível em: <http://arxiv.org/abs/1712.06179>. Citado 2 vezes nas páginas 32 e 33.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 24, n. 5, p. 513–523, ago. 1988. ISSN 0306-4573. Disponível em: <http://dx.doi.org/10.1016/0306-4573(88)90021-0>. Citado na página 31.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. *Commun. ACM*, ACM, New York, NY, USA, v. 18, n. 11, p. 613–620, nov. 1975. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/361219.361220>. Citado 2 vezes nas páginas 22 e 23.

SANTOS, L. F. D. *Explorando variedade em consultas por similaridade*. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2012. Citado 2 vezes nas páginas 27 e 62.

SHNEIDERMAN, B. The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings 1996 IEEE Symposium on Visual Languages*. [S.l.: s.n.], 1996. p. 336–343. ISSN 1049-2615. Citado 2 vezes nas páginas 20 e 71.

SIIRTOLA, H. et al. Text variation explorer: Towards interactive visualization tools for corpus linguistics. *International Journal of Corpus Linguistics*, v. 19, n. 3, p. 417–429, 2014. Disponível em: <http://www.jbe-platform.com/content/journals/10.1075/ijcl.19.3.05sii>. Citado na página 15.

SILVA, L. F.; ELER, D. M. Visual approach to boundary detection of clusters projected in 2d space. In: *14th International Conference on Information Technology: New Generations (ITNG 2017)*. Las Vegas, NV, USA: Springer International Publishing, 2017. (Advances in Intelligent Systems and Computing), p. 849–854. ISBN 978-3-319-54977-4. Citado na página 44.

SILVA, R. R. O. da. *Visualizing Multidimensional Data Similarities: Improviments and Applications*. Tese (Doutorado) — University of Groningen, 10 2016. Citado 3 vezes nas páginas 26, 28 e 29.

SMITH, G. *Tagging: People-powered Metadata for the Social Web*. [S.l.: s.n.], 2008. Citado na página 28.

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 17 e 27.

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 na página 55.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367. Citado na página 19.

TEJADA, E.; MINGHIM, R.; NONATO, L. G. On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization*, Palgrave Macmillan, v. 2, n. 4, p. 218–231, dez. 2003. ISSN 1473-8716. Disponível em: <http://dx.doi.org/10.1057/palgrave.ivs.9500054>. Citado na página 31.

VIÉGAS, F. B.; WATTENBERG, M. Timelines: Tag clouds and the case for vernacular visualization. *interactions*, ACM, New York, NY, USA, v. 15, n. 4, p. 49–52, jul. 2008. ISSN 1072-5520. Disponível em: <http://doi.acm.org/10.1145/1374489.1374501>. Citado na página 28.

WONG, P. C. Guest editor's introduction: Visual data mining. *IEEE Computer Graphics and Applications*, v. 19, p. 20–21, 09 1999. ISSN 0272-1716. Disponível em: <doi.ieeecomputersociety.org/10.1109/MCG.1999.788794>. Citado na página 19.

ZIPF, G. Human behaviour and the principle of least-effort. In: . Cambridge, MA: Addison-Wesley, 1949. Disponível em: </brokenurl#http://publication.wilsonwong.me/load.php?id=233281783>. Citado na página 22.