UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

PATRICIA SALLES ESCARASSATTI

Representação visual dos dados de produção bibliográfica da Plataforma Lattes

São Paulo
2022

PATRICIA SALLES ESCARASSATTI

Representação visual dos dados de produção bibliográfica da Plataforma Lattes

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo como parte dos requisitos para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 30 de Agosto de 2022. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Helton Hideraldo Bíscaro

São Paulo

2022

Ficha catalográfica elaborada pela Biblioteca da Escola de Artes, Ciências e Humanidades,
com os dados inseridos pelo(a) autor(a)
Brenda Fontes Malheiros de Castro CRB 8-7012; Sandra Tokarevicz CRB 8-4936

Salles Escarassatti, Patrícia
Representação visual dos dados de produção bibliográfica da Plataforma Lattes / Patrícia Salles Escarassatti; orientador, Helton Hideraldo Biscaro. -- São Paulo, 2022.
93 p: il.

Dissertacao (Mestrado em Ciencias) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2022.
Versão corrigida

1. Lattes. 2. Visualização de dados. 3. Projeção multidimensional. I. Biscaro, Helton Hideraldo, orient. II. Título.

Dissertação de autoria de Patrícia Salles Escarassatti, sob o título **"Representação visual dos dados de produção bibliográfica da Plataforma Lattes"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 30 de Agosto de 2022 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Helton Hideraldo Bíscaro
Universidade de São Paulo
Presidente

Prof. Dr. José de Jesus Pérez Alcazár
Universidade de São Paulo

Prof. Dr. Jesús Pascual Mena-Chalco
Universidade Federal do ABC

Agradecimentos

Ao meu orientador, Prof. Dr. Helton Hideraldo Bíscaro, pelos ensinamentos, apoio e dedicação oferecidos durante todo este tempo do mestrado. Sua contribuição foi fundamental para a realização e conclusão deste trabalho.

Aos professores da EACH-USP por todo o ensinamento transmitido ao longo desse período e, especialmente, ao Prof. Dr. Daniel de Angelis Cordeiro e ao Prof. Dr. Luciano Antonio Digiampietri, pelas orientações e correções no projeto de qualificação e ao Prof. Dr. José de Jesús Pérez Alcázar por fornecer os dados para realização deste trabalho.

Aos meus pais, por todo amor, apoio e por entender quando eu estava ausente ao longo desse período. Agradeço aos meus irmãos que sempre me apoiaram e torceram por mim.

Aos meus amigos e colegas que sempre me mandavam mensagens positivas e de incentivo e que entendiam meus momentos de ausência. Obrigada por se orgulharem das minhas decisões.

Infelizmente eu não consegui ter um convívio com os estudantes e a comunidade da EACH-USP, pois o momento vivido durante esse período foi muito solitário o que tornou a conclusão desse trabalho ainda mais desafiadora e com muitos ensinamentos.

"To educate as the practice of freedom is a way of teaching that anyone can learn."

(Bell Hooks)

"Simplicity is a great virtue but it requires hard work to achieve it and education to appreciate it. And to make matters worse: complexity sells better."

(Edsger Wybe Dijkstra)

Resumo

Escarassatti, Patrícia Salles. **Representação visual dos dados de produção bibliográfica da Plataforma Lattes**. 2022. 93 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2022.

A Plataforma Lattes é uma base de dados de currículos, grupos de pesquisas e instituições. O currículo Lattes é um repositório que contém informações dos estudantes e pesquisadores no Brasil e pode ser utilizada para gerar dados sobre os campos de pesquisa e pesquisadores. No entanto, as informações nem sempre são de fácil exploração e, por isso, torna-se necessário desenvolver ferramentas de visualização para auxiliar na identificação de autores e publicações em determinado campo de pesquisa. A utilização de visualização de coleção de documentos pode apoiar na exploração e análise visual de dados textuais. Técnicas de projeção criam representações visuais destacando a relação entre documentos com base no seu texto. Este trabalho propôs utilizar técnicas de visualização de projeção multidimensional para auxiliar na análise de dados bibliométricos que serão extraídos da plataforma Lattes. Por meio de coleta e análise de dados do Lattes, os dados serão preparados com o pré-processamento textual e serão aplicadas as técnicas de projeção multidimensional com a finalidade de verificar a existência de padrões, observando a distribuição geral dos dados e suas correlações. Foram avaliados 1038 artigos publicados de grupos de pesquisa da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH USP). Os grupos de pesquisa analisados nesse estudo são pertencentes ao programa de Pós-graduação em Sistemas de Informação, ao programa de Pós-graduação de Têxtil e Moda e o grupo de Astrofísica. Obtivemos resultados da projeção multidimensional que visualmente projetou artigos pertencentes a grupos de pesquisa que são relacionados entre si e de grupos de pesquisa que trabalham com temas distintos. De forma geral, o estudo evidenciou que as técnicas de projeção *Least Square Projection* e *Multidimensional Scaling - Isomap* apresentaram os melhores resultados para projetar e separar visualmente grupos de pesquisa que estudam temas distintos. Quando foram avaliados os grupos de pesquisa que estudam temas relacionados não houve claramente uma separação visual na projeção desses grupos. Dessa forma, essas técnicas de projeção podem ser utilizadas para avaliar, analisar e explorar visualmente os dados bibliométricos da plataforma Lattes.

Palavras-chaves: Lattes. Visualização de dados. Projeção multidimensional.

Abstract

Escarassatti, Patrícia Salles. **Visual representation of bibliographic production data from Lattes Platform**. 2022. 93 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2022.

The Lattes Platform is a database of curricula, research groups and institutions. The Lattes curriculum is a repository that contains information from students and researchers in Brazil and can be used to generate data about research fields and researchers. However, the information is not always easy to explore and, therefore, it is necessary to develop visualization tools to assist in the identification of authors and publications in a certain field of research. The use of document collection visualization can support visual exploration and analysis of textual data. Projection techniques create visual representations by highlighting the relationship between documents based on their text. This work proposed to use multidimensional projection visualization techniques to assist in the analysis of bibliometric data that will be extracted from the Lattes platform. By collecting and analyzing data from Lattes, the data will be prepared with text preprocessing and multidimensional projection techniques will be applied in order to verify the existence of patterns, observing the general distribution of the data and its correlations. Were evaluated 1038 published articles from research groups of the School of Arts, Sciences and Humanities of the University of São Paulo (EACH USP). The research groups analyzed in this study belong to the graduate program in Information Systems, the graduate program in Textiles and Fashion, and the Astrophysics group. Were obtained results from the multidimensional projection that visually projected articles belonging to research groups that are related to each other and from research groups that work with distinct themes. In general, the study showed that the Least Square Projection and Multidimensional Scaling - Isomap projection techniques presented the best results for projecting and visually separating research groups that study distinct themes. When research groups that study related topics were evaluated, there was clearly no visual separation in the projection of these groups. Thus, these projection techniques can be used to visually evaluate, analyze and explore the bibliometric data of the Lattes platform.

Keywords: Curriculum Lattes. Data visualization. Multidimensional projection.

Lista de figuras

Lista de quadros

Lista de tabelas

Sumário

1 Introdução

1.1 Contextualização

Com o rápido desenvolvimento da tecnologia e com dispositivos mais acessíveis, o número de documentos eletrônicos como artigos de notícias e artigos científicos aumentam continuamente. Esta explosão de documentos eletrônicos tornou difícil para um usuário selecionar os documentos que são úteis e extrair informações válidas deles (ALIGULIYEV, 2009).

Além disso, os avanços contemporâneos na área tecnológica têm elevado a necessidade de gestão do conhecimento organizacional disperso em diversas fontes de informação. Um desafio chave para sistemas de gestão do conhecimento é a descoberta eficaz e utilização dos conteúdos armazenados nas fontes de informação (ABBAS; ZHANG; KHAN, 2014).

O uso de visualização de coleção de documentos pode apoiar a recuperação, exploração e a análise de dados de texto. Esse é um tópico cada vez mais importante no campo de pesquisa de mineração visual de dados. Normalmente, o campo de visualização fornece suporte para outros domínios entenderem melhor seus dados (ISENBERG *et al.*, 2017).

O acompanhamento das áreas de pesquisa, especialmente quando existem múltiplas fontes de informação interdisciplinares, requer um esforço substancial de pesquisadores e programas de pesquisa para realizar a análise na área de produção científica. Os acadêmicos e cientistas contemporâneos dedicam substancial esforço para acompanhar os avanços em seus campos de atuação. O crescente número de publicações, combinado com fontes cada vez mais interdisciplinares, torna desafiador acompanhar as frentes de pesquisa emergentes e identificar os principais trabalhos. É ainda mais difícil começar a explorar um novo campo sem uma referência inicial (DUNNE *et al.*, 2012).

Em domínios específicos, como por exemplo, artigos científicos e patentes, a informação textual tem crescido na Internet e, consequentemente, a demanda por métodos eficientes de busca e recuperação de informação também cresceram. Pesquisadores têm dedicado esforços para propor ferramentas para lidar com a análise de documentos usando técnicas visuais e assim criando um campo de visualização conhecido como mineração de texto visual (ELER; GARCIA, 2013).

Para facilitar a atividade de análise dos dados de produções científicas e entender o que elas representam, a bibliometria pode ser utilizada para extrair e analisar esses dados. Essencialmente, bibliometria é um conjunto de técnicas utilizada para extrair dados de publicações e analisar esses dados de várias maneiras para responder as perguntas sobre a pesquisa que aquelas publicações representam. É um método de estudar os pesquisadores, processos e evolução da pesquisa (BELTER, 2015).

Nesse sentido, a bibliometria contribui para o progresso da ciência porque permite descobrirmos informações de muitas maneiras diferentes: avaliando o progresso a ser feito, identificando as mais confiáveis fontes de publicação científica, estabelecendo a base acadêmica para avaliação de novos empreendimentos, identificando principais atores científicos, desenvolvendo índices bibliométricos para avaliar a produção acadêmica e assim por diante (GUTIéRREZ-SALCEDO *et al.*, 2018).

As técnicas bibliométricas que empregam fundamentalmente análises quantitativas e índices estatísticos para avaliar a produção de pesquisa de indivíduos, instituições, periódicos, regiões ou países são instrumentos valiosos na medição e avaliação da produção de pesquisa científica. É possível utilizar essas técnicas para criar pronunciamentos sobre indicadores qualitativos das atividades científicas, além de seu alto potencial na realização de análises sistemáticas. Boas informações e conhecimentos úteis relacionados ao status das atividades de pesquisa em uma disciplina específica, que poderiam ajudar acadêmicos e pesquisadores na identificação e condução de novas dinâmicas de pesquisa, podem ser extraídos dos resultados e medições bibliométricas (ZYOUD; FUCHS-HANUSCH, 2017).

1.2 Justificativa

A comunidade científica brasileira tem disponível um sistema de informação curricular denominado Lattes mantido pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq. Por esse motivo, o chamado “Currículo Lattes” é considerado um padrão nacional de informação sobre realizações científicas e acadêmicas de alunos, professores, pesquisadores e profissionais envolvidos na ciência e tecnologia em geral (MENA-CHALCO; JUNIOR, 2009).

O currículo Lattes é um banco de dados rico e poderoso que apresenta inúmeras aplicações potenciais (científica, tecnológico, econômico, etc). Ele exibe informações apenas

de forma individual, ou seja, as informações cadastradas está individualmente associada a cada pessoa. Esta característica não fornece facilmente uma maneira de descobrir as produções bibliográficas, técnicas ou artísticas de um determinado grupo, como um grupo de pesquisa, professores de um departamento acadêmico ou membros de uma instituição brasileira (MENA-CHALCO; JUNIOR, 2009).

A maioria das instituições acadêmicas brasileiras costumam explorar os currículos Lattes a fim de elaborar relatórios sobre produções científicas, supervisões e projetos de grupos de pesquisa relacionados com essas instituições. Os relatórios são normalmente criados por análise manual dos dados do currículo Lattes de cada membro do grupo, a fim de obter um resumo completo de todas as produções científicas, supervisões e projetos do grupo. É importante notar que, apesar de ter informações estruturadas, esse procedimento é muito pesado e demorado, sendo altamente suscetível a erros causados pelo tratamento manual (MENA-CHALCO; JUNIOR, 2009).

Ferramentas para exploração rápida da literatura podem ajudar a diminuir as dificuldades em explorar e consolidar as informações do Lattes, fornecendo aos pesquisadores visões gerais concisas adaptados às suas necessidades e auxiliando na geração de pesquisas. Bibliotecas digitais e motores de busca são úteis para encontrar documentos específicos ou aqueles que correspondem a uma string de pesquisa, mas não fornecem as ferramentas de análise adicionais, como por exemplo análise de texto e visualização de dados textuais, necessárias para resumir rapidamente um campo. Usuários não familiarizados com o campo muitas vezes acham um desafio procurar pessoas influentes ou artigos, autores e periódicos inovadores (DUNNE *et al.*, 2012).

Técnicas de visualização e análise de texto podem ser usadas para fornecer visões gerais imediatas dos padrões de publicação e citação em um campo, mas são incomuns em ferramentas de exploração de literatura. Quando presente, eles geralmente não exibem muitos dados ou não fornecem as técnicas de interação necessárias para analisar as tendências de publicação e comunidades de pesquisa em um campo (DUNNE *et al.*, 2012).

A exploração visual de conjuntos de dados de alta dimensão tornou-se uma tarefa comum nos últimos anos e necessária para lidar com as complexidades de interpretação de grandes conjuntos de dados multidimensionais. A fim de tornar a exploração visual viável, diferentes abordagens de visualização de informações têm sido desenvolvido para lidar com a multidimensionalidade. Essas abordagens são conhecidas como técnicas de visualização de dados multidimensionais (TEJADA; NONATO; MINGHIM, 2003).

Esse trabalho pretende utilizar técnicas de visualização baseada na análise multidimensional dos dados que serão extraídos da plataforma Lattes. Com o objetivo de verificar se as técnicas de visualização de projeção multidimensional podem auxiliar a análise de dados bibliométricos da plataforma Lattes, verificando a existência de padrões e a distribuição geral dos dados.

A análise visual pode ser realizada pelo uso de técnicas de visualização de informações. Projeções multidimensionais são exemplos dessas técnicas, em que as dimensões originais são projetadas para um espaço dimensional inferior (normalmente bidimensional), e as instâncias são então exibidas em gráficos de dispersão. Esse processo de mapeamento pode levar à perda de informações, e diferentes estratégias podem ser aplicadas para criar a projeção, mas que preservam propriedades da distribuição de dados (ETEMADPOUR *et al.*, 2014).

1.3 Questão de pesquisa e objetivos

A questão de pesquisa desse trabalho é identificar as técnicas de visualização de dados direcionadas à exploração de coleção de documentos. O objetivo dessa pesquisa está descrito a seguir:

> *Verificar se as técnicas de visualização de projeção multidimensional podem auxiliar a análise de dados bibliométricos da plataforma Lattes. A visualização dessa coleção de documentos deve permitir a identificação visual de grupos de documentos relacionados por seus temas de pesquisa, bem como as fronteiras entre tais grupos.*

Além do objetivo principal, há também objetivos específicos desse mestrado que precisam ser atingidos e que estão descritos na sequência:

- Aplicar técnicas de pré-processamento de textos.
- Explorar as aplicações do modelo de espaço vetorial para análise bibliométrica.
- Mapear a coleção de documentos em um espaço visual usando técnicas de projeção multidimensional.
- Aplicar métricas de avaliação para avaliar os resultados das diferentes técnicas de visualização.

No próximo capítulo serão apresentadas técnicas de pré-processamento e as principais técnicas de projeção multidimensional para visualização de coleção de documentos.

2 Conceitos fundamentais

Uma enorme quantidade de material textual está aumentando a uma taxa exponencial, especialmente com o aumento do uso e aplicações da Internet. Dia após dia está se tornando muito difícil recuperar as informações relevantes dos documentos disponíveis eletronicamente (MANWAR *et al.*, 2012).

Projeções multidimensionais têm sido empregadas para gerar visões globais de conjuntos de dados, as projeções trabalham mapeando dados de alta dimensão em um espaço visual de baixa dimensão, normalmente duas dimensões, enquanto busca colocar pontos semelhantes próximos uns aos outros. Demonstrou-se que essas técnicas aplicadas a coleções de documentos podem gerar mapas de documentos perspicazes que são adequados para visualização e exploração intuitiva do assunto endereçado por essa coleção. Porque elas favorecem a percepção de similaridade e dissimilaridade de conteúdo, essas representações visuais permitem identificar visualmente grupos de documentos altamente relacionados (abordando temas semelhantes) e fronteiras entre grupos, favorecendo identificação de temas em geral, bem como focalização e exploração sobre temas de interesse (ALENCAR *et al.*, 2012).

Várias abordagens têm sido usadas pelos pesquisadores para retornar conhecimento relevante na recuperação de informações. O modelo de espaço vetorial tem sido o modelo mais popular na recuperação de informações entre os pesquisadores por fornecer uma estrutura formal para os sistemas de recuperação de informação (MANWAR *et al.*, 2012). As técnicas de visualização lidam com esses modelos vetorias para criar representações visuais que destacam as relações entre os documentos com base nas suas informações textuais.

2.1 Pré-processamento

A etapa de pré-processamento deve ser realizada com o objetivo de reduzir o número de termos no texto, selecionar as palavras relevantes e estruturar os dados para facilitar o processamento dos algoritmos e gerar uma boa representação visual da coleção de documentos. A lista abaixo apresenta as operações de pré-processamento que tipicamente são aplicadas para a criação de um modelo de espaço vetorial:

- **Tokenização:** o processo de tokenização consiste em representar cada palavra do texto em unidades distintas, chamadas de *tokens*. Esse processo se baseia em remover pontuação e espaços em branco no texto, sendo esses os pontos de separação de cada *token*. Dessa forma cada palavra do documento será representado como um *token* (ABASI *et al.*, 2020).
- **Remoção de *stopwords*:** *stopwords* são palavras que podem ser consideradas irrelevantes para a linguagem e que tem uma alta frequência, como por exemplo as palavras em português "o", "a", "e", "de", "em". É de extrema importância a remoção dessas palavras devido ao alto volume que afeta negativamente o agrupamento de documentos textuais, além de torná-lo mais demorado (ABASI *et al.*, 2020).
- ***Stemming*:** é uma técnica aplicada para remover prefixos e sufixos de palavras, dessa maneira essas palavras serão representadas pelo seu radical, por exemplo terra, terreno, terreiro, terrinha, terrestre, o radical será o termo "terr", que será utilizado como uma característica da coleção de documentos (ABASI *et al.*, 2020).
- **Lei de Zipf:** A Lei de Zipf é baseada na frequência dos termos que ocorrem em muitos documentos e que, por isso, não ajudam a distingui-los. A Curva de Zipf é desenhada considerando a ordenação da frequência de termos de modo decrescente e a partir dela podemos definir um limiar para excluir essas características menos significativas. No gráfico da Figura 1 o eixo cartesiano *Palavras* representa os termos em ordem decrescente de frequência e o eixo *Frequência* representa a frequência desses termos (ZIPF, 1949).
- **Corte de Luhn:** A partir da Curva de Zipf, Luhn especificou dois limiares para remover termos pouco representativos. A linha C na Figura 1 representa um desses cortes, palavras à esquerda seriam consideradas inadequadas, pois são os termos mais comuns por aparecer em qualquer tipo de documento. E uma vez que o grau de frequência foi proposto como um critério, um limite inferior, linha D, também seria estabelecida e os termos abaixo desse corte seriam considerados raros e, portanto, não indicariam significância em discriminar documentos distintos (LUHN, 1958). Para definir quais seriam esses limiares, Goffman estabeleceu um procedimento para eliminar os termos menos relevantes da base documental. Assim a zona de transição é dado pelo ponto onde a contagem de palavras tem a frequência próxima a um. Portanto, encontrando esse ponto n (Ponto de Goffman), ele nos serviria como ponto

de transição (SEQUERA; CASTILLO; SOTOS, 2009). O cálculo de n é realizado da seguinte maneira:

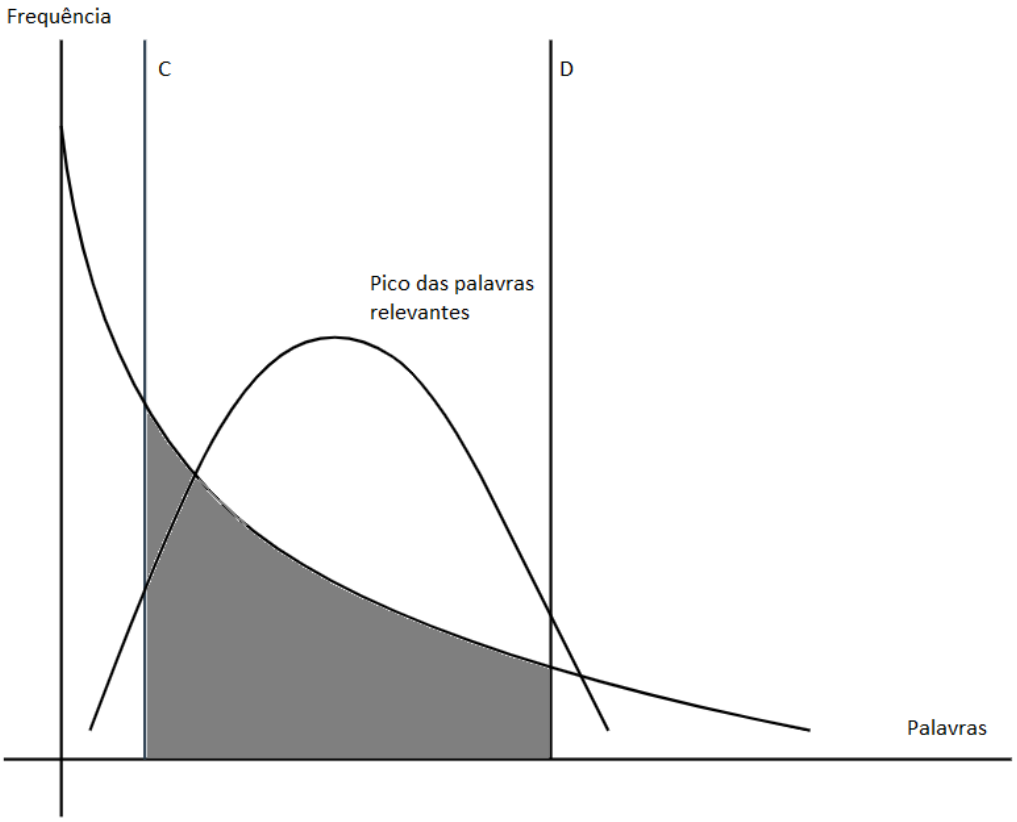
$$n = \frac{-1 + \sqrt{1 + 8 \times I_1}}{2} \tag{1}$$

onde:

I_1 - número de palavras com frequência igual a 1.

O objetivo é obter uma área de transição que proporciona maior número de sucessos a partir do ponto de Goffman.

Figura 1 – Corte de Luhn



Fonte: Elaborada pela autora com base em (LUHN, 1958)

2.2 Modelo de espaço vetorial

Em um modelo de espaço vetorial, cada documento é representado por um vetor de números que representa o quão bem cada palavra descreve o documento. Dessa forma, o significado de um documento pode ser obtido a partir de suas palavras. No modelo de espaço vetorial cada documento d_j é associado a um vetor no espaço dimensional

que consiste em todos os termos distintos dentro da coleção de documentos. Os termos, representados como k_i, no vetor de documentos são substituídos por seus pesos, com i variando de 0 a $m-1$,. Há várias maneiras de calcular os pesos dos termos, que consideram dois componentes básicos, o primeiro é o número de ocorrências de um termo em um determinado documento d_j e o segundo é o número de documentos que contém esse termo (KALMUKOV, 2020). A seguir será detalhado como calcular os pesos dos termos baseado nesses dois componentes básicos.

O peso $w_{j,i}$ é um valor real positivo associado ao par (k_i, d_j). No modelo de espaço vetorial, a frequência do termo k_i dentro do documento d_j é dado como tf e fornece um medida de quão bem esse termo descreve o documento, ou seja, quanto mais vezes um termo ocorre em um documento, mais importante esse termo é para esse documento. O fator df representa o número de documentos que contém k_i, quanto mais documentos contiverem um termo, mais comum e menos informativo ele será. Esta é uma medida inversa de informatividade. Mas os modelos de ponderação de termo consideram não a frequência do documento df, mas o inverso da frequência do documento, idf — quanto menos documentos contiverem um termo, mais informativo ele será. Idf é calculado da seguinte forma (ABASI *et al.*, 2020):

$$idf_i = log\frac{d}{df_i} \tag{2}$$

Onde:

idf_i - frequência inversa do documento do termo k_i.

d - número de documentos em toda coleção.

df_i - número de documentos que contém o termo k_i.

Intuitivamente, o esquema de ponderação tf-idf mais básico é a combinação desses dois termos:

$$w_{j,i} = tf_{j,i} * idf_i = tf_{j,i} * log\frac{d}{df_i} \tag{3}$$

Dessa forma, o vetor de um documento d_j é representado por $\vec{d_j}$ = (w$_{j,1}$, $w_{j,2}$, ..., $w_{j,n}$). O modelo de espaço vetorial representa os documentos como uma matriz $n \times m$ como segue:

$$\begin{pmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,(m-1)} & w_{1,m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \dots & \dots & \dots & \dots & \dots \\ w_{(n-1),1} & \dots & \dots & \dots & w_{(n-1),m} \\ w_{n,1} & w_{n,2} & \dots & \dots & w_{n,m} \end{pmatrix}$$

Para diminuir a dimensionalidade da matriz e melhorar a qualidade de representação vetorial é indicado realizar as técnicas de pré-processamento que foram citadas acima, com o objetivo de que somente os termos mais representativos da coleção de documentos possam ser utilizados. Além disso, pode ser observado uma discrepância nos valores numéricos da matriz, sendo necessário normalizá-la para que os vetores tenham norma euclidiana igual a 1. O processo de normalização segue a seguinte fórmula:

$$w'_{j,i} = \frac{w_{j,i}}{||(\vec{d_j})||} \tag{4}$$

Na Equação 4 cada coordenada do vetor de documentos é dividido pela norma $||(\vec{d_j})||$ do vetor de documento d_j da matriz $n \times m$ representada anteriormente.

O modelo de espaço vetorial consegue representar documentos de texto devido a seu esquema de ponderação de termos, sua estratégia de correspondência parcial e medida de similaridade. Porém a independência mútua de termos de índice é considerada uma desvantagem do modelo de espaço vetorial (MANWAR *et al.*, 2012).

2.3 Medidas de similaridade e dissimilaridade

Informalmente, a semelhança entre dois objetos é uma medida numérica em que dois objetos são semelhantes. Consequentemente, as semelhanças são mais altas para pares de objetos mais parecidos. As semelhanças não são negativas e geralmente estão entre 0 (sem similaridade) e 1 (similaridade completa). A dissimilaridade entre dois objetos é uma medida numérica para o qual os dois objetos são diferentes. Dissimilaridades são menores para pares mais semelhantes de objetos. Frequentemente, o termo distância é usado como sinônimo para a dissimilaridade, embora a distância seja frequentemente usada para se referir a uma classe especial de dissimilaridades (TAN; STEINBACH; KUMAR, 2005).

Dessa forma, o cálculo da distância euclidiana entre dois documentos é dado pela seguinte equação:

$$dist_{eucl}(d_1, d_2) = \sqrt{\sum_{k=1}^{m}(w_{1i} - w_{2i})^2} \tag{5}$$

onde m é o número de termos e $w_{j,i}$ refere-se a influência do termo i no documento j. Os valores de distância euclidiana só podem ser iguais ou maior que 0, quanto mais próximo de 0 for o valor da distância euclidiana mais semelhantes serão os documentos 1 e 2. A distância euclidiana pode ser generalizada pela métrica de distância Minkowski.

$$dist_{mink}(d_1, d_2) = \left(\sum_{k=1}^{m}|w_{1i} - w_{2i}|^r\right)^{\frac{1}{r}} \tag{6}$$

onde r é um parâmetro.

- $r = 1$ Distância de Manhattan
- $r = 2$ Distância Euclidiana

A distância euclidiana tem uma limitação quando os vetores, que representam os documentos, têm uma característica mais esparsa, ou seja, há muitos termos com valores nulos. Dessa forma, alguns documentos serão considerados similares por não apresentarem determinadas características. Uma medida que é utilizada para modelos vetoriais mais esparsos e de alta dimensionalidade é a similaridade do cosseno, pois ele não faz comparações com termos que não estão presentes nos documentos (TAN; STEINBACH; KUMAR, 2005).

Documentos representados como vetores podem ser comparados usando a similaridade de cosseno, que considera o ângulo entre eles: vetores idênticos com um ângulo de 0° produzem um valor 1, enquanto um par de vetores perpendiculares produz um valor 0 (DIAS; MILIOS; OLIVEIRA, 2019).

A similaridade do cosseno é calculada da seguinte maneira:

$$simi_{cos}(d_1, d_2) = \frac{\sum_{i=1}^{m}(w_{1i}.w_{2i})}{\sqrt{\sum_{i=1}^{m}(w_{1i}^2).\sum_{i=1}^{m}(w_{2i}^2)}} \tag{7}$$

Essa equação retorna valores entre -1 e 1, valores próximos de 1 indicam que os documentos compartilham termos similares, já se o valor é igual a 0 não há compartilha-

mento de características semelhantes entre os documentos e como os pesos dos termos de um documento é sempre não negativo, nesses casos não teremos valores menores que 0 (TAN; STEINBACH; KUMAR, 2005). A medida de cosseno pode ser transformada para uma medida de dissimilaridade aplicando o seguinte cálculo:

$$dist_{cos}(d_1, d_2) = 1 - simi_{cos}(d_1, d_2) \tag{8}$$

2.4 Técnicas de projeção multidimensional

Dados textuais, como coleções de documentos estão crescendo em número e tamanho. Esse mesmo crescimento significa que os dados costumam ser difíceis de entender e técnicas tradicionais de acesso a esses dados mostram pouco dos seus padrões, como por exemplo: como objetos individuais estão relacionados entre si e como os atributos de dados são distribuídos em tais padrões. Técnicas de visualização de informações geram uma representação gráfica em um espaço de baixa dimensão que representa bem os dados multidimensionais, o que permite a navegação e visão geral e podem ser produzidos de forma eficiente (CHALMERS, 1996).

Apresentaremos a seguir diferentes técnicas de projeção multidimensional que ajudam a tornar a análise de dados o mais visual possível.

2.4.1 *Force-directed placement* (FDP)

A ideia básica do algoritmmo de *Force-directed placement* (FRUCHTERMAN; REINGOLD, 1991) é baseada em grafos, substituindo os vértices por anéis de aço e substituindo cada aresta por uma mola para formar um sistema mecânico. Os vértices são colocados em um layout inicial de forma aleatória e as forças da mola nos anéis se movem até que o sistema atinja um estado mínimo de energia.

Esse modelo de grafo representado como um sistema físico de anéis e molas é baseado no trabalho de Eades (1984). Eades fez algumas considerações sobre as forças exercidas pela mola, por exemplo, as forças repulsivas são calculadas entre cada par de vértices, mas as forças de atração são calculadas apenas entre vizinhos. Isso reduz a complexidade do tempo porque calcular as forças de atração entre vizinhos é $\Theta(|E|)$,

embora o cálculo da força repulsiva seja $\Theta(|V|^2)$, onde V representa os vértices e E as arestas.

A configuração inicial do algoritmo pode ser total ou parcialmente especificado, mas normalmente os vértices são colocados aleatoriamente no layout. Diferentes funções podem ter sido escolhidas para calcular a força de repulsão e atração. Basicamente é calculado o efeito das forças de atração sobre cada vértice e o efeito das forças repulsivas. Deseja-se que os vértices fossem uniformemente distribuído no layout. Intuitivamente, quanto mais distantes dois vértices estão, ou quanto mais próximos o layout atual deve ser considerado e mais violenta a correção. Se f_a e f_r são as forças atrativas e repulsivas, respectivamente, com d a distância entre os dois vértices e k é a constante elástica, então:

$$f_a(d) = \frac{d^2}{k} \tag{9}$$

$$f_r(d) = -\frac{k^2}{d} \tag{10}$$

Para que o algoritmo *Force-directed placement* seja empregado para criação de projeções, as forças do sistema são proporcionais à diferença entre as dissimilaridades entre os objetos e as distâncias entre os pontos no espaço projetado.

Uma dissimilaridade em alta dimensão é calculada para pares de objetos, e então é aproximado o mais próximo possível no espaço de dimensão inferior do layout. Este último é geralmente medido como distância euclidiana. Simulações de forças físicas são usadas para conduzir o processo de layout. Cada par de objetos é considerado uma mola, cujas extremidades estão presas aos dois pontos. O comprimento relaxado da mola ou 'distância de repouso' é a proximidade ideal dos dois objetos, ou seja, sua distância em alta dimensão ou a dissimilaridade. Objetos semelhantes que estão muito distantes são puxados juntos e objetos diferentes que estão muito próximos são afastados. O layout final produzido pelo sistema refletirá o sistema de molas em equilíbrio.

Como cada obejto está sujeito às forças de todas os outros objetos e sendo necessário calcular $n(n-1)$ a força em cada iteração, o cálculo das forças é $O(n^2)$, onde n é a quantidade de objetos do sistema. Para produzir um layout são necessárias n iterações, o algoritmo resultante será $O(n^3)$. Portanto, essa técnica apesar de gerar layouts com precisão, sua aplicação é limitada a pequenos conjuntos de dados (MORRISON; ROSS; CHALMERS, 2011).

Algoritmo de Chalmers

Chalmers (1996) fez algumas considerações na técnica padrão de *Force-directed placement* para tornar o custo computacional em cada iteração linear em relação a n. Ao invés de fazer todos os cálculos de força $n(n-1)$ em cada iteração, será feito os cálculos de força entre cada objeto i e os membros de dois conjuntos cujo tamanho é limitado por uma constante. Desta forma, mantemos um custo computacional para cada iteração que é linear em relação a n.

O primeiro conjunto é armazenado como uma lista mantida dinamicamente de referências a objetos "vizinhos", V_i. Esta lista tem comprimento máximo $Vmax$ e as entradas em V_i são armazenados em ordem de distância do espaço multidimensional. Junto com V_i é mantido um valor $maxDist$ que é a distância máxima para qualquer membro da lista V_i. Enquanto o conjunto de vizinhos é mantido entre as iterações, o segundo conjunto é construído novamente a cada iteração. Este conjunto S_i é composto por objetos escolhido aleatoriamente, seu tamanho é uma constante $Smax$ e nenhum dos objetos pertencem a V_i.

Cada elemento candidato j para o conjunto S_i é selecionado e a distância d_{ij} é calculada. Se $d_{ij} < maxDist$ então j é inserido na posição apropriada no conjunto V_i em vez de $S - i$, esse processo pode alterar o valor da variável $maxDist$. Caso contrário, j é adicionado a S_i e uma vez que temos a quantidade de membros $Smax$, então as forças em i são calculadas usando uma série de cálculos restritos a serem menores ou iguais à constante $Vmax + Smax$:

$$F_i = \sum_{v \,\epsilon\, V_i} F_{iv} + \sum_{s \,\epsilon\, S_i} F_{is} \tag{11}$$

Conforme a amostragem continua, o conjunto S_i evolui em direção ao conjunto de objetos $Smax$ mais intimamente relacionados a i. Se adições ao conjunto de dados são feitas entre iterações, eles podem ser fácil, mas gradualmente acomodado - sem realizar um recálculo global de complexidade maior do que linear. Se não há mais adições ao conjunto, geralmente o processo de layout encerra depois de aproximadamente n a $3n$ iterações.

Force scheme

A abordagem também é baseada nos conceitos de força de atração e repulsão e usa o fato de que a relação entre as distâncias tanto no espaço original quanto no projetado deve ser constante para cada par de pontos de dados (x'_i, x'_j) A ideia é separar instâncias projetadas muito próximas e aproximar instâncias projetadas muito longe (TEJADA; NONATO; MINGHIM, 2003).

Essa técnica de melhoria de projeção baseada na força foi usada para melhorar a colocação de pontos, recuperando parte das informações perdidas durante o processo de projeção. A principal diferença aqui é que, uma vez que os pontos eram já projetados, utilizando técnicas rápidas, e com o esforço de preservar a distância, o número de iterações necessárias para convergir é muito pequeno (MINGHIM; PAULOVICH; LOPES, 2006).

A base do esquema de melhoria de projeção é a seguinte: para uma instância x'_i, é calculado o vetor $v_{ij} = (x'_i - x'_j)$, $\forall\ x'_j \neq x'_i$. Então, aplica-se uma perturbação x'_j na direção de v_{ij}. Esta perturbação depende das distâncias reais e ideais entre as instâncias projetadas. As distâncias são normalizadas para evitar inconsistências derivada da diferença entre os intervalos dos domínios bidimensional e multidimensional original. Uma vez que a normalização das distâncias é um processo caro, não pode ser executado em cada iteração para o espaço projetado. Para melhorar o desempenho, as distâncias são normalizadas apenas uma vez para o espaço original, e para cada iteração aplica-se uma normalização para as coordenadas das instâncias projetadas no espaço bidimensional, em vez das distâncias projetadas. O processo para cada iteração é apresentada no algoritmo abaixo (TEJADA; NONATO; MINGHIM, 2003).

Algoritmo 1 Force scheme

```
1: Para cada ponto de dado projetado x'
2:     Para cada ponto de dado projetado q' ≠ x'
3:         Calcule v como sendo o vetor de x' para q'
4:         Move q' na direção de v a uma fração de Δ
5: Normalize as coordenadas de projeção para o intervalo [0, 1] em ambas as dimensões
```

Fonte: (TEJADA; NONATO; MINGHIM, 2003)

Δ no Algoritmo 1 é uma aproximação para a diferença real entre a distância projetada e a distância no espaço original (ou seja, o erro nas posições relativas dos dois pontos x' e q').

A aproximação é dada por:

$$\Delta = \frac{d(x,q) - dmin}{dmax - dmin} - d_2(x',\, q') \tag{12}$$

onde $dmax$ e $dmin$ são as distâncias mínimas e máximas no espaço multidimensional, respectivamente e x' e q' representam as projeções dos pontos x e q.

A abordagem acima melhora as projeções bidimensionais de dados multidimensionais e em uma iteração cada ponto tem seu posicionamento alterado n - 1 vezes, assim uma iteração realiza muito mais trabalho sem ser mais complexa do que uma iteração do modelo original de molas (PAULOVICH, 2008).

2.4.2 *Multidimensional scaling* (MDS)

Multidimensional scaling (MDS) é uma das técnicas de redução de dimensionalidade que converte dados multidimensionais em um espaço de dimensão inferior, mantendo as informações intrínsecas dos dados. A principal razão para usar o MDS é obter uma exibição gráfica para os dados fornecidos, de modo que sejam muito mais fáceis de entender. A única suposição do MDS é que o número de dimensões deve ser um a menos que o número de pontos, o que também significa que pelo menos três variáveis devem ser inseridas no modelo e pelo menos duas dimensões devem ser especificadas (SAEED *et al.*, 2018).

MDS é a abordagem que mapeia os dados originais de alta dimensão (m dimensões) em um espaço dimensional inferior (d dimensões). Ele endereça o problema de construir uma configuração entre os n pontos de uma $\mathbf{D}$ matriz $k\,x\,k$, que é chamada de matriz de afinidade. MDS encontra n pontos de dados $y_1, ..., y_n$ de uma matriz de distância $\mathbf{D}$ em um espaço de dimensão d, de modo que se $\hat{d_{ij}}$ é a distância euclidiana entre y_i e y_j, então $\hat{\mathbf{D}}$ é semelhante a $\mathbf{D}$. O *Multidimensional scaling* pode ser considerado como:

$$\min_{Y} \sum_{i=1}^{k} \sum_{i=1}^{k} (d_{ij}^{X} - d_{ij}^{Y})^2 \tag{13}$$

onde $d_{ij}^{X} = ||x_i - x_j||^2$ and $d_{ij}^{Y} = ||y_i - y_j||^2$.

Várias técnicas diferentes de MDS foram propostas. A maioria delas tenta representar as coordenadas das dissimilaridades observadas no espaço m-dimensional. As dissimilaridades são mapeadas de forma a tentar coincidir com as distâncias euclidianas.

Assim, a dissimilaridade ρ_{ij} entre os pontos i e j é mapeada em sua distância d_{ij} com o mínimo de perda de informação. As dissimilaridades estão relacionadas à distância euclidiana pela função $f : \rho_{ij} \rightarrow d_{ij}(\mathbf{X})$, onde $d_{ij}(\mathbf{X})$ implica que a distância d_{ij} depende das coordenadas desconhecidas de $\mathbf{X}$. Onde $\mathbf{X}$ é uma matriz $n\,x\,m$, sendo n o número de pontos e m define o espaço dimensional.

Diferentes modelos de MDS podem ser definidos com base neste mapeamento de dissimilaridades para as distâncias. Portanto, todo modelo MDS começa com a Equação $f : \rho_{ij} = d_{ij}(\mathbf{X})$. O sinal de igualdade, no entanto, tem apenas valor teórico. O modo como essa aproximação entre as distâncias é realizada é que define as diferentes técnicas de MDS, sendo normalmente divididas em dois tipos: métricos MDS (MMDS) e não métrico MDS (NMDS). O MMDS tenta preservar as distâncias o mais próximo possível dos intervalos e proporções entre as proximidades. Enquanto o NMDS considera apenas a ordem de informações de proximidade. Existem cinco tipos de modelos MMDS, clássico MDS (CMDS), MDS replicado (RMDS), MDS generalizado (GMDS), MDS ponderado (WMDS) e Isomap. Nesse trabalho vamos descrever três tipos de modelos MMDS a seguir.

Classical MDS (CMDS)

O CMDS (SAEED *et al.*, 2018) procura encontrar uma isometria entre os pontos distribuídos em um espaço dimensional superior e em um espaço dimensional inferior. Se houver n pontos m-dimensionais, $\mathbf{X}$, então as dissimilaridades entre os pares de pontos é, p_{ij}. O CMDS tenta criar n projeções, x, dos pontos de alta dimensionalidade em um espaço linear d-dimensional, tentando organizar as projeções de modo que a distância euclidiana entre seus pares, d_{ij}, assemelhem-se às dissimilaridades entre os pontos de alta dimensão. Resumindo, o CMDS tenta minimizar:

$$\chi = \sum_{i \neq 1} (p_{ij} - d_{ij})^2 \tag{14}$$

onde p_{ij} é a dissimilaridade entre o ponto X_i e o ponto X_j, e d_{ij} é a distância entre a projeção de X_i, x_i e da projeção de X_j, x_j. O CMDS encontra a localização de pontos na forma de matriz $\mathbf{X}$, tomando a decomposição de autovalor da matriz duplamente centrada $\mathbf{B} = \mathbf{X}\mathbf{X}^T$. A matriz $\mathbf{B}$ de centro duplo é construída a partir da matriz de proximidade

$\mathbf{P}$ multiplicando-a pela matriz de centragem $\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T$, com $\mathbf{1} = (1, 1, 1, ..., 1)^T$ um vetor com n coordenadas iguais a 1 (SAEED *et al.*, 2018). A seguir estão as principais etapas para o CMDS:

- Calcular o quadrado da matriz de proximidade $\mathbf{P} = [p^2]$.
- Centralizar duplamente as informações de proximidade, ou seja, $\mathbf{B} = \frac{1}{2}\mathbf{JPJ}$ usando o operador de centragem $\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T$, onde n nos informa sobre o total de números de objetos. Centragem dupla é o método de subtrair as médias das colunas e linhas dos elementos de uma matriz e adicionando a média combinada.
- Extrair m autovetores $e_1, ..., e_m$ e os autovalores $\lambda_1, ..., \lambda_m$ correspondentes.
- As coordenadas dos n objetos no espaço m-dimensional são derivadas de:

$$X = \mathbf{E}_m \Lambda_m^{\frac{1}{2}} \tag{15}$$

  onde $\mathbf{E}_m$ são os m autovetores e Λ_m são os autovalores.

O CMDS pode ser aplicado para situações onde a dissimilaridade p_{ij} não é a distância euclidiana, nesse caso a matriz $\mathbf{B}$ deve ser semi-definida positiva ou a matriz de dissimilaridade $\mathbf{P}$ tem que ter propriedades euclidianas. Se a matriz $\mathbf{B}$ não for semi-definida positiva, então pode haver grande número de autovalores negativos. Nessas circunstâncias, a matriz de dissimilaridade $\mathbf{P}$ é considerada euclidiana por transformações, descartando autovalores muito pequenos.

O algoritmo *Classical scaling* não lida muito bem com grande quantidade de dados, pois sua complexidade aumenta quadraticamente, nesses casos podem ser utilizados outras variações baseado no CMDS (GROENEN; BORG, 2013).

Weighted MDS (WMDS)

No algoritmo WMDS deve ser calculado um parâmetro extra para ajustar os pontos e suas correspondentes dissimilaridades. As aplicações do WMDS incluem lidar com dados ausentes, ponderar os dados com base na sua confiabilidade e normalizar os dados (FRANCE; CARROLL, 2011). Uma vez que esses pesos são estimados, o resto do procedimento é semelhante ao CMDS. A função de perda para o WMDS é definida como:

$$\chi = \sum_{j=i+1}^{n} w_{ij}(p_{ij} - d_{ij}(\mathbf{X}))^2 \tag{16}$$

onde w_{ij} são os pesos associados que quantificam a precisão da dissimilaridade p_{ij}. Se não houver informação de dissimilaridade disponível entre os pontos i e j, então $w_{ij} = 0$. Os pesos w_{ij} nos informam sobre a precisão das proximidades, de modo que medições mais precisas recebem maior peso na função de perda. Há diferentes formulações propostas por autores e que usam a ideia de ponderação, uma das mais conhecidas é chamada de *Sammon's Mapping*.

Sammon's Mapping é popular no reconhecimento de padrões e utiliza a ideia de peso, onde o peso $w_{ij} = p_{ij}^{-1}$, dessa forma as pequenas dissimilaridades terão maior peso que as grandes. A função de erro para *Sammon's Mapping* é definido como:

$$\xi_{sam} = \sum_{j=i+1}^{n} \left(1 - \frac{d_{ij}(\mathbf{X})}{p_{ij}}\right)^2 = \sum_{i<j} p_{ij}^{-1}(p_{ij} - d_{ij}(X))^2 \tag{17}$$

Os pesos w_{ij} são especificados com base em algumas considerações formais. Uma maneira de escolher w_{ij} é equalizá-lo com a confiabilidade das informações de proximidade, o que significa que as proximidades mais confiáveis recebem mais peso, enquanto as proximidades não confiáveis têm menos peso (SAEED *et al.*, 2018).

Foi observado que o algoritmo ao lidar com *manifold* fechado, por exemplo loop ou esfera, o *Sammon's Mapping* tem problemas para desdobrar essas distribuições com loops levando a uma representação de "falsa vizinhança". Por exemplo, ao reduzir a dimensionalidade de uma forma de ferradura, as pontas da ferradura são susceptíveis a serem plotadas próximas uma da outra na dimensão inferior, apesar de estarem em extremos opostos da distribuição (FRANCE; CARROLL, 2011).

Isometric feature mapping (Isomap)

Isomap também é uma técnica de redução de dimensionalidade que mapeia as estruturas de alta dimensão em um espaço de baixa dimensão. O Isomap utiliza distâncias

geodésicas em vez da distância euclidiana entre cada par de pontos da sua distribuição (SAEED *et al.*, 2018).

Uma versão básica do algoritmo Isomap é descrita a seguir:

- Calcular as distâncias geodésicas (menor distância entre dois pontos) entre cada par de pontos para construir um grafo G. No grafo, há uma ligação entre o ponto i e j se a distância geodésica g_{ij} é menor que o limiar ϵ,$g_{ij} < \epsilon$, e o valor desta aresta é igual a g_{ij}.
- Uma vez que o valor para cada vértice é computado, usar um algoritmo, por exemplo Floyd-Warshall ou Dijkstra, para calcular os caminho mais curtos entre cada par de pontos no grafo.
- Uma vez que as distâncias para cada par de pontos estão disponíveis, executar o CMDS nas distâncias de caminhos mais curtos.

Uma possível função de perda para Isomap é definida como:

$$\xi_i = \sum_{i \neq j} (g_{ij} - d_{ij})^2 \tag{18}$$

O algoritmo Isomap é baseado no CMDS para problemas de *Multidimensional scaling* em grande escala. Em particular, ele se concentra em variedades não lineares e em dimensões mais altas. Como o CMDS não pode lidar com valores de dissimilaridades faltantes, eles são substituídos pelo caminho mais curto no grafo. Isso forma uma matriz totalmente preenchida de pseudossimilaridades nas quais o CMDS é executado (GROENEN; BORG, 2013).

Non-metric MDS (NMDS)

Non-metric MDS e CMDS, tenta calcular as coordenadas dos objetos no espaço m-dimensional, de forma que as proximidades coincidam com as distâncias entre os pontos. A base do NMDS é motivado por duas desvantagens do CMDS:

- Define uma função explicitamente de modo que as dissimilaridades dadas sejam transformadas em distâncias.

- A configuração do objeto é restrita a ser determinada na geometria euclidiana.

No entanto, o CMDS tem um desempenho melhor que a técnica de NMDS.

Basicamente, o NMDS exige uma relação menos restrita entre as proximidades e as distâncias. O NMDS depende da ordem de classificação das proximidades em vez de seus verdadeiros valores, portanto, é de natureza ordinal e também chamado de MDS ordinal. A distância da configuração final deve ser na mesma ordem de classificação que os dados originais. Dessa forma, o objetivo do NMDS é encontrar a configuração de pontos cujas distâncias refletem o mais próximo possível a ordem de classificação dos dados, por exemplo uma medida de intensidade e qualidade (SAEED *et al.*, 2018).

Em áreas de psicologia onde as medições dos dados são importantes, mas conjuntos de dados são pequenos, a técnica NMDS é frequentemente utilizada, mas para uso mais amplo, seria necessário aprimorar as técnicas de otimização do NMDS (FRANCE; CARROLL, 2011).

2.4.3 *Principal component analysis* (PCA)

A *Principal component analysis* (FODOR, 2002) é uma técnica de redução de dimensionalidade que combina as dimensões dos dados em um conjunto menor de dimensões. A dimensão dos dados é o número de variáveis que são medidos em cada observação da base de dados. Um dos problemas de conjunto de dados de alta dimensão é que nem todas as variáveis são importantes para a compreensão do conjunto como um todo.

PCA pode ser enunciado da seguinte forma: dado uma variável aleatória p-dimensional $\mathbf{x} = (x_i, ..., x_p)^T$, encontre uma representação em menor dimensão $\mathbf{s} = (s_i, ..., s_k)^T$ com $k \leq p$, que captura o conteúdo do dado original de acordo com algum critério. O componente $\mathbf{s}$ são chamados de componentes ocultos.

A redução da dimensão dos dados é baseada em combinações lineares ortogonais e são chamados de *Principal Components* (PC) das variáveis originais com a maior variância. O primeiro PC, s_1, é a combinação linear com a maior variância. Tem-se $s_1 = \mathbf{x}^T\mathbf{w}_1$ onde o vetor de coeficiente p-dimensional $\mathbf{w}_1 = (w_{1,1}, ..., w_{1,p})^T$ é resolvido com a seguinte equação:

$$\mathbf{w}_1 = argmax_{||w=1||} Var\, \mathbf{x}^T\mathbf{w} \tag{19}$$

O segundo PC é a combinação linear com a segunda maior variância e ortogonal ao primeiro PC. Para muitos conjuntos de dados, os primeiros PCs explicam a maior parte da variância, de modo que o resto pode ser desconsiderado com perda mínima de informações. Uma vez que a variância depende da escala das variáveis, é comum primeiro padronizar cada variável para obter média zero e desvio padrão igual a um de cada variável, dessa forma as variáveis originais que tinham escalas diferentes estarão todas em escalas comparáveis. Após a padronização dos dados a matriz de covariância dos dados $\mathbf{X}$ de n observações e p variáveis será dada por:

$$\sum_{pxp} = \frac{1}{n}\mathbf{X}\mathbf{X}^T \tag{20}$$

Realizado o cálculo da matriz de covariância, aplica-se uma decomposição espectral para encontrar seus autovetores.

$$\sum = \mathbf{U}\Lambda\mathbf{U}^T \tag{21}$$

onde $\Lambda = diag(\lambda_1, ..., \lambda_p)$ é uma matriz diagonal dos autovalores ordenados $\lambda_1 \leq ... \leq \lambda_p$ e $\mathbf{U}$ é uma matriz ortogonal pxp contendo os autovetores. Os PCs podem ser dados pelas linhas p de uma matriz $\mathbf{S}$ pxn.

$$\mathbf{S} = \mathbf{U}^T\mathbf{X} \tag{22}$$

Os pesos da matriz $\mathbf{W}$ são dados por $\mathbf{U}^T$. Dessa forma, o subespaço medido pelos primeiros k autovetores tem o menor desvio quadrático médio de $\mathbf{X}$ entre todos os subespaços de dimensão k.

Para determinar o número de PCs que serão mantidos pode se fixar um limite λ_0 e manter somente os autovetores onde seus autovalores correspondentes sejam maiores que esse limite λ_0. Os PCs são variáveis não correlacionadas construídas como combinações lineares das variáveis originais e sua interpretação não é simples e há perda de interpretabilidade nesse sentido.

As distâncias relativas entre os dados enquanto eles são projetados em um espaço de menor dimensão pelo PCA são preservadas o máximo possível e essa é uma característica da técnica *Classical* MDS, onde os resultados obtidos aplicando-se essa técnica são idênticos aos resultados apresentados usando o PCA.

Para dados que representam relações não-lineares há técnicas de PCA que introduzem análise não-linear aos componentes principais, mas os componentes resultantes ainda são combinações lineares das variáveis originais.

2.4.4 *Least square projection* (LSP)

Dado um conjunto de pontos $S = p_1, ..., p_n$ em R^m, o algoritmo visa representar os pontos de S em um espaço dimensional menor R^p, onde $p \leq m$, preservando a relação de vizinhança entre os pontos. Há dois passos envolvidos nesse processo de projeção. O primeiro, um subconjunto de pontos em S, chamado de "pontos de controle", são projetados em R^p pelo método MDS. Fazendo uso da relação de vizinhança dos pontos em R^m e as coordenadas cartesianas dos pontos de controle no R^p, é possível construir um sistema linear cujas soluções são coordenadas cartesianas dos pontos p_i em R^p (PAULOVICH *et al.*, 2008).

A Equação 23 representa um sistema linear, onde $V_i = p_{i1}, ..., p_{i_{k_i}}$ é um conjunto k_i pontos em uma vizinhança de um ponto p_i e $\tilde{p}_i$ são as coordenadas de p_i no R^p.

$$\tilde{p}_i - \sum_{p_j \in V_i} \alpha_{ij} \tilde{p}_j = 0 \tag{23}$$

onde $0 \leq \alpha_{ij} \leq 1; \sum \alpha_{ij} = 1$. Resolvendo a Equação 23 para os pontos em S então cada p_i será posicionado no fecho convexo dos pontos em V_i e quando $\alpha_{ij} = \frac{1}{k_i}$ teremos p_i no centróide dos pontos em V_i (PAULOVICH *et al.*, 2008).

Com esse conjunto de sistemas lineares é possível calcular as coordenadas dos pontos $\tilde{p}_i$, que é:

$$L\mathbf{x}_1 = 0, L\mathbf{x}_2 = 0, ..., L\mathbf{x}_p = 0 \tag{24}$$

onde L é a matriz $n \times n$ e $\mathbf{x_1}$, $\mathbf{x_2}$, ... ,$\mathbf{x_p}$ são os vetores das coordenadas cartesianas (x_1, ... , x_n). A matriz L é chamada de Laplaciana e é dada por:

$$l_{ij} = \begin{cases} 1 & i = j \\ -\alpha_{ij} & p_j \in V_i \\ 0 & \text{caso contrario} \end{cases} \tag{25}$$

O *rank* de L depende da relação de vizinhança entre os pontos e os pesos α_{ij}. Quando uma malha é fornecida, a vizinhança dos pontos pode ser obtido a partir da relação de incidência da malha. No entanto a matriz L não apresenta nenhuma informação geométrica, assim as soluções lineares não são tão úteis. Sendo necessário adicionar algumas informações geométricas ao sistema (PAULOVICH *et al.*, 2008).

Os pontos de controle nc são inseridos no sistema linear como novas linhas na matriz. Do lado direito do sistema linear são adicionados as coordenadas cartesianas dos pontos de controle, gerando um vetor não-nulo. Assim, considerando um conjunto de pontos de controle $S_c = p_{c1}, ..., p_{cnc}$, a Equação 24 poderá ser reescrita da seguinte maneira (PAULOVICH *et al.*, 2008):

$$A\mathbf{x} = \mathbf{b} \tag{26}$$

onde A é uma matriz retangular $(n + nc) \times n$ dada por:

$$A = \begin{pmatrix} L \\ C \end{pmatrix}, \; c_{ij} = \begin{cases} 1 & p_j \text{ é um ponto de controle} \\ 0 & \text{caso contrario} \end{cases} \tag{27}$$

e $\mathbf{b}$ é o vetor:

$$b_i = \begin{cases} 0 & i \leq n, \\ x_{p_{c_i}} & n < i \leq n + nc \end{cases} \tag{28}$$

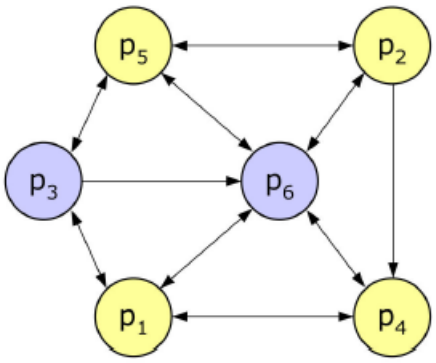
onde $x_{p_{c_i}}$ é uma das coordenadas cartesianas do ponto de controle p_{c_i} (PAULOVICH *et al.*, 2008).

A Figura 2 representa uma matriz A para um conjunto X com seis pontos. Os nós em azul são pontos de controle e os vizinhos de cada ponto são dado pelas relações de incidência no grafo direcionado.

Figura 2 – Matriz A resultante, relações de vizinhança e pontos de controle

$$A = \begin{bmatrix} 1 & 0 & -1/3 & -1/3 & 0 & -1/3 \\ 0 & 1 & 0 & -1/3 & -1/3 & -1/3 \\ -1/3 & 0 & 1 & 0 & -1/3 & -1/3 \\ -1/2 & 0 & 0 & 1 & 0 & -1/2 \\ 0 & -1/3 & -1/3 & 0 & 1 & -1/3 \\ -1/4 & -1/4 & 0 & -1/4 & -1/4 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \leftarrow L \\ \\ \leftarrow C \end{matrix}$$

(a) Matriz A (com pontos de controle p_3 e p_6).



(b) Relações de vizinhanças.

Fonte: (PAULOVICH, 2008)

Um sistema linear com pontos de controle pode ser resolvido aplicando-se o método de mínimos quadrados, que significa encontrar $\mathbf{x}$ que minimize $||Ax - b||^2$, ou seja, $\mathbf{x} = (A^T A)^{-1} A^T \mathbf{b}$. O sistema $A^T A\mathbf{x} = A^T \mathbf{b}$ que dever ser resolvido é simétrico e esparso o que facilita sua solução (SORKINE; COHEN-OR, 2004).

O conjunto de pontos de controle é determinado por meio de um algoritmo de agrupamento. Para executar esse algoritmo deve-se selecionar uma amostra de objetos do conjunto X que melhor representa a distribuição dos dados em R^m e os possíveis grupos de objetos existentes no espaço m-dimensional. O algoritmo de agrupamento é executado e irá gerar nc agrupamentos e um objeto representativo de cada grupo é escolhido como ponto de controle. Geralmente o medóide de cada cluster é escolhido como o ponto de controle, pois ele é o que mais se aproxima dos centróides (BERKHIN, 2006).

Qualquer método de agrupamento pode ser utilizado nesse processo de definição dos pontos de controle. Quando os dados tem uma representação vetorial se aplica o *bisecting k-means* e o método de *k-medoids* quando não existir uma representação vetorial (PAULOVICH, 2008).

Após a definição dos pontos de controle é necessário projetá-los em R^p utilizando um método de projeção multidimensional. O posicionamento dos pontos de controle tem um impacto importante na técnica LSP, uma vez que os objetos restantes serão interpolados para o layout final de acordo com esse layout inicial (PAULOVICH, 2008).

A complexidade computacional da LSP está relacionada com o número de pontos de controle que será escolhido e com a técnica de projeção aplicada. Se executarmos uma

técnica de projeção $O(n^2)$ e número de pontos de controle igual a $\sqrt{n}$, a complexidade de escolher e projetar os pontos de controle será dada por $O(n\sqrt{n})$ (PAULOVICH, 2008).

Com os pontos de controle já definidos é preciso encontrar os vizinhos mais próximos de cada ponto. Uma técnica utilizada é baseada em agrupamentos, onde primeiro procura-se os vizinhos mais próximos dos medóides dos agrupamentos, definindo os k agrupamentos. Dessa maneira, os vizinhos mais próximos do ponto de controle p_i estarão somente no agrupamento em que p_i pertence e nos agrupamentos vizinhos a esse grupo. A complexidade dessa técnica tem relação direta ao número de agrupamentos, sendo $\sqrt{n}$ agrupamentos ela terá complexidade $O(n\sqrt{n})$ (PAULOVICH *et al.*, 2008).

A complexidade final da LSP será $O(\max\{n\sqrt{n}, n\sqrt{k}\})$. Onde $n\sqrt{n}$ é a complexidade de definir os pontos de controle e o grafo de vizinhança, se $\sqrt{n}$ pontos de controle forem utilizados. E $n\sqrt{k}$ é a complexidade de resolver o sistema linear, sendo k o número de condiçaõ da matriz A^TA (PAULOVICH, 2008).

2.5 Métricas de avaliação de técnicas de projeção multidimensional

Há diversas técnicas de projeção multidimensional e cada uma tem diferentes características que retornam diferentes resultados. Dessa forma, é necessário avaliar objetivamente as projeções resultantes baseado em uma avaliação analítica. As principais medidas utilizadas foram o Coeficente da Silhueta e a *Neighborhood Hit*.

O Coeficiente de Silhueta é uma medida usada na análise de agrupamento e para avaliar a qualidade dos grupos gerados por técnicas de projeção. É uma medida calculada para cada instância, então calculamos o coeficiente médio de todas as instâncias como o Coeficiente de Silhueta da projeção. Este coeficiente varia de -1 a 1, onde os melhores resultados estão próximos de 1 (ELER; GARCIA, 2013).

A *Neighborhood Hit* é uma medida que representa a porcentagem de vizinhos mais próximos que pertencem à mesma classe de certa instância. A estratégia visa analisar a capacidade da visualização de preservar as classes em uma mesma vizinhança, favorecendo a percepção visual. Aplicando essa métrica para projeções, o cálculo é realizado em função de uma distância entre os pontos no plano de projeção. Quanto mais separados e agrupados estiverem os pontos maior será a precisão (ROMAN *et al.*, 2013).

2.6 Considerações finais

Neste capítulo foram apresentadas as principais técnicas de projeção multidimensional que mapeam objetos em um espaço m-dimensional em pontos em outro espaço p-dimensional, onde $p < m$ e técnicas de mineração de dados que auxiliam na captura de informação em coleções de documentos.

As técnicas foram dividas em três grupos: *Force-directed placement*, *Multidimensional scaling* e *Principal component analysis* representando as técnicas de redução de dimensionalidade. Além dessas, foi apresentada a técnica *Least square projection* que lida com dados de alta dimensionalidade.

A partir desse estudo foi observado que técnicas que apresentam maior complexidade computacional obtêm os melhores resultados. Porém sua aplicação fica limitada a conjunto de dados menores.

O próximo capítulo apresenta a revisão de trabalhos correlatos bem como os protocolos de revisão e a estratégia para realizar a análise dos dados desses trabalhos.

3 Trabalhos correlatos

Para a seleção dos trabalhos correlatos foi conduzida uma revisão sistemática com intuito de selecionar o estado da arte de trabalhos que melhor explicam as técnicas que são aplicadas para visualização de coleção de documentos. A revisão sistemática consiste em uma metodologia científica específica que vai um passo além da simples visão geral. A revisão sistemática é um método que permite especialistas obtenham resultados relevantes e quantificados. Isso pode levar à identificação, seleção e produção de evidências sobre a pesquisa em um determinado tema (MIAN *et al.*, 2005).

Normalmente, a condução de revisões sistemáticas é uma abordagem de três etapas. Os principais passos que compõem o processo de revisão são referentes ao planejamento, execução e análise de resultados. Durante a fase de planejamento, os objetivos da pesquisa são listados e um protocolo de revisão é definido. Tal protocolo especifica a questão central da pesquisa e os métodos que irão ser usados para executar a revisão. A fase de execução envolve a identificação dos estudos primários, seleção e avaliação de acordo com os critérios de inclusão e exclusão estabelecidos no protocolo de revisão. Uma vez selecionados os estudos, os dados dos artigos podem ser extraídos e sintetizados durante a fase de análise dos resultados. Enquanto essas fases são executadas, seus resultados devem ser armazenados. Nas próximas seções serão descritos como foram realizadas essas etapas durante o processo de revisão (MIAN *et al.*, 2005).

3.1 Protocolo da revisão de trabalhos correlatos

Para realizar o protocolo da revisão de trabalhos correlatos é necessário definir quais são as questões de pesquisa que precisam ser respondidas para condução da revisão. Definir qual será a estratégia e a *string* de busca dos artigos candidatos a estudos primários, juntamente com os critérios de inclusão e exclusão. Após selecionar os estudos primários é preciso extrair as informações desses estudos e selecionar os que serão prioritários para leitura e irão compor a revisão de trabalhos correlatos.

3.1.1 Questão de pesquisa

Foram definidas duas questões de pesquisas para este protocolo com propósito de conduzir a revisão de escopo. As questões de pesquisa têm o objetivo de descobrir quais são as técnicas utilizadas para visualizar coleção de documentos, com a finalidade de obter um entendimento mais amplo deste tema e quais trabalhos têm sido realizados com relação aos dados bibliométricos da plataforma Lattes.

Questão de pesquisa 1: Quais são as principais técnicas de projeção multidimensional utilizadas para visualização de coleções de documentos?

Questão de pesquisa 2: Quais são os principais trabalhos e análises visuais conduzidos com os dados bibliométricos da plataforma Lattes?

Para responder a primeira pergunta foi utilizada uma estratégia que visa realizar as buscas por técnicas de projeção multidimensional ou técnicas de visualização de documentos. O objetivo dessa questão é encontrar quais são as técnicas de projeção e visualização que são utilizadas para visualizar e mapear coleção de documentos. Coleção de documentos é um conjunto de documentos que contêm características comuns. Com relação à segunda questão de pesquisa foi realizada uma busca por trabalhos que estudam os dados bibliométricos do Lattes.

3.1.2 Estratégia e *string* de busca

O protocolo da revisão do estado da arte deve abordar a forma que foi realizada a pesquisa dos artigos candidatos a estudos primários, a seleção dos estudos primários relevantes e a análise dos dados para selecionar os estudos que irão responder as questões de pesquisa definidas anteriormente.

A fonte de dados utilizada para realizar a busca dos artigos foi o indexador de bases Scopus. Com relação a estratégia de busca para responder a primeira questão de pesquisa, foi construída uma *string* de busca apresentada no Quadro 1. A *string* é composta por palavras-chave presentes na questão de pesquisa e a utilização de sinônimos para cada um desses termos principais. O propósito da *string* foi pesquisar o maior número possível de trabalhos relacionados ao tema para responder a primeira questão de pesquisa desse projeto. A definição da *string* de busca foi um processo iterativo que envolveu vários ciclos

de tentativas e verificação dos artigos retornados até alcançar o resultado final da *string* de busca.

Quadro 1 – *String* de busca genérica para a primeira questão de pesquisa

TITLE-ABS-KEY (((“Multidimensional Projection Technique” OR “Multidimensional Projection”) OR (“visualization” OR “knowledge visualization” OR “visualization techniques”)) AND (“document collection” OR “collection of document” OR “document”))

Fonte: Autora do trabalho, 2020

A seleção dos estudos primários foi fundamentada em critérios de inclusão e exclusão que foram especificados para elencar os estudos primários mais relevantes. O objetivo desses critérios era garantir que somente estudos certamente relacionados ao contexto de visualização de coleção de documentos fossem selecionados. Considerando a questão de pesquisa, os seguintes critérios foram aplicados:

Critérios de inclusão:

- O trabalho aborda técnicas de visualização de coleção de documentos como parte principal do tema de estudo.
- O trabalho trata principalmente de técnicas de projeção multidimensional aplicadas a mapeamento de documentos.

Critérios de exclusão:

- O estudo não está escrito na língua inglesa.
- O estudo foi publicado antes de 2010.
- O tipo de documento do estudo é *Conference Review*, *Letter*, *Erratum*.
- A área de estudo é diferente das áreas de Ciência da Computação e Engenharia.

A *string* de busca resultante da aplicação dos critérios descritos acima está apresentada no Quadro 2.

Após a aplicação dos critérios de inclusão e exclusão tem-se um conjunto de artigos que passaram por uma coleta de dados. A extração dos dados desses estudos primários tem a finalidade de responder à primeira questão de pesquisa que foi levantada. Inicialmente é realizada a extração das informações mais gerais dos artigos, por exemplo, ano, autores, resumo, número de citações.

Para responder a segunda questão de pesquisa também foi utilizado o indexador de bases Scopus. Com relação a estratégia de busca para responder a segunda questão de

Quadro 2 – *String* de busca específica para a primeira questão de pesquisa

TITLE-ABS-KEY (((“Multidimensional Projection Technique” OR “Multidimensional Projection”) OR (“visualization” OR “knowledge visualization” OR “visualization techniques”)) AND (“document collection” OR “collection of document” OR “document”)) AND (EXCLUDE (DOCTYPE , “cr”) OR EXCLUDE (DOCTYPE , “er”) OR EXCLUDE (DOCTYPE , “le”)) AND (LIMIT-TO (SUBJAREA , “COMP”) OR LIMIT-TO (SUBJAREA , “ENGI”)) AND (LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-TO (PUBYEAR , 2013) OR LIMIT-TO (PUBYEAR , 2012) OR LIMIT-TO (PUBYEAR , 2011) OR LIMIT-TO (PUBYEAR , 2010)) AND (LIMIT-TO (LANGUAGE , “English”))

Fonte: Autora do trabalho, 2020

pesquisa, foi construída a seguinte *string* de busca: *TITLE-ABS-KEY ((“bibliometric” OR (“bibliographic production”) AND “lattes”)*. O propósito da *string* foi pesquisar os trabalhos que estudam os dados bibliométricos da plataforma Lattes e então responder a segunda questão de pesquisa desse projeto.

3.1.3 Estratégia planejada para extração de dados e síntese de resultados

Além da extração dessas informações mais diretas para cada estudo, foram elencados alguns outros dados que serão extraídos dos estudos primários e que serão utilizados como estratégia para chegar às respostas da questão de pesquisa. Os principais pontos que serão extraídos dos artigos são:

- Objetivo do artigo
- Quais técnicas de projeção foram utilizadas para visualização dos documentos
- Quais técnicas de pré-processamento textuais foram aplicadas
- Foi aplicada alguma técnica de extração de características? Quais?
- Base de dados utilizada
- Quais avaliações das técnicas de projeção e visualização de dados foram utilizadas?

Para operacionalizar essa extração de dados será utilizada uma planilha eletrônica que irá conter todas essas informações relativas aos estudos primários.

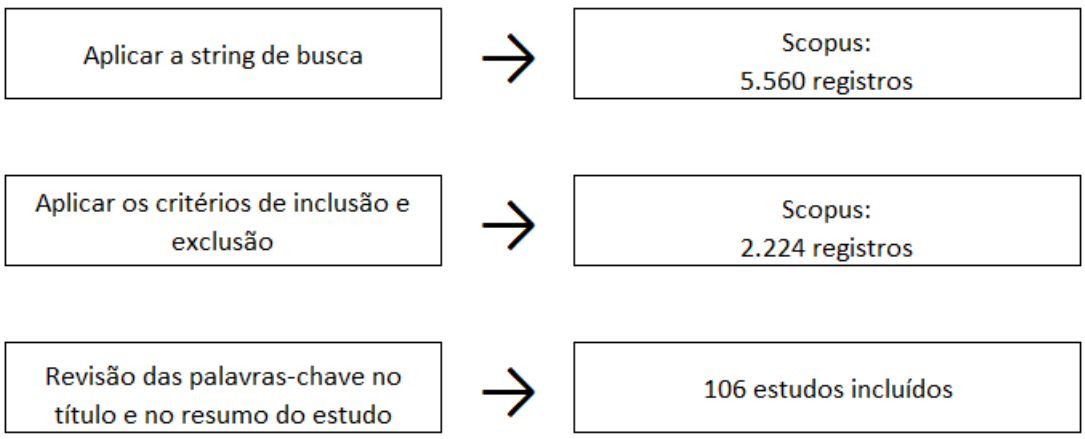
3.2 Condução da revisão de trabalhos correlatos para responder a primeira questão de pesquisa

Com a definição do protocolo da revisão de trabalhos correlatos, o próximo passo é executar todos os procedimentos que foram citados anteriormente para identificar os estudos primários. A *string* de busca que irá nos responder a primeira questão de pesquisa sobre as técnicas de projeção multidimensional foi executada na opção de busca avançada da biblioteca digital Scopus na primeira semana de outubro de 2020 e a busca retornou 5.560 registros que são os candidatos a estudos primários. Esse número pode ser considerado elevado, mas ele faz parte da estratégia de encontrar o maior número de candidatos a estudos primários.

Aplicando os critérios de inclusão e exclusão na opção de busca avançada do Scopus foram retornados 2.224 candidatos a estudos primários. Além desses critérios de inclusão e exclusão, foram realizadas buscas e leituras no título e resumo dos artigos selecionados. Essas buscas foram baseadas nas palavras-chave e termos sinônimos que compoem a questão de pesquisa desse trabalho. Por exemplo, artigos que apresentavam no seu título ou resumo os termos “projection”e “collection” ou “visualization” e “collection” ou “projection” e “document” ou “space” e “collection” foram incluídos como estudos primários após uma leitura mais detalhada do resumo desses estudos.

Após o término da aplicação dos critérios de inclusão e exclusão foram selecionados 106 estudos primários. Todo esse processo de seleção que foi descrito está esquematizado na Figura 3.

Figura 3 – Diagrama da seleção do estudo primário



Fonte: Autora do trabalho, 2020

Em posse dos estudos primários que foram selecionados foi realizada uma análise gráfica de algumas informções bibliográficas e de citação que são disponibilizadas no Scopus. A Tabela 1 mostra a distribuição dos estudos pelo tipo de documento, observa-se que a grande maioria dos estudos selecionados são artigos publicados em revista ou conferência.

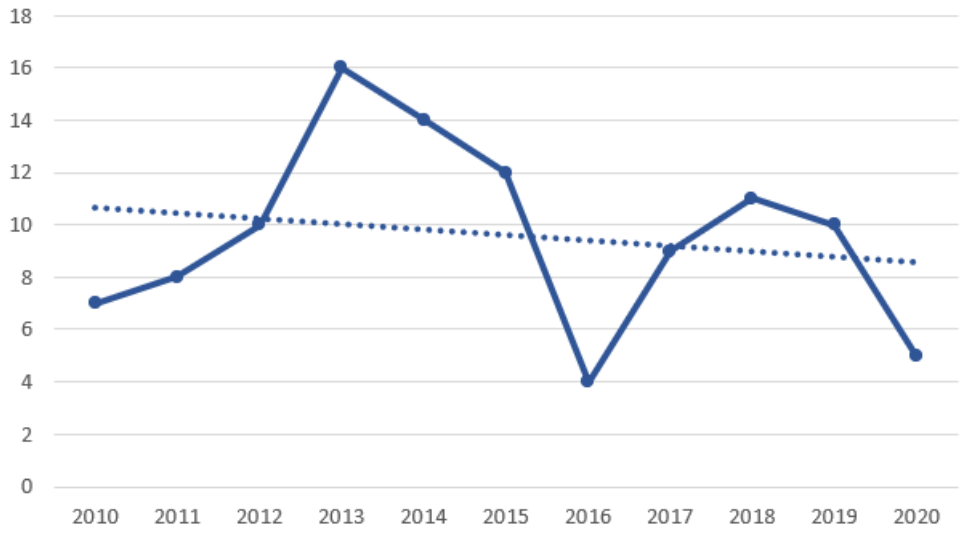
Tabela 1 – Tipo de documento dos estudos primários

Tipo de documento	# de estudos
Artigos em conferência	62
Artigos em revista	41
Revisão de artigos em revista	2
Capítulo de Livro	1

Fonte: Autora do trabalho, 2020

A Figura 4 mostra a distribuição dos artigos no período de tempo selecionado na busca, entre os anos de 2010 e 2020.

Figura 4 – Distribuição temporal dos estudos primários



Fonte: Autora do trabalho, 2020

Por fim, foi levantado os dados referentes às conferências ou revistas onde esses estudos foram publicados, como mostrado na Tabela 2.

A lista dos estudos primários que foram selecionados nesse trabalho com o objetivo de responder a questão de pesquisa sobre as técnicas de visualização de coleção de documentos é apresentada no Apêndice A. Todos os 106 documentos que estavam disponíveis eletronicamente foram baixados e, com a utilização de uma planilha eletrônica, foi realizada a extração de dados dos estudos para responder a questão de pesquisa.

Os estudos primários que tinham o maior número de citação foram os prioritários para leitura integral e para os estudos restantes foi realizado a leitura das seções de

Tabela 2 – Top 6 conferências e revistas

Título	# de estudos
IEEE Transactions on Visualization and Computer Graphics	10
Lecture Notes in Computer Science	8
Proceedings of the Annual Hawaii International Conference on System Sciences	3
ACM International Conference Proceeding Series	3
Proceedings of the International Conference on Information Visualisation	3
Computer Graphics Forum	3

Fonte: Autora do trabalho, 2020

introdução e conclusão. Toda essa análise para escolher os artigos que irão compor os trabalhos correlatos foi conduzida com base na leitura dos artigos e pela busca de palavras-chave dentro do texto dos estudos e organizados na planilha eletrônica conforme os pontos descritos anteriormente na estratégia de busca.

Após a definição do protocolo da revisão de estado da arte e definição da estratégia para realizar a análise e extração dos dados dos estudos primários foi necessário definir quantos artigos irão compor a presente revisão de estado da arte. Dentre os 106 estudos primários, os artigos selecionados e priorizados para a leitura integral foram aqueles que apresentaram o maior número de citação (maior que 50). Nos artigos restantes foram realizadas buscas de palavras-chave dentro do texto desses estudos primários. Por fim, foram selecionados 22 estudos primários para realizar leitura integral e compor a revisão de estado da arte.

No Quadro 3 temos um resumo das informações dos estudos selecionados que irão compor a revisão de estado da arte.

Quadro 3 – Estudos selecionados para a revisão de estado da arte sobre a primeira questão de pesquisa

Título do estudo	Autores	Ano
3D gesture-based exploration and search in document collections	De Antonio A., Moral C., Klepel D., Abente M.J.	2013
A literature review on the state-of-the-art in patent analysis	Abbas A., Zhang L., Khan S.U.	2014
A study on the role of similarity measures in visual text analytics	San Roman F.S., De Pinho R.D., Minghim R., De Oliveira M.C.F.	2013
Cite2vec: Citation-Driven Document Exploration via Word Embeddings	Berger M., McDonough K., Seversky L.M.	2017

Fonte: Patrícia Salles Escarassatti, 2020

Quadro 3 – Estudos selecionados para a revisão de estado da arte sobre a primeira questão de pesquisa

Título do estudo	Autores	Ano
Content Visualization of Scientific Corpora Using an Extensible Relational Database Implementation	Giannakopoulos T., Stamatogiannakis E., Foufoulas I., Dimitropoulos H., Manola N., Ioannidis Y.	2014
Dimensionality reduction for documents with nearest neighbor queries	Ingram S., Munzner T.	2015
Eye-tracking investigation during visual analysis of projected multidimensional data with 2D scatterplots	Etemadpour R., Olk B., Linsen L.	2014
Hybrid approach for visualization of documents clusters using GHSOM and sammon projection	Butka P., Pocsova J.	2013
Hybrid visualization approach to show documents similarity and content in a single view	Andreotti A.L.D., Silva L.F., Eler D.M.	2018
Interactive document clustering revisited: A visual analytics approach	Sherkat E., Nourashrafeddin S., Milios E.E., Minghim R.	2018
LabelTransfer-Integrating Static and Dynamic Label Representation for Focus+Context Text Exploration	Han Q., John M., Koch S., Assenov I., Ertl T.	2018
Scalable recursive top-down hierarchical clustering approach with implicit model selection for textual data sets	Muhr M., Sabol V., Granitzer M.	2010
Semantic wordification of document collections	Paulovich F.V., Toledo F.M.B., Telles G.P., Minghim R., Nonato L.G.	2012
Similarity preserving snippet-based visualization of web search results	Gomez-Nieto E., Roman F.S., Pagliosa P., Casaca W., Helou E.S., De Oliveira M.C.F., Nonato L.G.	2014
The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis	Chen C., Ibekwe-SanJuan F., Hou J.	2010
Time-aware visualization of document collections	Alencar A.B., Börner K., Paulovich F.V., De Oliveira M.C.F.	2012
Trivir: A visualization system to support document retrieval with high recall	Dias A.G., Milios E.E., Ferreira de Oliveira M.C.	2019
Two-stage framework for a topology-based projection and visualization of classified document collections	Oesterling P., Scheuermann G., Teresniak S., Heyer G., Koch S., Ertl T., Weber G.H.	2010
Using otsu's threshold selection method for eliminating terms in vector space model computation	Eler D.M., Garcia R.E.	2013
Vispubdata.org: A Metadata Collection about IEEE Visualization (VIS) Publications	Isenberg P., Heimerl F., Koch S., Isenberg T., Xu P., Stolper C.D., Sedlmair M., Chen J., Moller T., Stasko J.	2017

Quadro 3 – Estudos selecionados para a revisão de estado da arte sobre a primeira questão de pesquisa

Título do estudo	Autores	Ano
Visual abstraction and ordering in faceted browsing of text collections	Thai V., Rouille P.-Y., Handschuh S.	2012
Visual analysis and exploration of entity relations in document collections	John M., Heimerl F., Vu B.-A., Ertl T.	2018

A Tabela 3 apresenta as principais técnicas de projeção multidimensional e suas variações que foram aplicadas nos estudos primários e que auxilia na resposta da primeira questão de pesquisa desse trabalho. Observa-se que mais de uma técnica de projeção multidimensional pode ser aplicada em um mesmo artigo.

Tabela 3 – Técnicas de projeção multidimensional aplicadas nos estudos primários

Técnicas de projeção	# de estudos primários
Least Square Projection	6
t-Distributed Stochastic Neighbor Embedding	6
Multidimensional Scaling	4
Force-Direct Placement	3
Principal Component Analysis	3
Self Organizing Map	2

Fonte: Patrícia Salles Escarassatti, 2020

3.3 Condução da revisão de trabalhos correlatos para responder a segunda questão de pesquisa

Para realizar a identificação dos estudos primários que nos responderá a segunda questão de pesquisa referente aos artigos que estudam os dados bibliométricos da plataforma Lattes a *string* de busca foi executada na biblioteca digital Scopus e a busca retornou 43 registros que foram publicados entre os anos de 2010 e 2020.

Foram realizadas buscas e leituras no título e resumo dos artigos selecionados. Essas buscas foram realizadas com a finalidade de selecionar artigos que realizaram estudos e análises dos dados bibliométricos da plataforma Lattes. Após essa análise 12 artigos não foram selecionados para os estudos primários, pois não apresentam análises dos dados bibliométricos e de produção bibliográfica do Lattes. A lista dos 31 estudos primários que foram selecionados nesse trabalho com o objetivo de responder a segunda questão de

pesquisa sobre os estudos dos dados bibliométricos do Lattes é apresentada no Apêndice B.

Os resumos dos 31 artigos primários escolhidos foram lidos e 8 estudos foram selecionados para responder a segunda questão de pesquisa deste trabalho. Essa escolha foi baseada na análise do resumo desses artigos, os artigos escolhidos realizaram análises dos dados bibliométricos e de produção bibliográfica da plataforma Lattes. Os 23 artigos que não foram selecionados tinham como objetivo criar um perfil dos pesquisadores brasileiros e um perfil da produção científica de diferentes áreas de pesquisa, além de realizar estudos da participação de mulheres na ciência brasileira. Por fim, foram selecionados 8 estudos primários para responder a segunda questão de pesquisa deste trabalho.

No Quadro 4 temos um resumo das informações dos estudos selecionados que irá responder a segunda questão de pesquisa desse trabalho.

Quadro 4 – Estudos selecionados para a revisão de estado da arte sobre a segunda questão de pesquisa

Título do estudo	Autores	Ano
A bibliometric analysis of the scientific production and collaboration between graduate programs in manufacturing engineering in Brazil	Dutra S.T., Lezana Á.G.R., Dutra M.L., Pinto A.L.	2019
A System for Discovery of Knowledge in Data Repository Education	De Sousa Costa E., Rodrigues Dias T.M., Dias P.M.	2019
Analysis of an advisor-advisee relationship: An exploratory study of the area of Exact and Earth Sciences in Brazil	Tuesta E.F., Delgado K.V., Mugnaini R., Digiampietri L.A., Mena-Chalco J.P., Pérez-Alcázar J.J.	2015
Brazilian bibliometric coauthorship networks	Mena-Chalco J.P., Digiampietri L.A., Lopes F.M., Cesar Jr. R.M.	2014
Competitive intelligence in panorama of Brazil: The researchers and scientific production on lattes platform	Amaral R.M., Brito A.G.C., Rocha K.G.S., Quoniam L.M., de Faria L.I.L.	2016
Brazilian bibliometric coauthorship networks	Mena-Chalco J.P., Digiampietri L.A., Lopes F.M., Cesar Jr. R.M.	2014
Competitive intelligence in panorama of Brazil: The researchers and scientific production on lattes platform	Amaral R.M., Brito A.G.C., Rocha K.G.S., Quoniam L.M., de Faria L.I.L.	2016
Multi and interdisciplinarity in the brazilian postgraduate programs in information science	Lança T.A., Amaral R.M., Gracioso L.S.	2018

Fonte: Patrícia Salles Escarassatti, 2022

Quadro 4 – Estudos selecionados para a revisão de estado da arte sobre a segunda questão de pesquisa

Scientific collaboration in biotechnology: The case of the northeast region in Brazil	Costa B.M.G., da Silva Pedro E., de Macedo G.R.	2013
The Brazilian academic genealogy: evidence of advisor–advisee relationships through quantitative analysis	Damaceno R.J.P., Rossi L., Mugnaini R., Mena-Chalco J.P.	2019

Na sequência desse capítulo é apresentada uma análise mais aprofundada dos estudos primários selecionados para essa revisão de escopo.

3.4 Resultado da revisão de trabalhos correlatos

Nessa seção serão detalhados os resultados da revisão de trabalhos correlatos para responder a primeira e a segunda questão de pesquisa do presente trabalho.

3.4.1 Técnicas de pré-processamento

O estudo de Alencar *et al.* (2012) pontua que projeções multidimensionais têm sido empregadas para gerar visualizações globais de conjuntos de dados de alta dimensão. É realizado um mapeamento de dados de alta dimensão em um espaço visual de baixa dimensão, normalmente 2D, enquanto pontos semelhantes são colocados próximos um ao outro. Foi demonstrado que essas técnicas aplicadas a coleções de documentos podem gerar mapas de documentos perspicazes que são adequados para visualização e exploração intuitiva do conteúdo da coleção.

A maioria das técnicas de projeção multidimensional aplicadas nos estudos cria representações visuais destacando a relação entre documentos a partir de informações textuais, conforme Eler e Garcia (2013). Para isso, um modelo de espaço vetorial é calculado usando o conteúdo do documento e as técnicas de visualização lidam com tais modelos para estabelecer relação entre documentos. Ferramentas visuais de mineração de texto aplicam técnicas de redução de dimensionalidade ao modelo de espaço vetorial para representar coleções de documentos no espaço visual (espaço 2D). O pré-processamento para a criação do modelo de espaço vetorial é importante para obter informações que caracteriza os dados e geram representações visuais.

Com relação às técnicas de pré-processamento, podemos destacar que os artigos de Gomez-Nieto *et al.* (2014), Alencar *et al.* (2012), Eler e Garcia (2013), Giannakopoulos *et al.* (2013) e Butka e Pócsová (2013) usaram técnicas de tokenização, *stopwords*, *stemming* e a técnica de cálculo de frequência dos termos, conhecida como tf-idf, para pré-processar o conjunto de dados de entrada e produzir uma matriz que representa os documentos (a chamada representação baseada em um modelo vetorial).

Os estudos de Chen, Ibekwe-SanJuan e Hou (2010), Oesterling *et al.* (2010), Sherkat *et al.* (2018), Thai, Rouille e Handschuh (2012) e Dias, Milios e Oliveira (2019) usaram a métrica tf-idf para auxiliar na obtenção de uma visão geral do conteúdo do conjunto de documentos que estava sendo estudado.

Os estudos de Han *et al.* (2018) e de John *et al.* (2018) além de usar a técnica tf-idf utilizaram uma técnica de pré-processamento chamada de G^2. A medida G^2 classifica a importância dos termos comparando seu uso em um agrupamento e fora do agrupamento. Especificamente, ele compara a frequência relativa dos termos no grupo e fora do grupo. Termos que possuem um maior frequência relativa no agrupamento tem uma avaliação mais elevada. Assim, a métrica G^2 seleciona termos que podem diferenciar de forma mais eficaz os documentos em um agrupamento daqueles fora do agrupamento.

Dando continuidade as técnicas de pré-processamento, é necessário computar a similaridade entre dois documentos no modelo de espaço vetorial que foi definido para a coleção de documentos. Os artigos de Ingramn e Munzner (2015), Berger, McDonough e Seversky (2017), Antonio *et al.* (2013), Muhr, Sabol e Granitzer (2010), Sherkat *et al.* (2018), Dias, Milios e Oliveira (2019), Gomez-Nieto *et al.* (2014), Alencar *et al.* (2012) e Chen, Ibekwe-SanJuan e Hou (2010) utilizaram a similaridade do cosseno para recuperar informações dos documentos. Conforme reportou o artigo de Antonio *et al.* (2013) esta medida de similaridade, calculada para cada documento pareado, é de natureza estatística, pois reflete apenas a proporção de *tokens* que ambos os documentos têm em comum, independentemente de seu significado semântico. Portanto, não refletirá fielmente se eles compartilharem o mesmo tópico, mas apenas a taxa de palavras que eles compartilham. Mesmo que isso possa parecer uma desvantagem, o objetivo não era obter muito valores precisos de similitude, mas obter valores bons o suficiente no menor tempo possível.

O estudo de Abbas, Zhang e Khan (2014) usa a distância euclidiana para calcular a similaridade entre os documentos e suas relações de distância. Já o artigo de Isenberg *et al.* (2017) utilizou a distância de Levenshtein para quantificar a distância entre duas *strings*

expressa como o número de caracteres inseridos, excluídos e operações necessárias para converter uma em outra. Esse cálculo de distância ajuda a lidar com erros de digitação, variações ortográficas, variações em pontuação e ruído geral nos dados.

No trabalho de Roman *et al.* (2013) foi realizado uma comparação entre diferentes métricas de similaridade de textos para análise visual de coleção de documentos utilizando diferentes base de dados. As medidas que foram avaliadas são as seguintes: o Coeficiente de Dice, a Similaridade do Cosseno, o Coeficiente de Matching, o Coeficiente de Overlap, a medida Q-gram, e as medidas denominas *Normalized Compression Distance* e *scaled Normalized Compression Distance*. Essas medidas calculam a similaridade entre duas *strings*.

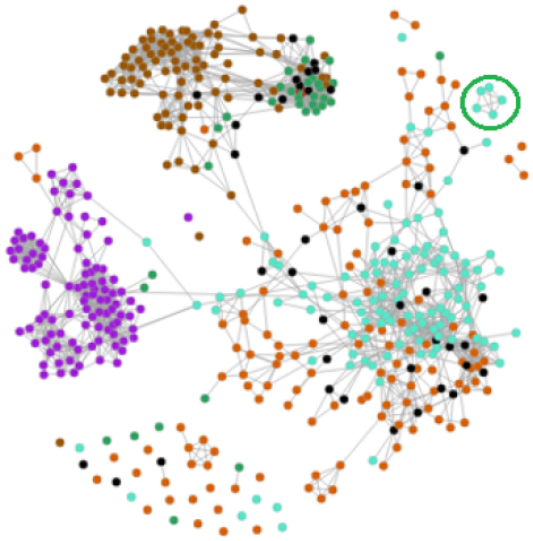
As medidas de similaridade que melhor representaram as coleções de documentos foram a Q-gram, a Similaridade do Cosseno e o Coeficiente de Overlap. Essas técnicas têm como principal vantagem a não representação intermediária dos textos, como os modelos de espaços vetoriais, no entanto os cálculos destas distâncias têm um custo computacional caro e o processo de contabilizar as dissimilaridades torna-se lento. As abordagens não consideram a análise semântica dos textos. Embora este tipo de processamento e o cálculo da dissimilaridade é suficiente para muitas aplicações, uma investigação mais aprofundada deve ser conduzida em distâncias baseadas na semântica, já que a semântica não pode ser ignorada em algumas análises de texto.

3.4.2 Técnicas de projeção multidimensional

Com relação às técnicas de projeção multidimensional, os estudos de Antonio *et al.* (2013), Sherkat *et al.* (2018) e Muhr, Sabol e Granitzer (2010) utilizaram a técnica de *Force-Direct Placement* para realizar a representação visual das coleções de documentos. Essa técnica foi utilizada pois ela tenta melhorar a proximidade de pontos de dados semelhantes e aumentar a separação para pontos de dados diferentes. Projetar documentos com base na abordagem de *Force-Direct Placement* coloca documentos com rótulos de grupos semelhantes juntos, enquanto projetam os nós isolados longe do centro do agrupamento, conforme a Figura 5.

A técnica *Force-Direct Placement* simula forças de atração e repulsão entre os documentos dependendo de suas medidas de similaridade. Além dessa abordagem ter uma

Figura 5 – *Force-Direct Placement*



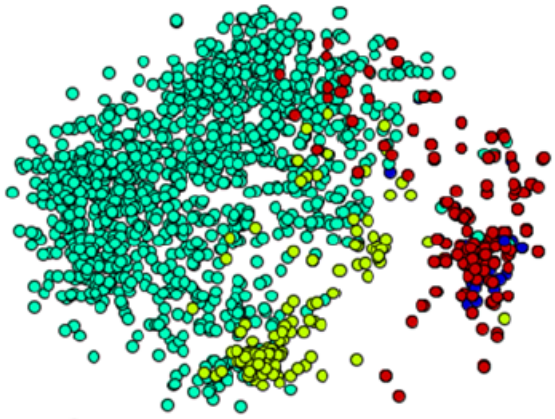
Fonte: (SHERKAT *et al.*, 2018)

grande escalabilidade, ela também tem muitas propriedades desejáveis para o objetivo desse estudo: boa qualidade do *layout* resultante, processo de posicionamento iterativo e em tempo real e capacidade de ser estendido incluindo outros fatores no processo de posicionamento.

Além da técnica *Force-Direct Placement* outra técnica de projeção multidimensional empregada nos estudos foi a *Multidimensional Scaling*. O artigo de Eler e Garcia (2013) aplicou essa técnica por apresentar melhores resultados visuais em comparação com a técnica *Least Square Projection*. No entanto, os experimentos utilizaram conjuntos de dados com poucos documentos, já que não foi considerado razoável utilizar *Multidimensional Scaling* em conjuntos de dados maiores devido a sua complexidade computacional. Com relação ao estudo Etemadpour, Olk e Linsen (2014), a técnica *Multidimensional Scaling* foi utilizada pois é uma alternativa capaz de lidar com conjuntos de dados não lineares. A projeção *Multidimensional Scaling* teve uma tendência em criar aglomerados de dados mais arredondados conforme Figura 6, essa técnica foi avaliada nesse estudo utilizando rastreadores oculares que analisa dados multidimensionais projetados, buscando relação, comparação de comportamento e identificação de padrão.

Outra técnica de projeção multidimensional que foi aplicada nos estudos selecionados é a *Least Square Projection*. Diversos estudos aplicaram essa técnica e compararam com outras projeções. Por exemplo, o estudo Gomez-Nieto *et al.* (2014) aplicou a técnica *Least Square Projection* devido à sua boa precisão em termos de preservação de distância e seu

Figura 6 – *Multidimensional Scaling*



Fonte: (ETEMADPOUR; OLK; LINSEN, 2014)

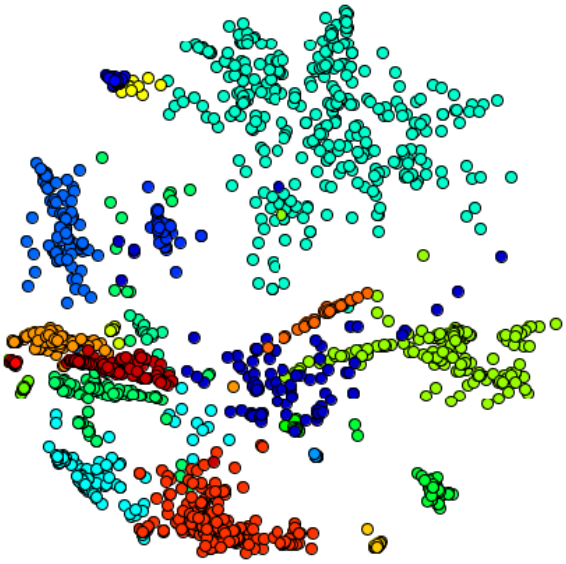
baixo custo computacional. A projeção preserva muito da estrutura de vizinhança original dos dados, garantindo que instâncias semelhantes sejam colocadas próximas umas das outras no espaço visual. O estudo de Andreotti, Silva e Eler (2018) e Paulovich *et al.* (2012) empregaram a técnica de projeção *Least Square Projection* que lida com grandes conjuntos de dados e tem um baixo custo computacional para projetar coleções de documentos no espaço 2D.

O artigo de Roman *et al.* (2013) aplicou *Least Square Projection* que gera um *layout* que preserva os agrupamentos de vizinhança no espaço de características, conforme mostrado na Figura 7. Primeiro é obtido uma subamostra dos dados, chamada de pontos de controle, que representa a distribuição espacial dos documentos analisados. Em seguida, é calculado as vizinhanças para esses pontos de controle. Esses pontos de controle então são projetados, mostrando uma visão global que representa grupos de textos com conteúdo semelhantes.

A técnica *Least Square Projection* também foi aplicada no estudo Alencar *et al.* (2012) com a finalidade de criar uma projeção com percepção temporal. O estudou gerou uma sequência de mapas baseados em similaridade que transmite a evolução de uma coleção de documentos ao longo do tempo.

Além das técnicas de projeção multidimensional, outros estudos utilizaram técnicas de redução de dimensionalidade para projetar documentos em um espaço visual 2D. O estudo de Sherkat *et al.* (2018) utilizou a técnica *t-Distributed Stochastic Neighbor Embedding* (t-SNE) combinada com a técnica de projeção *Force-Direct Placement*. O

Figura 7 – *Least Square Projection*



Fonte: (ROMAN *et al.*, 2013)

algoritmo t-SNE demonstra melhor desempenho na visualização de agrupamentos de pontos de dados do que a técnica *Principal Component Analysis* (PCA) que também é baseada na redução de dimensionalidade dos documentos avaliados. Os algoritmos t-SNE e PCA usam uma abordagem de sacos de palavras para representação de documentos com a finalidade de calcular a semelhança entre pares de documentos.

No estudo de Etemadpour, Olk e Linsen (2014) a técnica *Principal Component Analysis* foi aplicada para efeito de comparação com a *Multidimensional Scaling*. No entanto, o PCA teve maiores problemas para segregar os agrupamentos.

O estudo de Giannakopoulos *et al.* (2013) utilizou técnicas de agrupamento de documentos, o algoritmo utilizado foi o *k-means* que agrupa classes com conteúdos similares. Após o procedimento de agrupamento foi utilizada uma técnica de representação 2D chamada *Self Organizing Map* (SOM) que é um tipo de rede neural artificial. Cada classe é representada por um par de coordenadas discretizadas no espaço de recursos 2D.

3.4.3 Estudos dos dados bibliométricos da plataforma Lattes

Os 8 estudos primários selecionados para responder a segunda questão de pesquisa desse trabalho podem ser divididos entre dois grupos, o primeiro grupo é composto por artigos que realizaram análises visuais dos dados bibliométricos do Lattes com o objetivo de criar redes de colaboração entre pesquisadores, universidades e grupos de pesquisa. O

segundo grupo contém artigos que estudaram os dados bibliométricos do Lattes com a finalidade de criar mapeamentos dos tópicos estudados dentro das universidades, cursos e grupos de pesquisa.

Com relação aos estudos sobre redes de colaboração, o artigo Costa e Dias (2019) propõe utilizar técnicas de processamento de linguagem natural e teoria de grafos para criar uma visão macro de como ocorrem colaboração nas pesquisas interdisciplinares no Brasil. Os estudos de Dutra Álvaro Guillermo Rojas Lezana e Pinto (2019) e Costa e Macedo (2013) criaram redes de colaboração entre pesquisadores e instituições de ensino, respectivamente. O artigo de Damaceno Luciano Rossi e Mena-Chalco (2019) criou uma visualização de grafos para representar o relacionamento entre áreas de conhecimento e a rede de colaboração entre essas áreas. Por fim, o artigo de Mena-Chalco e Digiampietri (2014) e o artigo de Tuesta Karina Delgado e Pérez-Alcázar (2015) identificaram redes de coautoria de pesquisadores de algumas grandes áreas de conhecimento e a relação de colaboração entre os pesquisadores cadastrados na plataforma Lattes.

Os artigos que estudaram e mapearam tópicos de pesquisa foram o artigo de Lança e Gracioso (2018) e o artigo de Amaral Aline Grasiele Cardoso Brito e Faria (2016). O primeiro avaliou a multidisciplinaridade dos programas de pós-graduação em Ciência da Informação, essa avaliação ocorreu a partir de análises das áreas de atuação declaradas pelos docentes. O segundo artigo avaliou as palavras-chave presentes nos artigos científicos para elaborar quais são as principais temáticas de desenvolvimento na área de Inteligência Competitiva no Brasil.

3.5 Considerações finais

Nesse capítulo foi apresentada o protocolo da revisão do estado da arte que foi aplicado no presente trabalho. Foram definidas estratégias e *string* busca, além da estratégia para extrair os dados para chegar às respostas das questões de pesquisa.

Após essas definições, os estudos primários foram selecionados e a partir deles foi realizado a seleção dos artigos que compôs essa revisão. A revisão de estado da arte apresentou as principais técnicas de pré-processamento e de projeção multidimensional que são utilizadas na análise visual de coleção de documentos e quais as principais análises visuais realizadas com os dados bibliométricos da plataforma Lattes.

No que diz a respeito à primeira questão de pesquisa sobre as principais técnicas de projeção multidimensional utilizadas para visualização de coleção de documentos os estudos selecionados indicaram que a utilização da *Force-Direct Placement* tem uma boa qualidade na visualização resultante e boa escalabilidade. O *Multidimensional Scaling* cria agrupamentos de dados mais agrupados, porémm tem alta complexidade computacional. A técnica *Least Square Projection* apresenta um baixo custo computacional e garante que instâncias semelhantes sejam colocadas próximas umas das outras na visualização. Outras técnicas como os algoritmos de aprendizado de máquina *Self Organizing Map* e *t-Distributed Stochastic Neighbor Embedding* também foram utilizados para projetar documentos em um espaço visual de menor dimensão mas eles não serão o foco deste trabalho.

Com relação à segunda questão de pesquisa sobre os trabalhos e análises visuais conduzidos com os dados bibliométricos da plataforma Lattes os principais estudos tinham como finalidade criar redes de colaboração entre pesquisadores, universidades e grupos de pesquisa utilizando técnicas de processamento de linguagem natural e teoria de grafos. Além disso, outros artigos criaram mapeamentos dos tópicos estudados dentro das universidades e grupos de pesquisa a partir da análise dos dados bibliométricos da plataforma Lattes. Vale destacar que nenhum dos estudos selecionados nessa revisão de trabalhos correlatos aplicaram técnicas de projeção multidimensional para visualização de dados bibliométricos da Plataforma Lattes.

4 Metodologia

Há diferentes tipos de pesquisas do ponto de vista da abordagem da pesquisa, sua natureza, seus objetivos e seus procedimentos técnicos. Tendo isso em vista, o objetivo dessa seção é conceituar a metodologia de pesquisa que será adotada nesse trabalho de mestrado, ou seja, quais serão as modalidades de pesquisa adequadas para alcançar o fim proposto dessa pesquisa.

Do ponto de vista da abordagem de pesquisa, há dois tipos principais, uma pesquisa pode ser categorizada como qualitativa ou quantitativa. A pesquisa quantitativa é baseada na medição da quantidade. É aplicável a fenômenos que podem ser expressos em termos de quantidade. A pesquisa qualitativa, por outro lado, preocupa-se com o fenômeno qualitativo, ou seja, fenômenos relacionados ou envolvendo qualidade ou tipo.

Esse trabalho terá uma abordagem quantitativa, significa que os resultados alcançados podem ser replicados. A pesquisa quantitativa se centra na objetividade e considera que a realidade só pode ser compreendida com base na análise de dados brutos, recolhidos com o auxílio de instrumentos padronizados e neutros. A pesquisa quantitativa recorre à linguagem matemática para descrever as causas de um fenômeno e as relações entre variáveis (KOTHARI, 2004).

Com relação à natureza da pesquisa, ela pode ser classificada como pesquisa aplicada ou pesquisa básica, pura. A pesquisa aplicada visa encontrar uma solução para um problema imediato enfrentado por uma sociedade ou uma organização industrial e/ou empresarial, enquanto a pesquisa pura preocupa-se principalmente com generalizações e com a formulação de uma teoria (KOTHARI, 2004).

Quanto à natureza da pesquisa esse trabalho será aplicado, ou seja, o objetivo é gerar conhecimentos para aplicação prática, dirigidos à solução de problemas. O objetivo central da pesquisa aplicada é descobrir uma solução para um problema prático enfrentado por uma sociedade ou uma organização industrial.

Uma pesquisa pode ser categorizada também com relação ao seu objetivo, em que o objetivo da pesquisa pode ser dividido em descritivo ou analítico. Na pesquisa analítica o pesquisador deve usar fatos ou informações já disponíveis e analisar estes para fazer uma avaliação crítica do material. A pesquisa descritiva inclui pesquisas e investigações

para apuração de fatos de diferentes tipos. O principal objetivo da pesquisa descritiva é a descrição do estado de assuntos como existe no presente (KOTHARI, 2004).

Esse trabalho tem um objetivo de pesquisa descritivo, ele visa descrever os fatos, as características, os fenômenos de determinada realidade. Busca estabelecer relações entre as variáveis estudadas.

Para desenvolver uma pesquisa é necessário selecionar o procedimento técnico que será utilizado para alcançar os objetivos do estudo. Nesse trabalho será utilizado o procedimento técnico de modelagem. Esse procedimento usa técnicas matemáticas para descrever o funcionamento e características de um sistema.

Nas próximas seções serão apresentadas os materiais e métodos desse trabalho.

4.1 Materiais

O conjunto de dados desse trabalho é composto pelas informações da plataforma Lattes dos pesquisadores e alunos da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH USP), esses dados contém atualização até o ano de 2020. Os dados bibliométricos que serão analisados estão inseridos dentro do módulo de Produção da plataforma Lattes, especificamente será analisado o tópico de "artigos publicados".

O conjunto de dados está separado por grupos de pesquisa pertencentes à EACH USP e para cada grupo de pesquisa tem um conjunto de dados referente à produção bibliográfica. Os itens presentes dentro da produção bibliográfica onde o pesquisador insere as informações são artigos aceitos para publicação, artigos publicados, demais tipos de produção bibliográfica, livros e capítulos, textos em jornais ou revistas e trabalhos em eventos.

Nesse estudo foi analisado o item de artigos publicados, as informações principais que são cadastradas nesse item e que estão presentes no conjunto de dados são título do artigo, ano de publicação do artigo, idioma, país de publicação, título do periódico ou revista de publicação, nomes dos autores, entre outros. A análise multidimensional de dados realizada nesse trabalho foi composta pelas informações obtidas dos títulos dos artigos publicados para cada grupo de pesquisa analisado pertencente à EACH USP.

Os grupos de pesquisa analisados nesse estudo foram os grupos pertencentes ao programa de Pós-graduação em Sistemas de Informação, ao programa de Pós-graduação de

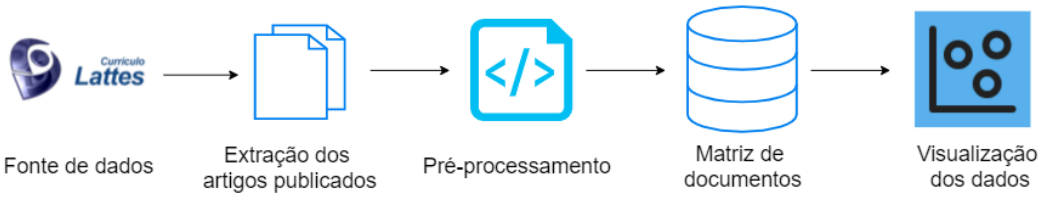
Têxtil e Moda e o grupo de Astrofísica. Os grupos de pesquisa de Sistemas de Informação contém 863 artigos publicados que pertencem a cinco grupos de pesquisa: Gestão de Sistemas de Informação (GESI), Grupo de Políticas Públicas de Acesso à Informação (GPOPAI), Grupo de Pesquisa em Inteligência Artificial (GrIA), Laboratório de Aplicações de Informática em Saúde (LApIS) e o Grupo de Pesquisa em Modelagem de Sistemas Complexos (GRIFE). O grupo de pesquisa de Têxtil e Moda contém 99 artigos publicados e o grupo de pesquisa de Astrofísica contém 76 artigos publicados. Considerando todos os grupos de pesquisa descritos anteriormente foram analisados o total de 1038 artigos publicados.

4.2 Métodos

O algoritmo para representação visual dos dados bibliométricos da Plataforma Lattes pode ser resumido em cinco etapas principais: obtenção dos currículos da plataforma Lattes, extração do módulo de artigos publicados, pré-processamento textual, geração de matriz de documentos e visualização da projeção multidimensional dos dados.

O conhecimento utilizado para a construção desta abordagem corresponde a um levantamento de várias abordagens utilizadas na literatura relacionada sobre projeção multidimensional dos dados. A Figura 8 representa graficamente as cinco etapas principais desse trabalho.

Figura 8 – Resumo do algoritmo



Fonte: Patrícia Salles Escarassatti, 2021

O algoritmo desenvolvido nesse trabalho foi implementado na linguagem de programação Python devido sua grande quantidade de bibliotecas disponíveis para análise e visualização de dados. O algoritmo inicia com a aplicação das etapas de pré-processamento, a análise textual dos artigos ocorrerá a partir do título dos artigos publicados. A primeira etapa realizada foi excluir os títulos dos artigos duplicados presentes em um mesmo grupo de pesquisa. Para que a análise textual seja realizada de uma forma adequada foi necessário

coletar os artigos publicados em uma mesma língua, no caso desse estudo a língua escolhida foi a língua inglesa, pois a grande maioria dos artigos publicados estão em inglês.

A segunda etapa do pré-processamento foi realizar a remoção das *stopwords* que são as palavras consideradas irrelevantes para a linguagem e que tem uma alta frequência. O objetivo da remoção dessas palavras era diminuir o alto volume dessas palavras que afeta negativamente o agrupamento de documentos textuais.

Na sequência foi realizado o terceiro passo denominado como processo de *stemming* que consiste em remover prefixos e sufixos de palavras, dessa maneira essas palavras serão representadas pelo seu radical. O objetivo era representar as palavras que são derivadas umas das outras pelo seu radical com a finalidade de mapear um grupo de palavras por um mesmo radical.

A quarta etapa do pré-processamento textual foi transformar as palavras em *tokens*, esse processo consiste em representar cada palavra do texto em unidades distintas, assim cada palavra do documento será representado como um *token*. O objetivo desse processo foi sumarizar a frequência dessas palavras em cada artigo para posteriormente gerar um histograma das palavras mais frequentes contidas na coleção dos títulos dos artigos publicados.

Com a construção do histograma de frequência dos termos foi possível obter a curva de Zipf. A curva de Zipf é desenhada considerando a ordenação das frequências das palavras de forma decrescente e a partir da curva podemos definir um limiar para excluir os termos menos significativos. Há dois limiares utilizados para remover essas palavras pouco representativas, Luhn especificou esses dois cortes. Para escolher onde esses cortes estarão definidos e serão realizados foi aplicado o ponto de Goffman.

Após a escolha das palavras mais significativas dos documentos foi realizada uma seleção dos artigos que contém essas palavras. Com a seleção desses artigos e após todo o pré-processamento textual realizado nesses documentos foi inciado a etapa da criação do modelo vetorial de documentos, onde esses documentos foram representados como uma matriz. O cálculo dos pesos de cada palavra contida nessa coleção de documentos foi realizado utilizando o método tf-idf.

Finalizado a etapa de pré-processamento e criação da matriz de documentos iniciou-se o processo para criação da visualização dos dados. A matriz de documentos obtida foi processada pelos algoritmos de projeção multidimensional com a finalidade de converter os

dados multidimensionais em um espaço de dimensão inferior, mantendo as características intrínsecas desses dados.

A estratégia de projeção multidimensional utiliza diferentes tipos de algoritmos e combinações que serão descritos na sequência.

O início na etapa de visualização dos dados começou com a utilização dos três algoritmos de *Multidimensional scaling* e suas derivações. O *Classical multidimensional scaling* foi implementado utilizando a distância de cosseno para o cálculo da matriz de distância e com a finalidade de gerar uma projeção com duas dimensões. O algoritmo *Weighted Multidimensional Scaling* foi implementado utilizando a formulação chamada *Sammon's mapping* que utiliza a matriz de documentos para gerar uma projeção com duas dimensões. O Isomap também foi implementado realizando uma combinação de diferentes parâmetros do modelo: o método utilizado para encontrar o caminho mais curto entre cada par de pontos (algoritmo de Floyd-Warshall e o algoritmo de Dijkstra), o algoritmo a ser utilizado para pesquisa de vizinhos mais próximos (*KD tree*, *ball tree* e *brute force*) e o método para cálculo de autovetores e autovalores (Método de Arnaldi e Lapack). A combinação escolhida foi a que retornou o menor valor para a função de perda do Isomap e nas duas análises realizadas utilizando os dados dos programas de Pós-graduação em Sistemas de Informação, do programa de Pós-graduação de Têxtil e Moda e do grupo de Astrofísica os parâmetros escolhidos foram o algoritmo de Dijkstra, o algoritmo utilizado para pesquisa de vizinhos mais próximos escolhido foi o *brute force* e a biblioteca de cálculo de autovetores e autovalores escolhida foi Lapack.

Dando continuidade a etapa de visualização dos dados o algoritmo *Force-directed placement* foi executado utilizando sua ideia básica que é baseada em grafos. O *Force-directed placement* foi executado utilizando como entrada a matriz de documentos obtida após as etapas de pré-processamento e utilizando a distância euclidiana como parâmetro de cálculo de distância. O algoritmo de Chalmers também foi executado tendo como dataset de entrada a matriz de documentos, e mantendo os parâmetros padrões do algoritmo como a distância sendo a euclidiana. O algoritmo *Force scheme* também utilizou como dados de entrada a matriz de documentos para gerar uma projeção com duas dimensões.

Para finalizar a etapa de visualização dos dados o algoritmo *Least square projection* foi implementado utilizando como dados de entrada a matriz de documentos gerada na etapa de pré-processamento e o dado de saída do algoritmo foi uma projeção com duas dimensões.

Na próxima seção serão apresentados os resultados obtidos por cada uma das abordagens.

5 Resultados e Discussões

As etapas de pré-processamento discutidas no capítulo de Metodologia são realizadas com o objetivo de gerar uma matriz de documentos que será utilizada nos algoritmos que irão gerar as projeções em um espaço dimensional de duas dimensões. Todo o processo de análise textual foi aplicado nos artigos publicados dos grupos de pesquisa dos programas de Pós-graduação em Sistemas de Informação, do programa de Pós-graduação de Têxtil e Moda e do grupo de Astrofísica.

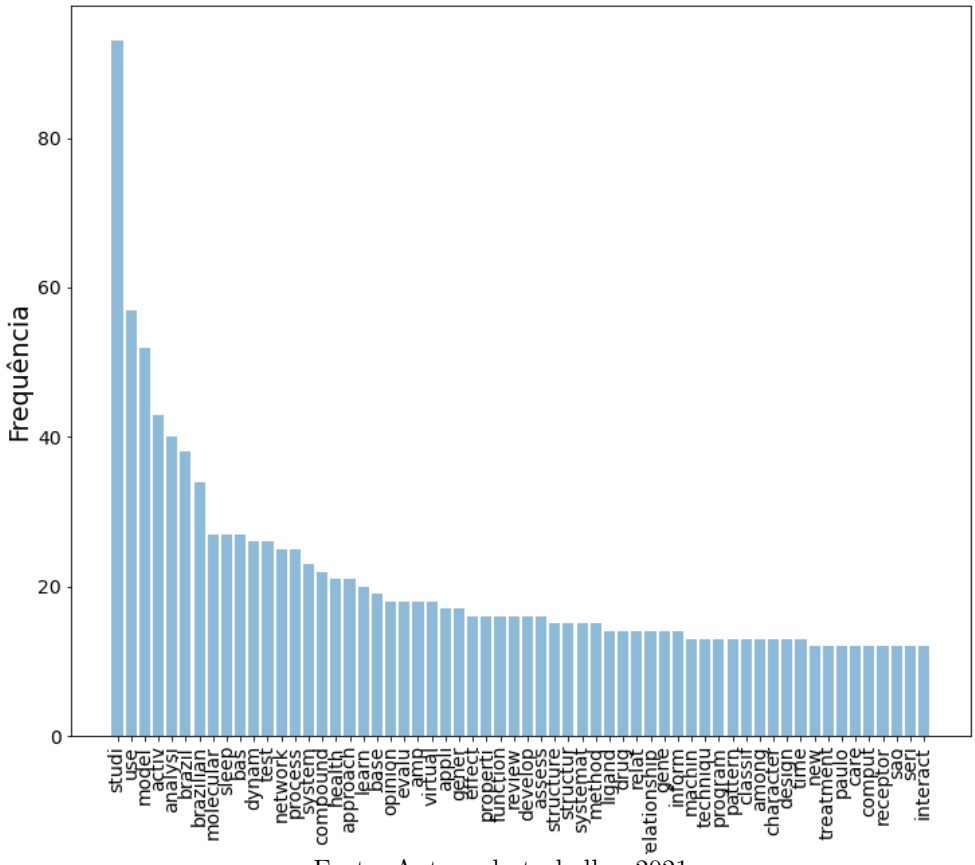
Realizadas as etapas de pré-processamento textual foi gerado um histograma das palavras mais frequentes contidas na coleção dos títulos dos artigos publicados dos grupos de pesquisa estudados nesse trabalho. E a partir desse histograma de palavras mais frequentes foi obtido a curva de Zipf e utilizando essa curva que podemos excluir os termos menos significativos. O resultado do histograma dos artigos dos grupos de pesquisa do programa de Pós-graduação em Sistemas de Informação está apresentado na Figura 9 e o resultado do histograma dos artigos dos grupos de pesquisa de Têxtil e Moda e o grupo de pesquisa de Astrofísica está apresentado na Figura 10.

Com o resultado do histograma dos artigos dos grupos de pesquisa do programa de Pós-graduação em Sistemas de Informação que está apresentado na Figura 9 observa-se que os radicais mais frequentes encontrados nos artigos em estudo tem relação com Programa de Pós-graduação em Sistemas de Informação e com seus temas estudo, por exemplo, "studi", "model", "network", "process", "system", "learn", entre outros.

Analisando o histograma dos artigos dos grupos de pesquisa de Têxtil e Moda e o grupo de pesquisa de Astrofísica que está apresentado na Figura 10 observa-se que há relação entre os radicais mais frequentes encontrados nos artigos e os temas de estudo de Têxtil e Moda e de Astrofísica, por exemplo, "textil", "product", "wind", "fiber", "star", "magnet".

Após a definição das palavras mais significativas dos documentos foi realizada uma seleção dos artigos que contém essas palavras. Após todo o pré-processamento textual realizado nesses documentos iniciou a etapa da criação do modelo vetorial de documentos, onde esses documentos serão representados como uma matriz. Foi realizado o cálculo dos pesos de cada palavra contida nessa coleção de documentos utilizando o método tf-idf.

Figura 9 – Histograma das palavras mais frequentes dos grupos de pesquisa do programa de Pós-graduação em Sistemas de Informação



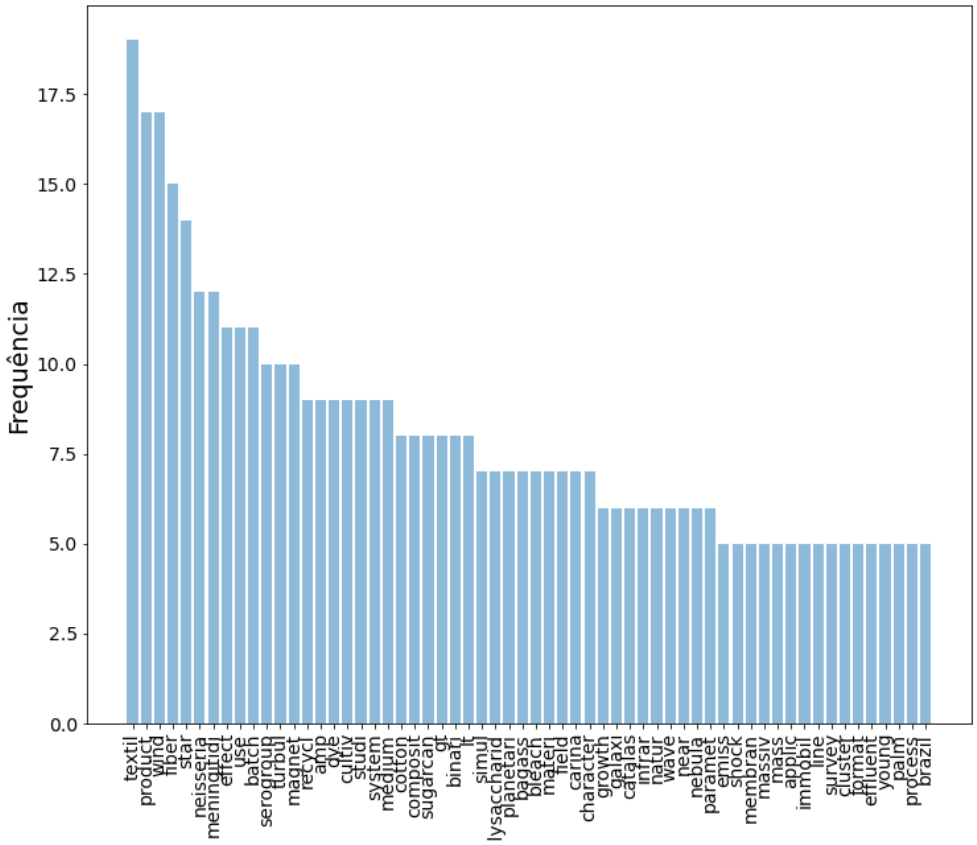
Fonte: Autora do trabalho, 2021

Finalizado a etapa de pré-processamento e criação da matriz de documentos se deu início ao processo para criação da visualização dos dados.

Para análise visual de dados multidimensionais é comum usar técnicas de redução de dimensionalidade que projetam os pontos multidimensionais para pontos em um espaço visual de baixa dimensão e, normalmente, os pontos projetados são exibidos na forma de gráficos de dispersão em duas dimensões. O método de projeção deve preservar as distribuições dos dados multidimensionais o tanto quanto possível com o objetivo de obter informações sobre esses dados (ETEMADPOUR; OLK; LINSEN, 2014). Será apresentado nessa seção alguns resultados das técnicas de projeção dos dados em espações visuais bidimensionais.

Os resultados da projeção dos dados multidimensionais dos artigos dos grupos de pesquisa do programa de Pós-graduação em Sistemas de Informação estão apresentados na Figura 11. Foram obtidos três gráficos de dispersão para cada uma das técnicas de *Multidimensional scaling*, o primeiro gráfico representa o resultado da técnica *Classical*

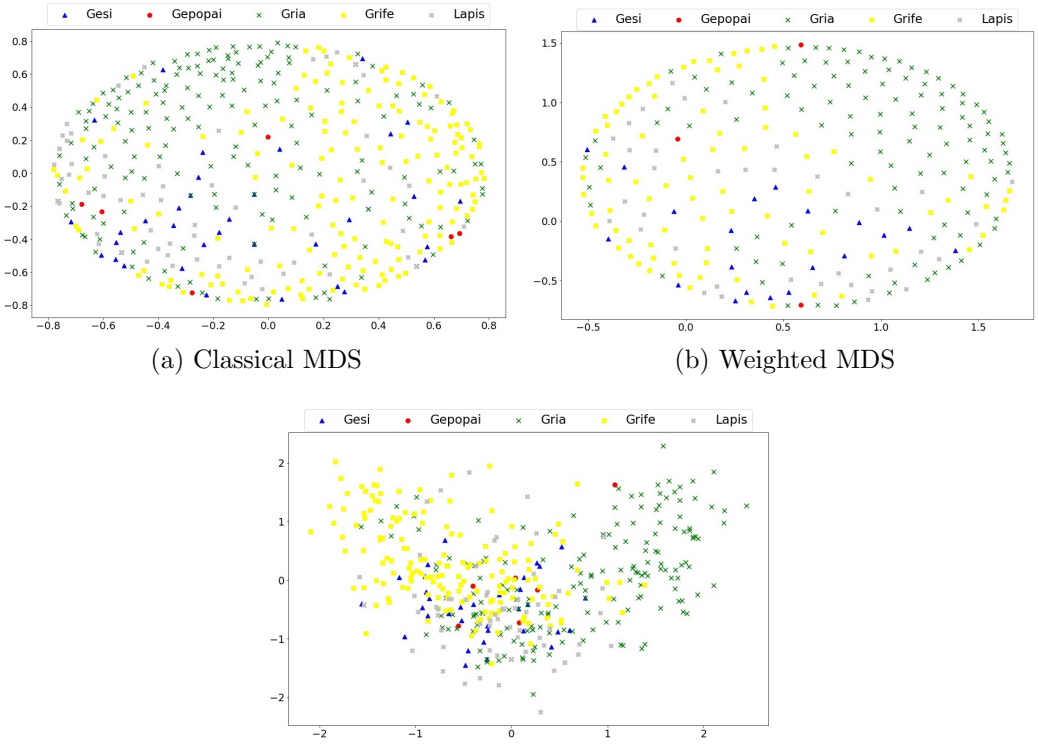
Figura 10 – Histograma das palavras mais frequentes dos grupos de pesquisa de Têxtil e Moda e o grupo de pesquisa de Astrofísica



Fonte: Autora do trabalho, 2021

MDS, o segundo gráfico representa o algoritmo *Weighted* MDS e o terceiro gráfico é a técnica Isomap. Cada cor dos pontos nos gráficos de dispersão representa um grupo de pesquisa de Sistemas de Informação e cada ponto no gráfico de dispersão representa um artigo publicado, além disso os valores das duas dimensões de cada ponto posicionado na projeção visual são representados pelos eixos X (horizontal) e Y (vertical) dos gráficos. Observa-se que os grupos de pesquisa não ficam segregados nas visualizações o que pode ser um indicativo desses grupos estudarem tópicos que são relacionados entre si. Outro ponto observado é que a técnica Isomap apresenta um gráfico de dispersão diferente das outras duas técnicas.

Outra análise realizada foi comparar conjuntamente os artigos dos grupos de pesquisa de Têxtil e Moda e o grupo de pesquisa de Astrofísica. Teoricamente esses grupos de pesquisa estudam tópicos de campos de pesquisa que não são muito relacionados entre si. Os resultados obtidos estão apresentados na Figura 12, a cor verde dos pontos nos gráficos de dispersão representa os artigos publicados do grupo de pesquisa de Têxtil e

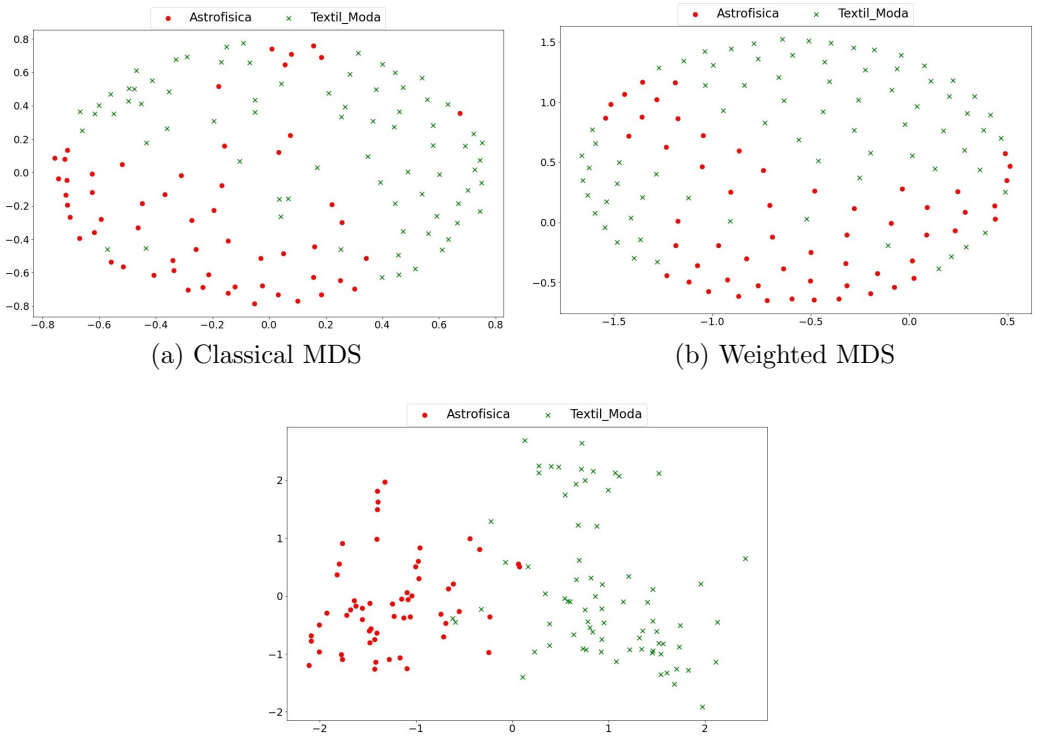


(a) Classical MDS

(b) Weighted MDS

(c) Isomap

Figura 11 – *Grupos de pesquisa de Sistemas de Informação*



(a) Classical MDS

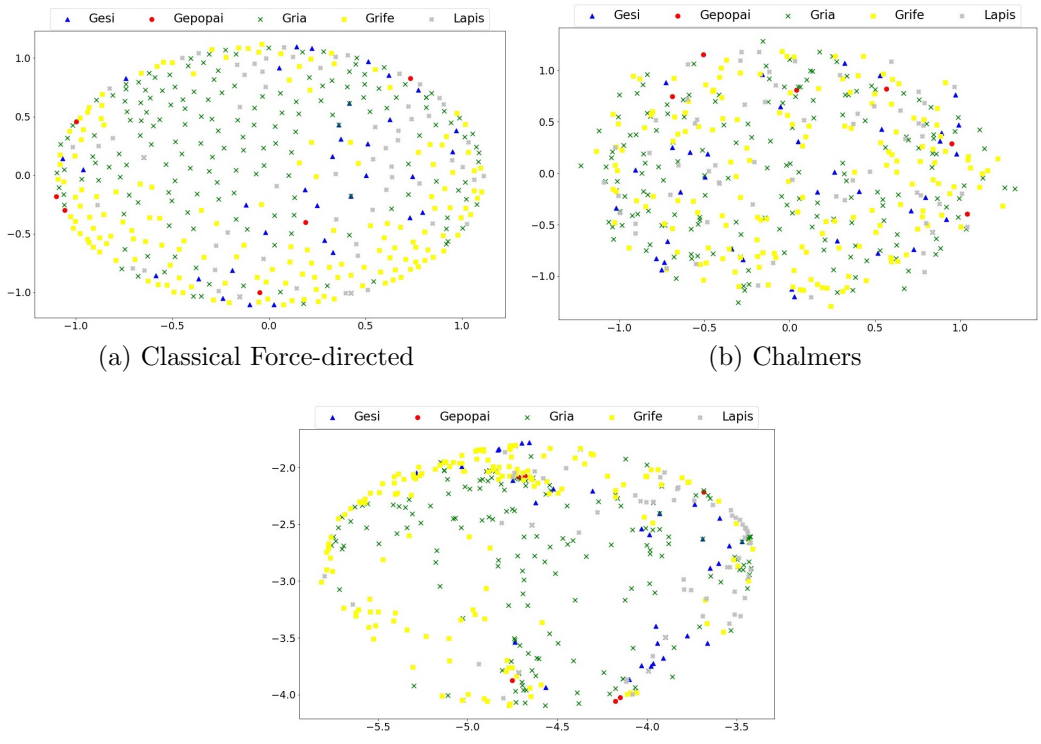
(b) Weighted MDS

(c) Isomap

Figura 12 – *Grupos de pesquisa de Têxtil e Moda e Astrofísica*

Moda e a cor vermelha representa os artigos publicados do grupo de pesquisa de Astrofísica. Observa-se que, diferentemente dos grupos de pesquisa de Sistemas de Informação, os artigos publicados de Têxtil e Moda e de Astrofísica tendem a se separar mais facilmente na visualização dos dados o que pode confirmar nossa hipótese inicial de que esses grupos de pesquisas estudam tópicos que não se relacionam entre si.

O algoritmo de *Force-directed placement* foi aplicado no mesmo conjunto de documentos que foi estudado anteriormente com a técnica *Multidimensional scaling*. Os resultados obtidos com essa técnica de projeção estão apresentados na Figura 13 e na Figura 14. Quando o algoritmo de *Force-directed placement* e suas variações foram aplicados nos artigos publicados dos grupos de pesquisa de Sistemas de Informação não é observado uma separação visual desses grupos na projeção obtida conforme demonstrado na Figura 13.



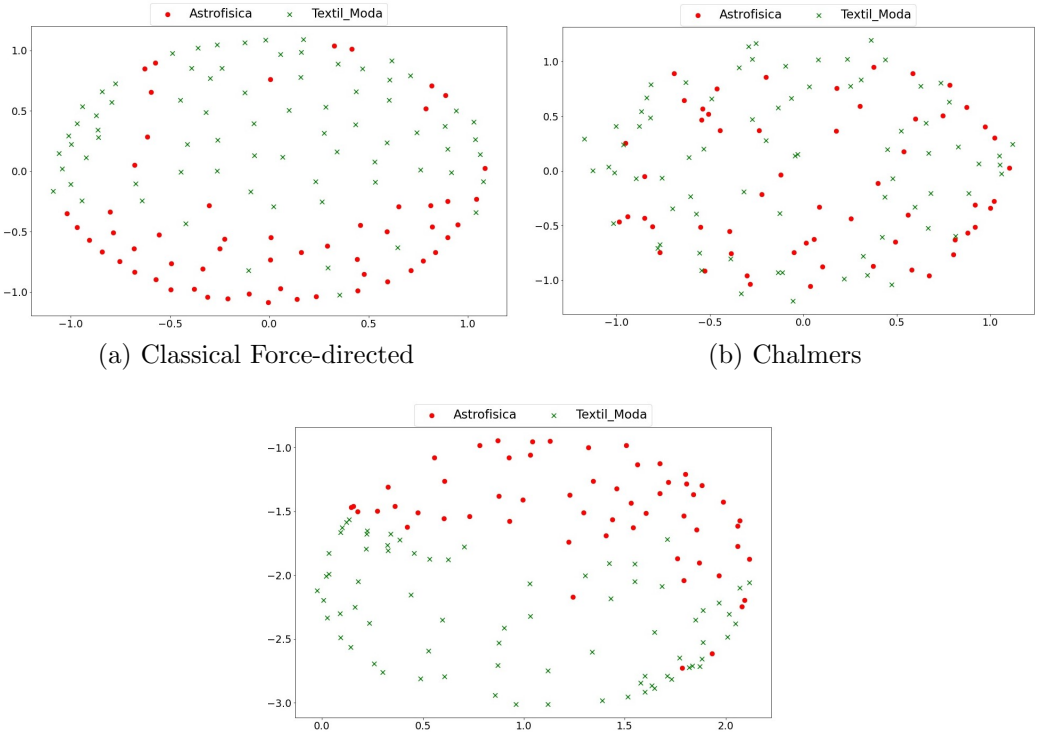
(a) Classical Force-directed

(b) Chalmers

(c) Force scheme

Figura 13 – *Grupos de pesquisa de Sistemas de Informação*

Com relação à aplicação das técnicas de *Force-directed placement* nos artigos publicados dos grupos de pesquisa de Têxtil e Moda e do grupo de pesquisa de Astrofísica é observado uma melhor segregação visual desses grupos conforme a Figura 14.



(a) Classical Force-directed (b) Chalmers

(c) Force scheme

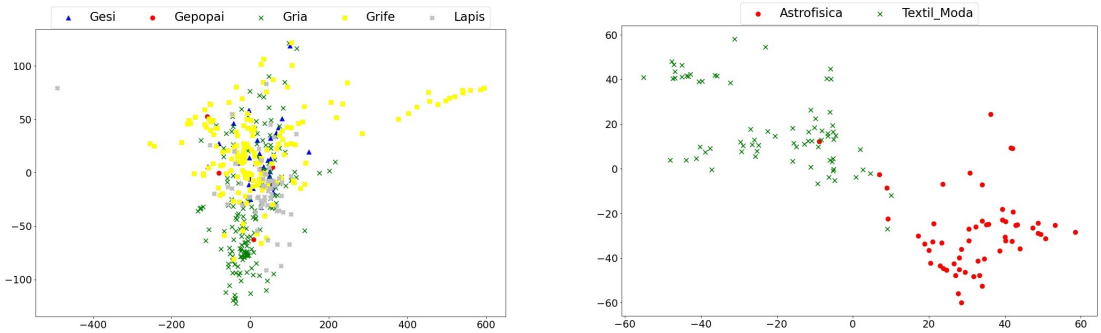
Figura 14 – *Grupos de pesquisa de Têxtil e Moda e Astrofísica*

No entanto, o algoritmo de Chalmers apresentado no item b da Figura 14 não conseguiu separar os grupos de Têxtil e Moda e Astrofísica possivelmente por ser uma abordagem que reduz a complexidade das iterações usando amostras de dados para determinar quais instâncias são ligadas entre si, essa amostragem pode ocasionar perda de informação dos artigos dos grupos de pesquisa (CHALMERS, 1996).

Por fim, foi aplicado o algoritmo *Least Square Projection* nos artigos publicados dos grupos de pesquisa dos programas de Pós-graduação em Sistemas de Informação e do programa de Pós-graduação de Têxtil e Moda e do grupo de pesquisa de Astrofísica, os resultados obtidos estão apresentados na Figura 15.

Novamente é observado que há separação visual apenas dos grupos de pesquisa de Têxtil e Moda e do grupo de pesquisa de Astrofísica. Visualmente a técnica *Least Square Projection* é a que melhor separa os dados desses grupos de pesquisa quando comparado com os outros algoritmos apresentados anteriormente.

Para finalizar a análise das técnicas de projeção multidimensional foi empregado as abordagens e métricas de avaliação *Neighborhood Hit* e o Coeficiente de Silhueta para comparar os diferentes layouts produzidos pelas técnicas de *Multidimensional Scaling*,



(a) Grupos de pesquisa de Sistemas de Informação

(b) Grupos de pesquisa de Têxtil e Moda e Astrofísica

Figura 15 – *Least Square Projection*

Force-directed placement e *Least Square Projection*. As Figuras 16 e 17 apresentam os resultados da avaliação de *Neighborhood Hit* e na Tabela 4 e na Tabela 5 são apresentados os resultados do Coeficiente de Silhueta.

Tabela 4 – Avaliação comparativa entre as técnicas de projeção utilizando o Coeficiente de Silhueta nos Grupos de Pesquisa de Sistemas de Informação

Técnicas de projeção	Coeficiente de Silhueta
Multidimensional Scaling	-0,037
Weighted Multidimensional Scaling	-0,044
Force-directed Placement - Force Scheme	-0,065
Force-directed Placement	-0,077
Multidimensional Scaling - Isomap	-0,086
Force-directed Placement - Algoritmo de Chalmers	-0,105
Least Square Projection	-0,178

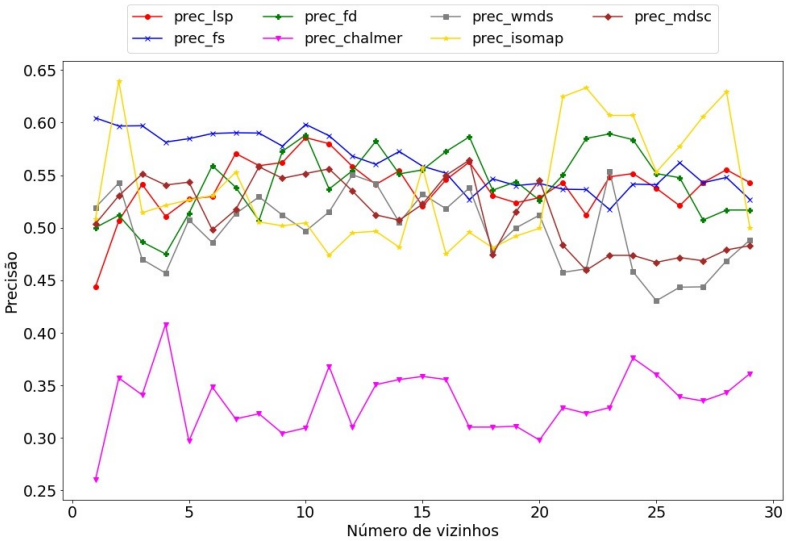
A avaliação comparativa das técnicas de projeção para os grupos de pesquisa de Sistemas de Informação não apresenta resultados satisfatórios conforme os valores do Coeficiente de Silhueta apresentados na Tabela 4. Coeficiente de Silhueta próximo de 0 significa que há uma baixa qualidade na formação de grupos gerados pelas técnicas de projeção. Os valores do Coeficiente de Silhueta apresentados na Tabela 5 que contém a avaliação comparativa entre as técnicas de projeção dos Grupos de Pesquisa de Têxtil e Moda e do Grupo de Pesquisa de Astrofísica apresentaram melhores resultados, com Coeficiente de Silhueta mais próximo do valor 1, para as técnicas *Least Square Projection* e *Multidimensional Scaling - Isomap* e esse resultado pode ser comprovado avaliando visualmente as Figuras 15 e 12 item c, respectivamente, onde o grupo de Astrofísica se separa mais facilmente do grupo de Têxtil e Moda.

Tabela 5 – Avaliação comparativa entre as técnicas de projeção utilizando o Coeficiente de Silhueta nos Grupos de Pesquisa de Têxtil e Moda e do Grupo de Pesquisa de Astrofísica

Técnicas de projeção	Coeficiente de Silhueta
Least Square Projection	0,587
Multidimensional Scaling - Isomap	0,424
Force-directed Placement - Force Scheme	0,232
Multidimensional Scaling	0,192
Force-directed Placement	0,164
Weighted Multidimensional Scaling	0,138
Force-directed Placement - Algoritmo de Chalmers	0,009

Observando o resultado da abordagem *Neighborhood Hit* apresentado na Figura 16 que avalia as projeções geradas para os Grupos de Pesquisa de Sistemas de Informação nenhuma das técnicas consegue ter um resultado de precisão maior que 70% o que sugere que os estudos de Sistemas de Informação não são facilmente separáveis visualmente e que possivelmente esses grupos estudam assuntos que são relacionados entre si. Uma taxa de precisão próxima de 70% significa que um artigo tem cerca de 70% de precisão de ser classificado corretamente para determinado grupo de pesquisa. O resultado da métrica de avaliação *Neighborhood Hit* para a projeção visual utilizando o algoritmo de Chalmers tem o pior resultado comparando com os outros algoritmos, com valores de precisão em torno de 35% e visualmente esse resultado pode ser comprovado avaliando a projeção na Figura 13 item b.

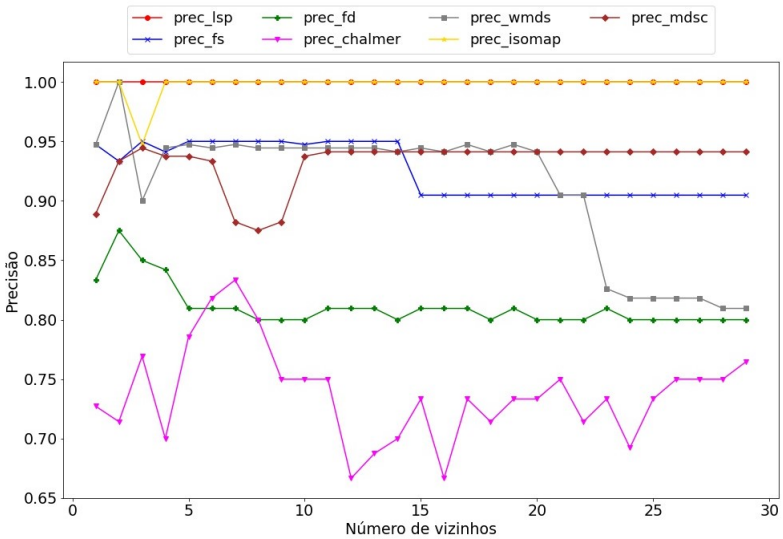
Figura 16 – Avaliação comparativa entre as técnicas de projeção utilizando a abordagem *Neighborhood Hit* nos Grupos de Pesquisa de Sistemas de Informação



Fonte: Autora do trabalho, 2021

Quando é realizada a avaliação comparando as técnicas de projeção para os artigos publicados dos grupos de pesquisa de Têxtil e Moda e do grupo de pesquisa de Astrofísica observa-se resultados mais satisfatórios. Os melhores resultados apresentados pela abordagem *Neighborhood Hit* pertencem às técnicas de projeção *Least Square Projection* e *Multidimensional Scaling - Isomap*, assim como o resultado da métrica de avaliação do Coeficiente de Silhueta para esses grupos de pesquisa. Esse resultado significa que um artigo tem praticamente 100% de precisão de ser classificado para o grupo de pesquisa correto quando ele é projetado por uma das técnicas de projeção *Least Square Projection* ou *Multidimensional Scaling - Isomap*, ou seja, os grupos formados por essa projeção estão visualmente distantes um dos outros e bem agrupados. Para as duas abordagens de avaliação o algoritmo de projeção de Chalmers apresenta um resultado inferior quando comparado aos outros algoritmos quando se trata em separar os grupos de pesquisa de Têxtil e Moda do grupo de pesquisa de Astrofísica, o que pode ser confirmado pela observação do resultado visual apresentado na Figura 14 item b.

Figura 17 – Avaliação comparativa entre as técnicas de projeção utilizando a abordagem *Neighborhood Hit* nos Grupos de Pesquisa de Têxtil e Moda e do Grupo de Pesquisa de Astrofísica



Fonte: Autora do trabalho, 2021

6 Conclusões e Trabalhos Futuros

O estudo mostrou que as técnicas de projeção *Least Square Projection* e *Multidimensional Scaling - Isomap* apresentaram os melhores resultados quando foram aplicados nos artigos publicados dos grupos de pesquisa de Têxtil e Moda e do grupo de pesquisa de Astrofísica que foram extraídos da plataforma Lattes. O Coeficiente de Silhueta para essas duas técnicas de projeção ficam próximas de 0,5 o que indica uma maior facilidade em separar visualmente os grupos de pesquisa. Com relação à metrica de avaliação *Neighborhood Hit* as duas técnicas de projeção citadas, *Least Square Projection* e *Multidimensional Scaling - Isomap*, apresentam um valor de precisão próximo de 100%. Esse valor de precisão nos mostra que um artigo tem cerca de 100% de precisão em ser classificado corretamente para um grupo de pesquisa.

Quando foram avaliados os grupos de pesquisa de Sistemas de Informação ficou evidente que não há uma separação visual desses grupos, o que indica que os artigos publicados pertencentes a este Programa de Pós-gradução têm temas muito relacionados uns aos outros. Esse resultado da projeção visual dos grupos de pesquisa de Sistemas de Informação são confirmados com os valores apresentados pelas métricas de avaliação do Coeficiente de Silhueta e *Neighborhood Hit*.

As informações do currículo Lattes são cadastradas e exibidas de forma individual e está associada a cada pessoa. Esta característica não fornece uma maneira de descobrir as produções bibliográficas de um determinado grupo. No entanto, a abordagem estudada nesse trabalho, principalmente as técnicas *Least Square Projection* e *Multidimensional Scaling - Isomap*, pode ser utilizada para avaliar e analisar dados bibliométricos, identificar como grupos de pesquisa podem estar relacionados entre si e auxiliar na elaboração de relatórios sobre as produções científicas e os projetos de diferentes grupos de pesquisa presentes no currículo Lattes.

A metodologia empregada para extração e processamento dos dados textuais se mostra adequada para análise dos dados bibliométricos do Lattes, as técnicas de projeção visual tiveram um bom desempenho para representar os grupos de pesquisa avaliados nesse trabalho. Dessa forma, os relatórios que são normalmente criados por análise manual dos dados do currículo Lattes de cada membro do grupo, a fim de obter uma correlação entre

produções bibliográficas de grupos de pesquisa distintos, podem ser gerados utilizando os procedimentos empregados nesse trabalho.

A limitação encontrada neste trabalho está relacionada ao escopo de pesquisa onde definiu-se que seriam avaliados apenas alguns grupos de pesquisa pertencentes à Escola de Artes, Ciência e Humnanidades da Universidade de São Paulo. Além disso, há uma limitação qualitativa pois as informações inseridas na plataforma Lattes é de inteira responsabilidade do usuário e esses dados não passam por nenhuma verificação com relação a integridade dessas informações. Outra limitação encontrada nesse trabalho se refere a utilização apenas dos títulos dos artigos que não contém tantos termos para criação das projeções multidimensionais. Poderiam ser utilizados outros termos para realizar a análise visual dos artigos publicados, como por exemplos, as palavras-chave dos artigos em estudo.

Ficam como trabalhos futuros a utilização das técnicas de projeção multidimensional presentes nesse estudo e/ou a utilização de outras técnicas de projeção multidimensional mais sofisticadas. Outra possibilidade futura é a expansão da análise de outros grupos de pesquisa presentes no currículo Lattes, além da utilização dessas técnicas de projeção para realizar análise de diferentes informações textuais presentes na plataforma Lattes. Além disso, os algoritmos estudados nesse trabalho podem ser aplicados em outras áreas de mineração de dados de texto.

Outra sugestão de trabalhos futuros poderiam ser aplicados controles de robustez para avaliar se há uma diferença significativa no resultado das projeções multidimensionais se não houvesse a aplicação da etapa de pré-processamento de texto, como a remoção de *stopwords* e realização do processo de *stemming*.

Referências

ABASI, A. K.; KHADER, A. T.; AL-BETAR, M. A.; NAIM, S.; MAKHADMEH, S. N.; ALYASSERI, Z. A. A. Link-based multi-verse optimizer for text documents clustering. *Applied Soft Computing Journal*, Elsevier, v. 87, n. 106002, 2020. Citado 2 vezes nas páginas 19 e 21.

ABBAS, A.; ZHANG, L.; KHAN, S. U. A literature review on the state-of-the-art in patent analysis. *World Patent Information*, Elsevier, v. 37, p. 3–13, 2014. Citado 2 vezes nas páginas 13 e 51.

ALENCAR, A. B.; PAULOVICH, F. V.; BöRNER, K.; OLIVEIRA, M. C. F. de. Time-aware visualization of document collections. *Proceedings of the ACM Symposium on Applied Computing*, Association for Computing Machinery, p. 997–1004, 2012. Citado 4 vezes nas páginas 18, 50, 51 e 54.

ALIGULIYEV, R. M. Clustering of document collection – a weighting approach. *Expert Systems with Applications*, Elsevier, v. 36, p. 7904–7916, 2009. Citado na página 13.

AMARAL ALINE GRASIELE CARDOSO BRITO, K. G. d. S. R. L. M. Q. R. M.; FARIA, L. I. L. de. Panorama da inteligência competitiva no brasil: os pesquisadores e a produção científica na plataforma lattes. *Perspectivas em Ciência da Informação*, v. 21, p. 97–120, 2016. Citado na página 56.

ANDREOTTI, A. L. D.; SILVA, L. F.; ELER, D. M. Hybrid visualization approach to show documents similarity and content in a single view. *Information (Switzerland)*, MDPI AG, v. 9, n. 129, 2018. Citado na página 54.

ANTONIO, A. de; MORAL, C.; KLEPEL, D.; ABENTE, M. J. 3d gesture-based exploration and search in document collectionsn. *17th International Conference on Electronic Publishing, ELPUB 2013*, IOS Press, p. 13–22, 2013. Citado 2 vezes nas páginas 51 e 52.

BELTER, C. W. Bibliometric indicators: opportunities and limits. *Journal of the Medical Library Association : JMLA*, v. 103,4, p. 219–221, 2015. Citado na página 14.

BERGER, M.; MCDONOUGH, K.; SEVERSKY, L. M. cite2vec: Citation-driven document exploration via word embeddings. *IEEE Transactions on Visualization and Computer Graphics*, IEEE Computer Society, v. 23, n. 7539398, p. 691–700, 2017. Citado na página 51.

BERKHIN, P. *A survey of clustering data mining techniques*. [S.l.]: Springer Berlin Heidelberg, 2006. Citado na página 37.

BUTKA, P.; PóCSOVá, J. Hybrid approach for visualization of documents clusters using ghsom and sammon projection. *8th IEEE International Symposium on Applied Computational Intelligence and Informatics, SACI 2013*, IEEE, n. 6608994, p. 337–342, 2013. Citado na página 51.

CHALMERS, M. A linear iteration time layout algorithm for visualising high-dimensional data. *Proceedings of Seventh Annual IEEE Visualization Conference*, IEEE, n. 5456773, 1996. Citado 3 vezes nas páginas 24, 26 e 69.

CHEN, C.; IBEKWE-SANJUAN, F.; HOU, J. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, Wiley Online Library, v. 61, p. 1386–1409, 2010. Citado na página 51.

COSTA, E. d. S. P. B. M. G.; MACEDO, G. R. de. Scientific collaboration in biotechnology: The case of the northeast region in brazil. *Scientometrics*, v. 95, 2013. Citado na página 56.

COSTA, T. D. E.; DIAS, P. A system for discovery of knowledge in data repository education. *International Journal of Information and Education Technology*, v. 9, p. 535–538, 2019. Citado na página 56.

DAMACENO LUCIANO ROSSI, R. M. R.; MENA-CHALCO, J. The brazilian academic genealogy: Evidence of advisor-advisee relationships through quantitative analysis. *Scientometrics*, v. 119, 2019. Citado na página 56.

DIAS, A. G.; MILIOS, E. E.; OLIVEIRA, M. C. F. de. Trivir: A visualization system to support document retrieval with high recall. *Proceedings of the ACM Symposium on Document Engineering, DocEng 2019*, Association for Computing Machinery, n. 3345401, 2019. Citado 2 vezes nas páginas 23 e 51.

DUNNE, C.; SHNEIDERMAN, B.; GOVE, R.; KLAVANS, J.; DORR, B. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, Wiley Online Library, v. 63, p. 2351–2369, 2012. Citado 2 vezes nas páginas 13 e 15.

DUTRA ÁLVARO GUILLERMO ROJAS LEZANA, M. L. D. S.; PINTO, A. L. A bibliometric analysis of the scientific production and collaboration between graduate programs in manufacturing engineering in brazil. *Informação Sociedade: Estudos*, v. 29, 2019. Citado na página 56.

EADES, P. A heuristic for graph drawing. In: *Congressus Numerantium*. [S.l.: s.n.], 1984. p. 149–160. Citado na página 24.

ELER, D. M.; GARCIA, R. E. Using otsu's threshold selection method for eliminating terms in vector space model computation. *17th International Conference on Information Visualisation, IV 2013*, IEEE, n. 6676566, p. 220–226, 2013. Citado 5 vezes nas páginas 13, 38, 50, 51 e 53.

ETEMADPOUR, R.; MOTTA, R. C. da; PAIVA, J. G. de S.; MINGHIM, R.; OLIVEIRA, M. C. F. de; LINSEN, L. Role of human perception in cluster-based visual analysis of multidimensional data projections. *5th International Conference on Information Visualization Theory and Applications, IVAPP 2014*, SciTePress, p. 276–283, 2014. Citado na página 16.

ETEMADPOUR, R.; OLK, B.; LINSEN, L. Eye-tracking investigation during visual analysis of projected multidimensional data with 2d scatterplots. *5th International Conference on Information Visualization Theory and Applications, IVAPP 2014*, SciTePress, p. 233–246, 2014. Citado 4 vezes nas páginas 53, 54, 55 e 65.

FODOR, I. K. A survey of dimension reduction techniques. *Center for Applied Scientific Computing*, 2002. Citado na página 33.

FRANCE, S. L.; CARROLL, D. Two-way multidimensional scaling: A review. *Systems, Man, and Cybernetics, Part C: Applications and Reviews*, IEEE Transactions on systems, v. 41, p. 644–661, 2011. Citado 3 vezes nas páginas 30, 31 e 33.

FRUCHTERMAN, T. M. J.; REINGOLD, E. M. Graph drawing by force-directed placement. *Software: Practice and Experience*, v. 21, p. 1129–1164, 1991. Citado na página 24.

GIANNAKOPOULOS, T.; STAMATOGIANNAKIS, E.; FOUFOULAS, I.; DIMITROPOULOS, H.; MANOLA, N.; IOANNIDIS, Y. Content visualization of scientific corpora using an extensible relational database implementation. *17th International Conference on Theory and Practice of Digital Libraries, TPDL 2013*, Springer, v. 416, p. 101–112, 2013. Citado 2 vezes nas páginas 51 e 55.

GOMEZ-NIETO, E.; ROMAN, F. S.; PAGLIOSA, P.; CASACA, W.; HELOU, E. S.; OLIVEIRA, M. C. F. de. Similarity preserving snippet-based visualization of web search results. *IEEE Transactions on Visualization and Computer Graphics*, IEEE, v. 20, n. 6629989, p. 457–470, 2014. Citado 2 vezes nas páginas 51 e 53.

GROENEN, P. J.; BORG, I. The past, present, and future of multidimensional scaling. *Econometric Institute Report*, 2013. Citado 2 vezes nas páginas 30 e 32.

GUTIéRREZ-SALCEDO, M.; MARTíNEZ, M. Ángeles; MORAL-MUNOZ, J. A.; HERRERA-VIEDMA, E.; COBO, M. Some bibliometric procedures for analyzing and evaluating research fields. *Applied Intelligence*, Springer, v. 48, p. 1275–1287, 2018. Citado na página 14.

HAN, Q.; JOHN, M.; KOCH, S.; ASSENOV, I.; ERTL, T. Labeltransfer - integrating static and dynamic label representation for focus+context text exploration. *International Symposium on Big Data Visual and Immersive Analytics, BDVA 2018*, Institute of Electrical and Electronics Engineers Inc., n. 8533897, 2018. Citado na página 51.

INGRAMN, S.; MUNZNER, T. Dimensionality reduction for documents with nearest neighbor queries. *Neurocomputing*, Elsevier, v. 150, p. 557–569, 2015. Citado na página 51.

ISENBERG, P.; HEIMERL, F.; KOCH, S.; ISENBERG, T.; XU, P.; STOLPER, C. D.; SEDLMAIR, M.; CHEN, J.; M€OLLER, T.; STASKO, J. Vispubdata.org: A metadata collection about ieee visualization (vis) publications. *IEEE Transactions on Visualization and Computer Graphics*, IEEE Computer Society, v. 23, n. 7583708, p. 2199–2206, 2017. Citado 2 vezes nas páginas 13 e 51.

JOHN, M.; HEIMERL, F.; VU, B.-A.; ERTL, T. Visual analysis and exploration of entity relations in document collections. *13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2018*, SciTePress, v. 3, p. 244–251, 2018. Citado na página 51.

KALMUKOV, Y. Automatic assignment of reviewers to papers based on vector space text analysis model. *ACM International Conference Proceeding Series*, Association for Computing Machinery, p. 229–235, 2020. Citado na página 21.

KOTHARI, C. R. *Research Methodology - Methods and Techniques*. 2. ed. [S.l.]: New Age International, 2004. Citado 2 vezes nas páginas 58 e 59.

LANçA, R. M. A. T. A.; GRACIOSO, L. S. Multi and interdisciplinarity in the brazilian postgraduate programs in information science. *Perspectivas em Ciência da Informação*, v. 23, p. 150–183, 2018. Citado na página 56.

LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, IBM, v. 2, p. 159–165, 1958. Citado 2 vezes nas páginas 19 e 20.

MANWAR, A. B.; MAHALLE, H. S.; CHINCHKHEDE, K. D.; CHAVAN, V. A vector space model for information retrieval: A matlab approach. *Indian Journal of Computer Science and Engineering*, v. 3, n. 2, p. 222–229, 2012. Citado 2 vezes nas páginas 18 e 22.

MENA-CHALCO, F. M. L. J.; DIGIAMPIETRI, L. Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, v. 65, p. 1424–1445, 2014. Citado na página 56.

MENA-CHALCO, J. P.; JUNIOR, R. M. C. Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, v. 15, p. 31–39, 2009. Citado 2 vezes nas páginas 14 e 15.

MIAN, P.; CONTE, T.; NATALI, A.; BIOLCHINI, J.; TRAVASSOS, G. A systematic review process to software engineering. v. 32, 2005. Citado na página 40.

MINGHIM, R.; PAULOVICH, F. V.; LOPES, A. de A. Content-based text mapping using multi-dimensional projections for exploration of document collections. *Proceedings of SPIE - The International Society for Optical Engineering*, SPIE Digital Library, v. 6060, n. 60600S, 2006. Citado na página 27.

MORRISON, A.; ROSS, G.; CHALMERS, M. Combining and comparing clustering and layout algorithms. *Department of Computing Science, University of Glasgow*, 2011. Citado na página 25.

MUHR, M.; SABOL, V.; GRANITZER, M. Scalable recursive top-down hierarchical clustering approach with implicit model selection for textual data sets. *21st International Workshop on Database and Expert Systems Applications, DEXA 2010*, IEEE, n. 5591979, p. 15–19, 2010. Citado 2 vezes nas páginas 51 e 52.

OESTERLING, P.; SCHEUERMANN, G.; TERESNIAK, S.; HEYER, G.; KOCH, S.; ERTL, T.; WEBER, G. H. Two-stage framework for a topology-based projection and visualization of classified document collections. *1st IEEE Conference on Visual Analytics Science and Technology, VAST 10*, IEEE, n. 5652940, p. 91–98, 2010. Citado na página 51.

PAULOVICH, F. V. *Mapeamento de dados multi-dimensionais - integrando mineração e visualização*. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, 2008. Citado 3 vezes nas páginas 28, 37 e 38.

PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, IEEE, v. 14, n. 4378370, p. 564–575, 2008. Citado 3 vezes nas páginas 35, 36 e 38.

PAULOVICH, F. V.; TOLEDO, F. M. B.; TELLES, G. P.; MINGHIM, R.; NONATO, L. G. Semantic wordification of document collections. *Computer Graphics Forum*, Blackwell Publishing Ltd, v. 31, p. 1145–1153, 2012. Citado na página 54.

ROMAN, F. S.; PINHO, R. D.; MINGHIM, R.; OLIVEIRA, M. C. F. de. A study on the role of similarity measures in visual text analytics. *International Conference on Computer Graphics Theory and Applications, GRAPP 2013*, p. 429–438, 2013. Citado 4 vezes nas páginas 38, 52, 54 e 55.

SAEED, N.; NAM, H.; HAQ, M.; BHATTI, D. M. A survey on multidimensional scaling. *ACM Computing Surveys*, Research gate, v. 51, n. 1, 2018. Citado 6 vezes nas páginas 28, 29, 30, 31, 32 e 33.

SEQUERA, J. L. C.; CASTILLO, J. R. F. D.; SOTOS, L. G. Cluster of reuters 21578 collections using genetic algorithms and nzipf method. *IADIS European Conference Data Mining 2009*, p. 174–176, 2009. Citado na página 20.

SHERKAT, E.; NOURASHRAFEDDIN, S.; MILIOS, E. E.; MINGHIM, R. Interactive document clustering revisited: A visual analytics approach. *23rd ACM International Conference on Intelligent User Interfaces, IUI 2018*, Association for Computing Machinery, p. 281–292, 2018. Citado 4 vezes nas páginas 51, 52, 53 e 54.

SORKINE, O.; COHEN-OR, D. Least-squares meshes. *Proceedings - Shape Modeling International SMI 2004*, p. 191–199, 2004. Citado na página 37.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to data mining*. 1. ed. [S.l.]: Addison Wesley, 2005. Citado 3 vezes nas páginas 22, 23 e 24.

TEJADA, E.; NONATO, L. G.; MINGHIM, R. On improved projection techniques to support visual exploration of multi-dimensional data sets. *Information Visualization*, v. 2, p. 218–231, 2003. Citado 2 vezes nas páginas 15 e 27.

THAI, V.; ROUILLE, P.-Y.; HANDSCHUH, S. Visual abstraction and ordering in faceted browsing of text collections. *ACM Transactions on Intelligent Systems and Technology*, Association for Computing Machinery, v. 3, n. 21, 2012. Citado na página 51.

TUESTA KARINA DELGADO, R. M. L. D. J. M.-C. E.; PéREZ-ALCáZAR, J. Analysis of an advisor-advisee relationship: an exploratory study of the area of exact and earth sciences in brazil. *PLos One*, 2015. Citado na página 56.

ZIPF, G. K. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. 1. ed. [S.l.]: Cambridge: Addison-Wesley, 1949. Citado na página 19.

ZYOUD, S. H.; FUCHS-HANUSCH, D. A bibliometric-based survey on ahp and topsis techniques. *Expert Systems with Applications*, Elsevier, v. 78, p. 158–181, 2017. Citado na página 14.

Apêndice A – Estudos primários da primeira questão de pesquisa

Na tabela a seguir é apresentada a lista dos estudos primários que irão apoiar a revisão de trabalhos correlatos.

Tabela 6 – Estudos primários da primeira questão de pesquisa

Título	Autores	Ano
D3 data-driven documents	Bostock M., Ogievetsky V., Heer J.	2011
Hierarchical attention networks for document classification	Yang Z., Yang D., Dyer C., He X., Smola A., Hovy E.	2016
The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis	Chen C., Ibekwe-SanJuan F., Hou J.	2010
A literature review on the state-of-the-art in patent analysis	Abbas A., Zhang L., Khan S.U.	2014
A bibliometric-based survey on AHP and TOPSIS techniques	Zyoud S.H., Fuchs-Hanusch D.	2017
Local Affine Multidimensional Projection	Joia P., Paulovich F.V., Coimbra D., Cuminato J.A., Nonato L.G.	2011
Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization	Dunne C., Shneiderman B., Gove R., Klavans J., Dorr B.	2012
Semantic wordification of document collections	Paulovich F.V., Toledo F.M.B., Telles G.P., Minghim R., Nonato L.G.	2012
An enhanced bag-of-visual word vector space model to represent visual content in athletics images	Kesorn K., Poslad S.	2012
Applications of topic models	Boyd-Graber J., Hu Y., Mimno D.	2017
Similarity preserving snippet-based visualization of web search results	Gomez-Nieto E., Roman F.S., Pagliosa P., Casaca W., Helou E.S., De Oliveira M.C.F., Nonato L.G.	2014

Continua na próxima página

Tabela 6 – Estudos primários

Título	Autores	Ano
Visualizing search results and document collections using topic maps	Newman D., Baldwin T., Cavedon L., Huang E., Karimi S., Martinez D., Scholer F., Zobel J.	2010
Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists	Brehmer M., Ingram S., Stray J., Munzner T.	2014
Vispubdata.org: A Metadata Collection about IEEE Visualization (VIS) Publications	Isenberg P., Heimerl F., Koch S., Isenberg T., Xu P., Stolper C.D., Sedlmair M., Chen J., Moller T., Stasko J.	2017
Morphable Word Clouds for Time-Varying Text Data Visualization	Chi M.-T., Lin S.-S., Chen S.-Y., Lin C.-H., Lee T.-Y.	2015
WordBridge: Using composite tag clouds in node-link diagrams for visualizing content and relations in text corpora	Kim K., Ko S., Elmqvist N., Ebert D.S.	2011
Dimensionality reduction for documents with nearest neighbor queries	Ingram S., Munzner T.	2015
Cite2vec: Citation-Driven Document Exploration via Word Embeddings	Berger M., McDonough K., Seversky L.M.	2017
Taking Word Clouds Apart: An Empirical Investigation of the Design Space for Keyword Summaries	Felix C., Franconeri S., Bertini E.	2018
Interactive visualization for opportunistic exploration of large document collections	Lehmann S., Schwanecke U., Dörner R.	2010
Trading consequences: A case study of combining text mining and visualization to facilitate document exploration	Hinrichs U., Alex B., Clifford J., Watson A., Quigley A., Klein E., Coates C.M.	2015
FacetScape: A visualization for exploring the search space	Seifert C., Jurgovsky J., Granitzer M.	2014

Continua na próxima página

Tabela 6 – Estudos primários

Título	Autores	Ano
Two-stage framework for a topology-based projection and visualization of classified document collections	Oesterling P., Scheuermann G., Teresniak S., Heyer G., Koch S., Ertl T., Weber G.H.	2010
An interactive visual testbed system for dimension reduction and clustering of large-scale high-dimensional data	Choo J., Lee H., Liu Z., Stasko J., Park H.	2013
Helping intelligence analysts make connections	Hossain M.S., Andrews C., Ramakrishnan N., North C.	2011
DocuCompass: Effective exploration of document landscapes	Heimerl F., John M., Han Q., Koch S., Ertl T.	2017
Time-aware visualization of document collections	Alencar A.B., Börner K., Paulovich F.V., De Oliveira M.C.F.	2012
Sparse machine learning methods for understanding large text corpora	El Ghaoui L., Li G.-C., Duong V.-A., Pham V., Srivastava A., Bhaduri K.	2011
Interactive document clustering revisited: A visual analytics approach	Sherkat E., Nourashrafeddin S., Milios E.E., Minghim R.	2018
Analysis of large digital collections with interactive visualization	Xu W., Esteva M., Jain S.D., Jain V.	2011
Topic-and Time-Oriented Visual Text Analysis	Dou W., Liu S.	2016
Eye-tracking investigation during visual analysis of projected multidimensional data with 2D scatterplots	Etemadpour R., Olk B., Linsen L.	2014
Iterative generation of insight from text collections through mutually reinforcing visualizations and fuzzy cognitive maps	Pillutla V.S., Giabbanelli P.J.	2019
Typograph: Multiscale spatial exploration of text documents	Endert A., Burtner R., Cramer N., Perko R., Hampton S., Cook K.	2013

Continua na próxima página

Tabela 6 – Estudos primários

Título	Autores	Ano
Understanding large text corpora via sparse machine learning	El Ghaoui L., Pham V., Li G.-C., Duong V.-A., Srivastava A., Bhaduri K.	2013
Evaluating exploratory visualization systems: A user study on how clustering-based visualization systems support information seeking from large document collections	Liu Y., Barlowe S., Feng Y., Yang J., Jiang M.	2013
Animated geooral clusters for exploratory search in event data document collections	Craig P., Seiler N.R., Cervantes A.D.O.	2014
Footprints: A visual search tool that supports discovery and coverage tracking	Isaacs E., Damico K., Ahern S., Bart E., Singhal M.	2014
Using otsu's threshold selection method for eliminating terms in vector space model computation	Eler D.M., Garcia R.E.	2013
Document summarization using semantic clouds	Rinaldi A.M.	2013
Exploring large digital library collections using a map-based visualisation	Hall M., Clough P.	2013
Visual abstraction and ordering in faceted browsing of text collections	Thai V., Rouille P.-Y., Handschuh S.	2012
Exploring Topic Models on Short Texts: A Case Study with Crisis Data	Manna S., Phongpanangam O.	2018
Big Text Visual Analytics in Sensemaking	Bradel L., Wycoff N., House L., North C.	2015
Exploratory visual analysis and interactive pattern extraction from semi-structured data	Soto A.J., Kiros R., Kešelj V., Milios E.	2015
Metadata enriched visualization of keywords in context	Fischl D., Scharl A.	2014

Continua na próxima página

Tabela 6 – Estudos primários

Título	Autores	Ano
INVISQUE: Technology and methodologies for interactive information visualization and analytics in large library collections	Wong B.L.W., Choudhury S., Rooney C., Chen R., Xu K.	2011
The Effect of Semantic Interaction on Foraging in Text Analysis	Wenskovitch J., Bradel L., Dowling M., House L., North C.	2018
Visual search analytics: Combining machine learning and interactive visualization to support human-centred search	Hoeber O.	2014
Designing map-based visualizations for collection understanding	Buchel O.	2011
Scalable recursive top-down hierarchical clustering approach with implicit model selection for textual data sets	Muhr M., Sabol V., Granitzer M.	2010
Visual topic models for healthcare data clustering	Rajendra Prasad K., Mohammed M., Noorullah R.M.	2019
LabelTransfer-Integrating Static and Dynamic Label Representation for Focus+Context Text Exploration	Han Q., John M., Koch S., Assenov I., Ertl T.	2018
Patterning of writing style evolution by means of dynamic similarity	Amelin K., Granichin O., Kizhaeva N., Volkovich Z.	2018
A visual approach for interactive keyterm-based clustering	Nourashrafeddin S., Sherkat E., Minghim R., Milios E.E.	2018
Multi-level mining and visualization of scientific text collections	Accuosto P., Ronzano F., Ferrés D., Saggion H.	2017
An online inference algorithm for Labeled Latent Dirichlet allocation	Zhou Q., Huang H., Mao X.-L.	2015

Continua na próxima página

Tabela 6 – Estudos primários

Título	Autores	Ano
Analysis of text cluster visualization in emergent self organizing maps using unigrams and its variations after introducing bigrams	Singh P.K., MacHavolu M., Bharti K., Suda R.	2012
TexTonic: Interactive visualization for exploration and discovery of very large text collections	Paul C.L., Chang J., Endert A., Cramer N., Gillen D., Hampton S., Burtner R., Perko R., Cook K.A.	2019
Contravis: Contrastive and visual topic modeling for comparing document collections	Le T.V.M., Akoglu L.	2019
Web summarization and browsing through semantic tag clouds	Rinaldi A.M.	2019
Reading through graphics: Interactive landscapes to explore dynamic topic spaces	Ulbrich E., Veas E., Singh S., Sabol V.	2015
Comparative exploration of document collections: A visual analytics approach	Oelke D., Strobelt H., Rohrdantz C., Gurevych I., Deussen O.	2014
Content Visualization of Scientific Corpora Using an Extensible Relational Database Implementation	Giannakopoulos T., Stamatogiannakis E., Foufoulas I., Dimitropoulos H., Manola N., Ioannidis Y.	2014
A focus + context technique for visualizing a document collection	Dunsmuir D., Lee E., Shaw C.D., Stone M., Woodbury R., Dill J.	2012
The hot research topics and the research fronts in the field of Web Data Mining(WDM) based on web of science	Chen L., Wei L.	2010
Science Mapping of Tunnel Fires: A Scientometric Analysis-Based Study	Li J., Liu J.	2020
Hybrid visualization approach to show documents similarity and content in a single view	Andreotti A.L.D., Silva L.F., Eler D.M.	2018
Visual analysis and exploration of entity relations in document collections	John M., Heimerl F., Vu B.-A., Ertl T.	2018

Continua na próxima página

Tabela 6 – Estudos primários

Título	Autores	Ano
Distribution features and intellectual structures of digital humanities: A bibliometric analysis	Wang Q.	2018
Construction inverted index for dynamic collections visualization in thematic virtual museums system	Anggai S., Blekanov I.S., Sergeev S.L.	2017
Exploring & summarizing document collections with multiple coordinated views	Di Sciascio C., Mayr L., Veas E.	2017
SEPIR: A semantic and personalised information retrieval tool for the public administration based on distributional semantics	Basile P., Caputo A., Di Ciano M., Grasso G., Rossiello G., Semeraro G.	2017
Visualizing document image collections using image-based word clouds	Wilkinson T., Brun A.	2015
Multi-focus cluster labeling	Eikvil L., Jenssen T.-K., Holden M.	2015
Exploring document collections with topic frames	Hinneburg A., Rosner F., Pessler S., Oberländer C.	2014
Hybrid approach for visualization of documents clusters using GHSOM and sammon projection	Butka P., Pocsova J.	2013
A study on the role of similarity measures in visual text analytics	San Roman F.S., De Pinho R.D., Minghim R., De Oliveira M.C.F.	2013
Search and graphical visualization of concepts in document collections using taxonomies	Schmidt A., Kimmig D., Dickerhof M.	2013
Visually summarizing semantic evolution in document streams with topic table	Gohr A., Spiliopoulou M., Hinneburg A.	2013
Visualizations for the spyglass ontology-based information analysis and retrieval system	Lin H., Rushing J., Berendes T., Stein C., Graves S.	2010

Continua na próxima página

Tabela 6 – Estudos primários

Título	Autores	Ano
StanceVis Prime: visual analysis of sentiment and stance in social media texts	Kucher K., Martins R.M., Paradis C., Kerren A.	2020
Exploring Collections of research publications with Human Steerable AI	González Martínez A., Wooton B.T., Kirshenbaum N., Kobayashi D., Leigh J.	2020
Semantic concept spaces: Guided topic model refinement using word-embedding projections	El-Assady M., Kehlbeck R., Collins C., Keim D., Deussen O.	2020
Research and application of space-time behavior maps: a review	Zhang X., Cheng Z., Tang L., Xi J.	2020
TopicSifter: Interactive Search Space Reduction through Targeted Topic Modeling	Kim H., Choi D., Drake B., Endert A., Park H.	2019
Trivir: A visualization system to support document retrieval with high recall	Dias A.G., Milios E.E., Ferreira de Oliveira M.C.	2019
Overview of trends in global epigenetic research (2009–2017)	Olmeda-Gómez C., Romá-Mateo C., Ovalle-Perandones M.-A.	2019
Designing effective knowledge presentation techniques for large digital collections	Wu Y., Yang S.	2019
Topic tomographies (Toptom): A visual approach to distill information from media streams	Gobbo B., Balsamo D., Mauri M., Bajardi P., Panisson A., Ciuccarelli P.	2019
Ontology coverage tool and document browser for learning material exploration	Grevisse C., Meder J., Botev J., Rothkugel S.	2018
Visualization of subtopics of the thematic document collection using the context-semantic graph	Sboev A., Moloshnikov I., Gudovskikh D., Rybka R.	2016
Exploratory analysis of text collections through visualization and hybrid biclustering	Médoc N., Ghoniem M., Nadif M.	2016

Continua na próxima página

Tabela 6 – Estudos primários

Título	Autores	Ano
AnnotatedTimeTree: Visualization and annotation of news text and other heterogeneous document collections	Xia J., Zhao J., Sheeley I., Christopher J., Wang Q., Guo C., Zhang J., Ebert D.S., Chen Y.V., Qian Z.C.	2015
User-centered text mining (invited tutorial)	Soto A.J., Milios E.E.	2015
LocLinkVis: A geographic information retrieval-based system for large-scale exploratory search	Olieman A., Kamps J., Claros R.M.	2015
Linked visual analysis of structured datasets and document collections	Kolman S., Galkina E., Dufilie A.S., Luo Y.F., Gupta V., Grinstein G.	2014
Aspect grid: A visualization for iteratively refining aspect-based queries on document collections	Haag F., Han Q., John M., Ertl T.	2014
3D gesture-based exploration and search in document collections	De Antonio A., Moral C., Klepel D., Abente M.J.	2013
Supervised content visualization of scientific publications: A case study on the ArXiv dataset	Giannakopoulos T., Dimitropoulos H., Metaxas O., Manola N., Ioannidis Y.	2013
Gesture-based control of the 3D visual representation of document collections for exploration and search	De Antonio A., Moral C., Klepel D., Abente M.J.	2013
Understanding collections and their implicit structures through information visualization	Alfredo Sánchez J.	2013
SENSE: Intelligent storage and exploration of large document sets	Wehner P.	2013
Visualization of records classified with the 1998 ACM CCS	Medina M.A., Sánchez J.A., De La Mora J.C., Benítez Ruiz A.	2012
Gesture-based interaction with 3D visualizations of document collections for exploration and search	De Antonio A., Moral C., Klepel D., Abente M.J.	2012

Continua na próxima página

Tabela 6 – Estudos primários

Título	Autores	Ano
Automatic indexing and information visualization: A study based on paraconsistent logic	Corrêa C.A., Kobashi N.Y.	2012

Fonte: Autora do trabalho, 2020

Apêndice B – Estudos primários da segunda questão de pesquisa

Na tabela a seguir é apresentada a lista dos estudos primários que irão apoiar a revisão de trabalhos correlatos.

Tabela 7 – Estudos primários da segunda questão de pesquisa

Título	Autores	Ano
Scientific production of women in Brazil	de Oliveira Santiago M., Affonso F., Dias T.M.R.	2020
Profile of women's guidelines and productions based on data from the lattes platform	Santiago M.O., Affonso F., Dias T.M.R.	2020
Analysis of intellectual production in Information Science in Postgraduate studies: A bibliometric study based in data from the Lattes Platfrom	Nascimento M.R., Pinto A.L., Dias T.M.R.	2020
Analysis of the technical-scientific production of the National Council for Scientific and Technological Development (CNPq) productivity fellows in Pediatrics	Klepa T.C., Pedroso B.	2020
Profile of neurophysiology research groups of Brazil	Vieira A.S., Welter M.R.T., Mello-Carpes P.B.	2014
Analysis of an advisor-advisee relationship: An exploratory study of the area of Exact and Earth Sciences in Brazil	Tuesta E.F., Delgado K.V., Mugnaini R., Digiampietri L.A., Mena-Chalco J.P., Pérez-Alcázar J.J.	2015
Open access data for understanding women's scientific production in Brazil	De Oliveira Santiago M., Dias T.M.R.	2019
Competitive intelligence in panorama of Brazil: The researchers and scientific production on lattes platform	Amaral R.M., Brito A.G.C., Rocha K.G.S., Quoniam L.M., de Faria L.I.L.	2016

Continua na próxima página

Tabela 7 – Estudos primários

Título	Autores	Ano
Profile and scientific output of researchers recipients of CNPq productivity grant in the field of medicine	Martelli D.R., Oliveira M.C.L., Pinheiro S.V., Santos M.L., Dias V., Silva A.C.S., Martelli-Júnior H., Oliveira E.A.	2019
Multi and interdisciplinarity in the brazilian postgraduate programs in information science	Lança T.A., Amaral R.M., Gracioso L.S.	2018
Analysis of the communities of Brazilian researchers in the area of philosophy: A study based on the juxtaposition between the data of the Lattes Platform and Web of Science (2007-2016)	Silva F.M., Sánchez M.L.L., López A.E.S., Casado E.S.	2018
Analysis of technological production in biotechnology in northeast Brazil	Gomes Costa B.M., Nannini da Silva Florencio M., Oliveira Junior A.M.D.	2018
Scientific production of researchers in the Nutrition field with productivity fellowships from the National Council for Scientific and Technological Development	de Pinho L., Martelli-Júnior H., Oliveira E.A., Martelli D.R.B.	2017
Analytical visualization of the keywords in scientific meeting: Proposed from the Lattes platform	Gomes J.O., Dias T.M.R., Pinto A.L., Moita G.F.	2016
A bibliometric analysis of the scientific production and collaboration between graduate programs in manufacturing engineering in Brazil	Dutra S.T., Lezana Á.G.R., Dutra M.L., Pinto A.L.	2019
Geosciences of CNPq from research productivity fellows	Cândido L.F.O., Santos N.C.F., da Rocha J.B.T.	2016

Continua na próxima página

Tabela 7 – Estudos primários

Título	Autores	Ano
Science in Brazilian regions: Development of scholarly production and research collaboration networks	Sidone O.J.G., Haddad E.A., Mena-Chalco J.P.	2016
The oswaldo cruz foundation and science on women: Women's participation in practice and research management in an educational and research institution	Rodrigues J.G., Guimarães M.C.S.	2016
A System for Discovery of Knowledge in Data Repository Education	De Sousa Costa E., Rodrigues Dias T.M., Dias P.M.	2019
Customer relationship management (CRM): State of the art, bibliometric review of high-quality Brazilian production, institutionalization of research in Brazil and research agenda	Demo G., Fogaça N., Ponte V., Fernandes T., Cardoso H.	2015
Scientific collaboration in biotechnology: The case of the northeast region in Brazil	Costa B.M.G., da Silva Pedro E., de Macedo G.R.	2013
The Brazilian academic genealogy: evidence of advisor–advisee relationships through quantitative analysis	Damaceno R.J.P., Rossi L., Mugnaini R., Mena-Chalco J.P.	2019
Brazilian bibliometric coauthorship networks	Mena-Chalco J.P., Digiampietri L.A., Lopes F.M., Cesar Jr. R.M.	2014
Profile and scientific output analysis of physical therapy researchers with research productivity fellowship from the Brazilian national council for scientific and technological development	Sturmer G., Viero C.C.M., Silveira M.N., Lukrafka J.L., Plentz R.D.M.	2013
Scientific research output evaluation of professors of Sao Paulo State University, Marília/SP	Herculano R.D., Norberto A.M.Q.	2012

Continua na próxima página

Tabela 7 – Estudos primários

Título	Autores	Ano
Profile and Scientific Production of CNPq Researchers in Cardiology	de Oliveira E.A., Ribeiro A.L.P., Quirino I.G., Oliveira M.C.L., Martelli D.R., Lima L.S., Colosimo E.A., Lopes T.J., e Silva A.C.S., Martelli-Junior H.	2011
Scientific research output of professors of Sao Paulo State Universty, Assis/SP	Herculano R.D., Norberto A.M.Q.	2011
Scientific research in nursing education: Rio de Janeiro and Minas Gerais research groups	Gomes D.C., Backes V.M., Lino M.M., Canever B.P., Ferraz F., Schveitzer M.C.	2011
CNPq researchers in medicine: A comparative study of research areas	Martelli-Junior H., Martelli D.R.B., Quirino I.G., Oliveira M.C.L.A., Lima L.S., de Oliveira E.A.	2010
Profile of scientific and technological production in nursing education research groups in the South of Brazil	Lino M.M., Backes V.M.S., Canever B.P., Ferraz F., Prado M.L.	2010
Research productivity of CNPq: Analysis of the chemistry researchers' profile	Santos N.C.F., De Cândido L.F.O., Kuppens C.L.	2010

Fonte: Autora do trabalho, 2022