UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Seleção de bibliografia científica baseada em relevância e similaridade utilizando redes complexas

Débora Fritscher Pires

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Débora Fritscher Pires

Seleção de bibliografia científica baseada em relevância e similaridade utilizando redes complexas

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Alneu de Andrade Lopes

Versão original

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Pires, Débora Fritscher
S856m Seleção de bibliografia científica baseada em relevância e similaridade utilizando redes complexas / Débora Fritscher Pires ; orientador Alneu de Andrade Lopes. – São Carlos, 2023.
40 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Lopes, Alneu de Andrade, orient. II. Título.

AGRADECIMENTOS

Ao meu marido José Paulo pelo companheirismo, compreensão, por compartilhar a mesma paixão pela computação e pela troca de idéias que muito sedimentaram meu conhecimento.

À minha filha Paula pela compreensão, paciência e interesse pelo assunto.

Ao Prof. Alneu de Andrade Lopes pelo incentivo, orientação e apoio ao longo do desenvolvimento deste trabalho.

Aos professores e tutores do MBA em Inteligência Artificial e Big Data por tantos conhecimentos transmitidos e tantas dúvidas esclarecidas.

À coordenação do curso por fazer tudo isso possível.

RESUMO

Pires, D. F. **Seleção de bibliografia científica baseada em relevância e similaridade utilizando redes complexas**. 2023. 40p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Uma grande quantidade de documentos científicos é produzida e pode ser disponibilizada atualmente com muita facilidade. Essa abundância, porém, gera uma sobrecarga de informação e torna difícil a análise de todo conteúdo. Consultas que usam a busca por palavras-chaves, como o Google, retornam muitas respostas, mas a maioria sem relação com o conteúdo desejado. É necessária uma seleção por um critério mais efetivo que retorne documentos significativos para o trabalho que se está fazendo. Quando artigos científicos são selecionados, se deseja ter como retorno documentos de qualidade e relacionados com a pesquisa em andamento. Um trabalho citado por muitos indica que os autores destes atestam a qualidade do primeiro. Quando os trabalhos que referenciam outros são relevantes em suas áreas, a qualidade dos referenciados é validada e reforçada. Já a similaridade entre textos de documentos mostra o grau de relação entre eles. Trabalhos que citam um mesmo conjunto de outros documentos também estão relacionados. Muitos algoritmos tradicionais de classificação que utilizam na sua avaliação a similaridade de textos treinam um modelo a partir de dados rotulados. A obtenção destes dados, porém, é difícil e tem um custo bastante elevado. Opções são algoritmos que utilizam abordagens baseadas em grafos com uma quantidade muito pequena de artigos de interesse na entrada. Constroem um grafo, identificam automaticamente documentos negativos e utilizam algum algoritmo de propagação para espalhar os rótulos positivos e negativos pelos nodos do grafo. Estes algoritmos avaliam apenas o grau de relação com a pesquisa sendo realizada, não contribuindo para uma seleção de qualidade. Este trabalho acrescenta, neste cenário, o critério de qualidade na seleção de artigos, adicionando a análise da rede de citações a esta técnica e utilizando PageRank na avaliação. A seleção por similaridade também é aprimorada com a utilização do número de artigos que cada documento tem em comum com os artigos de interesse como critério.

Palavras-chave: Rede de citações. Similaridade de texto. PageRank. Aprendizado transdutivo.

ABSTRACT

Pires, D. F. **Scientific paper selection based on relevance and similarity using complex networks**. 2023. 40p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

A vast amount of scientific documents is currently being produced and can be easily accessed. However, this abundance creates information overload, making it challenging to analyze the entire content. Queries using keyword searches, such as Google's, yield many responses, but most are unrelated to the desired content. There is a need for a more effective criterion that retrieves meaningful documents for ongoing research. When selecting scientific articles, the goal is to obtain high-quality documents related to the ongoing research. A work cited by many indicates that its authors attest to the quality of the original. When works referencing others are relevant in their areas, the quality of the referenced works is validated and reinforced. The similarity between texts in documents reveals the degree of relation between them. Works citing the same set of other documents are also related. Many classification algorithms used in text similarity assessment train models from labeled data, but obtaining such data is difficult and costly. Alternatives include algorithms using graph-based approaches with a very small set of input articles. They construct a graph, automatically identify negative documents, and use some label propagation algorithm to spread positive and negative labels across the graph nodes. However, these algorithms assess only the degree of relation to the ongoing research, not contributing to quality selection. This work introduces a quality criterion in article selection by adding citation network analysis to this technique, utilizing PageRank in the evaluation. Similarity selection is also enhanced by considering the number of articles each document shares with the articles of interest as a criterion.

Keywords: Network citation. Text similarity. PageRank. Transdutive learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização e Motivação

Atualmente o conhecimento é amplamente difundido no formato digital. Documentos são produzidos e disponibilizados nas mais variadas plataformas. No meio acadêmico isso não é diferente. Em nenhuma outra época foi possível o acesso a tantos documentos científicos. Essa abundância, no entanto, cria uma sobrecarga de informação e pesquisadores tem dificuldade em analisar e julgar todo esse conteúdo. Consultas feitas utilizando as tradicionais técnicas de busca baseadas em palavras-chaves, como *Google* e *Bing*, retornam um grande número de respostas, mas a maioria é irrelevante e pouco acrescenta à pesquisa. São necessárias técnicas mais efetivas para uma massiva quantidade de dados, que ajudem a localizar mais rapidamente artigos importantes e relevantes em um determinado campo de estudo (BAI *et al.*, 2019). Para isso, trabalhos relacionados precisam ser filtrados usando algum critério de limitação. Uma boa seleção precisa levar em conta a qualidade do trabalho e o grau de relação com a pesquisa sendo realizada (AMANCIO *et al.*, 2012). Um trabalho citado por muitos indica que os autores destes atestam a qualidade do primeiro. Quando os trabalhos que referenciam são relevantes em suas áreas, a qualidade dos referenciados é validada e reforçada. Já a similaridade entre textos de documentos mostra o grau de relação entre eles. Trabalhos que citam um mesmo conjunto de outros documentos também estão relacionados.

Muitos algoritmos tradicionais de classificação que utilizam na sua avaliação a similaridade de textos treinam um modelo a partir de dados rotulados. Porém, um conjunto consistente de documentos rotulados para induzir um classificador não está disponível na maioria das aplicações reais. Além disso, produzir documentos rotulados é uma tarefa que consome bastante tempo e esforço de especialistas (FALEIROS, 2016). Como opção, existem algoritmos que utilizam apenas uma quantidade muito pequena de artigos de interesse na entrada, identificados como da classe positiva. Neste cenário, para a classificação de textos, abordagens baseadas em grafos tem obtido desempenho superior aos modelos de espaço vetorial. Além disso, oferecem várias opções para representar uma coleção de documentos (CARNEVALI *et al.*, 2021). Alternativas, então, passam a ser algoritmos baseados em grafos que utilizam uma quantidade muito pequena de artigos de interesse em um corpus, constroem um grafo, identificam automaticamente documentos negativos e utilizam algum algoritmo de propagação para espalhar os rótulos positivos e negativos pelos nodos do grafo. Na propagação, atribuem pesos aos documentos tanto para a classe positiva quanto para a negativa e os documentos são classificados de acordo com a classe de maior peso. Estes algoritmos, porém, avaliam apenas o grau de relação com a pesquisa sendo realizada, não contribuindo para uma seleção de qualidade.

Este trabalho se propõe a acrescentar o critério de qualidade na seleção de artigos que serão indicados ao usuário, adicionando a análise da rede de citações a este método e utilizando para isso o indicador Pagerank que avalia a relevância dos textos. A rede de citações também aprimora a seleção por similaridade, utilizando como critério o número de artigos que cada documento tem em comum com os artigos de interesse.

1.2 Objetivos

O objetivo deste trabalho é, a partir de um pequeno conjunto de artigos científicos de interesse, retornar o conjunto de trabalhos mais relevantes relacionados a eles em um repositório, auxiliando no levantamento de referências bibliográficas. São utilizadas na investigação tanto a rede de citações quanto a similaridade entre os documentos, modeladas em redes complexas, com uma visão para cada um dos critérios.

1.3 Organização do Texto

O restante de deste trabalho está organizado da seguinte forma. O Capítulo 2 apresenta a fundamentação teórica e os trabalhos relacionados, abordando redes complexas, rede de citações, representação e similaridade de dados textuais, incluindo modelos de espaçõ vetorial, medidas de distância e grafos, e aprendizado transdutivo para classificação de textos. O Capítulo 3 apresenta a proposta para a seleção de bibliografia científica baseada em relevância e similaridade. Sua avaliação experimental é relatada no Capítulo 4, mostrando os dados, a configuração experimental, os resultados e discussões. Por fim, no Capítulo 5, são apresentadas as conclusões.

2 FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

2.1 Redes Complexas

Uma rede, também chamada de grafo na literatura matemática, é um conjunto de itens que recebem o nome de vértices ou nodos com conexões entre eles chamadas arestas ou ligações (NEWMAN, 2003). Sistemas com o formato de redes são bem conhecidos. Exemplos são redes sociais de amizade ou de trabalho, de energia elétrica, telefônica, neurais e a Internet. Uma rede simples pode ser vista na Figura 1.

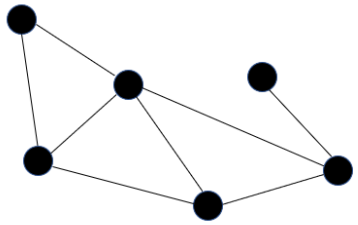

Figura 1 – Rede com seis vértices e oito arestas

Historicamente, as redes foram estudadas em uma área da matemática discreta chamada teoria dos grafos. Nos últimos anos, houve um movimento de interesse no estudo de redes com uma estrutura mais complexa e o foco principal se moveu da análise de pequenos grafos e das propriedades de seus vértices e arestas para análises estatísticas de redes com milhares ou milhões de nodos (BOCCALETTI *et al.*, 2006).

Em um grafo, vértices e arestas podem ter propriedades associadas a eles, numéricas ou não. Por exemplo, arestas podem ter pesos representando o grau de relação entre os nodos associados. Arestas também podem ser direcionadas ou não. Alguns sistemas são formados por arestas não direcionadas como uma rede de documentos, onde o peso das ligações indica a similaridade entre eles. Outros possuem arestas direcionadas. Um exemplo é uma rede onde artigos citam outros artigos, com o fluxo indo na direção de quem cita para quem é citado. Grafos direcionados podem ser cíclicos, quando contém laços fechados de arestas, ou acíclicos, no caso contrário. Um grafo é direcionado quando todas as suas ligações também o são e não direcionado quando todas arestas não o são. Alguns grafos possuem, simultaneamente, arestas dos dois tipos (BARABÁSI, 2013)

Grafos também podem conter mais de um tipo de vértice, aresta ou ambos. Grafos com mais de um tipo de vértice são chamados de heterogêneos. Existem redes heterogêneas com características específicas (ROSSI, 2016). Um exemplo é a rede heterogênea bipartida, composta por nodos de dois tipos diferentes, distribuídos em dois conjuntos disjuntos tal que cada aresta une um nodo de um conjunto a um nodo de outro. Tomando como

exemplo um conjunto de artigos científicos, os vértices podem representar os documentos e as palavras neles contidas e as arestas a ocorrência destas últimas nos artigos. O peso da aresta pode ser a frequência dos termos no respectivo artigo. Como mostra a Figura 2, podem ser geradas duas projeções para cada rede bipartida, onde cada projeção conecta os nodos de um conjunto por uma aresta se estiverem unidos ao mesmo nodo do outro conjunto na representação bipartida (BARABÁSI, 2013).

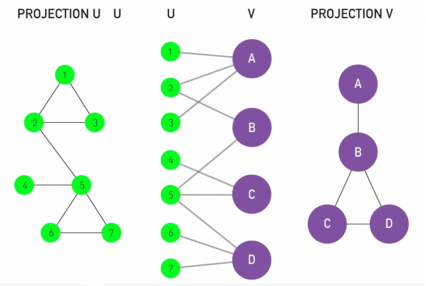

Figura 2 – Rede bipartida e suas projeções, extraído de (BARABÁSI, 2013)

O grau de um vértice é o número de arestas conectadas a ele. Um caminho é uma rota que corre através das arestas da rede. O comprimento do caminho representa o número de arestas contidas no caminho. O menor caminho entre os nodos x e y, ou a distância d_{xy} corresponde ao caminho com o menor número de arestas que conecta os nodos x e y. O diâmetro da rede é a maior distância na rede (BARABÁSI, 2013).

No mundo real são encontradas redes de diferentes tipos. Entre elas, pode-se destacar quatro categorias (NEWMAN, 2003):

Redes tecnológicas: Feitas pelo homem tipicamente para distribuição de alguma mercadoria ou recurso. Exemplos são redes de distribuição de energia, de telefonia e de computadores.

Redes biológicas: Conjunto de elementos biológicos e suas relações. Exemplos são redes neurais, de proteínas e de cadeia alimentar.

Redes sociais: Conjunto ou grupo de pessoas com algum padrão de contato ou conexão entre si, como relações de amizade ou de trabalho. Exemplos são *Instagram*, *Linkedin* e redes de coautoria entre pesquisadores, onde as relações representam trabalhos publicados em conjunto.

Redes de informação: Conjunto de dados com alguma forma de relacionamento ou troca de informação entre si (ROSSI, 2016). Também chamadas redes de conhecimento, tem como exemplos redes de páginas web, de similaridade e de citações entre artigos acadêmicos. Esta última, por serem de interesse para este trabalho, será detalhada a seguir.

2.2 Rede de Citações

São grafos onde os artigos citam trabalhos anteriores e essas citações formam uma rede onde os vértices são artigos e uma aresta direcionada do artigo A para o B indica que A cita B. Redes de citações são acíclicas, porque artigos só podem citar trabalhos já escritos, não os que ainda estão para ser elaborados. Com isso, todas as arestas da rede apontam para o passado, não permitindo, em princípio, a existência de laços fechados (NEWMAN, 2003).

A distribuição dos graus dos nodos em redes de citações é heterogênea, onde muitos nodos apresentam um grau baixo e muito poucos possuem um grau elevado. Isso significa que alguns poucos artigos, de destaque, são muito citados e a maioria é pouco referenciada (BARABÁSI; ALBERT, 1999). Artigos muito citados provavelmente são trabalhos de qualidade e tendem a se tornar uma boa referência para elaboração de outros documentos.

Apesar do número de citações ter relação com a importância de um artigo, ele pode não refletir o prestígio de um documento, uma vez que ser mencionado por um trabalho significativo tem mais valor do que por um com pouco interesse. O algoritmo PageRank (BRIN; PAGE, 1998) leva em consideração esses dois fatores. Proposto inicialmente como um mecanismo de busca que fazia uso da estrutura de hipertextos, passou a ser aplicado também para redes de citações (MA; GUAN; ZHAO, 2008). Ele não considera igualmente todas citações, mas dá um peso a cada uma relacionado com a importância e o número de referências dos artigos que o citam, ou seja, o prestígio de um trabalho é transferido a suas referências. O PageRank de uma página A, chamado de *PR(A)* é dado por:

$$PR(A) = (1 - d) + d \cdot \sum_{i} \frac{PR(T_i)}{C(T_i)} \tag{2.1}$$

onde d é um fator de amortecimento entre 0 e 1, normalmente definido como 0,85. PageRank forma uma distribuição de probabilidade sobre a rede de citações, de forma que a soma de seus valores para todos os artigos é 1. A Equação 2.1 mostra que um artigo pode ter um PageRank alto se muitos trabalhos o referenciam ou se alguns artigos que o citam tem um valor alto para este indicador.

Redes de citações também são úteis para aumentar o desempenho em tarefas de classificação quando documentos compartilham algumas referências bibliográficas (LAGUNA; LOPES, 2009). Particularmente para artigos científicos, quando eles tem um conjunto de citações em comum, a probabilidade de serem semelhantes é muito grande.

Apesar de citações representarem um bom indicador de qualidade e semelhança entre trabalhos, o fato de a maioria dos artigos não ser citada por outros faz com que as ligações em uma rede de citações sejam muito esparsas e existam poucas relações entre os documentos (LIU *et al.*, 2019). Com isso, o retorno de um classificador apenas por este

critério muitas vezes é pequeno ou até nulo, sugerindo sua utilização em conjunto com outro mecanismo de classificação.

2.3 Representação e Similaridade de Dados Textuais

O conceito de similar difere do de sinônimo, este último definido como um vocábulo que pode ser usado no lugar de outro sem alterar o significado da sentença. Similaridade é um conceito mais amplo. Está relacionado com as palavras em si e não apenas com seu significado. Mostra que uma palavra pode ocorrer no mesmo contexto que outra. Cachorro e gato são similares, mas não são sinônimos. Por outro lado, sinônimos são muito similares. Similaridade é um conceito que se estende também a textos. Dois documentos similares tendem a ter palavras similares. A representação computacional textos está bastante relacionada ao conceito de similaridade entre eles (JURAFSKY; MARTIN, 2023).

Modelos de espaço vetorial e grafos são duas formas de representar documentos computacionalmente.

2.3.1 Modelos de espaço vetorial

Em uma representação no modelo de espaço vetorial, documentos são representados por vetores cuja dimensão é o número de seus atributos. Estes vetores guardam em si o contexto onde o respectivo texto se encontra. Vetores que representam conteúdos com significados semelhantes se encontram próximos no espaço vetorial. Medidas de distância entre vetores conseguem então avaliar a maior ou menor similaridade entre os textos analisados, associando uma maior proximidade com uma maior semelhança.

Podem ser divididos em dois grupos (JURAFSKY; MARTIN, 2023). O primeiro é formado por modelos baseados em uma matriz de co-ocorrência, uma forma de representar o quanto frequentemente as palavras ocorrem juntas no contexto desejado. Cada palavra é representada por um vetor longo e esparso com dimensões correspondendo ao número de palavras do vocabulário. A matriz de co-ocorrência de um texto com vocabulário de n palavras será de ordem n x n, com cada linha e cada coluna correspondem a uma destas palavras. Estes vetores são esparsos pois a maioria das palavras não ocorre no contexto das outras. Neste grupo estão incluídos os modelos bag-of-words e TF-IDF. No primeiro, cada célula da matriz corresponde ao número de vezes que as palavras de sua linha e coluna aparecem juntas. Já o segundo, considera que palavras do vocabulário que aparecem muito em todos os textos, como artigos e preposições, não são significativas para comparar contextos, pois não dizem muito sobre eles. É formado por dois termos: TF que diz quantas vezes o termo aparece em um documento e IDF que é a razão entre o número total de documentos do corpus e o número de documentos onde um termo aparece, tendo seu menor valor quando está presente em todos os documentos. O peso TF-IDF para uma palavra em um documento é o produto destes dois termos. Este peso é aplicado à matriz

de co-ocorrência e o valor de cada dimensão passa a ser o peso TF-IDF da palavra aplicado ao número de vezes que ela ocorre simultaneamente com cada vizinho.

O segundo grupo é formado por embeddings, representações vetoriais formadas a partir de redes neurais, que recebem um conjunto muito grande de documentos chamado de corpus na entrada e transformam cada elemento em um vetor de menores dimensões, denso, onde a maioria dos valores é diferente de zero. Os atributos são aprendidos pelo modelo de forma a melhor representar o problema, mas o significado de cada um não é interpretável (JURAFSKY; MARTIN, 2023). Existem dois tipos de embeddings: não contextuais e contextuais (QIU *et al.*, 2020). Exemplos são, respectivamente, o Word2Vec (MIKOLOV *et al.*, 2013) e o Bert (DEVLIN *et al.*, 2019). A principal diferença entre eles é que no segundo caso, ao contrário do primeiro, a representação das palavras é alterada dinamicamente de acordo com o contexto onde aparece. Podem ser treinados pelo usuário ou podem ser utilizados modelos já pré-treinados com corpus muito grande. Neste último caso estão incluídos embeddings do Word2Vec treinados com textos do *Google News* e o SciBERT, treinado com artigos do corpus do *Semantic Scholar*.

2.3.2 Medidas de distância

Medidas de distância entre vetores são necessárias para avaliar a maior ou menor similaridade entre os textos analisados, associando uma maior proximidade com uma maior semelhança. Neste trabalho será usada a similaridade de cosseno (NEWMAN, 2003). Para os vetores u e w representando dois documentos, ela é obtida por:

$$\cos(u, w) = \frac{u \cdot w}{|u||w|} = \frac{\sum_{i=1}^{N} u_i w_i}{\sqrt{\sum_{i=1}^{N} u_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}} \tag{2.2}$$

2.3.3 Grafos

Coleções de textos também podem ser representadas por grafos. A representação pode ser um grafo de documentos, onde estes elementos são os nodos e a relação entre eles, por exemplo, de similaridade, citações ou coautoria, são as arestas. De maneira semelhante pode ser de palavras ou ainda de ambos. Neste último caso, a representação é um grafo bipartido, onde documentos e palavras são vértices de dois conjuntos disjuntos e as ocorrências de palavras nos documentos são as arestas (ANGELOVA; WEIKUM, 2006).

O grafo de documentos é a forma mais tradicional e a relação de similaridade entre textos é a que fornece melhores resultados (ANGELOVA; WEIKUM, 2006). Neste caso, a similaridade pode ser obtida a partir de uma representação em modelo de espaço vetorial,

utilizando a similaridade de cosseno definida na Equação 2.2 entre pares de vetores de artigos e, a partir dai, construir o grafo.

Em aprendizado semi-supervisionado para classificação de texto, abordagens baseadas em grafos tem obtido desempenho superior aos modelos de espaço vetorial. Além disso, oferecem várias alternativas para representar uma coleção de documentos (CARNEVALI *et al.*, 2021).

2.4 Aprendizado Transdutivo para Classificação de Textos

Em aprendizado de máquina, as tarefas supervisionadas constroem um classificador ou regressor a partir de um conjunto de pares de entrada e sua respectiva saída, o qual estima as saídas para entradas ainda não vistas. Nas tarefas não supervisionadas, nenhuma saída é fornecida e se procura inferir alguma estrutura a partir das entradas. Já as tarefas semi-supervisionadas procuram combinar estas duas tarefas (ENGELEN; HOOS, 2020). Em geral, elas procuram aumentar o desempenho em uma destas tarefas utilizando informação associada com a outra. Entre as tarefas semi-supervisionadas, uma das que se destaca é a de classificação, especialmente em cenários onde é difícil conseguir dados rotulados, como é o caso de textos em corpus com conteúdo muito técnico, diverso e extenso. Um número considerável de pares de entrada e saída é necessário para criar um modelo de classificação acurado. Entretanto, um conjunto consistente de documentos rotulados para induzir um classificador não está disponível na maioria das aplicações reais. Além disso, produzir documentos rotulados é uma tarefa que consome bastante tempo e esforço de especialistas (FALEIROS, 2016).

O aprendizado semi-supervisionado pode ser indutivo ou transdutivo (ENGELEN; HOOS, 2020). O primeiro utiliza dados rotulados e não rotulados para construir um modelo de classificação que pode fornecer predições de itens ainda não vistos. Já o segundo, a partir de um pequeno conjunto de valores rotulados, fornece predições para dados sem identificação fornecidos ao algoritmo. Como métodos transdutivos não geram um modelo, a informação precisa ser propagada através da conexão entre os dados, o que conduz naturalmente a uma abordagem baseada em grafo. Esta abordagem se identifica com o objetivo desta pesquisa, que é, a partir de um pequeno conjunto de documentos de interesse, encontrar referências bibliográficas em uma base de dados e, por isso, será utilizada no decorrer deste trabalho.

Estes métodos transdutivos se constituem em frameworks formados por etapas que podem ser construção do grafo, ponderação e inferência (ENGELEN; HOOS, 2020) ou, alternativamente, construção do grafo de similaridade, identificação de documentos negativos e propagação de rótulos (CARNEVALI *et al.*, 2021). Neste trabalho será adotada a segunda abordagem.

2.4.1 Construção do grafo de similaridade

Partindo de um modelo de representação no espaço vetorial da base de dados, o grafo de documentos é construído, utilizando uma medida de similaridade entre os pares de vetores que representam os artigos. Uma medida muito utilizada é a similaridade de cosseno definida na Equação 2.2. As conexões são feitas apenas entre vizinhos mais próximos. Para estabelecer estas conexões podem ser usados algoritmos como *k-Nearest Neighbors (kNN)*, *Mutual k-Nearest Neighbors (MkNN)*, *ϵ-Neighborhood (ϵN)* e *b-matching*. *kNN* conecta cada artigo com seus k documentos mais similares. *MkNN* é semelhante, mas a conexão é criada apenas se ambos os objetos são vizinhos mais próximos entre si. ϵN conecta a cada objeto cuja similaridade é igual ou maior que o parâmetro ϵ. *B-matching* garante que cada nodo tenha o mesmo número de vizinhos, priorizando a regularidade do grafo (ENGELEN; HOOS, 2020).

2.4.2 Identificação de documentos negativos

Documentos pertencentes à mesma classe tendem a ser vizinhos no grafo. A partir desta consideração é assumido que os mais distantes são negativos. A distância mínima entre cada documento positivo e cada um dos não rotulados é calculada utilizando um algoritmo que calcula distâncias na rede. Um dos mais utilizados é o algoritmo de Dijkstra (DIJKSTRA, 2022). A seguir, é calculada a distância média entre cada documento não rotulado e todos os positivos e os artigos são ordenados em ordem decrescente de distância. Os mais afastados são retornados como negativos (CARNEVALI *et al.*, 2021).

2.4.3 Propagação de rótulos

Classifica os documentos não rotulados restantes em positivos ou negativos. A classe associada a documentos precisa respeitar duas condições: (i) respeitar a topologia da rede, com vértices vizinhos recebendo rótulos similares, e (ii) ser similar aos rótulos originais. Este resultado é obtido com a minimização de uma função de regularização cujos dois termos correspondem respectivamente a cada uma das duas condições que devem ser respeitadas (ENGELEN; HOOS, 2020).

$$Q(F) = \frac{1}{2} \sum_{o_i o_j \in O} w_{o_i,o_j} \Omega(f_{o_i} - f_{o_j}) + \mu \sum_{o_i \in O^L} \Omega'(f_{o_i} - y_{o_i}), \tag{2.3}$$

onde f_{o_i} e f_{o_j} são respectivamente os vetores dos objetos o_i e o_j que armazenam os pesos associados a cada classe pelo algoritmo de regularização, y_{o_i} é a classe original fornecida ao texto, onde $y_i = 1$ se o objeto o_i pertence à classe c_j e 0 em caso contrário. Ω e Ω' são respectivamente as distâncias entre os vetores presumidos dos documentos f_{o_i} e f_{o_j} e entre os vetores f e y do mesmo texto (CARNEVALI *et al.*, 2021).

O objetivo do primeiro termo é minimizar a distância entre as classes de dois vetores estimados no processo, ponderado pelo peso da conexão destes objetos. A iteração é feita sobre todos os documentos da rede. Quanto maior o peso da conexão, mais próximos precisam ser os respectivos vetores f para minimizar a Equação 2.3, fazendo com que documentos similares tendam a pertencer a mesma classe. O segundo termo é iterado apenas sobre os documentos inicialmente rotulados. Seu objetivo é fazer com que a distância entre o vetor de rótulo original e o de rótulo estimado seja mínima para não penalizar a equação. μ é uma variável que controla o quanto se deseja respeitar esta condição (ENGELEN; HOOS, 2020). F é uma matriz que contém todos os vetores f e o objetivo é encontrar F que minimize todo o processo. Para isso são usadas soluções iterativas que propagam a informação de classe dos objetos pela rede proporcionalmente ao peso das conexões entre os vetores que representam cada par de documentos e ficam ajustando os valores dos vetores de F para minimizar a a Equação 2.3. Quando isso acontece, o processo entra em convergência e a solução da rede é encontrada. No final, cada objeto vai ter um vetor associado com pesos tanto para a classe positiva quanto negativa e os documentos são classificados de acordo com a classe de maior peso.

Existem diferentes algoritmos baseados em regularização. Os mais conhecidos são: *Gaussian Fields and Harmonic Functions (GFHF)* (ZHU; GHAHRAMANI; LAFFERTY, 2003) e *Learning With Local and Global Consistency (LLGC)* (ZHOU *et al.*, 2003). O primeiro atualiza o vetor f de cada documento usando a média harmônica entre os vetores de documentos vizinhos. O segundo considera que eventualmente possam ter havido erros na rotulação dos documentos de interesse por um engano no processo manual e permite a alteração destes rótulos, pois caso contrário degradariam o desempenho do algoritmo. Além disso, erros de propagação são aumentados se documentos com alto grau são rotulados erroneamente. *LLGC* diminui a influência de objetos de alto grau na atualização da informação de classe de seus vizinhos (CARNEVALI *et al.*, 2021). Neste trabalho, os documentos de interesse são selecionados pela sua relação com o tema a ser pesquisado, sendo assim positivos e não devem ser alterados. Com isso, a escolha recai sobre o algoritmo *GFHF*.

3 PROPOSTA PARA SELEÇÃO DE BIBLIOGRAFIA CIENTÍFICA BASEADA EM PRESTÍGIO E SIMILARIDADE

O objetivo deste trabalho é, a partir de um pequeno conjunto de artigos científicos de interesse, retornar o conjunto de trabalhos mais relevantes relacionados a eles em um repositório, auxiliando no levantamento de bibliográfica científica. Como apresentado nos capítulos anteriores, para uma boa referência bibliográfica é necessário um trabalho de qualidade e que tenha relação com a pesquisa sendo realizada. Para a qualidade, o critério de PageRank mostra a importância de um documento, não apenas pelo número de citações, mas também pela importância desta citações, refletindo melhor o prestígio deste trabalho. Para a similaridade, com os artigos de interesse constituindo os poucos textos rotulados, se configura um cenário de aprendizado semi-supervisionado para classificação de texto, onde abordagens baseadas em grafos tem obtido desempenho superior aos modelos de espaço vetorial e onde o aprendizado transdutivo é de grande utilidade para classificar textos quando existem poucos documentos rotulados. Baseado nisso, foi proposta uma solução para seleção de bibliografia científica que utiliza tanto a rede de citações quanto a similaridade de textos entre os documentos, modeladas em redes complexas, com uma visão para cada um dos critérios, cujos resultados se combinam para uma solução final.

3.1 Rede de Citações

A visão da rede de citações é utilizada tanto para prestígio quanto para similaridade. O prestígio é avaliado pelo PageRank e a similaridade pelo número de citações que cada nodo tem em comum com os artigos de interesse.

Um grafo dirigido é criado e são acrescentadas arestas direcionadas de cada artigo de interesse e de cada selecionado para os que eles referenciam. Não são adicionadas arestas partindo dos artigos apenas referenciados, pois suas referências não formam o corpus. Os artigos apenas referenciados tem os trabalhos que citam registrados apenas para cálculo do número de documentos em comum com os de interesse, valor este armazenado como atributo do respectivo nodo.

Para determinar a relevância dos artigos, é calculado o PageRank dos documentos da rede e acrescentado como atributo de seu respectivo nodo.

3.2 Rede de Similaridade de Textos

A análise de similaridade de textos é feita em três etapas: construção do grafo, extração dos documentos negativos e propagação de rótulos.

3.2.1 Construção do grafo

Neste artigo são usadas três técnicas para representar computacionalmente os textos:

- TF-IDF
- Word2Vec, sendo utilizados embeddings pré-treinados com textos do *Google News* (*GoogleNews-vectors-negative300.bin*)
- BERT, sendo utilizado o SciBERT, que são embeddings pré-treinados em textos científicos do *Semantic Scholar* (BELTAGY; LO; COHAN, 2019)

A escolha das técnicas se baseia na facilidade de sua utilização futura em temas diversos, não necessitando esforço adicional de treinamento. As técnicas são utilizadas separadamente e, a partir de cada uma, é dada continuidade ao trabalho.

Com a representação vetorial de cada documento é construído um grafo, utilizando o algoritmo kNN (k-Nearest Neighbors), que liga cada nodo a seus k vizinhos mais similares, diminuindo o esforço computacional ao não ligar todos os nodos entre si. Para medir a proximidade é utilizada a similaridade de cosseno.

3.2.1.1 Extração de documentos negativos

Cada nodo no grafo está ligado a seus vizinhos mais próximos pelo critério similaridade e documentos similares tendem a pertencer à mesma classe. Com isso, artigos próximos aos da classe positiva tendem a ser positivos. Fazendo uma inferência para a classe negativa, é possível assumir que estes sejam os mais distantes (CARNEVALI *et al.*, 2021), considerando a distância como sendo a menor soma de pesos obtidos ao percorrer os caminhos possíveis entre os dois nodos. Para calcular esta distância é usado o algoritmo de Dijkstra. Ao utilizar a similaridade de cosseno, quanto mais próximos, maior a similaridade, maior o valor do cosseno. No entanto, ao calcular o menor caminho pelo algoritmo, pares mais próximos possuem uma menor distância entre eles. Assim sendo, o valor da similaridade é invertido para o cálculo de Dijkstra. É calculada a distância de cada documento positivo para todos os não rotulados. A seguir é calculada a distância média de cada documento não rotulado para todos os positivos. Os nodos são ordenados em ordem descendente de sua distância aos positivos e os melhores ranqueados, na mesma quantidade dos positivos, são retornados como negativos. Esta implementação segue a utilizada por (CARNEVALI *et al.*, 2021).

3.2.2 Propagação de rótulos

Um algoritmo de propagação espalha rótulos pela rede para os documentos ainda não identificados, a partir dos nodos positivos e negativos já rotulados. Dos dois algoritmos

mais utilizados, GFHF e LLGC, citados em 2.4.3, a escolha recaiu sobre o primeiro (ZHU; GHAHRAMANI; LAFFERTY, 2003) por não alterar os rótulos positivos dos artigos de interesse, uma vez que, ao fornecê-los ao sistema, se tem certeza que são de interesse e, portanto, positivos.

Cada nodo recebe como atributo um valor para a classe positiva e um para a classe negativa, o qual é inicialmente zerado, à exceção dos já rotulados que recebem o valor 1 em sua respectiva classe e 0 na outra. Estes documentos positivos e negativos rotulados não tem o valor de suas respectivas classes alterado durante o processamento. Um laço que se repete até o número máximo de iterações é iniciado, onde, o valor de cada classe de cada elemento não rotulado inicialmente é atualizado pela média harmônica do valor da classe de cada um de seus vizinhos multiplicado pelo peso da ligação entre ele e seu respectivo vizinho. A diferença entre o valor antigo e o calculado na iteração para cada classe de cada nodo é adicionada a um contador geral que fornece, a final de cada iteração, o valor da diferença da iteração atual para a anterior. Se esta diferença for menor que determinado valor, o processo converge e pode ser encerrado antes de alcançar o número máximo de iterações.

A propagação de rótulos é um algoritmo de classificação. Os documentos são classificados como positivos ou negativos de acordo com a classe de maior valor associada a seu nodo. Para fornecer referência bibliográfica, não há interesse em conhecer todos os elementos positivos, apenas os mais significativos. Com isso, o algoritmo de classificação é computado e os valores associados às classes positiva e negativa de cada nodo são utilizados para outra tarefa, neste caso de recuperação de informação, onde informações são recuperadas em resposta a consultas formuladas pelos usuários.

3.3 Avaliação

São feitas três ordenações em ordem decrescente dos nodos, uma pelo PageRank, uma considerando o número de artigos em comum com os de interesse e a última pelo valor atribuído à classe positiva, incluindo apenas nodos onde o valor desta classe é superior ao da classe negativa. Para os três casos acima, a ordem na classificação é escolhida como o valor a ser utilizado no cálculo final, para evitar que valores muito extremos, mesmo normalizados, distorçam o resultado. Para a seleção final são utilizadas duas ponderações entre os três critérios: o mesmo peso para todos e um maior para o número de artigos em comum com os de interesse, tendo o dobro do peso dos outros dois critérios.

4 AVALIAÇÃO EXPERIMENTAL

4.1 Conjuntos de Dados

A escolha do repositório para extração dos dados levou em conta a disponibilização em forma de metadados de título, resumo e, principalmente, referências bibliográficas, essenciais para a execução do trabalho. A grande variação na estrutura dos textos de artigos bem como nos formatos utilizados nas referências bibliográficas tornaria sua extração diretamente do texto uma tarefa bastante complexa. Repositórios como *arXiv*, *Dblp*, *Web of Science*, entre outros foram testados sem sucesso. A escolha recaiu sobre o banco de dados *OpenAlex*. Seus dados são coletados de várias fontes, incluindo *Crossref*, *PubMed* e *arXiv*. Vários trabalhos não muito recentes vieram do *Microsoft Academic Graph (MAG)*, descontinuado em 2021. Eles podem ser acessados através de uma API gratuita e sem autenticação, sendo uma boa opção em relação a outras bases pagas. Também podem ser usadas bibliotecas de terceiros, como é o caso da *PyAlex*, utilizada neste trabalho.

OpenAlex (PRIEM; PIWOWAR; ORR, 2023) inclui entidades acadêmicas e as conexões destas entidades entre si, formando um grafo direcionado heterogêneo. Todas as entidades tem um id associado, o qual é a chave primária na base de dados. O repositório é formado atualmente por sete tipos de entidades acadêmicas:

- works: documentos acadêmicos como artigos, livros, bases de dados e teses.
- authors: pessoas que criaram os documentos.
- sources: onde os works estão hospedados, como revistas, periódicos, conferências e repositórios.
- institutions: universidades e outras organizações às quais os autores alegam estar afiliados.
- concepts: temas, assuntos, idéias abstratas sobre as quais tratam os documentos.
- publishers: companhias e organizações que distribuem os documentos.
- funders: organizações que financiam pesquisas.

Neste trabalho as entidades acadêmicas utilizadas foram concepts e works.

Concepts ou conceitos são hierárquicos, constituindo uma árvore de seis níveis, onde o primeiro nível é composto por 19 itens. São indexados aproximadamente 65000 concepts (PRIEM; PIWOWAR; ORR, 2022). Cada conceito possui um id no *OpenAlex*.

Aproximadamente 85% dos documentos são rotulados com ao menos um concept. Conceitos são ligados aos documentos pela propriedade concepts.

Works ou documentos é uma entidade particularmente importante, pois quase todas as outras tem conexão com ela (PRIEM; PIWOWAR; ORR, 2022). OpenAlex indexa aproximadamente 240.000.000 documentos, sendo que diariamente são anexados por volta de 50.000 trabalhos (PRIEM; PIWOWAR; ORR, 2023). Um objeto Work contém toda a informação que *OpenAlex* possui sobre determinado documento. Os atributos de interesse para este trabalho foram:

- id: a identificação do *OpenAlex* para o documento
- title: título do documento
- language: o idioma no qual foi escrito o documento no formato *ISO 639-1*. É automaticamente detectado usando as palavras do resumo ou do título se não houver resumo. Não é um método perfeito, podendo haver incoerências. Ele não é reportado se não existirem palavras suficientes para uma conclusão. No caso do texto ser em um idioma diferente do resumo ou título, é considerado apenas o idioma destes últimos.
- abstract_inverted_index: resumo do documento como um índice invertido, que guarda informação sobre as palavras do resumo e sua posição no texto. *OpenALex* não inclui o texto original do resumo por restrições legais. Trabalhos mais recentes tem uma tendência maior de possuir índice invertido.
- referenced_works: IDs do *OpenAlex* dos trabalhos citados pelo documento.
- referenced_works_count: número de trabalhos que o documento cita.
- cited_by_count: número de citações que o documento teve de outros.
- concepts: lista dos objetos concepts atribuídos ao documento. Cada documento é rotulado com múltiplos conceitos, tendo como base título e resumo. A rotulação é feita usando um classificador automático treinado com o corpus do *MAG*. Há uma pontuação para cada concept em um documento, a qual mostra a confiança do classificador em escolher aquele item. Conceitos com pontuação 0.3 ou maior são atribuídos ao documento. No entanto, os antecessores destes conceitos também são adicionados à lista, o que pode ocasionar valores muito baixos, eventualmente zero. (PRIEM; PIWOWAR; ORR, 2023).

4.1.1 Seleção dos dados

O corpus foi formado por artigos de interesse, trabalhos selecionados no *OpenAlex* e suas referências bibliográficas. Os artigos de interesse foram fornecidos pelo usuário e

tinham relação com o assunto para o qual se desejava obter bibliografia científica. Foram analisados três cenários com dois, quatro e sete artigos de interesse para avaliar a influência do tamanho da entrada no resultado final. Em um primeiro momento, o critério de seleção na base de dados foi a presença da expressão *semi-supervised learning* no título, visando garantir a relação do texto com essa expressão. Foi observado que, ao mesmo tempo em esta relação era garantida, era excluída uma série de documentos relevantes sobre o tema que não apresentavam esta expressão no título. Como a relevância do trabalho é importante para sua indicação bibliográfica, se partiu, então, para uma seleção mais ampla, escolhendo textos classificados pelo *OpenAlex* com conceito *semi-supervised learning* (concept id *C58973888*). Este critério não é determinístico, uma vez que a rotulação dos conceitos é feita por um classificador automático, podendo aparecer, eventualmente, textos não muito relacionados ao assunto. Completando o corpus foram incluídas as referências bibliográficas de cada um dos artigos acima citados.

Do corpus original foram excluídos os textos que o *OpenAlex* não reconheceu como sendo escritos na língua inglesa, pois o processamento foi feito neste idioma. Foram descartados também artigos sem título ou sem resumo registrados, pois ambos são concatenados para formar o texto a partir do qual a similaridade é calculada. Sem um deles há o risco de restarem poucas palavras para caracterizar corretamente o documento. Também não foram considerados documentos que possuíam um id no *OpenAlex*, mas esta identificação não era encontrada no repositório, trabalhos cujo resumo era um texto padrão e não o seu texto verdadeiro do resumo e artigos que foram DOIs criadas com erro e que, não podendo ser apagadas, permaneceram na base, mas sem metadados associados. Este texto resultante foi convertido para letras minúsculas, foram removidos stopwords, dígitos, caracteres especiais e pontuações e foi lematizado.

Também foi guardada a relação dos trabalhos referenciados por todos os artigos de interesse com o número de vezes em que foram citados.

4.2 Configuração Experimental

Foram selecionados inicialmente 1200 artigos da base de dados do *OpenAlex*. Após as exclusões elencadas na seção anterior, restaram 1040 textos. Acrescentando os artigos de interesse e referências bibliográficas, o corpus totalizou 13752, 13772 e 13864 documentos para 2, 4 e 7 artigos de referência, respectivamente.

O framework foi avaliado utilizando 2, 4 e 7 artigos de interesse. A seleção destes textos procurou ser direcionada para o conteúdo deste trabalho, como uma forma de auxiliar na sua elaboração. Como é um assunto sem muitas publicações específicas, a escolha dos textos procurou abranger os vários aspectos abordados, partindo do texto base (CARNEVALI *et al.*, 2021), juntamente com (STERLING; MONTEMORE, 2021), que foram utilizados como artigos de interesse em todas as seleções. A escolha dos artigos foi

feita de modo a não tender apenas para um assunto específico. Isso também limitou a quantidade de entradas, pois esse balanceamento se torna mais sensível a medida que o número de artigos cresce.

A avaliação também foi feita considerando cada uma das representações vetoriais de textos mencionadas em 3.2.1.

- TF-IDF: utilizou *min_df* de 3, ignorando termos que aparecem em menos de 3 documentos.
- Word2Vec: os embeddings pré-treinados foram construídos a partir de textos do *Google News* (*GoogleNews-vectors-negative300.bin*)
- BERT: utilizou o SciBERT, embeddings pré-treinados em textos científicos do *Semantic Scholar* (BELTAGY; LO; COHAN, 2019). Como vários textos tem mais de 512 tokens, limite do BERT, foi usada a técnica de dividi-los em segmentos, calcular o embedding para cada segmento e finalmente calcular a média dos embeddings do texto.

Para a construção de grafo foi utilizado o algoritmo kNN, com k=3. Na extração de documentos negativos são retornados nodos na mesma quantidade dos artigos de interesse que foram fornecidos como entrada. Para a propagação de rótulos, o número máximo de iterações é 100 e a diferença entre as iterações para definir a convergência é 0,0001.

4.3 Resultados e Discussões

Nesta seção são apresentados os resultados obtidos considerando a configuração experimental mostrada na seção anterior. A seleção considerou três critérios: similaridade de texto com os artigos de interesse, número de referências em comum com estes artigos e PageRank. Foram analisados nove cenários ligados à similaridade de textos, resultantes das variações na quantidade de artigos de interesse e na representação vetorial dos documentos. Os dois outros critérios se mantiveram fixos, não tendo alterações. Os critérios utilizados na seleção foram combinados de modos diferentes para a obtenção do resultado final. O primeiro modo teve uma ponderação igual para todos eles. Foram feitas tentativas ponderando cada um dos critérios com o dobro do valor dos outros. Um valor maior atribuído para similaridade ou PageRank tendeu o resultado, respectivamente, para documentos semelhantes, mas que pouco acrescentavam, ou relevantes, mas com pouca relação com o assunto. O último cenário se configurou ainda pior que primeiro, pois um documento sem relação com a pesquisa não consegue ser utilizado nela. Com isso, essas duas opções foram descartadas, restando a atribuição do peso maior para o número de artigos em comum com os de interesse. O valor de cada nodo utilizado nas análises foi o seu índice na ordenação decrescente dos resultados de cada critério. Cada cenário

retornou cinco documentos. Apesar de existirem nove cenários, retornando cada um cinco artigos, o que poderia fornecer até 45 documentos para cada combinação de critérios, foram retornados ao todo 14 artigos diferentes quando foi usada a mesma ponderação dos critérios e 16 quando a ponderação foi o dobro para o número de artigos em comum com os de interesse. Além disso, os 14 primeiros artigos estavam incluídos nos 16 últimos. Foi observada uma tendência, para uma mesma representação vetorial de textos, de retorno dos mesmos documentos, independente da quantidade de textos fornecidos na entrada.

Os resultados foram avaliados inicialmente pela similaridade por um especialista na área. Para os outros dois critérios foi feita a análise de seus valores e a sua compatibilidade com os resultados, principalmente onde o critério similaridade não se destacou. Todos os textos retornados foram considerados como tendo relação com os artigos de interesse, alguns com maior relação que outros. A Figura 3 mostra o resultado. Nesta análise, cada artigo retornado é considerado uma vez, independentemente do número de cenários em que foi escolhido.

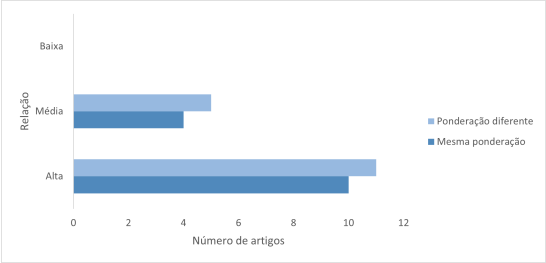

Figura 3 – Avaliação de Similaridade

Baseada nesta avaliação, foi feita uma análise dos classificados com uma relação média em relação aos de interesse, ressaltando que mesmos estes foram considerados relacionados. A Tabela 1 mostra a quantificação destes documentos para cada um dos cenários analisados para similaridade de textos, considerando o caso de uma ponderação igual para todos os critérios.

Metodo/Entrada	**2**	**4**	**7**
TF-IDF	1	0	1
Word2Vec	3	2	2
SciBERT	0	0	0

Tabela 1 – Documentos com relação média com os de interesse

Em relação às representações vetoriais de texto, pode ser observado pela Tabela 1 que Word2Vec concentra o maior número de retornos com sete ocorrências, seguido do TF-IDF com duas e nenhuma para o SciBERT. Para um mesmo número de artigos

de interesse e como o PageRank é uma característica do nodo e se mantém constante para cada documento, as variações ao longo de cada coluna mostram que em todos os casos o pior desempenho foi do Word2Vec. Este resultado é compatível com a fato do Word2Vec utilizado ser constituído por embeddings pré-treinadas do *Google News*, um corpus de notícias, não específico para a área científica. Já o SciBERT, específico para esta área, teve o melhor desempenho. Com relação ao número de artigos de interesse, o melhor desempenho foi com quatro entradas, com pouca variação em relação às outras duas opções. A composição destes grupos de documentos precisou de muito critério, abordando todos os tópicos da área de interesse em proporção semelhante. Esta composição se mostrou mais importante para o resultado final do que a quantidade de artigos fornecidos na entrada. Cabe ressaltar, que apesar da análise acima, todos documentos selecionados foram considerados relacionados aos de interesse.

Na análise dos outros dois critérios, todos os artigos selecionados estavam entre os 3% com maior PageRank e entre os 8% com mais referências em comum com os de interesse. A exceção de um deles, todos os classificados como relação média de similaridade de texto estavam entre os 2% melhores nestes dois critérios, o que justificou sua escolha.

5 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou uma abordagem para auxiliar o levantamento de referências bibliográficas, um problema recorrente entre pesquisadores e estudantes que gastam muito tempo e esforço nesta tarefa e muitas vezes não conseguem bons resultados. Uma primeira aprendizagem foi sobre a compreensão do que é uma boa referência. Precisa ter relação com o assunto que se deseja pesquisar, mas muitas vezes o que se tem de retorno não acrecenta nenhum valor adicional ao trabalho. Precisa ser relevante, mas sem relação com o assunto é descartado por não tem utilidade. O que se busca é uma ponderação entre os dois critérios, procurando textos com um bom grau de semelhança e relevância, mas sempre levando em conta que a semelhança não pode deixar de existir. Um trabalho similar e não muito importante pode ser útil, mas um relevante e sem relação com o assunto não tem nenhuma utilidade. Neste trabalho, estes critérios foram tratados da mesma forma. Para trabalhos futuros, uma linha a ser seguida é a exigência de similaridade com a relevância completando o critério.

Outra aprendizagem foi a importância dos algoritmos e, principalmente, dos dados. Entre tantos algoritmos de inteligência artificial, o desafio foi definir o que melhor se adaptava ao problema em questão, com poucos dados de entrada. Outros existem e podem ser seguidos no futuro, como, por exemplo, a utilização de GNNs. Uma vez definido o algoritmo, encontrar os dados necessários e conhecê-los foi uma tarefa que perdurou ao longo de todo o trabalho. Analisando estes dados, uma série de outras linhas de pesquisa podem ser seguidas. Utilizar a rede de autores e até a data, pois muitas vezes bons trabalhos recentes são mais interessantes. Aumentar e diversificar a base de dados também é uma opção.

A abordagem escolhida indicou vários textos similares e relevantes. Como o tema escolhido para os artigos de interesse foi o assunto deste trabalho, alguns documentos retornados foram úteis na sua elaboração, mas a grande maioria ainda não pode ser analisada. A intenção para o futuro é refazer esta pesquisa, utilizando a bibliografia indicada no trabalho atual.

REFERÊNCIAS

AMANCIO, D. *et al.* Using complex networks concepts to assess approaches for citations in scientific papers. **Scientometrics**, Springer, v. 91, p. 827–842, 2012.

ANGELOVA, R.; WEIKUM, G. Graph-based text classification: learn from your neighbors. *In*: **Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval**. [*S.l.: s.n.*], 2006. p. 485–492.

BAI, X. *et al.* Scientific paper recommendation: A survey. **IEEE Access**, v. 7, p. 9324–9339, 2019.

BARABÁSI, A.-L. Network science. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, The Royal Society Publishing, v. 371, n. 1987, p. 20120375, 2013.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999.

BELTAGY, I.; LO, K.; COHAN, A. Scibert: Pretrained language model for scientific text. *In*: **EMNLP**. [*S.l.: s.n.*], 2019.

BOCCALETTI, S. *et al.* Complex networks: Structure and dynamics. **Physics reports**, Elsevier, v. 424, n. 4-5, p. 175–308, 2006.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer networks and ISDN systems**, Elsevier, v. 30, n. 1-7, p. 107–117, 1998.

CARNEVALI, J. C. *et al.* A graph-based approach for positive and unlabeled learning. **Information Sciences**, Elsevier, v. 580, p. 655–672, 2021.

DEVLIN, J. *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019.

DIJKSTRA, E. W. A note on two problems in connexion with graphs. *In*: **Edsger Wybe Dijkstra: His Life, Work, and Legacy**. [*S.l.: s.n.*], 2022. p. 287–290.

ENGELEN, J. E. V.; HOOS, H. H. A survey on semi-supervised learning. **Machine learning**, Springer, v. 109, n. 2, p. 373–440, 2020.

FALEIROS, T. d. P. **Propagação em grafos bipartidos para extração de tópicos em fluxo de documentos textuais**. 2016. Tese (Doutorado) — Universidade de São Paulo, 2016.

JURAFSKY, D.; MARTIN, J. **Speech and language processing**. 2023. Available at: <https://web.stanford.edu/~jurafsky/slp3/>. Access at: 17 mar. 2023.

LAGUNA, V. A.; LOPES, A. de A. A multi-view approach for semi-supervised scientific paper classification. **WAAMD**, v. 9, p. 26–33, 2009.

LIU, H. *et al.* Link prediction in paper citation network to construct paper correlation graph. **EURASIP Journal on Wireless Communications and Networking**, Springer, v. 2019, n. 1, p. 1–12, 2019.

MA, N.; GUAN, J.; ZHAO, Y. Bringing pagerank to the citation analysis. **Information Processing & Management**, Elsevier, v. 44, n. 2, p. 800–810, 2008.

MIKOLOV, T. *et al.* Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

NEWMAN, M. E. J. The structure and function of complex networks. **SIAM Review**, v. 45, n. 2, p. 167–256, 2003.

PRIEM, J.; PIWOWAR, H.; ORR, R. **OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts**. 2022.

PRIEM, J.; PIWOWAR, H.; ORR, R. **OpenAlex API documentation**. 2023. Available at: <https://docs.openalex.org/>. Access at: 08 out. 2023.

QIU, X. *et al.* Pre-trained models for natural language processing: A survey. **Science China Technological Sciences**, Springer, v. 63, n. 10, p. 1872–1897, 2020.

ROSSI, R. G. **Classificaçao automática de textos por meio de aprendizado de máquina baseado em redes**. 2016. Tese (Doutorado) — Universidade de São Paulo, 2016.

STERLING, J. A.; MONTEMORE, M. M. Combining citation network information and text similarity for research article recommender systems. **IEEE Access**, IEEE, v. 10, p. 16–23, 2021.

ZHOU, D. *et al.* Learning with local and global consistency. **Advances in neural information processing systems**, v. 16, 2003.

ZHU, X.; GHAHRAMANI, Z.; LAFFERTY, J. D. Semi-supervised learning using gaussian fields and harmonic functions. *In*: **Proceedings of the 20th International conference on Machine learning (ICML-03)**. [*S.l.: s.n.*], 2003. p. 912–919.