UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE DO CAMPUS ARARANGUÁ
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

Patrick Davila Kochan

Um Método de Análise de Patentes Voltado à Previsão de Tecnologias

Araranguá
2022

Patrick Davila Kochan

Um Método de Análise de Patentes Voltado à Previsão de Tecnologias

Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia de Computação do Centro de Ciências, Tecnologias e Saúde do Campus Araranguá da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Engenharia de Computação.
Orientador: Prof. Alexandre Leopoldo Gonçalves, Dr.

Araranguá
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Kochan, Patrick Davila
Um Método de Análise de Patentes Voltado à Previsão de Tecnologias / Patrick Davila Kochan ; orientador, Alexandre Leopoldo Gonçalves, 2022.
38 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Engenharia de Computação, Araranguá, 2022.

Inclui referências.

1. Engenharia de Computação. 2. Predição de Ligações. 3. Previsão de Tecnologias. 4. Redes Neurais Rasas. I. Gonçalves, Alexandre Leopoldo. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Computação. III. Título.

Patrick Davila Kochan

Um Método de Análise de Patentes Voltado à Previsão de Tecnologias

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Engenharia de Computação e aprovado em sua forma final pelo Curso de Graduação em Engenharia de Computação.

Araranguá, 17 de março de 2022.

Documento assinado digitalmente
Analucia Schiaffino Morales
Data: 25/03/2022 09:30:00-0300
CPF: 622.256.420-87
Verifique as assinaturas em https://v.ufsc.br

Profª. Analúcia Schiaffino Morales, Drª.
Coordenadora do Curso

Banca Examinadora:

Documento assinado digitalmente
Alexandre Leopoldo Goncalves
Data: 25/03/2022 09:10:50-0300
CPF: 786.740.349-34
Verifique as assinaturas em https://v.ufsc.br

Prof. Alexandre Leopoldo Gonçalves, Dr.
Orientador

Documento assinado digitalmente
Andrea Sabedra Bordin
Data: 25/03/2022 08:43:22-0300
CPF: 626.193.220-49
Verifique as assinaturas em https://v.ufsc.br

Profª Andréa Sabedra Bordin, Drª.
Avaliadora
Universidade Federal de Santa Catarina

Documento assinado digitalmente
ALESSANDRO BOTELHO BOVO
Data: 25/03/2022 08:57:07-0300
CPF: 025.718.019-29
Verifique as assinaturas em https://v.ufsc.br

Prof. Alessandro Botelho Bovo, Dr.
Avaliador
Universidade Tecnológica Federal do Paraná

Prof. Fábio Rodrigues de la Rocha, Dr.
Avaliador Suplente
Universidade Federal de Santa Catarina

Um Método de Análise de Patentes Voltado à Previsão de Tecnologias

A Patent Analysis Method Toward Forecasting Technologies

Patrick Davila Kochan [*] Alexandre Leopoldo Gonçalves [†]

2022, Março

Resumo

Patentes são amplas e confiáveis fontes de dados de invenções tecnológicas. A fim de viabilizar a extração de informação desses documentos para aplicá-las a um objetivo, a análise de patentes abrange diversas tarefas. Entre elas, a tarefa de previsão de tecnologias é essencial para o processo de tomada de decisões estratégicas em organizações e no desenvolvimento de pesquisas, visto seu potencial de geração de inovação e vantagem competitiva. Neste contexto, o presente trabalho propõe um método de apoio à tomada de decisões capaz de identificar tendências tecnológicas e reduzir a extensiva análise manual de patentes por parte dos especialistas. Para cumprir este objetivo, explora-se o aprendizado de representação de redes através do algoritmo Node2vec e predição de ligações para identificar possíveis tendências de conexões entre tecnologias. Um estudo de caso sobre o domínio de carbono é apresentado para demonstrar a efetividade do método proposto considerando o conjunto de patentes de utilidade USPTO-2M®. Para tal, foram realizadas análises das predições em cenários de curto (um ano), médio (três anos) e longo (cinco anos) prazo. O cenário que melhor desempenhou foi o de longo prazo, onde as predições realizadas com o limiar de 85% alcançaram a média de 53,98% de acerto levando em conta predições realizadas no intervalo de 2006 a 2010. O modelo de predição de ligações atingiu uma média de 0,91 considerando a métrica ROC-AUC. Diante dos resultados, o modelo de predição construído demonstrou ser capaz de promover indícios de possíveis tendências tecnológicas e o método proposto mostrou-se viável para a conjuntura apresentada, sendo apto a contribuir com o processo de tomada de decisão sobre investimentos em Pesquisa, Desenvolvimento e Inovação.

Palavras-chaves: Predição de Ligações. Previsão de Tecnologias. Redes Neurais Rasas.

[*]patrick.kochan@grad.ufsc.br

[†]a.l.goncalves@ufsc.br

Um Método de Análise de Patentes Voltado à Previsão de Tecnologias

A Patent Analysis Method Toward Forecasting Technologies

Patrick Davila Kochan [*] Alexandre Leopoldo Gonçalves [†]

2022, Março

Abstract

Patents are extensive and reliable sources of data for technological inventions. In order to make it possible to extract information from these documents to apply them to a goal, patent analysis covers several tasks. Among them, the technology forecasting task is essential for the strategic decision-making process in organizations and in research development, given its potential to generate innovation and competitive advantage. In this context, the present work proposes a decision support method capable of identifying technological trends and reducing the extensive manual analysis of patents by specialists. To fulfill this objective, we explore the network learning representation through the Node2vec algorithm and link prediction to identify possible trends in connections between technologies. A case study on the carbon domain is presented to demonstrate the effectiveness of the proposed method considering the USPTO-2M® utility patent data set. To this end, analysis of the predictions were carried out in short (one year), medium (three years) and long (five years) term scenarios. The scenario that performed the best was the long-term one, where the predictions made with the threshold of 85% reached an average of 53.98% of accuracy, taking into account predictions made in the interval from 2006 to 2010. The link prediction model reached an average of 0.91 considering the ROC-AUC metric. Thus, the prediction model built proved to be capable of promoting indications of possible technological trends and the proposed method proved to be viable for the presented situation, being able to contribute to the decision-making process on investments in Research, Development and Innovation.

Key-words: Link Prediction. Technology Forecasting. Shallow Neural Networks.

[*]patrick.kochan@grad.ufsc.br

[†]a.l.goncalves@ufsc.br

1 Introdução

Com o advento da Indústria 4.0 o ambiente organizacional tornou-se mais complexo e competitivo. Essa revolução permitiu às instituições inovar em seus serviços, processos e produtos, de forma a estabelecer diferenciais fundamentais para a permanência no mercado (OTTONICAR; VALENTIM; MOSCONI, 2018). Para as organizações, a capacidade de inovar destaca-se como um dos fatores determinantes para alcançar desempenho superior aos concorrentes (CHEN; WANG; HUANG, 2020).

Nesta direção, a literatura de patentes se sobressai como uma ampla e valiosa fonte de conhecimento e invenções tecnológicas para pesquisadores, organizações e comunidades de inovação (KRESTEL et al., 2021; YUAN; CAI, 2021). Estes documentos oferecem dados confiáveis e refletem os avanços do desenvolvimento tecnológico da sociedade, o que tornou a análise de patentes uma ferramenta vital para a formulação de estratégias na área de Pesquisa, Desenvolvimento e Inovação (PD&I) (ZHANG; LIU, 2020). De acordo com Zhang e Liu (2020), aproximadamente 80% do conhecimento de tecnologia do mundo pode ser encontrado em patentes. Essa informação é relevante visto que tecnologias desempenham um importante papel ao se tratar de geração de inovação, pois a inovação pode ser alcançada através da convergência de tecnologias ou áreas tecnológicas (KIM; BAE, 2017).

O propósito da análise de patentes é diverso em termos de tomada de decisão técnica e econômica. Os dados resultantes são utilizados para fornecer indicadores de capacidade tecnológica, identificar novidades e violações de patentes, identificar concorrentes e parceiros, analisar tendências de tecnologias, determinar a qualidade de patentes e avaliar o processo de inovação (ZHANG; LIU, 2020). A partir disso, prioridades em investimento de PD&I podem ser definidas, assim como a construção de possíveis estratégias de aquisição e fusão de tecnologias (YOON; LEE, 2008; ZHANG; LIU, 2020).

No entanto, o rápido crescimento do número de patentes representa um grande desafio para a recuperação e análise de suas informações de maneira efetiva e, para este fim, a análise de patentes descreve um grupo de tarefas que podem ser parcialmente automatizadas. Uma dessas tarefas é a previsão de tecnologias, na qual as patentes são utilizadas para avaliar um determinado cenário tecnológico, auxiliando pesquisadores a identificar tendências tecnológicas ou novas tecnologias (CHOI; SONG, 2018; KRESTEL et al., 2021). Esta tarefa auxilia no gerenciamento de risco de inovações e tecnologias emergentes, assim como no processo de tomada de decisões orientado a dados provendo ferramental que tem se tornado cada vez mais necessário para economias, organizações e a sociedade em geral. Sendo assim, a previsão de tecnologias configura uma oportunidade para instituições públicas e privadas, as quais se beneficiam através do amparo de investimento de capital, potencialização da identificação de oportunidades de negócio e da redução de riscos no processo decisório (HALEEM et al., 2019; KRESTEL et al., 2021).

De modo geral, este cenário vem promovendo um aumento exponencial no volume de dados em decorrência dos desdobramentos da transformação digital, como a interconectividade e tecnologias como *Big Data* e Aprendizado de Máquina (do inglês *Machine*

Learning - ML), bem como tem evidenciado a Análise de Redes (do inglês *Network Analysis* - NA) como uma crescente área de pesquisa (DADU et al., 2019; OTTONICAR; VALENTIM; MOSCONI, 2018). Como suporte a esta área, grafos ocupam uma importante posição ao longo da história, devido a sua ampla capacidade de caracterizar problemas do mundo real. Sua representação rica dos dados revela as relações entre entidades de maneira simples, em que as entidades são modeladas como vértices e os relacionamentos entre elas são dispostos em forma de arestas (WANG et al., 2019). Essa estrutura tem sido utilizada para representar redes sociais, conexões entre páginas *web*, mapas geográficos e inúmeras outras possibilidades de interconexão entre dados (GUPTA; MATTA; PANT, 2021).

Ainda, tratando-se do campo de análise de redes, a predição de ligações é uma tarefa que vem recebendo destaque nas últimas décadas e abrange os mais diversos domínios de aplicação, como a recomendação de amigos em redes sociais, interações proteína-proteína no campo da bioinformática e sistemas de recomendações de produtos em *e-Commerces* (KUMAR et al., 2020). Para este fim, diferentes técnicas são utilizadas para mensurar a probabilidade de dois nós de uma rede se conectarem (WU et al., 2021).

Neste sentido, o presente trabalho propõe um método para realização de predição de ligações no contexto de análise de patentes, mais estritamente na tarefa de previsão de tecnologias. Em suma, as contribuições ocorrem em duas camadas:

1. Da perspectiva metodológica, a proposição de uma abordagem que busca amparar a tomada de decisão por parte de gestores de PD&I e reduzir o extensivo trabalho de análise de patentes realizado por especialistas. O grande volume de documentos analisados de forma automática garante a cobertura de mais oportunidades tecnológicas e as predições realizadas apoiam o processo decisório. Desta maneira, o método colabora com os três pilares das organizações: pessoas, processos e tecnologias.
2. Do ponto de vista de aplicação, a pesquisa apresenta grande potencial de negócio. O método proposto pode ser aplicado independentemente do porte da empresa, uma vez que o domínio e sua lista de tecnologias são definidos de acordo com os interesses da organização. Além disso, a abordagem auxilia a identificação de tendências tecnológicas e o gerenciamento de risco da tomada de decisões, o que é de interesse tanto de instituições privadas quanto públicas.

Sendo assim, esta pesquisa científica tem sua relevância ao apontar um possível método de identificação de tendências tecnológicas a partir do conceito de predição de ligações e por contribuir com o suporte no processo de tomada de decisão estratégica em ambientes organizacionais. Ademais, espera-se que os resultados deste estudo possam ajudar na promoção de novas pesquisas em temas correlatos.

Além desta seção, este trabalho é composto por cinco outras. A segunda seção introduz a fundamentação teórica desta pesquisa e abrange conceitos essenciais para a compreensão do tema em sua totalidade. Na terceira seção, discussões acerca de trabalhos relacionados à previsão de tecnologias e predição de ligações são realizadas. O método proposto é descrito na quarta seção e então seus resultados são analisados e discutidos na quinta seção. Por fim, na sexta seção são apresentadas as conclusões e sugestões de trabalhos futuros são levantadas.

2 Fundamentação Teórica

2.1 Análise de Patentes

Com o rápido desenvolvimento econômico da sociedade, a alta tecnologia exerce um papel crucial no desenvolvimento das empresas, dos países e de toda a sociedade (ABBAS; ZHANG; KHAN, 2014). Como tecnologias específicas aplicadas em projetos desempenham um papel importante, elas devem ser analisadas e avaliadas antes da realização dos projetos a fim de garantir que projetos futuros possuam boas perspectivas de PD&I. O risco atrelado à utilização de uma tecnologia é usado principalmente para identificar se a tecnologia aplicada em um projeto tem valor de PD&I, se há riscos de violação e assim por diante. Uma vez que as patentes são fontes úteis de conhecimento sobre o progresso técnico e a atividade inovadora, e constituem um recurso confiável que reflete os avanços no desenvolvimento tecnológico, a análise de patentes tem sido considerada uma ferramenta vital para a formulação de estratégias na área de PD&I (ZHANG; LIU, 2020).

Abbas, Zhang e Khan (2014) afirmam que o volume cada vez maior de dados técnicos relativos às invenções tecnológicas tornou extremamente trabalhosa a tarefa de análise. A quantidade de patentes existentes compreende milhões de documentos espalhados por diferentes bases de dados disponíveis em fontes na *web* (dados abertos) ou privadas. Os repositórios mais populares para esses documentos são o *United States Patent and Trademark Office*® (USPTO), o *European Patent Office*® (EPO) e o *Japan Patent Office*® (JPO). Segundo Yoon e Lee (2008), confiar unicamente nos conhecimentos e habilidades de especialistas para analisar patentes tornou-se inviável e, sendo assim, o uso de ferramentas auxiliares passa a ser indispensável. As ferramentas automatizadas empregadas não apenas reduzem a extensiva análise manual de patentes por parte dos especialistas, mas também aceleram as etapas do processo: extração de patentes de bancos de dados; extração de informações das patentes; e a análise das informações extraídas.

Pode-se classificar as informações extraídas de uma patente em dados estruturados e não estruturados. Dados estruturados têm um formato consistente e padronizado, como o número de patente, número da Classificação Internacional de Patentes (do inglês *International Patent Classification* - IPC) e a família da patente. Já os dados não estruturados compreendem texto narrativo, como o título da patente, resumo, descrição e reivindicações (ABBAS; ZHANG; KHAN, 2014). Sendo assim, pode-se realizar a análise dessas informações de maneira quantitativa ou qualitativa. A análise de redes é um popular exemplo de um método avançado de análise de dados estruturados e pode ser aplicada com objetivos variados, como monitorar o desenvolvimento tecnológico, identificar concorrência, sugerir a integração de tecnologias interdisciplinares e monitorar a violação de patentes (ZHANG; LIU, 2020).

De acordo com Krestel et al. (2021) e Lupu et al. (2017), o estudo de patentes requer um trabalho aprofundado de especialistas e a avaliação de um enorme volume de dados. Todo o trabalho que envolve a análise de patentes pode ser divido em tarefas, as quais podem ser parcialmente automatizadas. Essas tarefas, Krestel et al. (2021), podem ser divididas em oito principais: i) as tarefas de suporte possuem a função de extração de informação, segmentação e tradução dos documentos; ii) a classificação de patentes trata da categorização desses documentos em códigos definidos por esquemas; iii) a recuperação de patentes busca pelo estado da arte de uma patente e encontra outras relacionadas com o

intuito de identificar o cenário da patente; iv) a avaliação de mercado avalia a qualidade da patente com o intuito de determinar seu valor de mercado; v) a geração de dados de patentes é utilizada para automatizar a elaboração destes documentos; vi) a análise de conflito de interesses em patentes zela pelo processo legal de litígio, onde ocorre a disputa de patentes entre duas empresas; vii) a tarefa de visão computacional em patentes trabalha com as figuras, fluxogramas e fluxos de trabalho dos documentos a fim de representar a invenção; e por fim, viii) a tarefa de previsão de tecnologia, objetivo deste trabalho, auxilia no reconhecimento de tendências tecnológicas e será aprofundada na próxima subseção.

2.1.1 Previsão de Tecnologias

A previsão de tecnologias está entre as mais populares tarefas de análise de patentes. Seu objetivo é proporcionar a identificação de oportunidades tecnológicas e gerar percepções valiosas para governos, empresas e tomadores de decisão a fim de garantir amparo ao investimento de capital. (KIM et al., 2019; KRESTEL et al., 2021). Além disso, esta tarefa tem sido reconhecida como uma etapa essencial no processo de PD&I (KIM; BAE, 2017). Nesse contexto, as patentes surgem como um recurso valioso para a previsão e deliberações a cerca de tecnologia, pois fornecem conhecimento atualizado e confiável para a identificação de tendências tecnológicas (ALTUNTAS; DERELI; KUSIAK, 2015).

Lenz (1962), um dos pioneiros da previsão de tecnologias, definiu-a como a previsão de invenções, características ou desempenho de uma máquina servindo a um propósito útil, e apontou que as qualidades buscadas para os métodos de previsão são clareza, expressão quantitativa e reprodutibilidade dos resultados. Para Cho e Daim (2013), a previsão de tecnologias é a capacidade de analisar e avaliar os parâmetros de desempenho de um produto através de declarações de probabilidade com um nível de confiança relativamente alto, capturando oportunidades e ameaças de mudanças tecnológicas a fim de fornecer uma informação valiosa para a tomada de decisão em PD&I.

O termo “previsão de tecnologias” foi cunhado em meados de 1940 e, ao longo da história, alguns grandes acontecimentos marcaram a evolução da tarefa. As primeiras grandes tentativas de prever tendências tecnológicas e novas invenções foram desenvolvidas pela *National Resource Commission* (NRC) em 1935, durante a Grande Depressão nos EUA (HALEEM et al., 2019). Na sequência, o ataque realizado pelos japoneses a Pearl Harbor em dezembro de 1941, fez com que os Estados Unidos organizassem um grupo de cientistas para revisar a pesquisa aeronáutica e fazer recomendações sobre o futuro da Força Aérea através da identificação de oportunidades (CHO; DAIM, 2013).

Após a Segunda Guerra Mundial, a previsão de tecnologia foi impulsionada pela competição militar e espacial com a União Soviética, especificamente após o que ficou conhecido como a “Crise do Sputnik” em 1957, que forçou os EUA a melhorar suas defesas e antecipar a capacidade tecnológica dos inimigos (HALEEM et al., 2019). A partir de 1960 vários métodos de previsão de tecnologia foram desenvolvidos visando reduzir o risco e obter evidências confiáveis para a realização de previsões. Desde então, a tarefa de previsão de tecnologias desenvolve-se como disciplina e ferramenta de planejamento, principalmente por órgãos governamentais e consultorias especializadas (CHO; DAIM, 2013).

Mudanças tecnológicas afetam indivíduos, organizações e nações influenciando a alocação de recursos. A tarefa de previsão serve de apoio ao processo decisório, auxiliando a maximizar o ganho e minimizar a perda em condições futuras (FIRAT; WOON; MADNICK, 2008). Com a rápida e frequente mudança nos mais variados contextos tecnológicos, empresas estão cada vez mais integradas com outras áreas e políticas governamentais. O aumento da complexidade deste ecossistema faz com que os impactos causados através das tecnologias de previsão na construção de estratégias de negócio e inteligência competitiva se sobressaiam diante a precisão de previsões. Desta forma, prever o sucesso de uma tecnologia futura tornou-se um processo chave para os tomadores de decisão (ALTUNTAS; DERELI; KUSIAK, 2015; CHO; DAIM, 2013).

2.2 Predição de Ligações

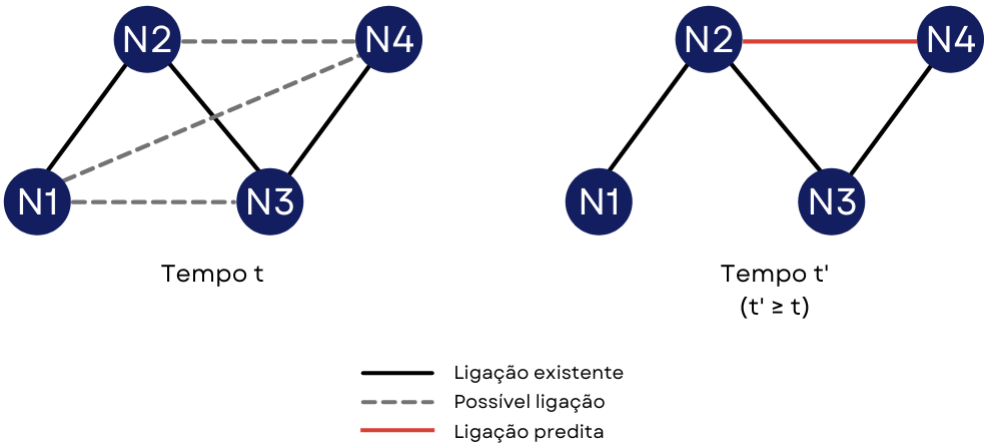
De acordo com Barabási e Pósfai (2016), a teoria de grafos é o pano de fundo matemático histórico da Ciência de Redes (do inglês *Network Science* - NS) moderna, assim como para a predição de ligações. Na literatura científica, os termos rede e grafo são usados indistintamente. A Ciência de Redes utiliza-se da terminologia rede, nó e ligação, enquanto a teoria de grafos emprega os termos grafo, vértice e aresta. No entanto, há uma distinção sutil entre as duas terminologias: a combinação rede, nó e ligação geralmente se refere a sistemas reais, como por exemplo a *World Wide Web* (WWW), que é uma rede de documentos da *web* vinculados por URLs (*Uniform Resource Locator*), assim como a própria sociedade, que é uma rede de indivíduos ligados por laços familiares, de amizade ou profissionais. Em contraste, utiliza-se grafo, vértice e aresta ao se referir a representação matemática destas redes.

Prever interações ausentes e que podem ocorrer no futuro em uma rede em evolução é o objetivo da predição de ligações. Esta tarefa é um dos problemas fundamentais da análise de redes e está profundamente associada a sistemas de recomendação (AMARA; TAIEB; AOUICHA, 2021). As primeiras tentativas de implementar a predição de ligações utilizavam-se heurísticas buscando capturar e explorar algumas poucas informações estruturais da rede, como por exemplo, os algoritmos de número de vizinhos comuns, o coeficiente de similaridade de Jaccard e o índice Adamic-Adar (MALEK et al., 2021). Atualmente, a maioria das técnicas de predição de ligações existentes considera um problema de classificação binária, onde os pares de nós desconectados recebem uma pontuação proporcional à probabilidade de existência de uma ligação entre eles. Estabelece-se um limiar de predição de maneira que todos os pares com uma pontuação acima ao limite arbitrário sejam considerados amostras positivas e todos os pares abaixo do limite são vistos como amostras negativas. Este limiar pode ser especificado pelo usuário ou determinado através de alguma estratégia (MARTÍNEZ; BERZAL; CUBERO, 2016).

Kumar et al. (2020) entendem a predição de ligações da seguinte maneira: dado um grafo não direcionado $G(V, E)$ onde V caracteriza o conjunto de vértices e E o conjunto de arestas, esse grafo possui um conjunto universo U que contém um total de $\frac{n\cdot(n-1)}{2}$ ligações, onde $n = |V|$. As ligações não existentes são determinadas por $|U| - |E|$, sendo que algumas dessas ligações podem surgir em um futuro próximo. Encontrar as ligações faltantes é o objetivo da predição de ligações. Formalmente, Liben-Nowell e Kleinberg (2003) definiram o problema da predição de ligações como: um grafo $G_{t0-t1}(V, E)$ representa um instante da rede durante o intervalo de tempo $[t0, t1]$ e E_{t0-t1}

representa o conjunto de ligações presentes nesse momento. A tarefa de predição de ligações é encontrar o conjunto $E_{t'0-t'1}$ durante um intervalo de tempo $[t'0, t'1]$, onde $[t'0, t'1] \geq [t0, t1]$. A Figura 1 representa graficamente a tarefa, onde, para fins de exemplo, dentre três possíveis ligações apenas uma foi predita como a mais provável de ocorrer.

Figura 1 – Tarefa de predição de ligações



Fonte: Elaborado pelo autor (2022).

Os métodos de predição de ligações existentes podem ser classificados como métodos baseados em similaridade e métodos baseados em aprendizado. Métodos baseados em similaridade assumem que quanto mais similares os nós são, maior é a possibilidade de uma ligação ocorrer entre eles. A partir dessa premissa, calculam a similaridade entre dois nós definindo uma função que pode utilizar informações como a topologia da rede ou os atributos dos nós. Por outro lado, os métodos baseados em aprendizado são capazes de extrair várias características de uma rede para então construir um modelo, treiná-lo com a informação existente e finalmente utilizá-lo para predizer se uma ligação entre dois nós irá ocorrer (WANG et al., 2019). Neste estudo, a predição de ligações recai sobre a classificação de métodos baseados em aprendizado e é abordada como um problema de classificação binária.

A predição de ligações é baseada na evidência empírica de que dois nós (ou duas entidades) são mais propensos a interagirem se forem semelhantes. Tratando-se de redes, a similaridade deve ser entendida como um conceito abstrato e que pode variar de acordo com o domínio. Dessa forma, a compreensão do domínio que a rede representa é crucial para definir a noção de similaridade que será aplicada a ele. Na maioria dos domínios, observa-se que os nós tendem a formar comunidades altamente conectadas. Isso levou a uma definição comum de similaridade como sendo a quantidade de caminhos diretos ou indiretos relevantes entre os nós (MARTÍNEZ; BERZAL; CUBERO, 2016).

Nos últimos anos, novos métodos e técnicas baseados em aprendizado têm sido propostos e podem ser divididos em duas categorias. Os métodos baseados em redes neurais rasas, como por exemplo, *DeepWalk*, *LINE* e *Node2vec*, baseiam-se na computação explicita da similaridade entre os nós. Neste caso, os nós possuem sua representação simplificada, obtida ao expressar cada nó em um espaço de baixa-dimensão (WU et al., 2021). Já os métodos baseados em redes neurais profundas, como *GNN*, *SEAL* e *GraphSAGE*,

demandam grande quantidade de recursos computacionais devido ao massivo aprendizado parametrizado, o que, dependendo do cenário, cria dificuldades para a aplicação desses algoritmos (WANG et al., 2019; WU et al., 2021).

2.3 Redes Neurais Artificiais

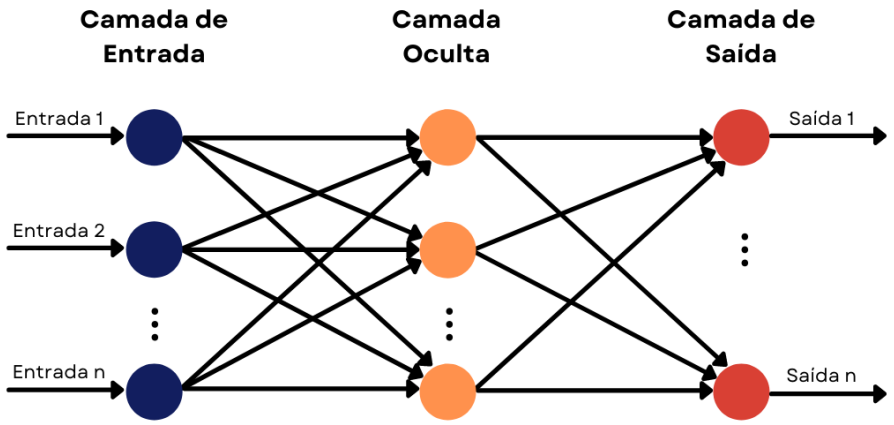
As ANNs surgiram ainda na década de 40, quando o neurofisiologista Warren McCulloch e o matemático Walter Pitts, da Universidade de Illinois, publicaram em 1943 uma pesquisa onde propuseram uma comparação entre a atividade nervosa e a lógica proposicional (YADAV; YADAV; KUMAR, 2015). O trabalho demonstrou que mesmo redes neurais simples poderiam computar funções aritméticas ou lógicas, e instigou outros pesquisadores a estudarem modelos inspirados no funcionamento do cérebro. Em 1949 foi publicado o livro "*The Organization of Behavior*", do biólogo e psicólogo Donald Hebb. Neste livro, Hebb reafirmava a ideia de que o condicionamento psicológico estava presente em todos os animais, pois esta é uma propriedade de neurônios individuais. Hebb propôs uma lei de aprendizado específica para as sinapses dos neurônios, o que preparou o cenário para desenvolvimentos posteriores na área de ANNs (HAENLEIN; KAPLAN, 2019; YADAV; YADAV; KUMAR, 2015).

Entre 1957 e 1958, Frank Rosenblatt propôs o modelo *Perceptron*, considerado a primeira rede neural verdadeira, documentado em sua pesquisa "*The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain*". Baseado nas linhas de pensamento de McCulloch e Pitts, Rosenblatt desenvolveu seu modelo matemático de sinapse humana com pesos continuamente ajustáveis (WU; FENG, 2018). Também ao final da década de 50, Bernard Widrow e alguns alunos da Universidade de *Stanford* desenvolveram um modelo de processamento de redes neurais chamado de ADALINE. O modelo destacava-se pela sua poderosa lei de aprendizado, a Regra Delta. Além dos progressos de Rosenblatt e Widrow, diversos outros avanços ocorreram no final dos anos 50 e meados dos anos 60 no desenvolvimento de arquiteturas de ANNs e conceitos de implementação (YADAV; YADAV; KUMAR, 2015). No entanto, em 1969, Marvin Minsky e Seymour Papert demonstraram em seu livro "*Perceptrons*" que os computadores da época não possuíam poder de processamento suficiente para lidar com o trabalho exigido pelas ANNs, o que praticamente paralisou o interesse na área até 1980 (HAENLEIN; KAPLAN, 2019; WU; FENG, 2018).

Com o aumento do poder computacional, em 1980 ressurgiram diversas propostas para o desenvolvimento de neurocomputadores e aplicações de redes neurais. John Hopfield, físico renomado, escreveu entre 1983 e 1986 dois artigos sobre o tema que instigaram um grande número de pesquisadores qualificados de diversas áreas a explorar o campo (YADAV; YADAV; KUMAR, 2015). As ANNs ganharam popularidade em 1986 com a publicação do livro "*Parallel Distributed Processing: Explorations in the Microstructure of Cognition*", de autoria dos psicólogos David Rumelhart e James McClelland das universidades de *Stanford* e *Carnegie Mellon*. Eles apresentaram um modelo matemático e computacional que propiciava o treinamento supervisionado dos neurônios artificiais: o algoritmo *backpropagation*. Este algoritmo de otimização global sem restrições foi aplicado a muitos problemas de aprendizagem em várias áreas e a ampla divulgação de seus resultados gerou grande entusiasmo (RUSSELL; NORVIG, 2020).

Pode-se definir uma ANN como um modelo computacional inspirado biologicamente na maneira em que o cérebro humano realiza o processamento de uma determinada tarefa. ANNs são capazes de resolver problemas não lineares ou mal definidos baseados em composição paralela e seu funcionamento consiste em: a) elementos de processamento (neurônios); b) conexões entre eles (sinapses); e c) coeficientes vinculados às conexões (pesos). As conexões constituem a estrutura neuronal e anexadas a essa estrutura estão os algoritmos de treinamento e recuperação (TKáč; VERNER, 2016; SHANMUGANATHAN, 2016). O advento das ANNs permitiu que sistemas fossem capazes de interpretar dados externos, aprender com esses dados e utilizar o aprendizado para alcançar um objetivo. Esse modelo foi apontado como alternativa para solucionar tarefas onde se necessita flexibilidade e adaptação (HAENLEIN; KAPLAN, 2019).

Figura 2 – Estrutura básica de uma ANN



Fonte: Elaborado pelo autor (2022).

Em geral, uma ANN simples é constituída de três camadas, entrada, oculta e saída, como ilustra a Figura 2. Todas as camadas são compostas por neurônios de forma a comporem uma rede. Além das camadas, função de ativação, algoritmo de aprendizado e pesos constituem os elementos básicos de uma ANN (DHARWAL; KAUR, 2016; SILVA et al., 2017), sendo:

1. Camada de entrada: responsável por receber dados, sinais, características ou medições do ambiente externo e os enviar para processamento posterior. Deve descrever as informações corretamente, sem redundâncias ou erros, para obter o melhor desempenho para a rede.
2. Camadas ocultas: recebem os dados da camada de entrada e executam a maior parte do processamento interno da rede. São responsáveis por extrair os padrões associados aos dados analisados e podem variar de rede para rede. O número de camadas ocultas depende da natureza e da dimensão do problema e deve passar por experimentos de forma a obter o resultado adequado.
3. Camada de saída: recebe as informações processadas das camadas anteriores e, a partir disso, produz e apresenta os resultados da rede. Esses resultados variam de acordo com o problema e podem ser, por exemplo, classes ou valores contínuos.

4. Função de ativação: seu papel é dimensionar a saída de uma dada camada da rede neural para um intervalo adequado. Existem diversas funções de ativação e cada uma delas produz diferentes impactos sobre o comportamento da rede neural.

A capacidade de aprender como um sistema se comporta através de um conjunto de amostras é, sem dúvidas, uma das características mais relevantes das redes neurais. Essa capacidade de aprendizado é derivada a partir de três abordagens: i) aprendizado supervisionado, onde cada amostra de treinamento é composta por sinais de entrada e suas respectivas saídas, que representam o processo e seu comportamento; ii) não supervisionado, que não requer conhecimento da saída esperada e ajusta os pesos e limites sinápticos da rede para refletir subconjuntos de amostras semelhantes; e iii) aprendizado por reforço, um processo de tentativa e erro onde se calcula a saída para uma determinada entrada e, caso seja satisfatória, recompensa-se a conexão através do incremento dos pesos das sinapses, do contrário pune-se a conexão (SHANMUGANATHAN, 2016; SILVA et al., 2017). Quanto à arquitetura das ANNs, as duas mais comuns são as redes neurais *feedforward*, onde as informações são transmitidas apenas na direção da camada de entrada para a saída, e as redes neurais *feedback*, que permitem que os sinais viajem em ambas as direções, introduzindo laços na rede (ABIODUN et al., 2018; DHARWAL; KAUR, 2016).

As soluções de redes neurais artificiais são aplicáveis a incontáveis problemas do mundo real em um extenso conjunto de áreas, que englobam de negócios e engenharia até agricultura e educação. Alguns exemplos de sua aplicação são o processamento de imagens, a análise de tumores cerebrais por ressonância magnética e sensores de temperatura e umidade do solo (DHARWAL; KAUR, 2016). As ANNs são excelentes ferramentas para identificar tendências e padrões em dados, o que é propício para suprir as necessidades das tarefas de previsão e predição. Neste âmbito, ANNs têm sido utilizadas na previsão do sucesso ou fracasso de negócios, estimativas do mercado de ações e também na previsão de mudanças climáticas e do tempo (ABIODUN et al., 2018). Tkáč e Verner (2016) afirmam que características das ANNs, como adaptabilidade e robustez, as tornam um instrumento valioso para o suporte à decisão, cujo sucesso pode ser observado por meio do número crescente de publicações e aplicações.

2.3.1 Node2vec

Uma representação concisa que permita que as tarefas baseadas em aprendizado de máquina, como a predição de ligações, possam ser aplicadas de forma eficiente em termos de complexidade de espaço e tempo se faz necessária (AMARA; TAIEB; AOUICHA, 2021). Frente a esse propósito, o aprendizado de representação de redes tem gerado grande interesse de pesquisa. Seu objetivo é aprender representações latentes e de baixa dimensão dos nós da rede, de maneira a preservar a topologia da rede, capturar a diversidade de padrões de conectividade e outras informações secundárias (ZHANG et al., 2020).

De modo geral, podem-se dividir os algoritmos de aprendizado de representação de rede em duas categorias: métodos baseados em redes neurais rasas, onde a função de mapeamento dos nós para a representação de vetores é um mapeamento chave-valor, e métodos baseados em redes neurais profundas, que usam codificadores complexos (MALEK et al., 2021; WANG et al., 2019). As abordagens profundas geralmente alcançam resultados mais precisos na maioria dos problemas de aprendizado, principalmente em tarefas que lidam com representações mais complexas como processamento de imagens

e de linguagem natural. Isso se deve ao fato de que esses algoritmos são capazes de tirar vantagem de uma grande quantidade de dados de treinamento, pois as camadas ocultas adicionais combinam e transformam os recursos extensivamente, permitindo a modelagem de dados complexos com maior eficácia do que uma rede rasa (AL-ASWADI; CHAN; GAN, 2020; WINKLER; LE, 2017).

No entanto, a alta precisão oferecida pela arquitetura de redes neurais profundas conduz a um custo de escalabilidade. Isto se deve aos altos requisitos de tempo de treinamento e da necessidade de recursos computacionais substanciais por conta do aprendizado massivo de parâmetros (GROVER; LESKOVEC, 2016; WU et al., 2021). Em contraposição, os métodos baseados em redes neurais rasas são computacionalmente mais eficientes e, em particular, para a tarefa de predição de ligações produzem resultados comparáveis aos métodos profundos (MALEK et al., 2021; WINKLER; LE, 2017). Ao considerar a semelhança de poder preditivo de ambos os métodos e a utilização de recursos computacionais, este trabalho utiliza o Node2vec, um proeminente algoritmo de aprendizado de representação de rede baseado em redes neurais rasas.

Segundo Grover e Leskovec (2016), para resolver problemas de predição em redes, é necessário construir uma representação vetorial de recursos para os nós e ligações e, para este propósito, foi desenvolvido o Node2vec. O Node2vec é um algoritmo semissupervisionado para aprendizado de recursos escalável em redes. Ele define uma noção flexível de vizinhança de um nó e explora estas vizinhanças eficientemente através de um percurso aleatório tendencioso de segunda ordem. A flexibilidade na exploração das vizinhanças é considerada a chave para aprender representações mais ricas. O algoritmo estende a representação de recursos de nós individuais para ligações através da composição simples dos recursos dos nós presentes no par, permitindo que tarefas de previsão envolvendo tanto nós quanto ligações sejam realizadas.

O aprendizado do Node2vec foi inspirado no modelo de rede neural rasa *Skip-Gram*, originalmente aplicado à tarefa de processamento de linguagem natural. Seu modelo, *Skip-Gram with Negative Sampling* (SGNS), estabelece uma analogia que representa uma rede como uma espécie de "documento". Assim como um documento pode ser interpretado como uma sequência ordenada de palavras, o SGNS entende a rede como uma sequência ordenada de nós (GROVER; LESKOVEC, 2016). Além disso, o SGNS implementa o uso de amostras negativas para aumentar a velocidade do processo de treinamento e evitar o aprendizado de pesos triviais (AGGARWAL, 2018; PENG et al., 2020). O Node2vec utiliza o conceito de percurso aleatório tendencioso para aprender os vetores de representação dos nós através do modelo SGNS, o que melhora o desempenho da tarefa de análise de rede (WANG et al., 2019).

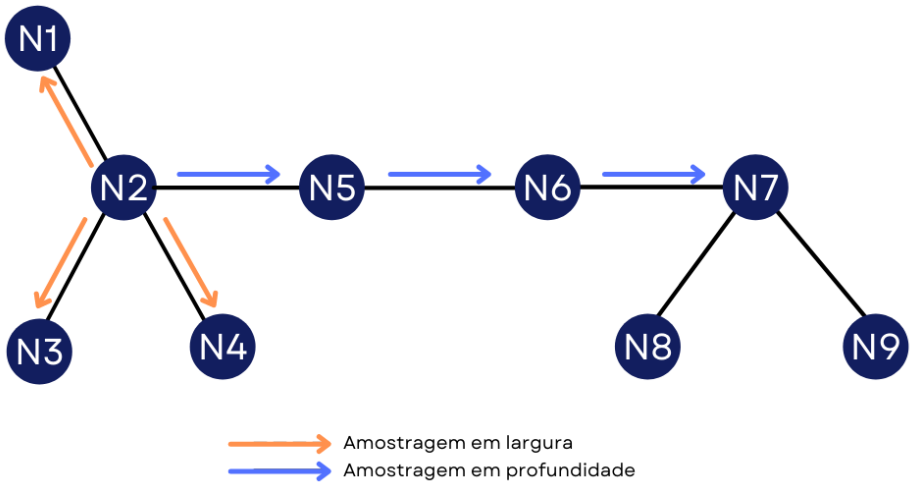
Contudo, existem diversas estratégias de amostragem da vizinhança de nós em uma rede, o que resulta em diferentes aprendizados de representação do recurso. As estratégias utilizadas pelo Node2vec levam em consideração os objetivos das tarefas de predição e a organização dos nós da rede, sendo baseadas em dois conceitos de similaridade, a homofilia e a equivalência estrutural. A hipótese de homofilia trata como similares nós altamente interconectados e que pertencem a comunidades de rede semelhantes (GROVER; LESKOVEC, 2016). Em contraposição, a equivalência estrutural considera semelhantes os nós que desempenham papéis estruturais similares, como *hubs* ou pontes de suas respectivas comunidades. Redes do mundo real normalmente apresentam ambos os tipos de

equivalência e, por esse motivo, as estratégias de amostragem da vizinhança implementadas no Node2vec buscam alternar entre elas (KUMAR et al., 2020).

A Figura 3 ilustra as estratégias de amostragem de vizinhança de nós do Node2vec tendo como ponto de partida o nó "N2". De acordo com Grover e Leskovec (2016) e Zhang et al. (2020), podemos definir estas estratégias da seguinte forma:

1. Amostragem em largura: a vizinhança é restrita aos nós que são vizinhos imediatos da fonte. Corresponde à equivalência estrutural e obtém uma visão microscópica da vizinhança de cada nó.
2. Amostragem em profundidade: a vizinhança é composta por nós amostrados sequencialmente a distâncias crescentes do nó fonte. Equivale a hipótese de homofilia e reflete com mais precisão uma visão macro da vizinhança. Uma importante observação é que essa estratégia infere a respeito das dependências entre nós. Portanto, deve-se atentar ao tamanho da amostra pois nós mais distantes da origem podem ser potencialmente menos representativo e impactar no desempenho do aprendizado.

Figura 3 – Estratégias de amostragem de vizinhança utilizadas pelo Node2vec



Fonte: Elaborado pelo autor (2022).

O algoritmo do Node2vec inicia a partir do cálculo das probabilidades de transição entre os nós da rede. Estas probabilidades são guiadas por dois parâmetros, um de retorno e outro de entrada e saída, que permitem que o algoritmo alterne entre as estratégias de amostragem de vizinhança. O parâmetro de retorno indica a probabilidade de um nó ser revisitado durante o percurso, já o parâmetro de entrada e saída indica a probabilidade do percurso passar por nós ainda não visitados. Em seguida, os percursos aleatórios enviesados pelos parâmetros citados são executados, construindo as sequências de nós que alimentam o SGNS. Por fim, o SGNS produz as representações vetoriais de cada um dos nós e ligações presentes na rede (AMARA; TAIEB; AOUICHA, 2021; GROVER; LESKOVEC, 2016).

3 Trabalhos Correlatos

Através de buscas na literatura científica foram selecionados trabalhos a fim de identificar métodos e abordagens de análise de patentes utilizadas na previsão de tecnologias que empregam técnicas de predição e identificação de tendências. A pesquisa foi realizada nas bases de artigos acadêmicos ACM Digital Library®, ScienceDirect®, Scopus® e Web of Science®. Foram considerados artigos científicos em língua inglesa publicados entre 2017 e 2021 que continham combinações dos termos: "*Patent**", "*Link Prediction*", "*Forecasting or Prediction*", "*Emerging Technolog**", onde "*" representa as possíveis variações dos termos.

Tabela 1 – Resultados da revisão de literatura

Base de Artigos Acadêmicos	Quantidade de artigos resultantes
ACM Digital Library®	4
ScienceDirect®	52
Scopus®	123
Web of Science®	5
Total de resultados	184

Fonte: Elaborado pelo autor (2022).

A Tabela 1 apresenta a síntese dos resultados obtidos pela expressão de busca. Do total de 184 trabalhos resultantes, 30 apresentaram título adequado ao objetivo da busca. Seus resumos foram então lidos na íntegra e 13 artigos foram selecionados para leitura da introdução. Por fim, os 8 trabalhos que possuem relação com o tema desta pesquisa e que passaram por todos os critérios de exclusão são descritos a seguir.

O trabalho de Choi e Song (2018) propõe uma abordagem baseada em patentes para explorar tendências tecnológicas no domínio da logística. Utilizando patentes relacionadas ao setor disponíveis no USPTO®, foi aplicada a técnica de modelagem de tópicos com uso de *Latent Dirichlet Allocation* (LDA). O LDA é um método estatístico o qual assume que um documento é composto por vários tópicos e estes, por sua vez, são compostos por conjuntos de palavras que frequentemente ocorrem juntas. Os tópicos identificados foram então investigados em relação às tendências na atividade de patenteamento e classificados em quatro grupos (tópicos dominantes, emergentes, saturados e em declínio) de forma a auxiliar o entendimento do cenário logístico em níveis tecnológico e empresarial.

Kim et al. (2019) realizaram um estudo de caso na área de transferência de energia sem fio voltado à identificação de tecnologias emergentes e lacunas tecnológicas. Os autores conduzem uma análise de patentes utilizando a extração de tópicos via mineração de texto e LDA, de maneira a agrupar tópicos com semântica similar e, a partir disso, formar agrupamentos. Com o intuito de viabilizar a identificação de lacunas no desenvolvimento de tecnologias, uma etapa que correlaciona os agrupamentos resultantes com a análise de séries temporais e o ciclo de inovação de tecnologias foi aplicada. Essa etapa, além de proporcionar a identificação de lacunas tecnológicas, teve como objetivo refinar o método proposto e melhorar a precisão e validação do processo de identificação de tecnologias. Como resultados da pesquisa, a análise revelou uma área de tecnologia emergente e duas lacunas tecnológicas existentes no contexto de transferência de energia sem fio.

A pesquisa de Zhou et al. (2020) aplicou aprendizado profundo para prever tecnologias emergentes com base em dados de patentes. Primeiramente, construiu-se o conjunto de dados de amostra utilizando o ciclo de expectativa de Gartner e indicadores de patentes. Em seguida, para contornar a grande quantidade de dados necessários para o conjunto de treinamento do aprendizado profundo, os autores utilizaram uma rede adversária generativa (do inglês *Generative Adversarial Network* - GAN) para gerar amostras sintéticas e aumentar a quantidade de dados do conjunto. Por fim, um classificador binário baseado em redes neurais profundas foi treinado com o conjunto de dados com o objetivo de prever tecnologias emergentes. O método proposto foi capaz de prever mais de 77% das tecnologias emergentes em um determinado ano com alta precisão, mesmo em cenários em que as amostras de dados eram limitadas. A validade e eficácia da abordagem foi verificada através da previsão de tecnologias emergentes do ciclo de expectativa de Gartner de 2017 com base em dados de patentes de 2000 a 2016, onde quatro em cada seis tecnologias foram previstas corretamente.

Com o objetivo de descobrir oportunidades de diversificação de negócios em empresas, Jeong et al. (2021) propuseram uma estrutura baseada em marcas registradas. O trabalho enfatiza que estudos na área de identificação de oportunidades focam majoritariamente em uma perspectiva tecnológica e não de negócio, e apontam como limitação fundamental dessa perspectiva a incapacidade de ser aplicada em negócios oferecidos como serviços que são complexos de se patentear. A estrutura proposta compreendeu três fases: i) construção de um conjunto de dados de treinamento e validação que usou rede de coocorrência de marcas registradas e construção de um modelo de predição de ligações; ii) previsão de diversificação de negócios e construção de um portfólio de negócios expansível para a empresa alvo; e iii) análise de inteligência competitiva para estabelecer as estratégias de diversificação. O modelo de predição de ligações foi capaz de aprender a dinâmica do negócio e de identificar oportunidades de diversificação com base em dados objetivos. A precisão da validação do modelo construído foi de 81,87%.

Kim e Bae (2017) abordaram o tema de previsão de tecnologias promissoras e propuseram uma abordagem que fez uso da análise de patentes. Utilizaram a classificação cooperativa de patentes (do inglês *Cooperative Patent Classification* - CPC) e o algoritmo *K-means* com implementação baseada em distância euclidiana para separar os dados em três agrupamentos contendo documentos de patentes de acordo com as características semelhantes. Na sequência, foi realizada a avaliação de quão promissores eram os agrupamentos. A eleição do agrupamento com tecnologias mais promissoras foi realizada por meio da análise de três indicadores numéricos de patentes, citações futuras, famílias de patentes e reivindicações independentes. Para verificar a metodologia proposta, foi efetuado um experimento na área da indústria de cuidados com o bem-estar a partir de dados de patentes coletados do USPTO® de 2002 a 2014, onde o agrupamento com tecnologias relacionadas à telemedicina foi apontado como promissor.

Lee et al. (2018) propuseram um método baseado em aprendizado de máquina para a identificação antecipada de tecnologias emergentes. O método consiste em extrair indicadores de patentes do USPTO® e aplicar uma rede neural multicamadas *feed-forward* para capturar as relações complexas entre os indicadores no período de tempo de interesse. A rede neural foi utilizada para classificar patentes de acordo com o número previsto de citações futuras nos próximos três, cinco e dez anos imediatos à emissão da patente.

Um estudo de caso sobre o tópico "tecnologia farmacêutica" foi apresentado e de acordo com os autores a conclusão do trabalho foi que, devido aos valores de precisão obtidos serem maiores do que a medida de revocação, o método proposto forneceu resultados conservadores.

Baseados no conjunto de tecnologias utilizadas por determinada empresa, Lee et al. (2021) sugeriram uma abordagem de descoberta de oportunidades de tecnologias com o emprego de predição de ligações sobre redes de coocorência baseadas na classificação *File forming Term* (F-term) de patentes. Foram geradas duas redes de coocorrências F-Term: uma com um conjunto universal de patentes e outra centrada na empresa. A predição de ligações foi aplicada às redes sobrepostas para identificar as tecnologias oportunas e então os resultados foram avaliados através de um mapa visual com índices de impacto da tecnologia. Como resultado da predição de ligações, as 100 ligações mais prováveis foram assumidas como tecnologias oportunas para a pesquisa. Um estudo de caso foi conduzido em um fabricante japonês de componentes de ciclismo para demonstrar a validade da abordagem, onde entre as 100 oportunidades descobertas, 39 corresponderam aos resultados reais de P&D da empresa alvo.

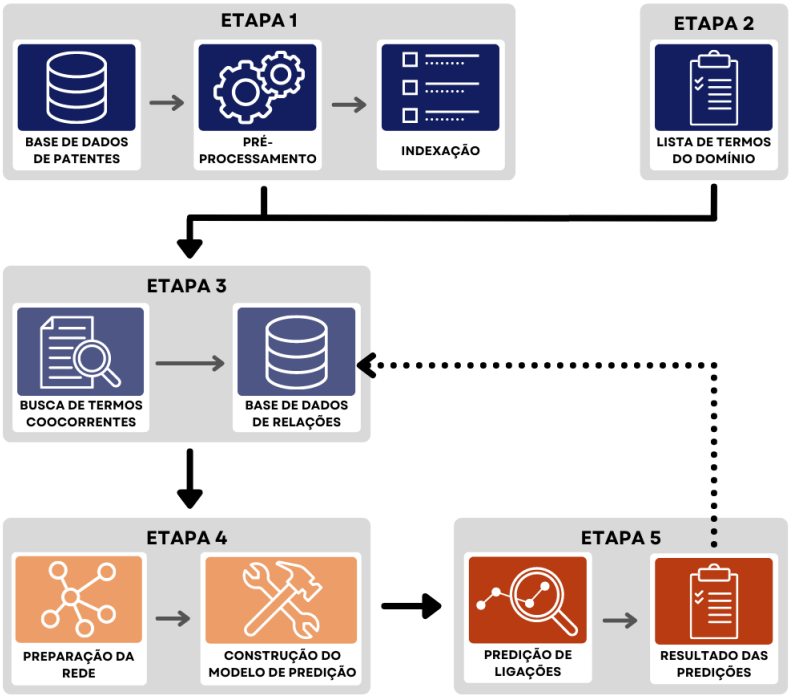
Por fim, o estudo realizado por Park e Yoon (2018) apresentou uma proposta para a descoberta de oportunidades tecnológicas voltadas à convergência de tecnologias utilizando informações de patentes. Mediante a técnica de predição de ligações em uma rede direcionada que adaptou o conceito de acoplamento bibliográfico e centralidade entre arestas, as potenciais oportunidades tecnológicas foram preditas. A técnica de predição de ligações foi proposta com o intuito de refletir o contexto de convergência e fluxo de conhecimento de tecnologia à pesquisa. A fim de ilustrar o método proposto, oportunidades tecnológicas entre as áreas de biotecnologia e tecnologia da informação foram investigadas. Foram utilizadas patentes coletadas do banco de dados *online* do USPTO® entre os anos de 2006 e 2015. Entre os 96 pares de termos de patentes previstos, 86 pares ocorreram no futuro, resultando em uma precisão de 89%. O valor de revocação obtido foi de 96%.

As pesquisas apresentadas refletem o atual cenário de competitividade em que as empresas estão expostas no mercado e também evidenciam a demanda por ampliar o potencial de geração de inovações. Através da análise dos trabalhos presentes na literatura, constatou-se que três principais métodos foram incorporados às abordagens a fim de realizar a previsão de tecnologias, a saber: modelagem de tópicos, agrupamento de documentos e redes neurais. As abordagens que fizeram uso dos métodos de modelagem de tópicos e agrupamento de documentos necessitaram de outras formas de análise para auxiliar a obtenção dos resultados, característica que, a depender do cenário estudado, pode apresentar um resultado que não corresponde à realidade. Já os trabalhos que aplicaram métodos baseados em redes neurais foram capazes de fornecer resultados coerentes independente do cenário, uma vez que a técnica é capaz de aprender e se adaptar ao contexto. Além disso, os métodos baseados em redes neurais forneceram resultados expressivos e validações mais extensivas.

4 Método Proposto

Este trabalho apresenta um método de análise de patentes voltado à previsão de tecnologias para apoio a tomada de decisões estratégicas em ambientes organizacionais. Para alcançar este objetivo, utiliza técnicas de aprendizado de representação de rede e predição de ligações no contexto de análise de patentes. A Figura 4 fornece uma visão geral do método e das cinco etapas que o compõe: 1) Coleta, pré-processamento e indexação; 2) Definição do domínio e termos de interesse; 3) Formação da base de relações; 4) Construção do modelo de predição; e 5) Predição e apresentação dos resultados. As primeiras duas etapas são responsáveis pela coleta e preparação dos documentos de patente de forma que possam ser consultados pelos termos do domínio definido. Na terceira etapa são realizadas as buscas de termos coocorrentes que definem a tecnologia e a geração do banco de dados de relações. A quarta etapa consiste em preparar as representações da rede e os conjuntos de dados para, a partir destas informações, construir o modelo de predição. Por fim, na quinta etapa é realizada a predição de ligações e a apresentação dos resultados na forma de pares de termos com suas probabilidades. A seguir cada uma das etapas é descrita em detalhes.

Figura 4 – Processo geral do método proposto



Fonte: Elaborado pelo autor (2022).

4.1 Etapa 1: Coleta, Pré-Processamento e Indexação

A primeira etapa do método proposto está relacionada à coleta e a indexação dos documentos de patentes. Cada documento de patente é pré-processado, por exemplo, retirando-se as pontuações e levando-se em conta os campos mais relevantes, como título, o resumo e o ano. Após, as patentes são indexadas em um banco de dados orientado a

documentos. Este tipo de armazenamento facilita a manipulação e consulta dos dados e, pela flexibilidade de estrutura, permite a evolução dos dados conforme a necessidade de desenvolvimento. Como resultado desta etapa, tem-se uma base de patentes indexadas onde é possível realizar buscas por palavras-chave.

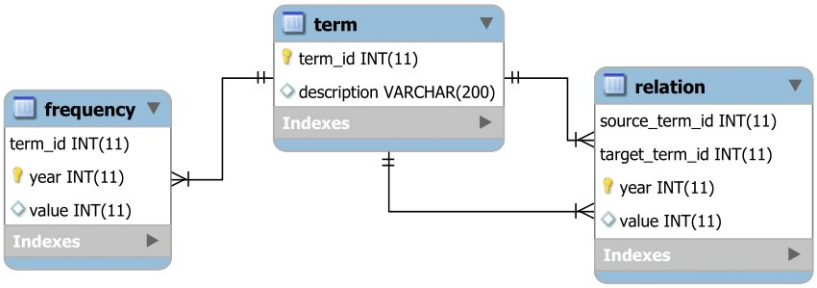
4.2 Etapa 2: Definição do Domínio e Termos de Interesse

Para a segunda etapa, inicialmente deve-se eleger um domínio de interesse. Este domínio é o elemento chave desta etapa, pois a partir dele é formado todo o cenário de tecnologias e conceitos que farão parte de determinada análise. Em seguida, fontes de dados que apresentem termos intrinsecamente relacionados ao tema de interesse são localizadas. Os termos capazes de descrever o domínio são então coletados e integrados de forma a produzir uma lista que servirá de base para o estudo alvo. Esta etapa ocorre de maneira independente da primeira.

4.3 Etapa 3: Formação da Base de Relações

A terceira etapa é responsável por unir os resultados das duas primeiras com o objetivo de formar a base de dados de relações. Para tal, são realizadas consultas na base de patentes indexadas a partir dos termos da lista obtida na segunda etapa. Primeiramente, geram-se todas as possibilidades distintas de relações entre os termos. Então, efetuam-se as buscas de modo que as frequências de menções conjuntas entre eles sejam armazenadas ano a ano. Por fim, armazenam-se a quantidade de coocorrências, o ano e os termos coocorrentes no formato origem-destino em uma base de dados relacional. Devido a natureza da tarefa de previsão de tecnologias e predição de ligações, neste trabalho, uma rede homogênea é constituída. Como consultas mais especializadas sobre nós e ligações não são efetuadas, uma base de dados relacional atende as necessidades do estudo e foi empregada nesta etapa (Figura 5).

Figura 5 – Modelo lógico da base de dados de relações de termos



Fonte: Elaborado pelo autor (2022).

4.4 Etapa 4: Construção do Modelo de Predição

Na sequência, a etapa quatro prepara a rede e constrói o modelo de predição de ligações. Inicialmente, utiliza-se a base de dados de relações para gerar a rede propriamente dita. A seguir, forma-se um conjunto de amostras positivas e negativas que alimentarão o

modelo de predição. Neste ponto, algumas ligações são removidas da rede para compor as amostras positivas, permitindo que o modelo seja capaz de aprender como ocorreram estas ligações verdadeiras. Isto é feito através da remoção aleatória de ligações e da verificação de que o número de componentes conectados da rede se mantém o mesmo. Em contrapartida, as amostras negativas são compostas por todas as ligações inexistentes da rede, obtidas através de uma matriz de adjacência. O aprendizado de representação de rede é então aplicado e, por último, o modelo de predição é gerado a partir das representações das ligações e do conjunto de amostras.

4.5 Etapa 5: Predição e Apresentação dos Resultados

Por fim, na quinta etapa as predições de ligações são realizadas através do modelo construído. Define-se uma probabilidade mínima aceitável para considerar uma predição como verdadeira. Como resultado, obtém-se uma lista de tecnologias ou conceitos atrelados ao domínio escolhido que provavelmente irão se conectar em um futuro próximo. Utiliza-se então esta lista para verificar se as predições ocorreram à medida que novas patentes são indexadas e novas relações são incluídas na base de relações.

5 Resultados Experimentais

5.1 Apresentação do Cenário de Estudo

Para demonstrar a viabilidade do método proposto, um estudo de caso foi realizado. O conjunto de dados foi constituído a partir de patentes obtidas do *United States Patent and Trademark Office*® (USPTO) entre os anos 2006 a 2015 publicados no estudo de Li et al. (2018). Este conjunto de dados, nomeado como USPTO-2M®, é utilizado em testes e validações de tarefas que envolvem patentes, em especial a tarefa de classificação, e conta com 2.000.147 documentos. O USPTO-2M® é composto por patentes de utilidade, isto é, patentes que cobrem a criação de um produto, processo ou máquina novo ou melhorado, indo de encontro ao objetivo deste trabalho.

Neste estudo de caso, o tema foi definido a partir do termo “carbono”, pois este tem sido alvo de um grande volume de pesquisas científicas. Uma breve consulta no mecanismo de busca de publicações acadêmicas Semantic Scholar® revelou que, nos últimos dez anos, aproximadamente dois milhões de resultados apresentam o termo. Para a avaliação do método proposto levou-se em conta predições em cenários de curto, médio e longo prazo, considerando respectivamente períodos de um, três e cinco anos. Assim, obteve-se uma avaliação de desempenho do método proposto com base em dados do mundo real.

5.2 Instanciação do Método Proposto

Esta seção objetiva detalhar o método proposto apresentando os componentes tecnológicos utilizados e como estes se interconectam nas etapas de maneira que ao final seja possível realizar a predição de ligações.

5.2.1 Etapa 1

A fim de constituir a base de patentes do estudo de caso coletou-se manualmente o conjunto de dados USPTO-2M® composto por patentes de 2006 a 2015. Os conjuntos estão disponíveis *online*, em formato JSON (do inglês *JavaScript Object Notation*) e separados por ano. Iniciou-se então o pré-processamento das patentes, sendo utilizados os campos "título" e "resumo" devido sua importância para o estudo. Além disso, adicionou-se o campo "ano", que não consta na estrutura JSON dos documentos. Este campo pode ser obtido a partir do próprio nome do arquivo disponibilizado. O ano é uma característica fundamental, pois permite a aplicação de filtros em passos futuros de avaliação do método. Para cada texto foram realizadas operações simples de pré-processamento, removendo as pontuações. A partir disso, cada documento foi indexado no banco de dados Apache Solr®. Como resultado, obteve-se uma base de patentes indexada onde é possível realizar buscas em texto completo por palavras-chave.

5.2.2 Etapa 2

Como apontado na apresentação do cenário de estudo, "carbono" foi o domínio escolhido devido a sua notoriedade em pesquisa recentes, facilmente identificada ao observar-se o volume de publicações acadêmicas que incluem o termo. Após a definição do domínio, diversos glossários foram consultados de maneira *online* com o intuito de formar uma lista com cerca de 500 termos. Destes, 150 termos que mais coocorreram entre os glossários foram considerados, sendo 2 termos removidos por serem muito genéricos, o que poderia enviesar os resultados. Os 148 termos restantes foram selecionados e integrados de forma a compor a lista do domínio. A Tabela 2 apresenta uma amostra da lista que serviu de base para o estudo alvo. A lista completa de termos selecionados pode ser consultado no Apêndice I.

Tabela 2 – Amostra da lista de termos que representa o domínio de "carbono"

Termo
Carbon Footprint
Net Zero Carbon Business
Biofuel
Fugitive Emissions
Carbon Nanotube
Global Warming
Recycling
Tipping Point
Carbon Fiber
Circular Economy
Biocapacity

Fonte: Elaborado pelo autor (2022).

5.2.3 Etapa 3

A partir da lista de termos do domínio, todas as possíveis combinações distintas de relações entre os termos foram geradas. Em seguida, através da plataforma Apache Solr®, realizou-se a busca dos termos coocorrentes, ou seja, de pares de termos. Ao todo, foram geradas 3719 coocorrências entre 79 termos da lista do domínio de "carbono" ao se considerar todo o período do estudo de caso. A partir dos resultados das consultas foi gerada a base de relações suportada por um sistema gerenciador de banco de dados MySQL®. Foram armazenadas as frequências de menções individuais de termos e os termos coocorrentes no formato origem-destino, ano a ano, para todo o intervalo de 2006 a 2015. O modelo de dados foi descrito na subseção 4.3.

5.2.4 Etapa 4

Nesta etapa, primeiramente definiu-se o período de treinamento do modelo de predição de forma que a avaliação pudesse ocorrer. O intervalo de 2006 a 2010 foi selecionado com o objetivo de que todos os cenários citados na subseção 5.1 pudessem ser avaliados em relação ao ano em que o modelo de predição foi treinado. A partir do ano determinado, por exemplo, 2006 (Figura 6), a respectiva rede foi construída através da biblioteca NetworkX® da linguagem de programação Python®.

Formou-se então, a partir da rede em questão, o conjunto de amostras positivas e negativas. A Tabela 3 apresenta exemplos destas amostras coletadas da rede de 2006, onde o valor "0" na coluna ligações simboliza a ausência da ligação entre os nós origem-destino e "1" a presença. Para realizar o aprendizado de representação da rede, o algoritmo Node2vec® foi aplicado e as representações vetoriais das ligações foram obtidas. Em seguida, separou-se o conjunto das representações das ligações e o conjunto de amostras em dois pares representação-alvo: um par para o teste, com 30% dos dados, e outro para treinamento, com o restante dos dados. Por fim, o modelo de predição para o ano determinado foi gerado com base na regressão logística por meio da biblioteca de aprendizado de máquina de código aberto do Scikit-learn®.

Tabela 3 – Exemplos de amostras negativas e positivas

Tipo da amostra	**2006**		
	Nó origem	**Nó destino**	**Ligação**
Negativa	fugitive emissions	carbon fiber	0
	carbon black	bioprocessing	0
	biofuel	recycling	0
Positiva	carbon composite	carbon material	1
	recycling	biogas	1
	biomass	atmosphere	1

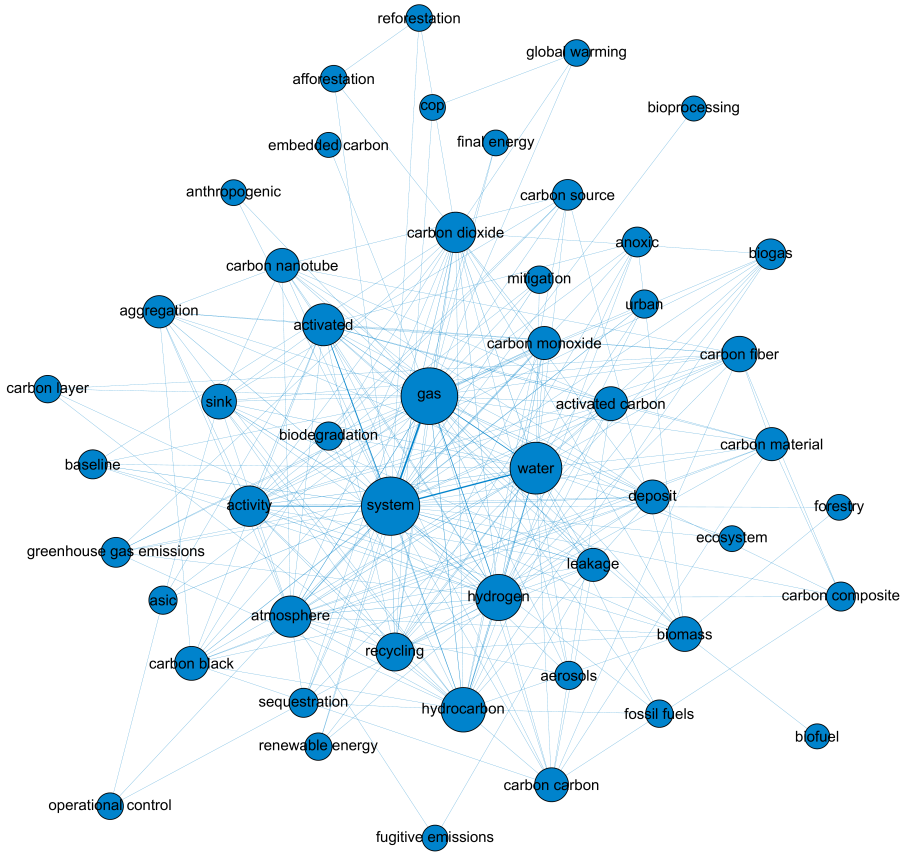
Fonte: Elaborado pelo autor (2022).

5.2.5 Etapa 5

Por fim, na quinta etapa, as predições através do modelo treinado no ano em questão sobre todas as ligações inexistentes de determinada rede foram calculadas obtendo a lista de termos (tecnologias e conceitos) prováveis de se conectarem nos anos seguintes.

A avaliação do método proposto ocorreu da seguinte maneira: i) a partir de cada modelo treinado no intervalo de 2006 a 2010, 100 execuções foram realizadas para três limiares de predição, sendo 50%, 70% e 85% de probabilidade de ocorrer a ligação em três cenários temporais, curto (um ano), médio (três anos) e longo (cinco anos); ii) obteve-se a média das métricas de avaliação do modelo e das predições em relação às 100 execuções, a fim de garantir uma maior fidelidade dos resultados.

Figura 6 – Rede de termos relacionados ao domínio “carbono” no ano de 2006



Fonte: Elaborado pelo autor (2022).

5.3 Análise e Discussão dos Resultados

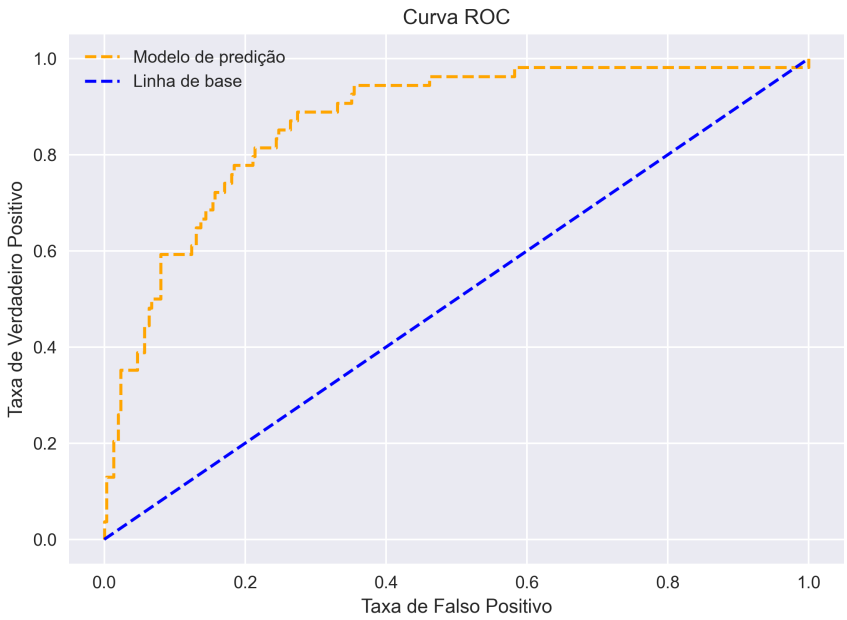
Esta seção foi elaborada com o objetivo de avaliar os resultados obtidos através do método proposto aplicado ao cenário do estudo de caso apresentado na subseção 5.1. Discute-se sobre o desempenho do modelo de predição e a viabilidade de aplicação do método proposto.

A métrica empregada para avaliar o modelo de predição foi a Área Sob a Curva de Característica de Operação do Receptor (do inglês *Area Under the Receiver Operating Characteristic Curve* - ROC-AUC). Esta métrica é amplamente utilizada como indicador de desempenho para problemas de predição de ligações e de classificação binária (LEE et al., 2021). A métrica ROC-AUC calcula a área sob a curva ROC, uma curva de probabilidade que apresenta a taxa de verdadeiros positivos contra a taxa de falsos positivos em diferentes

limiares de classificação. De modo geral, simplifica a análise da curva ROC ao agregar todos seus limiares e resumí-la a um único valor. Basicamente, quanto mais próximo de 1 a ROC-AUC, melhor o desempenho do classificador ao distinguir entre as ligações que ocorreram e as que não.

A Figura 7 exibe a curva ROC do modelo treinado para o ano de 2006, onde se pode observar que o modelo de predição se comporta da maneira desejada, com baixa taxa de falsos positivos, ou seja, ligações preditas que não ocorreram e boa taxa de verdadeiros positivos, ligações preditas que de fato ocorreram. Este diagnóstico é auxiliado pela linha de base, que representa um classificador incapaz de distinguir entre as classes. Ao considerar a ROC-AUC de todos os modelos treinados no intervalo de 2006 a 2010 (Tabela 4) o modelo de predição foi capaz de atingir um desempenho médio de 0,91, o que demonstra que o classificador é altamente capaz de distinguir entre as amostras positivas e negativas de ligações do conjunto de teste.

Figura 7 – Curva ROC do modelo de predição treinado no ano de 2006



Fonte: Elaborado pelo autor (2022).

Tabela 4 – ROC-AUC obtido por cada modelo de predição treinado

ROC-AUC				
2006	2007	2008	2009	2010
0,89	0,91	0,91	0,93	0,91

Fonte: Elaborado pelo autor (2022).

Após a avaliação inicial do modelo de predição, iniciou-se a avaliação do método proposto a partir do cenário do estudo de caso com o objetivo de determinar seu desempenho real. Para isto, os passos descritos na subseção 5.2.5 foram considerados e os resultados sumarizados na Tabela 5 e Tabela 6.

Com o intuito de facilitar a visualização dos dados e acompanhamento das análises, a Tabela 5 categoriza a porcentagem de acerto das predições realizadas em: i) ruim, [0%, 33%), destacada na cor vermelha; ii) satisfatória, [33%, 66%), destacada em amarelo; e iii) boa, [66%, 100%], destacada em verde. As tabelas dividem-se ainda em três seções, representando o total de predições em cada ano considerando limiares maiores ou iguais a 50%, 70% ou 85%. Ou seja, o limiar de 50% indica que somente as predições de ligações em que a probabilidade for maior ou igual a 0,5 devem ser consideradas para serem avaliadas nos cenários de curto, médio e longo prazo. A quantidade de predições é indicada na coluna "Quantidade média de predições", visto que foram realizadas 100 execuções. As taxas percentuais de acertos aparecem nas colunas "Porcentagem de acerto".

Tabela 5 – Métricas obtidas a partir das predições realizadas no intervalo de 2006 a 2010

2006						
Cenário (prazo)	**Limiar de Predição**					
	50%		**70%**		**85%**	
	Quantidade média de predições	Porcentagem de acerto	Quantidade média de predições	Porcentagem de acerto	Quantidade média de predições	Porcentagem de acerto
Curto	173	16,71%	51	22,92%	11	25,26%
Médio	172	36,76%	51	52,60%	12	68,43%
Longo	173	50,27%	51	65,58%	11	78,06%
2007						
Cenário (prazo)	**Limiar de Predição**					
	50%		**70%**		**85%**	
	Quantidade média de predições	Porcentagem de acerto	Quantidade média de predições	Porcentagem de acerto	Quantidade média de predições	Porcentagem de acerto
Curto	189	11,43%	48	19,50%	6	44,37%
Médio	188	32,24%	47	44,74%	6	57,26%
Longo	188	40,67%	48	55,21%	7	76,36%
2008						
Cenário (prazo)	**Limiar de Predição**					
	50%		**70%**		**85%**	
	Quantidade média de predições	Porcentagem de acerto	Quantidade média de predições	Porcentagem de acerto	Quantidade média de predições	Porcentagem de acerto
Curto	183	8,61%	52	11,90%	7	17,24%
Médio	183	26,64%	52	33,59%	7	38,98%
Longo	184	33,56%	53	42,44%	7	50,08%
2009						
Cenário (prazo)	**Limiar de Predição**					
	50%		**70%**		**85%**	
	Quantidade média de predições	Porcentagem de acerto	Quantidade média de predições	Porcentagem de acerto	Quantidade média de predições	Porcentagem de acerto
Curto	207	12,46%	77	15,92%	18	14,87%
Médio	205	22,00%	77	27,88%	18	29,16%
Longo	205	32,00%	75	40,66%	19	41,46%
2010						
Cenário (prazo)	**Limiar de Predição**					
	50%		**70%**		**85%**	
	Quantidade média de predições	Porcentagem de acerto	Quantidade média de predições	Porcentagem de acerto	Quantidade média de predições	Porcentagem de acerto
Curto	234	6,85%	91	8,72%	25	8,58%
Médio	234	14,31%	92	17,13%	25	14,97%
Longo	238	20,07%	90	24,00%	25	23,92%

Fonte: Elaborado pelo autor (2022).

Tabela 6 – Informações da rede

Ano	Quantidade de nós	Quantidade de ligações
2006	50	261
2007	52	325
2008	54	359
2009	59	397
2010	63	461

Fonte: Elaborado pelo autor (2022).

A partir da Tabela 5, constatou-se que o cenário que melhor desempenhou quanto a porcentagem de acerto foi o cenário de longo prazo, independente do ano de treinamento do modelo e do limiar de predição. Este resultado evidencia que as predições realizadas apresentam, de fato, potencial de ocorrerem futuramente.

Atentando-se ao limiar de predição e considerando todo o intervalo de 2006 a 2010, o limiar de 85% forneceu melhores resultados em 4 dos 5 anos avaliados. A exceção foi o ano de 2010, onde a maior taxa de acerto foi de 24% com o limiar de 70%, muito próxima da acurácia de 23,92% obtida pelo limiar de 85%. De forma abrangente, com a exceção do exemplo recém-citado, a interpretação dos resultados deve-se ao fato de que a quantidade de predições realizadas diminui à medida que a probabilidade da ligação ocorrer aumenta, o que consequentemente eleva a porcentagem de acerto das predições. Ao levar em conta todo o período de análise, a média da quantidade de predições realizadas foi de 188 para o limiar de 50%, 52 para o de 70%, e 11 predições para o limiar de 85%, o que influencia diretamente nas taxas de acerto.

Outro fator relevante na discussão é a complexidade da rede. Ao confrontar os dados da Tabela 5 e Tabela 6, observa-se que a porcentagem de acerto das predições decai ano a ano e, em contrapartida, a quantidade de nós e ligações presentes na rede cresce. Este comportamento revela que o aspecto dinâmico das redes tem um grande impacto sobre a tarefa de previsão de tecnologias. A tendência de evolução da rede presume que ela fique maior e mais complexa ao longo do tempo. Isto torna mais difícil a atribuição de representá-la e predizer quais ligações irão ocorrer na prática, pois a inclusão de novas entidades implica no surgimento de inúmeras novas possíveis ligações. Sob esta ótica, torna-se necessário aumentar a capacidade de aprendizado de representação de rede e a robustez do modelo preditivo a fim de obter melhores resultados em redes complexas.

Ao considerar o cenário de longo prazo e o limiar de predição 85%, o método proposto foi capaz de alcançar uma porcentagem de acerto média de 53,98% para o período analisado no estudo de caso. Em redes de menor complexidade, ou seja, com menor quantidade de nós e consequentemente de possíveis ligações, o método obteve um desempenho melhor, atingindo até 78,06% de acerto nas predições realizadas. Diante disto, o modelo de predição construído demonstra ser capaz de cumprir sua tarefa, promovendo indícios, por meio da predição de ligações, de possíveis tendências tecnológicas. Da perspectiva de aplicação, o método proposto é capaz de apoiar o processo de tomada de decisões estratégicas, uma vez que seu impacto em termos de inteligência competitiva e construção de estratégias de negócios se sobressai diante a precisão da predição.

6 Considerações Finais e Trabalhos Futuros

Este trabalho apresentou um método de análise de patentes para desempenhar a tarefa de previsão de tecnologias a fim de apoiar a tomada de decisões estratégicas em ambientes organizacionais. Com este objetivo, o trabalho empregou aprendizado de representação de rede baseado em redes neurais rasas e predição de ligações levando em conta a semelhança de poder preditivo e a utilização de recursos computacionais em comparação aos métodos de aprendizado profundo. Optou-se por adotar uma abordagem baseada em redes neurais com a finalidade de garantir ao método proposto neutralidade em termos de domínio, uma vez que a técnica é capaz de adaptar-se ao contexto de aplicação.

Nesta pesquisa, 2.000.147 documentos de patente do conjunto de dados USPTO-2M® foram coletados para realização de um estudo de caso a respeito do domínio de "carbono". As avaliações do método proposto levaram em conta predições em cenários de curto (um ano), médio (três anos) e longo (cinco anos) prazo sob limiares de predição de 50%, 70% e 85%. Os resultados demonstraram que, independente do limiar e do ano de treinamento do modelo, o melhor cenário de predições foi o de longo prazo. Em média, a taxa de acerto a longo prazo das predições realizadas sob o limiar de 85% foi de 53,98%. Além disso, o modelo de predição de ligações atingiu uma ROC-AUC média de 0,91. Assim, o método proposto mostra-se viável para a conjuntura apresentada.

Apesar das contribuições do método proposto, este estudo está sujeito a limitações que requerem trabalhos futuros. Primeiro, a fim de lidar com redes mais complexas, é imprescindível trabalhar a capacidade de aprendizado de representação de rede e a robustez do modelo preditivo. Segundo, ainda que o domínio e a lista de tecnologias sejam definidos pela organização conforme seus interesses, as predições não levam em consideração os objetivos específicos e a capacidade tecnológica da instituição. Ademais, é necessário que os termos que descrevem o domínio e a base de patentes sejam atualizados periodicamente visando refletir as tecnologias mais recentes, o que pode causar atraso indesejado na identificação de oportunidades.

Pesquisas futuras podem explorar o portfólio de tecnologias das organizações para identificar suas estratégias de PD&I e fornecer predições personalizadas. Com o intuito de automatizar a coleta de documentos de patentes e manter a base atualizada, técnicas de coleta de dados podem ser adotadas. Tendo em vista a evolução do estudo, reconhecimento de entidades nomeadas poderia ser aplicado sobre as ligações preditas no ano subsequente para aumentar a representatividade do conhecimento daquele domínio. Além disso, o conceito de sinais fracos poderia ser integrado à predição de ligações para potencializar a capacidade preditiva do método.

Referências

ABBAS, A.; ZHANG, L.; KHAN, S. U. A literature review on the state-of-the-art in patent analysis. *World Patent Information*, v. 37, p. 3–13, 2014. ISSN 0172-2190. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0172219013001634>. Citado na página 9.

ABIODUN, O. I. et al. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, v. 4, n. 11, p. e00938, 2018. ISSN 2405-8440. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2405844018332067>. Citado na página 15.

AGGARWAL, C. C. Machine learning with shallow neural networks. In: *Neural networks and deep learning*. [S.l.]: Springer, Cham, 2018. p. 53–104. ISBN 978-3-319-94463-0. Citado na página 16.

AL-ASWADI, F. N.; CHAN, H. Y.; GAN, K. H. Automatic ontology construction from text: a review from shallow to deep learning trend. *Artificial Intelligence Review*, v. 53, n. 6, p. 3901–3928, Aug 2020. ISSN 1573-7462. Disponível em: <https://doi.org/10.1007/s10462-019-09782-9>. Citado na página 16.

ALTUNTAS, S.; DERELI, T.; KUSIAK, A. Forecasting technology success based on patent data. *Technological Forecasting and Social Change*, v. 96, p. 202–214, 2015. ISSN 0040-1625. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0040162515000700>. Citado 2 vezes nas páginas 10 e 11.

AMARA, A.; TAIEB, M. A. H.; AOUICHA, M. B. Network representation learning systematic review: Ancestors and current development state. *Machine Learning with Applications*, v. 6, p. 100130, 2021. ISSN 2666-8270. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666827021000657>. Citado 3 vezes nas páginas 11, 15 e 17.

BARABÁSI, A.-L.; PÓSFAI, M. *Network science*. Cambridge: Cambridge University Press, 2016. ISBN 9781107076266. Disponível em: <http://networksciencebook.com/>. Citado na página 11.

CHEN, Q.; WANG, C.-H.; HUANG, S.-Z. Effects of organizational innovation and technological innovation capabilities on firm performance: evidence from firms in china's pearl river delta. *Asia Pacific Business Review*, Routledge, v. 26, n. 1, p. 72–96, 2020. Disponível em: <https://doi.org/10.1080/13602381.2019.1592339>. Citado na página 7.

CHO, Y.; DAIM, T. Technology forecasting methods. In: DAIM, T.; OLIVER, T.; KIM, J. (Ed.). *Research and Technology Management in the Electricity Industry: Methods, Tools and Case Studies*. London: Springer London, 2013. p. 67–112. ISBN 978-1-4471-5097-8. Disponível em: <https://doi.org/10.1007/978-1-4471-5097-8_4>. Citado 2 vezes nas páginas 10 e 11.

CHOI, D.; SONG, B. Exploring technological trends in logistics: Topic modeling-based patent analysis. *Sustainability*, v. 10, n. 8, 2018. ISSN 2071-1050. Disponível em: <https://www.mdpi.com/2071-1050/10/8/2810>. Citado 2 vezes nas páginas 7 e 18.

DADU, A. et al. A study of link prediction using deep learning. *Communications in Computer and Information Science*, v. 955, p. 377–385, 2019. Cited By 1. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059036251&doi=10.1007%2f978-981-13-3140-4_34&partnerID=40&md5=e7614381a2d797d118a8183773f2d25e>. Citado na página 8.

DHARWAL, R.; KAUR, L. Applications of artificial neural networks: a review. *Indian J. Sci. Technol*, v. 9, n. 47, p. 1–8, 2016. Citado 2 vezes nas páginas 14 e 15.

FIRAT, A. K.; WOON, W. L.; MADNICK, S. Technological forecasting–a review. *Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology*, p. 1–19, 2008. Citado na página 11.

GROVER, A.; LESKOVEC, J. Node2vec: Scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 855–864. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939754>. Citado 2 vezes nas páginas 16 e 17.

GUPTA, A.; MATTA, P.; PANT, B. Graph neural network: Current state of art, challenges and applications. *Materials Today: Proceedings*, v. 46, p. 10927–10932, 2021. ISSN 2214-7853. International Conference on Technological Advancements in Materials Science and Manufacturing. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2214785321010543>. Citado na página 8.

HAENLEIN, M.; KAPLAN, A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, v. 61, n. 4, p. 5–14, 2019. Disponível em: <https://doi.org/10.1177/0008125619864925>. Citado 2 vezes nas páginas 13 e 14.

HALEEM, A. et al. Technology forecasting (tf) and technology assessment (ta) methodologies: a conceptual review. *Benchmarking: An International Journal*, Emerald Publishing Limited, v. 26, n. 1, p. 48–72, Jan 2019. ISSN 1463-5771. Disponível em: <https://doi.org/10.1108/BIJ-04-2018-0090>. Citado 2 vezes nas páginas 7 e 10.

JEONG, B. et al. Trademark-based framework to uncover business diversification opportunities: Application of deep link prediction and competitive intelligence analysis. *Computers in Industry*, v. 124, p. 103356, 2021. ISSN 0166-3615. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016636152030590X>. Citado na página 19.

KIM, G.; BAE, J. A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change*, v. 117, p. 228–237, 2017. ISSN 0040-1625. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0040162516307661>. Citado 3 vezes nas páginas 7, 10 e 19.

KIM, K. H. et al. Text mining for patent analysis to forecast emerging technologies in wireless power transfer. *Sustainability*, v. 11, n. 22, 2019. ISSN 2071-1050. Disponível em: <https://www.mdpi.com/2071-1050/11/22/6240>. Citado 2 vezes nas páginas 10 e 18.

KRESTEL, R. et al. A survey on deep learning for patent analysis. *World Patent Information*, v. 65, p. 102035, 2021. ISSN 0172-2190. Disponível em: <https://www.sciencedirect.com/science/article/pii/S017221902100017X>. Citado 3 vezes nas páginas 7, 9 e 10.

KUMAR, A. et al. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, v. 553, p. 124289, 2020. ISSN 0378-4371. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0378437120300856>. Citado 3 vezes nas páginas 8, 11 e 17.

LEE, C. et al. Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change*, v. 127, p. 291–303, 2018. ISSN 0040-1625. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0040162517304778>. Citado na página 19.

LEE, J. et al. An approach for discovering firm-specific technology opportunities: Application of link prediction to f-term networks. *Technological Forecasting and Social Change*, v. 168, p. 120746, 2021. ISSN 0040-1625. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0040162521001785>. Citado 2 vezes nas páginas 20 e 26.

LENZ, R. C. *Technological forecasting*. [S.l.], 1962. Citado na página 10.

LI, S. et al. Deeppatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, v. 117, n. 2, p. 721–744, Nov 2018. ISSN 1588-2861. Disponível em: <https://doi.org/10.1007/s11192-018-2905-5>. Citado na página 23.

LIBEN-NOWELL, D.; KLEINBERG, J. The link prediction problem for social networks. In: *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2003. (CIKM '03), p. 556–559. ISBN 1581137230. Disponível em: <https://doi.org/10.1145/956863.956972>. Citado na página 11.

LUPU, M. et al. Patent-related tasks at ntcir. In: LUPU, M. et al. (Ed.). *Current Challenges in Patent Information Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017. p. 77–111. ISBN 978-3-662-53817-3. Disponível em: <https://doi.org/10.1007/978-3-662-53817-3_3>. Citado na página 9.

MALEK, M. et al. Shallow node representation learning using centrality indices. In: *2021 IEEE International Conference on Big Data (Big Data)*. [S.l.: s.n.], 2021. p. 5209–5214. Citado 3 vezes nas páginas 11, 15 e 16.

MARTÍNEZ, V.; BERZAL, F.; CUBERO, J.-C. A survey of link prediction in complex networks. Association for Computing Machinery, New York, NY, USA, v. 49, n. 4, dec 2016. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3012704>. Citado 2 vezes nas páginas 11 e 12.

OTTONICAR, S.; VALENTIM, M.; MOSCONI, E. A competitive intelligence model based on information literacy: Organizational competitiveness in the context of the 4th industrial revolution. *Journal of Intelligence Studies in Business*, v. 8, p. 55–65, 12 2018. Citado 2 vezes nas páginas 7 e 8.

PARK, I.; YOON, B. Technological opportunity discovery for technological convergence based on the prediction of technology knowledge flow in a citation network. *Journal of Informetrics*, v. 12, n. 4, p. 1199–1222, 2018. ISSN 1751-1577. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1751157718300907>. Citado na página 20.

PENG, H. et al. Dynamic network embedding via incremental skip-gram with negative sampling. *Science China Information Sciences*, v. 63, n. 10, p. 202103, Sep 2020. ISSN 1869-1919. Disponível em: <https://doi.org/10.1007/s11432-018-9943-9>. Citado na página 16.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence, A Modern Approach*. Hoboken, NJ, USA: Pearson Education, 2020. 2145 p. ISBN 9780134610993. Citado na página 13.

SHANMUGANATHAN, S. Artificial neural network modelling: An introduction. In: *Artificial neural network modelling*. [S.l.]: Springer, Cham, 2016. p. 1–14. ISBN 978-3-319-28495-8. Citado 2 vezes nas páginas 14 e 15.

SILVA, I. N. D. et al. Artificial neural network architectures and training processes. In: *Artificial neural networks*. [S.l.]: Springer, Cham, 2017. p. 21–28. ISBN 978-3-319-43162-8. Citado 2 vezes nas páginas 14 e 15.

TKáč, M.; VERNER, R. Artificial neural networks in business: Two decades of research. *Applied Soft Computing*, v. 38, p. 788–804, 2016. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494615006122>. Citado 2 vezes nas páginas 14 e 15.

WANG, W. et al. Link prediction based on deep convolutional neural network. *Information (Switzerland)*, v. 10, n. 5, 2019. Cited By 5. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065883407&doi=10.3390%2finfo10050172&partnerID=40&md5=8b5bc15482957fa7df98a5a5259606b6>. Citado 5 vezes nas páginas 8, 12, 13, 15 e 16.

WINKLER, D. A.; LE, T. C. Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and qsar. *Molecular Informatics*, v. 36, n. 1-2, p. 1600118, 2017. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201600118>. Citado na página 16.

WU, W. et al. Hashing-accelerated graph neural networks for link prediction. In: *Proceedings of the Web Conference 2021*. New York, NY, USA: Association for Computing Machinery, 2021. (WWW '21), p. 2910–2920. ISBN 9781450383127. Disponível em: <https://doi.org/10.1145/3442381.3449884>. Citado 4 vezes nas páginas 8, 12, 13 e 16.

WU, Y.-c.; FENG, J.-w. Development and application of artificial neural network. *Wireless Personal Communications*, v. 102, n. 2, p. 1645–1656, Sep 2018. ISSN 1572-834X. Disponível em: <https://doi.org/10.1007/s11277-017-5224-x>. Citado na página 13.

YADAV, N.; YADAV, A.; KUMAR, M. History of neural networks. In: *An Introduction to Neural Network Methods for Differential Equations*. [S.l.]: Springer, Dordrecht, 2015. p. 13–15. ISBN 978-94-017-9816-7. Citado na página 13.

YOON, B.; LEE, S. Patent analysis for technology forecasting: Sector-specific applications. In: *2008 IEEE International Engineering Management Conference*. [S.l.: s.n.], 2008. p. 1–5. Citado 2 vezes nas páginas 7 e 9.

YUAN, X.; CAI, Y. Forecasting the development trend of low emission vehicle technologies: Based on patent data. *Technological Forecasting and Social Change*, v. 166, p. 120651, 2021. ISSN 0040-1625. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0040162521000834>. Citado na página 7.

ZHANG, D. et al. Network representation learning: A survey. *IEEE Transactions on Big Data*, v. 6, n. 1, p. 3–28, 2020. Citado 2 vezes nas páginas 15 e 17.

ZHANG, L.; LIU, Z. Research on technology prospect risk of high-tech projects based on patent analysis. *PLOS ONE*, Public Library of Science, v. 15, n. 10, p. 1–19, 10 2020. Disponível em: <https://doi.org/10.1371/journal.pone.0240050>. Citado 2 vezes nas páginas 7 e 9.

ZHOU, Y. et al. Forecasting emerging technologies using data augmentation and deep learning. *Scientometrics*, v. 123, n. 1, p. 1–29, Apr 2020. ISSN 1588-2861. Disponível em: <https://doi.org/10.1007/s11192-020-03351-6>. Citado na página 19.

Apêndice I

Lista de termos do domínio “carbono” utilizadas neste trabalho. Termos separados por “;” indicam sinônimos e também são considerandos na busca por meio do operador OR. Por exemplo, considerando o primeiro e o segundo termos, a busca na base de patentes será (“carbon dioxide equivalent” OR co2e) AND “carbon offset”.

Índice	Termo	Índice	Termo
1	carbon dioxide equivalent;co2e	75	emission factor;emissions factor
2	carbon offset	76	emission reduction;emission reduction
3	carbon footprint	77	emission trading;emissions trading
4	climate change	78	esg
5	greenhouse gas;ghg	79	extreme event
6	biomass	80	final energy
7	carbon neutral	81	forestry
8	kyoto protocol	82	fugitive emissions
9	afforestation	83	greenhouse gas emissions
10	carbon budget	84	greenwashing
11	carbon capture and sequestration;ccs;carbon capture;carbon sequestration	85	megawatt hour;mwh
12	carbon cycle	86	national greenhouse and energy reporting;nger
13	carbon dioxid;co2	87	net primary production;npp
14	carbon sequestration	88	net zero carbon business
15	carbon sink	89	operational control
16	global warming potential;gwp	90	paris agreement
17	leakage	91	regional greenhouse gas initiative;rggi
18	net zero	92	soil carbon
19	primary production;gpp	93	sustainable development goals;sdg
20	reduced emissions from deforestation and forest	94	urban

	degradation;redd+		
21	scope 1 emissions	95	verra
22	scope 2 emissions	96	water
23	scope 3 emissions	97	aau
24	united nations framework convention on climate change;unfccc	98	ab32
25	biofuel	99	accc
26	carbon capture and storage;ccs	100	acquittal
27	carbon credit	101	activated carbon
28	carbon market	102	activity
29	clean development mechanism;cdm;	103	additional offsets
30	deforestation	104	aerosols
31	fossil fuels	105	aeu
32	global warming	106	afsl
33	gold standard verified carbon standard;gs ver	107	aggregation
34	joint implementation;ji	108	agriculture, forestry, and other land use;afolu
35	land use, land use change and forestry;lulucf	109	airport carbon accreditation
36	life cycle assessment;lca	110	albedo
37	mitigation	111	algal bloom
38	negative emissions	112	altered carbon
39	net zero emissions	113	bioprocessing
40	recycling	114	biogas
41	reforestation	115	anoxic
42	removal unit;rmu	116	anreu
43	renewable energy	117	anthropogenic climate change
44	sequestration	118	asic
45	sink	119	atmosphere
46	tipping point	120	atmospheric column
47	verified carbon unit;vcu	121	australian emissions unit;aeu
48	verified emission reduction;ver	122	australian national greenhouse accounts
49	voluntary carbon	123	base year

	market;vcm		
50	accu	124	basis of preparation;bop
51	anthropogenic	125	bio-based
52	baseline	126	biochar
53	biocapacity	127	biodegradation
54	biodiversity	128	bioeconomy
55	bioenergy with carbon capture and storage;beccs	129	bioenergy
56	biosphere	130	biogenic emissions
57	blue carbon	131	biogeochemical cycles
58	cap and trade	132	biome
59	carbon capture and use;ccu	133	carbon nanotube
60	carbon dioxide removal;cdr	134	carbon dioxide
61	carbon flux	135	carbon fiber
62	carbon intensity	136	gas
63	carbon source	137	activated
64	carbon stock	138	carbon material
65	carbon units	139	system
66	certified emission reduction;cer	140	activated carbon
67	circular economy	141	carbon monoxide
68	compliance carbon market	142	hydrogen
69	cop	143	carbon black
70	direct air capture;dac	144	deposit
71	decarbonisation	145	carbon layer
72	ecosystem	146	carbon carbon
73	embedded carbon	147	carbon composite
74	embodied carbon	148	hydrocarbon