UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Mineração de texto usando algoritmos de redes complexas. Um estudo de caso em identificação de autores influentes

Gabriela Batista Silva

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Gabriela Batista Silva

Mineração de texto usando algoritmos de redes complexas. Um estudo de caso em identificação de autores influentes

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

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Alneu de Andrade Lopes

Versão original

São Carlos

2023

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

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

S856m

Silva, Gabriela Batista
Mineração de texto usando algoritmos de redes complexas. Um estudo de caso em identificação de autores influentes / Gabriela Batista Silva ; orientador Alneu de Andrade Lopes. – São Carlos, 2023.
56 p. : il. (algumas color.) ; 30 cm.

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

1. Mineração de Texto. 2. Redes Complexas. 3. Inteligência Artificial. 4. Identificação de Autores. 5. Colaboração em Pesquisa. 6. Rede de Coautoria. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Lopes, Alneu de Andrade, orient. II. Título.

Gabriela Batista Silva

Text mining using complex network algorithms. A case study on influential author identification.

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Alneu de Andrade Lopes

Original version

São Carlos

2023

"Mantenha-se faminto.
Mantenha-se tolo."
Steve Jobs

RESUMO

Silva, G. B. **Mineração de texto usando algoritmos de redes complexas. Um estudo de caso em identificação de autores influentes** . 2023. 56p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

A integração da mineração de texto e análise de redes complexas apresenta uma abordagem promissora para compreender as interações entre entidades em diversos campos. Este artigo explora a identificação de autores influentes na pesquisa em Inteligência Artificial (IA) nas áreas de engenharia elétrica e eletrônica, por meio do desenvolvimento de uma abordagem em redes complexas e mineração de texto. Uma coleção abrangente de artigos relevantes foi processada usando técnicas de mineração de texto, seguida pela criação de uma rede de coautoria. Os pesos das arestas foram baseados na frequência de coautoria, e métricas de centralidade foram utilizadas para identificar autores influentes. Além disso, a análise de palavras-chave proporcionou *insights* sobre tendências e padrões. O estudo visa contribuir para a identificação de autores influentes, oferecendo informações valiosas para a academia e indústria. Os objetivos gerais incluem o desenvolvimento de algoritmo, análise abrangente de artigos, e validação de técnicas de redes complexas e mineração de texto no contexto da pesquisa em IA. Espera-se que os resultados informem estratégias de tomada de decisão, colaboração e investimento no campo da pesquisa em IA.

Palavras-chave: Mineração de texto. Redes complexas. Inteligência Artificial. Identificação de autores. Rede de coautoria.

ABSTRACT

Silva, G. B. **Text mining using complex network algorithms. A case study on influential author identification.** . 2023. 56p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

The integration of text mining and complex network analysis presents a promising approach for understanding interactions among entities in various fields. This paper explores the identification of influential authors in Artificial Intelligence (AI) research within the electrical and electronic engineering domains through an approach based on complex networks and text mining. A comprehensive collection of relevant articles was processed using text mining techniques, followed by the creation of a co-authorship network. Edge weights were based on co-authorship frequency, and centrality metrics were employed to identify influential authors. Additionally, keyword analysis provided insights into trends and patterns. The study aims to contribute to the identification of influential authors, offering valuable insights for academia and industry. General objectives include algorithm development, comprehensive article analysis, and validation of complex network and text mining techniques in the AI research context. The outcomes are expected to inform decision-making, collaboration, and investment strategies in the field of AI research. T

Keywords: Text mining. Complex networks. Artificial Intelligence. Author Identification. Co-authorship Network.

LISTA DE FIGURAS

Figura 1 – Representação de um grafo. Fonte: Autor. 27
Figura 2 – Representação de uma rede complexa. Fonte: (SILVA; MACHADO; MACIEL, 2020). 28
Figura 3 – Quantidade de trabalhos por ano de publicação. Fonte: Autor. 43
Figura 4 – Rede de coautoria. Fonte: Autor. 44
Figura 5 – Rede de coautoria com os pesos da medida de centralidade de Katz por nó (autor). Fonte: Autor. 46
Figura 6 – Rede de coautoria com os pesos da medida de centralidade de grau por nó (autor). Fonte: Autor. 48
Figura 7 – Rede de coautoria com os pesos da medida de centralidade de proximidade por nó (autor). Fonte: Autor. 49
Figura 8 – Rede de coautoria com os pesos da medida de centralidade de proximidade por nó (autor). Fonte: Autor. 51
Figura 9 – Nuvem de palavras a partir das palavras-chave dos artigos. Fonte: Autor. 52
Figura 10 – Nuvem de palavras a partir do resumo pré-processado dos artigos. Fonte: Autor. 52

LISTA DE TABELAS

Tabela 1 – Autores influentes e sua posição no ranking de cada uma das medidas . 45
Tabela 2 – 10 autores com maior Centralidade de Katz 47
Tabela 3 – 10 autores com maior Centralidade de Grau 47
Tabela 4 – 10 autores com maior Centralidade de Proximidade 50
Tabela 5 – 10 autores com maior Centralidade de Autovetor 50

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

IA	Inteligência Artificial
RC	Redes Complexas
USP	Universidade de São Paulo

SUMÁRIO

1 INTRODUÇÃO . . . 23
1.1 Contextualização . . . 23
1.2 Objetivos . . . 25
1.3 Organização do texto . . . 26

2 FUNDAMENTAÇÃO TEÓRICA . . . 27
2.1 Redes Complexas . . . 27
2.2 Medidas de Centralidade . . . 30
2.2.1 Centralidade de Grau . . . 30
2.2.2 Centralidade de Autovetor . . . 30
2.2.3 Centralidade de Intermediação . . . 31
2.2.4 Centralidade de Proximidade . . . 32
2.2.5 Centralidade de Katz . . . 32

3 TRABALHOS RELACIONADOS . . . 33

4 METODOLOGIA . . . 37
4.1 Extração da Base de Dados . . . 37
4.2 Preparação da Base . . . 40
4.3 Construção da Rede de Coautoria . . . 40
4.4 Medidas de Centralidade . . . 41
4.5 Análise de Tendências e Padrões . . . 41

5 AVALIAÇÃO EXPERIMENTAL . . . 43
5.1 Conhecendo o Conjunto de Dados . . . 43
5.2 Rede de Coautoria . . . 43
5.3 Medidas de Centralidade e Definição dos Autores Influentes . . . 43
5.3.1 Centralidade de Katz . . . 43
5.3.2 Centralidade de Grau . . . 44
5.3.3 Centralidade de Proximidade . . . 45
5.3.4 Centralidade de Autovetor . . . 46
5.4 Tendências e Padrões . . . 47
5.5 Resultados e Discussões . . . 50

6 CONCLUSÕES . . . 53

Referências . . . 55

1 INTRODUÇÃO

1.1 Contextualização

A área de *text mining* é uma ferramenta poderosa utilizada em pesquisas acadêmicas e empresariais para extrair informações valiosas a partir de grandes volumes de texto. As técnicas de *text mining* são usadas para analisar dados não estruturados, como documentos, e-mails, posts em mídias sociais e outros tipos de conteúdo textual. Essas técnicas permitem a identificação de padrões, tópicos, opiniões, sentimentos e outras informações relevantes contidas nos textos. As pesquisas em *text mining* têm sido aplicadas em diversas áreas, como saúde, finanças, marketing, análise de mídias sociais, entre outras, permitindo uma melhor compreensão dos dados e uma tomada de decisão mais informada. A mineração de texto é uma extensão da técnica de mineração de dados. As técnicas de mineração de dados são principalmente usadas para a extração de padrões de um banco de dados estruturado. As técnicas de mineração de texto se tornam mais complexas em comparação com a mineração de dados devido à natureza não estruturada do texto em linguagem natural (IRFAN *et al.*, 2015).

O uso de *text mining* em conjunto com análise de redes complexas (RCs) é uma abordagem promissora para a compreensão de interações entre diferentes entidades em diversas áreas de aplicação, devido à complexidade e riqueza das informações disponíveis em redes. As redes complexas podem ser visualizadas em grafos que exibem suas topologias configuradas por nós interligados por arestas. O objetivo é descobrir novos conhecimentos úteis a partir de dados representados em forma de grafo (OSMAN; BARUKUB, 2020) Os grafos são recursos matemáticos utilizados nas pesquisas de redes por possuírem propriedades relevantes além de serem adequados para representar a topologia das mesmas (MELO, 2017)

Grafos são vistos como uma forma natural de representar as conexões entre dados e ao aumento no número de ferramentas disponíveis para lidar com essas estruturas complexas. Os grafos têm a capacidade de mapear diferentes níveis de linguagem em uma estrutura de dados mais rica, que de outra forma não poderia ser facilmente e de forma integrada utilizada. As técnicas de *text mining* podem ser utilizadas em associação com as redes complexas para extrair informações valiosas dos textos relacionados a cada nó, como por exemplo, a frequência de palavras-chave, tópicos recorrentes e sentimentos associados. Além disso, as informações extraídas a partir dos textos podem ser combinadas com as informações da rede para análises mais avançadas, como identificação de comunidades e detecção de *hubs*. Isso permite uma compreensão mais completa das interações entre os nós, possibilitando a identificação de padrões e comportamentos interessantes que podem ajudar a melhorar a tomada de decisões em diferentes contextos. A análise de redes complexas em

conjunto com *text mining* é uma abordagem cada vez mais relevante e valiosa em diversas áreas de pesquisa e aplicação (JUAREZ OFELIA CERVANTES VILLAGóMEZ, 2017).

No presente estudo, busca-se explorar a identificação de autores influentes em pesquisas sobre Inteligência Artificial (IA) no contexto de estudos das áreas de engenharia elétrica e eletrônica, por meio do desenvolvimento de um algoritmo baseado em redes complexas e *text mining*. Foi realizado um processo de coleta de artigos relevantes na área de interesse, abrangendo informações detalhadas dos trabalhos, que foram pré processadas utilizando técnicas de *text mining*. Posteriormente, foi criada uma rede de coautoria onde os nós representam autores e as arestas representam colaborações entre eles. A ponderação das arestas foi feita com base na frequência de coautoria. Métricas de centralidade foram aplicadas para ajudar a identificar autores que desempenham papéis importantes na rede de coautoria. Finalmente, o resumo as palavras-chave de cada artigo foram analisadas para extrair *insights* sobre tendências e padrões.

Este tema aborda várias áreas importantes na área de Big Data e IA, como processamento de texto, análise de redes complexas, mineração de dados e aprendizado de máquina. Além disso, o tema é relevante em várias áreas de pesquisa e pode ter aplicações práticas na indústria e na academia, nas quais espera-se poder contribuir com os resultados desta pesquisa, dentre elas:

- A identificação de autores influentes em uma área de pesquisa pode ajudar a identificar tendências e tópicos emergentes em uma determinada área, permitindo que as organizações e pesquisadores se mantenham atualizados com as últimas descobertas e desenvolvimentos.

- Identificar autores influentes em uma área de pesquisa também pode ajudar a identificar colaboradores potenciais para novos projetos ou pesquisas. Ao conhecer os principais especialistas em uma determinada área, os pesquisadores podem formar equipes de colaboração que aumentem a qualidade e a eficácia do seu trabalho.

- Em ambientes altamente competitivos, como a indústria de tecnologia, a identificação de autores influentes em uma área de pesquisa pode ajudar as organizações a entender melhor a concorrência e as tendências do mercado.

- A análise de redes complexas também pode ser expandida em contextos de redes sociais, como identificar os principais influenciadores em uma determinada comunidade online ou mídia social.

- A identificação de autores influentes em uma área de pesquisa pode ajudar os investidores a tomar decisões informadas sobre onde investir seu dinheiro. Por exemplo, se uma determinada empresa possui um grande número de autores influentes

trabalhando em uma área de pesquisa, isso pode indicar um investimento promissor nessa empresa.

1.2 Objetivos

Os objetivos gerais deste projeto são:

- Desenvolver um algoritmo de identificação de autores influentes: O principal objetivo é criar um algoritmo capaz de identificar os autores mais influentes nas pesquisas relacionadas à Inteligência Artificial (IA) nas áreas de elétrica e eletrônica. Isso será alcançado por meio do uso de técnicas de redes complexas e *text mining*, que permitirão analisar a estrutura e os padrões de colaboração entre os autores, bem como a similaridade de textos.

- Coletar e analisar artigos na área de interesse: Será realizada uma coleta abrangente de artigos científicos relacionados à IA no contexto estudado, com informações detalhadas sobre os autores. Os artigos serão analisados quanto aos temas, áreas de interesse e colaborações entre os autores, visando compreender a dinâmica da produção científica nessa área e identificar os autores mais relevantes.

- Contribuir para a identificação de autores influentes na área de IA nas áreas de elétrica e eletrônica: O projeto tem como objetivo fornecer *insights* e informações relevantes para a identificação de autores influentes na área de IA, contribuindo para o avanço do conhecimento nessa área. Os resultados obtidos podem ser utilizados por pesquisadores, instituições de pesquisa e agências de fomento para tomar decisões informadas sobre colaborações, financiamentos e identificação de líderes de pesquisa na área de IA no contexto brasileiro.

- Aplicar e validar técnicas de redes complexas e *text mining*: O projeto também busca aplicar e validar técnicas de redes complexas e *text mining* na área de IA, mostrando sua aplicabilidade e relevância na identificação de autores influentes em pesquisas científicas. Isso contribuirá para o desenvolvimento de metodologias e abordagens inovadoras na área de análise de redes complexas e mineração de textos aplicadas à identificação de autores influentes em diferentes contextos.

Esses são os objetivos gerais do projeto, que visam desenvolver um algoritmo de identificação de autores influentes em pesquisas de IA, utilizando técnicas de redes complexas e *text mining*, e contribuir para o avanço do conhecimento na área de análise de redes complexas e mineração de textos aplicadas à pesquisa científica em IA.

1.3 Organização do texto

No Capítulo 2 é apresentada uma revisão bibliográfica sobre RCs aplicadas em *text mining*‘. No Capítulo 3 é apresentada a metodologia utilizada neste trabalho. No Capítulo 4 são apresentados os resultados e discussões. Por fim, o Capítulo 5 reúne as conclusões deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Redes Complexas

O estudo de redes representadas por grafos é considerado um dos princípios básicos da matemática discreta e teve início em 1735, quando Euler apresentou uma solução para o problema das pontes de Königsberg, marcando o início da teoria dos grafos. As redes complexas são um tipo de grafo que apresentam propriedades topográficas bastante particulares, não encontradas em grafos mais simples. Elas são caracterizadas por serem grafos que apresentam uma estrutura topológica não trivial, composta por vértices (ou nós) interconectados por meio de arestas (AMANCIO, 2013). Um representação de um grafo é mostrada na figura 1

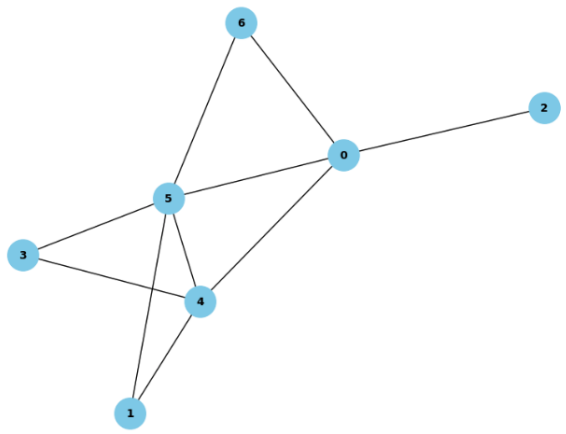

Figura 1 – Representação de um grafo. Fonte: Autor.

Uma alternativa para representar grafos é por meio de uma matriz de adjacências, na qual as linhas e colunas estão associadas aos nós da rede. Se houver uma conexão entre dois nós, o elemento correspondente na matriz é preenchido com o valor 1. Em redes não direcionadas, a matriz de adjacências é sempre simétrica. Quando as conexões na rede têm pesos, são utilizadas duas matrizes para representá-la: a matriz de adjacência, indicando as conexões entre os nós, e a matriz de pesos, que contém os valores dos pesos das conexões (RONQUI, 2014).

Os estudos iniciais das redes complexas, na década de 1930, eram voltados para a análise do comportamento social e das relações entre indivíduos, com base em características como centralidade e conectividade dos vértices. Com o avanço da tecnologia e análise de grandes volumes de dados, houve uma mudança de paradigma para o estudo de redes em larga escala, com milhões ou bilhões de vértices. Essa mudança revelou que as redes do mundo real têm características distintas das redes aleatórias, anteriormente consideradas

como principal modelo. Essa diferença na topologia e evolução das redes do mundo real levou ao termo "redes complexas"para descrevê-las (METZ *et al.*, 2007).

Uma rede é um grafo composto por vértices (ou nós) e arestas (ou arcos) que estabelecem relações entre esses vértices, de acordo com o problema modelado. O grafo pode ser direcionado ou não, onde em um grafo direcionado (dígrafo), cada aresta possui uma direção que conecta um vértice de origem a um vértice de destino. Na figura 2 é apresentada uma representação de uma rede complexa.

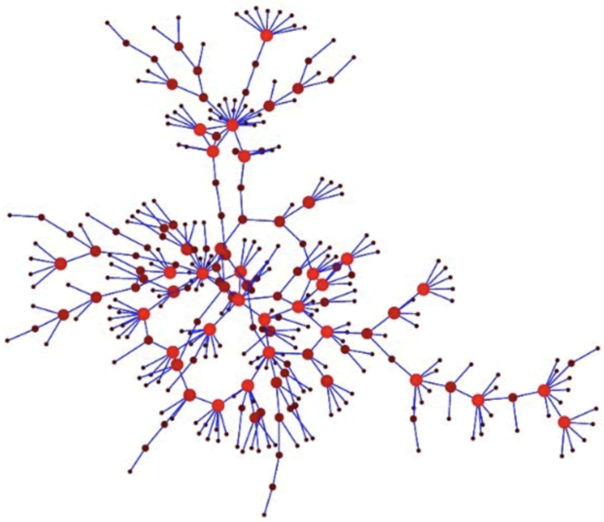

Figura 2 – Representação de uma rede complexa. Fonte: (SILVA; MACHADO; MACIEL, 2020).

Nem todo grafo é uma rede complexa, pois essa classificação requer a presença de propriedades topográficas específicas que não estão presentes em grafos simples (METZ *et al.*, 2007). Essas propriedades incluem:

- Estrutura topológica não trivial: As redes complexas apresentam uma organização estrutural que não é aleatória, com padrões de conexões específicos entre os vértices.
- Conectividade: Os vértices em uma rede complexa estão interconectados por meio de arestas, criando uma teia complexa de conexões.
- Centralidade: Alguns vértices têm maior importância ou influência na rede do que outros, sendo considerados mais centrais. Isso pode ser medido através de várias métricas, como centralidade de grau, centralidade de intermediação (*betweenness*) e centralidade de proximidade.
- *Clustering*: Agrupamentos de vértices interconectados formam regiões densas na rede, conhecidos como *clusters* ou comunidades, onde os vértices têm mais conexões entre si do que com o restante da rede.

- Hierarquia: A rede complexa pode apresentar uma hierarquia de conexões, com alguns vértices atuando como intermediários ou de ligação entre diferentes regiões da rede.

- Evolução dinâmica: As redes complexas podem mudar ao longo do tempo, com adição ou remoção de vértices e arestas, ou reorganização de sua estrutura. O estudo da dinâmica das redes é fundamental em áreas como a análise de redes sociais e a modelagem de redes biológicas.

- Resiliência: As redes complexas podem ser resilientes a falhas ou ataques, mantendo sua conectividade e funcionamento mesmo diante de perturbações. No entanto, também podem ser vulneráveis a determinadas falhas ou ataques, sendo importante entender sua resiliência para garantir sua robustez.

- Propriedades emergentes: A rede como um todo pode apresentar características que emergem das interações locais entre os vértices, resultando em propriedades globais complexas. Por exemplo, a propagação de informações ou influências em uma rede social pode ser uma propriedade emergente.

- Distribuição de graus: A distribuição do número de conexões dos vértices na rede pode seguir uma lei de potência, com alguns vértices possuindo muitas conexões (*hubs*) e a maioria possuindo poucas conexões. Isso é conhecido como distribuição de graus livre de escala e é uma característica importante de muitas redes complexas.

Existem muitos outros modelos desenvolvidos para estudar a estrutura e dinâmica das redes em diferentes contextos, como modelos de redes de mundo pequeno, modelos de redes hierárquicas, modelos de redes temporalmente evolutivas, entre outros. A escolha do modelo de rede complexa apropriado depende do contexto de estudo e dos objetivos de pesquisa específicos. Os três principais tipos de RCs são Redes Aleatórias, Redes Livres de Escala e Redes Pequeno Mundo. Redes Aleatórias são redes em que as conexões entre os nós são estabelecidas de forma aleatória, sem seguir um padrão específico. Um exemplo de modelo de rede aleatória é o Modelo de Erdős-Rényi (ER), em que cada par de nós é conectado com uma probabilidade fixa. Redes Livres de Escala são redes em que a distribuição de grau (número de conexões de um nó) segue uma distribuição de probabilidade de preferência, ou seja, alguns nós têm muitas conexões (*hubs*) enquanto a maioria dos nós tem poucas conexões. Um exemplo de modelo de rede livre de escala é o Modelo de Barabási-Albert (BA), em que novos nós são adicionados à rede com probabilidade proporcional ao seu grau existente. Por fim, Redes Pequeno Mundo são redes em que os nós estão altamente interconectados localmente, formando agrupamentos (*clusters*), mas também têm caminhos curtos de conexão globalmente. Um exemplo de modelo de rede pequeno mundo é o Modelo de Watts-Strogatz (WS), em que uma rede

regular é "rewirada"com uma probabilidade para criar conexões aleatórias, mantendo altos níveis de aglomeração local e curtos comprimentos médios de caminho (METZ *et al.*, 2007).

2.2 Medidas de Centralidade

As medidas de centralidade são métricas usadas em análise de redes para quantificar a importância ou centralidade de nós em um grafo. Essas medidas ajudam a identificar quais nós são mais "centrais"ou influentes em uma rede (RONQUI, 2014). Essas medidas são úteis em diferentes contextos e podem oferecer *insights* sobre a estrutura e a dinâmica de uma rede. A escolha da medida de centralidade depende do tipo de rede e das questões específicas que se deseja responder. Existem diversos tipos de medidas de centralidade, e algumas delas são exploradas a seguir.

2.2.1 Centralidade de Grau

A centralidade de grau é talvez a mais simples de todas as medidas de centralidade, ela mede o número de arestas conectadas a um nó. Um nó com um grande número de conexões tem uma centralidade de grau alta. O grau de um nó pode ser calculado utilizando a expressão a seguir:

$$k_i = \sum_{j=1}^{n} A_{ij} \tag{2.1}$$

onde k_i é o grau do nó i e A_ij são os elementos da matriz de adjacências da rede complexa e n é o número de vértices na rede complexa. Em alguns casos, pode-se utilizar uma variação da centralidade de grau dividindo-se o valor de k_i pelo maior grau possível, fazendo com que o valor da centralidade de grau de cada nó fique no intervalo entre 0 e 1. Para redes com direção, deve-se levar em consideração duas medidas de centralidade de grau: grau de entrada, que representa quantas ligações chegam em um nó, e grau de saída, que indica a quantidade de ligações o nó possui para outros nós (RONQUI, 2014).

2.2.2 Centralidade de Autovetor

A centralidade de autovetor é uma medida de influência de um nó em uma rede. Essa medida atribui pontuações relativas a todos os nós da rede com base no conceito de que conexões com nós de pontuação alta contribuem mais para a pontuação do nó em questão do que conexões iguais a nós de pontuação baixa. Assim, a centralidade de autovetor leva em consideração não apenas as conexões que o nó em si possui, como também quantas os seus vizinhos possuem (RONQUI, 2014).

O cálculo envolve um processo iterativo que pode ser expresso da seguinte forma:

- Definir a Matriz de Adjacência: Inciaremos supondo que possuímos uma matriz de adjacências A para o grafo em questão. Se A é uma matriz de nxn para um grafo com n nós, onde $A_ij = 1$ se há uma aresta entre os nós i e j, e $A_ij = 0$ caso contrário.

- Inciar o vetor de centralidade: Criaremos então um vetor inicial de centralidade $x^{(0)}$ para cada nó do grafo. Pode-se começar ocm um vetor uniforme ou com valores iniciais arbitrários.

- iteração até convergência: A expressão iterativa para o cálculo da centralidade de autovetor é dada por
$$x^{(t+1)} = \frac{Ax^{(t)}}{\|Ax^{(t)}\|}$$
onde t é o número da iteração e $||\ .\ ||$ representa a norma euclidiana.

- Normalização: Neste passo, vamos normalizar o vetor resultante após cada iteração. Este passo é importante para evitar que os valores cresçam indefinidamente.

- Convergência: Por fim, repete-se o processo até que o vetor x convirja para o autovetor dominante da matriz de adjacência A. A convergência pode ser verificada observando mudanças pequenas entre iterações sucessivas.

2.2.3 Centralidade de Intermediação

Também conhecida como *Betweenness Centrality*, mede a frequência com que um nó aparece nos caminhos mais curtos entre todos os pares de nós na rede. Um nó com alta centralidade de intermediação está frequentemente no caminho entre outros nós. Pode ser utilizada para avaliar a importância de um nó na transmissão de mensagens ou eventos entre os outros nós da rede se estes quiserem trocar informações (RONQUI, 2014).

A centralidade de intermediação para nós pode ser definida por:

$$x_i = \sum_{s,t} \frac{n^i_{st}}{g_{st}} \tag{2.2}$$

onde $\frac{n^i_{st}}{}$ é o número de menores caminhos entre os vértices s e t que passa pelo vértice i e g_{st} é o número total de menores caminhos entre os vértices s e t.

A centralidade de intermediação também pode ser definida para ligações da rede complexa. A ideia desta medida é conseguir distinguir quais ligações são mais importantes para a rede. A formulação matemática é similar à centralidade de interposição para nós:

$$x_e = \sum_{s,t} \frac{n^e_{st}}{g_{st}} \tag{2.3}$$

onde $\frac{n_{st}^{e}}{}$ é o número de menores caminhos entre os vértices s e t que passam pela ligação e e x_e é o valor da centralidade de interposição para a ligação e.

2.2.4 Centralidade de Proximidade

Denominada também *Closeness Centrality*, esta medida mede a distância média de um nó para todos os outros nós na rede. Um nó com alta centralidade de proximidade está mais próximo de todos os outros nós. Assim, nós que possuírem uma menor distância média comparados com os demais, receberão um valor alto para a centralidade de proximidade. Esses nós são importantes para a rede pois as informações contidas neles atingem os demais elementos da rede em um tempo menor do que os outros (RONQUI, 2014). A centralidade de proximidade pode ser calculada pela equação:

$$C_i = \frac{1}{l_i} \quad onde \quad l_i = \frac{1}{n-1} \sum_{j \neq i} d_{ij} \tag{2.4}$$

onde n representa o número total de nós na rede e d_{ij} é o comprimento do menor caminho entre os nós i e j, e l_i representa a média do comprimento das menores distâncias entre i e todos os outros nós da rede.

2.2.5 Centralidade de Katz

A centralidade de Katz leva em consideração não apenas os vizinhos diretos, mas também os vizinhos dos vizinhos e assim por diante (LARANJEIRA; CAVIQUE, 2018). A equação matemática para o cálculo da centralidade de Katz é dada por:

$$C_i = \sum_{k=1}^{\infty} \beta^k \sum_j A_{ij}^{(k)} \tag{2.5}$$

Onde C_i é a centralidade de Katz do nó i, β é um fator de atenuação com $0 < \beta < \frac{1}{\lambda_{max}}$ (onde λ_{max} é o maior autovalor da matriz de adjacência A) e A^k é a matriz de adjacência elevada à k-ésima potência.

3 TRABALHOS RELACIONADOS

As redes complexas têm sido aplicadas nas mais diversas áreas, para a resolução dos mais variados tipos de problemas. O objetivo deste trabalho é explorar a importância de uma estrutura de dados mais rica, como redes complexas, em tarefas de classificação de texto, por isso, os trabalhos relacionados foram examinados a partir da perspectiva do uso de grafos em problemas de texto em evidência.

A tarefa de classificação de textos com redes complexas refere-se ao uso de métodos e técnicas de análise de redes complexas para classificar textos em categorias ou classes pré-definidas. Essa tarefa envolve a representação do texto como uma rede complexa, em que os elementos do texto, como palavras ou frases, são tratados como nós, e as relações entre esses elementos, como co-ocorrência ou coocorrência de palavras, são tratadas como arestas ou conexões entre os nós. A análise de redes complexas pode ajudar a identificar características importantes dos textos, como palavras-chave, temas, padrões de co-ocorrência, influência de palavras ou elementos específicos e estrutura global do texto. Amancio (2013) investiga a aplicabilidade de atributos de estilo em tarefas tradicionais de classificação, usando modelagem de textos como redes complexas e estuda como métricas topológicas podem ser úteis no processamento de linguagem natural, com o apoio de métodos de aprendizado de máquina supervisionado e não supervisionado. A técnicas utilizadas são aplicadas para reconhecimento de autoria e desambiguação de sentido de palavras.

Em Amancio *et al.* (2012) os autores representaram trechos de texto com diferentes níveis de simplificação em redes de coocorrência e constataram que a regularidade topológica correlacionou negativamente com a complexidade textual. Além disso, em textos menos complexos, a distância entre os conceitos, representados como nós, tendeu a diminuir. As métricas de redes complexas foram tratadas com técnicas de reconhecimento de padrões multivariados, o que permitiu aos autores distinguir entre os textos originais e suas versões simplificadas. O objetivo do estudo é realizar uma triagem rápida de textos sempre que a avaliação da complexidade for essencial para garantir a acessibilidade a leitores com habilidades limitadas de leitura.

O problema de Atribuição de Autoria consiste em identificar o autor de um texto desconhecido a partir de características linguísticas presentes no texto. A ideia é criar um modelo de autor baseado em um conjunto de textos conhecidos e, em seguida, usar esse modelo para identificar o autor de um texto desconhecido (CASTILLO *et al.*, 2015). Nesse contexto, os grafos podem ser usados para representar os textos de forma a capturar as características importantes para a identificação do autor. Cada nó do grafo representa uma palavra, e as arestas representam as relações entre as palavras, como co-ocorrências, similaridade semântica, entre outras. Com essa representação em grafo, é possível aplicar

algoritmos de análise de rede para identificar padrões e características que possam ajudar na identificação do autor. Além disso, a abordagem com grafos permite a integração de diferentes fontes de informação, como metadados dos textos, informações sobre os autores, entre outros, o que pode melhorar a qualidade da atribuição de autoria (JUAREZ OFELIA CERVANTES VILLAGóMEZ, 2017)

Castillo *et al.* (2015) apresenta uma metodologia formal para mapear textos em grafos, onde o processo de mapeamento considera a estrutura sintática das sentenças do texto. Além disso, integra-se as etiquetas de Partes do Discurso (PoS) como características estilométricas do autor, isto é, são utilizados grafos para identificar o estilo de escrita do autor. Os experimentos apresentados no trabalho alcançam uma taxa de precisão de cerca de 79%, demonstrando que a representação baseada em grafos pode ser uma forma de encapsular diversos níveis de descrições de linguagem natural em uma estrutura simples.

Macke (2015) propõe uma rede neural que utiliza a estrutura gramatical e o vocabulário de documentos de texto como base para encontrar pistas que possam determinar o autor do documento. Stamatatos *et al.* (2014) aplicam uma rede neural que combina recursos lexicais e sintáticos como camadas de entrada para classificar um autor.

Barbosa (2016) apresenta um método que utiliza redes complexas heterogêneas para representar e analisar metadados de artigos científicos. A abordagem proposta extrai dados relevantes, como atributos, por meio de técnicas de mineração de texto, e os insere como elementos em uma rede complexa armazenada em um banco de dados orientado a grafos. O objetivo principal é analisar como conceitos, instituições de ensino e indivíduos estão relacionados em um conjunto local de artigos científicos, e compreender como a colaboração é estabelecida sob diferentes perspectivas, incluindo parcerias entre autores e países, relação entre temas de pesquisa, entre outros tópicos relacionados à colaboração científica.

É comum encontrar diferentes formas de citação de um mesmo autor em coleções de artigos científicos, o que pode resultar em registros duplicados que precisam ser identificados em muitas aplicações. Isso é conhecido como o problema de identificação de duplicatas, para o qual ainda não foram alcançados resultados satisfatórios. No estudo desenvolvido em Silva Breno Júnio V e Motta (2011), foi investigada a utilização de redes de cooperação científica e técnicas de detecção de comunidade para lidar com esse problema. Os resultados indicam que essa abordagem não só melhora a precisão na identificação de duplicatas, mas também reduz o custo computacional associado a essa tarefa em comparação com abordagens em que cada registro é comparado com todos os outros.

Em Antiqueira *et al.* (2009) A sumarização automática de textos com redes complexas é explorada. A sumarização automática se mostra importante mediante à quantidade de informações disponíveis em mídias digitais, o que aumentou a demanda por estratégias de sumarização simples e independentes de linguagem. Neste trabalho, utiliza-se conceitos

e métricas de redes complexas para selecionar frases para um resumo. O grafo ou rede representa o texto, com nós correspondentes a frases e arestas conectando frases que compartilham substantivos em comum.

Existem duas abordagens principais na análise de textos com fins de identificar o autor: a Atribuição de Autoria e o Perfil de Autor. A primeira se concentra na análise do estilo de escrita individual do autor para determinar a autoria de um texto desconhecido. Já a segunda considera informações demográficas do autor, como idade, gênero ou traços de personalidade, para entender melhor o contexto em que o texto foi produzido. Ambas as abordagens são valiosas para analisar como as pessoas compartilham textos e os tipos de padrões textuais que emergem dessas interações (JUAREZ OFELIA CERVANTES VILLAGóMEZ, 2017).

Uma abordagem comum para a definição do perfil do autor é classificação supervisionada com um grafo de co-ocorrência para detectar idade e gênero. Rangel and Rosso (2016) modelam as seis emoções básicas com uma abordagem baseada em grafo inovadora, eles investigam o impacto das emoções na caracterização de autores, identificando concretamente a idade e o gênero. Alemán, Loya and Vilariño (2013) descreve duas metodologias aplicadas à tarefa de caracterização de autores. A primeira metodologia considera uma abordagem clássica de classificação, usando diversas características extraídas dos textos para alimentar um classificador baseado em florestas aleatórias. A segunda metodologia, baseada em técnicas de mineração de grafos, obteve um desempenho muito ruim na tarefa de caracterização de autores.

A Verificação de Autoria tem como objetivo confirmar se um autor específico escreveu um documento desconhecido com base em algumas amostras do seu estilo de escrita. Nesse contexto, é importante destacar que há um número limitado de artigos que utilizam uma abordagem baseada em grafos para solucionar essa tarefa. Ayala *et al.* (2013) utiliza características lexicais, sintáticas e baseadas em grafos para construir um modelo de representação de autores de documentos e utilizam como modelo de classificação a técnica de Máquinas de Vetores de Suporte (SVM, do inglês *Support Vector Machine*).

4 METODOLOGIA

4.1 Extração da Base de Dados

Neste estudo, almeja-se investigar a identificação de autores influentes em pesquisas relacionadas à Inteligência Artificial (IA) no âmbito dos estudos em engenharia elétrica e eletrônica. Todos os algoritmos deste trabalho foram desenvolvidos utilizando a linguagem de programação Python.

Para a realização desta tarefa, foi selecionado como base dados o repositório de artigos do Instituto de Engenheiros Eletricistas e Eletrônicos (*Institute of Electrical and Electronics Engineers*) ou IEEE. Estabelecido em 1884 nos Estados Unidos, o IEEE é uma sociedade técnico-profissional internacional com sede em Nova Jersey. Comprometido com o progresso da teoria e prática da engenharia nos domínios da eletricidade, eletrônica e computação, o IEEE reúne mais de 410.000 membros, incluindo engenheiros, cientistas, pesquisadores e outros profissionais, distribuídos em aproximadamente 160 países. O instituto possui uma biblioteca digital, o IEEE Xplore, que é um banco de dados de pesquisa para descoberta e acesso a artigos de periódicos, anais de conferências, normas técnicas e materiais relacionados em ciência da computação, engenharia elétrica e eletrônica e áreas afins. É possível acessar os metadados de cada trabalho na biblioteca, que possuem 45 campos que contém, dentre outras informações, dados dos autores, sua afiliação, o resumo do trabalho, palavras-chave, quantidade de citações, etc. Os campos retornados são:

- abstract: Resumo breve ou declaração do conteúdo de um artigo de periódico, paper de conferência, padrão, livro/capítulo de livro e/ou curso;
- abstract_url: URL do IEEE Xplore que irá retornar o resumo;
- author_url: URL do IEEE Xplore que irá retornar detalhes do autor;
- accessType: Tipo de acesso. Acesso Aberto (Disponível gratuitamente pela IEEE), Efémero (Disponível gratuitamente pela IEEE), Bloqueado (Faça login ou saiba sobre opções de assinatura), Plagiado (Marcado como um artigo plagiado);
- article_number: Identificador exclusivo do IEEE para um artigo específico;
- author_order: Quando vários autores estão listados, a ordem do autor fornece um número para cada autor;
- author_terms: Termos fornecidos pelo autor que descrevem os tópicos ou assuntos do documento;

- authors: Nome do(s) autor(es) listado(s) no documento. Os nomes dos autores são fornecidos por completo, juntamente com a ordem em que estão listados;
- affiliation: Nome da(s) afiliação(ões) listada(s) no documento. Os nomes das afiliações são fornecidos por extenso, juntamente com a ordem em que estão listados;
- citing_paper_count: Número de artigos citando o artigo fornecido;
- citing_patent_count: Número de patentes que citam o artigo fornecido;
- conference_dates: Data do evento da conferência. O formato da data não é padronizado e varia de acordo com a conferência;
- conference_location: Local do evento da conferência (pode ser uma ou mais cidades, estados, países);
- content_type: Especifica o tipo de publicação ao qual o conteúdo pertence. Os tipos de conteúdo podem incluir: Livros / ebooks, Conferências, Cursos, Acesso Antecipado, Revistas, Revistas Científicas, Normas;
- start_record: Um espaço reservado dentro do processo de entrega para várias interações, buscando grandes volumes de dados;
- d_au: Open Author Facet;
- doi: O Identificador de Objeto Digital (DOI) é o identificador único atribuído a um artigo/documento pela CrossRef;
- d_publisher: Open Publisher's Facet;
- d_pubtype: Faceta de Tipo de Conteúdo Aberto;
- d_year: Open Publication Facet;
- end_page: O número da última página na versão impressa do artigo;
- facet: Open/Facet dimension;
- full_name: Nome completo do autor;
- html_url: URL do IEEE Xplore que retornará o HTML do texto completo;
- ieee_terms: Palavras-chave atribuídas a artigos de periódicos e papers de conferências da IEEE a partir de um vocabulário controlado criado pela IEEE;
- index_terms: Esta é uma área combinada que retorna Palavras-chave do Autor e Termos IEEE;

- insert_date: Data da última atualização;
- is_number: Identificador interno (apenas para periódicos);
- isbn: Número Padrão Internacional de Livro. Um número usado para identificar de forma única um livro ou uma obra não-serializada;
- issn: Número Internacional Normalizado para Publicações Seriadas. Um número de 8 dígitos utilizado para identificar de forma única uma publicação periódica (revista ou série);
- issue: Número da edição da revista na qual o artigo foi publicado;
- pdf_url: URL do IEEE Xplore que retorna o PDF em texto completo;
- publication_date: Data de publicação do artigo; essa informação pode ser representada como data completa, mês e ano, ou trimestre e ano. O formato varia de acordo com a publicação;
- publication_year: Ano em que o artigo foi publicado; esses dados podem ser representados como data completa, mês e ano, ou trimestre e ano. O formato varia de acordo com a publicação;
- publication_number: Um número de registro único do IEEE atribuído a uma publicação;
- publication_title: Título de uma publicação
- publisher: Nome da editora da publicação específica mencionada;
- rank: Indica a hierarquia dos documentos retornados com base em um algoritmo da IEEE utilizado;
- standard_number: Designação norma (por exemplo, IEEE 802.11u-2011). As normas padrão são atribuídas pelo Administrador do Comitê de Novos Padrões da IEEE-SA Standards Board (NesCom);
- standard_status: Status da norma (por exemplo, Ativo, Rascunho, etc.); aplica-se apenas a Normas;
- start_page: Número da página inicial na versão impressa do artigo;
- title: Título de um documento;
- totalfound: Número total de documentos encontrados que correspondem aos critérios da consulta;
- totalsearched: Número total de documentos pesquisados;

- volume: Detalhe associado apenas a Revistas e Conferências.

Os dados dos artigos foram extraídos em formato JSON através de uma API disponibilizada pelo próprio instituto. As consultas foram feitas utilizando filtros de palavras-chave no resumo e no título dos trabalhos, a fim de se obter trabalhos relacionados às áreas pertinentes à IA. As palavras-chave selecionadas foram:

- *artificial intelligence*;
- *machine learning*;
- *complex networks*;
- *big data*;
- *neural networks*;
- *computer vision*;
- *natural language processing*

4.2 Preparação da Base

Após a extração, os dados foram armazenados em uma tabela para facilitar o seu preprocessamento. Os campos nulos foram preenchidos com elementos textuais vazios. Utilizando a biblioteca NLKT foi criada função *preprocess_text* que remove pontuações, converte o texto para minúsculas, realiza a tokenização do texto, remove as chamadas *"stopwords"* (palavras comuns que geralmente são removidas em etapas de pré-processamento) e, por fim, aplica o *stemming* às palavras, reduzindo-as ao seu radical. Essa função foi aplicada ao título, resumo e palavras-chave de cada trabalho.

Como as informações dos autores estavam aninhadas em dicionários contento dados como ID, url, nome completo, afiliação, dentre outras informações, foi feito um tratamento para extrair os nomes completos dos autores de cada trabalho, transformado todos os caracteres em letras minúsculas.

4.3 Construção da Rede de Coautoria

A rede de coautoria foi construída utilizando a biblioteca networkx, uma biblioteca Python para construção e análise de redes. Para a construção da rede, iterou-se sobre as linhas da tabela extraindo a lista de autores de cada artigo. Cada autor da lista é então adicionado como um nó no grafo. Para cada par de autores que colaboraram no mesmo artigo (combinatória de 2 autores da lista), adiciona-se uma aresta entre eles e atribui-se um peso às arestas. Se a aresta já existe, o peso é incrementado; caso contrário, uma

nova aresta é criada com peso 1. Assim, o peso nas arestas pode representar o número de colaborações entre os autores. Essa representação de rede é valiosa para análises mais avançadas, como identificação de autores influentes, comunidades de coautoria, entre outros.

Algoritmo 1: Construir Grafo de Coautoria

```
Passo 1: Criar um grafo vazio
G = GrafoVazio()
Passo 2: Adicionar nós e arestas ao grafo
    Para Cada Artigo em df:
        Extrair informações dos autores
        autores = Artigo['authors_list']
        Adicionar nós ao grafo
        Para Cada Autor em autores:
            G.adicionarNo(Autor)
            Adicionar arestas entre autores que colaboraram no mesmo artigo
            Para Cada Par de Autores em Combinações(autores, 2):
                Se G.tiverAresta(Par):
                    Incrementar o peso da aresta se ela já existir
                    G.incrementarPesoAresta(Par)
                Senão:
                    Criar uma nova aresta com peso 1 se ela não existir
                    G.adicionarAresta(Par, Peso=1)
```

Retorne G

4.4 Medidas de Centralidade

Com o gráfico de coautoria construído, foram calculas as medidas de centralidade de Katz, centralidade de grau, centralidade de proximidade e centralidade de autovetor, a partir das quais foram construídos rankings para mostrar os autores mais influentes de acordo coma rede de coautoria para cada metodologia aplicada.

4.5 Análise de Tendências e Padrões

A fim de se avaliar o conteúdo dos trabalhos para extrair *insights* sobre tendências e padrões, foi feita uma análise das palavras-chave e do resumo já pré-processado através da construção de nuvens de palavras. O código analisa a frequência das palavras no texto processado e cria uma nuvem de palavras para visualizar as palavras-chave mais comuns. A nuvem de palavras é uma representação visual em que o tamanho das palavras é proporcional à sua frequência no texto. Essa análise pode fornecer *insights* sobre os tópicos ou temas mais recorrentes nos trabalhos.

5 AVALIAÇÃO EXPERIMENTAL

5.1 Conhecendo o Conjunto de Dados

A consulta feita à API do IEEE retornou um conjunto de 1400 trabalhos nas áreas de engenharia elétrica e eletrônica relacionados com Inteligência Artificial através das palavras-chave listadas no Capítulo 4. Nestes 1400 trabalhos, é possível encontrar informações de 3591 autores. Como a pesquisa não foi restringida a um período específico, tem-se trabalhos publicados de 1983 a 2023. A distribuição dos trabalho por ano de publicação pode ser visualizada na figura 3.

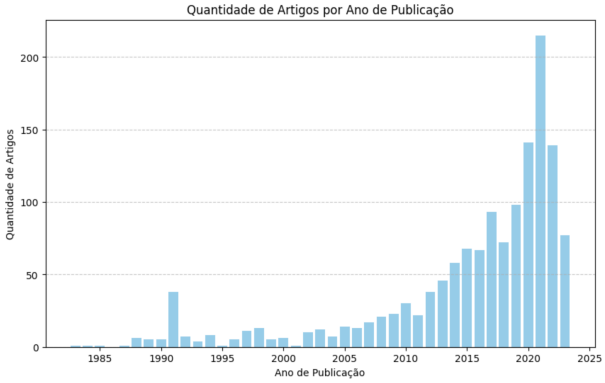

Figura 3 – Quantidade de trabalhos por ano de publicação. Fonte: Autor.

5.2 Rede de Coautoria

A rede de coautoria construída pode ser visualizada em 2D na figura 4. Os *labels* dos nós foram omitidos para facilitar a visualização do grafo. A rede contém 3591 nós e 4989 arestas. A aresta de menor peso tem peso 2, e a de maior peso, tem peso 21.

5.3 Medidas de Centralidade e Definição dos Autores Influentes

A tabela 1 traz a comparação de todos os autores que compõem o ranking dos 10 auotres mais influentes identificados por cada medida de centralidade e sua posição em cada um dos rankings das demais métricas.

5.3.1 Centralidade de Katz

A centralidade de Katz foi calculada com $beta = 0,1$. Aplicando os valores calculados para cada autor, é possível redesenhar a rede com uma escala de cores marcando os autores com maiores medidas em tons mais escuros. O grafo com a medida de centralidade pode ser visualizado na figura 5.

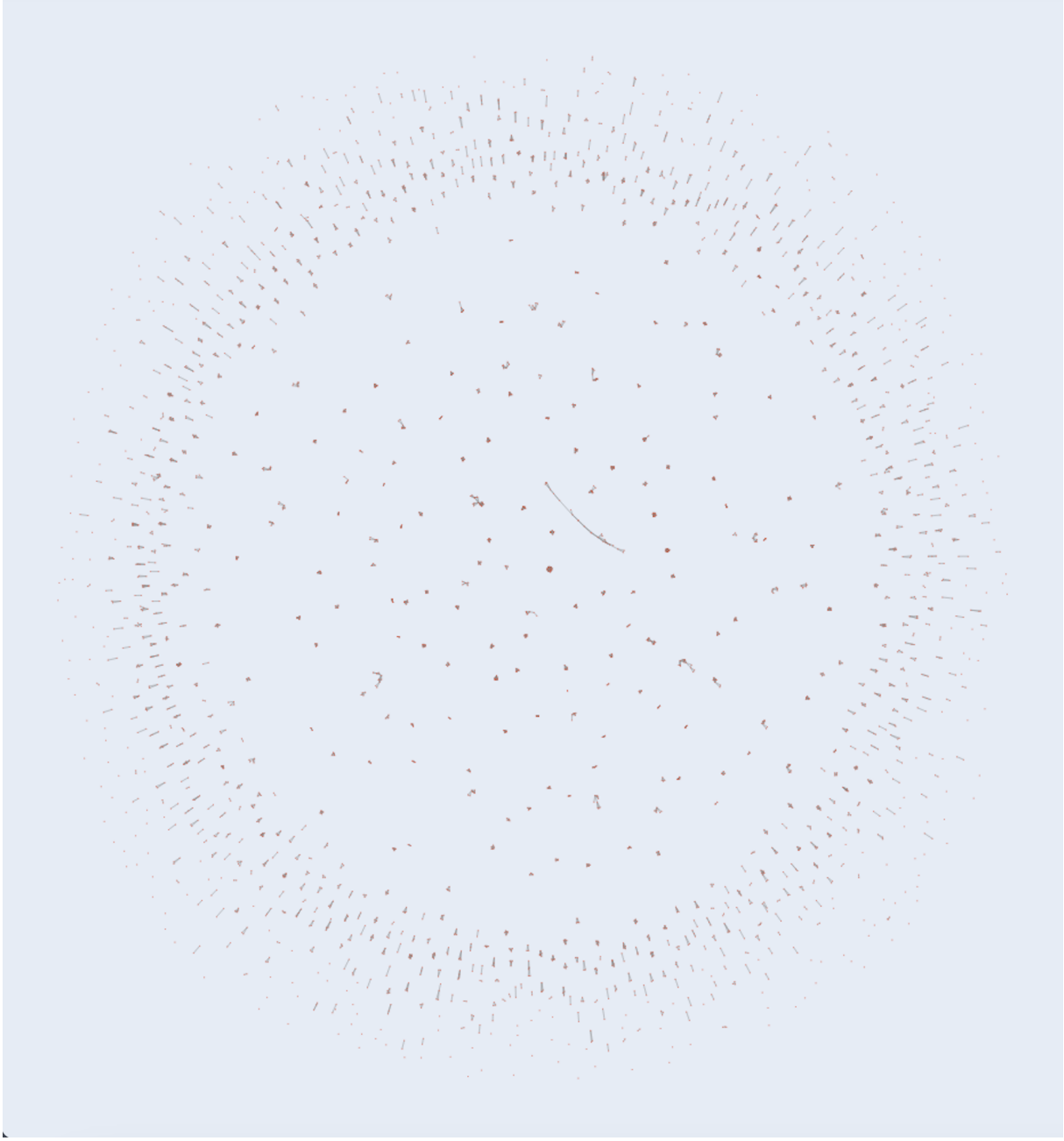

Figura 4 – Rede de coautoria. Fonte: Autor.

A partir da medida de centralidade, foi possível definir os 10 autores mais influentes, conforme listado na tabela 2.

5.3.2 Centralidade de Grau

Aplicando os valores calculados para cada autor, é possível redesenhar a rede com uma escala de cores marcando os autores com maiores medidas em tons mais escuros. O grafo com a medida de centralidade pode ser visualizado na figura 6.

A partir da medida de de centralidade, foi possível definir os 10 autores mais

Autor	Centralidade de Katz	Centralidade de Grau	Centralidade de Proximidade	Centralidade de Auto-vetor
markus koch	1	32	44	31
christoph balada	2	35	47	34
sheraz ahmed	3	36	48	35
stefan agne	4	37	49	36
andreas dengel	5	38	50	37
max bondorf	6	31	43	30
markus zdrallek	7	34	46	33
fabian karl	8	39	51	38
ulrik dietzler	9	40	52	39
markus krampf	10	41	53	40
hirotaka murakami	32	1	1	1
eric bohannon	16	2	2	2
john childs	33	3	3	3
grace gui	40	4	4	4
eric moule	34	5	5	5
katsuhiko hanzawa	35	6	6	6
tomofumi koda	37	7	7	7
chiaki takano	38	8	8	8
toshimasa shimizu	39	9	9	9
yuki takizawa	26	10	10	10

Tabela 1 – Autores influentes e sua posição no ranking de cada uma das medidas

influentes, conforme listado na tabela 3.

5.3.3 Centralidade de Proximidade

Aplicando os valores calculados para cada autor, é possível redesenhar a rede com uma escala de cores marcando os autores com maiores medidas em tons mais escuros. O grafo com a medida de centralidade pode ser visualizado na figura 7.

A partir da medida de centralidade, foi possível definir os 10 autores mais influentes, conforme listado na tabela 4.

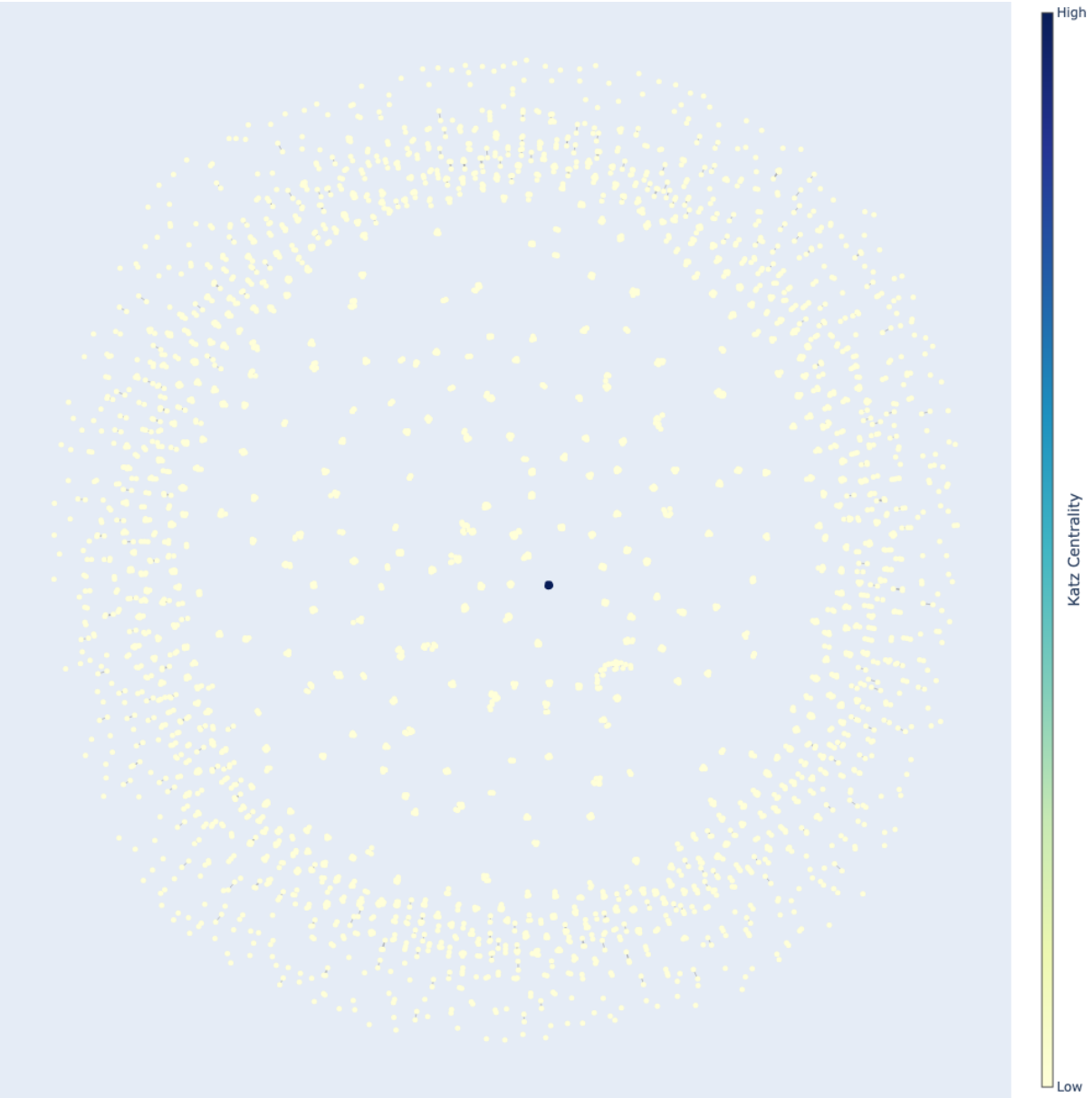

Figura 5 – Rede de coautoria com os pesos da medida de centralidade de Katz por nó (autor). Fonte: Autor.

5.3.4 Centralidade de Autovetor

Aplicando os valores calculados para cada autor, é possível redesenhar a rede com uma escala de cores marcando os autores com maiores medidas em tons mais escuros. O grafo com a medida de centralidade pode ser visualizado na figura 8.

A partir da medida de proximidade definir os 10 autores mais influentes, conforme listado na tabela 5.

Autor	Centralidade de Katz
markus koch	0.3015113445777637
christoph balada	0.30151134457776363
sheraz ahmed	0.30151134457776363
stefan agne	0.30151134457776363
andreas dengel	0.30151134457776363
max bondorf	0.3015113445777636
markus zdrallek	0.3015113445777636
fabian karl	0.3015113445777636
ulrik dietzler	0.3015113445777636
markus krampf	0.3015113445777636

Tabela 2 – 10 autores com maior Centralidade de Katz

Autor	Medida de Centralidade de Grau
hirotaka murakami	0.004456824512534819
eric bohannon	0.004456824512534819
john childs	0.004456824512534819
grace gui	0.004456824512534819
eric moule	0.004456824512534819
katsuhiko hanzawa	0.004456824512534819
tomofumi koda	0.004456824512534819
chiaki takano	0.004456824512534819
toshimasa shimizu	0.004456824512534819
yuki takizawa	0.004456824512534819

Tabela 3 – 10 autores com maior Centralidade de Grau

5.4 Tendências e Padrões

Nas figuras 9 e 10 são apresentadas as nuvens de palavras feitas a partir das palavras-chave e dos resumos dos artigos, respectivamente.

A análise das 100 palavras mais frequentes nos resumos e palavras-chave, as que mais se destacam na nuvem de palavras, traz a percepção de pontos significativos sobre as principais tendências e tópicos abordados nas pesquisas relacionadas à Inteligência

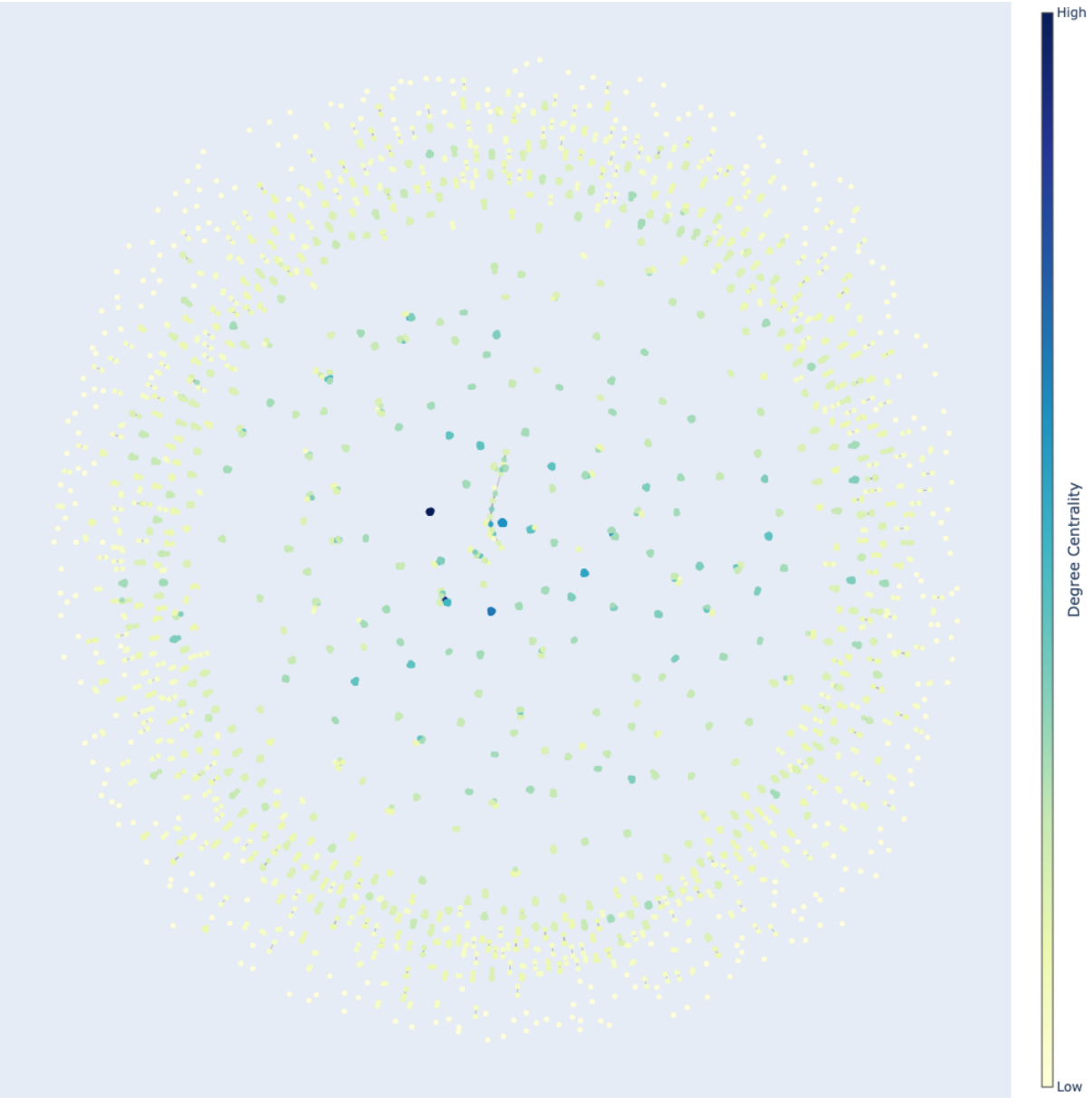

Figura 6 – Rede de coautoria com os pesos da medida de centralidade de grau por nó (autor). Fonte: Autor.

Artificial (IA) no campo da engenharia elétrica e eletrônica.

As palavras *"data"* e *"learning"* indicam uma ênfase marcante em métodos de aprendizado de máquina e destacam a importância crucial dos dados nessas pesquisas. A presença recorrente de termos como *"networks"*, *"neural"*, *"network"* e *"algorithm"* aponta para um foco substancial em redes neurais e algoritmos no contexto da IA.

A inclusão de *"computer"* e *"technology"* destaca a estreita relação entre IA e tecnologia da informação. Além disso, termos como *"processing"e "image"* indicam uma associação robusta com processamento de imagem, enquanto *"natural"* e *"language"* revelam uma forte presença em pesquisas de processamento de linguagem natural (NLP).

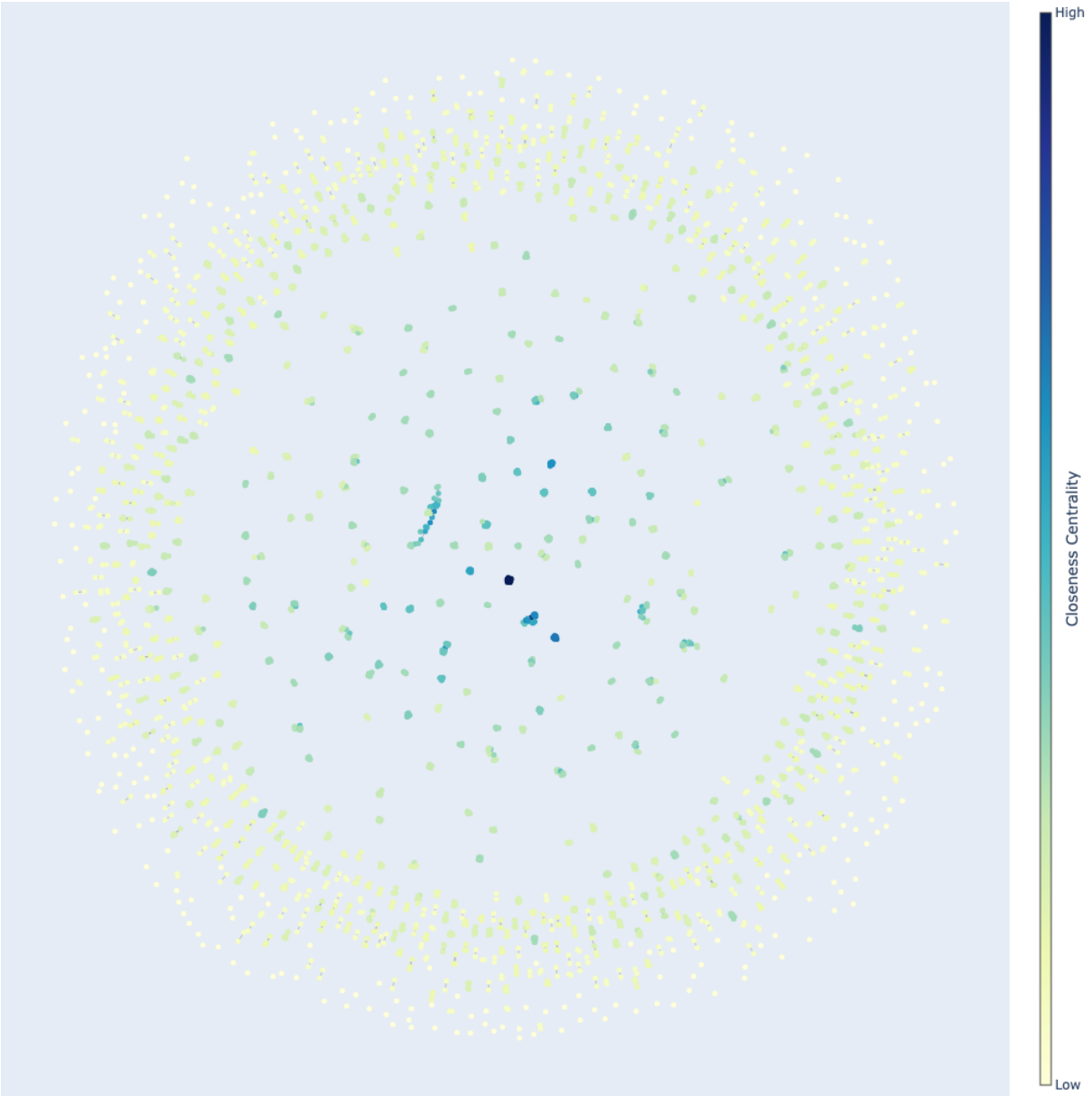

Figura 7 – Rede de coautoria com os pesos da medida de centralidade de proximidade por nó (autor). Fonte: Autor.

"Deep", *"cloud"*, *"predictive"*, *"privacy"*, *"internet"*, *"biological"*, *"quantum"*, *"fuzzy"*, *"mobile"*, *"topology"*, *"robotics"*, *"automation"* destacam áreas emergentes e tendências, como aprendizado profundo (*deep learning*), computação em nuvem, previsão, privacidade, internet das coisas, biologia computacional, computação quântica, robótica e automação.

A presença de palavras como *"challenges"*, *"security"* e *"ethics"* sugere que os pesquisadores estão atentos aos desafios e questões éticas associados à IA.

Autor	Medida de Centralidade de Proximidade
hirotaka murakami	0.004456824512534819
eric bohannon	0.004456824512534819
john childs	0.004456824512534819
grace gui	0.004456824512534819
eric moule	0.004456824512534819
katsuhiko hanzawa	0.004456824512534819
tomofumi koda	0.004456824512534819
chiaki takano	0.004456824512534819
toshimasa shimizu	0.004456824512534819
yuki takizawa	0.004456824512534819

Tabela 4 – 10 autores com maior Centralidade de Proximidade

Autor	Medida de Centralidade de Autovetor
hirotaka murakami	0.2425353056014203
eric bohannon	0.24253530560142034
john childs	0.24253530560142034
grace gui	0.24253530560142034
eric moule	0.24253530560142034
katsuhiko hanzawa	0.24253530560142034
tomofumi koda	0.24253530560142034
chiaki takano	0.24253530560142034
toshimasa shimizu	0.24253530560142034
yuki takizawa	0.24253530560142034

Tabela 5 – 10 autores com maior Centralidade de Autovetor

5.5 Resultados e Discussões

A análise da rede de coautoria fornece insights sobre a colaboração entre autores na área de engenharia elétrica e eletrônica relacionada à Inteligência Artificial. Diferentes medidas de centralidade identificaram diferentes conjuntos de autores influentes, indicando diferentes aspectos de sua importância na rede, mas ainda assim é possível notar nomes em comum nesses diferentes cenários. Autores como Hirotaka Murakami, Eric Bohannon, e John Childs são consistentemente destacados como influentes em quase todas as medidas. A

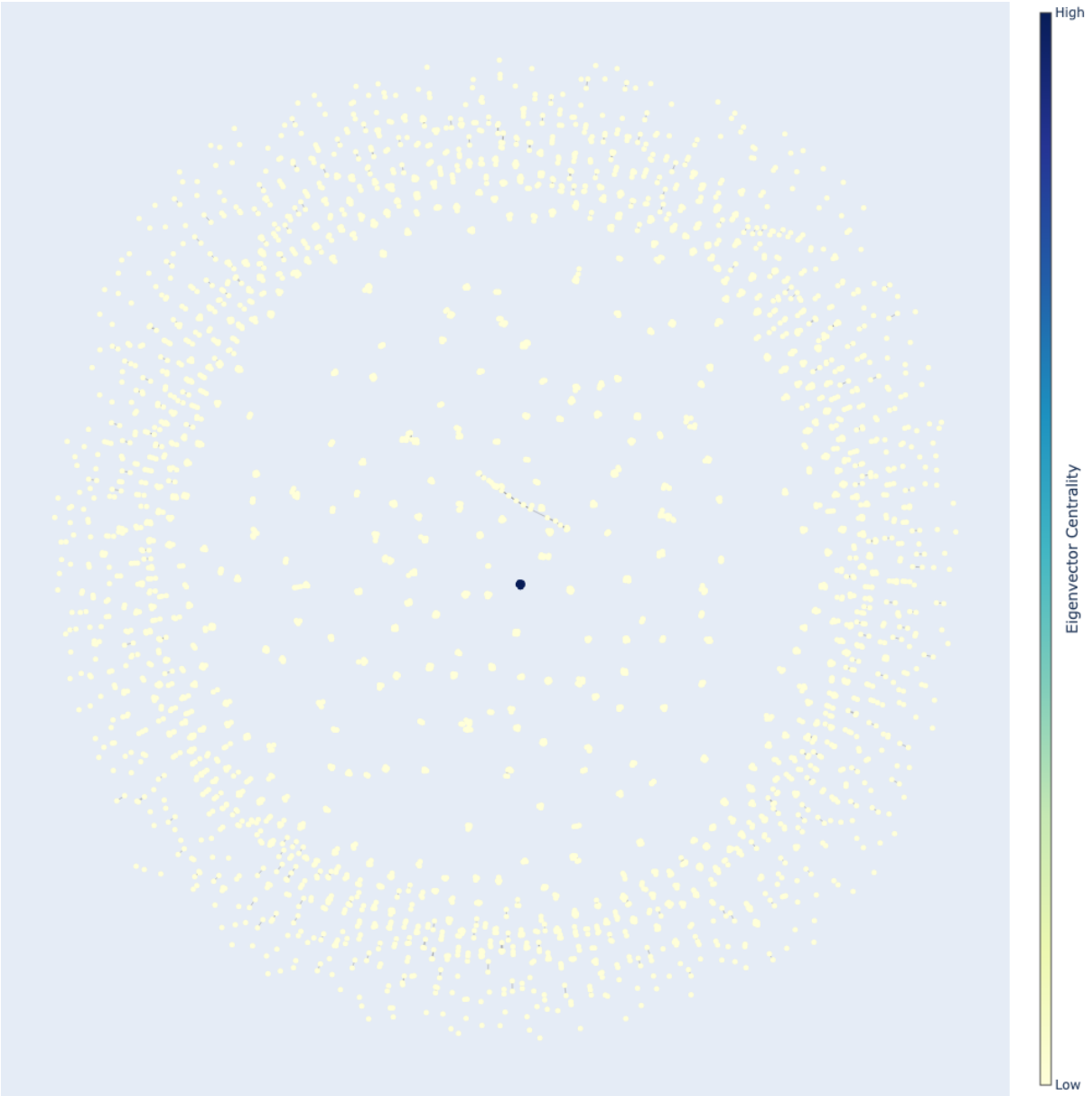

Figura 8 – Rede de coautoria com os pesos da medida de centralidade de proximidade por nó (autor). Fonte: Autor.

consistência nas listas sugere que esses autores não são apenas centralizados em termos de conexões (centralidade de grau), mas também têm importância em termos de proximidade, autovetor e a medida de Katz.

Cada medida de centralidade oferece uma perspectiva única sobre a importância dos autores na rede. A centralidade de grau destaca autores com muitas colaborações diretas. A centralidade de proximidade destaca aqueles que estão mais próximos de outros na rede. A centralidade de autovetor destaca autores que estão conectados a outros autores importantes. A centralidade de Katz considera não apenas as conexões diretas, mas também as conexões indiretas, atribuindo importância a autores conectados a outros

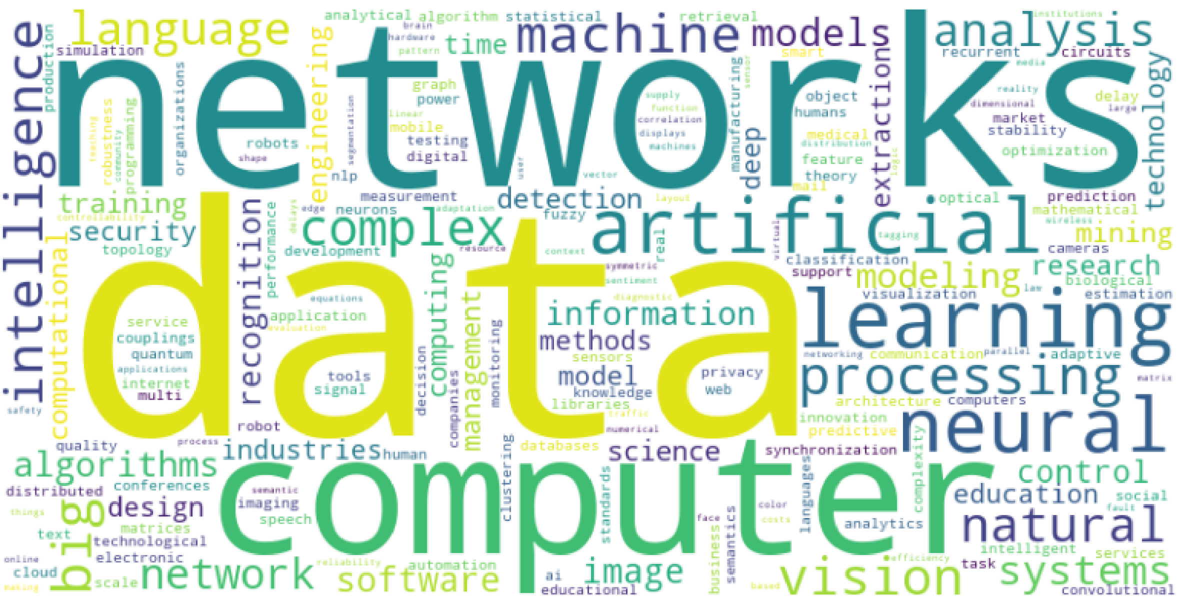

Figura 9 – Nuvem de palavras a partir das palavras-chave dos artigos. Fonte: Autor.

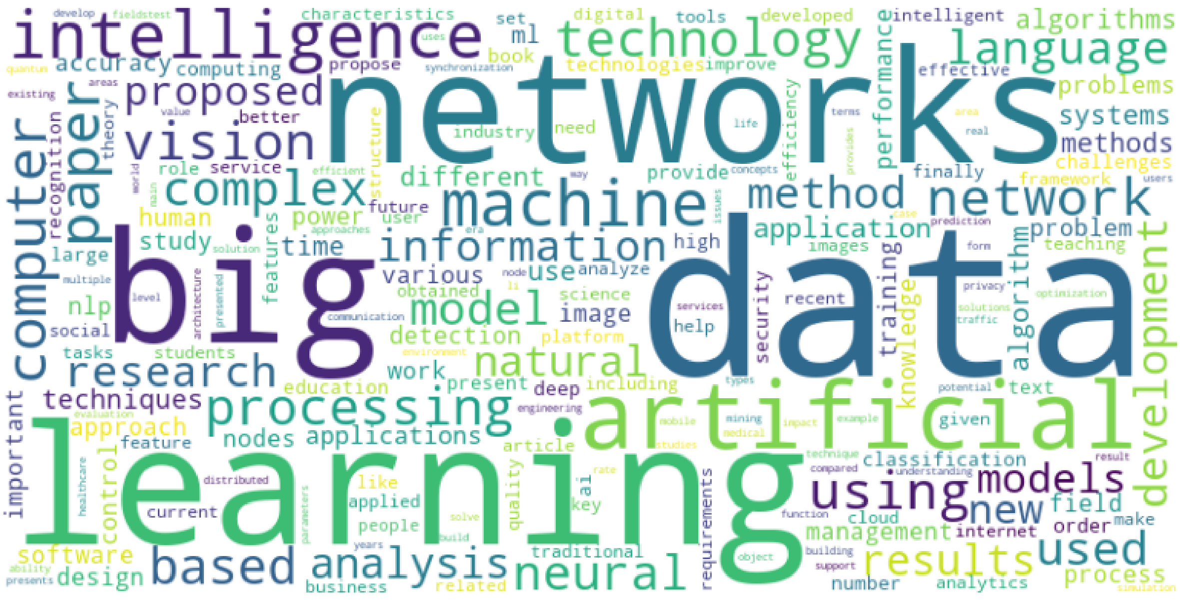

Figura 10 – Nuvem de palavras a partir do resumo pré-processado dos artigos. Fonte: Autor.

autores importantes.

O experimento realizado forneceu uma visão geral do cenário de colaboração entre autores no conjunto de dados analisado, destacando a diversidade de métodos utilizados para identificar a influência dos autores na rede.

6 CONCLUSÕES

A análise da colaboração entre autores na área de engenharia elétrica e eletrônica, especificamente relacionada à Inteligência Artificial, fornece *insights* valiosos ao examinar a rede de coautoria. Diferentes métricas de centralidade revelaram conjuntos distintos de autores influentes, destacando diferentes aspectos de sua importância na rede. No entanto, é notável que alguns nomes permaneçam consistentemente proeminentes em várias métricas, como Hirotaka Murakami, Eric Bohannon e John Childs. Isso sugere que esses autores não apenas possuem conexões significativas (centralidade de grau) mas também desempenham papéis importantes em termos de proximidade, autovetor e a medida de Katz.

Cada métrica de centralidade oferece uma perspectiva única sobre a relevância dos autores na rede. A centralidade de grau destaca aqueles com muitas colaborações diretas, enquanto a centralidade de proximidade destaca aqueles mais próximos de outros na rede. Já a centralidade de autovetor destaca autores conectados a outros influentes, enquanto a centralidade de Katz considera tanto conexões diretas quanto indiretas, atribuindo importância aos autores conectados a outros autores de destaque.

Devido a limitação de tempo para a realização desta trabalho de conclusão de curso, optou-se por explorar a rede de coautoria para realizar a identificação de autores influentes, mas, para trabalhos futuros, o estudo pode ser repetido usando uma rede de citações. A análise de uma rede de citação pode revelar padrões, tendências e relações entre diferentes áreas de pesquisa. Os pesquisadores podem usar essas informações para entender a evolução de um campo, identificar lacunas no conhecimento ou explorar a influência de determinados trabalhos. Uma rede de citação será uma rede mais conectada e as medidas farão sentido em um contexto maior. Por exemplo, o autor será importante não apenas dentro de uma comunidade de coautores, mas dentro de uma subárea. O mesmo ocorrerá se usar ainda uma rede de coautoria, porém, de apenas uma subárea. A análise de uma rede de citação pode revelar padrões, tendências e relações entre diferentes áreas de pesquisa. Os pesquisadores podem usar essas informações para entender a evolução de um campo, identificar lacunas no conhecimento ou explorar a influência de determinados trabalhos.

Como sugestões de aprimoramento para futuros trabalhos, destaca-se ainda explorar a utilização de múltiplas bases de dados ou repositórios para validar e complementar os resultados obtidos. Além disso, em vez de apenas extrair os nomes completos dos autores, considerar a normalização de nomes para lidar com variações, como abreviações diferentes ou representações distintas dos nomes. No contexto da construção da rede de coautoria, a inclusão de informações temporais poderia ser considerada para explorar a evolução das colaborações ao longo do tempo.

Outras melhorias propostas incluem a identificação de hubs na rede, ou seja, autores com muitas conexões, e a aplicação de algoritmos de detecção de comunidades para encontrar grupos altamente conectados. Além disso, a divisão da rede em sub-redes correspondentes a diferentes subáreas de IA, como machine learning, visão computacional e processamento de linguagem natural, poderia proporcionar insights mais específicos sobre dinâmicas de colaboração. A análise da evolução temporal das comunidades e hubs também seria valiosa para identificar tendências e mudanças na colaboração ao longo do tempo, enriquecendo a compreensão da dinâmica da rede de coautoria.

REFERÊNCIAS

ALEMáN, Y.; LOYA, N.; VILARIñO, D. Two methodologies applied to the author profiling task. *In*: . [*S.l.: s.n.*], 2013.

AMANCIO, D. R. Classificação de textos com redes complexas. **Instituto de Física de São Carlos, USP**, 10 2013.

AMANCIO, D. R. *et al.* Complex networks analysis of language complexity. **Europhysics Letters**, EDP Sciences, IOP Publishing and Società Italiana di Fisica, v. 100, n. 5, p. 58002, dec 2012. Available at: <https://dx.doi.org/10.1209/0295-5075/100/58002>.

ANTIQUEIRA, L. *et al.* A complex network approach to text summarization. **Information Sciences**, v. 179, n. 5, p. 584–599, 2009. ISSN 0020-0255. Special Section - Quantum Structures: Theory and Applications. Available at: <https://www.sciencedirect.com/science/article/pii/S0020025508004520>.

AYALA, D. V. *et al.* Lexical-syntactic and graph-based features for authorship verification notebook for pan at clef 2013. *In*: **Conference and Labs of the Evaluation Forum**. [*S.l.: s.n.*], 2013.

BARBOSA, L. M. **Um modelo para extrair conhecimento de artigos científicos utilizando redes complexas**. 2016. Tese (Doutorado) — [sn], 2016.

CASTILLO, E. *et al.* Author attribution using a graph based representation. *In*: **2015 International Conference on Electronics, Communications and Computers (CONIELECOMP)**. [*S.l.: s.n.*], 2015. p. 135–142.

IRFAN, R. *et al.* A survey on text mining in social networks. **The Knowledge Engineering Review**, Cambridge University Press, v. 30, n. 2, p. 157–170, 2015.

JUAREZ OFELIA CERVANTES VILLAGóMEZ, D. V. A. E. C. Text analysis using different graph-based representations. **Computación y Sistemas**, v. 21, n. 4, p. 581–599, 2017.

LARANJEIRA, P. A.; CAVIQUE, L. Métricas de centralidade em redes sociais. 2018.

MACKE, S. Deep sentence-level authorship attribution. *In*: . [*S.l.: s.n.*], 2015.

MELO, R. D. R. AplicaÇÃo da teoria das redes complexas e da anÁlise de redes sociais para avaliaÇÃo da produÇÃo intelectual cientÍfica dos programas de pÓs-graduaÇÃo stricto sensu acadÊmicos em engenharia de produÇÃo 2013-2015. *In*: **Dissertação (Programa de Pós-Graduação STRICTO SENSU em Engenharia de Produção e Sistemas) - Pontifícia Universidade Católica de Goiás, Goiânia-GO.** [*S.l.: s.n.*], 2017.

METZ, J. *et al.* Redes complexas: conceitos e aplicações. 2007.

OSMAN, A. H.; BARUKUB, O. M. Graph-based text representation and matching: A review of the state of the art and future challenges. **IEEE Access**, v. 8, p. 87562–87583, 2020.

RANGEL, F.; ROSSO, P. On the impact of emotions on author profiling. **Information Processing Management**, v. 52, n. 1, p. 73–92, 2016. ISSN 0306-4573. Emotion and Sentiment in Social and Expressive Media. Available at: <https://www.sciencedirect.com/science/article/pii/S0306457315000783>.

RONQUI, J. R. F. **Medidas de centralidade em redes complexas: correlações, efetividade e caracterização de sistemas**. 2014. Tese (Doutorado) — Universidade de São Paulo, 2014.

SILVA BRENO JÚNIO V E MOTTA, R. e. d. A. L. A. da. Detecçao de autores duplicados utilizando estrutura de comunidades em redes de cooperaçao cientıfica. 2011.

SILVA, T.; MACHADO, C. J. R.; MACIEL, A. Identificação de possíveis influenciadores digitais em fóruns de discussão em cursos a distância. **Revista de Engenharia e Pesquisa Aplicada**, v. 5, p. 117–123, 04 2020.

STAMATATOS, E. *et al.* Overview of the author identification task at pan 2014. **CEUR Workshop Proceedings**, v. 1180, p. 877–897, 01 2014.