A APLICAÇÃO DE REDES DE SIMILARIDADE NA CONSTRUÇÃO DO CONHECIMENTO

THE APPLICATION OF SIMILARITY NETWORKS IN KNOWLEDGE CONSTRUCTION

LA APLICACIÓN DE REDES DE SIMILITUD EN LA CONSTRUCCIÓN DEL CONOCIMIENTO

Recebido em: 27/09/2023. Aprovado em: 14/01/2024

Eneida Santana[1]
Tereza Kelly Carneiro[2]
Roberto Monteiro[3]

RESUMO: O artigo apresenta um Modelo de Redes de Similaridades (MRS) que se fundamenta em ciclos de conexões estruturadas por atributos. A metodologia adotada foi a da teoria de redes e incorporaram-se os princípios das medidas para avaliar as similaridades entre os atores da rede gerada a partir de dados do Índice de Desenvolvimento da Educação Básica (IDEB) . Através de uma revisão sistemática de literatura (RSL), este trabalho cataloga algumas medidas e métodos utilizados na análise dessas redes. Como resultado foi possível a aplicação do MRS para identificação Clusters em uma rede de unidades escolares, utilizando a métrica de coeficiente de clustering para identificação de similaridades, contribuindo, assim, para a compreensão dos métodos e métricas aplicáveis na criação e análise de redes de similaridade em contextos científicos.

Palavras-chave: Redes de Similaridade; Medidas de Similaridade; Modelagem do Conhecimento.

ABSTRACT: The article presents a Similarity Network Model (MRS) that is based on cycles of connections structured by attributes. The methodology adopted was network theory and the principles of measurements were incorporated to evaluate the similarities between the actors in the network generated from data from the Basic Education Development Index (IDEB). Through a systematic literature review (RSL), this work catalogs some measures and methods used in the analysis of these networks. As a result,

[1] Doutora em Difusão do Conhecimento. Bibliotecária-documentalista no Instituto Federal da Bahia. Email: eneida@ifba.edu.br
[2] Doutora em Difusão do Conhecimento . Professora do Instituto Federal de Educação, Ciência e Tecnologia da Bahia -Campus Camaçari. Email: terezakelly1@gmail.com
[3] Doutor em Difusão do Conhecimento . Atualmente é professor da Universidade do Estado da Bahia e do Centro Universitário SENAI CIMATEC. Email: robertolsmonteiro@gmail.com

it was possible to apply MRS to identify clusters in a network of school units, using the clustering coefficient metric to identify similarities, thus contributing to the understanding of methods and metrics applicable in the creation and analysis of similarity networks. in scientific contexts.

Keywords: Similarity Networks; Similarity Measures; Knowledge Modeling

RESUMEN: El artículo presenta un Modelo de Red de Similitud (MRS) que se basa en ciclos de conexiones estructuradas por atributos. La metodología adoptada fue la teoría de redes y se incorporaron principios de mediciones para evaluar las similitudes entre los actores de la red generados a partir de datos del Índice de Desarrollo de la Educación Básica (IDEB). A través de una revisión sistemática de la literatura (RSL), este trabajo cataloga algunas medidas y métodos utilizados en el análisis de estas redes. Como resultado, fue posible aplicar MRS para identificar conglomerados en una red de unidades escolares, utilizando la métrica del coeficiente de agrupamiento para identificar similitudes, contribuyendo así a la comprensión de métodos y métricas aplicables en la creación y análisis de redes de similitud en el ámbito científico. contextos.
Palabras clave: Redes de Similitud; Medidas de Similitud; Modelado del Conocimiento.

1. Introdução

Este artigo tem por objetivo apresentar as possibilidades de aplicação de redes de similaridade para estudos no campo da construção do conhecimento, identificados em estudos de teses e dissertações implicadas, entre os anos de 2012 a 2019. Além de apresentar a criação de um MRS conceitual e empírico baseado em ciclos de conexão de similares por atributos.

A discussão sobre o conceito de similaridade na teoria do conhecimento traz à cena autores como Lin (1998, p.296) que identifica três intuições principais sobre similaridade, que são: (a) a similaridade é proporcional às correlações que dois objetos compartilham; (b) a similaridade é inversamente proporcional às diferenças entre os objetos; e (c) dois objetos são maximamente similares quando são idênticos, independentemente de quantas correlações eles possam compartilhar.

Nesta perspectiva Lin (1998) propõe que a similaridade entre dois objetos possa ser medida em termos da informação que eles compartilham e das informações que são distintas para cada um. Para fazer isso, ele utiliza a teoria da informação, particularmente entropia e informação mútua(IM).

Autores como Jeon e Park (2016) e Yao et al (2022), acrescentam que em redes onde vértices representam entidades e arestas representam suas relações ou co-

ocorrências, através das informações entre elas é possível medir a similaridade entre entidades. Barabási (2009) também indica que a teoria das redes possibilita medir as informações compartilhadas entre dois vértices ou grupos.

Compreender as pesquisas que tenham a identificação de similaridades em diferentes vieses permite que cientistas e pesquisadores identifiquem padrões e lacunas de conhecimento e desenvolvam estratégias eficazes para a transmissão e absorção de novas informações, promovendo assim a colaboração efetiva.

Portanto, motivados pelo questionamento de quais medidas de similaridade são empregadas no estudo de redes de afinidade ou similaridade na área de pesquisa da construção do conhecimento, realizamos uma revisão sistemática da literatura e, identificando uma lacuna na aplicação dessas métricas para a análise de dados governamentais abertos, desenvolvemos um modelo de redes de similaridade (MRS) para processamento de redes de unidades escolares similares da microrregião de Salvador (Bahia), no ano de 2019.

2. Teoria de Redes e Métricas de Similaridades

Para darmos início à incursão na teoria de redes é necessário revisitar os princípios da teoria dos grafos, uma vez que os grafos constituem representações matemáticas das redes.

A teoria dos grafos foi desenvolvida através dos trabalhos do matemático suíço Leonard Euler, que em 1736, mostrou que a presença de um caminho é uma característica intrínseca do grafo, através da sua teoria sobre as pontes de Königsberg. Por definição, um grafo G é formado por dois conjuntos: V, que é finito e não vazio, e E, que representa uma relação binária entre os elementos de V (Gross e Yellen 1999, p. 25). Nesse contexto, os elementos de V são chamados de vértices ou nós, e os de E são chamados de arestas, sendo que cada aresta está vinculada a um ou dois vértices.

O avanço na elaboração de algoritmos, muitas vezes ocorreu concomitantemente com os avanços dos estudos sobre a teoria de grafos, o que demonstra que a expansão da teoria deve muito a essa sinergia. Com isso, percebemos que as redes complexas surgem como derivações dessa interação dual entre grafos e algoritmos.

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

Com o avanço de seus estudos, a teoria de grafos evoluiu para abranger os grafos de conhecimento. Estes possuem a propriedade de aprimorar os detalhes das entidades de um grafo e extrair relações significativas que podem potencialmente enriquecer as respectivas explanações. Wang *et al*. (2019) argumentam que "uma das razões para isso é que a caracterização da conectividade usuário-item é atingida de maneira bastante indireta".

De acordo com Barabási (2009), os grafos são representados como redes, e possuem propriedades que são fundamentais para entendermos os cenários e situações possíveis no mundo real.

Os aspectos e agentes sociais que integram o mundo são elementos possíveis de serem representados através das redes compreendidas como redes sociais.

As redes sociais, de acordo com Newman (2010), são estruturas de conexão onde os vértices representam indivíduos ou agrupamentos de pessoas e as arestas simbolizam algum tipo de interação ou relação social entre eles. A construção terminológica para os estudos que permeiam as redes sociais é abordada pelos estudos de Moreno (1954). Dada a complexidade da sua estrutura, a rede é considerada um sistema.

Apresentada a questão, é importante ressaltar que a teoria de redes, conforme argumenta Newman (2010), tem como primeiro objetivo encontrar e destacar propriedades estatísticas que caracterizam a estrutura e o comportamento de sistemas baseados em rede e sugerir meios adequados para medir essas propriedades.

Ainda para Newman (2010), o segundo objetivo da teoria de redes é a criação de modelos de redes que favoreçam a compreensão do significado dessas propriedades. E o terceiro objetivo, teria como finalidade predizer qual será o comportamento de sistemas baseados em rede, considerando as propriedades estruturais medidas e as regras locais que vigoram sobre as partes do sistema.

Deste modo, a compreensão das propriedades das medidas de uma rede é essencial para compreensão das estruturas de grafos construídas a partir das redes e as principais medidas para a interpretação de uma rede complexa. Estas medidas estão divididas em medidas descritivas e são estatísticas que caracterizam a estrutura e o comportamento dessas redes, são elas: a) grau médio (<k>): refere-se ao número de conexões que, em média, os vértices da rede possuem; b) densidade (Δ): expressa o quão perto o grafo está de se tornar completo; c) diâmetro (D): representa o caminho mais longo de todos os

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

caminhos mais curtos calculados entre dois vértices; d) caminho mínimo médio (L): denota a distância média do caminho entre dois vértices da rede. Já as medidas de aglomeração são usadas para quantificar a tendência de formação de grupos ou comunidades dentro da rede, representadas pela coeficiente de aglomeração (C) que estabelece se um vértice está conectado a dois outros vértices, (C) indica a probabilidade desses dois vértices também estarem conectados e a Modularidade a mensuração da formação de comunidades na rede. Já as medidas de centralidade são métodos para identificar os vértices mais importantes dentro da rede, são elas: a) centralidade de grau: avalia o quanto um vértice está conectado aos outros vértices da rede; b) centralidade de proximidade: o vértice é considerado mais próximo quanto menor é o caminho que precisa percorrer para alcançar outros vértices da rede e c) centralidade de intermediação: o vértice é considerado mais intermediário quando ocorre entre muitos caminhos entre outros vértices da rede.

As medidas apresentadas correspondem a leitura geral de uma rede complexa, no entanto redes com características específicas como uma rede de similaridades necessitam de medidas específicas. Por exemplo, Easley e Kleinberg (2010) indicam que a similaridade é uma medida de quão semelhantes dois nós são em uma rede, com base em suas conexões ou atributos. E por isso, segundo esses autores a ideia de similaridade entre nós em uma rede é o princípio de que a conexão implica semelhança. Logo, se dois nós estão ligados a muitos nós em comum, ou seja, se eles têm muitos vizinhos em comum, então eles mesmos são semelhantes aos demais.

3. Redes de similaridades: revisão sistemática da literatura

Para aprofundar a compreensão sobre o tema optamos pela metodologia de Revisão Sistemática da Literatura (RSL), guiados pela pergunta: "Quais medidas de similaridade são empregadas no estudo de redes de afinidade ou similaridade no campo de estudo da construção do conhecimento?". Esta questão serviu de norte para a nossa investigação e, a fim de conduzi-la de maneira organizada e rigorosa, optamos por seguir o protocolo PRISMA (Prisma, 2020).

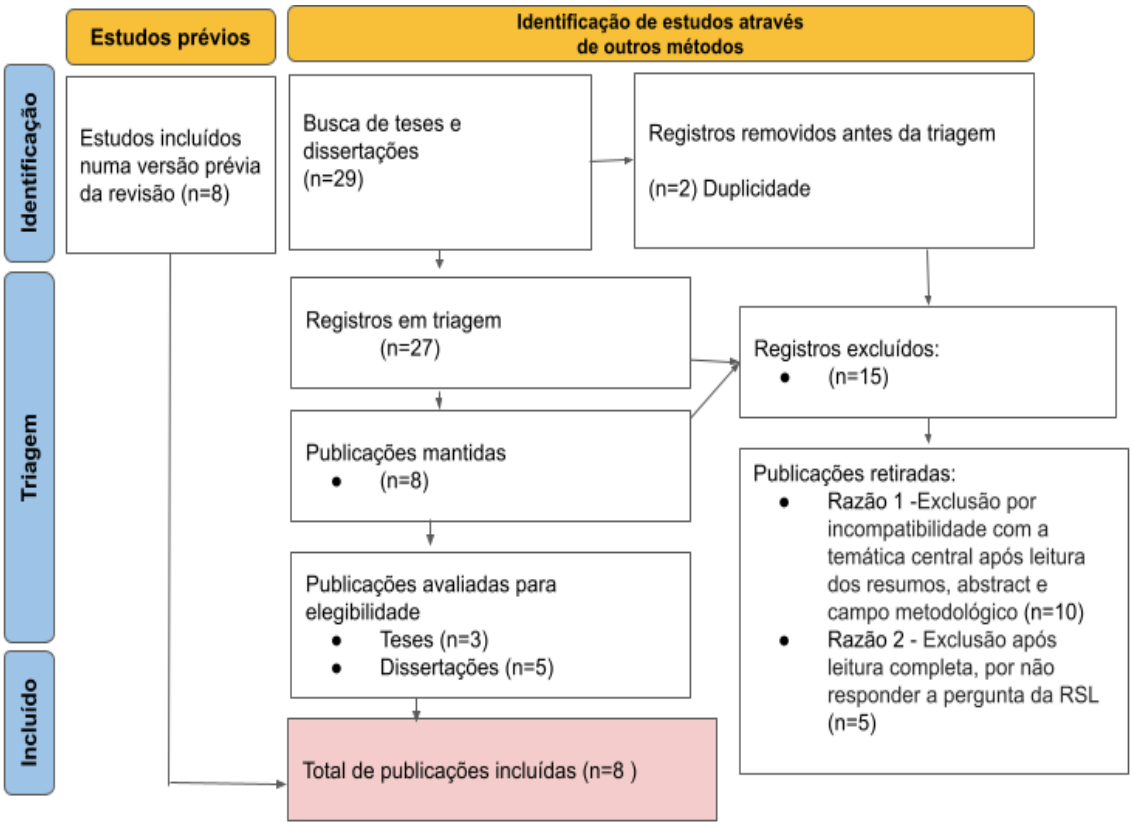
Como estratégia de pesquisa, decidimos focar nossa RSL nas teses e dissertações publicadas entre 2000 e 2023, todas indexadas na Biblioteca Digital Brasileira de Teses

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

e Dissertações. E estabelecemos que os termos de busca empregados na recuperação seriam: “redes de similaridades”; “medidas de similaridade”; “redes de afinidade”; “índice de similaridade”. A opção por esses termos foi realizada tendo em vista a busca por entender, a partir de outros estudos, como a aplicação de medidas de similaridade vêm sendo utilizadas por outros pesquisadores em diferentes propostas de pesquisas.

O protocolo PRISMA nos levou à seleção de 8 estudos relevantes para a nossa Revisão Sistemática da Literatura, como pode ser visualizado na Figura 1:

Figura 1 - Fluxograma para revisões sistemáticas que incluam buscas de teses e dissertações.



Fonte: Modelo Prisma 2020 - Adaptado pelos autores

A seleção final dos trabalhos para essa revisão consiste em 3 teses de doutorado e 5 dissertações de mestrado, cujos detalhes são apresentados no Quadro 1. É relevante salientar que as teses incluídas nesta revisão foram publicadas sob os auspícios do Programa de Pós-Graduação em Difusão do Conhecimento (PPGDC). Estes trabalhos

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

representam esforços de pesquisa centrados em objetivos relacionados à construção, difusão e gestão do conhecimento.

Quadro 1 - Detalhamento das teses e dissertações utilizadas na RSL

Título	**Autoria/ Instituição**	**Ano**	**Documento**
Um modelo evolutivo para simulação de redes de afinidade	Roberto Luiz Souza Monteiro/UFBA	2012	Tese
Redes de afinidade como estratégia de gestão pedagógica e difusão do conhecimento em cursos na modalidade a distância	Tereza Kelly Gomes Carneiro/UFBA	2014	Tese
Estudo de padrões em sinais musicais sob a perspectiva dos grafos de visibilidade	Dirceu de Freitas Piedade Melo/UFBA	2017	Tese
Refinamento multinível em redes complexas baseado em similaridade de vizinhança	Alan Demetrius Baria Valejo/USP	2014	Dissertação
Técnicas e algoritmos de Link Analysis na geração de medidas de similaridade	Rodrigo Carvalho Rezende/Unicamp	2012	Dissertação
Predição de links em redes complexas utilizando informações de estruturas de comunidades	Jorge Carlos Valverde Rebaza/USP	2013	Dissertação
Similaridade semântica de atributos para dados em nuvem: um estudo de caso no MIDAS	Witã dos Santos Rocha/UFBA	2019	Dissertação
Predição temporal de links baseada na evolução de tríades	Hugo Neiva de MELO/UFPE	2016	Dissertação

Fonte: Autoria própria

Para responder à pergunta da RSL sobre medidas de similaridade empregadas nos estudos de redes de similaridade ou afinidade, identificamos nas teses e dissertações quais foram a abordagem adotada e as medidas de similaridade implementadas pelos autores.

Ao analisarmos o estudo de Monteiro (2012), observamos que o autor implementa uma abordagem de análise da estrutura topológica de Arranjos Produtivos Locais (APLs), buscando promover a difusão do conhecimento e reforçar a competitividade das empresas

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

integrantes dos APLs. Neste processo, o autor recolhe dados relacionados às relações estabelecidas entre as empresas do APL em questão, determina as características das topologias das redes sociais formadas, e concebe uma solução computacional que possibilita a visualização e a previsão da evolução dessas redes ao longo do tempo. Monteiro (2012) utiliza como métricas para a caracterização dessas redes as medidas de: densidade, coeficiente de aglomeração, caminho mínimo médio, distribuição de graus, centralidade e eficiência. Entre essas métricas, o autor destaca a importância da eficiência como métrica de análise topológica, validando a sua aplicabilidade no contexto do estudo.

Com base na análise dessas métricas e na observação das dinâmicas existentes entre as empresas, o autor consegue identificar a topologia da rede, permitindo a realização de um diagnóstico detalhado de sua estrutura e processo evolutivo. Neste contexto, uma das contribuições centrais do estudo de Monteiro (2012) é o desenvolvimento de um modelo de redes de afinidade, concebido a partir do conceito de semelhança ou "afinidade" representado entre as respostas das empresas pesquisadas.

Já a tese defendida por Carneiro (2014), também realizado no âmbito do PPGDC, é um modelo de gestão pedagógica para cursos de formação profissional à distância. Este modelo utiliza as redes de afinidade criadas a partir do perfil tecnológico dos estudantes. O conceito de similaridade ou afinidade é aqui determinado pelas habilidades individuais no uso do computador.

Para estruturar e analisar essas redes de afinidade, a autora emprega uma variedade de medidas que incluem: a centralidade de grau, a centralidade de proximidade, e a centralidade de intermediação. Além disso, Carneiro também utiliza a medida de eficiência global, que avalia a eficiência de comunicação em toda a rede, e a eficiência local, que foca na eficiência de comunicação dentro dos subconjuntos de nós. O coeficiente de aglomeração é outra medida empregada, que se refere à tendência de nós, no caso do estudo de Carneiro (2014), são os estudantes de uma rede e como eles se agrupam.

Ainda no seguimento dos estudos orientados pelo PPGDC, analisamos a tese defendida por Melo (2018), onde o autor introduz uma metodologia para identificar padrões de autossimilaridade em sinais de áudio, usando propriedades topológicas de redes, conhecida como Descritor de Visibilidade em Flutuações de Variância (DVFV).

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

Este descritor é composto por: modularidade, número de comunidades, grau médio e densidade (Delta).

No contexto do estudo de Melo (2018), o conceito de similaridade é abordado por meio da análise de autossimilaridade em sinais de áudio. Essa autossimilaridade se refere à medida em que os padrões dentro de um sinal de áudio se repetem ou são semelhantes a si mesmos ao longo do tempo.

Nos estudos que resultaram em sua dissertação de mestrado, Rezende (2012) explora técnicas de *Link Analysis* (Análise de Link); é uma técnica de análise que se concentra em conexões e relacionamentos em um conjunto de dados, para calcular a similaridade entre artigos acadêmicos organizados em uma biblioteca digital. O autor investiga técnicas de amostragem de grafos, avaliando a qualidade das amostras geradas por estes métodos. Uma nova abordagem de amostragem baseada na técnica *Forest Fire* é proposta e, através de experimentos, demonstra-se a superioridade deste novo algoritmo.

Rezende (2012) também propõe uma meta-função, apresentando como conceito de similaridade os artigos acadêmicos que consideram apenas as informações de citação entre os artigos, sem levar em conta o conteúdo textual e seus metadados. Esta meta-função transforma medidas de similaridade locais, como o coeficiente Jaccard e assortividade, em medidas recursivas. A ideia é de que dois artigos são mais similares à medida que estão associados a artigos que também são similares. No estudo, são apresentadas as principais medidas de similaridade encontradas na literatura, incluindo SimRank, Katz, Adamic/Adar, Weighted Paths, Common Neighbors, coeficiente Jaccard e assortividade. A dissertação apresentada por Rebaza (2013) converge com os estudos de Rezende (2012), ambos os autores adotam a abordagem da *Link Analysis*, além de apresentarem as mesmas medidas para identificar a similaridade.

O estudo de Rebaza (2013) apresenta duas propostas alinhadas aos métodos que se baseiam nas informações das comunidades para a predição de links. A primeira proposta apresenta um novo índice de similaridade que utiliza as informações dos vértices pertencentes à mesma comunidade na vizinhança de um par de vértices analisados, assim como as informações dos vértices pertencentes a diferentes comunidades nessa mesma vizinhança. A segunda proposta consiste em um conjunto de índices, desenvolvidos a partir da reformulação de algumas propostas já existentes, porém, inserindo nessas

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

informações dos vértices pertencentes unicamente à mesma comunidade na vizinhança topológica de um par de vértices analisados.

Já o autor Melo (2016) realiza um estudo direcionado para a ARS no contexto virtual, concentrando-se particularmente na tarefa de predição de links. Este estudo se insere na crescente área de inteligência artificial e ARS, onde se busca descobrir padrões estruturais, semelhanças entre indivíduos, e dados estatísticos, para entender e prever a formação de conexões em uma rede. O conceito de similaridade, utilizado por Melo (2016), é baseado na similaridade entre os nós i e j, calculada considerando a quantidade de informação que descreve as semelhanças entre i e j e a quantidade de informação necessária para descrever i e j separadamente. Essa medida serve de base para o cálculo de métricas que representam padrões de maneira mensurável, valorando o grau de similaridade ou proximidade entre dois indivíduos dentro da rede. A metodologia proposta por Melo faz uso do algoritmo SimRank para calcular a semelhança entre vértices na rede. Esse algoritmo baseia-se na ideia de que dois nós têm uma maior probabilidade de se relacionarem se seus vizinhos compartilham características comuns.

Para encerrar a RSL desta seção, recorremos à leitura da dissertação de Rocha (2019), que propõe uma solução para facilitar o acesso a dados distribuídos em diferentes níveis de serviços, como Data-as-a-Service (DaaS) e Database-as-a-Service (DBaaS), através de uma camada intermediária denominada *Middleware for Interoperability Between SaaS and DaaS* (MIDAS). O estudo aborda o desafio da ambiguidade dos dados em ambientes de nuvem e propõe uma abordagem automatizada para resolver esse problema. A metodologia envolve a implementação do modelo no MIDAS, testando critérios como sobrecarga, desempenho e correção, por meio de simulações em um ambiente com vários algoritmos de medida de distância e parâmetros de provedores de DaaS.

O autor utiliza o conceito de similaridade ou afinidade para definir o contexto da manutenção da confiabilidade das solicitações de dados originais, mesmo com atualizações dos parâmetros dos níveis de serviços DaaS. Duas avaliações de similaridade são propostas: a contagem de arestas (Cosseno e Jaccard) para medir a similaridade entre dois parâmetros, e o Information Content (IC) para medir a similaridade com base no conhecimento dos parâmetros no corpus da *WordNet*.

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

Após a identificação das métricas de similaridades apresentadas pelos estudos que compõem a RSL que compreendemos irão colaborar com a proposta deste estudo, passamos a detalhar as métricas de análise mais presentes no campo da construção do conhecimento, as quais também irão contribuir com a discussão que se pretende desenvolver. São elas:

(1) **SimRank:** Esta é uma medida de similaridade que compara a similaridade dos vizinhos de dois nós. A fórmula do SimRank para dois nós i e j é dada por:

$$S(i,j) = \frac{C}{N(i).N(j)} \sum \square\, a \in N(i) \sum \square\, b \in N(j)\, S\,(a,b)$$

onde:

- i e j são nós distintos
- $S(i,j)$ representa a similaridade entre dois nós i e j.
- C é uma constante que diminui a contribuição dos vizinhos distantes na similaridade.
- N(i) e N(j) são os números de vizinhos dos nós i e j, respectivamente.
- Σ é o somatório, a e b são todos os pares de vizinhos de i e j, respectivamente.

(2) **Coeficiente de Clustering (Clustering Coefficient):** É uma medida que indica a probabilidade de quanto os vizinhos de um nó estão conectados entre si. A fórmula é:

$$C_{(i)} = \frac{2 - E(i)}{K(i).\,(K(i) - 1)}$$

onde:

- i é um nó.
- O coeficiente de clustering de um nó específico *i* é representado por $C_{(i)}$
- K(i) é o grau do nó *i* , que indica o número de conexões que o nó possui.
- E(i) é o número de triângulos formados pelos vizinhos do nó *i*.

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

(3) **Centralidade de Proximidade (Closeness Centrality)**: Esta é uma medida da proximidade média de um nó para todos os outros nós na rede. É calculada como o inverso da soma das distâncias mais curtas de um nó para todos os outros.

$$C(i) = \frac{1}{\Sigma_{j \neq i} d(i,j)}$$

onde:

- i e j são nós distintos
- A centralidade de proximidade para um nó específico i é representada por C(i).
- d (i,j) representa a distância geodésica i e j.

(4) **Centralidade de Intermediação (Betweenness Centrality):** Esta medida quantifica o número de vezes que um nó atua como elo mais curto entre dois outros nós. A fórmula é:

$$C_B(i) = \Sigma_{j \neq i \neq k} \frac{g(j,k|i)}{g(j,k)}$$

onde:

- i, j e k são nós distintos.
- A centralidade de intermediação para um nó i é representada por $C_B(i)$.
- g(j,k) é a função que retorna o número de caminhos mínimos (geodésicos) entre os nós j e k.
- g(j,k|i) representa o número de caminhos mínimos entre os nós j e k que passam pelo nó i.

(5) **Eficiência Local/Global**: Eficiência é uma medida de quão eficientemente a informação é trocada através de um nó (local) ou da rede (global).

(a) A eficiência global (E_{global}) de uma rede é uma medida que avalia a rapidez com que a informação se propaga entre todos os pares de nós na rede.

$$E_{global} = \frac{1}{N(N-1)} \Sigma_{i \neq j} \frac{1}{d(i,j)}$$

onde:

- i ej são nós distintos.

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

- N é o número total de nós (ou vértices) na rede.
- d(i,j) representa a distância geodésica entre os nós i e j.

(b) A eficiência local (E_{local}) entre dois nós específicos em uma rede é uma medida que avalia quão eficientemente esses nós trocam informações.

$$E_{local}(i) = \frac{1}{N-1}\Sigma_{i\neq j}\frac{1}{d(i,j)}$$

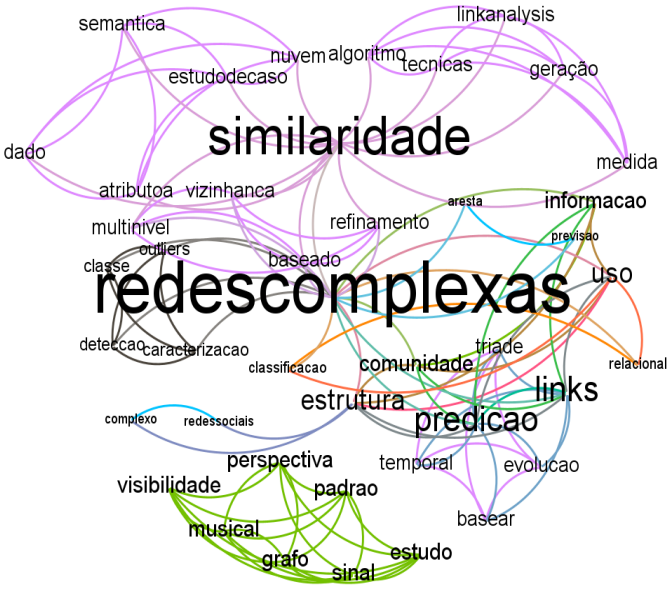
onde:

- i ej são nós distintos.
- N é o número total de nós (ou vértices) na rede.
- d(i,j) representa a distância geodésica entre os nós i e j.

Durante o processo de análise das teses e dissertações, além da identificação das medidas, realizamos o levantamento das palavras-chaves que mostravam relação com os conceitos de similaridade apresentados pelos estudos. A RSL permitiu-nos estabelecer uma relação intrínseca entre os temas abordados e as palavras-chave usadas na pesquisa.

Com base nessas observações, desenvolvemos um grafo de conhecimento semântico. Esse recurso permitiu a ampliação da visualização das principais temáticas associadas aos estudos, resultando na Figura 2:

Figura 2 – Rede Semântica das palavras-chaves das teses e dissertações da RSL



Fonte: Autoria própria

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

Nela, são destacados termos como "predição de links", "redes complexas", "comunidade", "estrutura", "informação" e "uso". Esses termos emergem como elementos centrais das pesquisas, estabelecendo ligações fortes e significativas com o conceito geral de similaridade nos estudos em questão.

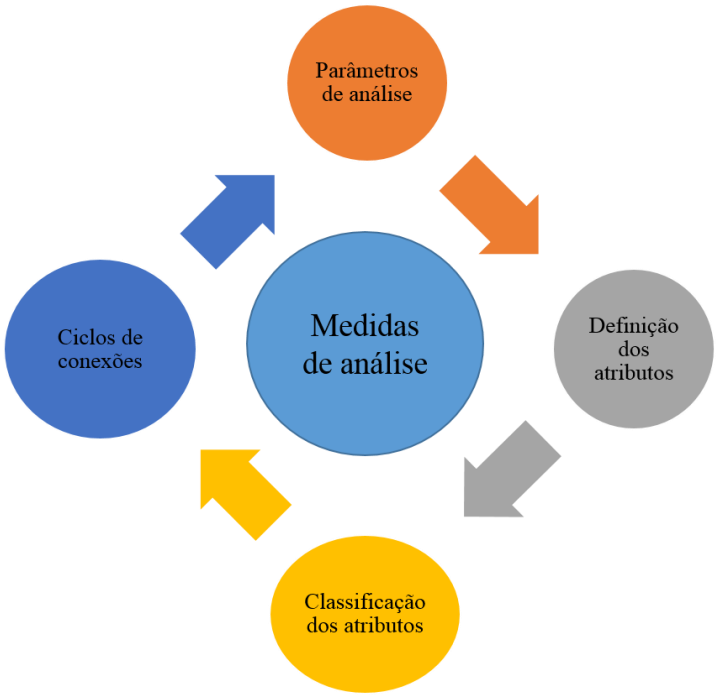
Após a revisão e as definições apresentadas nesta seção sobre teoria das redes e medidas de similaridade, necessárias ao entendimento e interpretação dos resultados obtidos das redes criadas para este estudo. Na próxima seção, apresentaremos a aplicação prática da teoria discutida, a partir da visualização das medidas em um modelo real de rede de similaridades.

4. Modelo de Redes de Similaridade (MRS)

O MRS para a criação de redes de similaridade é um processo sistemático que envolve a construção de ciclos para o estabelecimento das conexões das informações que irão compor o campo da similaridade entre os atores de uma rede. A figura 3 detalha os passos para aplicação de um MRS:

Figura 3 - Fluxo conceitual do MRS

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719



Fonte: Autoria própria

A primeira etapa envolve o mapeamento inicial dos atores ou elementos que comporão a rede, definido como parâmetros de análise. Nesta fase serão definidos quem são os vértices que irão compor o conjunto de atores desta rede , e é estabelecida a dimensão da rede a ser analisada. Os atores podem ser indivíduos, instituições, publicações ou qualquer outra entidade que possa ser caracterizada por um conjunto de atributos. Em seguida, os atributos que serão usados para medir as similaridades entre os atores são definidos. Os atributos podem variar desde características demográficas e sociais até métricas mais complexas, como o impacto científico. Na etapa seguinte, os atributos definidos são então classificados com base em sua relevância e aplicabilidade ao modelo. A classificação pode ser feita por meio de técnicas estatísticas ou de aprendizado de máquina para determinar o peso de cada atributo na análise subsequente. Adotamos o peso das arestas ou criação de código para identificação dos atributos.

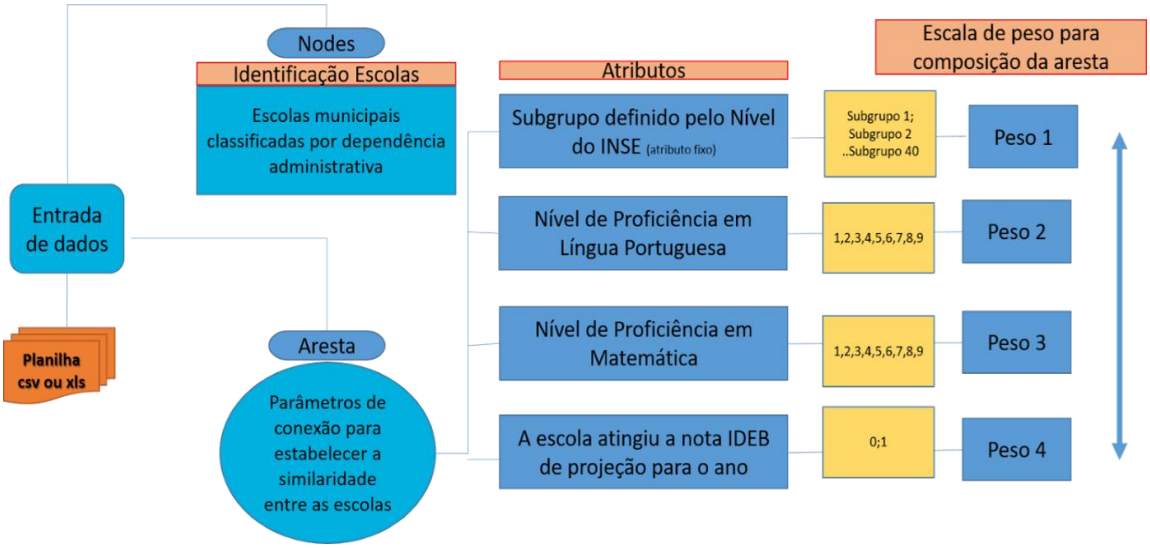
Por fim, é iniciada a etapa da criação dos ciclos de conexões, é nesta fase onde o modelo proposto é aplicado para criar a rede de similaridades. Utilizando os atributos e suas respectivas classificações, os ciclos de conexões são formados. Esses ciclos

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

permitem a criação de clusters dentro da rede, onde atores com características similares são agrupados.

Como exemplo de aplicação, apresentamos na figura 4, o MRS utilizado para identificar Perfis de Unidades Escolares Similares:

Figura 4 - MRS em aplicação para identificação de Perfis de Unidades Escolares Similares



Fonte: Autoria própria

Nesta aplicação os ciclos de conexões são etapas subsequentes que refinam a formação de redes de similaridade. Primeiro: estabelece a conexão entre as escolas que compõem o mesmo subgrupo; Segundo: estabelece a conexão entre as escolas que não pertencem ao mesmo subgrupo, mas estão enquadradas no mesmo nível de proficiência em Língua Portuguesa e/ou Matemática e entre as escolas que atingiram a nota Índice de Desenvolvimento da Educação Básica (IDEB) correspondente à projeção para aquele ano; e por fim, na terceira etapa, atribui peso à aresta de acordo com o número de atributos similares entre as escolas. Após essas três etapas, inicia-se a fase de criação e visualização do grafo de uma rede não dirigida que representa a rede de unidades escolares similares.

5. Aplicação do MRS na identificação de *Clusters* em Redes de Unidades Escolares Similares

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

Neste estudo, o ponto de partida para a aplicação do MRS foi a seleção do coeficiente de clustering como a métrica de similaridade adotada. Isso foi feito para identificar perfis de unidades escolares semelhantes, com a capacidade de representar *clusters* em uma rede de escolas que abrange diferentes subgrupos.

O conceito de *cluster* utilizado neste estudo, fundamentado com base no estudo de Pereira (2019), que identifica seis características centrais de um *cluster*: proximidade geográfica, cooperação e competição, compartilhamento de valores e conhecimento, conexões e redes, dinamismo e a presença de atores de apoio.

Essa estrutura conceitual, trazida para o contexto das redes de unidades escolares similares, apoia o conceito de *cluster* relacionado a redes de escolas com perfil de similaridade, como um agrupamento de escolas que, devido à sua semelhança geográfica, cultural ou pedagógica, compartilham práticas e valores educacionais, além de problemas sociais derivados do campo e do hábito em que se constituiu a unidade escolar.

Neste estudo, os clusters, bem como seus subgrupos e inter-relações, representam a estrutura subjacente que guia a formação e a interação dessas escolas. Através da análise da rede de similaridade, é possível identificar os "hubs" que seriam as escolas centrais ou conjuntos de escolas que servem como principais impulsionadores ou núcleos de formação de um cluster. Esses hubs, no contexto educacional, simbolizam instituições que, de alguma forma, exercem uma influência significativa ou desempenham um papel de liderança dentro de seu *cluster*, moldando práticas e promovendo a colaboração.

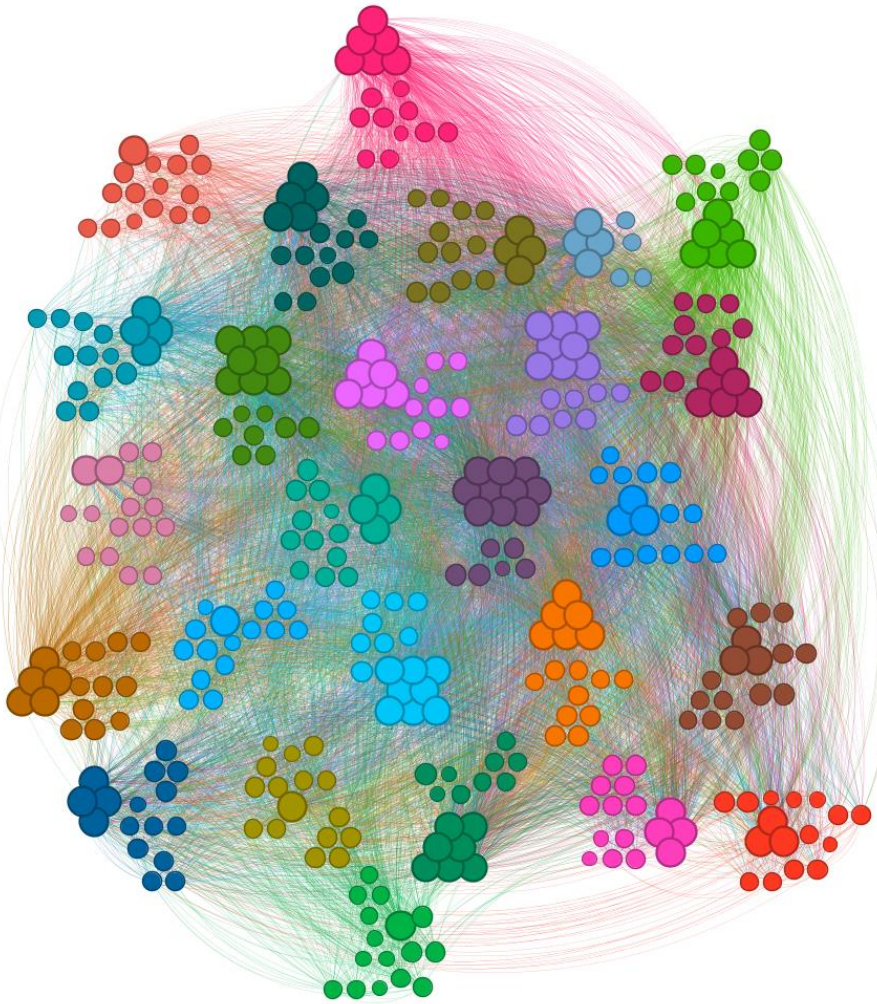
Na perspectiva teórica da aplicação do pensamento de Castells (1999) à análise das redes de unidades escolas similares, os *cluster* são categorizados como cluster do conhecimento, que pode ser entendido como um agrupamento geográfico de escolas que compartilham conhecimentos e práticas similares, facilitando a disseminação de informações e recursos entre elas. Nesse sentido, os hubs podem ser vistos como escolas dentro desses clusters que desempenham um papel central na distribuição de informações e na conexão com outras escolas.

O coeficiente de clustering é uma métrica que mede a densidade das conexões em um grupo de escolas, ou seja, o quanto essas escolas estão interconectadas. Um coeficiente de clustering elevado indica que as escolas estão fortemente conectadas entre si e compartilham muitas características comuns, ou seja ampla similaridade.

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

Neste estudo, a métrica é aplicada à rede de escolas similares localizadas na microrregião de Salvador, Bahia. Todas as instituições educacionais estão situadas na região urbana e oferecem à comunidade cursos voltados para as séries iniciais do ensino fundamental. A representação visual dessa rede, conforme apresentado na figura 5, é composta por 408 nós, cada um representando uma unidade escolar, totalizando 14.481 conexões distribuídas em 26 subgrupos distintos.

Figura 5 - Rede de Escolas Similares da Microrregião de Salvador – Bahia – Localização: Urbana - Ensino Fundamental - Séries Iniciais (2019)

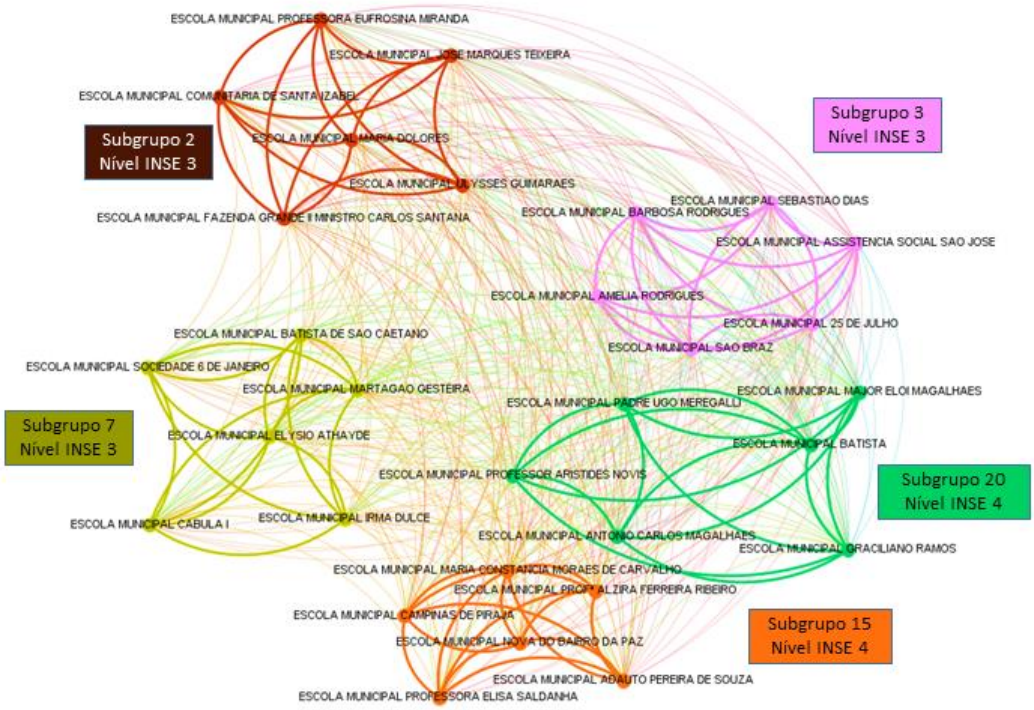


Fonte: Autoria própria

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

Ao calcular o coeficiente de clustering na rede representada pela figura 5, observou-se que aproximadamente 7,85% das unidades escolares formam *clusters* distintos. Esses *clusters* estão distribuídos em cinco dos subgrupos da rede, totalizando 30 unidades escolares que serão consideradas como hubs da rede, conforme ilustrado na figura 6.

Figura 6 - Identificação de unidades escolares com maior Coeficiente de Clustering na Rede de unidades escolares Similares da Microrregião de Salvador – Bahia – Localização: Urbana - Ensino Fundamental - Séries Iniciais (2019)



Fonte: Autoria própria

A presença de subgrupos de unidades escolares hubs com um coeficiente de clustering elevado indica que essas unidades escolas têm um alto grau de similaridade em termos de características, práticas e resultados educacionais. Estas unidades escolas desempenham um papel central na rede, servindo como pontos de referência para outras unidades. O fato de que esses subgrupos de unidades escolares hubs estão inseridos em

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

diferentes níveis do Índice de Nível Socioeconômico (INSE) sugere que, independentemente de seu desempenho educacional, essas unidades têm uma capacidade significativa de influenciar outras unidades escolares em sua rede. Isso pode ser atribuído à sua posição central na rede, à sua capacidade de compartilhar práticas bem-sucedidas e ao fato de que outras escolas as veem como exemplos a serem seguidos.

6. Considerações finais

O estudo teve como caminho metodológico a RSL. Por este motivo foi possível identificar e categorizar as diversas medidas de similaridade empregadas em teses e dissertações ao longo dos anos, proporcionando uma visão detalhada da evolução e tendências na área. A proposta do Modelo de Redes de Similaridade (MRS) evidenciou uma metodologia prática e sistemática para a análise e interpretação de redes de similaridade. A aplicação do MRS na identificação de clusters em uma rede de unidades escolares similares ilustrou sua eficácia e possibilidade de replicabilidade do MRS, destacando sua relevância e potencial de aplicação no campo de estudo da construção do conhecimento.

7. Referências

BARABÁSI, Albert-László. Linked (conectado): a nova ciência das networks. Tradução: Jonas Pereira dos Santos. São Paulo: Leopardo, 2009.

CARNEIRO, Tereza. Redes de afinidade como estratégia de gestão pedagógica e difusão do conhecimento em cursos na modalidade a distância. 2014. Tese (Doutorado Multi-institucional e Multidisciplinar em Difusão do Conhecimento) - Universidade Federal da Bahia, Salvador, 2014. Disponível em: http://repositorio.ufba.br/ri/handle/ri/16842. Acesso em: 05 jun. 2023.

CASTELLS, Manuel. A sociedade em rede. São Paulo: Paz e Terra, 1999. v. 1.

EASLEY, David; KLEINBERG, Jon. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge: Cambridge University Press, 2010.

GROSS, J.; YELLEN, J. Graph Theory and its Applications. Boca Raton: CRC Press, 1999.

Revista Educação e Ciências Sociais, UNEB, Salvador.v.6, n.11, jul – dez. de 2023 - ISSN 2595-9980
DOI: https://doi.org/10.38090/recs.vol6.n11.a18719

JEON, G.; PARK, J. Jury-Contestant Bipartite Competition Network: Identifying Biased Scores and Their Impact on Network Structure Inference. arXiv: Physics and Society, [s.l.], n. 1608.02326, 8 ago. 2016. Disponível em: https://doi.org/10.48550/arXiv.1608.02326. Acesso em: 21 de janeiro de 2023.

LIN, D. An information-theoretic definition of similarity. In Proceedings of the 15th international conference on machine learning. 1998. p. 296-304. Disponível em: https://www.cse.iitb.ac.in/~cs626-449/Papers/WordSimilarity/3.pdf . Acesso em: 25 de março de 2022.

MELO, D. F. P. Estudo de padrões em sinais musicais sob a perspectiva dos grafos de visibilidade. 2018. Tese (Doutorado em Difusão do Conhecimento) - Universidade Federal da Bahia, Salvador, 2018. Disponível em: https://repositorio.ufba.br/handle/ri/25713. Acesso em: 05 jun. 2023.

MELO, H. N. Predição temporal de links baseada na evolução de tríades. 2016. Tese (Doutorado em Ciência da Computação) - Universidade Federal de Pernambuco, Recife, 2016. Disponível em: https://repositorio.ufpe.br/handle/123456789/20829. Acesso em: 05 jun. 2023.

MONTEIRO, Roberto Luiz Souza. Um modelo evolutivo para simulação de redes de afinidade. 2012. Tese (Doutorado Multi-institucional e Multidisciplinar em Difusão do Conhecimento) - Universidade Federal da Bahia, Salvador, 2012. Disponível em: http://www.repositorio.ufba.br/ri/handle/ri/12961. Acesso em: 05 jun. 2023.

MORENO, J. L. Fondements de la sociométrie. Tradução de H. Lesage e P.-H. Maucorps. Revue française de science politique, [s.l.], v. 5, n. 3, p. 641-646, 1954. Disponível em: https://www.persee.fr/doc/rfsp_0035-2950_1955_num_5_3_402631_t1_0641_0000_002. Acesso em: 01 de março de 2021.

NEWMAN, M. Networks: An Introduction. Oxford: Oxford University Press, 2010.

PEREIRA, Aliger dos Santos. Clusters de veículo em Salvador: geoprocessamento e gestão de negócio para micro, pequenas e médias empresas (MPMEs). Salvador: EDUFBA; EDUNEB, 2019. Disponível em: http://repositorio.ufba.br/ri/handle/ri/30945. Acesso em: 18 jul. 2023.

PRISMA: ATUALIZANDO ORIENTAÇÕES PARA RELATAR REVISÕES SISTEMÁTICAS: DESENVOLVIMENTO DA DECLARAÇÃO PRISMA 2020. Journal of Clinical Epidemiology, [s.l.], v. 134, p. 103-112, jun. 2021. https://doi.org/10.1016/j.jclinepi.2021.02.003. Acesso em: 25 ago. 2022.

REBAZA, Jorge Carlos Valverde. Predição de links em redes complexas utilizando informações de estruturas de comunidades. 2013. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) - Universidade de São Paulo, São Paulo, 2013. Disponível em: https://doi.org/10.11606/D.55.2013.tde-05062013-104308. Acesso em: 05 jun. 2023.

REZENDE, Rodrigo Carvalho. Técnicas e algoritmos de Link Analysis na geração de medidas de similaridade. 2012. Dissertação (Mestrado) - Universidade Estadual de Campinas, Campinas, 2012. Disponível em: https://hdl.handle.net/20.500.12733/1619626. Acesso em: 05 jun. 2023.

ROCHA, W. S. Similaridade Semântica de Atributos para Dados em Nuvem: um estudo de caso no MIDAS. 2019. Dissertação (Mestrado em Matemática e Estatística) - Universidade Federal da Bahia, Salvador, 2019. Disponível em: http://repositorio.ufba.br/ri/handle/ri/33461. Acesso em: 05 jun. 2023.

VALEJO, Alan Demetrius Baria. Refinamento multinível em redes complexas baseado em similaridade de vizinhança. 2014. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) - Universidade de São Paulo, São Paulo, 2014. Disponível em: https://doi.org/10.11606/D.55.2014.tde-14042015-142526. Acesso em: 05 jun. 2023.

WANG, X.; HE, X.; CAO, Y.; LIU, M.; CHUA, T.-S. Kgat: Knowledge graph attention network for recommendation. Em: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery, 2019. Disponível na Internet: https://doi.org/10.1145/3292500.3330989. Acesso em: 02 jul. 2023.

YAO, K.; CHANG, L.; YU, J. X. Identifying similar-bicliques in bipartite graphs. Proceedings of the VLDB Endowment, v. 15, n. 11, p. 3085-3097, jul. 2022. Disponível em: https://doi.org/10.14778/3551793.3551854. Acesso em: 21 de janeiro de 2023.