Mecanismo de Busca Semântica Baseado em Word Embeddings em Dados do Currículo Lattes, Programas de Pós-Graduação e Grupos de Pesquisa

João Vítor Café dos R. Batista [1], Gleidson de Meireles Costa [2], Eduardo Manuel de Freitas Jorge[1]

[1]Departamento de Ciência Exatas e da Terra – Universidade do Estado da Bahia (UNEB)
Salvador – BA – Brasil

[2]Universidade Federal do Recôncavo da Bahia (UFRB)
– Feira de Santana, BA – Brasil

batistajv2012@gmail.com, geu_costa@outlook.com, ejorge@uneb.br

Abstract. *The search for researchers and scientific publications is essential for access to academic knowledge. However, keyword-based search mechanisms might fail to capture the semantics of queries, which can lead to less relevant results. This research proposes the implementation and analysis of a semantic search mechanism, using Word Embeddings to provide more relevant answers in the academic context. The study presents an architecture and implementation that allows efficient semantic searches in scientific databases through the transformation and indexing of Word Embeddings.*

Resumo. *A busca por pesquisadores e publicações científicas é fundamental para o acesso ao conhecimento acadêmico. No entanto, os mecanismos de busca baseados em correspondência de palavras-chave podem ignorar a semântica das consultas, o que pode resultar em respostas pouco relevantes. Esta pesquisa propõe a implementação e análise de um mecanismo de busca semântica, utilizando Word Embeddings para fornecer respostas mais relevantes no contexto acadêmico. O estudo apresenta uma arquitetura e implementação que permite a realização de buscas semânticas em bases de dados científicas de maneira eficiente, por meio da transformação e indexação de Word Embeddings.*

1. Introdução

Os mecanismos de busca desempenham um papel crucial no processo de descoberta do conhecimento e no acesso à informação. De acordo com [Sheela and Jayakumar 2019], mecanismos de busca são programas de computador utilizados para pesquisar e recuperar diversos tipos de informação em repositórios de informação, como, por exemplo, a internet. Esses sistemas têm como objetivo retornar os resultados mais relevantes que correspondam às consultas feitas por um determinado usuário [Sharma and Kumar 2022].

Historicamente, os motores de busca continham uma abordagem léxica e eram empregados de forma que os resultados eram obtidos apenas com base nas palavras-chave da consulta, o que poderia fornecer resultados irrelevantes como resposta à consulta do usuário [Sharma et al. 2021]. As abordagens tradicionais de *Recuperação de*

Informação (RI) funcionam com base em termos e ignoram o contexto semântico dos conteúdos textuais [Sharma and Kumar 2022]. Isso resulta em uma lacuna vocabular quando as consultas e os documentos usam termos diferentes para descrever o mesmo conceito [Sharma and Kumar 2022].

Abordagens semânticas impulsionadas por elementos da *Inteligência Artificial* (IA) começaram a ser utilizadas para aumento da eficácia nos mecanismos de busca. Nessa linha de raciocínio, surge uma técnica conhecida como *Word Embeddings*, que vem sendo utilizada para enfrentar muitas tarefas desafiadoras de *Processamento de Linguagem Natural* (PLN) que necessitam capturar as relações semânticas em dados textuais [Jbene et al. 2021]. Os *Word Embeddings* codificam palavras ao transformá-las em vetores, de modo que palavras com significados semelhantes tenham representações vetoriais parecidas [Forgues et al. 2014]. Dessa forma, esses *embeddings* são gerados com o objetivo de capturar a semântica lexical e a relação contextual para lidar com o conceito semântico das palavras [Jbene et al. 2021]. Com o surgimento de uma abordagem semântica, os motores de busca ganharam potencial para resolver os problemas de lacuna vocabular e imprecisão que a busca baseada apenas no contexto léxico traz à tona [Sharma et al. 2021]. A busca semântica representa a busca com significado, que neste caso, pode referir-se à compreensão do contexto da consulta, à interpretação dos dados ou a representação dos resultados de uma maneira mais assertiva de acordo com a consulta realizada [Gundyreva et al. 2022].

Quando aplicados ao contexto de dados acadêmicos, os mecanismos de busca podem contribuir significativamente para a disseminação do conhecimento, fornecendo meios para explorar diversas fontes de informação. Isso permite que indivíduos, pesquisadores, profissionais e empresas conduzam estudos, desenvolvam projetos, encontrem pesquisadores que são referência em uma determinada área e tomem decisões fundamentadas em informações relevantes. De acordo com [Gupta 2017], buscar artigos, teses, livros, pesquisadores, resumos de editoras acadêmicas, sociedades profissionais, repositórios online, universidades e outros sites na web pode se tornar um processo muito mais fácil ao utilizar mecanismos de busca acadêmicos que oferecem uma busca direcionada especificamente para esse tipo de conteúdo.

Diante disso, o objetivo da pesquisa é a implementação e análise de um mecanismo de busca semântica que utiliza *Word Embeddings* como ferramenta para facilitar e proporcionar resultados mais relevantes para os usuários no contexto de buscas em dados acadêmicos. Esse mecanismo será integrado a uma base de dados que engloba informações dos currículos Lattes, programas de pós-graduação e grupos de pesquisa das universidades estaduais da Bahia (UEBA), contribuindo para o avanço da pesquisa científica e facilitando a identificação de oportunidades de colaboração.

2. Trabalhos correlatos

Diversas estratégias de busca semântica têm sido implementadas para aprimorar a recuperação de informações na web. [Rastogi et al. 2021] propõe o algoritmo WeOnto, que utiliza um processo em duas etapas combinando ontologias e modelos de Word Embeddings para reformular consultas com maior precisão, enriquecendo palavras-chave com informações semânticas. Da mesma forma, [Deepak and Santhanavijayan 2022] apresenta o UQSCM-RFD, que também utiliza Word Embeddings gerados pelo modelo

Word2Vec para expandir consultas com termos semanticamente semelhantes. Além disso, [Sharma and Kumar 2022] desenvolveram uma rede neural e um método baseado em ontologia para indexação semântica, integrando embeddings e recursos de conhecimento externo para melhorar a eficiência na recuperação de informações.

A necessidade de refinar modelos de *embeddings* para domínios específicos resulta em estudos como o de [Farmanbar et al. 2020], que sugere expandir consultas com palavras similares com base em Word Embeddings treinados em domínios específicos. [Tuncer et al. 2021] propõem treinar modelos como Word2Vec e FastText em textos relacionados a vagas de emprego, utilizando algoritmos de clusterização para melhorar a precisão na recuperação de informações.

Além disso, alguns artigos abordam mecanismos de indexação e mensuração de similaridades. [Jbene et al. 2021] apresenta o uso de embeddings para melhorar a busca em e-commerce, destacando BERT e FastText, além da similaridade do cosseno como medida de similaridade. Por fim, [Ta et al. 2022] propõem uma pipeline para recuperação de argumentos usando Doc2Vec e Sentence-BERT, com o Elastic Search para indexação e o modelo de similaridade DirichletLM para pontuação de relevância entre documentos e consultas.

3. Metodologia

A metodologia utilizada para o desenvolvimento dessa pesquisa é baseada na abordagem *Design Science Research* (DSR). A DSR é o método de pesquisa orientado a problemas que fundamenta e operacionaliza a condução da pesquisa quando o objetivo a ser alcançado é um artefato ou prescrição [Dresch et al. 2020]. A DSR parte do princípio, de que a partir do entendimento do problema, deve-se construir e avaliar artefatos que melhorem a situação para um estado melhor ou desejável [Dresch et al. 2020].

No contexto dessa pesquisa o DSR pode ser explorado da seguinte maneira:

- **Identificação do problema:** A necessidade de superar as limitações impostas pelos mecanismos de busca acadêmicos que utilizam métodos tradicionais baseados em correspondências de palavras-chave para retornar resultados relevantes dado um conjunto de dados.
- **Definição dos objetivos do artefato:** Proporcionar um mecanismo de busca que atue em dados acadêmicos, e que considere a semântica entre a consulta do usuário e os dados armazenados para retornar os resultados relevantes.
- **Desenvolvimento do artefato:** Projetar e desenvolver um mecanismo de busca semântico que permita a realização de consultas em dados que contém informações sobre produções científicas e pesquisadores. Isso engloba a seleção de uma base de dados que contenha as informações acadêmicas necessárias para realização da busca, a seleção de tecnologias que serão utilizadas, a implementação de uma pipeline de dados para transformação dos dados relevantes em *Word Embeddings* e por fim a implementação do mecanismo de busca, formado por uma camada de *back-end* responsável pelo processamento, e uma camada de *front-end*, responsável pela apresentação dos resultados.
- **Avaliação do artefato:** Aplicar métricas de avaliação comumente utilizadas em mecanismo de busca, como *Precision*, *Recall* e *F1-Score*, para verificação da eficácia do artefato desenvolvido.

- **Apresentação dos resultados obtidos:** Apresentar os resultados obtidos, e com base nas métricas de avaliação definidas, compara-los com um método de busca tradicional baseado em correspondência de palavras-chave com o objetivo de realizar a validação do artefato construído.

4. Desenvolvimento

4.1. Base de dados

A base de dados utilizada pelo *Sistema de Mapeamento de Consultas* (SIMCC) foi escolhida para a realização das consultas. Essa base de dados possui dados extraídos da Plataforma Lattes, detentora do formato-padrão de coleta de informações curriculares, instituições e pesquisadores da área de ciência e tecnologia na maioria das universidades do Brasil, da plataforma Sucupira responsável por coletar, analisar e avaliar as informações utilizadas como base padronizadora do *Sistema Nacional de Pós-Graduação* (SNPG) e do *Journal Citation Reports* (JCR) para avaliar a qualidade e o impacto de revistas científicas [dos Santos et al. 2024].

4.2. *Embedding Model*

O *Massive Text Embedding Benchmark* (MTEB) foi utilizado para comparar o desempenho dos *embeddings* gerados por diversos *Embedding Models* em diversas tarefas relacionadas a RI. O *Embedding Model* escolhido para a transformação de dados textuais em *Word Embeddings* foi o modelo *embeddings-001* pertencente a *Google*. A escolha se dá pelo fato da utilização do modelo ser possível de forma gratuita e de fácil acesso via *Application Programming Interface* (API), além de possuir uma velocidade considerável, visto que não é necessário executar o *Embedding Model* localmente.

4.3. *Pipeline e camada de Back-end*

Para a construção do pipeline de dados e para a camada de *back-end* foi utilizada a linguagem Python, que possui um conjunto de bibliotecas para pré-processamento de texto e outras técnicas de PLN.

O primeiro passo para a realização de uma busca semântica baseada em embeddings, é a transformação dos elementos textuais selecionados em *embeddings*. Os elementos textuais relacionados ao contexto científico como o titulo de produções científicas e currículo dos pesquisadores foram extraídos, pré-processados, transformados em *embeddings* e armazenados em tabelas criadas para a construção da solução.

O banco de dados PostgreSQL foi utilizado para armazenamento dos dados utilizados na aplicação, visto que ele possui suporte à indexação vetorial, de modo que é possível armazenar os *Word Embeddings* e realizar uma comparação de similaridade entre esses vetores através de uma função de similaridade embutida, como a *cosine similarity*, que é comumente utilizada em tarefas de *Similaridade Semântica Textual* (STS). Além disso, o banco de dados PostgreSQL, também possui um amplo número de extensões voltadas para tarefas de processamento de linguagem natural. No contexto do artefato desenvolvido, a extensão *pgvector*, foi utilizada para possibilitar o armazenamento de *embeddings* e a busca por similaridade de vetor eficiente.

O framework *Flask* foi utilizado na camada de back-end para construção da *API* responsável por processar a consulta e retornar os dados relevantes ao usuário. A consulta então é transformada pré-processada, transformada em *embeddings*, de modo que seja feita uma consulta SQL no banco de dados utilizando-se do operador <=> do PostgreSQL para calcular a *similaridade do cosseno* entre o *embedding* da consulta e todos os *embedding* referentes a coluna que possui os embeddings gerados em uma tabela no banco de dados. Os resultados são ordenados de forma decrescente e filtrados de acordo com o valor obtido no cálculo da similaridade do cosseno. Após a realização da consulta, os dados são formatados e enviados para exibição na camada de interface, que foi construída com base nas principais tecnologias fundamentais da web como *HyperText Markup Language* (HTML), *Cascading Style Sheets* (CSS) e Javascript.

4.4. *Arquitetura da solução*

Na figura 1 é possível visualizar a arquitetura da solução, bem como processo que ocorre internamente, desde a consulta do usuário até o retorno dos resultados relevantes.

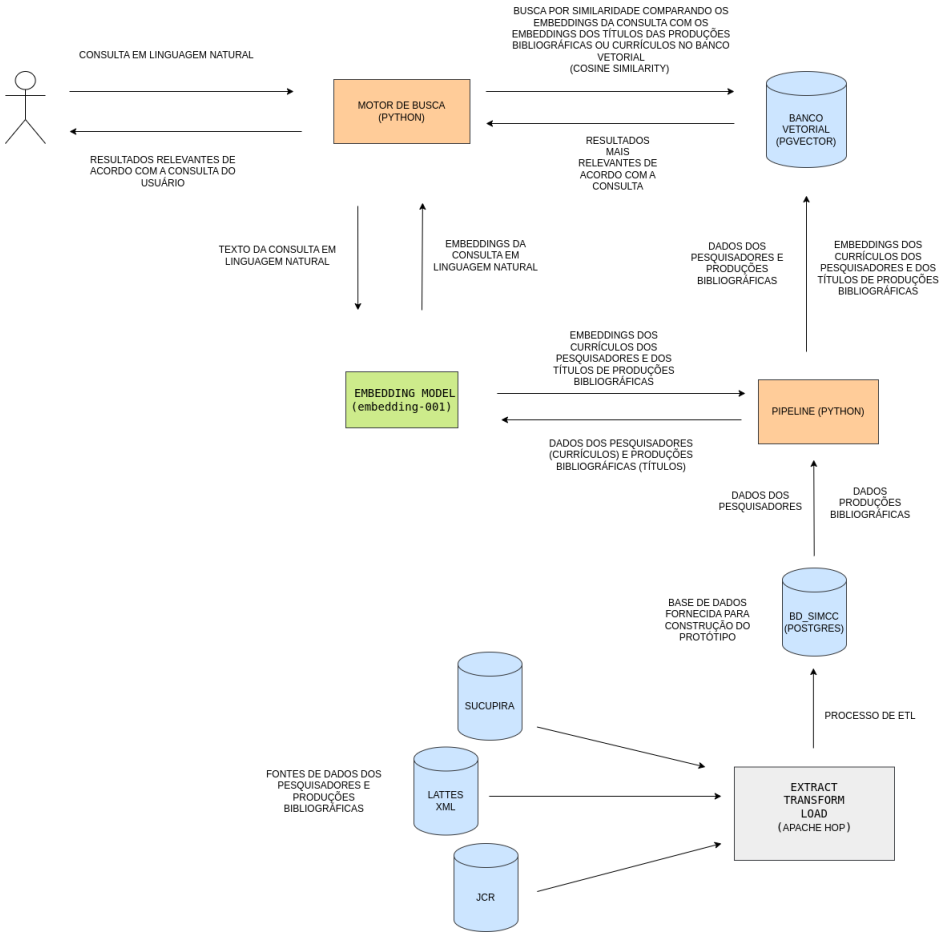


Figura 1. Representação da arquitetura do projeto

5. Experimentos

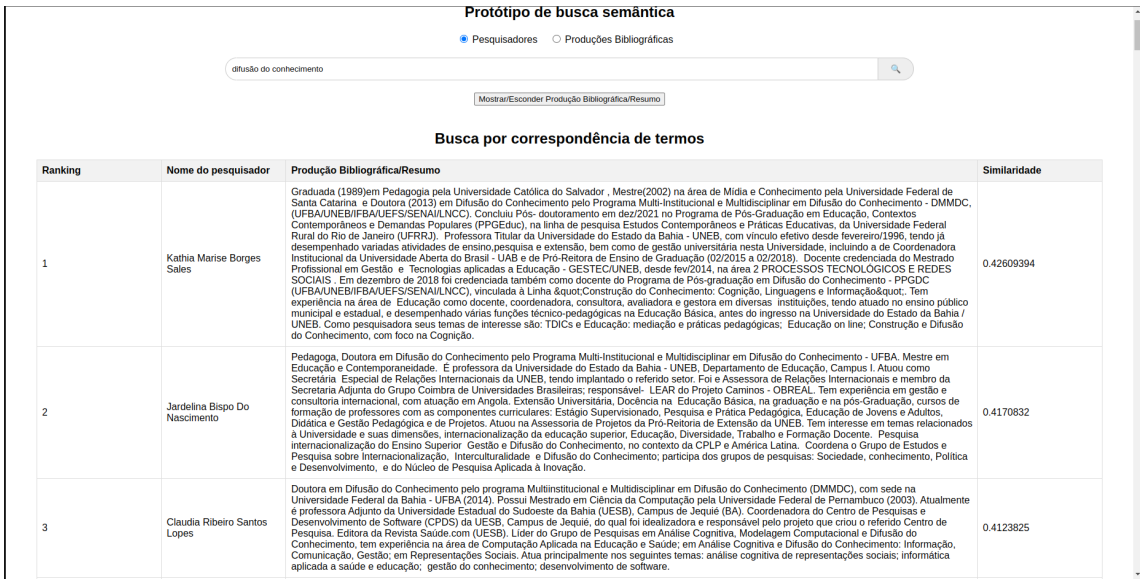
Visto que a base de dados utilizada não possui resultados relevantes pré-definidos ou conhecidos para uma determinada consulta ainda não é possível utilizar as métricas de

avaliação, como *Precision*, *Recall* e *F1-Score* destacadas na metodologia. Diante desse cenário, foram realizadas comparações empíricas entre o método de busca semântico desenvolvido, baseado em *Word Embeddings*, o método de busca tradicional, baseado em correspondências de palavras chave e um método de busca híbrido, que utiliza uma abordagem semântica em conjunto com uma abordagem léxica, ou seja, a pontuação de similaridade é definida pela soma da *similaridade do cosseno* com uma pontuação gerada por uma função do PostgreSQL chamada *websearch_to_tsquery*, que também determina a similaridade de uma consulta com base em correspondência de palavras. Embora uma busca baseada em *embeddings* funcione bem ao considerar a semântica das palavras, ela pode gerar um ranqueamento de resultados ineficaz. Para mitigar essa limitação, a abordagem híbrida combina a proximidade semântica com a correspondência de palavras, assegurando que os resultados com termos da consulta e alta relevância semântica sejam melhor ranqueados, ou seja, classificados como mais relevantes.

5.1. Consultas referentes a pesquisadores

A busca por pesquisadores visa trazer os pesquisadores mais relevantes com base na similaridade entre a consulta e o seu currículo.

Na figuras 2, 3 e 4 é possível observar os três primeiros resultados retornados para a consulta “difusão do conhecimento”, visando buscar pesquisadores que tem uma formação em difusão do conhecimento, ou uma formação semânticamente próxima.



Ranking	Nome do pesquisador	Produção Bibliográfica/Resumo	Similaridade
1	Kathia Marise Borges Sales	Graduada (1989)em Pedagogia pela Universidade Católica do Salvador , Mestre(2002) na área de Mídia e Conhecimento pela Universidade Federal de Santa Catarina e Doutora (2013) em Difusão do Conhecimento pelo Programa Multi-Institucional e Multidisciplinar em Difusão do Conhecimento - DMMDC, (UFBA/UNEB/IFBA/UEFS/SENAI/LNCC). Concluiu Pós- doutoramento em dez/2021 no Programa de Pós-Graduação em Educação, Contextos Contemporâneos e Demandas Populares (PPGEduc), na linha de pesquisa Estudos Contemporâneos e Práticas Educativas, da Universidade Federal Rural do Rio de Janeiro (UFRRJ). Professora Titular da Universidade do Estado da Bahia - UNEB, com vínculo efetivo desde fevereiro/1996, tendo já desempenhado variadas atividades de ensino,pesquisa e extensão, bem como de gestão universitária nesta Universidade, incluindo a de Coordenadora Institucional da Universidade Aberta do Brasil - UAB e de Pró-Reitora de Ensino de Graduação (02/2015 a 02/2018). Docente credenciada do Mestrado Profissional em Gestão e Tecnologias aplicadas a Educação - GESTEC/UNEB, desde fev/2014, na área 2 PROCESSOS TECNOLÓGICOS E REDES SOCIAIS . Em dezembro de 2018 foi credenciada também como docente do Programa de Pós-graduação em Difusão do Conhecimento - PPGDC (UFBA/UNEB/IFBA/UEFS/SENAI/LNCC), vinculada à Linha "Construção do Conhecimento: Cognição, Linguagens e Informação". Tem experiência na área de Educação como docente, coordenadora, consultora, avaliadora e gestora em diversas instituições, tendo atuado no ensino público municipal e estadual, e desempenhado várias funções técnico-pedagógicas na Educação Básica, antes do ingresso na Universidade do Estado da Bahia / UNEB. Como pesquisadora seus temas de interesse são: TDICs e Educação: mediação e práticas pedagógicas; Educação on line; Construção e Difusão do Conhecimento, com foco na Cognição.	0.42609394
2	Jardelina Bispo Do Nascimento	Pedagoga, Doutora em Difusão do Conhecimento pelo Programa Multi-Institucional e Multidisciplinar em Difusão do Conhecimento - UFBA. Mestre em Educação e Contemporaneidade. É professora da Universidade do Estado da Bahia - UNEB, Departamento de Educação, Campus I. Atuou como Secretária Especial de Relações Internacionais da UNEB, tendo implantado o referido setor. Foi e Assessora de Relações Internacionais e membro da Secretaria Adjunta do Grupo Coimbra de Universidades Brasileiras; responsável- LEAR do Projeto Caminos - OBEREAL. Tem experiência em gestão e consultoria internacional, com atuação em Angola. Extensão Universitária, Docência na Educação Básica, na graduação e na pós-Graduação, cursos de formação de professores com as componentes curriculares: Estágio Supervisionado, Pesquisa e Prática Pedagógica, Educação de Jovens e Adultos, Didática e Gestão Pedagógica e de Projetos. Atuou na Assessoria de Projetos da Pró-Reitoria de Extensão da UNEB. Tem interesse em temas relacionados à Universidade e suas dimensões, internacionalização da educação superior, Educação, Diversidade, Trabalho e Formação Docente. Pesquisa internacionalização do Ensino Superior Gestão e Difusão do Conhecimento, no contexto da CPLP e América Latina. Coordena o Grupo de Estudos e Pesquisa sobre Internacionalização, Interculturalidade e Difusão do Conhecimento; participa dos grupos de pesquisas: Sociedade, conhecimento, Política e Desenvolvimento, e do Núcleo de Pesquisa Aplicada à Inovação.	0.4170832
3	Claudia Ribeiro Santos Lopes	Doutora em Difusão do Conhecimento pelo programa Multiinstitucional e Multidisciplinar em Difusão do Conhecimento (DMMDC), com sede na Universidade Federal da Bahia - UFBA (2014). Possui Mestrado em Ciência da Computação pela Universidade Federal de Pernambuco (2003). Atualmente é professora Adjunto da Universidade Estadual do Sudoeste da Bahia (UESB), Campus de Jequié (BA). Coordenadora do Centro de Pesquisas e Desenvolvimento de Software (CPDS) da UESB, Campus de Jequié, do qual foi idealizadora e responsável pelo projeto que criou o referido Centro de Pesquisa. Editora da Revista Saúde.com (UESB). Líder do Grupo de Pesquisas em Análise Cognitiva, Modelagem Computacional e Difusão do Conhecimento, tem experiência na área de Computação Aplicada na Educação e Saúde; em Análise Cognitiva e Difusão do Conhecimento: Informação, Comunicação, Gestão; em Representações Sociais. Atua principalmente nos seguintes temas: análise cognitiva de representações sociais; informática aplicada a saúde e educação; gestão do conhecimento; desenvolvimento de software.	0.4123825

Figura 2. Resultados para a busca baseada em correspondência de termos

Figura 3. Resultados para a busca baseada em semântica utilizando *Word Embeddings*

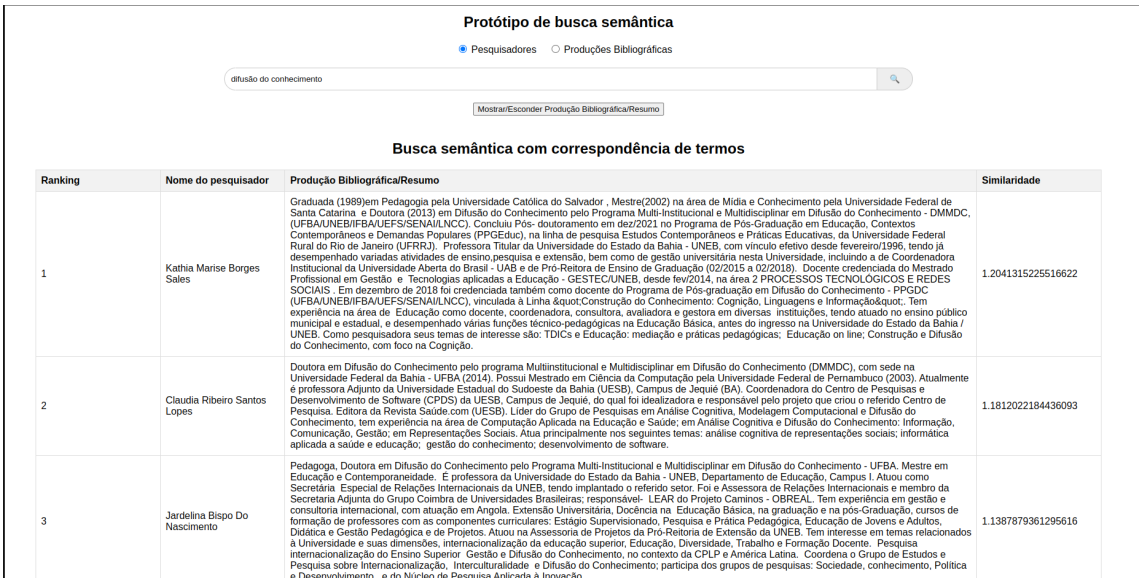


Figura 4. Resultados para a busca baseada em semântica utilizando *Word Embeddings* com correspondência de termos

5.2. Consultas referentes a produções bibliográficas

Na figuras 5, 6 , 7 é possível observar os resultados retornados para a consulta “doenças infantis’. É possível observar os resultados para 3 tipos de mecanismos de busca construídos. Essa consulta é feita com o objetivo de buscar por produções científicas que estejam relacionados a doenças infantis, ou seja, doenças que afetam crianças.

Protótipo Busca Semântica SIMCC

Protótipo de busca semântica

Pesquisadores Produções Bibliográficas

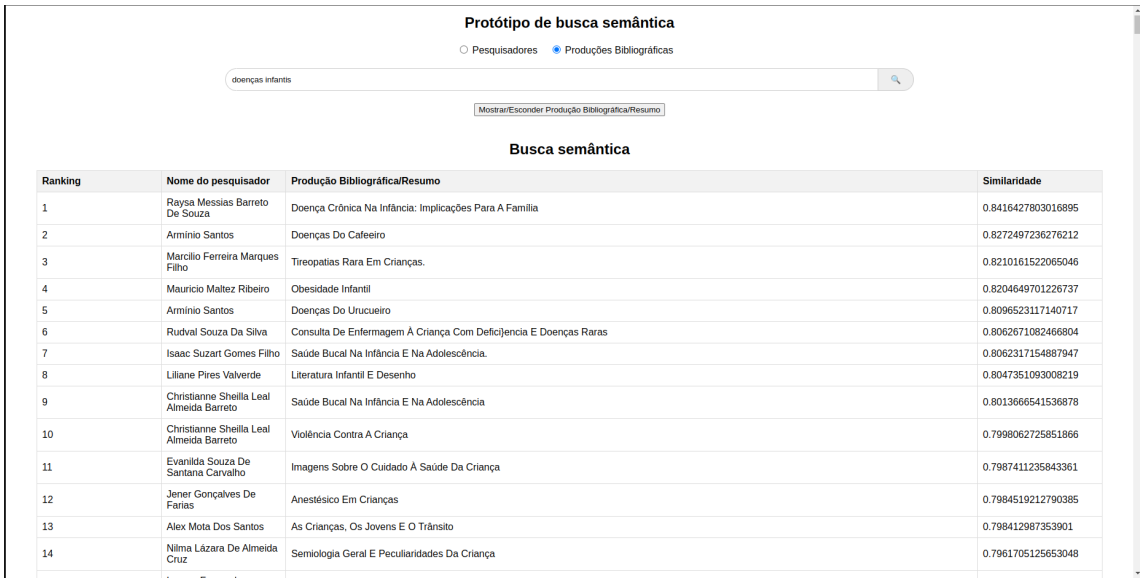
doenças infantis

Mostrar/Esconder Produção Bibliográfica/Resumo

Busca por correspondência de termos

Ranking	Nome do pesquisador	Produção Bibliográfica/Resumo	Similaridade
1	Gerenice Ribeiro De Oliveira Cortes	As Doenças Negligenciadas Nas Mídias Digitais: O Processo Social Saúde-Doença, Imaginário E Efeitos-Sentidos	0.075990885
2	Cristiana Da Costa Libório Lago	Associação Entre A Doença Periodontal E A Doença Pulmonar Obstrutiva Crônica: Uma Revisão De Literatura	0.075990885
3	Evanilda Souza De Santana Carvalho	Representações Da Doença Falciforme E Do Corpo Para Os Adolescentes Com Doença Falciforme	0.075990885
4	Rodolfo Macedo Cruz Pimenta	Relação Da Doença Periodontal Com Doenças Sistêmicas E Alterações Bucais: Uma Revisão De Literatura	0.075990885
5	Artur Gomes Dias Lima	Água, Saúde E Doença: Uma Revisão Sistemática Sobre Doenças De Veiculação Hídrica Em Comunidades Indígenas Brasileiras	0.075990885
6	Edval Gomes Dos Santos Júnior	Biomarcadores Em Cardiologia - Parte 2: Na Doença Coronária, Doença Valvar E Situações Especiais	0.075990885
7	Eulina Patrícia Oliveira Ramos Pires	Percepção Da Qualidade De Vida De Idosas Com Doença Crônica Em Terapia Aquática.	0.06079271
8	Fabio Ornellas Prado	Doença Periodontal: Gravidez E Parto Prematuro	0.06079271
9	Vanner Boere Souza	Uma Intervenção Sobre A Toxoplasmose Em Três Municípios Do Sul Da Bahia: A Percepção Popular Da Doença	0.06079271
10	Arminio Santos	Doenças Do Urucueiro	0.06079271
11	Cláudio Bispo De Almeida	Atividade Física E Doenças Respiratórias	0.06079271

Figura 5. Resultados para a busca baseado em correspondência de termos



Protótipo de busca semântica

Pesquisadores Produções Bibliográficas

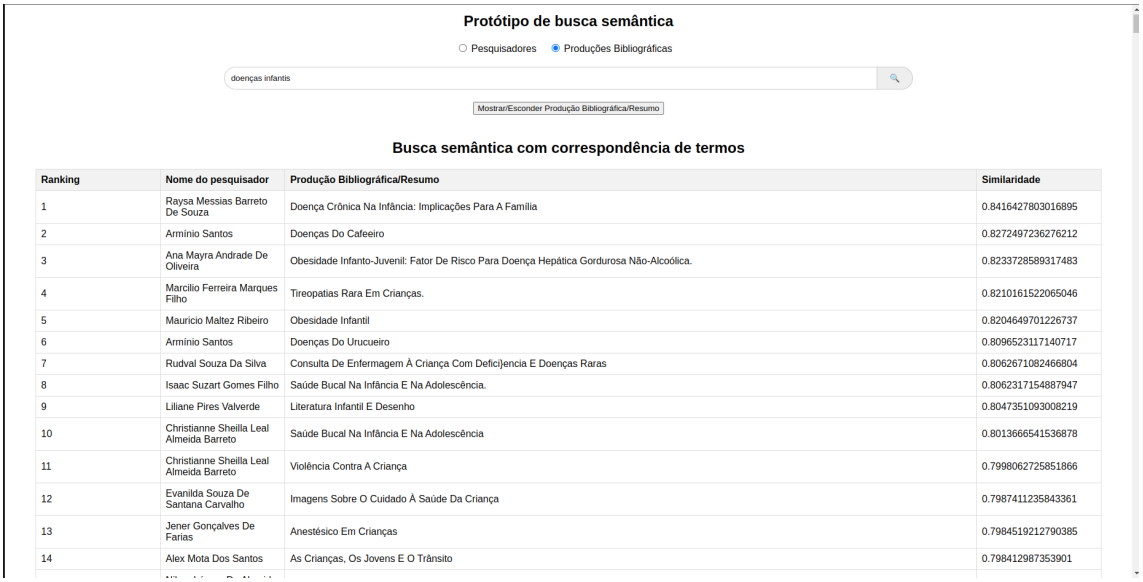
doenças infantis

Mostrar/Esconder Produção Bibliográfica/Resumo

Busca semântica

Ranking	Nome do pesquisador	Produção Bibliográfica/Resumo	Similaridade
1	Raysa Messias Barreto De Souza	Doença Crônica Na Infância: Implicações Para A Família	0.8416427803016895
2	Arminio Santos	Doenças Do Cafeeiro	0.8272497236276212
3	Marcílio Ferreira Marques Filho	Tireopatias Rara Em Crianças.	0.8210161522065046
4	Mauricio Maltez Ribeiro	Obesidade Infantil	0.8204649701226737
5	Arminio Santos	Doenças Do Urucueiro	0.8096523117140717
6	Rudval Souza Da Silva	Consulta De Enfermagem À Criança Com Defici}encia E Doenças Raras	0.8062671082466804
7	Isaac Suzart Gomes Filho	Saúde Bucal Na Infância E Na Adolescência.	0.8062317154887947
8	Liliane Pires Valverde	Literatura Infantil E Desenho	0.8047351093008219
9	Christianne Sheilla Leal Almeida Barreto	Saúde Bucal Na Infância E Na Adolescência	0.8013666541536878
10	Christianne Sheilla Leal Almeida Barreto	Violência Contra A Criança	0.7998062725851866
11	Evanilda Souza De Santana Carvalho	Imagens Sobre O Cuidado À Saúde Da Criança	0.7987411235843361
12	Jener Gonçalves De Farias	Anestésico Em Crianças	0.7984519212790385
13	Alex Mota Dos Santos	As Crianças, Os Jovens E O Trânsito	0.798412987353901
14	Nilma Lázara De Almeida Cruz	Semiologia Geral E Peculiaridades Da Criança	0.7961705125653048

Figura 6. Resultados para a busca baseada em semântica utilizando *Word Embeddings*



Ranking	Nome do pesquisador	Produção Bibliográfica/Resumo	Similaridade
1	Raysa Messias Barreto De Souza	Doença Crônica Na Infância: Implicações Para A Família	0.8416427803016895
2	Arminio Santos	Doenças Do Cafeeiro	0.8272497236276212
3	Ana Mayra Andrade De Oliveira	Obesidade Infanto-Juvenil: Fator De Risco Para Doença Hepática Gordurosa Não-Alcoólica.	0.8233728589317483
4	Marcílio Ferreira Marques Filho	Tireopatias Rara Em Crianças.	0.8210161522065046
5	Mauricio Maltez Ribeiro	Obesidade Infantil	0.8204649701226737
6	Arminio Santos	Doenças Do Urucueiro	0.8096523117140717
7	Rudval Souza Da Silva	Consulta De Enfermagem À Criança Com Defici}encia E Doenças Raras	0.8062671082466804
8	Isaac Suzart Gomes Filho	Saúde Bucal Na Infância E Na Adolescência.	0.8062317154887947
9	Liliane Pires Valverde	Literatura Infantil E Desenho	0.8047351093008219
10	Christianne Sheilla Leal Almeida Barreto	Saúde Bucal Na Infância E Na Adolescência	0.8013666541536878
11	Christianne Sheilla Leal Almeida Barreto	Violência Contra A Criança	0.7998062725851866
12	Evanilda Souza De Santana Carvalho	Imagens Sobre O Cuidado À Saúde Da Criança	0.7987411235843361
13	Jener Gonçalves De Farias	Anestésico Em Crianças	0.7984519212790385
14	Alex Mota Dos Santos	As Crianças, Os Jovens E O Trânsito	0.798412987353901

Figura 7. Resultados para a busca baseada em semântica utilizando *Word Embeddings* com correspondência de termos

6. Conclusões preliminares e trabalhos futuros

A pesquisa realizada permite a apresentação de algumas conclusões parciais. Estas conclusões são baseadas nas observações e análises empíricas dos resultados da busca. O método de busca semântica se mostra eficiente e retorna resultados compatíveis com a consulta realizada pelo usuário na maioria das consultas. Entretanto, às vezes, enfrenta o mesmo problema do método de busca léxico, baseado em correspondência de termos, ao retornar resultados que não são relevantes. Para mitigar esse problema, foi utilizado um método de busca híbrido que combina ambos os métodos. Contudo, é necessário o uso de uma base de dados que contenha resultados relevantes conhecidos para a aplicação das métricas de precision, recall e F1-score, para validação dos resultados.

Em trabalhos futuros, considera-se a possibilidade de a busca levar em conta não apenas a semântica, mas também metadados, como o qualis, autores e veículos de publicação, que podem estar expressos em linguagem natural na consulta do usuário e servir como filtros durante a pesquisa. A extração desses metadados pode ser realizada por meio de um modelo de IA generativa , permitindo que os filtros sejam gerados e a busca semântica ocorra apenas sobre os registros filtrados.

Referências

Deepak, G. and Santhanavijayan, A. (2022). Uqscm-rfd: A query–knowledge interfacing approach for diversified query recommendation in semantic search based on river flow dynamics and dynamic user interaction. *Neural Computing and Applications*, 34(1):651–675.

dos Santos, M. S., de Jesus Oliveira, V. H., de Freitas Jorge, E. M., and de Meireles Costa, G. (2024). Solução para mapeamento e consulta das competências dos pesquisadores: uma arquitetura para extração, integração e consultas de informações acadêmicas. *Cadernos de Prospecção*, 17(2):671–688.

Dresch, A., Lacerda, D. P., and Junior, J. A. V. A. (2020). *Design science research: método de pesquisa para avanço da ciência e tecnologia*. Bookman Editora.

Farmanbar, M., Van Ommeren, N., and Zhao, B. (2020). Semantic search with domain-specific word-embedding and production monitoring in fintech. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 28–33.

Forgues, G., Pineau, J., Larchevêque, J.-M., and Tremblay, R. (2014). Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2, page 168.

Gundyreva, E., Pivovarova, L., and Zosa, E. (2022). Unsupervised linking of scientific articles to food systems taxonomies.

Gupta, S. (2017). A survey on search engines. *Journal for Research— Volume*, 2(11).

Jbene, M., Tigani, S., Saadane, R., and Chehri, A. (2021). Deep neural network and boosting based hybrid quality ranking for e-commerce product search. *Big Data and Cognitive Computing*, 5(3):35.

Rastogi, N., Verma, P., and Kumar, P. (2021). Query expansion based on word embeddings and ontologies for efficient information retrieval. *International Journal of Advanced Computer Science and Applications*, 12(11).

Sharma, A. and Kumar, S. (2022). Shallow neural network and ontology-based novel semantic document indexing for information retrieval. *Intelligent Automation & Soft Computing*, 34(3):1989–2005.

Sharma, D. K., Pamula, R., and Chauhan, D. (2021). Semantic approaches for query expansion. *Evolutionary Intelligence*, 14(2):1101–1116.

Sheela, A. S. and Jayakumar, C. (2019). Comparative study of syntactic search engine and semantic search engine: A survey. In *2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, volume 1, pages 1–4.

Ta, C. V., Reiner, F., von Detten, I., and Stöhr, F. (2022). Touché-task 1-team korg: Finding pairs of argumentative sentences using embeddings. In *CLEF (Working Notes)*, pages 3131–3148.

Tuncer, I., Kara, K. C., and Karakaş, A. (2021). Improving search relevance with word embedding based clusters. In *Trends in Data Engineering Methods for Intelligent Systems: Proceedings of the International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2020)*, pages 15–24. Springer.