Modelo semântico de recuperação da informação para recomendação de revistas científicas

Renan C. Batista[1,2], Fabio L. Canto[2], Washington L. R. C. Segundo[2], Thiago M. R. Dias[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Maracanaú – CE – Brazil

[2]Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
Brasília – DF – Brazil

{renancarbat,thiagomagela}@gmail.com,

{fabiocanto,washingtonsegundo}@ibict.br

Abstract. *This paper describes the approach employed in the development of a scientific journal recommendation system. The system leverages natural language processing and machine learning techniques applied to a large volume of scientific article data. Textual data were extracted from the OpenAlex repository, comprising over 10 million articles published across approximately 140,000 journals. The system architecture integrates textual pre-processing, semantic embedding generation, and optimized vector-based retrieval, enabling journal recommendations based on similarity scores with the abstract of a user-provided manuscript.*

Resumo. *Este trabalho descreve a abordagem empregada no desenvolvimento de um sistema de recomendação de revistas científicas. São utilizadas técnicas de processamento de linguagem natural e aprendizado de máquina aplicadas a um grande volume de dados de artigos científicos. Foram extraídos do repositório OpenAlex dados textuais de mais de 10 milhões de artigos publicados aproximadamente 140 mil revistas. A arquitetura do sistema combina recursos de pré-processamento textual, geração de embeddings semânticos e recuperação vetorial otimizada, permitindo recomendações de revistas a partir do escore de similaridade com o resumo de um manuscrito fornecido pelo usuário.*

1. Introdução

O crescimento da produção científica global pode dificultar o processo de escolha de revistas científicas adequadas para submissão de manuscritos, especialmente entre pesquisadores iniciantes ou que desejam publicar em outras áreas do conhecimento. É um processo que consome um tempo significativo dos pesquisadores, e uma escolha equivocada pode levar à rejeição do manuscrito, atrasando a publicação dos resultados da pesquisa e, talvez, do projeto com um todo [Entrup et al. 2022, Gündoğan et al. 2023].

Sistemas de recomendação de revistas científicas podem facilitar o processo de submissão de manuscritos, auxiliando autores na seleção de revistas relevantes entre milhares de títulos disponíveis a partir da análise de padrões de temas de milhões de artigos publicados [Ogunde et al. 2020].

Para construção de sistemas desse tipo vêm sendo utilizadas técnicas avançadas de processamento de linguagem natural e aprendizado de máquina, capazes de identificar similaridades semânticas entre textos e mapear padrões temáticos em grandes volumes de dados textuais. Essas técnicas permitem gerar recomendações mais precisas, considerando a similaridade temática entre o manuscrito e a lista de revistas recomendadas [Gündoğan et al. 2023].

As abordagens mais recentes vêm incluindo o uso de técnicas de aprendizado profundo. Modelos como o SBERT (Sentence-BERT) e o uso de *embeddings* densos possibilitam representar sentenças como vetores em espaços semânticos de alta dimensionalidade, permitindo cálculos mais precisos de similaridade semântica entre textos [Kamalloo et al. 2023].

O presente trabalho descreve a metodologia empregada para desenvolvimento de um sistema de recomendação de revistas científicas baseado em técnicas de processamento de linguagem natural e aprendizado de máquina aplicadas a um grande conjunto de dados de artigos e de revistas científicas. Foi utilizado o repositório OpenAlex para extração de dados de aproximadamente 10 milhões de artigos publicados em mais de 140 mil revistas científicas.

A arquitetura do sistema combina recursos de pré-processamento textual, geração de *embeddings* semânticos e recuperação vetorial otimizada, permitindo recomendações precisas e contextualizadas a partir do resumo de um manuscrito fornecido pelo usuário. Para a geração dos *embeddings* dos artigos, foram utilizados modelos pré-treinados para inferência em tempo real e suporte a múltiplos idiomas, resultando em um bom desempenho em tarefas de similaridade semântica com baixa latência.

2. Fundamentação Teórica

O crescimento da produção científica mundial tem tornado cada vez mais desafiadora a seleção de revistas científicas adequadas para submissão de manuscritos, sobretudo por pesquisadores iniciantes ou que desejam atuar em novas áreas do conhecimento. Os sistemas de recomendação surgem como ferramentas úteis que auxiliam os autores a selecionar as fontes mais indicadas para publicação dos resultados de suas pesquisas com base em variados critérios, como aderência temática, impacto da revista, modelo de publicação entre outros [Rollins et al. 2017, Ogunde et al. 2020, Gündoğan et al. 2023].

Sistemas de recomendação de conteúdo científico empregam diferentes abordagens, especialmente a baseada na análise de assuntos de pesquisa (recomendação baseada em conteúdo) [Rollins et al. 2017]. Essa abordagem se baseia em informações textuais dos manuscritos, como título, resumo e palavras-chave, para calcular a similaridade entre o conteúdo do artigo e o escopo temático das revistas, sugerindo aquelas mais alinhadas ao tema da pesquisa [Entrup et al. 2022].

A evolução dos sistemas de recomendação de revistas baseados em conteúdo foi impulsionada pelos avanços nos modelos de representação textual no campo da recuperação da informação [Gündoğan et al. 2023]. Inicialmente, abordagens como o Modelo de Espaço Vetorial [Salton et al. 1975] permitiram representar documentos por vetores numéricos, mas enfrentavam limitações para representação do sentido de termos indexados. Para superar essas restrições, novos modelos passaram a incorporar relações

semânticas entre termos, como no Modelo Vetorial Generalizado [Wong et al. 1985] e na Análise Semântica Latente [Deerwester et al. 1990].

Com o crescimento de dados textuais disponíveis em rede, surgiram modelos mais amplos, como o Word2Vec [Mikolov et al. 2013] e o GloVe [Pennington et al. 2014], que modelam significados com base na coocorrência de palavras em grandes conjuntos de dados, estabelecendo as bases para os atuais *embeddings*: vetores numéricos que representam a essência semântica das palavras. Esses modelos são marcos na representação distribuída de palavras, possibilitando a criação de métodos ainda mais avançados como a geração de *embeddings* contextuais, que ampliou ainda mais o campo da recuperação semântica.

Nesse sentido, o modelo ELMo introduziu vetores derivados de estados internos de modelos bidirecionais baseados em LSTMs (biLMs), permitindo a construção de representações sensíveis ao contexto sintático e semântico de cada ocorrência de palavra. Em vez de gerar uma única representação por palavra, o modelo ELMo cria significados mais completos e flexíveis. Para isso, ele combina informações de todas as suas camadas de processamento, adaptando-se melhor a diferentes tarefas [Peters et al. 2018].

O desenvolvimento do modelo BERT levou esse paradigma adiante ao utilizar a arquitetura Transformer com atenção bidirecional mascarada (*masked self-attention*), permitindo capturar dependências contextuais em qualquer direção do texto. Pré-treinado com tarefas como *masked language modeling* e *next sentence prediction*, o BERT gera *embeddings* complexos que impulsionaram o desempenho do processamento de linguagem natural. Cada *token* no BERT é representado por um vetor contextualizado extraído das camadas do Transformer, e esses *embeddings* são ajustáveis via *fine-tuning* em tarefas como análise de sentimento, resposta a perguntas ou reconhecimento de entidades [Devlin et al. 2019].

Para aplicações práticas na recuperação de informação esses *embeddings* podem ser utilizados para indexação vetorial em bancos de dados como FAISS ou VectorSearch, permitindo busca aproximada por similaridade semântica. Recentemente, [Kamalloo et al. 2023] avaliaram o desempenho de APIs modernas de *embeddings* (como Cohere, OpenAI e Hugging Face) em tarefas de recuperação e rerank, revelando que modelos como all-MiniLM e bge-base atingem alta performance com baixo custo computacional.

A escalabilidade dessas soluções também tem sido abordada com o desenvolvimento de bancos vetoriais otimizados para tarefas de recuperação, como o VectorSearch, que integra compactação de índices e estratégias de balanceamento semântico em grandes bases [Monir et al. 2024].

A geração de *embeddings* distribuídos e contextuais baseados em grandes conjuntos de dados e arquiteturas profundas tem revolucionado os sistemas de recuperação semântica, permitindo representações que refletem com precisão a riqueza semântica de textos científicos. Esses avanços são fundamentais para o desenvolvimento de sistemas de recomendação de conteúdo científico e outras aplicações que exigem compreensão semântica profunda e em larga escala.

Com base nos avanços em *embeddings* contextuais e busca vetorial, esta pesquisa propõe o desenvolvimento de um sistema de recomendação de revistas. A abordagem

consistirá em gerar vetores de alta fidelidade semântica a partir de resumos e, em seguida, indexá-los para permitir buscas por similaridade em larga escala. O objetivo é, portanto, construir e validar uma solução prática que aplica o estado da arte da recuperação de informação ao desafio da seleção de periódicos.

3. Procedimentos metodológicos

Foi realizada a carga (*snapshot*) do conjunto de dados de trabalhos (*works*) e revistas (*sources*) do OpenAlex [Priem et al. 2022] em uma infraestrutura em nuvem da Amazon Web Services (AWS), com gerenciamento por meio de uma arquitetura *serverless* e x86_64 e processamento de dados em *cluster* Spark com ferramenta EMR.

Foi utilizado um *script* em linguagem Python (PySpark) para filtrar um amplo conjunto de artigos publicados em aproximadamente 140 mil revistas científicas entre os anos de 2022 e 2024. Foram extraídos os dados textuais dos campos título, palavras-chave e tópicos. O campo tópicos (*topics*) do OpenAlex é resultado de um processo automatizado de classificação temática que combina o uso de um modelo de linguagem de larga escala com uma técnica de clusterização baseada em redes de co-citação [Van Eck and Waltman 2024]. A Figura 1 apresenta uma síntese dos procedimentos metodológicos.

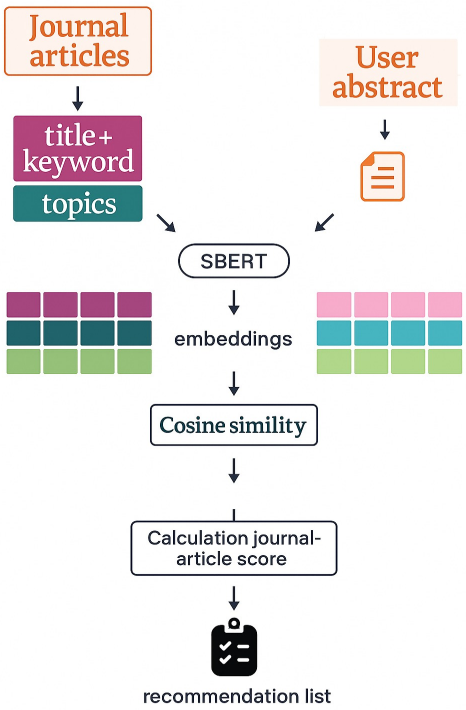

Figura 1. Síntese dos procedimentos metodológicos.

Esse conjunto de dados textuais, provenientes da base OpenAlex, passou por um processo de curadoria para selecionar os campos textuais chave: título (*work_title*), palavras-chave (*keywords*) e tópicos (*topics*). Optou-se por não incluir os resumos (*abstracts*) nesta etapa inicial devido a inconsistências na sua disponibilidade na base de dados, embora a sua incorporação esteja planejada para enriquecer a representação tex-

tual dos artigos em trabalhos futuros. O conteúdo combinado desses campos foi então vetorizado utilizando o modelo paraphrase-multilingual-MiniLM-L12-v2 da arquitetura SBERT, gerando para cada artigo um único *embedding*. Por fim, esses *embeddings* foram armazenados em um banco de dados PostgreSQL para viabilizar a futura recuperação e análise de similaridade.

Para representar tematicamente cada revista, os *embeddings* de todos os artigos publicados em cada revista foram agrupados, calculando-se a média aritmética elemento a elemento dos vetores. Esse procedimento gera um vetor único que sintetiza o perfil semântico de todos os artigos publicados por cada revista.

Esses vetores representativos, que encapsulam o escopo temático de cada revista, foram finalmente armazenados no banco de dados e indexados com HNSW (*Hierarchical Navigable Small World*) para otimizar e acelerar as consultas de similaridade. Este método é conhecido por sua performance em buscas de similaridade em grande escala [Malkov and Yashunin 2020].

O sistema de recomendação foi implementado através de um *backend* em linguagem Python utilizando o *framework* FastAPI. Para garantir respostas de baixa latência, o sistema adota uma estratégia de carregamento em memória, onde os vetores médios das revistas e o modelo são pré-carregados em memória RAM. Ao receber o texto de um resumo de manuscrito fornecido pelo usuário, este é primeiro vetorizado em tempo real para ter a sua similaridade temática comparada com os *embeddings* de cada revista através da Similaridade de Cossenos [Salton et al. 1975], definida pela Equação 1:

$$\text{Similaridade}(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|} \tag{1}$$

Onde $\vec{A}$ é o vetor do resumo passado pelo usuário e $\vec{B}$ é o vetor médio da revista. Para refinar o ranqueamento, o sistema emprega uma métrica de similaridade ajustada, detalhada na Equação 2. Ela utiliza o resultado da Equação 1 e o pondera por um fator que considera a produtividade da revista:

$$\text{Pontuação Final} = \text{Similaridade}(\vec{A}, \vec{B}) \cdot (1 + C) \tag{2}$$

Onde C representa a contagem total de trabalhos publicados pela revista e a aplicação do logaritmo atenua o impacto desse valor, para que revistas com maior volume de artigos publicados não distorçam a relevância temática da base. O ranqueamento final das revistas é, portanto, baseado nesta pontuação composta.

4. Resultados

A eficiência do sistema foi quantificada através de um teste de carga simulado, no qual foram enviadas 50 requisições sequenciais para o *endpoint* responsável pelo ranqueamento. Conforme a Figura 2, a execução completa do teste realizado em um ambiente com os perfis de aproximadamente 140 mil revistas pré-carregadas totalizou 33.98 segundos, resultando em um tempo médio de resposta por requisição de 679.57 milissegundos.

Figura 2. Resultado do teste de desempenho do sistema.

Este desempenho sub-segundo para uma busca em larga escala é atribuído à estratégia de carregamento em memória e ao uso do índice HNSW para a busca de similaridade. A Figura 3 apresenta os resultados do processamento de 611 artigos válidos, ocorrendo a uma velocidade de 2,45 iterações por segundo. A análise dos resultados demonstra uma melhoria progressiva na taxa de acerto, que foi de 20,13% ao considerar os 10 principais periódicos (Top 10) , subiu para 33,55% no Top 30 e atingiu 42,06% no Top 50.

Figura 3. Resultados do processamento de 611 artigos.

Para avaliar a precisão das recomendações foi realizado um teste com um resumo sobre o diagnóstico de retinopatia diabética via Redes Neurais Convolucionais. Os resultados apresentados na Figura 4 mostram revistas multidisciplinares de grande volume, como PLoS ONE e Scientific Reports, nas primeiras posições. Simultaneamente, o sistema recomendou revistas com alta relevância temática, como Diabetes, que se alinha diretamente ao tema do resumo. A presença dessa revista demonstra a capacidade do sistema em capturar o contexto temático específico do resumo e recomendar uma revista especializada no tema.

A classificação de diferentes tipos de revistas no topo da lista evidencia o impacto da pontuação ajustada pela Equação 2. Ao utilizar a contagem de trabalhos como um fator de ponderação para similaridade semântica, o modelo produz um ranqueamento que inclui tanto publicações de grande alcance quanto publicações tematicamente especializadas. Esta abordagem híbrida resulta em uma lista de recomendações balanceada que vai desde publicações altamente especializadas até revistas com um escopo mais amplo.

A capacidade do sistema de operar com tempo de resposta inferior a um segundo, enquanto oferece recomendações que incluem tanto periódicos de alto volume quanto de nicho especializado, é notável, tendo este desempenho sido atingido utilizando o modelo de linguagem em seu estado original, sem qualquer tipo de ajuste fino (*fine-tuning*).

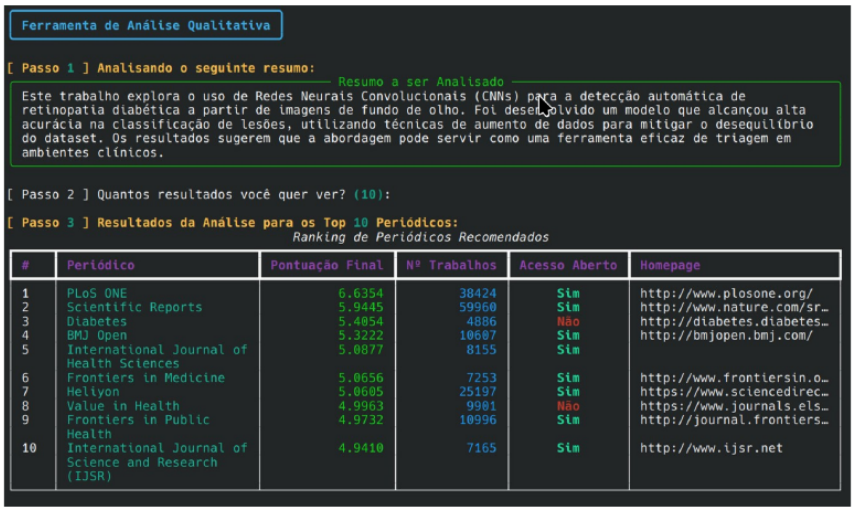

Figura 4. Resultado da análise quantitativa do sistema.

5. Conclusões

Este trabalho detalhou o desenvolvimento e a avaliação de um sistema de recomendação de revistas científicas, que utiliza *embeddings* semânticos e técnicas de processamento de linguagem natural sobre o repositório OpenAlex. A arquitetura demonstrou ser eficiente e precisa, validando sua capacidade de sugerir revistas relevantes a partir do conteúdo semântico de um resumo. Os resultados reforçam o potencial da ferramenta para auxiliar pesquisadores a otimizar o processo de submissão científica e a disseminar suas pesquisas de forma mais estratégica.

Como trabalhos futuros, está em desenvolvimento uma interface interativa em React que permitirá a submissão de resumos e a visualização das recomendações, incluindo suporte para mais de 50 idiomas, filtros avançados por métricas como modelo de acesso, índice-h e custos de publicação. Adicionalmente, planeja-se realizar o *fine-tuning* do modelo de linguagem para aprofundar sua compreensão semântica no domínio científico. O objetivo é aprimorar a capacidade do sistema de capturar as nuances entre diferentes campos de estudo, aumentando assim a acurácia e a relevância contextual das sugestões.

Referências

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Entrup, E., Eppelin, A., Ewerth, R., Hartwig, J., Tullney, M., Wohlgemuth, M., and Hoppe, A. (2022). B!SON: A Tool for Open Access Journal Recommendation. In Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G. M., Golub, K., Ferro, N., and

Poggi, A., editors, *Linking Theory and Practice of Digital Libraries*, volume 13541, pages 357–364. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Gündoğan, E., Kaya, M., and Daud, A. (2023). Deep learning for journal recommendation system of research papers. *Scientometrics*, 128(1):461–481.

Kamalloo, E., Zhang, X., Ogundepo, O., Thakur, N., Alfonso-Hermelo, D., Rezagholizadeh, M., and Lin, J. (2023). Evaluating Embedding APIs for Information Retrieval. arXiv:2305.06300 [cs].

Malkov, Y. A. and Yashunin, D. A. (2020). Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs].

Monir, S. S., Lau, I., Yang, S., and Zhao, D. (2024). VectorSearch: Enhancing Document Retrieval with Semantic Embeddings and Optimized Search. arXiv:2409.17383 [cs].

Ogunde, A. O., Odim, M. O., Olaniyan, O. O., Ojewumi, T. O., Oguntunde, A. O., Fayemiwo, M. A., Olowookere, T. A., and Bolanle, T. H. (2020). The Design of a Hybrid Model-Based Journal Recommendation System. *Advances in Science, Technology and Engineering Systems Journal*, 5(6):1153–1162.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Priem, J., Piwowar, H., and Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833 [cs].

Rollins, J., McCusker, M., Carlson, J., and Stroll, J. (2017). Manuscript Matcher: A Content and Bibliometrics-based Scholarly Journal Recommendation System.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Van Eck, N. J. and Waltman, L. (2024). An open approach for classifying research publications.

Wong, S. K. M., Ziarko, W., and Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval. In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '85*, pages 18–25, Montreal, Quebec, Canada. ACM Press.