



Evento: XXXIV Seminário de Iniciação Científica

RECOMENDAPROF: INTEGRAÇÃO DE PROCESSAMENTO DE LINGUAGEM NATURAL E MODELAGEM MATEMÁTICA NA RECOMENDAÇÃO DE ORIENTADORES DE PÓS-GRADUAÇÃO¹

Luan Vitor Casali Dallabrida², Fabricia Carneiro Roos Frantz³

¹ Projeto de pesquisa desenvolvido na Unijui; com financiamento da agência CNPq.

² Estudante do curso Ciência da Computação, bolsista do Programa Institucional de Bolsas de Iniciação Científica do Conselho Nacional de Desenvolvimento Científico e Tecnológico - PIBIC/CNPq.

³ Professora orientadora do projeto Análise e Integração de Grandes Volumes de Dados

INTRODUÇÃO

A escolha de um orientador em programas de pós-graduação *stricto sensu* trata-se de um processo complexo, exigindo que haja um alinhamento entre os interesses de pesquisa do discente e a trajetória acadêmica do docente. No Brasil, a Plataforma Lattes centraliza esses dados (CNPq, 2026); contudo, a ausência de ferramentas auxiliares e a vasta quantidade de informações não estruturadas dificultam a tomada de decisão. A busca manual força o aluno a dedicar bastante de seu tempo e energia, frequentemente resultando em escolhas baseadas em buscas simples por palavras-chave.

Para preencher esta lacuna, desenvolveu-se o sistema "RecomendaProf", uma aplicação web interativa baseada em um algoritmo híbrido que integra Processamento de Linguagem Natural (PLN) (JURAFSKY; MARTIN, 2023) e modelagem matemática multicritério voltada a sistemas de recomendação (RICCI et al., 2015), para a recomendação personalizada de orientadores de pós-graduação. O presente trabalho consolida a evolução tecnológica das pesquisas do grupo. Iniciativas anteriores basearam-se no uso de Geração Aumentada de Recuperação (RAG) e bancos de dados vetoriais puros (BECKMANN; FRANTZ, 2025). Para resolver o problema da dependência exclusiva de Modelos de Linguagem de Grande Escala (LLMs) e garantir maior portabilidade, o RecomendaProf adota uma arquitetura híbrida e relacional. Esta nova abordagem integra a síntese semântica da Inteligência Artificial a um modelo matemático multicritério estruturado e transparente. A ferramenta objetiva democratizar e otimizar o acesso à informação acadêmica, alinhando-se



ao Objetivo de Desenvolvimento Sustentável 4 (Educação de Qualidade) da Agenda 2030 da ONU, garantindo uma educação inclusiva e equitativa.

METODOLOGIA

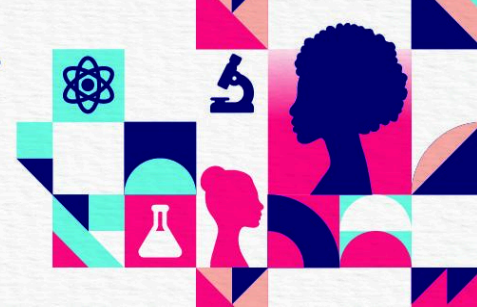
O estudo caracteriza-se como uma pesquisa aplicada e experimental. A amostra de dados foi construída por currículos Lattes, em formato XML, de professores vinculados a programas de pós-graduação das universidades do Consórcio das Universidades Comunitárias Gaúchas (COMUNG).

A arquitetura do software foi desenvolvida em linguagem Python, com uma clara e estrita separação entre a camada de apresentação (Streamlit), o processamento lógico e a persistência de dados. Para garantir a portabilidade e eliminar a necessidade de infraestrutura de servidores complexos, adotou-se o banco de dados relacional SQLite, substituindo iterações anteriores (PostgreSQL e ChromaDB).

O processamento das variáveis e a geração dos dados de recomendação ocorrem em três estágios algorítmicos: O primeiro estágio aplica PLN, usando a biblioteca spaCy, para analisar a descrição em linguagem natural fornecida pelo discente. O texto é submetido à lematização e convertido em matrizes vetoriais esparsas via TF-IDF (Term Frequency-Inverse Document Frequency). Aplicam-se algoritmos de clusterização (K-Means e Birch) para a filtragem semântica inicial.

O segundo estágio, por sua vez, aplica o modelo matemático proposto por Martins (2025). O Índice de Recomendação (IR) é calculado a partir de seis variáveis quantitativas e qualitativas extraídas via *proxies* do currículo XML: Área (P_{Area} , coincidência hierárquica do CNPq), Experiência (P_{Exp} , volume de orientações), Produção (P_{Prod} , densidade bibliométrica), Eficiência (P_{Efi} , taxa de conclusão), Colaboração (P_{Colab} , densidade de rede) e Pesquisa Ativa (P_{Pesq} , produção recente de 4 anos). Os dados brutos de cada variável são normalizados dinamicamente para um intervalo $[0, 1]$. O cálculo segue uma soma ponderada, onde o próprio usuário pode calibrar o peso (α) de cada dimensão conforme seu interesse.

O terceiro estágio, por fim, integra Modelos de Linguagem de Grande Escala (LLMs), operando localmente (Ollama) ou em nuvem (Gemini). Esta camada atua



exclusivamente na síntese de perfis e na explicabilidade da recomendação, sem interferir no cálculo matemático do IR. Diferente da versão anterior do projeto, que exigia bancos vetoriais mais robustos, o sistema atual emprega mecanismos de caching em memória e rotinas de *fallback* automático para os LLMs, garantindo a resiliência e a disponibilidade da aplicação, evitando o recálculo redundante de vetores ou indisponibilidades de API externa.

RESULTADOS E DISCUSSÃO

O desenvolvimento resultou em um sistema funcional operado por uma interface conversacional e um painel de parametrização. A interface assíncrona processa o *input* textual do discente e executa as métricas de similaridade. A análise multifatorial implementada superou tecnicamente a busca tradicional por palavras-chave. Essa abordagem mitigou a redundância informacional dos currículos e gerou recomendações precisas, adaptáveis aos pesos definidos pelo discente (Figura 1).

Figura 1 - Página principal do RecomendaProf no final

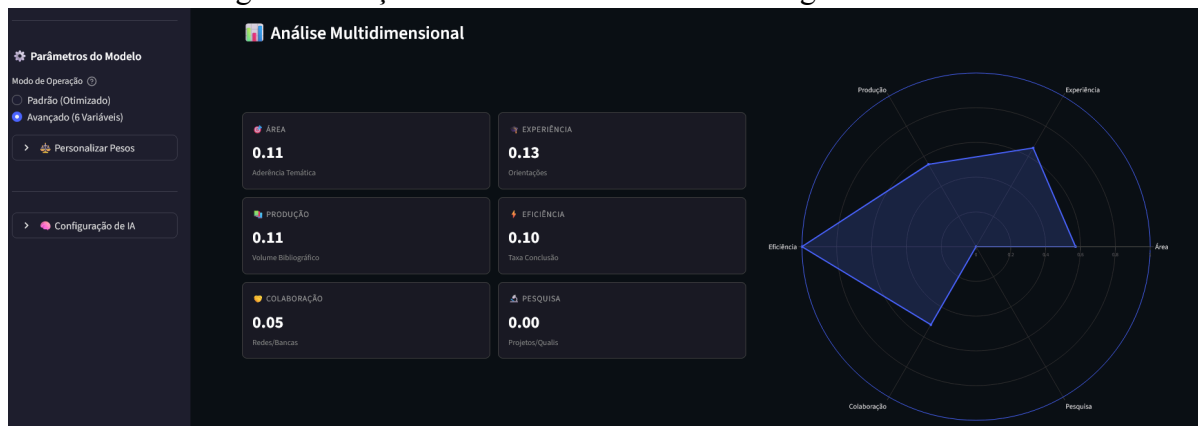


Fonte: Própria do autor (2026).

Diferenciou-se estritamente as variáveis (ex.: publicações, número de orientados) dos dados gerados (a pontuação final normalizada do docente no cluster). A aplicação resolve o problema de sistemas baseados puramente em Inteligência Artificial, conhecidos como "caixa preta", ao implementar um módulo de auditoria visual. Gráficos de radar expõem o desempenho multidimensional do docente nas seis dimensões analisadas, exibindo a matemática por trás da recomendação e garantindo transparência, conforme visto na Figura 2.



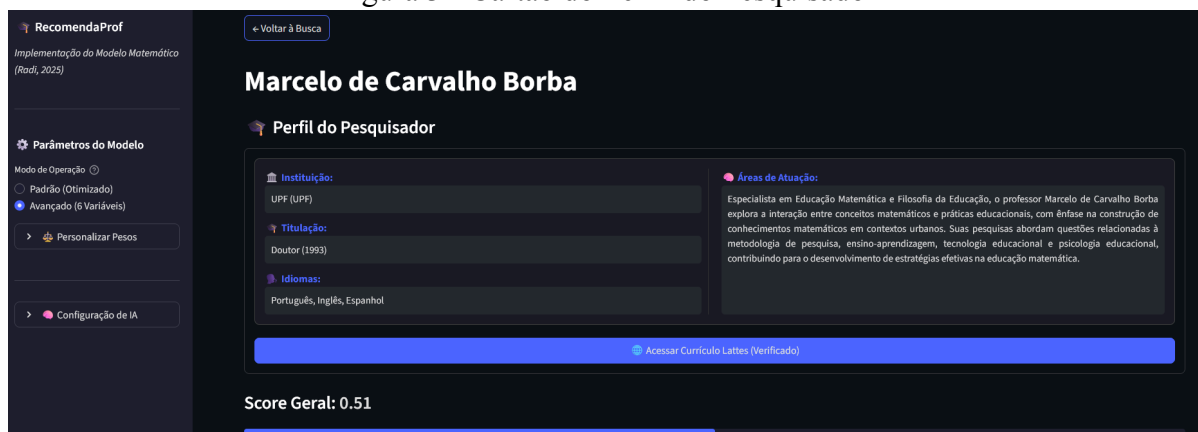
Figura 2 - Seção de análise das variáveis com gráfico de radar



Fonte: Própria do autor (2026).

A avaliação crítica do sistema identifica dependências ligadas à base de dados. A precisão do modelo matemático depende da periódica atualização do Currículo Lattes pelos docentes. Adicionalmente, a utilização de LLMs demanda rigoroso controle de estado de sessão e armazenamento temporário (*caching*) para impedir a degradação do tempo de resposta da interface gráfica. Com uma alternativa, colocada em prática, sendo a de fornecer um botão para levar o discente direto para o Currículo Lattes do docente escolhido, logo abaixo da descrição inicial, conforme demonstrado na Figura 3.

Figura 3 - Cartão do Perfil do Pesquisador



Fonte: Própria do autor (2026).

Comparado à livros que falam de sistemas de recomendação clássicos, que priorizam filtragem colaborativa, o RecomendaProf utiliza uma outra abordagem, baseada em conteúdo e regras multicritério. Esta escolha é mais adequada ao domínio acadêmico, onde a adequação técnica supera o volume de interações de outros usuários.



CONSIDERAÇÕES FINAIS

A integração de Processamento de Linguagem Natural com modelagem matemática provou-se uma solução viável e robusta para a recomendação de orientadores. A substituição do banco de dados para SQLite, documentação completa e executáveis conferiram a portabilidade exigida para a distribuição da ferramenta para um público mais amplo.

A experiência do projeto foi marcada por práticas de engenharia de software e análise de grandes volumes de dados (Big Data). Planos futuros incluem a testagem do modelo computacional com estudantes em formato de ensaio cego, onde atribuirão notas às recomendações, permitindo o cálculo da taxa de acerto do algoritmo em um ambiente real.

Palavras-chave: Big Data. Engenharia de Software. Processamento de Linguagem Natural. Sistemas de Recomendação.

AGRADECIMENTOS

Agradeço ao Grupo de Pesquisa em Computação Aplicada (GCA), à orientadora Fabricia Carneiro Roos Frantz e ao professor Rafael Frantz pela oportunidade e orientação durante a bolsa de Iniciação Científica (PIBIC/CNPq).

REFERÊNCIAS BIBLIOGRÁFICAS

- CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO (CNPq). **Plataforma Lattes**. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 15 jul. 2026.
- JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 3. ed. [S.l.: s.n.], 2023.
- MARTINS, R. M. **Modelagem matemática para recomendação de orientadores em programas de pós-graduação stricto sensu**: uma abordagem baseada na análise de métricas acadêmicas e científicas. 2025. 212 f. Tese (Doutorado em Modelagem Matemática e Computacional) - Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí, 2025.
- RICCI, F. et al. **Recommender Systems Handbook**. Nova York: Springer, 2015.
- BECKMANN, ISADORA; FRANTZ, FABRICIA CARNEIRO ROOS. **CHATBOT ACADÊMICO BASEADO EM RETRIEVAL-AUGMENTED GENERATION PARA INDICAÇÃO DE ORIENTADORES EM PROGRAMAS DE PÓS GRADUAÇÃO**. Salão do Conhecimento, [S. l.], v. 11, n. 11, 2025. Disponível em: <https://publicacoeseventos.unijui.edu.br/index.php/salaconhecimento/article/view/27888>. Acesso em: 5 set. 2026.