

ANÁLISE DE MÉTODOS ESPARSOS E DENSOS DE RECUPERAÇÃO DA INFORMAÇÃO PARA TRIAGEM AUTOMATIZADA DE CURRÍCULOS CIENTÍFICOS

ANALYSIS OF SPARSE AND DENSE IR METHODS FOR AUTOMATED SCREENING OF SCIENTIFIC RESUMES

Gabriel Filipe Possela

Faculdade de Tecnologia de Porto Ferreira, FATEC

E-mail: gabrielfilipeposs@hotmail.com

Nicolas Alves de Oliveira

Faculdade de Tecnologia, FATEC

E-mail: nickoliver152637@gmail.com

Felipe Ivo da Silva

Universidade Federal de São Carlos, UFSCar

Faculdade de Tecnologia de São Paulo, FATEC

E-mail: felipe_ivodasilva@hotmail.com

RESUMO

As transformações digitais nos processos de recrutamento geraram um volume substancial de currículos, criando novos desafios para os departamentos de recursos humanos. Com isso, as abordagens tradicionais de busca baseadas apenas em palavras-chave exatas tornaram-se ineficientes, pois sofrem com limitações linguísticas inerentes, como polissemia e sinonímia, introduzindo uma lacuna semântica entre a intenção do recrutador e os documentos recuperados. No cenário de cidades de pequeno porte, esse problema é ainda mais crítico devido à forte dependência de recomendações informais de terceiros, o que dificulta a inserção de novos talentos recém-qualificados no mercado de trabalho. Diante disso, esta pesquisa apresenta o uso de métodos avançados de Recuperação de Informação (RI), combinando sistemas híbridos de busca e Grandes Modelos de Linguagem (LLMs) para a estruturação de um banco de currículos inteligente e automatizado. Por meio de análise bibliográfica e da avaliação de benchmarks públicos de larga escala — como BEIR, MS MARCO e Natural Questions (NQ) —, demonstra-se que a integração de métodos esparsos (baseados em termos) e densos (baseados em vetores) mitiga as limitações semânticas. Essa abordagem garante maior exatidão e precisão na triagem de candidatos. Por fim, o uso dessas ferramentas otimiza os processos

seletivos locais, gerando economia de tempo para recrutadores e candidatos, retendo talentos e impulsionando o desenvolvimento do mercado de trabalho regional.

Palavras-chave: Recuperação de Informação. Triagem de Currículos. Grandes Modelos de Linguagem. Sistemas Híbridos.

ABSTRACT

The digital transformations in recruitment processes have generated a substantial volume of resumes, creating new challenges for human resources departments. Consequently, traditional search approaches based solely on exact keywords have become inefficient, as they suffer from linguistic limitations such as polysemy and synonymy, introducing a semantic gap between the recruiter's intent and the retrieved documents. In the scenario of small towns, this problem becomes even more critical due to a strong reliance on informal third-party recommendations, which hinders the insertion of newly qualified talents into the local market. Given this context, this research presents the application of advanced Information Retrieval (IR) methods, combining hybrid search systems and Large Language Models (LLMs) for the structuring of an intelligent and automated resume database. Through a comprehensive bibliographical analysis and the evaluation of large-scale public benchmarks—specifically BEIR, MS MARCO, and Natural Questions (NQ)—it is demonstrated that the integration of sparse (term-based) and dense (vector-based) methods mitigates these semantic limitations. This hybrid approach guarantees greater accuracy and precision in candidate screening. Ultimately, the use of these tools optimizes local selection processes, generating significant time savings for both recruiters and job seekers, retaining talent, and boosting the development of the regional job market.

Keywords: Information Retrieval. Resume Screening. Large Language Models. Hybrid Systems.

REFERÊNCIAS

DEEKSHITHA, N.; REDDY, B. R. An LLM-Based Resume Intelligence System for Recruiters and Job Seekers. **Research Digest on Engineering Management and Social Innovations**, v. 2, n. 3, p. 144–151, 2026.

FORMAL, T.; PIWOWARSKI, B.; CLINCHANT, S. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2288–2292, 2021.

GAO, L.; DAI, Z.; CALLAN, J. **COIL**: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3030–3042, 2021.

GAO, Y.; XIONG, Y.; GAO, X.; JIA, K.; PAN, J.; BI, Y.; WANG, H. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, v. 2, n. 1, p. 32, 2023.

GUNJAL, A. Bridging the Lexical Semantic Gap: A Comprehensive Review of Hybrid Information Retrieval, 2025, dezembro 25.

JÄRVELIN, K.; KEKÄLÄINEN, J. Cumulated gain-based evaluation of IR techniques. **ACM Transactions on Information Systems (TOIS)**, v. 20, n. 4, p. 422–446, 2002.

KHATTAB, O.; ZAHARIA, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 39–48, 2020.

MALKOV, Y. A.; YASHUNIN, D. A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 42, n. 4, p. 824–836, 2018.

MENEZES JÚNIOR, A. A. D. Um método para busca de competências a partir de currículos lattes [Dissertação (Mestrado em Informática)]. Universidade Federal do Amazonas [61 f.], 2012.

ZHANG, H.; WANG, Y.; CHEN, Q.; CHANG, R.; ZHANG, T.; MIAO, Z.; CUI, B. Model-enhanced vector index. **Advances in Neural Information Processing Systems**, v. 36, p. 54903–54917, 2023.