

Da Pergunta à Evidência: busca científica híbrida

From Question to Evidence: hybrid scientific search

Rayssa Gomides Marconato – rayssagomidesmarconato@gmail.com
Victor Hugo Barros Baptista Pereira – victorhugobbpbaptista02@gmail.com

INTRODUÇÃO

A busca por referências é uma das etapas fundamentais de uma pesquisa científica, mas encontrar trabalhos realmente relevantes pode se tornar uma tarefa complexa diante do crescente volume de publicações disponíveis. Parte dessa dificuldade está no fato de que uma mesma ideia pode ser descrita de diferentes formas: o pesquisador formula sua questão utilizando determinados termos, enquanto um artigo potencialmente relevante pode empregar outro vocabulário para abordar o mesmo conceito. Métodos tradicionais de Recuperação de Informação, como o BM25, são eficientes na correspondência entre os termos presentes na busca e nos documentos (Robertson; Zaragoza, 2009), enquanto modelos de recuperação semântica buscam identificar relações de significado mesmo quando não há correspondência direta entre as palavras (Reimers; Gurevych, 2019). Entretanto, estudos demonstram que abordagens semânticas não necessariamente superam métodos lexicais em todos os domínios (Thakur et al., 2021), o que evidencia características potencialmente complementares entre essas estratégias. A partir desse cenário, este trabalho busca investigar se a combinação entre recuperação lexical e semântica pode melhorar a identificação e o ranqueamento de artigos relevantes para uma questão de pesquisa, comparando experimentalmente métodos lexicais, semânticos e híbridos em coleções científicas de referência.

DESENVOLVIMENTO

A avaliação experimental utilizou três coleções científicas do benchmark BEIR (Thakur et al., 2021): NFCorpus, SciFact e TREC-COVID, totalizando mais de 180 mil documentos. Foram comparadas quatro estratégias de recuperação: BM25, baseada em correspondência lexical (Robertson; Zaragoza, 2009); recuperação semântica por representações vetoriais; uma abordagem híbrida, combinando sinais lexicais e semânticos; e o reranqueamento neural dos resultados recuperados (Nogueira; Cho, 2019). O desempenho foi avaliado principalmente pelo $nDCG@10$, complementado por métricas de precisão e cobertura. Os resultados mostraram que a abordagem híbrida obteve o melhor $nDCG@10$ no NFCorpus (0,3351) e no SciFact (0,6698), enquanto no TREC-COVID o híbrido associado ao reranqueamento alcançou 0,7083, frente a 0,5678 do BM25, representando ganho de 24,74%.

CONCLUSÃO

Os resultados demonstram que a combinação entre recuperação lexical e semântica pode melhorar a identificação e o ranqueamento de literatura científica, confirmando o potencial de complementaridade entre essas abordagens. O método híbrido apresentou o melhor desempenho em duas das três coleções avaliadas, enquanto sua associação ao reranqueamento obteve o maior ganho no TREC-COVID. Como continuidade, pretende-se investigar modelos ajustados a domínios científicos específicos e aplicar a abordagem em uma ferramenta de apoio à busca por referências

acadêmicas.

REFERÊNCIAS

NOGUEIRA, R.; CHO, K. Passage re-ranking with BERT. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1901.04085>. Acesso em: 13 setembro 2026.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND THE INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (EMNLP-IJCNLP), 2019, Hong Kong. Proceedings [...]. Hong Kong: Association for Computational Linguistics, 2019. p. 3982-3992. Disponível em: <https://doi.org/10.18653/v1/D19-1410>.

ROBERTSON, S.; ZARAGOZA, H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, v. 3, n. 4, p. 333-389, 2009. Disponível em: <https://doi.org/10.1561/15000000019>.

THAKUR, N. et al. BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS DATASETS AND BENCHMARKS TRACK, 35., 2021, Online. Proceedings [...]. [S.l.: s.n.], 2021. Disponível em: <https://openreview.net/forum?id=wCu6T5RI5Z>. Acesso em: 13 setembro 2026.