

Desempenho da Mesclagem de Bases Bibliográficas via NBVIZ

Performance of Bibliographic Database Merging via NBVIZ

Leonardo Neves Bolfarini – leonardobolfarini@gmail.com

Rafael Gutierrez Castanha – rafaelcastanha@unimar.br

INTRODUÇÃO

A análise cienciométrica depende de metadados bibliográficos organizados e comparáveis. Entretanto, a cobertura e a estrutura dos registros variam entre fontes como Scopus, Web of Science e OpenAlex, o que torna necessária a padronização antes da produção de indicadores e redes científicas (Visser; Van Eck; Waltman, 2021; Priem; Piwowar; Orr, 2022). Este trabalho avaliou o desempenho do fluxo já implementado no NBVIZ, ecossistema de análises de dados científicos, para integrar registros dessas três bases. O objetivo foi verificar o comportamento da mesclagem e da deduplicação em diferentes volumes de dados, identificando a permanência de registros e o tempo necessário para gerar os arquivos de saída.

DESENVOLVIMENTO

Foram definidos três cenários de teste a partir da expressão de busca “graph”: 1.000, 10.000 e 20.000 registros de cada base, que somados correspondem a 3.000, 30.000 e 60.000 registros de entrada. Cada cenário foi executado cinco vezes online, no servidor de produção do NBVIZ, pela própria interface web. Assim, cada execução percorreu o fluxo real da aplicação: o front-end enviou os dados à API, o backend realizou leitura e padronização e a biblioteca nbviz_scientometric_tools executou a mesclagem e a deduplicação. Foram medidos o tempo da requisição no endpoint e, separadamente, o processamento isolado da mesclagem e da gravação dos arquivos de saída.

A deduplicação foi estável nas cinco repetições. Nos cenários de 3.000, 30.000 e 60.000 registros, foram preservados, respectivamente, 2.907, 27.645 e 55.720 registros; as remoções corresponderam a 3,10%, 7,85% e 7,13%. Para reportar a variação, foram calculadas média e mediana. O processamento isolado no backend, composto por mesclagem e gravação, apresentou médias e medianas de 0,045/0,043 s, 0,216/0,219 s e 0,363/0,370 s. Já o tempo da requisição apresentou médias/medianas de 11,877/11,604 s, 95,003/94,800 s e 231,105/189,726 s.

A média e a mediana são apresentadas conjuntamente porque respondem a papéis diferentes: a média evidencia a variação de todas as execuções, enquanto a mediana representa melhor a execução típica quando há valores extremos. No cenário de 60.000 registros, uma execução de 401,973 s elevou a média do tempo de requisição, mas não alterou de modo relevante o processamento isolado do backend. A diferença entre cerca de 0,37 s de processamento e 189,726 s na requisição mediana indica que o gargalo está no restante do fluxo da requisição, e não na mesclagem. Esse tempo inclui leitura e padronização, além da consulta paginada à OpenAlex; como as fases ainda não foram temporizadas individualmente, não é correto atribuir toda a latência a uma única fonte.

A continuidade proposta não é reescrever a lógica de mesclagem, que já se mostrou eficiente. Pretende-se instrumentar separadamente consulta, leitura, padronização, deduplicação e gravação. Só então serão avaliadas medidas como reaproveitamento temporário de resultados de buscas repetidas e processamento assíncrono de consultas extensas, preservando atualização e rastreabilidade dos dados.

CONCLUSÃO

O fluxo completo do NBVIZ, executado online pela interface e pelo backend com apoio da biblioteca de processamento, manteve resultados de deduplicação consistentes nos três volumes. A mesclagem e a gravação permaneceram abaixo de 0,4 s mesmo com 60.000 registros, enquanto a requisição completa apresentou maior tempo e variação. Os resultados indicam que a prioridade é diagnosticar as etapas externas à mesclagem e não substituir o algoritmo atual.

REFERÊNCIAS

PRIEM, J.; PIWOWAR, H.; ORR, R. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2205.01833>. Acesso em: 10 setembro 2026.

VISSER, M.; VAN ECK, N. J.; WALTMAN, L. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies, v. 2, n. 1, p. 20-41, 2021. Disponível em: https://doi.org/10.1162/qss_a_00112.