ISBN: 9786588884546

ANDSOL: PESQUISA E DESENVOLVIMENTO DE SOLUÇÕES PARA DESAMBIGUAÇÃO DE NOMES DE AUTORES EM REPOSITÓRIOS BIBLIOGRÁFICOS DIGITAIS

NATAN DE SOUZA RODRIGUES; GUSTAVO HENRIQUE COSTA PINTO; BARTOLOMEU SPEGIORIN GUSELLA

RESUMO

A ambiguidade de nomes de autores é um problema comum em repositórios bibliográficos digitais, impactando a identificação precisa de pesquisadores e a integridade das informações armazenadas. Este projeto tem como objetivo desenvolver soluções inovadoras para a Desambiguação de Nomes de Autores (AND), abordando a criação de ferramentas e abordagens práticas que buscam facilitar e permitir a resolução do problema de AND. Inicialmente, foi realizada uma revisão da literatura para identificar técnicas consolidadas, resultando no desenvolvimento de uma abordagem híbrida que combina Processamento de Linguagem Natural (PLN), Redes Convolucionais de Grafos (RCG) e Clusterização Hierárquica. Essa abordagem permite integrar informações semânticas e estruturais para aumentar a precisão na identificação de autores. Também foi desenvolvida uma Interface Gráfica de Usuário (GUI), implementada com a biblioteca Java Swing, para facilitar a seleção de conjunto de dados e formatos de entrada, ampliando a acessibilidade e compatibilidade com métodos existentes. Paralelamente, está em desenvolvimento um software para padronização de conjunto de dados (*datasets*), essencial para garantir comparações robustas entre diferentes métodos e a reprodutibilidade das análises. Resultados preliminares demonstram que a abordagem híbrida apresenta alta precisão ao combinar técnicas avançadas de aprendizado de máquina e análise de redes, com métricas de validação comparáveis ao estado da arte. As próximas etapas incluem a validação do método proposto utilizando métricas como Precisão, Revocação, Medida F1 e Métrica K, além da divulgação dos resultados em eventos científicos. Conclui-se que a combinação de técnicas avançadas e ferramentas acessíveis pode oferecer soluções eficazes para os desafios de AND, promovendo avanços na organização e recuperação de dados em repositórios bibliográficos digitais.

Palavras-chave: AND; Repositórios Bibliográficos Digitais; Ambiguidade de Nomes de Autores

1 INTRODUÇÃO

Repositórios bibliográficos digitais, como o DBLP, ArnetMiner e CiteSeerX (DBLP, 1993-2025; AMiner, 2005-2025; CiteSeerX, 2007-2019), desempenham um papel crucial na organização, disseminação e recuperação de informações acadêmicas em diversas áreas do conhecimento. Essas plataformas fornecem acesso a grandes volumes de dados científicos, permitindo que pesquisadores localizem publicações relevantes e acompanhem tendências globais de pesquisa. Contudo, o crescimento exponencial das publicações científicas intensificou o problema da ambiguidade de nomes de autores, dificultando a identificação precisa de pesquisadores e suas contribuições, o que afeta a qualidade e confiabilidade das análises realizadas nesses repositórios.

A resolução desse problema é um desafio significativo. A literatura propõe uma ampla

DOI: 10.51189/iv-ensipex/49524

ISBN: 9786588884546

gama de abordagens, variando de métodos baseados em heurísticas a soluções avançadas que utilizam inteligência artificial, incluindo aprendizado supervisionado e não supervisionado. Essas técnicas compõem um campo de estudo conhecido como Desambiguação de Nomes de Autores (*Author Name Disambiguation* - AND) (Ferreira et al., 2020). Análises recentes de co-citação e acoplamento bibliográfico destacam também que técnicas de agrupamento de autores têm se mostrado eficazes em grandes bases de dados, enquanto o uso de *embeddings* de palavras e aprendizado supervisionado emergem como tendências promissoras, com plataformas como AMiner e DBLP amplamente utilizadas para extração de informações (Rodrigues et al., 2024a). Além da diversidade de abordagens, a uniformização de datasets utilizados na tarefa de AND surge como um elemento essencial para garantir comparabilidade e confiabilidade dos resultados. A padronização facilita uma avaliação mais robusta e transparente das diferentes metodologias, contribuindo para avanços consistentes na área. Essas soluções ilustram a diversidade de métodos adotados para enfrentar os desafios da AND, ressaltando a necessidade de ferramentas e técnicas integradas que aprimorem essa tarefa em repositórios bibliográficos digitais.

Seguindo essa perspectiva, este projeto de pesquisa concentra-se no estudo, desenvolvimento e otimização de novas técnicas para AND. Como etapa inicial, estamos criando uma Interface Gráfica de Usuário (GUI) para aprimorar uma abordagem já existente, tornando-a mais eficiente e flexível. Além disso, já desenvolvemos uma abordagem híbrida para AND que combina Processamento de Linguagem Natural (PLN), Redes Convolucionais de Grafos (RCG) e técnicas de clusterização. Paralelamente, estamos trabalhando no desenvolvimento de um software voltado à padronização de conjunto de dados (*datasets*) utilizados na tarefa de AND. A expectativa é que essa GUI e as demais ferramentas desenvolvidas possam ser integradas a futuras soluções para AND, ampliando sua aplicabilidade e impacto.

Objetivos Gerais

- Criar softwares e ferramentas voltados à resolução de problemas de AND em repositórios bibliográficos digitais.
- Realizar avaliações de desempenho das soluções propostas, comparando-as com metodologias existentes na literatura.
- Divulgar os resultados em conferências e periódicos especializados, além de apoiar trabalhos de conclusão de curso relacionados ao tema.

2 MATERIAL E MÉTODOS

Este projeto iniciou-se com uma revisão da literatura para identificar técnicas e abordagens empregadas na tarefa de AND, fundamentando-se em estudos como (Rodrigues et al., 2024a) e (Sanyal et al., 2019), que analisaram os principais fronts de pesquisa na área. Essa etapa incluiu também a análise de taxonomias e métodos automáticos propostos por (Ferreira et al., 2020), que classificou as principais abordagens voltadas à resolução do problema de AND.

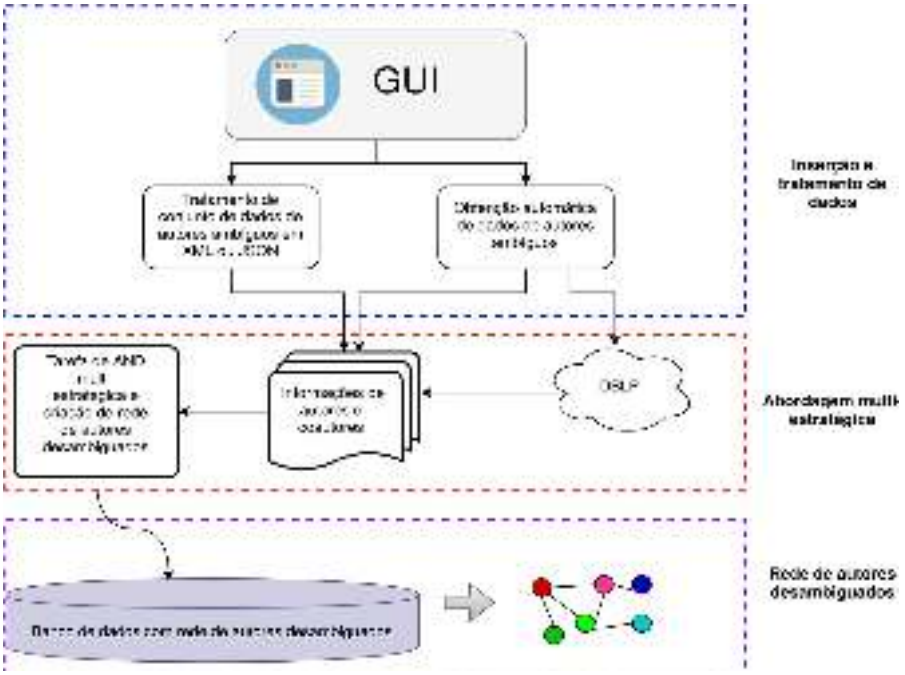
Na segunda etapa, foi desenvolvida uma abordagem híbrida que combina técnicas de PLN, RCG e Clusterização Hierárquica. O método híbrido utiliza um modelo de PLN para extração de embeddings textuais de títulos e resumos, enquanto a RCG e a Clusterização Hierárquica analisam as interações globais em redes heterogêneas. Essa abordagem permitiu integrar informações semânticas e estruturais, aumentando a precisão na tarefa de AND (Rodrigues et al., 2024b). Paralelamente, desenvolvemos uma GUI que facilita a seleção de datasets e formatos de entrada, como JSON e XML, para aplicação em métodos de AND. A GUI foi implementada utilizando a biblioteca Java Swing e integrada a uma abordagem multi-estratégica já existente na literatura (Rodrigues et al., 2021), ampliando a acessibilidade e flexibilidade das técnicas de AND e garantindo compatibilidade com métodos anteriores e

DOI: 10.51189/iv-ensipex/49524

ISBN: 9786588884546

novos desenvolvimentos do projeto. A Figura 1 ilustra o fluxo de trabalho da integração da abordagem multi-estratégica com a GUI, que inicialmente permite a escolha entre formatos de arquivos ou a obtenção automática do repositório DBLP.

***Figura 1**: Fluxo de trabalho da implementação da GUI com a abordagem multi-estratégica.*



Além disso, foi iniciado o desenvolvimento de um *software* para padronização de *datasets*, com o objetivo de uniformizar os dados utilizados em AND. A padronização é fundamental para permitir comparações mais robustas entre diferentes métodos e assegurar a reprodutibilidade das análises realizadas.

As etapas subsequentes incluem:

- Coleta e preparação de *datasets* para validação das soluções desenvolvidas.
- Avaliação das soluções utilizando métricas como Precisão, Revocação, Medida F1, ACP, AAP e Métrica K.
- Publicação dos resultados em conferências e periódicos relevantes, contribuindo para o avanço do estado da arte na área de AND.

Essa combinação de técnicas e ferramentas visa oferecer soluções inovadoras para os desafios de AND, promovendo avanços significativos para resolução desse problema em repositórios bibliográficos digitais.

3 RESULTADOS E DISCUSSÃO

Os resultados parciais obtidos até o momento demonstram avanços significativos no desenvolvimento de soluções para a tarefa de AND. Um dos principais marcos foi a criação de uma GUI que aprimora uma abordagem consolidada na literatura. Essa GUI permite que o usuário selecione *datasets* e formatos de entrada, como JSON e XML, facilitando a integração de diferentes fontes de dados e aumentando a acessibilidade para usuários finais. O fluxo de trabalho foi implementado com sucesso, proporcionando uma experiência prática e intuitiva. A Figura 2 ilustra a interface desenvolvida, e testes preliminares indicam que a GUI atende plenamente aos requisitos funcionais propostos.

DOI: 10.51189/iv-ensipex/49524

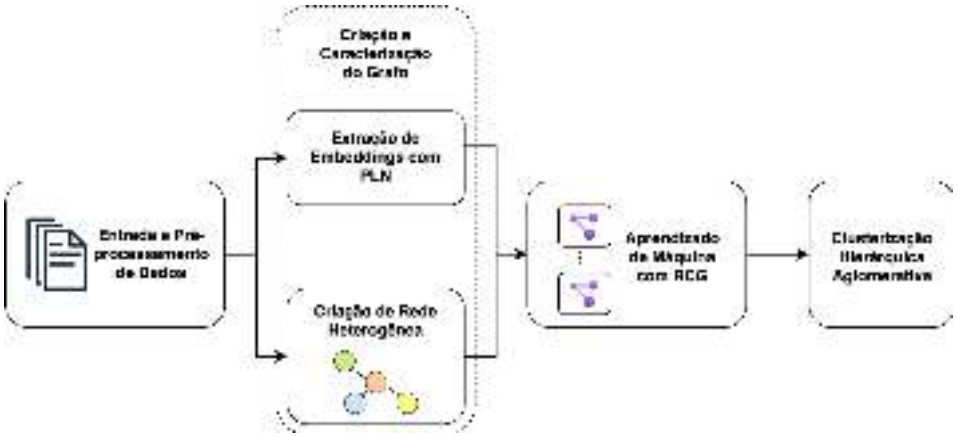
ISBN: 9786588884546

***Figura 2**: Interface GUI desenvolvida.*

Adicionalmente, está em desenvolvimento um software para padronização de *datasets*, projetado para uniformizar dados utilizados na tarefa de AND. Esse software visa garantir a comparabilidade entre diferentes métodos, formatando e processando os dados de entrada de maneira consistente e eficiente. Essa ferramenta desempenha um papel essencial na melhoria da qualidade das análises realizadas e no aumento da reprodutibilidade de estudos na área.

Outro ponto importante deste projeto foi a formulação de uma abordagem híbrida que combina PLN, RCG e Clusterização Hierárquica, publicada em (Rodrigues et al., 2024b). A Figura 3 apresenta o fluxo de trabalho da integração dessa abordagem híbrida que possibilita a captura de informações semânticas e estruturais de redes heterogêneas, resultando em uma precisão superior na identificação de autores.

***Figura 3**: Fluxo de Trabalho da Abordagem Híbrida para AND (adaptado).*



Conforme a Tabela 1, resultados preliminares evidenciam a eficácia da abordagem híbrida, com métricas como Precisão, Revocação, Medida F1 e Métrica K atingindo valores elevados. Grupos como "A Choudhary" e "J Martin" alcançaram 100% em todas as métricas, demonstrando alta precisão na desambiguação. O método de AND, (Waqas e Qadir., 2021), foi utilizado como linha de base. Em comparação ao método de linha de base, a abordagem híbrida apresentou melhorias no Medida F1 (0.95 vs. 0.935) e Métrica K (0.969 vs. 0.912), destacando-se como uma solução eficiente para AND em repositórios bibliográficos digitais.

DOI: 10.51189/iv-ensipex/49524

ISBN: 9786588884546

***Tabela 1**: Métricas de validação entre grupos de nomes ambíguos de autores.*

Grupo de nomes ambíguos	Autores	Precisão	Revocação	Medida F1	ACP	AAP	Métrica K
A Choudhary	12	1.0	1.0	1.0	1.0	1.0	1.0
J Martin	9	1.0	1.0	1.0	1.0	1.0	1.0
M A Qadir	15	1.0	1.0	1.0	1.0	1.0	1.0
J Mitchell	10	1.0	1.0	1.0	1.0	1.0	1.0
A Gupta	8	0.853	0.878	0.865	0.875	0.875	0.875
J Robinson	12	1.0	1.0	1.0	1.0	1.0	1.0
A Kumar	9	1.0	1.0	1.0	1.0	1.0	1.0
J Smith	12	0.938	0.988	0.964	0.972	0.972	0.972
Bin Li	8	0.592	0.671	0.632	0.763	0.889	0.826
S Kim	10	0.754	0.944	0.839	0.897	0.895	0.896
D Eppstein	3	1.0	1.0	1.0	1.0	1.0	1.0
Z Zhang	10	1.0	1.0	1.0	1.0	1.0	1.0
J Lee	8	1.0	1.0	1.0	1.0	1.0	1.0
K Tanaka	11	1.0	1.0	1.0	1.0	1.0	1.0
Linha de base [Waqas and Qadir 2021]	137	0.946	0.925	0.935	0.958	0.87	0.912
Our Method	137	0.938	0.963	0.95	0.965	0.974	0.969

4 CONCLUSÃO

Este projeto de pesquisa tem apresentado resultados significativos para a tarefa de AND, conforme proposto inicialmente, destacando-se pela criação de uma GUI funcional e de um software para padronização de *datasets*, que aprimoram a acessibilidade e a reprodutibilidade das análises. Além disso, foi desenvolvida uma abordagem híbrida para AND, que combina PLN, RCG e Clusterização Hierárquica, superando métodos de linha de base na literatura em métricas como Medida F1 e Métrica K, demonstrando alta precisão.

Entre as limitações estão a dependência de dados bem estruturados e o alto custo computacional para grandes redes. Futuras etapas incluem validações com *datasets* diversificados, otimização dos modelos e a exploração de técnicas avançadas, como *Graph Attention Networks* (GAT) e *Large Language Models* (LLMs). As soluções desenvolvidas demonstram grande potencial para superar os desafios de AND, promovendo avanços na organização e recuperação de informações em repositórios bibliográficos digitais.

REFERÊNCIAS

AMINER. Search and mining of academic social networks. Tsinghua University, Beijing, 100084, China, 2005-2025. Disponível em: https://www.aminer.org/. Acesso em: 8 jan. 2025.

CITESEERX. Scientific literature digital library and search engine. Pennsylvania State University, University Park, PA, USA, 2007-2019. Disponível em: https://citeseerx.ist.psu.edu. Acesso em: 8 jan. 2025.

DBLP. The digital bibliography & library project. Schloss Dagstuhl, Leibniz-Zentrum für Informatik, LZI GmbH, 1993-2025. Disponível em: https://dblp.uni-trier.de/. Acesso em: 8 jan. 2025.

FERREIRA, A. A.; GONÇALVES, M. A.; LAENDER, A. H. Automatic disambiguation of author names in bibliographic repositories. Morgan & Claypool Publishers, 2020.

DOI: 10.51189/iv-ensipex/49524

ISBN: 9786588884546

RODRIGUES, N. S.; COSTA, A. R.; LEMOS, L. C.; RALHA, C. G. Multi-strategic approach for author name disambiguation in bibliography repositories. In: LOSSIO-VENTURA, J. A.; VALVERDE-REBAZA, J. C.; DÍAZ, E.; ALATRISTA-SALAS, H. (Ed.). Information Management and Big Data. SIMBig 2020. Communications in Computer and Information Science, v. 1410. Cham: Springer, 2021.

RODRIGUES, N. S.; MARIANO, A. M.; RALHA, C. G. Author name disambiguation literature review with consolidated meta-analytic approach. International Journal on Digital Libraries, Cham, p. 1–21, 2024.

RODRIGUES, N. S.; RALHA, C. G. A hybrid machine learning method to author name disambiguation. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 15., 2024, Belém/PA. Anais [...]. Porto Alegre: SBC, 2024. p. 108–117. DOI: 10.5753/stil.2024.245440. Disponível em: https://sol.sbc.org.br/index.php/stil/article/view/31122. Acesso em: 8 jan. 2025.

SANYAL, D. K.; BHOWMICK, P. K.; DAS, P. P. A review of author name disambiguation techniques for the PubMed bibliographic database. Journal of Information Science, v. 47, n. 2, p. 227-254, 2021. DOI: 10.1177/0165551519888605. Disponível em: https://doi.org/10.1177/0165551519888605. Acesso em: 8 jan. 2025.

WAQAS, H.; QADIR, M. A. Multilayer heuristics based clustering framework (MHCF) for author name disambiguation. Scientometrics, Dordrecht, v. 126, n. 9, p. 7637–7678, 2021.