



ACOPLAMENTO BIBLIOGRÁFICO ENTRE AUTORES ENQUANTO CRITÉRIO ADICIONAL PARA DESAMBIGUAÇÃO DE NOMES DE PESQUISADORES: primeiras aproximações

Rafael Gutierrez Castanha

 <https://orcid.org/0000-0002-3117-1780>.

✉ fr.castanha@gmail.com.

🏢 Universidade de Marília (Unimar) |

ROR: <https://ror.org/01w96sp43> | Marília, Brasil.

Raidan Cruz Silveira

 <https://orcid.org/0000-0002-8976-3391>.

✉ raidan.silveira@unesp.br.

🏢 Universidade Estadual Paulista (Unesp) |

ROR: <https://ror.org/00987cb86> | Marília, Brasil.

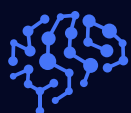
Eixo temático: Métodos Bibliométricos e de Citação

Modalidade: Resumo expandido

DOI: 10.22477/ix.ebbc.236

Resumo: Avalia o acoplamento bibliográfico entre autores como critério adicional para desambiguação de nomes de pesquisadores. Utilizando um conjunto de 31 pesquisadores vencedores da medalha *Solla Price* na *Web of Science*, dos quais 11 apresentam ambiguidades, identificaram-se 19 perfis ambíguos. Ao realizar o acoplamento bibliográfico entre perfis oficiais e ambíguos, constatou-se que em 16 casos os autores se acoplaram aos seus perfis ambíguos, enquanto em três não. Desses 16, 62,5% correspondiam ao mesmo pesquisador. Nos três acoplamentos nulos, nenhum correspondia à mesma pessoa. Conclui-se que o acoplamento bibliográfico é um potencial critério adicional para desambiguação de nomes de pesquisadores.

Palavras-Chave: Acoplamento bibliográfico. Ambiguidade. Bibliometria. Desambiguação.



1 INTRODUÇÃO

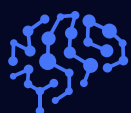
A ambiguação de dados em pesquisas científicas pode ser considerada um entrave substancial para seu desenvolvimento, visto que, cada vez mais, a fidedignidade dos dados é necessária. No âmbito bibliométrico, isso não é diferente. A busca por autores em bases de dados científicas, especialmente novos pesquisadores, frequentemente recuperam artigos de autores de outras áreas em que, a ambiguidade de nomes, emerge como um problema inevitável na vida acadêmica deste pesquisador impactando a recuperação de dados (Zhang; Ban, 2020).

Assim sendo, a desambiguação de nomes de autores é um caso especial de desambiguação de entidades que não são rotuladas com identificadores únicos (D'Angelo; Giuffrida; Abramo, 2011). Ainda segundo D'Angelo, Giuffrida e Abramo (2011), o problema de desambiguação de nomes de autores é geralmente dividido em duas frentes: a) separar as instâncias de múltiplos autores que compartilham o mesmo nome e; b) identificar instâncias de um autor com nomes diferentes (por exemplo: "Maria Ippoliti", "Valeria Ippoliti," e "Maria Valeria Ippoliti). Ambos os problemas são igualmente relevantes em bancos de dados bibliométricos e podem ter várias causas, incluindo nomes idênticos, variações de nome, pseudônimos ou grafias alternativas, abreviações de nome, erros ortográficos, erros tipográficos e erros originados do Reconhecimento Óptico de Caracteres (OCR) (D'Angelo; Giuffrida; Abramo, 2011).

Com isso, a desambiguação do nome do autor pode ser considerado um pré-requisito importante para as análises dos dados de publicação em nível de autor em que os algoritmos de desambiguação geralmente se baseiam em metadados para além dos nomes dos autores para distinguir entre autores com o mesmo nome (Schulz *et al.*, 2014). Nesse sentido, Digiampietri e Ferreira (2018) entendem que para enfrentar alguns dos desafios da desambiguação de nomes de autores é comum a utilização de informações complementares, além do nome dos autores, e entre as informações mais utilizadas estão o título das publicações, o ano das publicações, os veículos nos quais as publicações foram realizadas, informações sobre a vinculação do autor, além de identificadores únicos como ORCID e *ResearchID*.

Para além destes elementos, a depender da base de dados em que as publicações dos autores estão indexadas, é possível ter acesso ao conjunto de referências citadas. Admitindo que o conjunto de referências citadas por um determinado pesquisador, especificamente autores citados, pode configurar, segundo White (2001), sua identidade de citação, é plausível admitir que esta identidade, representada pelo conjunto de autores citados, pode ser um critério adicional para desambiguação de nomes de autores. Isto é, comparar a semelhança da lista de autores citados por pesquisadores com nomes similares poderá configurar um método para desambiguação de nomes de autores. Contudo, a comparação por semelhança de listas de autores recitados nada mais é, no âmbito bibliométrico, do que um processo de Acoplamento Bibliográfico de Autores (ABA), como proposto por Zhao e Strotmann (2008).

Diante do exposto, esta pesquisa pretende responder ao seguinte problema: de que maneira o método de ABA pode contribuir, enquanto critério adicional, para a desambiguação de nomes de pesquisadores? Tendo por objetivo, avaliar o uso do ABA como critério adicional para a desambiguação de nomes de pesquisadores acadêmicos em bases de dados. Como prova de conceito, utiliza-se o conjunto dos vencedores da medalha Solla Price com produções indexadas na base de dados *Web of Science* (WoS). Como pressuposto, assume-se que a intersecção entre referências (autores citados em comum;



acoplamento entre autores) das produções de dois ou mais autores na base pode ser um indício de que esses autores correspondem a um único autor.

2 METODOLOGIA

A partir do website da *International Society of Scientometrics and Informetrics* (ISSI)¹, foram selecionados os vencedores da medalha Solla Price (de 1984 a 2023) que tiveram seus perfis verificados na base WoS. Para a busca foi utilizado o padrão “Sobrenome, Nome” e selecionados somente os autores com perfis verificados (perfil oficial) pelo próprio pesquisador e chancelado pela WoS. Foram considerados oficiais os perfis reivindicados pelos próprios pesquisadores em que a base atribui a chancela (“selo”) de verificação. Este perfil reúne informações sobre suas publicações, métricas de produção e citação, financiamento de pesquisa e revisões por pares².

Os que não atendiam a esse critério foram excluídos. A coleta foi realizada em 8 de dezembro de 2023. Dos 32 medalhistas, 13 possuem perfil oficial, de modo que 11 perfis possuem ambigüações. Assim, compõem a população de análise os seguintes autores: Ludo Waltman, Judit Bar-Ilan, Mike Thelwall, Olle Persson, Péter Vinkler, White, Howard White, Loet Leydesdorff, John Irvine, Andras Schubert, Francis Narin, Henry Small. Para cada perfil oficial, foi realizado acoplamento entre o perfil oficial e os respectivos perfis que sugerem ambigüação.

Para cada perfil, oficial ou possivelmente ambíguo, foram exportados os dados e extraídas as referências citadas (somente autores citados) via *VosViewer* e realizado acoplamento via software *the coupler* de Castanha (2022) identificando a intensidade de acoplamento entre cada perfil oficial e suas ambigüações (Tabela 1). Para verificação da acurácia do método proposto, foram utilizados os cálculos de sensibilidade e especificidade (quinta e sexta colunas da Tabela 1), amplamente aplicados à teste clínico nas Ciências da Saúde, e aqui adaptados ao ABA.

A sensibilidade, refere-se a razão entre os verdadeiros resultados positivos, neste caso, a quantidade de perfis com frequência de acoplamento maior ou igual a 1 que de fato referem-se aos mesmos pesquisadores, e a quantidade de positivos, representado neste cenário pelos perfis com acoplamento entre si. Já a especificidade refere-se a razão entre a quantidade de verdadeiros negativos e a quantidade total de negativos, sendo representado neste método pela divisão entre a quantidade de não acoplamentos, em que os autores não acoplados não são a mesma pessoa e o total de acoplamentos nulos (não acoplamentos; ABA=0) entre os perfis. A verificação dos perfis serem de fato as mesmas pessoas se deu de maneira qualitativa a partir da análise manual dos metadados de cada perfil disponibilizado pela WoS, a saber: instituição, localização, departamento, nomes publicados, organizações publicadas, principais revistas, categorias de assuntos. Todos os dados de pesquisa foram disponibilizados via re-

1 INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS. **Derek de Solla Price Memorial Medal**: Price award winners. Leuven, Belgium: ISSI, 2024. Disponível em: <https://www.issi-society.org/awards/derek-de-solla-price-memorial-medal>. Acesso em: 03 mar. 2024.

2 WEB OF SCIENCE. **About Web of Science Researcher Profiles**. Clarivate: Estados Unidos, 2023. Disponível em: <https://webofscience.help.clarivate.com/en-us/Content/wos-researcher-profile.html>. Acesso em: 03 mar. 2024.

positório digital Zenodo em Silveira e Castanha³ (2024).

3 RESULTADOS

A partir da coleta realizada, tem-se na Tabela 1, o acoplamento entre os 11 perfis de pesquisadores vencedores da medalha Solla Price e seus perfis ambíguos. Os nomes dos pesquisadores com perfis oficiais e não oficiais, foram transcritos à Tabela 1 da mesma maneira com que são apresentados na WoS, mas incluídas numerações a frente dos autores que possuem nomes idênticos a fim de diferenciá-los nas análises realizadas. Dentre estes pesquisadores, verificou-se que em média, cada um possui 1,8 perfil ambíguo ao oficial, em que os pesquisadores com maior número de variações foram *John Irvine*, *Mike Thelwall* e *Howard White* com quatro variações de seus nomes. Ao acoplar cada perfil de cada pesquisador às suas possíveis ambiguações, foram realizados 19 acoplamentos, dos quais, três são nulos: entre John Irvine e dois perfis não, e, entre *Mike Thelwall* e o perfil não oficial *Thelwall, Mike* (*Mike, Tonnie Chanell*). É interessante observar que *Mike Thelwall* apresentou mais de uma variação de seu nome em dois perfis não oficiais: *Mike, Tonnie Chanell* e *Mike, Toko*. Este tipo de situação se assemelha ao segundo caso mencionado anteriormente por D'Angelo, Giuffrida e Abramo (2011) de identificar instâncias de um autor com nomes diferentes.

Tabela 1 – Acoplamento bibliográfico entre perfis oficiais e perfis não oficiais (ambíguos)

Nome dos perfis oficiais	Perfis não oficiais (ambíguos)	#ABA	Correspon- dem ao mes- mo pesquisa- dor?	Sensibilidade (serem os mesmos au- tores quando há acopla- mento)	Especificidade (não serem os mesmos autores quan- do não há aco- plamento)
Bar-Ilan, Judit	Bar-Ilan, Judit	224	Sim		-
Irvine, John (Irvine, J)	Irvine, John (1)	1	Não	x	-
Irvine, John (Irvine, J)	Irvine, John (1)	0	Não	-	ok
Irvine, John (Irvine, J)	Irvine, John (1)	0	Não	-	ok
Ludo Waltman	Waltman, Ludo	40	Sim	ok	-
Narin, Francis (Narin, F)	Narin, Francis	20	Sim	ok	-
Persson, Olle	Persson, Olle	48	Sim	ok	-
Schubert, Andras	Schubert, Andras	60	Sim	ok	-
Small, Henry G (Small, Henry)	Small, Henry	60	Sim	ok	-
Vinkler, Péter (Vinkler, Peter)	Vinkler, P (Vinkler, Peter)	37	Sim	ok	-

³ CASTANHA, Rafael Gutierrez; SILVEIRA, Raidan Cruz. Conjunto de dados utilizados no artigo “**Acoplamento bibliográfico entre autores enquanto critério adicional para desambiguação de nomes de pesquisadores: primeiras aproximações**”. Conjunto de dados. Zenodo [Site], publicado em 2 fev. 2024, version v1. DOI: <https://doi.org/10.5281/zenodo.10611614>. Disponível em: <https://zenodo.org/records/10611615>. Acesso em: 02 fev. 2024.



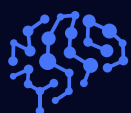
Nome dos perfis oficiais	Perfis não oficiais (ambíguos)	#ABA	Correspondem ao mesmo pesquisador?	Sensibilidade (serem os mesmos autores quando há acoplamento)	Especificidade (não serem os mesmos autores quando não há acoplamento)
Leydesdorff, Loet	Leydesdorff, Loet (Leydesdorff, L)	27	Sim	ok	-
Thelwall, Mike	Thelwall, Mike (1)	469	Sim	ok	-
Thelwall, Mike	Thelwall, Mike (2)	67	Sim	ok	-
Thelwall, Mike-	Thelwall, Mike (Mike, Tonnie Chanell)	0	Não	-	ok
Thelwall, Mike-	Thelwall, Mike (Mike, Toko)	1	Não	x	-
White, Howard D (White, Howard D.)	White, Howard D.	29	Não	x	-
White, Howard D (White, Howard D.)	White, Howard (1)	24	Não	x	-
White, Howard D (White, Howard D.)	White, Howard (2)	13	Não	x	-
White, Howard D (White, Howard D.)	White, Howard (3)	9	Não	x	-
				sensibilidade = 62,5%	especificidade = 100%

Legenda: "ok": denota acerto da operação; "x": denota erro da operação.

Fonte: Elaborado pelos autores (2024).

Os demais cálculos de acoplamento (16 do total) resultaram em valores não nulos (diferentes e maiores que zero) de modo que dois foram iguais a 1, sendo estes os menores valores não nulos. A seguir, tendo realizado todo processo de acoplamento entre perfis, passou-se para avaliação da acurácia do método em que, primeiramente, foi verificado se os perfis correspondiam ou não aos mesmos autores. Esta avaliação se deu pelos demais metadados disponibilizados pela base em cada perfil analisado de modo que, foram considerados que dois perfis correspondiam ao mesmo autor caso ambos compartilhassem informações muito próximas, principalmente, com relação aos vínculos institucionais, localização, departamento, nomes publicados, organizações publicadas, principais revistas, categorias de assuntos.

Dos acoplamentos nulos (iguais a zero), verificou-se que os perfis ambíguos de *John Irvine* com nomes de *Irvine, John* (2) e *Irvine, John* (3) possuem características bastantes diferentes ao perfil oficial do pesquisador, de modo que ambos perfis não oficiais correspondem: à localização "*Christchurch, New Zealand*", tem como organização publicada a "*Christchurch Hospital New Zealand*", e ambos possuem como áreas *Urology & Nephrology* como categorias de assuntos. A partir disso é possível inferir que, os

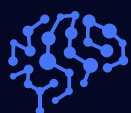


dois perfis ambíguos não correspondem ao perfil oficial, pois, o oficial corresponde a instituição "*University of Sussex*", localização "*England*", organização publicada "*University of Sussex*" e categoria *Science & Technology - Other Topics Business, Economics Physics Social Sciences - Other Topics e History & Philosophy of Science*. Já o perfil ambíguo *Invine, John* (1) correspondeu a instituição "*Queensland University of Technology (QUT)*", localização "*Austrália*", organização publicada "*Queensland University of Technology (QUT)*" e categoria de assuntos *Computer Science Imaging Science & Photographic, Technology, Engineering, Optics Remote Sensing*. Assim, verificou-se que os autores ambíguos possuem características distintas ao perfil oficial de *John Irvine*. Ademais, o perfil ambíguo *Invine, John* (1) acoplou-se somente por um autor citado em comum em que este autor acoplador (unidade de acoplamento) é o próprio nome *John Irvine*, sugerindo um acoplamento por autocitação. Em caso de cálculos de acoplamento, cujo as autocitações não fossem consideradas, o acoplamento entre *John Irvine* e *Invine, John* (1) não ocorreria.

Em relação ao outro perfil que apresentou acoplamento nulo, entre *Mike Thelwall* e *Thelwall, Mike* (*Mike, Tonnie Chanell*), nota-se que dentre todas as características (metadados), não se verificou nenhuma semelhança que fosse possível atribuir vínculo entre eles, visto que o perfil oficial tem como instituição a "*University of Wolverhampton*", localização "*Wolverhampton, England*" e categorias *Information Science & Library Science, Computer Science e Social Sciences*, enquanto este perfil ambíguo possui como vínculo institucional a *Marshall University*, localizada em Huntington e possui como categoria *Medical Informatics Computer Science*. Além deste acoplamento nulo, o perfil ambíguo *Thelwall, Mike* (*Mike, Toko*) com ABA de 1 também não sugeriu nenhum outro vínculo com o perfil oficial, visto que, suas principais categorias de assuntos foram *Nursing, Psychiatry e Biophysics*. Já os outros dois perfis ambíguos, *Thelwall, Mike* (1) e (2), corresponderam igualmente aos vínculos do perfil oficial. Nesse caso, infere-se que os perfis ambíguos *Thelwall, Mike* (1) e (2) correspondem ao perfil oficial.

Nesse sentido, os demais perfis que apresentaram ambiguidade e acoplamento não nulo, com exceção aos de *Howard White*, possuem características que estabelecem vínculos iguais ou próximos à instituição, local, departamento, organizações publicadas, principais revistas e categorias de assuntos próximos. Nesses casos, deduz-se que os perfis ambíguos correspondem aos perfis oficiais para cada autor. A exceção remete ao perfil oficial *Howard White* por apresentar o comportamento mais adverso perante o método proposto, visto que, dentre quatro ambiguações, todas se acoplaram ao perfil oficial. Ao avaliar os metadados disponibilizados pela WoS, não foram encontradas evidências que sugerissem que estes quatro pesquisadores correspondem ao perfil oficial, dado a dessemelhança entre estas informações, como por exemplo as áreas encontrados relacionadas aos quatro perfis ambíguos: *Biophysics, Biochemistry & Molecular Biology, Endocrinology & Metabolism, General & Internal Medicine, Science & Technology, Social Sciences - Other Topics, Government & Law, Business & Economics, Development Studies, Public Administration, Environmental & Occupational Health e Health Care Sciences & Services*.

Ao verificar as áreas referentes ao perfil original, tem-se: *Information Science & Library Science, Computer Science, Communication, Linguistics e Business & Economics*, nota-se somente uma área em comum. Assim sendo, é possível considerar que o acoplamento entre *Howard White* e os demais perfis ambíguos, como um potencial falso positivo, em que demonstra acoplamento, porém, não se referem ao mesmo pesquisador. Desta maneira, ao verificar os cálculos de sensibilidade e especificidade, notou-se que, dos 16 acoplamentos não nulos, em 10 deles, de fato o perfil ambíguo corresponde ao mesmo perfil oficial do pesquisador, gerando uma sensibilidade do método de 62,5% ($10/16 = 0,625$). Porém, ao



observar o não acoplamento ($ABA = 0$), observou-se que, dentre os 3 acoplamentos nulos, em nenhum destes casos os perfis ambíguos correspondem ao mesmo pesquisador, gerando uma especificidade de 100%. Ou seja, foi capaz de verificar que, perante um acoplamento nulo, em todos os casos, os pesquisadores de fato não correspondem à mesma pessoa.

4 CONSIDERAÇÕES FINAIS

Esta pesquisa apresentou as primeiras discussões para avaliar o uso do ABA como critério adicional para a desambiguação de nomes de pesquisadores em bases de dados utilizando, enquanto corpus para prova de conceito, o conjunto de pesquisadores vencedores da medalha Solla Price com perfil oficial na WoS. Todo processo de avaliação gerou alguns pontos de destaque: a) todos os autores ambíguos com acoplamento nulo ($ABA = 0$) de fato não correspondem aos autores dos perfis oficiais; b) com exceção do perfil oficial de Howard White (potencial falso positivo), todos os outros autores que tiveram acoplamento maior que um ($ABA > 1$) apresentaram vínculos correspondentes e/ou similares a instituição, localização, departamento, nomes publicados, organizações publicadas, principais revistas e categorias. Com isso, o método proposto, apresentou, perante este corpus de análise, valores de sensibilidade e especificidade iguais a 62,5% e 100%, respectivamente. Neste caso, o método comportou-se melhor no processo de desambiguação (dessemelhança), aferindo que nenhum dos acoplamentos nulos correspondem aos mesmos autores. Para verificação e possivelmente combinação de perfis, o método foi menos preciso (acertamento em 10 dos 16 acoplamentos).

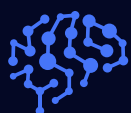
Sendo assim, pode-se dizer que o ABA pode contribuir como critério adicional para o processo de desambiguação de perfis, sobretudo, ao constatar acoplamento nulo entre perfis com possível ambiguidade. Por fim, compreendendo o desafio imposto pelo fenômeno da ambiguidade de nomes, espera-se aplicar este método a outros contextos adicionando mais complexidade ao processo de ABA, como a relativização dos cálculos via cosseno de Salton ou índice de Jaccard, identificação de unidades de acoplamento e número de citações recebidas pelos autores citados.

REFERÊNCIAS

CASTANHA, Rafael Gutierrez. The Coupler: uma nova ferramenta bibliométrica para análises relacionais de citação, acoplamento bibliográfico e cocitação. **RDBCi: Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 20, p. 1-29, 1 dez. 2022. Universidade Estadual de Campinas. DOI: <https://doi.org/10.20396/rdbci.v20i00.8671208>. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8671208>. Acesso em: 20 jan. 2024.

D'ANGELO, Ciriaco Andrea; GIUFFRIDA, Cristiano; ABRAMO, Giovanni. A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. **Journal of the American Society for Information Science and Technology**, United States, v. 62, n. 2, p. 257-269, 23 nov. 2010. DOI: <http://dx.doi.org/10.1002/asi.21460>. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/asi.21460>. Acesso em: 05 jan. 2024.

DIGIAMPIETRI, Luciano Antonio; FERREIRA, João Eduardo Desambiguação de nomes de autores para a



identificação automática de perfis acadêmicos. **Em Questão**, Porto Alegre, p. 37-54, 19 abr. 2018. DOI: <http://dx.doi.org/10.19132/1808-5245242.37-54>. Disponível em: <https://seer.ufrgs.br/index.php/Em-Questao/article/view/74064/45895>. Acesso em: 01 jan. 24.

SCHULZ, Christian; MAZLOUMIAN, Amin; PETERSEN, Alexander M.; PENNER, Orion; HELBING, Dirk. Exploiting citation networks for large-scale author name disambiguation. **EPJ Data Science**, Germany, v. 3, n. 1, p. 1-14, 25 set. 2014. DOI: <http://dx.doi.org/10.1140/epjds/s13688-014-0011-3>. Disponível em: <https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-014-0011-3>. Acesso em: 15 jan. 2024.

WHITE, Howard D. Authors as citers over time. **Journal of the American Society for Information Science and Technology**, United States, v. 52, n. 2, p. 87-108, 2001. DOI: [http://dx.doi.org/10.1002/1097-4571\(2000\)9999:99993.0.co;2-t](http://dx.doi.org/10.1002/1097-4571(2000)9999:99993.0.co;2-t). Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-4571%282000%299999%3A9999%3C%3A%3AAID-ASI1542%3E3.0.CO%3B2-T>. Acesso em: 02 jan. 2024.

ZHANG, Lizhi.; BAN, Zhijie. Author Name Disambiguation Based on Rule and Graph Model. **Natural Language Processing and Chinese Computing, Switzerland**, p. 617-628, 2020. DOI: http://dx.doi.org/10.1007/978-3-030-60450-9_49. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-60450-9_49. Acesso em: 01 jan. 2024.

ZHAO, Dangzhi.; STROTMANN, Andreas. Evolution of research activities and intellectual influences in information science 1996–2005: introducing author bibliographic coupling analysis. **Journal of the American Society for Information Science and Technology**, United States, v. 59, n. 13, p. 2070-2086, 16 jul. 2008. DOI: <http://dx.doi.org/10.1002/asi.20910>. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20910>. Acesso em: 10 jan. 2024.