O DESAFIO DA DEDUPLICAÇÃO DE PUBLICAÇÕES:

criação e avaliação de um *benchmark*

Jesús P. Mena-Chalco
https://orcid.org/0000-0001-7509-5532.
jesus.mena@ufabc.edu.br.
Universidade Federal do ABC (UFABC) |
https://ror.org/028kg9j04 | São Paulo, Brasil.

Fabio Lorensi do Canto
https://orcid.org/0000-0002-8338-1931.
fabio.lc@ufsc.br.
Universidade Federal de Santa Catarina (UFSC) |
https://ror.org/041akq887 | Florianópolis, Brasil.

Washington Luís Ribeiro de Carvalho Segundo
https://orcid.org/0000-0003-3635-9384.
washingtonsegundo@ibict.br.
Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) |
https://ror.org/006c42y96 | Brasília, Brasil.

Thiago Magela Rodrigues Dias
https://orcid.org/0000-0001-5057-9936.
thiagomagela@cefetmg.br.
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) |
https://ror.org/04ch49185 | Belo Horizonte, Brasil.

Tales Henrique José Moreira
https://orcid.org/0000-0002-9865-6918.
tales.info@gmail.com.
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) |
https://ror.org/04ch49185 | Belo Horizonte, Brasil.

Eixo temático: Bases e Fontes de Dados
Modalidade: Resumo expandido
DOI: 10.22477/ix.ebbc.411

Resumo: O objetivo deste artigo é apresentar um conjunto de 10 regras usadas para a criação de um *benchmark* e sua avaliação usando casamentro aproximado baseado na similaridade de Levenshtein. A finalidade prática é de trazer insumos para investigar o desafio da deduplicação de publicações. Após avaliação, algumas regras apresentaram desafios, ressaltando a complexidade da deduplicação e a necessidade por estratégias mais sofisticadas à de casamento aproximado. A análise das publicações não deduplicadas revela uma queda acentuada com diferentes valores de similaridade, enfatizando a necessidade de ajustar parâmetros conforme o contexto. Este trabalho caminha na direção da definição de estratégias eficazes e abrangentes para a deduplicação de artigos científicos.

Palavras-Chave: Deduplicação. Publicações científicas. *Benchmark*. Regras.

1 INTRODUÇÃO

Nas aplicações ou bases que consolidam dados acadêmicos de diversas fontes, é comum encontrar múltiplas instâncias de um mesmo elemento, como um artigo catalogado simultaneamente na *Web of Science* e na *Scopus*. Devido à divergência nos protocolos de registro de dados entre diferentes fontes, é frequente deparar-se com publicações duplicadas que apresentam características e detalhes distintos. Embora existam iniciativas para a implementação de identificadores persistentes, como o *Digital Object Identifier* (DOI) nas publicações, o desafio da duplicação está presente, especialmente em relação a publicações anteriores a 2000.

A deduplicação (He; Li; Zhang, 2010), concebida como a identificação de instâncias únicas em uma base de dados permeada por redundâncias, é uma questão central no cenário da Ciência da Informação. No contexto das publicações científicas, esse desafio assume uma complexidade relevante e permanece relativamente subexplorado. A identificação eficaz de artigos científicos duplicados não apenas aprimora a integridade, completude e a qualidade de bases de dados acadêmicas, mas também impacta diretamente a confiabilidade e a relevância das pesquisas conduzidas. Nesse sentido, pode-se entender que, a deduplicação de publicações bibliográficas é fundamental para a qualidade, integridade e eficiência da gestão de informações em bases de dados acadêmicas. Ela combate a repetição e o erro nos registros de artigos, possibilitando buscas mais precisas e análises mais confiáveis.

Ao eliminar duplicatas, os pesquisadores garantem resultados mais relevantes e eficientes, facilitando revisões sistemáticas com a contagem única de cada artigo. Isso contribui para a precisão dos dados e a recuperação de informações nas bases de dados, otimizando o tempo e o esforço dos pesquisadores (Jiang et al., 2014). Surge, assim, a indagação: qual percentagem de similaridade é mais apropriada para a deduplicação de artigos científicos e em quais contextos essa decisão é mais relevante?

O objetivo deste trabalho é apresentar um conjunto abrangente de 10 regras elaboradas para a criação de um *benchmark* (Saavedra; Smith, 1996) voltado à deduplicação de publicações científicas. Esse *benchmark* não apenas visa fornecer um conjunto diversificado de instâncias para avaliação de algoritmos de deduplicação, mas também estabelece um protocolo sistemático para testes futuros.

Nas próximas seções são apresentadas as regras, delineando suas implicações práticas. Além disso, foi testada a eficácia desse *benchmark* utilizando o algoritmo de casamento aproximado de Levenshtein (Ukkonen, 1985), um método consagrado na identificação de similaridade entre cadeias de caracteres que indica o quão próximo estão duas cadeias de texto.

Neste trabalho, busca-se preencher a lacuna existente na compreensão da deduplicação de publicações científicas. Este trabalho aborda não somente a aplicação prática de regras específicas, mas também contribui para a construção de uma base de referência concreta (aqui no artigo, denominado de *benchmark*) que pode ser útil para futuros desenvolvimentos de estratégias de deduplicação.

2 REGRAS PARA A CRIAÇÃO DO *BENCHMARK*

A criação de um *benchmark* para a deduplicação de publicações é uma etapa importante para

garantir a eficácia e a confiabilidade dos métodos utilizados para descoberta das publicações únicas. É importante que as regras estabelecidas para a construção desse *benchmark* sejam sistemáticas, seguindo um protocolo determinístico que minimize vieses e assegure a consistência nas interpretações e resultados.

Na Tabela 1 são apresentadas 10 regras elaboradas para proporcionar um guia robusto para a criação de instâncias do *benchmark*, focando especificamente nos títulos das publicações.

Tabela 1 - Regras usadas para a geração do *benchmark*

Regra	Descrição	Exemplo correto	Exemplo com erro
1	Exclusão de prefixo no título	A Strategy for Identifying Specialists in Scientific Data Repositories 2022	Strategy for Identifying Specialists in Scientific Data Repositories 2022
2	Variação no número de substantivos no título (singular/plural)	Analyzing Brazilian technical production: an approach considering patent records 2023	Analyzing Brazilian technical production: an approach considering patent record 2023
3	Variação nos anos de publicação	Journal self-citation on the h5-index of Ibero-American journals 2023	Journal self-citation on the h5-index of Ibero-American journals 2022
4	Erro de codificação de caracteres (ISO / UTF-8 / ASCII)	A musicobiografização como intriga narrativa: um ensaio teórico entre pesquisa (auto) biográfica e educação musical	A musicobiografizaÃ§Ã£o como intriga narrativa: um ensaio teÃ³rico entre pesquisa (auto)biogrÃ¡fica e educaÃ§Ã£o musical
5	Caracteres não imprimíveis	Gestión del conocimiento y capital intelectual según variables sociodemográficas en docentes Universitarios	[Non-Printable Characters] Gestión del conocimiento y capital intelectual según variables sociodemográficas en docentes Universitarios [Non-Printable Characters]
6	Erros tipográficos aleatórios	Scientific Discoveries in the Field of Biology	Scientific Discoveries in the Fiel of Biolgy
7	Inserção de código HTML no texto	Analyzing Scientific Data: A Comprehensive Approach	Analyzing Scientific Data: A Comprehensive Approach
8	Presença apenas do título, sem subtítulo após ":"	Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective	Data Science and Analytics

Regra	Descrição	Exemplo correto	Exemplo com erro
9	Presença de subtítulo, mas sem título após ":"	Improve the Security of Industrial Control System: A Fine-Grained Classification Method for DoS Attacks on Modbus/TCP	A Fine-Grained Classification Method for DoS Attacks on Modbus/TCP
10	Caracteres especiais como aspas, pontos, ponto e vírgula	Exploring "Big Data": Challenges and Opportunities	Exploring Big Data Challenges and Opportunities

Fonte: Elaborado pelos autores, 2024.

Essas regras, que consideram títulos de artigos como entrada, são um ponto de partida para a criação de benchmarks mais completos que considerem outros elementos, como lista de coautores, palavras-chave, veículo (nome do evento ou periódico) e ano de publicação.

Finalmente, a construção de um *benchmark* de qualidade é um trabalho em andamento que requer a colaboração de diferentes especialistas (e.g., Bibliotecários, Cientistas da Informação e da Computação) e ainda está sendo realizado no contexto de um projeto de pesquisa maior.

As 10 regras permitem alterar a grafia do título (não o formato, por exemplo, variação de maiúsculas e minúsculas). Estas regras estabelecem uma base inicial significativa para a elaboração de um *benchmark* que será de grande utilidade para a comunidade científica.

3 *BENCHMARK* PARA INVESTIGAR ESTRATÉGIAS DE DEDUPLICAÇÃO

O *benchmark* desenvolvido para investigar estratégias de deduplicação foi concebido mediante um protocolo rigoroso. Para a criação do *benchmark*, foi imprescindível contar com uma lista de publicações que servisse como padrão-ouro ou gabarito, proporcionando uma base para a avaliação dos algoritmos de deduplicação.

Em janeiro de 2024, foi realizada uma consulta à Web of Science com o objetivo de obter uma lista representativa de publicações científicas. A busca **(string = (AB=(Brasil OR Brazil) OR TI=(Brasil OR Brazil)) AND PY=(2010-2023))** foi direcionada a produções que contivessem a palavra “Brasil” em seus títulos ou resumos, publicadas no período entre 2010 e 2013 e que, adicionalmente, fossem consideradas como altamente citadas.

Essa abordagem estratégica permitiu não apenas abranger um intervalo temporal de 24 anos, mas também garantir que as publicações continham todas as informações completas (observa-se uma correlação positiva entre a completude dos dados de uma publicação e a sua probabilidade de ser altamente citada).

A consulta resultou em uma lista inicial composta por 544 publicações que atendiam aos critérios estabelecidos. Essa seleção inicial, embora de ordem menor, foi importante para garantir a representatividade das publicações no contexto brasileiro e assegurar a diversidade de temas e abordagens. A aplicação das 10 regras (definidas na seção anterior) permitiu a geração de um conjunto expandido de

3843 publicações. A introdução de variações linguísticas e a inclusão de ruído proporcionam um ambiente diversificado e realista, refletindo as complexidades encontradas em bases bibliométricas reais.

Para uma análise detalhada e consulta mais aprofundada, o leitor é convidado a acessar a planilha *Excel*, que contém os resultados da consulta à *Web of Science*, além das publicações geradas de acordo com as diretrizes das 10 regras apresentadas na Tabela 1[1].

Finalmente, é importante destacar que a escolha da *Web of Science*, como fonte primária para a formação do padrão-ouro, conferiu ao *benchmark* uma base de qualidade, uma vez que essa base é reconhecida por sua abrangência e rigor na indexação de publicações científicas.

4 AVALIAÇÃO

A avaliação do *benchmark*, composto por 3843 publicações geradas através das 10 regras previamente estabelecidas, desempenha um papel importante na compreensão da relevância de um referencial importante para a definição das melhores estratégias de deduplicação de dados. Nesta seção, exploramos a importância prática do *benchmark* ao empregar o algoritmo de casamento aproximado, com base na similaridade de Levenshtein, considerando intervalos de similaridade que variam de 60% a 100%.

Cada uma das 3843 publicações foi submetida a uma comparação com o gabarito, e o casamento foi avaliado se a similaridade atingisse valores dentro do intervalo estabelecido. A Tabela 2 apresenta a porcentagem de identificação de similaridade, onde as células destacadas em verde indicam uma identificação perfeita (i.e., 100%). Esse cenário ocorre quando todas as publicações geradas pelas regras são identificadas de forma precisa.

O algoritmo de casamento aproximado, fundamentado na similaridade de Levenshtein, permitiu uma análise inicial da capacidade do *benchmark* em identificar e deduplicar publicações de maneira precisa. Ao considerar valores de similaridade em uma faixa de 60% a 100%, abordamos uma grande variedade de cenários que simulam diferentes níveis de correspondência entre as instâncias. Na Tabela 2, são destacados resultados que revelam uma superfície de Pareto na relação entre acertos e perdas (Kaufman; Klevs, 2022). Isso significa que, para as regras 1-6 e regra 10, a maioria dos erros pode ser identificada por uma abordagem de casamento aproximado com um valor próximo aos 93%.

As células em azul na Tabela 2 representam casos em que poucas publicações geradas pelas regras foram identificadas de maneira completa, proporcionando uma visão clara de onde os algoritmos de depublicação podem se concentrar para obter uma lista de instâncias duplicadas mais abrangente possível.

Observe que, na Tabela 2, a linha correspondente à similaridade de 0.93 evidencia o desempenho mais destacado para as regras 1-6 e 10. Nessas instâncias, a identificação de elementos deduplicados alcançou um patamar ótimo. No entanto, é importante notar que, para as regras 7, 8 e 9, a identificação

[1] Disponível em: https://docs.google.com/spreadsheets/d/1ulz9T3ItCSdFhWzZ2bmD3q17EOQTnvT9/edit?usp=sharing&ouid=101704755271986939546&rtpof=true&sd=true. Acesso em: 21 maio 2024.

de elementos deduplicados ainda representa um desafio substancial. Estas regras, particularmente, introduzem alterações textuais significativas nos títulos das publicações, envolvendo a inserção de Tags HTML, a remoção de títulos e a eliminação de subtítulos, respectivamente.

Tabela 2 - Porcentagem de identificação das publicações, discretizado por regra

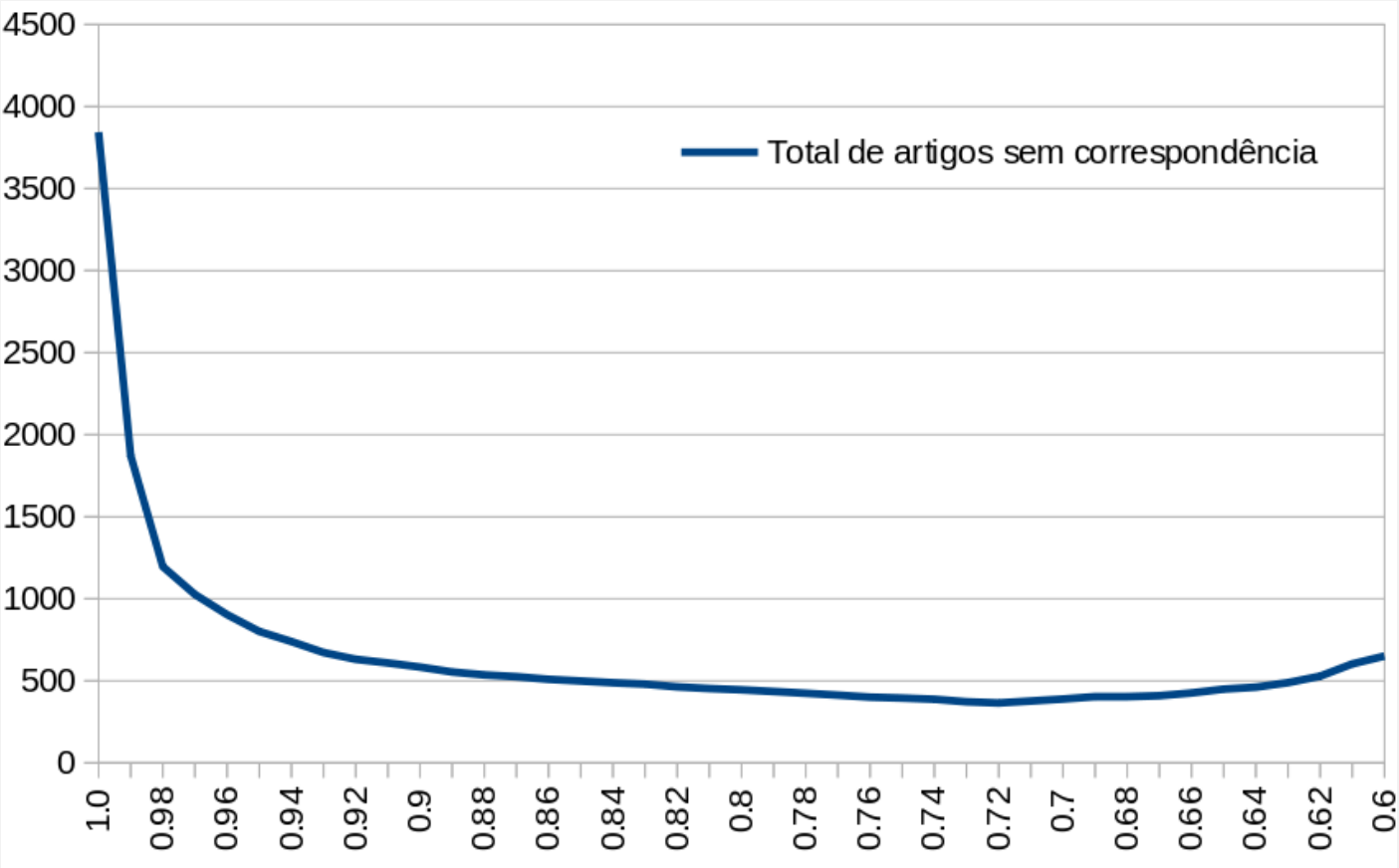
	Regra									
Similaridade	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1.0	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
0.99	8,4%	78,7%	40,0%	52,0%	95,2%	46,7%	0,0%	0,0%	0,0%	95,2%
0.98	55,4%	93,1%	92,5%	95,2%	99,8%	93,2%	3,7%	0,4%	0,0%	99,8%
0.97	80,7%	98,0%	100,0%	99,3%	100,0%	99,0%	16,2%	0,4%	0,0%	100,0%
0.96	90,4%	99,2%	100,0%	99,8%	100,0%	99,8%	34,7%	0,4%	0,4%	100,0%
0.95	96,4%	99,8%	100,0%	100,0%	100,0%	100,0%	51,5%	0,7%	0,4%	100,0%
0.94	98,8%	100,0%	100,0%	100,0%	100,0%	100,0%	61,4%	2,2%	0,7%	100,0%
0.93	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	73,2%	2,6%	1,1%	100,0%
0.92	100,0%	100,0%	100,0%	100,0%	99,6%	100,0%	80,0%	5,6%	1,1%	99,6%
0.91	100,0%	99,6%	100,0%	99,6%	99,6%	99,6%	84,4%	6,7%	1,5%	99,6%
0.9	100,0%	99,6%	100,0%	99,6%	99,6%	99,6%	88,1%	8,2%	1,9%	99,6%
0.89	100,0%	99,6%	100,0%	99,6%	99,6%	99,6%	91,7%	10,8%	3,0%	99,6%
0.88	100,0%	99,6%	100,0%	99,6%	99,6%	99,6%	93,0%	14,6%	3,4%	99,6%
0.87	100,0%	99,6%	100,0%	99,6%	99,6%	99,6%	93,9%	16,0%	4,1%	99,6%
0.86	100,0%	99,6%	100,0%	99,6%	99,6%	99,6%	95,2%	18,7%	4,9%	99,6%
0.85	100,0%	99,6%	100,0%	99,6%	99,6%	99,6%	95,8%	20,5%	5,6%	99,6%
0.84	100,0%	99,6%	100,0%	99,6%	99,6%	99,6%	95,8%	23,5%	6,7%	99,6%
0.83	100,0%	99,2%	100,0%	99,4%	99,3%	99,4%	96,3%	27,6%	7,5%	99,3%
0.82	100,0%	99,2%	100,0%	99,3%	99,3%	99,2%	96,5%	33,6%	8,2%	99,3%
0.81	100,0%	99,2%	100,0%	99,3%	99,3%	99,2%	96,7%	35,8%	9,3%	99,3%
0.8	100,0%	99,2%	100,0%	99,3%	99,3%	99,2%	96,7%	38,1%	10,1%	99,3%
0.79	100,0%	99,2%	100,0%	99,3%	99,3%	99,2%	97,1%	39,6%	11,6%	99,3%
0.78	100,0%	99,2%	100,0%	99,3%	99,3%	99,2%	97,1%	43,3%	11,9%	99,3%
0.77	100,0%	99,2%	100,0%	99,3%	99,3%	99,2%	97,4%	45,5%	13,1%	99,3%
0.76	100,0%	99,2%	100,0%	99,3%	99,3%	99,2%	97,4%	47,4%	16,0%	99,3%
0.75	100,0%	99,2%	100,0%	99,3%	98,9%	99,0%	97,6%	50,0%	16,8%	98,9%
0.74	100,0%	99,0%	100,0%	98,9%	98,9%	98,8%	98,0%	51,1%	19,0%	98,9%
0.73	100,0%	99,0%	100,0%	98,9%	98,9%	98,8%	98,3%	54,9%	20,5%	98,9%
0.72	97,6%	99,0%	100,0%	98,9%	98,9%	98,8%	98,0%	57,1%	22,4%	98,9%
0.71	96,4%	98,2%	95,0%	98,7%	98,2%	98,6%	98,0%	58,2%	23,1%	98,2%
0.7	96,4%	97,2%	95,0%	97,6%	97,4%	97,7%	98,2%	59,7%	25,7%	97,4%
0.69	96,4%	96,4%	95,0%	97,1%	96,5%	97,1%	98,2%	61,2%	26,5%	96,5%
0.68	96,4%	96,4%	95,0%	96,3%	96,5%	96,1%	97,8%	63,8%	28,0%	96,5%
0.67	96,4%	95,5%	95,0%	96,1%	96,0%	95,9%	96,7%	65,3%	31,0%	96,0%
0.66	95,2%	94,7%	95,0%	95,4%	95,0%	95,1%	96,3%	66,8%	32,5%	95,0%
0.65	94,0%	93,7%	95,0%	94,5%	94,1%	94,2%	95,8%	67,5%	34,0%	94,1%
0.64	92,8%	93,1%	92,5%	94,1%	93,8%	94,2%	94,9%	68,7%	34,3%	93,8%
0.63	91,6%	91,5%	90,0%	93,2%	92,3%	93,2%	94,3%	69,4%	37,7%	92,3%

Fonte: Elaborado pelos autores, 2024.

As regras 7, 8 e 9 introduzem variações consideráveis nos títulos originais, tornando a correspondência mais difícil de ser alcançada com base em critérios de similaridade. Este resultado destaca a necessidade de estratégias de deduplicação mais sofisticadas e adaptáveis, capazes de lidar com alterações drásticas nos títulos das publicações impostas por tais regras específicas.

Para explorar a extensão total de publicações não deduplicadas em todo o *benchmark*, foi realizada uma contagem abrangente do número de artigos após a aplicação do algoritmo de Levenshtein com diferentes valores de similaridade. A Figura 1 ilustra uma curva que delineia o comportamento do número de artigos deduplicados em relação a diferentes valores de similaridade. A observação dessa curva revela um declínio acentuado, fornecendo evidências de que, nos algoritmos de deduplicação, não existe um valor único que garanta os melhores resultados; em vez disso, há uma faixa de valores que resultam em desempenhos semelhantes.

Figura 1 - Total de publicações sem correspondência (i.e., não deduplicados).



Fonte: Elaborado pelos autores, 2024.

Destaca-se que a curva reflete a complexidade inerente à deduplicação, indicando que não há uma solução universalmente ideal, mas sim um conjunto de valores que proporcionam resultados comparáveis. Foi notado que, em particular, que a similaridade de 73% resultou no menor número de artigos não deduplicados. Contudo, é importante interpretar esse valor com cautela, dado que 73% representa uma similaridade relativamente baixa e pouco comum em aplicações do mundo real.

Este achado ressalta a importância de adaptar os parâmetros de similaridade conforme o contexto específico da deduplicação, considerando a natureza dos dados e os requisitos da aplicação. Foi observado que, na prática, as porcentagens de similaridade mais frequentemente utilizadas residem na faixa de 85% a 95%, sugerindo que, para muitos casos, esses valores proporcionam um equilíbrio eficaz entre a precisão na deduplicação e a tolerância a variações nos títulos das publicações. Essa evidência é valiosa ao direcionar a configuração de algoritmos de deduplicação para situações do mundo real, onde a eficácia e a praticidade desempenham papeis fundamentais.

5 CONCLUSÕES

Este estudo revela a complexidade inerente à deduplicação de publicações científicas e destaca a importância de estratégias adaptáveis para lidar com desafios específicos, como alterações drásticas nos títulos introduzidas por certas regras. As 10 regras propostas fornecem um guia relevante para a criação de *benchmarks*, contribuindo para avaliações cientométricas mais realistas. A análise do *ben-*

chmark usando o algoritmo de Levenshtein demonstra a necessidade de ajustar os parâmetros de similaridade de acordo com o contexto. Este estudo não apenas permite estudar do ponto de vista prático, a deduplicação de artigos, mas também oferece dados relevantes (e.g., faixa de valores interessantes para a similaridade de Levenshtein) para avanços futuros na deduplicação de dados, em particular, de publicações científicas.

Finalmente, é importante destacar que uma validação mais abrangente e mais exata para os algoritmos de deduplicação, pode ser realizada considerando um conjunto de dados extraída de uma base bibliográfica com informações adicionais como, por exemplo, ano de publicação, nome do veículo, lista de coautores e identificadores DOI.

REFERÊNCIAS

HE, Qinlu; LI, Zhanhuai; ZHANG, Xiao. Data deduplication techniques. *In*: INTERNATIONAL CONFERENCE ON FUTURE INFORMATION TECHNOLOGY AND MANAGEMENT ENGINEERING, 2010, Changzhou, China. **Proceedings [...].** Changzhou, China: IEEE, 2010. p. 430-433. DOI: https://doi.org/10.1109/FITME.2010.5656539. Disponível em: http://ieeexplore.ieee.org/document/5656539. Acesso em: 14 mar. 2024.

JIANG, Yu *et al.* Rule-based deduplication of article records from bibliographic databases. **Database:** The Journal of Biological Databases and Curation, [*S. l.*], v. 2014, article bat086, p. 1-7, 2014. DOI: https://doi.org/10.1093/database/bat086. Disponível em: https://academic.oup.com/database/article/doi/10.1093/database/bat086/2633762. Acesso em: 14 mar. 2024.

KAUFMAN, Aaron. R.; KLEVS, AJA. Adaptive fuzzy string matching: how to merge datasets with only one (messy) identifying field. **Political Analysis,** [*S. l.*], v. 30, n. 4, p. 590-596, Oct. 2022. DOI: https://doi.org/10.1017/pan.2021.38. Disponível em: https://www.cambridge.org/core/journals/political-analysis/article/adaptive-fuzzy-string-matching-how-to-merge-datasets-with-only-one-messy-identifying-field/275D7890548359215AC728C1E35B53CE. Acesso em: 14 mar. 2024.

SAAVEDRA, Rafael H.; SMITH, Alan J. Analysis of benchmark characteristics and benchmark performance prediction. **ACM Transactions on Computer Systems,** [*S. l.*], v. 14, n. 4, p. 344-384, Nov. 1996. DOI: https://doi.org/10.1145/235543.235545. Disponível em: https://dl.acm.org/doi/10.1145/235543.235545. Acesso em: 14 mar. 2024.

UKKONEN, Esko. Algorithms for approximate string matching. **Information and Control,** [*S. l.*], v. 64, n. 1/3, p. 100–118, Jan./Mar. 1985. DOI: https://doi.org/10.1016/S0019-9958(85)80046-2. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0019995885800462. Acesso em: 14 mar. 2024.