Aplicação de Mineração na Elicitação de Citações em Scholar: o Caso do WERpapers

Roxana L. Q. Portugal[1,2][0000-0001-7693-5353], Lyrene Silva[3][0000-0003-1772-6062], Julio Cesar Sampaio do Prado Leite[4][0000-0002-0355-0265], Dennis Farfán Lovón[1], Daniel Mosqueira Obando[1]

[1]Universidad Nacional de San Antonio Abad del Cusco, Perú
[2]LMU, Germany
[3]Universidade Federal de Rio Grande do norte, Brazil
[4]Universidad Federal da Bahia, Brazil
{roxana.quintanilla,101061,154634}@unsaac.edu.pe, lyrene@dimap.ufrn.br, julioleite@ufba.br

Abstract. Este estudo investiga as limitações das métricas reportadas pelo Scopus e do h5 do Google Scholar, que são usadas para qualificar eventos científicos como o WER. Por meio de uma mineração de citações em ambas as bases de dados, o estudo identifica inconsistências e imprecisões nessas métricas e discute possíveis maneiras de diminuir esses problemas. As descobertas deste estudo destacam a importância de uma avaliação precisa e justa de eventos científicos e sugerem a necessidade de aprimorar as métricas utilizadas.

Keywords: Google Scholar, Scopus, H5, digital libraries.

1 Introdução

O artigo tem por objetivo tratar a questão da contabilização de citações da Biblioteca Digital WERpapers[1], um repositório que armazena e descreve os artigos aceitos e revisados apresentados nas edições anuais do Workshop de Engenharia de Requisitos, que há vinte e cinco anos reúne pesquisadores ibero-americanos que investigam sobre o tema.

Com a proliferação das facilidades de referência cruzada fornecida pela digitalização e disponibilização de literatura científica, a contabilização de referências passou a ser uma demonstração de relevância dos meios de publicação, como também dos autores que passaram também a serem avaliados considerando o critério de citações.

Esse assunto já foi discutido em vários meios científicos, apontando a existência de problemas na tecnologia de busca e classificação utilizada por agregadores de literatura [1]. Nesse contexto o surgimento do projeto Scholar da Google trouxe um impacto considerável, na medida que passou a adotar uma visão mais holística das publicações [2] e também com aspectos de precisão de suas buscas ao contrário dos agregadores vinculados a editoras que fundamentalmente utilizavam indexação via banco de dados, mas que depois começaram a utilizar busca em texto, como o Google.

Nesse contexto, uma biblioteca livre e aberta como WERpapers teve a princípio uma vantagem, mas com o passar do tempo a organização de publicações com o sistema DOI[2] e depois com o próprio Scholar sistematizando as publicações de eventos[3], trouxeram novos desafios [5]. Esses desafios fundamentalmente estão relacionados com a infraestrutura necessária e suporte financeiro para manter a biblioteca pública livre e aberta nos níveis de sistematização que hoje são requeridos pelos indexadores/agregadores.

[1] http://wer.inf.puc-rio.br/WERpapers/
[2] Home Page (doi.org)
[3] Google Scholar Support for Publishers

DOI: 10.29327/1298356.26-13

O fato de que o WERpapers tem toda a sua coleção listada no DBLP[4] também foi importante durante uma época, mas não nos parece que o Scholar use essa biblioteca digital na sistematização de seu repositório, haja vista a discrepância muitas vezes notada no retorno do Scholar que indica o artigo para outros repositórios além do WERpapers.

Estes aspectos impactam a elicitação do real número de citações dos artigos do WERpapers. Neste sentido vamos tratar do assunto de maneira geral e reportar sobre uma estratégia de mineração para apoiar a identificação mais próxima da realidade das citações aos artigos publicados no WERpapers, comparando as citações recuperadas no Scopus, um agregador do grupo Elsevier.

Esse artigo está estruturado em cinco Seções. A segunda Seção trata da elicitação de citações nos indexadores Scopus e Google Scholar mostrando a diferença entre ambos ao fazer o cálculo h5 para os resultados reportados. A terceira Seção reporta a nossa análise inicial sobre os dados elicitados. A quarta Seção expõe os trabalhos futuros para motivar a comunidade de engenharia de software sobre a importância da transparência das métricas usadas pelas agências avaliadoras. Concluímos, refletindo sobre as motivações por trás das métricas expostas que podem não ser apropriadas para a comunidade ibero-americana e, portanto, podendo trazer um viés de avaliação para nossa produção científica.

2 Elicitação de Citações

Nesta seção tratamos como cada agregadora de literatura trata a questão de citações, utilizando o Google Scholar e a Scopus.

2.1 Citações Usando Scopus

Scopus disponibiliza o número de citações de artigos indexados pela plataforma. Esse número de citações é definido com base no número de vezes que aquele artigo foi citado em outros artigos também indexados pela Scopus.

O levantamento de citações usando Scopus foi realizado no próprio site da ferramenta. A *string* de busca utilizada para coletar os artigos do WER incluiu os nomes do workshop em português, inglês e espanhol, e excluiu algumas palavras que são usadas em outras conferências de requisitos:

“CONFNAME (("workshop on requirements engineering" OR "workshop em engenharia de requisitos" OR "workshop de ingeniería de requerimientos" OR "workshop de engenharia de requisitos") AND NOT (international OR education OR testing OR law OR voting OR system OR aging OR quality OR visualization OR ieee)) OR SRCTITLE (("workshop on requirements engineering" OR "workshop em engenharia de requisitos" OR "workshop de ingeniería de requerimientos" OR "workshop de engenharia de requisitos") AND NOT (international OR education OR test OR law OR voting OR system OR aging OR quality OR visualization OR ieee))”

Um arquivo CSV com 328 resultados foi baixado. Destes resultados foram removidos aqueles referentes ao editorial, aos sem título, aos sem autor, de outras conferências e duplicados. Assim, 88 foram removidos, restando um total de 240. Esses correspondem aos artigos publicados no WER entre os anos 2005 e 2019.

O WER é indexado na Scopus desde 2005, porém os anos 2020, 2021 e 2022 não foram indexados, possivelmente porque não haviam sido gerados ISBN para os anais do evento nesses anos. Há também divergências entre as quantidades de artigos na base Scopus e a quantidade realmente publicada no WERpapers.

[4] https://dblp.org/

Na tabela 1, listamos os resultados encontrados no Scopus usando o índice h5. O índice h5 do Google Scholar é calculado com base nas citações dos artigos publicados nos últimos cinco anos. O cálculo exige que os artigos sejam ordenados em ordem decrescente com base no número de citações que receberam. Procura-se a posição "h" em que o artigo ocupa na lista, e esse artigo deve ter pelo menos "h" citações[5].

No caso da Tabela 1 o maior número h tal que sua posição na lista tenha pelo menos h citações é 4. No caso a lista seria 5, 5, 4, 4, 4, 3, portanto a quarta posição da lista tem pelo menos 4 citações, ou seja, o h index é 4. Como estamos lidando com 5 anos, o índice é chamado de h5.

É importante lembrar que o cálculo do h5 muda temporalmente e que sua intenção é basicamente ser calculado justamente no intervalo do ano em curso menos 1 e os outros anos anteriores a este (ano em curso - 1) para refletir recência e evitar acúmulo de autocitações.

Tabela 1. h5=4 de WERpapers no horizonte 2015 - 2019 segundo o Scopus. Destacados em cinza os artigos que se repetem na tabela 3.

Título	Autores	Citações	Ano
Trends and needs in requirements engineering research in Ibero-America: Insights from a panel	de la Vara J.L., Brito I.S., Condori-Fernández N., Araújo J.	5	2016
Integrating the E4J editor to the JGOOSE tool	Merlin L.P., de Borba Silva A.L., Santander et al.	5	2015
Privacy and security in requirements engineering: results from a systematic literature mapping	Netto D., Peixoto M., Silva C.	4	2019
A catalogue of istar extensions	Gonçalves E., Heineck T., Araújo J., Castro J.	4	2018
Scalability of iStar: A systematic mapping study	Lima P., Vilela J., Gonçalves E., Pimentel J., Holanda A., Castro J., Alencar F., et al.	4	2016
Requirements engineering practice and problems in agile projects: Results from an international survey	Wagner S., Fernández D.M., Felderer M., Kalinowski M.	3	2017

2.2 Citações usando a API de Google Scholar

O Google Acadêmico disponibiliza sua API desde 2004, porém com algumas limitações como a quantidade de requisições que podem ser feitas mensalmente, atualmente são 100 buscas por mês. No caso do WERpapers temos 25 edições com uma média de 18 artigos por edição. Assim, várias contas do Gmail tiveram que ser usadas para cobrir todos os artigos. O procedimento de mineração foi o seguinte:

1. Os títulos de cada artigo de cada edição do WERpapers foram utilizados como parâmetro de busca. Os autores foram usados para descartar um resultado com várias correspondências.
2. Alguns títulos combinam perfeitamente com os títulos já indexados no Google Scholar, em alguns casos diferem em alguns caracteres.
3. Solicitamos os seguintes parâmetros da API como resposta: título no Google Scholar, autores no Google Scholar, e número de citações.

Foi gerado um arquivo CSV onde foram apagados os títulos dos prólogos que foram

[5] *The h-index of a publication is the largest number h such that at least h articles in that publication were cited at least h times each. For example, a publication with five articles cited by, respectively, 17, 9, 6, 3, and 2, has the h-index of 3.* (https://scholar.google.com/intl/en/scholar/metrics.html#metrics)

tirados automaticamente pelo algoritmo que faz o *Scraping*[6] das edições no WERpapers[7]. 457 artigos foram encontrados, 17 desses não foram indexados pelo Google Scholar. Após fazer um ranking geral pelo número de citações temos na tabela 2 os *top* 10 dos artigos mais citados do WERpapers em 25 anos. A tabela 3 mostra o h5 para o período de cinco anos 2015-2019. A tabela 4 mostra o h5 para o período de cinco anos 2018-2022.

Tabela 2. *Top* 10 do WERpapers em 25 anos

Título	**Autores**	**Citações**	**Ano**
Requirements for Tools for Ambiguity Identification and Measurement in Natural Language Requirements Specifications	Nadzeya Kiyavitskaya, Nicola Zeni, Luisa Mich, Daniel M. Berry.	189	2007
A Requirements Elicitation Approach Based in Templates and Patterns	A. Durán Toro, B. Bernárdez Jiménez, A., et al.	154	1999
Business Process Monitoring and Alignment: An Approach Based on the User Requirements Notation and Business Intelligence Tools	Alireza Pourshahid, Daniel Amyot, Pengfei Chen Michael Weiss, Alan J. Forster.	65	2007
From Early Requirements Modeled by the i* Technique to Later Requirements Modeled in Precise UML	Fernanda Alencar, Jaelson Castro, Gilberto Cysneiros, John Mylopoulos.	55	2000
Study of Elicitation Techniques Adequacy	Dante Carrizo, Oscar Dieste, Natalia Juristo.	54	2008
Formal and Informal Aspects of Requirements Tracing	Francisco A. C. Pinheiro.	51	2000
Evaluating the Effectiveness of Using Catalogues to Elicit Non-Functional Requirements	Luiz Marcio Cysneiros.	51	2007
Requirements Elicitation Using a Combination of Prototypes and Scenarios	Markus Mannio, Uolevi Nikula.	50	2001
Inspección del Léxico Extendido Del Lenguaje	Gladys N. Kaplan, Graciela D.S. Hadad, et al.	45	2000
A Goal-Oriented Approach for Variability in BPMN	Emanuel Santos, Jaelson Castro, et al.	45	2010

Tabela 3. h5=7 de WERpapers no horizonte 2015 - 2019 segundo o Google Scholar[8]

Título	**Autores**	**Citações**	**Ano**
Requirements engineering practice and problems in agile projects: results from an international survey	Stefan Wagner, Daniel Méndez Fernández, et al.	35	2017
Towards an Ontology of Goal-Oriented Requirements.	Pedro Pignaton Negri, Vítor Silva Souza, et al.	34	2017
Retrospective and Trends in Requirements Engineering for Embedded Systems: A Systematic Literature Review.	Tarcísio Pereira, Deivson Albuquerque, et al.	10	2017
Scalability of istar: a Systematic Mapping Study	Paulo Lima , Jéssyka Vilela, et al.	8	2016
Requirements Smells como indicadores de má qualidade na especificação de requisitos: Um Mapeamento Sistemático da Literatura	Rafael Nascimento, Eduardo Aranha, Uirá Kulesza, e Marcia Lucena.	8	2018
Integrating the E4J editor to the JGOOSE tool	Leonardo Pereira Merlin, et al.	8	2015
Elicitação e Especificação de Requisitos em Sistemas Embarcados: Uma Revisão Sistemática	Aêda Souza, Josenildo Melo, et al.	7	2015

[6] The process of extracting data from a digital source for automated replication, formatting, or manipulation by a computer program, as in data mining or website data analysis.

[7] http://wer.inf.puc-rio.br/WERpapers/index.lp

[8] No sítio https://jcspl.net/2021/11/20/wer-h5-2016-2020/ são listados outros intervalos do h5 do WER, calculados manualmente.

Tabela 4. h5=5 de acordo com as citações do Google Scholar (2018 - 2022)

Título	Autores	Citações	Ano
Requirements Smells como indicadores de má qualidade na especificação de requisitos: Um Mapeamento Sistemático da Literatura	Rafael Nascimento, Eduardo Aranha, et al.	8	2018
Uma Ferramenta para Construção de Catálogos de Padrões de Requisitos com Comportamento	Taciana N. Kudo, Renato F. Bulcão-Neto, Auri M. R. Vincenzi.	7	2020
Privacy and Security in Requirements Engineering: Results from a Systematic Literature Mapping	Dorgival Netto, Mariana Peixoto, Carla Silva.	6	2019
NFR4ES: Um Catálogo de Requisitos Não-Funcionais para Sistemas Embarcados	Reinaldo Silva, Jaelson Castro, João Pimentel.	6	2020
Organizing the TD Management Landscape for Requirements and Requirements Documentation Debt	Larissa Barbosa, Sávio Freire, Nicolli Rios, et al.	6	2022
RASO: an Ontology on Requirements for the Development of Adaptive Systems	Cássio Capucho Peçanha, Bruno Borlini Duarte, Vítor E. Silva Souza.	5	2018
Experimentando o SPIDe aplicado à Elicitação de Requisitos	Jean C. S. Rosa, Ecivaldo Matos, Fiama S. Santos, Gilton J. F. Silva.	5	2018

3 Análise de Resultados

Três tipos de análise podem ser feitas: a) sobre os números dos artigos que aparecem nas Tabelas de cada agregadora, b) diferenças (ausências) de títulos entre as Tabelas dos agregadores, c) impacto da língua em que está escrito o artigo.

Conforme observado nota-se, como esperado, uma diferença entre os resultados do Scholar e os do Scopus. Vale lembrar que isso era esperado, mas é importante notar a real diferença, ou seja na comparação entre artigos presentes em cada um dos agregadores. Comparando as Tabelas 1 e 3, que deveriam apresentar os mesmos artigos, temos que coincidem apenas em 2 artigos realçados em cinza nas Tabelas (tipo b). Ambos os artigos têm menos citações no Scopus (tipo a).

Porém, o que mais se destaca dessas duas Tabelas é o primeiro artigo da tabela 3, o trabalho de Wagner et al. “Requirements engineering practice and problems in agile projects: results from an international survey” possui 35 citações segundo o Scholar. Aqui o caso é perfeito para exemplificar como os indexadores podem ter as contas erradas (tipo a). O artigo de Wagner et al. foi subido no conhecido repositório arXiv.org, o qual não deveria ser problema, mas, no caso deste artigo, ganhou posicionamento por ser o mais acessado. Se um entrar no link das outras 14 versões, se encontra o ponteiro ao repositório WERpapers. Ainda esse não é um problema. O problema é o número de citações que foram consideradas pelo Scopus, seja de apenas 3. Se o Scopus considera somente citações de artigos que foram também indexados pelo Scopus, então temos 32 artigos que não são relevantes para Scopus e teríamos que fazer essa análise de quais sim ou não foram considerados.

Aqui são relevantes os trabalhos [6] [7], onde se discute se na Ibero-América devemos utilizar os indicadores Scopus como métrica de avaliação, já que este indexador deixa de fora eventos importantes em nossa região, mas que não se enquadram em seus critérios de indexação, onde prevalecem as publicações em Inglês (tipo c).

Como relata [3] existe um viés claro do agregador Scopus que deixa de computar áreas importantes da produção científica nacional, principalmente Agricultura e Saúde. Em [4] há o relato de que autores já identificaram o viés das bases Scopus e Web of Science que privilegiam as áreas de Ciências Naturais, Engenharia e Pesquisa Biomédica, nesse mesmo artigo [4] há o relato de que outros autores estudaram os erros desses agregadores e que a distribuição desses erros era diferente entre esses agregadores. Vale notar que as referências sobre uso do índice h ou h-5

encontradas tratam fundamentalmente de periódicos, ao passo que estamos tratando de uma conferência.

Estas análises nos ajudam a entender melhor as diferenças como também sugerir políticas para serem seguidas por eventos acadêmicos que disponibilizam de maneira livre seus conteúdos. A vantagem de ser livre e sem custo, no caso do WER, pode deixar de ser fator importante para os autores, na medida em que os índices calculados pelos agregadores desviem da realidade.

Sobre políticas a serem consideradas a principal é procurar fazer com que pelo menos o índice de mais fácil acesso e o com um pouco mais de transparência, o índice h5 do Google Scholar, seja entendido pelos humanos que serão alvo desta avaliação. Cabe aos avaliadores darem mais transparência ao significado do índice e cabe aos humanos submetidos ao índice que procurem saber se os índices estão realmente refletindo a realidade e como mitigar os problemas.

Este artigo pretende atuar em trazer mais transparência ao h5 aplicado ao WERpapers e conscientizar a comunidade para o debate de como valorizar o WER. Uma política direta é a consciência de que a leitura de artigos do WER é valorosa para a pesquisa e que, sendo assim, deve ser referida em seu trabalho. Veja, por exemplo, o trabalho de Valaski et al. [8], que ainda em 2013, fez uma avaliação da área usando o WER como base. Outra política é a de deixar claro para as agências de avaliação que o índice h5 tem problemas, não só como favorecendo áreas de maior procura nos últimos cinco anos (áreas *pop*), como também a limitação das máquinas de busca na correta identificação de artigos para eventos que não tenham sido reconhecidos pelo agregador Google Scholar ou pelo Scopus.

4 Trabalhos Futuros

Entendemos que esse é um artigo inicial que ajudará a comunidade do WERpapers entender melhor os instrumentos utilizados pelos agregadores para calcular métricas de avaliação, com especial ênfase em citações.

Para o trabalho futuro, não só outros agregadores não associados a editoras podem ser utilizados, como o Semantic Scholar (https://www.semanticscholar.org/), por exemplo, mas também os resultados aqui reportados, de maneira transparente, podem ser replicados.

Outra sugestão de trabalho futuro é através de consulta à comunidade mais ampla, por exemplo Engenharia de Software ou Sistemas de Informação, sobre o nível de consciência sobre a questão de como esses agregadores produzem essas métricas e como cada pesquisador entende como isso afeta sua carreira e o da sobrevivência de sua área de pesquisa.

Para este trabalho, temos dois arquivos CSV com dados que podem ser analisados com maior profundidade. Para tal, disponibilizamos estes itens no link[9][9], como também, disponibilizamos o código utilizado para consumir os dados da API do Google Scholar.

5 Conclusão

A discussão pela comunidade de requisitos sobre como a avaliação de agências de fomento ou de seus pares é efetuada é, sem dúvida, um interesse de todos. Neste artigo procuramos aclarar um aspecto da avaliação, citações, como feito por dois grandes agregadores. Um ligado a um grupo editorial e outro sem ligação com grupos editoriais e com a perspectiva de que se trata de dados gratuitos. Esperamos que a discussão do artigo no WER 2023 possa ajudar no sentido de lidar com este assunto.

Analisamos as diferenças e listamos algumas políticas que podem ser adotadas para evitar uma avaliação parcial de eventos livres, ou seja, sem divulgação por grandes

[9] https://github.com/daniel2019-max/PaperScraper

editoras. Em particular, tratamos do caso do WERpapers para conscientizar a comunidade no sentido de valorizar um fórum essencial para a discussão das oportunidades de pesquisa na área.

Observamos através da literatura que existe uma preocupação crescente em utilizar agregadores que tem viés tanto por área como por regiões geográficas e culturais. Canto [5] relata: "Por outro lado, a maioria das publicações indexadas nessas fontes são de origem anglo-saxônica, de idioma inglês e com aderência temática às áreas de ciências exatas e naturais, engenharias e medicina (WALTMAN, 2016)". Essa citação à Waltman[10] reflete algo que se observa no item c) da Seção 4.

Outra constatação do trabalho de Canto [5] é que "A escolha da região Ibero-Americana foi motivada pela menor visibilidade em bases de dados internacionais. Os sistemas de avaliação do Brasil, Espanha, México e Colômbia definiram o índice h5 como indicador de impacto para periódicos não indexados no Journal Citation Reports e na Scopus." Esse fato aponta que apesar das vantagens do h5 da GSM (Google Scholar Metrics), indíce que privilegia a recência, existem também dúvidas quanto a sua qualidade, como relata o próprio Canto: "Mas o GSM também possui disfunções, as quais colocam em dúvida a sua adequação com fonte oficial de avaliação científica." [5].

Por outro lado, temos que considerar que o próprio GSM diz: "Overall, Scholar Metrics cover a substantial fraction of scholarly articles published in the last five years. However, they don't currently cover a large number of articles from smaller publications.". Além disso, diz o GSM: "If you can't find the journal you're looking for, try searching by its abbreviated title or alternate title. There're sometimes several ways to refer to the same publication. (Fun fact: we've seen 959 ways to refer to PNAS.)". Sobre essa última observação vale lembrar que uma das dificuldades de indexação de artigos do WERpapers pelo GSM é que o WER é referido de diversas maneiras: WER, WERpapers, Workshop de Engenharia de Requisitos, Workshop en Ingeniería de Requerimientos, Workshop en Ingeniería de Requerimientos e variações.

Referencias

1. Caregnato, Sonia Elisa. "Google Acadêmico como ferramenta para os estudos de citações: avaliação da precisão das buscas por autor." Pontodeacesso 5.3 (2011): 72-86.
2. Silva, Deise Deolindo, and Maria Cláudia Cabrini Grácio. "Índice h de Hirsch: análise comparativa entre as bases de dados Scopus, Web of Science e Google Acadêmico." Em questão 23.5 (2017): 196-212.
3. Mugnaini, Rogério, et al. "Panorama da produção científica do Brasil além da indexação: uma análise exploratória da comunicação em periódicos." Transinformação 31 (2019).
4. Huang, Chun-Kai, et al. "Comparison of bibliographic data sources: Implications for the robustness of university rankings." Quantitative Science Studies 1.2 (2020): 445-478.
5. Canto, Fábio Lorensi do. "Avaliação de impacto de periódicos ibero-americanos com base no índice h5 do Google Scholar Metrics.", Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa Pós-Graduação em Ciência da Informação, Florianópolis, (2022).
6. Costa, Heloisa, Fabio Lorensi do CANTO, and Adilson Luiz Pinto. "Google Scholar Metrics e a proposta do novo Qualis: impacto dos periódicos brasileiros de Ciência da Informação." Informação & Sociedade: Estudos; v. 30 n. 1 (2020) 24.2 (2020).
7. Rozemblum, Cecilia, Juan Pablo Alperin, and Carolina Unzurrunzaga. "Las limitaciones de Scopus como fuente de indicadores: Buscando una visibilidad integral para revistas argentinas en ciencias sociales." E-Ciencias de la Información 11.2 (2021): 35-58.
8. Valaski, J., Stancke, W., Reinehr, S. S., & Malucelli, A. Retrospective and Trends in Requirement Engineering through WER. In WERpapers conf/wer/ValaskiSRM13 (2013)
9. Daniel, dennis721, & dfarfanl. (2023). nitanilla/PaperScraper: Eliciting citations from WERpapers and GoogleScholar (Version releaseWER23). Zenodo. https://doi.org/10.5281/zenodo.8173190

[10] WALTMAN, L. A review of the literature on citation impact indicators. Journal of Informetrics, v. 10, n. 2, p. 365–391, maio 2016. Disponível em: http://dx.doi.org/10.1016/j.joi.2016.02.007