MÉTODOS DE SELEÇÃO DE AUTORES PARA ESTUDOS DE COCITAÇÃO: como definir um ponto de corte

Methods for selecting authors for co-citation studies: how to define a cutoff point

Rodrigo A. de Carvalho (1), Francieli A. L. Muck (2), Sabrina S. Corrêa (3), Catarina P. de Carvalho (4), Sônia Elisa Caregnato (5)

(1) Universidade Federal do Rio Grande do Sul, Brasil, racfurg@gmail.com. (2) Brasil, francieli.muck@hotmail.com. (3) Brasil, sabrinascsc@gmail.com. (4) Universidade Federal de Pelotas, Brasil, catarinaprestes@yahoo.com.br. (5) Brasil, sonia.caregnato@ufrgs.br.

Resumo

Trata da análise de cocitação de autores – ACA, mais especificamente apresenta uma proposta para definir a seleção de autores. Objetivos: a) comparar dados de citação total, citação por documento citante e cocitação de autor com ele mesmo, para definir o indicador mais apropriado para a seleção; e b) propor uma forma para estabelecer um ponto de corte para a criação das matrizes. Dois conjuntos de dados, recuperados na *Web of Science*, foram utilizados: conjunto A, composto de 17.992 referências de 421 artigos, Conjunto B. composto de 5.771 referências de 151 artigos. A citação por documento citante é mais apropriada para classificar os autores do que as citações totais e a cocitação do autor com ele mesmo. Incluir autores cujas citações por documentos citantes somadas atingem aproximadamente 20% mostrou-se apropriado, mas autores citados uma única vez não devem fazer parte dos cálculos. Nos dados do Conjunto A, foram definidos 180 autores (20,6% da soma das citações) e 72 autores (20,1% da soma das citações), considerando as todas as posições de autoria das referências e as primeiras posições, respectivamente. Conclui-se pela validade da proposta por contextualizar os dados analisados diante do todo e considerar a melhor distribuição dos autores citados no *corpus*.

Keywords: Análise de cocitação de autores - ACA; Estudos de citação; Cientometria.

Abstract

The paper focuses on author co-citation analysis - ACA, particularly presenting a proposal to define the selection of authors. Objectives: a) to compare total citation, citation per citing document and author's co-citation with himself, to define the most appropriate indicator for inclusion of authors; and b) to propose a

Carvalho, R. A. de, et al. "Métodos de seleção de autores para estudos de cocitação: como definer um ponto de corte". *Brazilian Journal of Information Science: Research trends*, vol.15, publicação continuada, 2021, e02109. doi10.36311/1981.1640.2001.v15.e02109

way to establish a cut-off point for creating of the matrices. Two data sets, retrieved from the Web of Science, were used: set A, composed of 17,992 references from 421 articles; Set B, composed of 5,771 references from 151 articles. The citation per citing document is more appropriate to classify authors than the total number of citations and author's co-citation with himself. To include authors whose citations by citing documents reach approximately 20% proved to be appropriate, but authors cited only once, which constitute absolute dispersion data, should not be included in the calculations. Regarding the data in Set A, 180 authors (20.6% of the sum of citations) and 72 authors (20.1% of the sum of citations) were defined, considering all the positions of the references and the first positions, respectively. It concludes by the validity of the proposal for contextualizing the data analyzed and considering the best distribution of cited authors mentioned in the corpus.

Keywords2: Author co-citation analysis – ACA; Citation studies; Scientometrics.

1 Introdução

Estudos de cocitação são geralmente utilizados como técnica de análise de domínio (Hjørland 2002; Grácio e Oliveira 2013b) ou para visualização e mapeamento de uma área específica do conhecimento produzido e publicado (Eom 2009; Schneider *et al*., 2009) em teses e dissertações, artigos de periódicos, comunicações em eventos etc. Apresentam, portanto, um olhar para o passado, pois são estudos realizados a partir das referências, ou seja, com base em um *corpus* de documentos citantes.

Uma série de questões se apresenta para a realização de um estudo tradicional de cocitação. Cada escolha, desde a unidade de contagem (documento, autor e periódico), que define o tipo de análise de cocitação, até a técnica de estatística multivariada utilizada e a forma empregada para rotular os agrupamentos, implica em perda, ambiguidade, omissão e/ou ruído informacional, que interfere diretamente na análise do contexto estudado. Ou seja, emerge uma relação entre os dados que serão utilizados e os que não serão utilizados, ou que deixam de ser utilizados a partir de cada etapa, uma vez que as técnicas buscam apresentar as ligações mais frequentes entre autores, documentos ou periódicos (generalizações).

Uma discussão sobre as informações que um estudo dessa natureza gera como resultados, analisados ou não, é necessária. Assim, este trabalho coloca como questão qual o melhor indicador para a seleção de autores que serão analisados nesse tipo de estudo, ou seja, como definir o ponto de corte e o que é desconsiderado após essa escolha. Os objetivos da pesquisa são: a) comparar dados de citação total, citação por documento citante e cocitação de autor com ele mesmo para definir o indicador mais apropriado para a seleção de autores em ACA; e b)

propor uma forma para estabelecer um ponto de corte para a criação das matrizes de cocitação de autores.

Desde a sua concepção, os estudos de cocitação de autores (White, e Griffith 1981) são usados não apenas para o mapeamento de disciplinas científicas, mas também para discussões metodológicas que criticam a técnica e propõem variações, novas formas de interpretar e agregar informações aos agrupamentos, desenvolver e avaliar as matrizes, etc. O presente estudo tem a intenção de contribuir metodologicamente com esse domínio dos estudos de citação (EC).

2 Aspectos teóricos e trabalhos relacionados

Pode-se definir a cocitação como a contagem das citações de dois documentos ou autores em um mesmo texto, ou seja, conta-se o par de citações como uma unidade, e essa contagem determina uma potencial relação bibliográfica entre essas entidades (documentos ou autores) (Spinak 1996; Reitz c2014), a partir do contexto do documento citante. A relevância da relação bibliográfica é analisada a partir de sua ocorrência em um grande conjunto de documentos. As análises de cocitação se concentram, portanto, em identificar, medir e interpretar um grupo de entidades cocitadas em um contexto específico, como já destacado. Trata-se de um indicador de relacionamento derivado da citação, mas efetivado por estudos realizados a partir das referências. Small (1973) publica as bases conceituais desse indicador de relacionamento bibliográfico, diferenciando-o de um antecessor, o acoplamento bibliográfico (Kessler 1963).

Os tipos de estudos de cocitação se desenvolvem a partir de duas perspectivas. A primeira é a unidade de contagem dos pares, que advém dos elementos que formam uma referência bibliográfica ou outro metadado identificável a partir desses elementos, como os autores (White e Griffith 1981; McCain 1990), os títulos, que representam os documentos (e que também podem representar a relação autor/data) (Small 1973) e os periódicos (McCain 1991). A segunda perspectiva se refere ao que os autores Liu, e Chen (2012) chamam de proximidade da ocorrência da cocitação, que é dividida em quatro níveis, da menor para a maior distância, a saber: na sentença ou frase; no parágrafo; na seção do documento; e no artigo/documento. Os autores definem que a cocitação por proximidade em artigo ocorre quando duas referências são citadas em seções distintas de um mesmo documento e não necessariamente na lista de

referências. Essa distinção é apresentada porque os autores retiram as ocorrências de cocitação do texto completo do *corpus* analisado, sem considerar a lista de referências. Assim, é importante salientar que o presente estudo é específico sobre a análise de cocitação de autores (ACA), com dados utilizados a partir da lista de referências dos artigos, que formam o *corpus* de análise, ou seja, trata-se de um estudo tradicional.

A ACA consiste em seis etapas para uma operacionalização completa: a) seleção dos autores; b) recuperação das frequências de cocitação; c) compilação da matriz simétrica com os valores absolutos; d) conversão da matriz anterior em uma matriz de correlação ou normalizada; e) análise multivariada da matriz normalizada; e f) interpretação e validação dos agrupamentos identificados (McCain 1990). Esse trabalho se concentra nas duas primeiras etapas de operacionalização das técnicas, que define o número de entidades que entrarão nas análises e de que forma essa entrada será determinada, principalmente em relação aos dados de citação.

Os trabalhos relacionados se concentram especialmente na ACA. Parte desses estudos opta pela seleção dos autores pelo número total de citações recebidas, sem uma explicação sobre o número de autores que farão parte da matriz final com dados absolutos em relação a todos os autores que formam os dados de análise. Bu, *et al*. (2016) utilizam os 100 autores mais citados e Bu, *et al*. (2017) utilizam os 500 autores mais citados para criarem as matrizes. White, e McCain (1998) e Zhao, e Strotmann (2014) analisam os 120 e os 250 autores mais citados na área da Ciência da Informação, respectivamente, a partir das referências de 12 periódicos.

McCain (1990) cita alguns critérios para essa situação, identificados a partir das práticas de alguns pesquisadores de ACA[1], como utilizar taxas médias de cocitação acima de nove em um período específico de tempo e cocitação com pelo menos um terço dos autores que fazem parte da análise. Penan[2] (1989 *apud* McCain 1990 p. 435) estabeleceu, a partir de alguns testes de recuperação de dados, que deveriam ser computados 20% dos autores com mais citações e cocitações. McCain (1990) fez testes com cortes menores e indicou que os resultados foram satisfatórios em termos de interpretação dos agrupamentos gerados.

[1] Trecho: “ACA researchers, concerned with the possible instability of small cocitation counts, have selected for further analysis only those authors meeting certain ad hoc criteria, such as mean cocitation rates above nine (for 10 years of Social Scisearch

[2] Penan, H. “Pour une gestion bibliometrique de l' information scientifique et technique des entreprises. Application en theorie microeconomique et financiere”. 1989. Universite des Sciences Sociales, Doctoral dissertation [Texto completo não localizado].

Diversos aspectos surgem para ponderar as propostas dos autores citados acima. Um corte muito alto, por exemplo, pode acarretar na não identificação de uma área emergente, com dados de autores que estejam um pouco abaixo do ponto de corte estabelecido. Da mesma forma, um corte muito baixo pode apresentar relações não efetivamente temáticas ou consolidadas entre os cocitados e atrapalhar a interpretação adequada de um determinado agrupamento. Além disso, a diferença nas datas de publicação dos textos impacta em termos tecnológicos a capacidade de recuperação e manuseio dos dados. Assim, justifica-se a relevância dessa discussão.

Quanto ao indicador utilizado, os trabalhos metodológicos de Grácio, e Oliveira (2013a, 2013b, 2015) optam pelo uso do número de documentos que citam os autores para a formação das matrizes e não o número total de citações. O trabalho de Carvalho, *et al*. (2018) traz dados que evidenciam que o número de documentos citantes de um autor é mais apropriado para ACA do que a contagem total de citações, pois os principais autores são mencionados várias vezes em documentos citantes. Isso ocorre, por exemplo, com Birger Hjørland, que possui 22 citações em apenas oito documentos citantes nos dados apresentados pelos autores. Dois pequenos *rankings* com 30 posições, organizados por citações totais e número de documentos citantes, mostram que 23 autores fazem parte das duas listas e 14 autores distintos diferenciam os *rankings* (Carvalho, *et al*. 2018). Certamente essa condição interfere na melhor identificação das relações de cocitação.

Eom, e Farris (1996) utilizam a cocitação do autor com ele mesmo, o que facilita a definição dos dados inseridos na diagonal da matriz, mas eles alertam que utilizar essa informação pode implicar em limitações, principalmente em áreas multidisciplinares. Os autores aplicaram uma ACA no campo de pesquisa "*Decision Support Systems*" (DSS) e a escolha dos autores para a análise se deu em dois processos, pois os dados foram retirados da *Web of Science* (WoS), que indexa apenas o primeiro autor das referências, e houve um trabalho adicional para a identificação das coautorias. Assim, os 67 autores com pelo menos 25 cocitações consigo próprios entraram na análise.

Empiricamente é possível um autor ser citado por todos os documentos citantes de um *corpus* de análise e não ter nenhuma cocitação com ele mesmo, sem contar que estudos que utilizam a primeira posição das referências limitam ainda mais o uso dessa informação, como será apresentado nos resultados desta pesquisa.

3 Procedimentos metodológicos

A pesquisa se caracteriza como descritiva, de caráter metodológico e cientométrico. Dois conjuntos de dados, recuperados na base de dados multidisciplinar *Web of Science* (WoS), foram utilizados para cumprir os objetivos propostos: conjunto a) referências de 421 artigos, coletados em fevereiro de 2018, publicados entre 2015 e 2016, no campo da Ciência da Informação, subcampos Organização do Conhecimento e Recuperação da Informação[3]; conjunto b) referências de 151 artigos recuperados em maio de 2017, publicados entre 2011 e 2015, no campo da Ciência da Informação, subcampo Organização do Conhecimento[4].

As temáticas foram escolhidas pela familiaridade dos autores desta pesquisa. Os recortes temporais, dois e cinco anos, respectivamente, foram determinados para permitir inferências metodológicas razoáveis em conjuntos de dados robustos, mas que pudessem ser tratados manualmente, uma vez que a WoS não oferecia o detalhamento necessário para o problema proposto. Não houve limitação no número de artigos, ou seja, os dados refletem as temáticas nos recortes temporais indicados, mas artigos duplicados ou sem referências foram desconsiderados.

Os dados das referências foram retirados dos arquivos digitais do texto completo de cada artigo que formam os dois conjuntos de dados e não dos metadados da WoS, devido as referências com autorias múltiplas. A estruturação e a limpeza dos dados foram realizadas manualmente. Importante salientar que os dados das referências dos 151 artigos, coletados em 2017, foram utilizados em trabalhos anteriores (Carvalho, e Caregnato 2017; Carvalho, *et al*. 2019) com outros propósitos e serviram para os testes necessários para o desenvolvimento da proposta do presente trabalho.

Os 421 artigos do Conjunto A foram publicados por 69 periódicos distintos, sendo que os títulos que se destacam são o *Knowledge Organization* (17,10%) e o *Journal of The Association for Information Science and Technology* (14,25%), que publicaram juntos 132 artigos. Alguns periódicos brasileiros se destacam: TransInformação (cinco artigos), Informação & Sociedade

[3] A estratégia de busca consistiu no uso de quatro termos compostos utilizados no campo 'TÓPICO', como segue: "TOPIC: ("knowledge organization" OR "information organization" OR "knowledge representation" OR "information retrieval"). Os dados recuperados foram refinados da seguinte forma: "WEB OF SCIENCE CATEGORIES: (INFORMATION SCIENCE LIBRARY SCIENCE) AND DOCUMENT TYPES: (ARTICLE); Timespan: 2015-2016. Indexes: SCI-EXPANDED, SSCI, A&HCI, ESCI".

[4] A busca foi realizada pelo termo "Knowledge organization" no campo "TÓPICO" da base e refinado pela categoria "INFORMATION SCIENCE LIBRARY SCIENCE", foram considerados apenas artigos.

(quatro artigos), Perspectivas em Ciência da Informação (quatro artigos) e Revista Ibero-Americana de Ciência da Informação (quatro artigos).

Os dados extraídos das referências dos 421 artigos serviram ao desenvolvimento de três conjuntos de dados que interessam a este trabalho, a saber: dados de caracterização das referências; *ranking* de autores incluindo todas as posições das referências; e *ranking* de autores considerando a primeira posição e autoria única.

A Tabela 1 apresenta as estatísticas descritivas dos principais indicadores que caracterizam as referências retiradas dos 421 artigos. Nota-se, na última linha da tabela, o número de autores distintos citados uma única vez no *corpus* analisado, sendo essa informação relevante para entender a dispersão dos dados para se estabelecer uma forma de ponto de corte na seleção dos autores.

A partir das referências do Conjunto A (Tabela 1) foram elaborados dois *rankings* de autores, um que inclui todas as autorias das referências e outro apenas com autores na primeira posição das referências, incluindo autoria única. O *ranking* que inclui todas as posições de autoria ficou com 19.590 autores distintos (a Tabela 6 apresenta um recorte desse *ranking*), a partir da limpeza realizada, e o *ranking* que inclui a primeira posição ficou com 9.338 posições ou autores distintos.

Tabela 1 – Estatísticas descritivas dos indicadores das referências do Conjunto A (N = 421 artigos)

Indicadores	Total	Medidas de centralidade			Medidas de dispersão		
		Média	Mediana	Moda	Desvio Padrão	Mínimo	Máximo
Nº de referências	17992	42,74	38	38	27,37	02	236
Nº de referências desconsideradas	18	0,04	00	00	0,25	00	03
Nº de referências de autoria NÃO pessoal	862	2,05	00	00	3,95	00	34
Nº de referências de autoria pessoal	17114	40,65	36	46	26,89	02	221
Nº de referências com um autor pessoal	6822	16,20	13	14	15,18	00	164
Nº de referências com dois ou mais autores pessoais	10292	24,45	22	31	20,56	00	187
Nº de autorias pessoais (inclui repetições)	39060	92,78	81	33	70,30	02	621
Nº de coautorias NÃO pessoais	07	0,02	--	--	0,13	00	01
Nº de autores com 01 citação	13934	33,10	24	08	32,62	00	364

Fonte: dados da pesquisa, 2020.

Os indicadores relacionados para cada autor no *ranking*, com dados do Conjunto A, que inclui todas as posições e que servem para esse trabalho são os seguintes: i) número total de citações no *corpus*; ii) número de artigos do *corpus* que citam o autor (documento citante); e iii) número de cocitação do autor com ele mesmo (quando o autor tem dois ou mais documentos citados no mesmo documento citante). Um *ranking* com 107 autores é apresentado na Tabela 6, na seção dos resultados. Os indicadores relacionados para cada autor no *ranking* que inclui a primeira posição e autoria única do Conjunto A de referências são: i) número total de citações no *corpus*; e ii) número de artigos do *corpus* que citam o autor (documento citante).

A Tabela 2 apresenta as estatísticas descritivas dos principais indicadores que caracterizam as referências retiradas dos 151 artigos que formam o Conjunto B. Esses dados sustentaram alguns testes quanto aos objetivos desse trabalho e foram repetidos no conjunto maior de dados.

Tabela 2 – Estatísticas descritivas dos indicadores das referências do Conjunto B (N = 151 artigos)

Indicadores	**Total**	**Medidas de centralidade**			**Medidas de dispersão**		
		Média	**Moda**	**Mediana**	**Desvio padrão**	Mínimo	Máximo
Nº de referências	5771	38,22	28	34	22,50	05	120
Nº de referências desconsideradas	04	0,03	00	00	0,26	00	03
Nº de referências de autoria NÃO pessoal	318	2,11	00	01	5,29	00	55
Nº de referências de autoria pessoal	5449	36,06	21	32	21,81	04	111
Nº de referências com um autor pessoal	3469	23,08	07	20	17,09	02	88
Nº de referências com dois ou mais autores pessoais	1980	13,11	02	09	11,72	00	59
Nº de autorias pessoais (incluindo todas as posições)	9311	61,66	33	49	43,00	06	213

Fonte: dados da pesquisa, 2020.

A partir das referências do Conjunto B (Tabela 2) foram elaborados dois *rankings* de autores, como no conjunto apresentado anteriormente. O primeiro *ranking* que inclui todas as autorias das referências ficou com 5.333 posições e os indicadores foram: (i) número total de citação; e (ii) citações contadas pelo número de documentos citantes. A Tabela 3, apresentada na seção dos resultados, exibe 153 posições desse *ranking*. O segundo *ranking*, que considera apenas as primeiras posições das referências, ficou com 2.938 autores distintos e os indicadores foram os mesmos do *ranking* anterior.

4 Resultados e discussão

Para atingir o primeiro objetivo deste trabalho é necessário entender a relação entre os indicadores: citações; documentos citantes (número de documentos que citam um autor); e cocitação do autor com ele mesmo. Qualquer autor pode ser referenciado/citado mais de uma vez por um documento citante, desde que sua produção bibliográfica permita essa condição. Assim, contar o número de documentos que citam um autor permite entender melhor a distribuição de suas citações e o seu real alcance em relação a um *corpus* de análise. Mensurar se um autor é citado mais de uma vez pelo mesmo documento citante permite a contagem da "cocitação" do autor com ele mesmo. Esses dois indicadores, mais o número de citações totais, são utilizados na

literatura sobre cocitação para ordenar os autores (McCain 1990; White, e McCain 1998; Grácio, e Oliveira 2013b; Bu, *et al*. 2016; BU *et al*. 2017; Carvalho, e Caregnato 2017).

A Tabela 3 apresenta um *ranking* com 153 autores, elaborado com dados de referências do Conjunto B, que relaciona citações totais e documentos citantes (duas primeiras colunas de valores depois dos nomes dos autores). Observa-se que poucos possuem diferença (terceira coluna depois dos nomes dos autores) entre número de citações totais e número de citantes igual a zero (apenas 11 – onde a citação total é igual à distribuição dos autores pelos documentos citantes). O autor "Hjørland, B" (destacado com bordas vermelhas), que é o primeiro do *ranking*, tem 228 citações distribuídas por 70 citantes dos 151 artigos que formam o *corpus*.

Os autores em destaque na tabela (**negrito**, *itálico* e <u>sublinhado</u>) evidenciam que há uma discrepância entre as citações totais e o número de documentos citantes. Isso pode interferir na inclusão desses autores em uma matriz de cocitação, uma vez que a distribuição dos autores pelos documentos citantes pode oferecer mais possibilidades de ocorrências de cocitação com os outros autores, e isso é uma preocupação nos estudos, inclusive como critério de seleção dos autores (McCain 1990). Ou seja, um autor muito citado e pouco cocitado tem pouca relevância para um estudo tradicional de cocitação.

O autor "Ostwald, W" (destacado com bordas verdes) possui 26 citações e seria aproveitado em uma análise de cocitação que considerasse o número de citações totais como indicador de inclusão, mas esse dado é oriundo de três documentos citantes e isso interfere nas possíveis relações de cocitação. O autor "Fahlman, SE" (destacado com bordas azuis) é um caso muito significativo, pois possui nove citações oriundas de apenas um documento e é outro exemplo de que a distribuição dos autores pelo *corpus* de análise é mais significativa para estudos de cocitação do que o número total de citações.

Tabela 3 – *Ranking* de autores comparando citações totais com documentos citantes do *corpus* do Conjunto B (organizado por citações totais e com 153 posições)

AUTORES	CITAÇÕES	CITANTES	DIFERENÇA	AUTORES	CITAÇÕES	CITANTES	DIFERENÇA	AUTORES	CITAÇÕES	CITANTES	DIFERENÇA
Hjørland, B	228	70	158	La Barre, Kathryn	12	09	03	Miles, Alistair	08	06	02
Smiraglia, RP	68	29	39	Ingwersen, Peter	12	08	04	Moya-Anegón, F	08	06	02
Olson, HÁ	61	31	30	***Lykke Nielsen, M***	***12***	***05***	***07***	Saracevic, Tefko	08	06	02
Szostak, R	59	13	46	***Pinho, FA***	***12***	***05***	***07***	Andersen, Hanne	08	05	03
Mai, J-E	47	25	22	Bates, MJ	11	11	-	Wilson, TD	08	05	03
Dahlberg, I	44	28	16	Hendler, James	11	11	-	Bizer, Christian	08	04	04
Gnoli, Claudio	35	18	17	Lassila, Ora	11	11	-	Klein, JT	08	04	04
Tennis, JT	34	23	11	Jacob, EK	11	10	01	Maltese, Vincenzo	08	04	04
Berners-Lee, Tim	32	18	14	Kuhn, TS	11	10	01	Milani, SO	08	04	04
Guimarães, JAC	32	13	19	Mazzocchi, F	11	10	01	Welty, CA	08	04	04
Ranganathan, SR	31	19	12	Spiteri, LF	11	10	01	***Cahier, JP***	***08***	***03***	***05***
Zeng, ML	30	21	09	Furner, Jonathan	11	09	02	***Rosch, E***	***08***	***03***	***05***
García Gutiérrez, A	30	08	22	Taylor, AG	11	09	02	***Sørensen, Bent***	***08***	***02***	***06***
Thellefsen, TL	***27***	***09***	***18***	Keizer, Johannes	11	07	04	Belkin, NJ	07	07	-
Soergel, Dagobert	26	16	10	Pejtersen, AM	11	05	06	Börner, Katy	07	07	-
Ostwald, W	***26***	***03***	***23***	Poli, Roberto	11	05	06	Cronin, Blaise	07	07	-
Hodge, Gail	25	23	02	***Ding, Y***	***11***	***04***	***07***	Foskett, AC	07	07	-
Svenonius, Elaine	25	18	07	***Pepper, Steve***	***11***	***04***	***07***	Foskett, DJ	07	07	-
Thellefsen, MM	24	11	13	Berman, Sanford	10	09	01	Rowley, JE	07	07	-
Albrechtsen, H	23	19	04	Gilchrist, Alan	10	08	02	Sowa, JF	07	07	-
Broughton, Vanda	22	17	05	Hartel, Jenna	10	08	02	Day, RE	07	06	01
Chan, LM	21	16	05	Miksa, FL	10	08	02	Doerr, Martin	07	06	01
López-Huertas, MJ	21	16	05	Losee, RM	10	06	04	Frické, Martin	07	06	01
Frohmann, Bernd	21	12	09	Van Harmelen, F	10	06	04	Golder, SA	07	06	01
Vickery, BC	20	15	05	Giunchiglia, F	10	04	06	Lauser, Boris	07	06	01
Beghtol, Clare	20	14	06	Tognoli, NB	10	04	06	Rorty, Richard	07	06	01
Gruber, TR	17	14	03	van den Heuvel, C	10	04	06	Vizine-Goetz, D	07	06	01
McGuinness, DL	17	14	03	***Tudhope, D***	***10***	***02***	***08***	Willis, Craig	07	06	01
Star, SL	17	13	04	Bawden, David	09	09	-	Bufrem, LS	07	05	02
Smith, Barry	17	11	06	Liang, AC	09	08	01	Chen, Chaomei	07	05	02
Peirce, CS	***17***	***08***	***09***	Dousa, TM	09	07	02	Dumais, ST	07	05	02
Rayward, WB	***17***	***08***	***09***	Järvelin, K	09	07	02	Farrow, JF	07	05	02
Brier, Søren	***17***	***05***	***12***	McCulloch, Emma	09	07	02	Fox, MJ	07	05	02
Lancaster, FW	16	13	03	Panzer, Michael	09	07	02	Jiménez-Contreras, E	07	05	02
Feinberg, M	16	08	08	Salton, G	09	07	02	Miller, GA	07	05	02
Martínez-Ávila, D	16	06	10	Chowdhury, GG	09	06	03	Brickley, Dan	07	04	03
Fujita, MSL	***16***	***03***	***13***	Garshol, LM	09	06	03	Cabré, MT	07	04	03
Friedman, Alon	15	10	05	Lakoff, G	09	05	04	Campos, MLA	07	04	03

Continua...

Tabela 3 – Ranking de autores comparando citações totais com documentos citantes do *corpus* do Conjunto B (organizado por citações totais e com 153 posições) - **continuação**

AUTORES	CITAÇÕES	CITANTES	DIFERENÇA	AUTORES	CITAÇÕES	CITANTES	DIFERENÇA	AUTORES	CITAÇÕES	CITANTES	DIFERENÇA
Green, Rebecca	15	10	05	***Bourdieu, Pierre***	***09***	***03***	***06***	Dutta, Biswanath	07	04	03
Buckland, MK	14	12	02	***Golub, K***	***09***	***03***	***06***	Fernández-Molina, JC	07	04	03
Budd, JM	***14***	***08***	***06***	***Witten, IH***	***09***	***03***	***06***	Hapke, T	07	04	03
Small, HG	14	06	08	***Herre, Heinrich***	***09***	***02***	***07***	Marteleto, RM	07	04	03
Wittgenstein, L	13	08	05	***Fahlman, SE***	***09***	***01***	***08***	Tudhope, Douglas	07	04	03
Guarino, Nicola	13	06	07	Aitchison, Jean	08	07	01	White, HD	07	04	03
Otlet, P	13	06	07	Eco, Umberto	08	07	01	***Bakhtin, Mikhail***	***07***	***03***	***04***
Bowker, GC	12	12	-	Ibekwe-SanJuan, F	08	07	01	***Blei, DM***	***07***	***03***	***04***
Noy, NF	12	11	01	Shiri, AA	08	07	01	***Tillett, BB***	***07***	***03***	***04***
Andersen, Jack	12	10	02	Stock, WG	08	07	01	***Weller, K***	***07***	***03***	***04***
Greenberg, Jane	12	10	02	Bechhofer, Sean	08	06	02	***Beak, Jihee***	***07***	***02***	***05***
Hansson, Joacim	12	10	02	Ellis, David	08	06	02	***Sigel, Alexander***	***07***	***02***	***05***
Foucault, Michel	12	09	03	Lambe, Patrick	08	06	02	***Zacklad, Manuel***	***07***	***02***	***05***

Fonte: dados da pesquisa, 2020.

Vale salientar que esses indicadores são próximos e sua distinção só é possível com um domínio adequado dos dados dos documentos citantes e que, dependendo da forma de recuperação, nem sempre é possível estabelecer essa diferença.

Para identificar a proximidade dos indicadores foi utilizada a medida de correlação "r de Pearson" em alguns arranjos do ranking de autores que inclui todas as posições das referências. Esses dados evidenciam que os resultados apresentam forte correlação mesmo com as suas especificidades. O Quadro 1 apresenta os valores de correlação para cinco arranjos desses dados.

Quadro 1 – Medidas de correlação do ranking de autores do corpus do Conjunto B de dados, considerando "totais de citações" e "citações por documentos citantes"

ARRANJO / Valor da correlação (*r* de Pearson)		
1º	5333 posições - *Ranking* considerando **todos os autores** das referências	,903
2º	933 posições - Considerando os autores a partir de **dois documentos citantes**	,905
3º	1249 posições - Considerando os autores a partir de **duas citações**	,893
4º	153 posições - Recorte do *ranking* apresentado na Tabela 3 desse texto	,889
5º	156 posições - Recorte do *ranking* com número de autores próximos ao da Tabela 3 organizados pelo número de documentos citantes (quatro ou mais)	,913

Fonte: dados da pesquisa, 2020.

A força de correlação se mostra próxima e alta para todos os arranjos, mas os maiores valores estão ligados à organização do ranking pelo número de documentos citantes (2º e 5º itens

do Quadro 1), que são valores mais consistentes e regulares que os valores de citações totais. Além da questão da classificação dos autores para a inclusão em estudos, é preciso salientar que a escolha do indicador pode ter uma relação com o valor atribuído na diagonal da matriz quadrada e isso influencia diretamente nos resultados da aplicação de algum índice de similaridade, etapas posteriores em uma ACA. Assim, o valor de citação medido pelo total de documentos citantes pode ser considerado um indicador confiável para definir o número de autores que farão parte de um estudo de cocitação, dependendo dos objetivos propostos.

Estabelecido um indicador que aproveita melhor a distribuição dos autores em um corpus de análise, é momento de efetivamente estabelecer um ponto de corte, ou seja, definir um número de autores. O Princípio de Pareto (Becker 2015) se apresenta como uma alternativa, pois determina uma proporção entre dados triviais e relevantes para explicar alguns fenômenos. Embora essa proporção não seja absoluta, o mais comum é a utilização da relação 80-20, onde aproximadamente 80% dos dados são considerados triviais ou de dispersão e 20% são mais significativos. Vale salientar que mais do que determinar um número de autores para a inclusão em uma matriz de cocitação, é necessário relacionar de alguma forma essa definição com os dados que não serão utilizados ou considerados como dispersão. Penan5 (1989 apud McCain 1990 p. 435) utilizou essa proporção de autores, mas usando dados de citação e cocitação.

Utilizando os valores de citação por documentos citantes, foi identificado um número específico de autores cuja soma das citações ficasse próximo de 20%. A Tabela 4 apresenta os dados dessa proposta, a partir dos valores do Conjunto B de referências, especificamente do ranking que inclui todos os autores das referências (5.333 posições). Assim, 227 autores (destaque), citados em, pelo menos, quatro artigos, seriam analisados em uma ACA, e a soma das suas citações corresponde a 21,89% do total de 7651 citações, ou seja, 1675 citações contadas a partir dos documentos citantes, como evidenciam os resultados parciais na linha em destaque.

[5] Penan, H. "*Pour une gestion bibliometrique de l' information scientifique et technique des entreprises. Application en theorie microeconomique et financiere*". 1989. Universite des Sciences Sociales, Doctoral dissertation [Texto completo não localizado].

Tabela 4 – Distribuição do número de autores e suas citações por documento citante (N = 151 artigo, considerando todos os autores das referências)

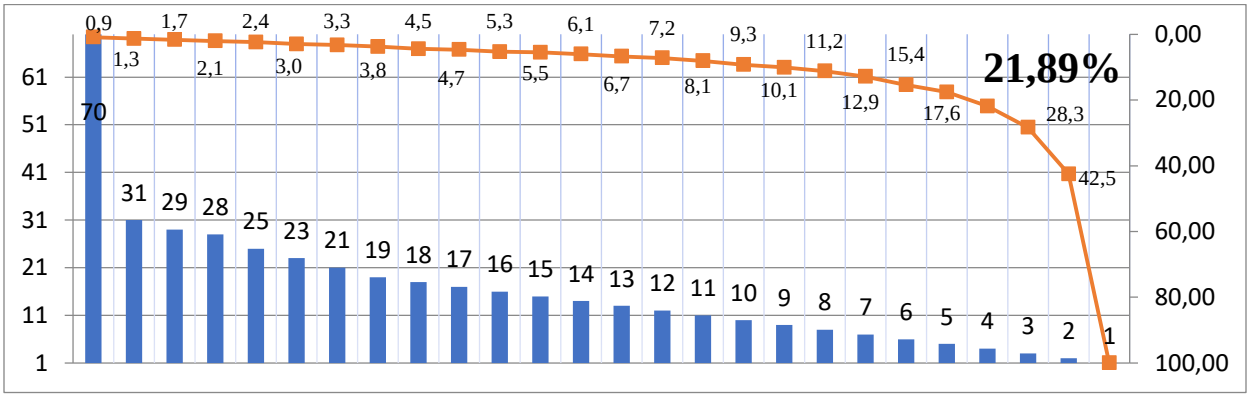
			Nº de autores	**Citações a partir dos citantes**	**% Citações**	**% Acumulado**
Autores citados em	70	artigos	01	70	0,91	0,91
Autores citados em	31	artigos	01	31	0,41	1,32
Autores citados em	29	artigos	01	29	0,38	1,70
Autores citados em	28	artigos	01	28	0,37	2,07
Autores citados em	25	artigos	01	25	0,33	2,39
Autores citados em	23	artigos	02	46	0,60	2,99
Autores citados em	21	artigos	01	21	0,27	3,27
Autores citados em	19	artigos	02	38	0,50	3,76
Autores citados em	18	artigos	03	54	0,71	4,47
Autores citados em	17	artigos	01	17	0,22	4,69
Autores citados em	16	artigos	03	48	0,63	5,32
Autores citados em	15	artigos	01	15	0,20	5,52
Autores citados em	14	artigos	03	42	0,55	6,06
Autores citados em	13	artigos	04	52	0,68	6,74
Autores citados em	12	artigos	03	36	0,47	7,21
Autores citados em	11	artigos	06	66	0,86	8,08
Autores citados em	10	artigos	09	90	1,18	9,25
Autores citados em	09	artigos	07	63	0,82	10,08
Autores citados em	08	artigos	11	88	1,15	11,23
Autores citados em	07	artigos	18	126	1,65	12,87
Autores citados em	06	artigos	32	192	2,51	15,38
Autores citados em	05	artigos	34	170	2,22	17,61
Autores citados em	***04***	***artigos***	***82***	***328***	***4,29***	***21,89***
Totais parciais	--	--	***227 autores***	***1675 citações***	--	--
Autores citados em	03	artigos	164	492	6,43	28,32
Autores citados em	02	artigos	542	1084	14,17	42,49
Autores citados em	01	artigos	4400	4400	57,51	100,00
Total de citações por documentos citantes				7651	100,00	

Fonte: dados da pesquisa, 2020.

A Figura 1 ilustra esses resultados apresentando as citações por documentos citantes como categoria (colunas em azul) e o percentual acumulado de citações em uma curva invertida (vermelho), destacando os 21,89%, que sugerem que autores citados, em pelo menos, quatro documentos são elegíveis para uma análise. No caso do *ranking* com os primeiros autores, o

ponto de corte é estabelecido a partir de cinco documentos citantes, totalizando 96 autores, que correspondem a 21,6% do total das citações contadas a partir dos documentos citantes. A intenção foi selecionar os autores pelo rastro dos documentos onde são citados e o número total de autores não precisa ser redondo, como ocorre em outros estudos (White, e McCain 1998; Zhao, e Strotmann 2014; Bu, *et al*. 2016; Bu, *et al*. 2017).

Figura 1 – Distribuição de valores de documentos citantes como categoria para definição de ponto de corte (todos os autores das referências – ranking com 5.333 autores – Conjunto B)



Fonte: dados da pesquisa, 2020.

Esse primeiro teste se mostrou parcialmente equivocado por dois motivos: o número de documentos citantes por autor ficou baixo e a quantidade de autores parecia muito grande em relação ao tamanho dos rankings. Ainda que McCain (1990) tenha tido bons resultados com valores baixos, estudos com grandes quantidades de dados (White, e McCain 1998; Zhao, e Strotmann 2014; Khasseh, et al. 2018; Zhao, et al. 2018) selecionaram menos autores. Mas a preocupação principal foi o número limite de documentos citantes por autor.

Um segundo teste com os mesmos dados descartou todos os autores citados em apenas um documento citante nos dois rankings. Esses autores foram considerados sem um alcance mínimo de relacionamento no corpus, ou seja, não foram citados em, pelo menos, dois documentos. Essa condição de ser citado em apenas um documento, no contexto do corpus, caracteriza esses autores mais pela relação com o documento citante do que pela relação com os outros autores citados. Para os estudos de cocitação, no entendimento desta pesquisa, são dados de natureza dispersiva absoluta.

Tabela 5 – Distribuição do número de autores citados em pelo menos dois documentos e suas citações por documento citante (N = 151 artigos, considerando todos os autores das referências)

			Nº de autores	**Citações a partir dos citantes**	**% Citações**	**% Acumulado**
Autores citados em	70	Artigos	01	70	2,15	2,15
Autores citados em	31	artigos	01	31	0,95	3,11
Autores citados em	29	artigos	01	29	0,89	4,00
Autores citados em	28	artigos	01	28	0,86	4,86
Autores citados em	25	artigos	01	25	0,77	5,63
Autores citados em	23	artigos	02	46	1,41	7,04
Autores citados em	21	artigos	01	21	0,65	7,69
Autores citados em	19	artigos	02	38	1,17	8,86
Autores citados em	18	artigos	03	54	1,66	10,52
Autores citados em	17	artigos	01	17	0,52	11,04
Autores citados em	16	artigos	03	48	1,48	12,52
Autores citados em	15	artigos	01	15	0,46	12,98
Autores citados em	14	artigos	03	42	1,29	14,27
Autores citados em	13	artigos	04	52	1,60	15,87
Autores citados em	12	artigos	03	36	1,11	16,98
Autores citados em	11	artigos	06	66	2,03	19,01
Autores citados em	***10***	***artigos***	***09***	***90***	***2,77***	***21,78***
Totais parciais	--	--	***43 autores***	***708 citações***	--	--
Autores citados em	09	artigos	07	63	1,94	23,72
Autores citados em	08	artigos	11	88	2,71	26,42
Autores citados em	07	artigos	18	126	3,88	30,30
Autores citados em	06	artigos	32	192	5,91	36,20
Autores citados em	05	artigos	34	170	5,23	41,43
Autores citados em	04	artigos	82	328	10,09	51,52
Autores citados em	03	artigos	164	492	15,13	66,66
Autores citados em	02	artigos	542	1084	33,34	100,00
Total de citações por documentos citantes				3251	100,00	

Fonte: dados da pesquisa, 2020.

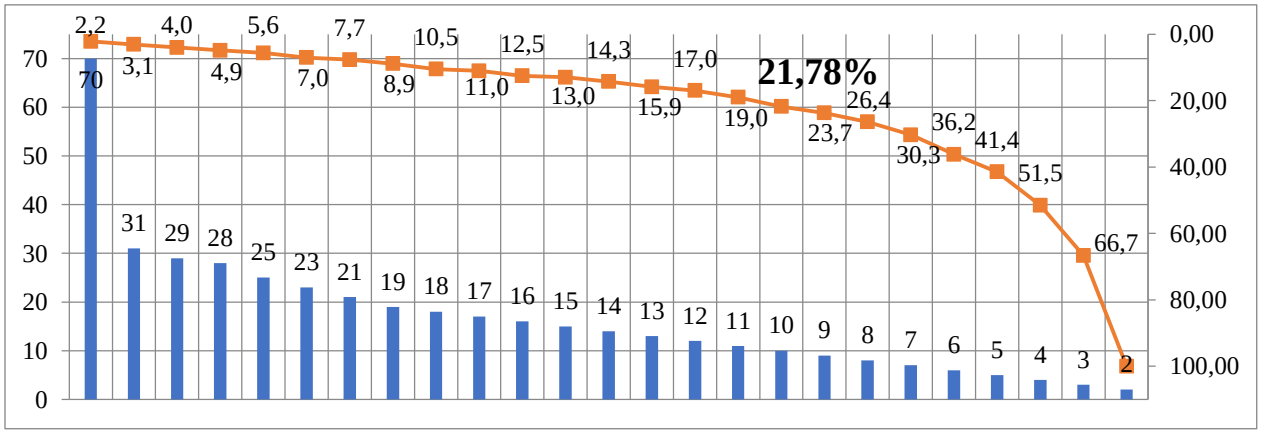
Assim, utilizando os valores de citações por documentos citantes, mas desconsiderando 4.400 autores (ver Tabela 4), citados em um único documento, foi identificado o número específico de autores, citados em dois documentos ou mais, cuja soma das citações ficasse próximo de 20%. A Tabela 5 apresenta os dados, que indicaram 43 autores (destaque), citados

em pelo menos 10 artigos, cuja soma das suas citações corresponde a 21,78% do total de 3251 citações, ou seja, 708 citações contadas a partir dos documentos citantes, como evidenciado.

A Figura 2 ilustra os resultados do segundo teste. Nessas condições o ranking ficou com 933 posições, sendo que 4.400 autores foram retirados dessa fase da análise, como já salientado. A aplicação no ranking com os primeiros autores das referências definiu o ponto de corte em autores citados em pelo menos 12 documentos (22,4% das citações oriundas de documentos citantes). O ranking nas condições estabelecidas ficou com 519 posições, sendo que 2.419 autores foram retirados dessa fase da análise. Considerando as posições analisadas, 22 autores fariam parte de um estudo de cocitação.

A inclusão de mais autores para esse caso não seria arbitrária, pois o que parece efetivamente relevante na proposta apresentada é fazer a relação adequada dos dados analisados com os dados não analisados.

Figura 2 – Distribuição de valores de documentos citantes como categoria para definição de ponto de corte (autores citados em dois documentos ou mais – ranking com 933 autores)



Fonte: dados da pesquisa, 2020.

Propõem-se, portanto, uma taxonomia para os dados a partir da proposta estabelecida para definição do ponto de corte para a seleção dos autores em ACA, a saber: **dados analisados**: incluindo autores que, somados os valores de citação por documentos citantes, correspondam a aproximadamente 20% dos dados do *corpus*, excluindo valores de dispersão absoluta; **dados intermediários de dispersão**: inclui autores citados em, pelo menos, dois documentos até o limite estabelecido para a análise e pode ser dividido em dois subconjuntos: metade próxima aos

dados analisados: autores com alta potencialidade de uso em estudo de cocitação (nos dados ilustrados na Figura 2, seriam autores que receberam entre seis e nove citações oriundas de documentos citantes); e metade próxima aos dados de dispersão absoluta: autores com menor potencialidade de uso em estudos de cocitação (nos dados apresentados na Figura 2, seriam autores citados entre dois e cinco documentos citantes); e, por fim, **dados de dispersão absoluta**: autores citados em apenas um documento.

Essa forma para selecionar os autores em um estudo de cocitação não tem a pretensão de ser definitiva, mas ressalta-se que os estudos dessa natureza deveriam apresentar dados descritivos que mostram a proporcionalidade dos autores selecionados e não selecionados no contexto de um estudo. Para esse fim, esta abordagem parece válida.

Para complementar a validação da proposta, segue a apresentação dos resultados com os dados do Conjunto A de referências. Ainda que sejam de natureza temática similar, os dois conjuntos são de tamanhos diferentes: 421 *versus* 151 artigos citantes; e 17.992 referências totais *versus* 5.771.

A Tabela 6 apresenta um *ranking* com 107 posições que compara valores de citações por documentos citantes, citações totais e cocitação do autor com ele mesmo. Os dados são organizados a partir de autores citados em pelo menos 15 artigos e consideram todas as posições de autorias das referências.

Os autores “Raghavan, P”, “Ng, AY”, “Hendler, JA" e “Park, JR” (destacado com bordas vermelhas) demonstram que a “cocitação do autor com ele mesmo” tem pouco impacto na distribuição dos autores pelos documentos citantes e mesmo sendo uma informação de natureza relacional, assim como a cocitação entre autores distintos, parece um dado arbitrário para definir a seleção de autores em ACA, tanto quanto o número total de citações, como indicam Ahlgren *et at*. (2003). Os autores mencionados, “Raghavan, P”, “Ng, AY”, “Hendler, JA" e “Park, JR”, não apresentam nenhuma cocitação consigo mesmo, mas possuem um alcance significativo em termos de artigos citantes, com valores iguais a 30, 26, 19 e 17, respectivamente.

Tabela 6 – Ranking de autores, comparando valores de documentos citantes, citações totais e cocitação do autor com ele mesmo (107 posições – 15 ou mais documentos citantes).

AUTORES	DC	CT	CoA	AUTORES	DC	CT	CoA
Hjørland, B	68	191	34	***Bawden, D***	***20***	***26***	***03***
Croft, WB	60	136	28	***Li, Y***	***20***	***25***	***04***
Järvelin, K	45	77	14	***Harman, DK***	***20***	***23***	***02***
Baeza-Yates, R	45	60	09	Guimarães, JAC	19	40	08
Salton, G	44	62	11	Wilson, TD	19	38	08
Robertson, SE	42	69	16	Callan, JP	19	38	07
Spink, A	41	74	21	***Ruthven, I***	***19***	***34***	***02***
Saracevic, T	40	68	15	***Hu, X***	***19***	***32***	***04***
Belkin, NJ	38	95	20	Garfield, E	19	30	06
Jansen, BJ	37	66	15	***Sparck Jones, K***	***19***	***26***	***05***
Manning, CD	36	49	08	***Ellis, D***	***19***	***25***	***05***
Voorhees, EM	34	59	16	Hendler, JA	19	19	00
Olson, HA	33	106	17	***Li, H***	***18***	***36***	***05***
Dumais, ST	33	65	16	***Fox, EA***	***18***	***35***	***04***
Smiraglia, RP	32	73	18	Lafferty, JD	18	34	10
Ingwersen, P	32	53	11	Nie, JY	18	33	08
Marchionini, G	32	49	13	***Zobel, J***	***18***	***29***	***05***
Buckley, C	32	49	09	***Moffat, A***	***18***	***27***	***05***
Zhang, J	32	46	06	***Jones, Rosie***	***18***	***23***	***05***
Ribeiro-Neto, B	***32***	***37***	***03***	***MacFarlane, A***	***18***	***22***	***04***
Bates, MJ	30	47	10	Soboroff, I	17	30	07
Schutze, H	***30***	***35***	***05***	Macdonald, C	17	29	06
Raghavan, P	30	30	00	***Wildemuth, BM***	***17***	***28***	***05***
Tennis, JT	29	54	12	***Bar-Ilan, J***	***17***	***24***	***04***
Blei, DM	28	51	10	***Svenonius, E***	***17***	***21***	***04***
Metzler, D	28	40	08	***Chen, H***	***17***	***21***	***01***
Sanderson, M	28	37	06	***Page, L***	***17***	***19***	***02***
Vakkari, P	27	47	08	***Jones, SA***	***17***	***18***	***01***
White, RW	26	70	12	Park, JR	17	17	00
Ng, AY	26	26	00	Leydesdorff, L	16	51	09
Lancaster, FW	25	36	07	Glänzel, W	16	44	10
Van Rijsbergen, CJ	***25***	***29***	***03***	***Small, H***	***16***	***33***	***05***
Ding, Y	24	46	09	***Wolfram, D***	***16***	***28***	***03***

Continua

Tabela 6 – Ranking de autores, comparando valores de documentos citantes, citações totais e cocitação do autor com ele mesmo (107 posições – 15 ou mais documentos citantes) (continuação)

AUTORES	DC	CT	CoA	AUTORES	DC	CT	CoA
Berners-Lee, T	24	39	10	Xu, J	16	27	06
Ounis, I	24	39	07	***Han, J***	***16***	***25***	***03***
Furner, J	24	32	07	***McCallum, A***	***16***	***25***	***02***
Zeng, ML	24	32	07	***Singhal, A***	***16***	***19***	***03***
Jordan, MI	***24***	***31***	***03***	***Buckland, MK***	***16***	***18***	***02***
Witten, IH	23	43	07	***Chan, LM***	***16***	***16***	***00***
De Rijke, M	23	41	09	Yan, E	15	29	07
Allan, J	23	39	06	***Dervin, B***	***15***	***21***	***04***
Li, X	***23***	***28***	***05***	***Chowdhury, GG***	***15***	***21***	***02***
Zhai, CX	22	56	10	***Toms, EG***	***15***	***20***	***02***
Kelly, D	22	40	08	***Fidel, R***	***15***	***19***	***02***
Beghtol, C	22	32	06	***Kamps, J***	***15***	***19***	***02***
Teevan, J	***22***	***31***	***05***	***Van Raan, AFJ***	***15***	***18***	***02***
Walker, S	***22***	***24***	***02***	***Brin, S***	***15***	***17***	***02***
White, HD	21	50	11	***Zhang, M***	***15***	***17***	***02***
Kuhlthau, CC	***21***	***36***	***05***	***Broder, AZ***	***15***	***16***	***01***
Joachims, T	21	34	06	***Cronin, B***	***15***	***16***	***01***
Dahlberg, I	21	30	07	***Yu, Y***	***15***	***16***	***01***
Borlund, P	20	35	06	***Nejdl, W***	***15***	***15***	***00***
Mai, JE	20	32	08	***Strohman, T***	***15***	***15***	***00***
Chen, C	***20***	***28***	***04***				

Legendas: DC: documentos citantes; CT: citações totais; CoA: cocitação do autor com ele mesmo.

Fonte: dados da pesquisa, 2020.

Os outros autores em destaque (**negrito**, *itálico* e sublinhado), com cocitação consigo mesmo igual ou menor a cinco, evidenciam a distância dos dados de cocitação do autor com ele mesmo para dados de citações totais e de citações por documentos citantes. Salienta-se a relevância dessa informação, mas não se recomenda seu uso para definir a seleção de autores em ACA, a não ser com um objetivo estritamente relacionado com a natureza dessa informação.

Os autores com destaque em verde nas bordas evidenciam a diferença dos indicadores "citação por documentos citantes" e "citações totais", e reforçam os dados já apresentados na Tabela 3, que está organizada pelas citações totais. Essas informações reforçam que o uso dos valores de citação a partir de documentos citantes é mais adequado para definir a seleção dos autores em estudos de cocitação, principalmente se o objetivo for entender um domínio.

Quanto aos valores de cocitação do autor com ele mesmo, os dados evidenciam que mais de 85% dos autores, nos dois *rankings* (considerando todas as posições e as primeiras), não possuem cocitação desse tipo e que uma única ocorrência fica em 11,18% para o *ranking* com todos os autores (N=19.590) e 10,44%, considerando as primeiras posições das referências (N=9.339).

O Quadro 2 apresenta as medidas de correlação (*r* de Pearson) dos três indicadores em alguns arranjos para os dois *rankings*. Os dados mostram uma relação próxima dos indicadores nos quatro cenários propostos: todas as posições dos *rankings*; considerando autores citados em dois ou mais documentos; considerando autores a partir de uma cocitação com ele mesmo; e autores a partir de duas citações totais. Os destaques no Quadro 2 (fundo cinza) mostram valores de correlação mais baixos entre os indicadores "documentos citantes (DC)" e "cocitação do autor com ele mesmo (CoA)".

Quanto ao ponto de corte utilizando os dados do Conjunto A de referências, os testes iniciais para definir os autores partiram dos valores de citação por documentos citantes, e incluíram os dados de dispersão absoluta, ou seja, todos os autores dos dois rankings. Para o ranking com todos os autores das referências, o ponto de corte ficou estabelecido inicialmente em 606 autores citados em, pelo menos, seis artigos, e os valores de citação a partir dos documentos citantes corresponderam a 20,90% do total das 31.849 citações por documentos citantes. Para o ranking com os primeiros autores das referências, o ponto de corte ficou definido inicialmente em 331 autores citados em, pelos menos, cinco artigos, e a soma dos valores de citações a partir de documentos citantes corresponderam a 21,55% do total das 14.326 citações contadas a partir dos documentos citantes.

Esses dados foram ao encontro dos resultados obtidos na análise com dados do Conjunto B de referências, ou seja, muitos autores para um número de documentos citantes baixo. Os dados de dispersão absoluta interferem de maneira significativa na análise e são retirados para realizar a segunda rodada de testes, que acabam por definir o número de autores para as análises.

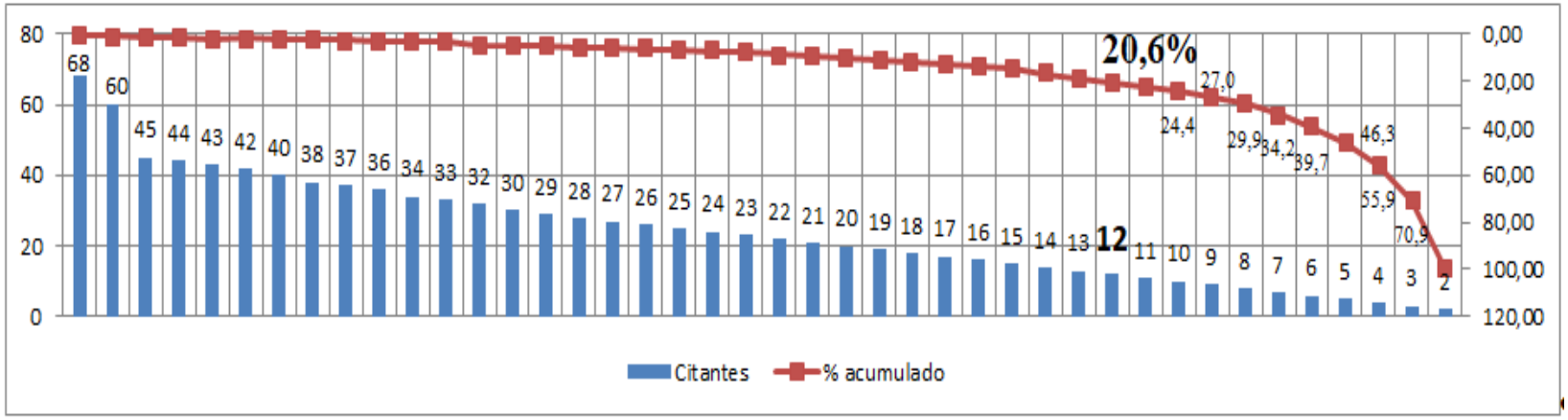
Quadro 2 – Medidas de correlação dos rankings de autores, do corpus do Conjunto A, considerando documentos citantes, citações totais e cocitação do autor com ele mesmo

Todos os autores das referências				**Primeiros autores das referências**			
19590 posições		*DC*	*CT*	9339 posições		*DC*	*CT*
	DC	1			DC	1	
	CT	0,933	1		CT	0,922	1
	CoA	**0,791**	0,91		**CoA**	**0,77**	0,902
4509 posições (dois ou mais citantes)		*DC*	*CT*	1891 posições (dois ou mais citantes)		*DC*	*CT*
	DC	1			DC	1	
	CT	0,928	1		CT	0,92	1
	CoA	0,81	0,923		CoA	0,81	0,924
2759 posições (uma ou mais cocitação)		*DC*	*CT*	1179 posições (uma ou mais cocitação)		*DC*	*CT*
	DC	1			DC	1	
	CT	0,93	1		CT	0,92	1
	CoA	0,836	0,931		CoA	0,831	0,936
5610 posições (duas ou mais citações totais)		*DC*	*CT*	2426 posições (duas ou mais citações totais)		*DC*	*CT*
	DC	1			DC	1	
	CT	0,92	1		CT	0,911	1
	CoA	**0,747**	0,893		**CoA**	**0,732**	0,888
Legenda: DC - documentos citantes; CT - citações totais; CoA - Cocitação do autor com ele mesmo							

Fonte: dados da pesquisa, 2020.

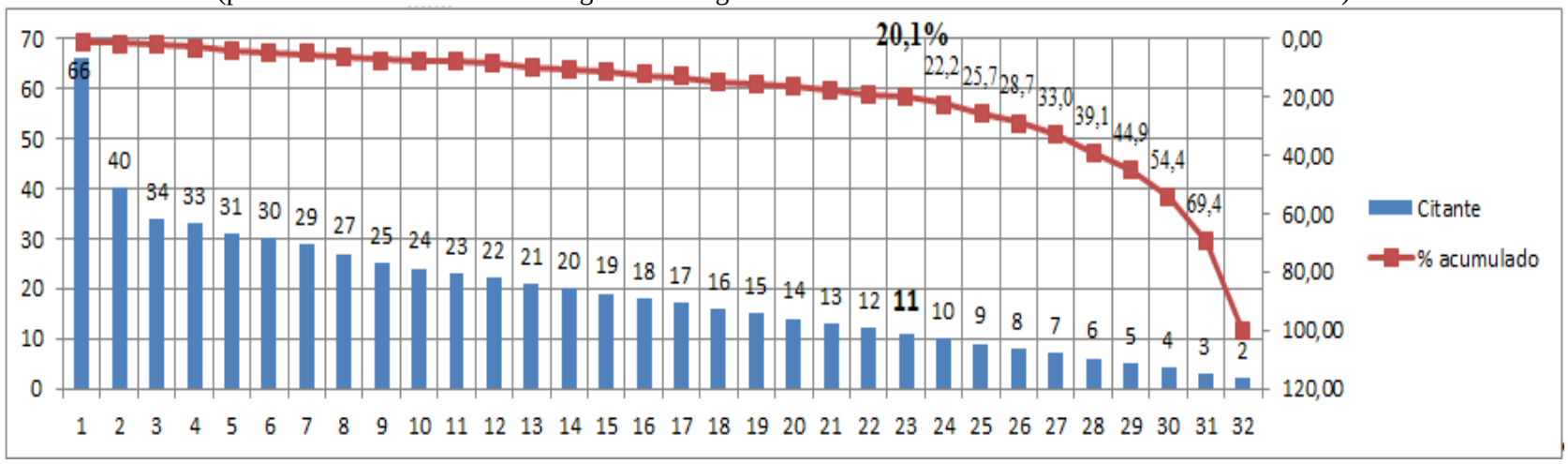
A Figura 3 apresenta a distribuição dos valores de documentos citantes como categoria (em azul) e o percentual acumulado (em vermelho) da soma das citações por documentos citantes, considerando todos os autores das referências. O gráfico destaca que os autores citados em pelo menos 12 artigos, que correspondem a 20,6% dos valores de citações por documentos citantes, excluindo os dados de dispersão absoluta, fariam parte de uma ACA e o total de autores é 180, número próximo às análises de White, e McCain (1998) e Zhao, e Strotmann (2014). A inclusão de autores citados em 10 e 11 documentos, por exemplo, que fazem parte dos dados intermediários de dispersão, acrescentaria 27 e 33 autores, respectivamente. Cabe ao pesquisador definir essa necessidade e justificar a escolha.

Figura 3 – Distribuição de valores de documentos citantes (dois ou mais) como categoria para definição de ponto de corte
(todos os autores – N = 421 artigos – ranking com 4.509 autores citados em dois ou mais documentos)



Fonte: dados da pesquisa, 2020.

Figura 4 – Distribuição de valores de documentos citantes (dois ou mais) como categoria para definição de ponto de corte
(primeiros autores – N = 421 artigos – ranking com 1891 autores citados em dois ou mais documentos)



Fonte: dados da pesquisa, 2020.

A Figura 4 segue o esquema da figura anterior e apresenta a distribuição dos valores de citação por documentos citantes como categoria (em azul) e o percentual acumulado (em vermelho) da soma dos valores das citações, considerando as primeiras posições das referências. A figura destaca que os autores citados em pelo menos 11 documentos, que representam 20,1% dos valores somados de citações oriundas de documentos citantes, excluindo os dados de dispersão absoluta, fariam parte de uma ACA, ou seja, 72 autores. Como já salientado, a inclusão de autores que fazem parte dos dados intermediários de dispersão é possível. Autores citados em 10 documentos acrescentariam 15 pesquisadores à análise, totalizando 87.

A ACA é uma técnica complexa, desde a extração dos dados até a utilização dos indicadores relacionados aos autores. Descrever, mesmo que exaustivamente, algumas escolhas metodológicas enriquece a literatura sobre o tema. Isso precisa ser planejado mesmo antes da recuperação das informações, já que a forma de extração é crucial para alcançar certo nível de detalhamento.

A definição do ponto de corte, independente da proporção, requer um domínio amplo dos dados e é fundamental que seja estabelecido a partir do todo. No contexto deste estudo, os dados que consideram todos os autores, do Conjunto A de referências, resultam na análise da relação de 180 autores, que correspondem a 0,92% do total de autores distintos (19.590). Por esse motivo, é aconselhável ponderar os autores pela sua presença no corpus, seja pela contagem da citação total ou pela contagem do número de documentos que os citam, como é apresentado na Figura 3. Considerando os primeiros autores das referências essa proporção de 0,77%, ou seja, 72 sujeitos fariam parte de uma análise diante do total de 9.338 autores distintos do *ranking*.

Porém, mais significativo do que essa relação genérica é a identificação dos rastros dos autores diante dos documentos citantes. A Tabela 7 apresenta os valores de citação e de cocitação dos autores em relação aos 421 artigos citantes, considerando os dois *rankings*: todos os autores, que foi definido em 180, e primeiros autores das referências (incluindo autoria única), que foi definido em 72, a partir da proposta de ponto de corte.

Carvalho, R. A. de, et al. “Métodos de seleção de autores para estudos de cocitação: como definer um ponto de corte”. *Brazilian Journal of Information Science: Research trends*, vol.15, publicação continuada, 2021, e02109. doi10.36311/1981.1640.2001.v15.e02109

Tabela 7 – Valores de citação e de cocitação dos autores diante dos artigos citantes (N=421)

	Citação			Cocitação		
	Citam	**Não citam**	**% ***	**Há cocitação**	**Não há cocitação**	**%***
Todos os autores (n=180)	369	52	87,65	334	87	79,33
Primeiros autores (n=072)	331	90	78,62	260	161	61,76

* percentual de alcance para um total de 421 artigos citantes.

Fontes: dados da pesquisa, 2020.

Esse alcance em relação aos documentos citantes é um elemento que pode interferir na utilização dos dados intermediários de dispersão, ou seja, aumentar o número de autores na análise. Isso ocorre principalmente para o caso da utilização dos primeiros autores, que atinge pouco mais de 60% dos artigos citantes, com relações de cocitação entre os 72 autores definidos pela proposta de ponto de corte, ainda que sejam citados em quase 80% dos 421 documentos.

As informações da Tabela 7 evidenciam que as relações de cocitação analisadas não chegam a ocorrer em 80% dos artigos citantes, no melhor cenário, e essa proporção diminui conforme as escolhas metodológicas. A ordem parece clara: o uso de todos os autores das referências torna um estudo de cocitação mais abrangente e o uso dos primeiros autores tem um alcance intermediário, considerando, evidentemente, o mesmo número de documentos citantes, ainda que os rankings sejam de grandezas diferentes. Oliveira e Grácio (2011) apresentam uma situação onde 15 documentos citantes não citam nenhum dos autores que formam a rede de cocitação apresentada pelas autoras.

Levando-se em consideração os 421 artigos dessa segunda parte da análise, ainda que o foco do estudo seja metodológico, pode-se afirmar que essa enorme dispersão torna a produção científica da área de Organização do Conhecimento e da Recuperação da Informação muito rica pela diversidade de fontes e autores citados.

4 Considerações finais

As análises de cocitação se sustentam como técnicas de mapeamento da ciência que se estabelecem mais por escolhas em cada etapa da operacionalização, que podem implicar em

omissões informacionais, do que por aplicações que possam ser consideradas erradas ou equivocadas, principalmente diante dos objetivos de uma determinada pesquisa.

O presente estudo, no entanto, mostrou que a contagem da citação por documento citante distribui melhor os autores em um corpus de análise do que as citações totais. A cocitação do autor com ele mesmo mostrou-se um indicador pouco apropriado, pois não tem relação direta com a influência de um autor em um conjunto de documentos, principalmente para estudos tradicionais de cocitação de autores.

O princípio de Pareto, a relação 80-20, se mostrou válida para definir o ponto de corte, mas a influência dos autores citados uma única vez nos corpus analisados foi preponderante e exigiu a retirada desses dados dos cálculos da proporção, o que ocasionou a proposta de uma "taxonomia" dos dados de cocitação: os dados analisados, dados intermediários de dispersão e os dados de dispersão absoluta. Contudo, o mais significativo no uso da proporção 80-20 é entender o todo dos dados para definir os autores analisados. Fica evidente, no entanto, que isso não é comum e nem simples na coleta de dados, principalmente utilizando apenas os metadados de bases de dados.

Outros dois elementos significativos que merecem destaque são: i) a proposta exige que os dois rankings, o que leva em consideração todos os autores das referências e o inclui apenas a primeira posição, sejam tratados como grandezas distintas, e estudos futuros devem comparar pontos de cortes com o mesmo número de autores para identificar, principalmente, o alcance desses dados diante dos documentos citantes, como apresentado na Tabela 7; e ii) ampliar a descrição dos padrões estruturais (corpus de análise/citantes) que alicerçam uma ACA, mensurando e comparando dados desconsiderados. No caso deste estudo em específico, são os dados de dispersão e o alcance do recorte diante dos documentos citantes (Tabela 7). Esta parece uma alternativa válida de pesquisa e discussão empírico-teórica no campo dos EC, e oferece mais credibilidade aos estudos de cocitação.

Por fim, espera-se que o presente estudo ofereça subsídios empíricos que ajudem em pesquisas metodológicas e aplicadas sobre o tema, e sugere-se para futuras pesquisas a aplicação dessa proposta em estudos de cocitação de documentos e de acoplamento bibliográfico.

Referências

Ahlgren, P., et al. “Requirements for a cocitation similarity measure, with special reference to Pearson’s Correlation Coefficient”. *Journal of the American Society for Information Science And Technology*, vol. 54, no. 6, 2003. pp. 550–60, doi: 10.1002/asi.10242.

Becker, J. L. *Estatística básica: transformando dados em informação*. Bookman, 2015.

Bu, Y., et al. “MFTACA: an author co-citation analysis method combined with metadata in full text”. *Proceedings of 16, International Conference On Scientometrics and Informetrics*, Wuhan, ISSI, 2017.

Bu, Y., et al. “MACA: a modified author co-citation analysis method combined with general descriptive metadata of citations”. *Scientometrics*, vol. 108, no. 1, 2016, pp. 143-166, doi: 10.1007/s11192-016-1959-5.

Carvalho, R. A., and Caregnato, S. E. “Análise de cocitação de autores – ACA: estudo exploratório comparando proximidade nas referências, seção do artigo e parágrafo”. *Anais do 18, Encontro Nacional de Pesquisa em Ciência da Informação*, Marília – SP. ANCIB, 2017, http://brapci.inf.br/index.php/article/download/59098.

Carvalho, R. A., et al. “A relação entre referências e menções: estudo exploratório em artigos na base de dados BRAPCI”. *Prisma Com*, no. 36, 2018, pp. 99- 115, doi: 10.21747/16463153/36a6.

Carvalho, R. A., et al. “Interpretação e validação de agrupamentos em análise de cocitação de autores: estudo exploratório e metodológico”. *Em Questão*, v.25, no. 2, 2019, pp. 89-116, doi: https://doi.org/10.19132/1808-5245252.89-116.

Eom, S. B. *Author cocitation analysis: quantitative methods for mapping the intellectual structure of an academic discipline*. Information Science Reference, 2009.

Eom, S. B., and Farris, R. S. “The contributions of organizational science to the development of decision support systems research subspecialties”. *Journal of the American Society for Information Science*, vol. 47, no. 12, 1996, pp. 941-52, doi: 10.1002/(SICI)1097-4571(199612)47:12<941::AID-ASI7>3.0.CO;2-2.

Grácio, M. C. C., and Oliveira, E. F. T. “Análise de cocitação de autores: um estudo teórico-metodológico dos indicadores de proximidade, aplicados ao GT7 da ANCIB”. *Liinc em Revista*, vol. 9, no. 1, 2013a, pp. 196-213, doi: 10.18617/liinc.v9i1.527.

Grácio, M. C. C., and Oliveira, E. F. T. “Estudos de análise de cocitação de autores: uma abordagem teórico-metodológica para a compreensão de um domínio”. *Anais do 14, Encontro Nacional de*

Pesquisa em Ciência da Informação, Florianópolis, ANCIB, 2013b, http://enancib.ibict.br/index.php/enancib/xivenancib/paper/viewFile/4331/3454.

Grácio, M. C. C., and Oliveira, E. F. T. "Indicadores de proximidades em análise de cocitação de autores: um estudo comparativo entre coeficiente de Correlação de Pearson e Cosseno de Salton". *Informação & Sociedade: Estudos*, vol.25, no. 2, 2015, pp. 105-16, http://www.ies.ufpb.br/ojs/index.php/ies/article/view/105.

Hjørland, B. "Domain analysis in Information Science: eleven approaches – traditional as well as innovative". *Jornal of Documention*, vol. 58, no. 4, 2002, pp.422-462, doi: 10.1108/00220410210431136.

Kessler, M. M. "Bibliographic coupling between scientific papers". *American Documentation*, vol.14, no.1, 1963, pp.10-25, doi: 10.1002/asi.5090140103.

Khasseh, A. A., et al. "An author co-citation analysis of 37 years of iMetrics". *The Electronic Library*, vol. 36, no. 2, 2018, pp.319-337, doi: 10.1108/EL-09-2016-0191.

Liu, S., and Chen, C. "The proximity of co-citation". *Scientometrics*, no. 91, 2012, pp.495-511, doi: 10.1007/s11192-011-0575-7.

McCain, K. "Mapping author intellectual space: a technical overview". *Journal o the American Society for Information Science*, vol. 41, no. 6, 1990, pp.433-443, doi: 10.1002/(SICI)1097-4571(199009)41:6433::AID-ASI113.0.CO;2-Q.

McCain, K. "Mapping economics through the journal literature: an experiment in journal cocitation analysis". *Journal of the American Society for Information Science*, vol. 42, no. 4, 1991, pp.290-296, doi: 10.1002/(SICI)1097-4571(199105)42:4290::AID-ASI53.0.CO;2-9.

Oliveira, E. F. T., and Grácio, M. C. C. "Visibilidade dos pesquisadores no GT7 da ANCIB: um estudo de cocitações". *Anais do 12, Encontro Nacional de Pesquisa em Ciência da Informação*, Brasília – DF. ANCIB, 2011, http://enancib.ibict.br/index.php/enancib/enancibXII/paper/view/605.

Reitz, J. M. *ODLIS - Online Dictionary for Library and Information Science*. c2014, http://www.abc-clio.com/ODLIS/odlis_c.aspx.

Schneider, J. W., et al. "A comparative study of first and all author co-citation counting, and two different matrix generation approaches applied for author co-citation analyses". *Scientometrics*, vol. 80, no. 1, 2009, pp. 105–132, doi: 10.1007/s11192-007-2019-y.

Small, H. G. "Co-citation in the scientific literature: a new measure of the relationship between two documents". *Journal of the American Society for Information Science*. vol. 24, no. 4, 1973, pp. 265-69, doi: 10.1002/asi.4630240406.

Spinak, E. *Diccionario enciclopédico de bibliometría, cientometría e informetría*. UNESCO, 1996.

White, H. D., and Griffith, B. C. "Author cocitation: a literature measure of intellectual structure". *Journal of the American Society for Information Science*. vol. 32, no. 3, 1981, pp. 163-171, doi: 10.1002/asi.4630320302.

White, H. D., and McCain, K. W. "Visualizing a discipline: an author co-citation analysis of Information Science, 1972–1995". *Journal of the American Society for Information Science*, vol. 49, no. 4, 1998, pp. 327–355, doi: doi.org/10.1002/(SICI)1097-4571(19980401)49:4327::AID-ASI43.0.CO;2-4.

Zhao, D., and Strotmann, A. "The knowledge base and research front of Information Science 2006–2010: an author cocitation and bibliographic coupling analysis". *Journal of the Association for Information Science and Technology*, vol. 65, no. 5, 2014, pp.995–1006, doi: 10.1002/asi.23027.

Zhao, H., et al. "Corporate social responsibility research in international business journals: an author co-citation analysis". *International Business Review*. vol. 27, no. 2, 2018, pp. 389-400, doi: 10.1016/j.ibusrev.2017.09.006.

Dados da pesquisa

Declaramos que os dois conjuntos de dados utilizados nesta pesquisa estão preservados e podem ser consultados a partir do contato com o primeiro autor do artigo, através de solicitação por e-mail. Os dados estão distribuídos em seis planilhas, a saber: dados gerais das referências dos dois conjuntos, incluindo informações dos documentos citantes; e quatro rankings de autores citados, dois para cada conjunto, considerando as posições das autorias (primeiros/todos os autores das referências). Os autores

Received: 2020/11/07 Accepted: 2021/04/20