em questão
E-ISSN 1808-5245

Cosseno de Salton, Índice de Jaccard e Correlação de Pearson: comparando índices normalizados e absolutos em análise de cocitação de autores

Ely Francina Tannuri de Oliveira
Doutora; Universidade Estadual Paulista "Júlio de Mesquita Filho", Marília, SP, Brasil;
etannuri@gmail.com

Bruno Henrique Alves
Doutorando; Universidade Estadual Paulista "Júlio de Mesquita Filho", Marília, SP, Brasil;
bruninkmkt@hotmail.com

Resumo: Esta pesquisa objetiva realizar um estudo comparativo entre os indicadores absolutos e os normalizados, para Análise de Cocitação de Autores, a saber: Cosseno de Salton (C_S), Índice de Jaccard (IJ) e Correlação de Pearson (r). , Visa apresentar os três índices normalizados, apontar as diferenças no uso entre os indicadores absolutos e normalizados, avaliar questões sobre a escolha dos três indicadores e apresentar um estudo teórico-aplicado. Ainda, calcular a Correlação de Spearman entre a matriz com os valores absolutos de cocitação e os três diferentes índices normalizados - Cosseno de Salton (C_S), Índice de Jaccard (IJ) e Correlação de Pearson (r). Como fonte de dados, utilizaram-se os artigos do periódico Scientometrics, pertencente à base de dados Scopus na temática Estudos Métricos. Recuperaram-se 234 artigos do período de 2013 e 2014, em junho de 2015. Identificaram-se 9.327 pesquisadores citados. Gerou-se a matriz absoluta dos autores mais cocitados e procederam-se às normalizações pelos três processos. No sentido de comparar os resultados da matriz absoluta com os respectivos índices normalizados Cosseno de Salton (C_S), Índice de Jaccard (IJ) e Correlação de Pearson (r), calculou-se a Correlação de Spearman, a partir do pareamento dos dados absolutos e dos normalizados de cada uma das matrizes, ordenados em postos, com a finalidade de sugerir o melhor indicador normalizado. Apresentaram-se os gráficos de dispersão e concluiu-se pelo uso de preferencial do Cosseno de Salton (Cs), a partir dos objetivos da pesquisa, da natureza dos dados e da maior significância relativa à Correlação de Spearman.

Palavras-chave: Indicadores normalizados de Cocitação. Cosseno de Salton. Índice de Jaccard. Correlação de Pearson.

1 Introdução

A pesquisa bibliométrica atual, segundo Glänzel (2003), é destinada a três grupos-alvo principais da Bibliometria contemporânea, a saber: Bibliometria para

profissionais da bibliometria; Bibliometria aplicada às disciplinas científicas; Bibliometria para a política científica e gestão.

O primeiro grupo é próprio da pesquisa bibliométrica "de base", que está preocupada com o seu desenvolvimento conceitual-teórico-metodológico. Contribuem para o avanço do conhecimento da área, propondo a criação de novos conceitos e indicadores, bem como reflexões e análises relativas à área. O segundo grupo, de natureza metodológica, se propõe a dar sustentação às áreas em que se aplicam os procedimentos da Bibliometria, constituindo a grande parte da própria Cientometria. O terceiro grupo se presta a subsidiar pesquisas destinadas à construção de indicadores voltados para a política científica e gestão.

Este estudo se situa no primeiro grupo, Bibliometria para profissionais da bibliometria, na medida em que objetiva contribuir, de forma analítica, para o desenvolvimento da temática "Análise de Cocitação de Autores". Constitui também um estudo aplicado, de natureza metodológica, na medida em que se propõe a dar sustentação às diferentes áreas da ciência em que se aplicam os procedimentos da Bibliometria, situando-se também no segundo grupo-alvo. Trata dos indicadores necessários e implícitos a essa metodologia, cujo uso é relevante, especialmente por abordar as diferentes subáreas da Cientometria, fortalecendo sua compreensão e seu uso.

Hjørland (2002) e Smiraglia (2011) destacam a contribuição dos estudos de citação e cocitação, dentre as metodologias bibliométricas, para a compreensão de um contexto e das relações que acontecem em uma comunidade científica. Nos estudos de citação e cocitação, destacam-se os estudos de Small (1973), White e Griffith (1981), entre outros.

A cocitação é definida por Small (1973, p. 265) como "[...] a frequência com que dois documentos citados estão juntos em um artigo.". O número de vezes que os autores ou documentos foram citados juntos determina a força da cocitação entre eles, bem como o núcleo da literatura da área. As citações indicam os paradigmas das comunidades formadas, seus procedimentos metodológicos, os grupos de cientistas, suas publicações, e evidenciam os pesquisadores de impacto de uma área. Associada à análise de citações, a Análise de Cocitações de Autores (ACA) foi aprofundada por White e Griffith, em 1981 (WHITE; GRIFFITH, 1981). A frequência de cocitação entre dois autores determina como a estrutura de

conhecimento da área é percebida pelos pesquisadores (GMÜR, 2003). Observa-se, ainda, que o conjunto de referências de uma área constitui a representação de sua estrutura intelectual.

Os estudos de coautoria e cocitação são, em geral, realizados a partir da construção de matrizes de valores absolutos. No entanto, quando se pretende realizar estudos comparativos, dadas as especificidades e peculiaridades de cada área, destaca-se a relevância dos indicadores normalizados. Estes possibilitam avaliações comparativas, uma vez que padronizam as unidades de medida, além de revelar aspectos não explicitados nos dados brutos obtidos nas matrizes com valores absolutos (GLÄNZEL et al., 2009), tais como a intensidade e a proximidade de relações entre os autores cocitados.

Segundo os estudos de Luukkonen et al. (1993), os índices absolutos e normalizados trazem tipos diferentes de informações: os absolutos mostram os autores centrais ou os mais periféricos nas redes, enquanto os índices normalizados mostram a intensidade das relações de colaboração ou de cocitação entre os pares de pesquisadores cocitados.

Apresentam-se vários procedimentos para a normalização dessas medidas, tais como: Cosseno de Salton (C_S), Índice de Jaccard (IJ) e Coeficiente de Correlação de Pearson, denotado por Correlação de Pearson (r), indicadores estes que são o foco desta pesquisa, utilizados aqui para análise de cocitações de autores. Há outros procedimentos matemáticos e estatísticos para normalização, não abordados no âmbito deste trabalho.

Considerando o exposto, objetiva-se nesta pesquisa realizar um estudo comparativo entre os indicadores absolutos e os normalizados, a saber: Cosseno de Salton (C_S), Índice de Jaccard (IJ) e Correlação de Pearson (r) para Análise de Cocitação de Autores. De forma mais específica, propõe-se apresentar os três índices normalizados, apontar as diferenças no uso entre os indicadores absolutos e os normalizados, relativos ao C_S, IJ e Correlação de Pearson (r), analisar algumas questões sobre a escolha desses diferentes procedimentos e apresentar um estudo teórico-aplicado desses indicadores, comparando resultados extremados.

Esta pesquisa se justifica, considerando as especificidades das diferentes áreas, relativas aos diferentes comportamentos no que se refere às citações. Ela se

propõe a analisar alguns procedimentos e técnicas para estudos relativos à normalização de cocitações, uma vez que esta padroniza as unidades de medida nas matrizes de cocitação de autores. Também pela necessidade de evidenciar algumas questões existentes no âmbito da comunidade científica relativas ao procedimento mais adequado a ser utilizado nas diferentes situações.

Ao comparar os três resultados obtidos pela Correlação de Postos, podem-se obter indicadores que contribuirão para destacar se algum dos índices normalizados deve ser utilizado de forma mais diferenciada ou se eles podem ou não ser usados indiscriminadamente.

2 Normalização dos indicadores absolutos e discussão

Os estudos de cocitação de autores utilizam matrizes de proximidade na medida em que deixam explícita a intensidade da proximidade ou o maior distanciamento entre os pares de autores cocitados, sob a ótica dos autores citantes. Nas matrizes de proximidade, quanto maior for o valor do indicador de cocitação, mais similares e próximos serão os dois autores cocitados, nos aspectos teóricos e/ou metodológico (LEYDESDORFF; VAUGHAN, 2006). Assim, todas as matrizes de frequência absoluta ou normalizada de cocitação são chamadas de similaridade.

O C_S é um índice normalizado, calculado a partir da matriz de coocorrência dos dados absolutos, tanto para a coautoria quanto para cocitações, conforme Luukkonen et al. (1993). É expresso pela frequência de coautorias ou cocitações de dois autores *i* e *j*. A fórmula do $C_{S,}$ presente nos estudos de Hamers et al. (1989), nos quais as coocorrências representam cocitações, pode ser assim expressa:

$$CS_{(i,j)} = \frac{coc_{(i,j)}}{(cit_{(i)}.\ cit_{(j)})^{1/2}}$$

Onde,

$coc_{(i,j)}$ = total de cocitações entre o autor i e j;

$cit_{(i)}$ = total de citações recebidas pelo autor i;

$cit_{(j)}$ = total de citações recebidas pelo autor j.

Small (1973) foi o primeiro a apontar o IJ como índice de normalização, utilizado somente para cocitações. Mede a intensidade da relação entre dois autores, revistas ou documentos, e é definido como a frequência de cocitação de dois autores A e B, chamados aqui de Cocit (A, B). É compreendido como o número de vezes que os autores A e B são citados conjuntamente, representados por Cocit (A, B), dividido pelo número total de citações recebidas pelos dois autores A e B, representadas conceitualmente por Cit (A) + Cit (B) – Cocit (A∩B).

Segundo Luukkonen et al. (1993), a fórmula, reelaborada, pode ser expressa por:

$$IJ = \frac{Cocit\ (A,B)}{Cit\ (A) + Cit\ (B) - Cocit(A \cap B)}$$

O Índice de Jaccard (IJ) é utilizado somente para cocitações e mede a intensidade da relação entre dois autores, revistas ou documentos; é definido como a frequência de cocitação de dois autores A e B, chamados aqui de Cocit (A, B) (VANZ, 2009).

Tanto o C_S quanto o IJ apresentam valores variando entre zero e um: quanto mais próximo de um estiver o valor desses indicadores, mais similares são os dois autores; quanto mais próximo de zero, mais distante (menos intensa) é a associação dos dois autores, seja como coautores ou na percepção do conjunto de autores citantes. Valor zero para C_S ou IJ indica ausência de coautorias ou cocitações entre dois autores.

Destaque-se que, tanto para C_S ou IJ, o valor zero indica ausência de coautorias ou cocitações entre os dois autores. Caso ocorra, para esses indicadores, o valor um, significa que os dois autores, no caso de cocitação, foram citados exatamente nos mesmos trabalhos; no caso de coautoria, foram coautores em todos os trabalhos produzidos.

A correlação mede a associação de pares de autores cocitados, considerando o conjunto de valores de cocitações de cada um deles com os demais em estudo. O r de Pearson foi a medida padrão na ACA ($-1 \leq r \leq +1$) e, nesse

intervalo, a correlação varia de sentido (de negativo para positivo) e de força (fraca, moderada ou forte).

Considerando x_i e y_i as frequências de cocitações de um autor X e de um autor Y, com os demais autores, para i variando de 1 a n, com n igual à quantidade de autores em estudo, o r de Pearson é definido por:

$$r = \frac{\sum x_i . y_i - \frac{\sum x_i . \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right) . \left(\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right)}}$$

A escolha do uso do índice de similaridade para a normalização C_S ou IJ ou r de Pearson, no caso da ACA, é controversa. Em 2003, Ahlgren et al. criticaram o uso de *r* de Pearson, mostrando que o mesmo não satisfaz isoladamente como medida de similaridade. Os autores citados fornecem argumentos para o uso do C_S, em vez de r de Pearson, especialmente quando se pretende analisar a visualização da estrutura das redes sociais ou Multi-Dimensional Scaling (MDS). Porém, Bensman (2004) e White (2003) consideram que, ao se propor uma análise estatística (multivariada, por exemplo), deve-se usar a correlação de Pearson.

Quando o foco da análise for a frequência de cocitação entre dois autores, independentemente do total do volume de cocitações, seja frequência absoluta ou normalizada pelo C_S ou IJ, tem-se uma medida local, também chamada bilateral de similaridade. Quando a análise de cocitação entre dois autores estiver relacionada com todos os outros autores do conjunto de dados analisados, tem-se uma medida global ou multilateral, como é o caso de *r* de Pearson. Ahlgren et al. (2003) apontam que o uso de um ou outro método (local ou global) leva a resultados totalmente diferentes.

O C_S é sempre maior que o IJ. O resultado do índice para o "[...] Cosseno de Salton é, em geral, o dobro do valor para o Índice de Jaccard." (HAMERS, 1989, p. 315), mas tem-se com o IJ uma medida tão boa quanto C_S. Egghe e Leydesdorff (2009) preferem utilizar o C_S, considerando que as diferenças entre os resultados são mínimas.

Não há consenso a respeito do uso de um ou outro indicador, mesmo porque eles estão em função das possíveis amostras nas quais se trabalha, sejam valores com menor dispersão, portanto mais concentrados, sejam valores extremados.

Na área da Ciência da Informação, estudos aplicados de análise de cocitação têm tido visibilidade, oriundos das matrizes de cocitação e de índice de cocitação relativa. Estes têm dado origem a uma significativa literatura na área, especialmente em âmbito internacional. Podem-se citar os estudos de White e McCain (1998), Moya Anegón, Jiménez Contreras e Moneda Corrochano (1998), Liberatore, Herrero-Solana e Guimarães (2007) e Pinheiro e Silva (2008) com o objetivo de evidenciar os pesquisadores mais destacados e/ou citados nas áreas estudadas e apresentar indicadores que contribuam para o delineamento do panorama da atividade científica.

Acrescente-se que, nos "Encontros Brasileiros de Bibliometria e Cientometria", eventos específicos da área de Estudos Métricos, alguns estudos dessa natureza têm sido apresentados, objetivando contribuir para a visibilidade da temática. Destaca-se, ainda, a temática presente nos eventos de âmbito internacional, tais como Indicadores de Ciência e Tecnologia (STI) e Sociedade Internacional para Cientometria e Informetria (ISSI), entre outros.

Contudo, a literatura não apresenta um posicionamento conclusivo sobre a questão apontada, e, por consequência, o estabelecimento da metodologia adequada para a ACA não se encontra plenamente consolidado, demandando muitos estudos a fim de se ratificar a necessidade de conversão da frequência de cocitação em índices de cocitação relativa (VANZ, 2009).

3 Metodologia

Conceituaram-se os três índices abordados, com questões sobre seu uso. Utilizaram-se os artigos publicados no periódico Scientometrics, sobre a temática "Estudos Métricos", na base de dados *SCOPUS*, com os termos de busca "*Bibliometr**" OR "*Cientometr**", recuperando-se 234 artigos publicados no período de 2013 e 2014, em junho de 2015. Identificaram-se 9.327 pesquisadores citados, em um total de 8.610 referências.

Construiu-se uma tabela em ordem decrescente de citações, desconsiderando-se as autocitações. O corte na tabela foi feito para os pesquisadores citados em pelo menos 28 artigos, totalizando o grupo-alvo de 21 pesquisadores. Construiu-se a matriz de dados absolutos (21x21). Para o cálculo do C_S e IJ, utilizou-se o Microsoft Excel, e para o cálculo de *r* de Pearson, o software *SPSS*. Destacaram-se os índices que apresentaram valores notáveis, apontando algumas contribuições desses três indicadores para a visualização das proximidades entre os autores cocitados, tendo como referência a matriz de valores absolutos.

No sentido de comparar melhor os resultados das matrizes com os índices normalizados — C_S, IJ e r de Pearson — com os valores da matriz absoluta, calculou-se a Correlação de Postos (Correlação de Spearman) por meio do *Excel* na opção suplementos, no recurso *Action*, a partir dos dados de cada uma das matrizes normalizadas ordenadas em postos, calculadas a Correlação com a matriz de valor absoluto, com os presentes dados, a fim de sugerir o melhor indicador normalizado.

Representaram-se as relações pareadas da matriz de valor absoluto com os resultados normalizados das outras três matrizes, construindo-se o(s) gráfico(s) de dispersão, por meio do *software Excel*. Objetivou-se, assim, a visualização das relações absolutas e normalizadas pelos três procedimentos utilizados, a fim de verificar qual delas é mais significativa, mostrando uma localização de pares ordenados mais próximos à organização linear.

4 Apresentação e análise dos dados

Considerando que os pesquisadores mais cocitados advêm de várias áreas, com diferentes padrões de citações, foi necessária a normalização dos mesmos. A partir dos maiores produtores existentes na base de dados *Scopus*, destacam-se as seguintes grandes áreas e subáreas do conhecimento, a partir da classificação da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (2012): *Ciências Biológicas* — Bioquímica, Genética, Biologia Celular, Agricultura e Ciências Biológicas; Ciências da Saúde — Medicina, Enfermagem, entre outras; Engenharias — Engenharias, entre outras; *Ciências Exatas e da Terra* — Ciências

da Computação, Matemática, entre outras; *Ciências Humanas* — Psicologia, Ciências Sociais, entre outras grandes áreas e subáreas.

Os procedimentos de normalização padronizam as unidades de medida nas matrizes. Apresentam-se a Figura 1, com a matriz absoluta, e as Figuras 2, 3 e 4, com as matrizes normalizadas pelo C_S, IJ e *r* de Pearson.

Figura 1- Matriz de valores absolutos.

	Leydesdorff, L.	van Raan, A.	Moed, H.	Schubert, A.	Garfield, E.	Bornmann, L.	Hirsch, J.	Egghe, L.	Rousseau, R.	Waltman, L.	Van Eck, N.	Braun, T.	Ho, Y.	Van Leeuwen, T.N	Moya -Anegón, F.,	Bordons, M.	Martin, B.	Persson, O.	Ding, Y.	Wagner, C.	Aksnes, D.,	Total de citações
Leydesdorff, L.	77	25	21	25	22	24	9	10	13	17	16	10	8	15	14	8	15	14	13	18	11	77
Van Raan, A.		70	34	23	26	24	18	17	12	21	20	12	5	20	12	13	14	6	10	10	12	70
Moed, H.			65	21	29	19	18	16	12	13	13	11	2	18	15	11	10	9	7	5	15	65
Schubert, A.				67	25	23	21	23	18	16	15	35	8	12	11	9	10	16	13	12	9	67
Garfield, E.					65	22	19	19	16	10	8	14	6	8	13	14	7	7	7	7	13	65
Bornmann, L.						50	18	15	13	19	16	12	4	13	11	13	5	5	4	7	9	50
Hirsch, J.							50	35	21	11	10	14	4	9	6	9	7	4	8	2	8	50
Egghe, L.								47	27	8	9	13	3	7	9	10	5	4	11	5	7	47
Rousseau, R.									42	9	9	6	2	4	8	5	3	8	9	8	5	42
Waltman, L.										36	32	8	5	14	9	8	6	3	5	7	3	36
Van Eck, N.											33	8	4	13	9	7	5	3	5	6	3	33
Braun, T.												35	7	6	5	5	4	6	2	3	5	35
Ho, Y.													27	1	3	1	1	5	7	3	3	27
V. Leeuwen, T.														28	4	6	5	2	2	5	7	28
MoyaAnegón,F															29	4	3	4	5	4	4	29
Bordons, M.																30	6	3	2	6	7	30
Martin, B.																	31	6	5	8	4	31
Persson, O.																		31	9	7	5	31
Ding, Y.																			29	5	4	29
Wagner, C.																				29	5	29
Aksnes, D.,																					28	28
Total de citações	**77**	**70**	**65**	**67**	**65**	**50**	**50**	**47**	**42**	**36**	**33**	**35**	**27**	**28**	**29**	**30**	**31**	**31**	**29**	**29**	**28**	

Fonte: Elaborado pelos autores.

A matriz da Figura 1 mostra os pesquisadores mais cocitados em sua primeira linha e coluna, e, na diagonal principal, os valores de citação de cada pesquisador, representados tanto pela linha quanto pela coluna, respectivamente. Destaca-se que a última linha e coluna são os totais marginais das citações de cada autor. Cada célula no interior da matriz representa o valor de Cocit (A, B), significando o valor de cocitação entre os autores (A, B). A seguir apresentam-se as três matrizes normalizadas por C_S, IJ e r de Pearson.

Figura 2 - Matriz normalizada pelo Cosseno de Salton (Cs).

	Leydesdorff, L.	Van Raan, A.	Moed, H.	Schubert, A.	Garfield, E.	Bornmann, L.	Hirsch, J.	Egghe, L.	Rousseau, R.	Waltman, L.	Van Eck, N.	Braun, T.	Ho, Y.	Van Leeuwen, T.N	Moya -Anegón, F.,	Bordons, M.	Martin, B.	Persson, O.	Ding, Y.	Wagner, C.	Aksnes, D.,
Leydesdorff, L.	1,00	0,34	0,30	0,35	0,31	0,39	0,15	0,17	0,23	0,32	0,32	0,19	0,18	0,32	0,30	0,17	0,31	0,29	0,28	0,38	0,24
Van Raan, A.		1,00	0,50	0,34	0,39	0,41	0,30	0,30	0,22	0,42	0,42	0,24	0,12	0,45	0,27	0,28	0,30	0,13	0,22	0,22	0,27
Moed, H.			1,00	0,32	0,45	0,33	0,32	0,29	0,23	0,27	0,28	0,23	0,05	0,42	0,35	0,25	0,22	0,20	0,16	0,12	0,35
Schubert, A.				1,00	0,38	0,40	0,36	0,41	0,34	0,33	0,32	0,72	0,19	0,28	0,25	0,20	0,22	0,35	0,29	0,27	0,21
Garfield, E.					1,00	0,39	0,33	0,34	0,31	0,21	0,17	0,29	0,14	0,19	0,30	0,32	0,16	0,16	0,16	0,16	0,30
Bornmann, L.						1,00	0,36	0,31	0,28	0,45	0,39	0,29	0,11	0,35	0,29	0,34	0,13	0,13	0,11	0,18	0,24
Hirsch, J.							1,00	0,72	0,46	0,26	0,25	0,33	0,11	0,24	0,16	0,23	0,18	0,10	0,21	0,05	0,21
Egghe, L.								1,00	0,61	0,19	0,23	0,32	0,08	0,19	0,24	0,27	0,13	0,10	0,30	0,14	0,19
Rousseau, R.									1,00	0,23	0,24	0,16	0,06	0,12	0,23	0,14	0,08	0,22	0,26	0,23	0,15
Waltman, L.										1,00	0,93	0,23	0,16	0,44	0,28	0,24	0,18	0,09	0,15	0,22	0,09
Van Eck, N.											1,00	0,24	0,13	0,43	0,29	0,22	0,16	0,09	0,16	0,19	0,10
Braun, T.												1,00	0,23	0,19	0,16	0,15	0,12	0,18	0,06	0,09	0,16
Ho, Y.													1,00	0,04	0,11	0,04	0,03	0,17	0,25	0,11	0,11
Van Leeuwen, T.N														1,00	0,14	0,21	0,17	0,07	0,07	0,18	0,25
Moya -Anegón, F.,															1,00	0,14	0,10	0,13	0,17	0,14	0,14
Bordons, M.																1,00	0,20	0,10	0,07	0,20	0,24
Martin, B.																	1,00	0,19	0,17	0,27	0,14
Persson, O.																		1,00	0,30	0,23	0,17
Ding, Y.																			1,00	0,17	0,14
Wagner, C.																				1,00	0,18
Aksnes, D.,																					1,00

Fonte: Elaborado pelos autores.

A Figura 2 apresenta a matriz normalizada de cocitação pelo C_S, com destaque para valores mais próximos ao valor 1. Observam-se valores significativos de índices de C_S, variando entre Van Raan e Moed, igual a 0,50, e entre Waltman e Van Eck, igual a 0,93, em amarelo, advindos de proximidade temática identificada pelos citantes, que explicitam a concepção de cocitação ou podem indicar coautoria entre os pares de autores cocitados.

A matriz normalizada pelo IJ se apresenta na Figura 3.

Figura 3 - Matriz normalizada pelo Índice Jaccard.

	Leydesdorff, L.	van Raan, A.	Moed, H.	Schubert, A.	Garfield, E.	Bornmann, L.	Hirsch, J.	Egghe, L.	Rousseau, R.	Waltman, L.	Van Eck, N.	Braun, T.	Ho, Y.	Van Leeuwen, T.N	Moya -Anegón, F.,	Bordons, M.	Martin, B.	Persson, O.	Ding, Y.	Wagner, C.	Aksnes, D.,	Total de citações
Leydesdorff, L.	1,00	0,20	0,17	0,21	0,18	0,23	0,08	0,09	0,12	0,18	0,17	0,10	0,08	0,17	0,15	0,08	0,16	0,15	0,14	0,20	0,12	**77**
van Raan, A.		1,00	0,34	0,20	0,24	0,25	0,18	0,17	0,12	0,25	0,24	0,13	0,05	0,26	0,14	0,15	0,16	0,06	0,11	0,11	0,14	**70**
Moed, H.			1,00	0,19	0,29	0,20	0,19	0,17	0,13	0,15	0,15	0,12	0,02	0,24	0,19	0,13	0,12	0,10	0,08	0,06	0,19	**65**
Schubert, A.				1,00	0,23	0,24	0,22	0,25	0,20	0,18	0,18	0,52	0,09	0,14	0,13	0,10	0,11	0,20	0,16	0,14	0,10	**67**
Garfield, E.					1,00	0,24	0,20	0,20	0,18	0,11	0,09	0,16	0,07	0,09	0,16	0,17	0,08	0,08	0,08	0,08	0,16	**65**
Bornmann, L.						1,00	0,22	0,18	0,16	0,28	0,24	0,16	0,05	0,20	0,16	0,19	0,07	0,07	0,05	0,10	0,13	**50**
Hirsch, J.							1,00	0,56	0,30	0,15	0,14	0,20	0,05	0,13	0,08	0,13	0,09	0,05	0,11	0,03	0,11	**50**
Egghe, L.								1,00	0,44	0,11	0,13	0,19	0,04	0,10	0,13	0,15	0,07	0,05	0,17	0,07	0,10	**47**
Rousseau, R.									1,00	0,13	0,14	0,08	0,03	0,06	0,13	0,07	0,04	0,12	0,15	0,13	0,08	**42**
Waltman, L.										1.00	0.86	0.13	0.09	0.28	0.16	0.14	0.10	0.05	0.08	0.12	0.05	**36**
Van Eck, N.											1.00	0.13	0.07	0.27	0.17	0.13	0.08	0.05	0.09	0.11	0.05	**33**
Braun, T.												1.00	0.13	0.11	0.08	0.08	0.06	0.10	0.03	0.05	0.09	**35**
Ho, Y.													1.00	0.02	0.06	0.02	0.02	0.09	0.14	0.06	0.06	**27**
Van Leeuwen, T.N														1.00	0.08	0.12	0.09	0.04	0.04	0.10	0.14	**28**
Moya -Anegón, F.,															1.00	0.07	0.05	0.07	0.09	0.07	0.08	**29**
Bordons, M.																1.00	0.11	0.05	0.04	0.11	0.14	**30**
Martin, B.																	1.00	0.11	0.09	0.15	0.07	**31**
Persson, O.																		1.00	0.18	0.13	0.09	**31**
Ding, Y.																			1.00	0.09	0.08	**29**
Wagner, C.																				1.00	0.10	**29**
Aksnes, D.,																					1.00	**28**

Fonte: Elaborado pelos autores.

Na Figura 3, os valores maiores, próximos a 1, em amarelo, representam as cocitações normalizadas entre Hirsch e Egghe, igual a 0,56, e entre Waltman e Van Eck, com valor máximo igual 0,86, maiores valores extremados. Por outro lado, valores normalizados próximos a zero, tal como Moed e Ho e, ainda, Ho e Van Leeuwen iguais 0,02 são os menores valores extremados próximos a zero.

Figura 4 - Matriz normalizada pela Correlação Linear de r Pearson.

	Leydesdorff, L.	Van Raan, A.	Moed, H.	Schubert, A.	Garfield, E.	Bornmann, L.	Hirsch, J.	Egghe, L.	Rousseau, R.	Waltman, L.	Van Eck, N.	Braun, T.	Ho, Y.	VanLeeuwen, T.	MoyaAnegón,F	Bordons, M.	Martin, B.	Persson, O.	Ding, Y.	Wagner, C.	Aksnes, D.,
Leydesdorff, L.	1	0,68	0,64	0,29	0,53	0,61	0,27	0,18	0,24	0,53	0,51	0,51	0,32	0,68	0,73	0,68	0,69	0,48	0,31	0,72	0,59
Van Raan, A.		1	0,86	0,45	0,75	0,85	0,48	0,37	0,46	0,61	0,60	0,50	0,12	0,91	0,87	0,83	0,62	0,3	0,21	0,36	0,79
Moed, H.			1	0,50	0,82	0,84	0,51	0,47	0,49	0,50	0,65	0,52	0,28	0,75	0,77	0,84	0,63	0,19	0,36	0,43	0,87
Schubert, A.				1	0,61	0,60	0,59	0,49	0,54	0,28	0,30	0,84	0,60	0,44	0,60	0,53	0,39	0,34	0,30	0,20	0,51
Garfield, E.					1	0,80	0,67	0,62	0,62	0,38	0,42	0,64	0,22	0,74	0,83	0,80	0,63	0,47	0,50	0,38	0,88
Bornmann, L.						1	0,56	0,54	0,58	0,65	0,69	0,65	0,37	0,84	0,86	0,83	0.71	0.37	0,48	0,52	0,70
Hirsch, J.							1	0,93	0,87	0,23	0,27	0,57	0,08	0,36	0,58	0,62	0,14	0,15	0,43	0,10	0,58
Egghe, L.								1	0,88	0,22	0,18	0,55	0,17	0,27	0,41	0,46	0,20	0,26	0,39	0,01	0,43
Rousseau, R.									1	0,22	0,24	0,62	0,31	0,34	0,57	0,60	0,35	0,26	0,64	0,16	0,46
Waltman, L.										1	0,99	0,34	0,16	0,76	0,58	0,48	0,40	0,05	0,14	0,34	0,27
Van Eck, N.											1	0,32	0,23	0,76	0,57	0,48	0,44	0,04	0,16	0,37	0,23
Braun, T.												1	0,50	0,43	0,53	0,50	0,39	0,56	0,60	0,35	0,46
Ho, Y.													1	0,19	0,38	0,06	0,41	0,62	0,52	0,44	0,14
VanLeeuwen, T.														1	0,81	0,75	0,74	0,21	0,27	0,45	0,65
MoyaAnegón,F.															1	0,76	0,66	0,47	0,53	0,54	0,78
Bordons, M.																1	0,54	0,08	0,21	0,27	0,78
Martin, B.																	1	0,57	0,55	0,75	0,70
Persson, O.																		1	0,71	0,70	0,40
Ding, Y.																			1	0,62	0,36
Wagner, C.																				1	0,36
Aksnes, D.,																					1

Fonte: Elaborado pelos autores.

Comparando-se os cinco maiores valores absolutos com suas respectivas normalizações, C_S, IJ e r de Pearson, em cor amarela, destacam-se as cocitações entre: Van Raan e Moed; Schubert e Braun; Hirsch e Egghe; Egghe e Rousseau; e Waltman e Van Eck. Tomaram-se os quatro menores valores de cocitação entre os autores, em verde, na matriz absoluta e suas respectivas normalizações, a saber: Moed e Ho; Ho e Van Leeuwen; Ho e Bordons; e Ho e Martin.

Analisando-se e comparando-se os valores de C_S e IJ nos cinco pares de cocitações destacados, observa-se que os valores de C_S são maiores que IJ (HAMERS, 1989).

Nas normalizações, analisando-se e comparando-se as Figuras 1, 2, 3 e 4índices absolutos, C_S, IJ e r de Pearson —, os maiores valores são para Waltman e Van Eck. Embora não tenham o maior valor absoluto para cocitação, eles possuem os maiores valores normalizados, pois diferenciam a intensidade de proximidade, não discriminada pelas frequências absolutas (valor 32), com o maior valor igual a 0,99 para *r* de Pearson, significando correlação quase perfeita entre os dois autores. Esse resultado advém de 32 cocitações absolutas relativizadas,

para 36 citações do Waltman e 33 citações do Van Eck. As demais correlações seguem igual raciocínio, mas com menor intensidade de cocitação.

Por outro lado, compararam-se os quatro menores valores absolutos com suas respectivas normalizações, C_S, IJ e r de Pearson (em cor verde), tendo Ho e Van Leeuwen; Ho e Bordons; e Ho e Martin apenas uma cocitação.

Para evidenciar algumas questões existentes no âmbito da comunidade científica relativas ao procedimento mais adequado a ser utilizado nas diferentes normalizações, destaque-se que C_S e IJ estão próximos para os nove valores destacados nas matrizes, porém distantes dos valores de *r* de Pearson, explicado pelo fato de serem índices de correlação local e o r de Pearson ser um índice global, influenciado por todo o conjunto (AHLGREN et al., 2003).

No sentido de aprofundar as análises e obter conclusões mais consistentes, calculou-se, a partir dos dados da matriz de valores absolutos pareada com seus valores normalizados — C_S, IJ e r de Pearson —, a Correlação de Postos (Correlação de Spearman), a fim de sugerir o indicador normalizado com maior significância a ser utilizado.

Assim, por meio do recurso Action do Excel, utilizaram-se a Estatística básica, a Matriz de Correlação e o Spearman Rank-order para os valores absolutos e as três matrizes normalizadas, obtendo-se os seguintes valores de Correlação de Postos: para os valores da matriz absoluta e os índices de C_S, obteve-se o valor 0,96; para a matriz absoluta e IJ, obteve-se o valor 0,94; e para a matriz absoluta e *r* de Pearson, obteve-se o valor 0,48.

Os valores altos de correlação mais próximos a 1 (inteiro) são as relações estatisticamente significantes, indicando que, a partir dos dados desta pesquisa, o mais conveniente é a utilização dos índices normalizados de C_S, por ter se apresentado com a maior correlação 0,96, mostrando, assim, forte grau de associação com os valores absolutos.

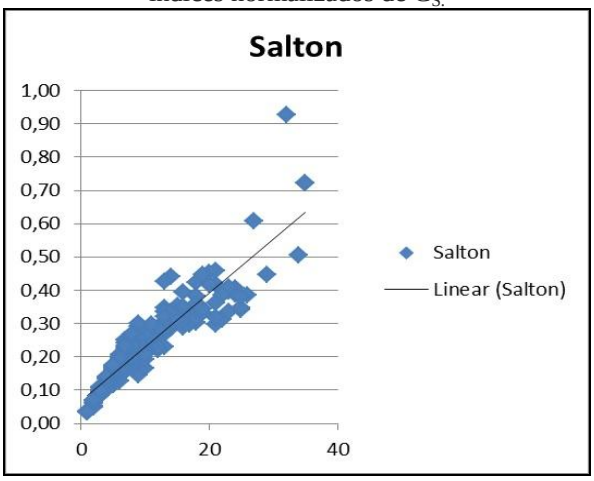
Ainda, a fim de ratificar o procedimento mais adequado a ser utilizado nas diferentes normalizações, construíram-se três diagramas de dispersão entre os valores da matriz absoluta no eixo das abscissas e os valores normalizados, para Cs, IJ e *r* de Pearson, no eixo das ordenadas.

A simples inspeção dos três diagramas de dispersão demonstra, pelo Gráfico 1, a pertinência do uso da normalização pelo C_S, pelo fato de a maioria dos pontos se aglutinarem com maior proximidade no entorno da linha reta. Por outro lado, a análise do Gráfico 3 mostra a dispersão dos valores no entorno da reta, ou seja, a escolha do Coeficiente Linear de Pearson para a normalização da matriz de valores absolutos seria a *menos* desejável, considerando os dados aqui apresentados.

Apresentam-se, a seguir, os respectivos cálculos de Correlação de Spearman e os gráficos de dispersão correspondentes.

Correlação: Spearman		
	Absoluta	Salton
Absoluta	1	0,964262324
Salton	0,964262324	1

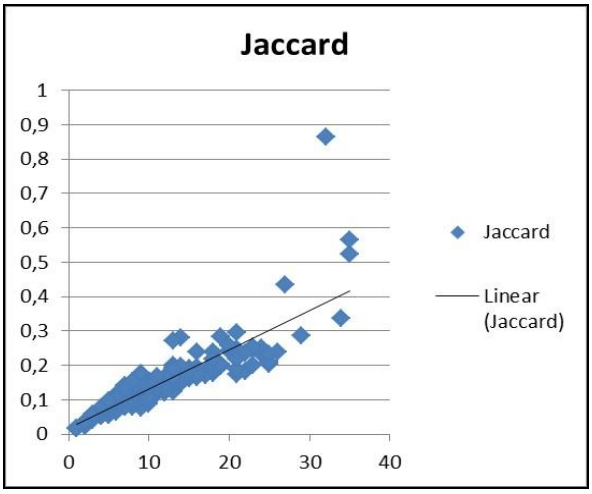
Gráfico 1 – Diagrama de dispersão entre matriz de valores absolutos e índices normalizados de C_S.



Fonte: Dados gerados pelos autores por meio do *Excel*.

Correlação: Spearman		
	Absoluta	Jaccard
Absoluta	1	0,942426436
Jaccard	0,942426436	1

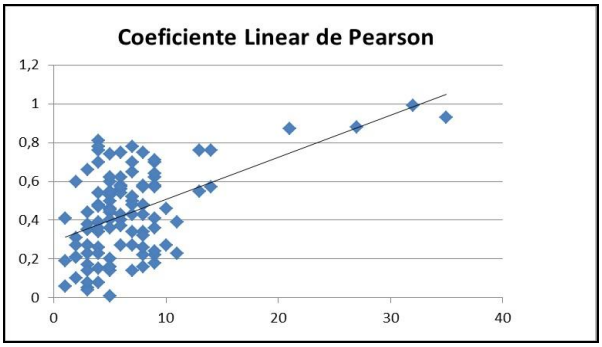
Gráfico 2 – Diagrama de dispersão entre matriz de valores absolutos e índices normalizados por IJ.



Fonte: Dados gerados pelos autores por meio do *Excel*.

Correlação: Spearman		
	Absoluta	Pearson
Absoluta	1	0,480213945
Pearson	0,480213945	1

Gráfico 3 – Diagrama de dispersão entre matriz de valores absolutos e índices normalizados pelo *r* de Pearson.



Fonte: Dados gerados pelos autores por meio do *Excel*.

Considerando as discussões existentes entre a utilização de um ou outro procedimento de normalização de matrizes, deve acontecer sempre uma análise cuidadosa na indicação do índice a ser utilizado de forma mais discriminativa. Pode ocorrer que o recorte da amostra apresente resultados dispersos, para os quais os índices indicam comportamentos bastante distintos, que deverão ser analisados.

5 Considerações finais

Apesar de não existir uma indicação unívoca para o uso dos índices de normalização, as pesquisas na temática sugerem que a escolha do procedimento está articulada com os objetivos da pesquisa e a natureza dos dados, para a qual se propõe a normalização. Assim, quando se pretende visualizar uma rede de cocitações, tanto o C_S como o IJ são índices bem aceitos. Quando o objetivo do estudo se encaminha para uma análise estatística mais avançada, a exemplo da multivariada, o r de Pearson é mais aconselhado. Observe-se que as informações utilizadas para C_S e IJ diferem daquelas usadas para o cálculo do *r* de Pearson.

Neste estudo, após análises e cálculos por meio da Correlação de Spearman, definiu-se pelo uso do C_S, como o melhor procedimento normalizado, considerado o mais adequado. Se a opção, sem análises prévias fosse IJ, também seria uma normalização aceita, porém, menos adequada que C_S.

Propõe-se a continuidade e o aprofundamento de estudos nas questões relativas às normalizações de índices, em amostras que diferem em natureza dos presentes dados, tanto relativos às análises de cocitações ou coautorias.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

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**, North Carolina, v. 54, n. 6, p. 550-560, 2003.

BENSMAN, S. J. Person's r and author cocitation analysis: a commentary on the controversy. **Journal of the American Society of Information Science & Technology**, North Carolina, v. 55, n. 10, p. 935-936, 2004.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR. Tabelas de áreas do conhecimento. [Brasília, 2012.] Disponível em:<http://www.capes.gov.br/images/stories/download/avaliacao/TabelaAreasConhecimento_072012.pdf>. Acesso em: 30 ago. 2016.

EGGHE, L.; LEYDESDORFF, L. The relation between Pearson's correlation coefficient r and Salton's cosine measure. **Journal of the American Society for Information Science and Techology**, North Carolina, v. 60, n. 5, p. 1027-1036, 2009.

GLÄNZEL, W. **Bibliometrics as a research field**: a course on theory and application of bibliometric indicators. Bélgica: [s.n.], 2003. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.5311&rep=rep1&type=pdf>. Last Accessed: 30 August. 2016.

GLÄNZEL, W. et al. Subfield-specific normalized relative indicators and a new generation of relational charts: methodological foundations illustrated on the assessment of institutional research performance. **Scientometrics**, Dordrecht, v. 78, n. 1, p. 165-188, 2009.

GMÜR, M. Co-citation analysis and the search for invisible colleges: a methodological evaluation. **Scientometrics**, Dordrecht, v. 57, n. 1, p. 27-57, 2003.

HAMERS, L. et al. Similarity measures in scientometric research: the Jaccard Index versus Salton´s Cosine formula. **Information Processing & Management**, Elmsford, v. 25, n. 3, p. 315-318, 1989.

HJØRLAND, B. Domain analysis in information science: eleven approaches traditional as well as innovative. **Journal of Documentation**, London, v. 58, n. 4, p. 422-462, 2002.

LEYDESDORFF, L.; VAUGHAN, L. Co-occurrence Matrices and their applications in Information Science: Extending ACA to the Web environment. **Journal of the American Society for Information Science and Technology**, North Carolina, v. 57, n. 12, p. 1616-1628, 2006.

LIBERATORE, G.; HERRERO-SOLANA, V.; GUIMARAES, J. A. C. Análise bibliometrica do periódico brasileiro Ciência da Informação durante o período 2000-2004. **Brazilian Journal of Information Science**, Marília, v. 1, n. 2, p. 3-21, 2007.

LUUKKONEN, T. et al. The measurement of international scientific collaboration. **Scientometrics**, Dordrecht, v. 28, n. 1, p. 15-36, 1993.

MOYA ANEGÓN, F.; JIMÉNEZ CONTRERAS, E.; MONEDA CORROCHANO, M. Research fronts in library and information science in Spain (1985-1994). **Scientometrics**, Dordrecht, v. 42, n. 2, p. 229-246, 1998.

PINHEIRO, L. V. R.; SILVA, G. S. Cartografia histórica e conceitual da bibliometria/informetria no Brasil. In: CONFERÊNCIA IBERO-AMERICANA DE PUBLICAÇÕES ELETRÔNICAS NO CONTEXTO DA COMUNICAÇÃO CIENTÍFICA (CIPECC), 2., 2008, Rio de Janeiro. Sub-Tema 2–Metrias da comunicação científica: da bibliometria/informetria à Webmetria. **Anais eletrônicos**...Rio de Janeiro: [s.n.], 2008. Disponível em: <http://ridi.ibict.br/handle/123456789/67>.Acesso em: 30 ago. 2016.

SMALL, H. Co-citation in the scientific literature: a new measure of the relationship between two documents. **Journal of the American Society for Information Science**, North Carolina, v. 24, n. 4, p. 265-269, 1973.

SMIRAGLIA, R. P. Domain coherence within Knowledge Organization: people, interacting theoretically, across geopolitical and cultural boundaries. In: ANNUAL CAIS/ACSI CONFERENCE, 39, June 2-4, 2011, Canada. **Proceedings...** Canada: University of New Brunswick, 2011. p. 1-6. Disponível em:<https://journals.library.ualberta.ca/ojs.cais-acsi.ca/index.php/cais-asci/article/view/601/551>. Acesso em: 30 ago. 2016.

VANZ, S. A. S. **As redes de colaboração no Brasil (2004-2006)**. 2009. 204 f. Tese (Doutorado em Comunicação e Informação) – Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

WHITE, H. D. Author cocitacion analysis and Pearson's *r*. **Journal of the American Society of Information Science and Technology**, North Carolina, v.54, n.13, p. 1250-1259, 2003.

WHITE, H. D.; GRIFFITH, B. Author co-citation: a literature measure of intellectual structure. **Journal of the American Society for Information Science and Technology**, North Carolina, v. 32, n. 2, p. 163-171, 1981.

WHITE, H.D.; MCCAIN, K.W. Visualizing a discipline: an author co-citation analysis of Information Science, 1972-1995. **Journal of the American Society for Information Science and Technology,** North Carolina, v. 49, n. 4, p. 327-355, 1998.

Salton's Cosine, Jaccard Index and Pearson's Correlation: comparing normalized and absolute indexed in author co-citation analysis

Abstract: This research aims to conduct a comparative study between the absolute and normalized indicators for Author Co-citation Analysis, namely:

Salton's Cosine (Sc), Jaccard Index (JI) and Pearson's correlation (*r*). As data source, we used the articles from the journal Scientometrics on Scopus database in the subject Metric Studies. We retrieved 234 articles in the 2013-2014 period, in June 2015. We identified 9,327 cited researchers. We generated the absolute matrix with the most co-cited authors, and proceeded to normalizations, through the three processes. In order to compare the results of the absolute matrix with their respective normalized indexes — Salton's Cosine (Sc), Jaccard Index (JI) and Pearson's correlation (*r*) — we calculated Spearman's correlation. We present the scatter plots and concluded that Salton's Cosine is preferred, from the research objectives, the nature of the data and greater significance in relation to Spearman's correlation.

Keywords: Normalized Co-citation indicators. Salton's Cosine. Jaccard index. Pearson's correlation.

Recebido em: 06/09/2016

Aceito em: 21/10/2016