

INDICADORES DE PROXIMIDADES EM ANÁLISE DE COCITAÇÃO DE AUTORES: um estudo comparativo entre coeficiente de Correlação de Pearson e Cosseno de Salton

Maria Cláudia Cabrini Grácio*
Ely Francina Tannuri de Oliveira**

RESUMO

Este trabalho propõe-se a apresentar e a analisar os indicadores de proximidade de Coeficiente de Correlação de Pearson (r) e Cosseno de Salton (CS), para a análise de cocitação de autores, e comparar a contribuição destes dois indicadores para a visualização das proximidades entre os autores cocitados, avaliando a correlação entre os indicadores e identificando os agrupamentos dos pesquisadores cocitados formados a partir dos dois índices. Para isso, utilizam-se como universo os artigos publicados sobre a temática “Estudos Métricos”, na base de dados BRAPCI. A partir de pesquisa anterior, recuperaram-se as matrizes com as frequências absolutas de cocitação e normalizadas por Cosseno de Salton entre os 38 autores mais citados. A seguir, foi construída a matriz com os valores de Correlação de Pearson entre os autores cocitados, utilizando-se o *software* SPSS; analisaram-se os índices em destaque para o Cosseno de Salton e a Correlação de Pearson; e calculou-se a correlação entre os 703 pares dos dois índices, representada por meio de um gráfico de dispersão de valores. A partir de cada matriz— Cosseno de Salton e Correlação de Pearson—, realizaram-se duas análises de *clusters*, gerando-se dois dendogramas, que se prestaram à análise comparativa dos agrupamentos formados. Como conclusões, destaca-se que coautores apresentam altos índices de CS e de r . Ainda, observou-se uma tendência de baixos valores para o CS por estarem pareados com baixos valores de r , do que se conclui que, embora CS meça a proximidade local e r meça a proximidade global, esses índices apresentaram as mesmas tendências de proximidade.

Palavras-chave: Análise de cocitação de autores. Indicadores normalizados de proximidade. Coeficiente de Correlação de Pearson. Cosseno de Salton.

* Doutora em Lógica pela Universidade Estadual de Campinas, Brasil. Docente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Júlio de Mesquita Filho - Campus de Marília, Brasil.
E-mail: cabrini@marilia.unesp.br.

** Doutora em Educação pela Universidade Estadual Paulista Júlio de Mesquita Filho, Brasil. Livre-Docente em Estudos Bibliométricos pela Universidade Estadual Paulista Júlio de Mesquita Filho, Brasil. Docente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, Brasil.
E-mail: etannuri@gmail.com.

I INTRODUÇÃO

Os estudos sobre análise de cocitação de autores (ACA) objetivam identificar os autores influentes e mostrar suas inter-relações, a partir das coocorrências de citações (White; Griffith, 1981; White; McCain, 1998). A ACA analisa a estrutura intelectual e social

de uma determinada área, campo científico ou conjunto de pesquisadores, evidenciando as proximidades teóricas e metodológicas entre eles.

Segundo os estudos de Luukkonen et al. (1993), as medidas absolutas e normalizadas de análise de coocorrência de autores trazem tipos diferentes de informações: as primeiras mostram os “atores” centrais nas redes, enquanto as normalizadas mostram a intensidade das

relações e revelam aspectos não identificáveis nas frequências absolutas.

Quando se pretende realizar estudos comparativos, dadas as especificidades de cada área, destaca-se a relevância dos indicadores normalizados, que padronizam as unidades de medida e revelam aspectos não explicitados nos absolutos. Entre os índices relativos, citam-se especialmente o Coeficiente de Correlação de Pearson, Cosseno de Salton e Índice de Jaccard. (LEYDERSDOFF; VAUGHAM, 2006).

Grácio e Oliveira (2013) apresentam um estudo em que se propõem a integrar a informação que os indicadores absolutos e relativos oferecem à Análise de Cocitação de Autores, a fim de explicitar, de forma sintética e analítica, as contribuições advindas de ambos os índices, por meio de uma aplicação ao universo dos “Estudos Métricos”, tendo como fonte de dados a base BRAPCI.

Em continuidade à pesquisa citada, esta pesquisa objetiva realizar um estudo comparativo da contribuição do coeficiente de Correlação de Pearson e do Cosseno de Salton como indicadores de proximidade para análise de cocitação de autores com vistas à compreensão do uso adequado de cada um deles, considerando a finalidade de um estudo de Análise de Cocitação de Autores e as particularidades do contexto do estudo em apreço.

De forma mais específica, propõe-se a apresentar e a analisar os indicadores de proximidade de Coeficiente de Correlação de Pearson (r) e Cosseno de Salton (CS), para a análise de cocitação de autores, e comparar a contribuição destes dois indicadores para a visualização das proximidades entre os autores cocitados, identificando os agrupamentos dos pesquisadores cocitados formados a partir dos dois índices e avaliando a correlação entre os indicadores. Para isso, utilizaram-se como universo os artigos publicados sobre a temática “Estudos Métricos”, na base de dados BRAPCI.

Apesar do corpo significativo da literatura existente, justifica-se a presente pesquisa por persistirem questões relativas à natureza dos indicadores de proximidade e do tipo de análise para os quais são apropriados. Embora em âmbito internacional existam diferentes estudos na temática, em especial no periódico *Journal of the American Society for Information Science and Technology*, há a necessidade de estudos

nacionais que insiram o Brasil nesta temática e que contribuam para o aprofundamento das questões relativas à pertinência e conveniência na utilização das diferentes medidas de proximidade nos estudos de cocitação de autores. Considera-se que estes estudos darão maior consistência à visualização das similaridades entre os autores cocitados e a estrutura subjacente a uma literatura científica analisada.

2 MEDIDAS DE PROXIMIDADE RELATIVA NA ANÁLISE DE COCITAÇÃO DE AUTORES: COEFICIENTE DE CORRELAÇÃO DE PEARSON E COSSENO DE SALTON

Os estudos de Análise de Cocitação de Autores baseiam-se em matrizes de valores absolutos ou relativos que mensuram a cocorrência entre os autores citados. Essas matrizes são um tipo de matriz de proximidade, que indicam quão próximos ou quão distantes estão os autores analisados, sob a perspectiva dos pesquisadores citantes.

As matrizes de proximidade podem ser matrizes de similaridade ou dissimilaridade (LEYDERSDOFF; VAUGHAM, 2006). Nas matrizes de similaridade, quanto maior o valor, seja indicador absoluto ou relativo de cocitação, mais similares são os dois autores. Observe-se desse modo que todas as matrizes de frequência absoluta ou relativa de cocitação são matrizes de similaridade.

As frequências absolutas de cocitação apresentam contribuição significativa na construção das redes de cocitação, medindo a conectividade científica entre os autores cocitados. Entretanto, elas contêm menos informação que as matrizes de índices relativos de cocitação. Esses índices – entre eles, o Coeficiente de Correlação de Pearson, Cosseno de Salton, Índice de Jaccard e Qui-Quadrado – têm sido, preferencialmente, utilizados nos procedimentos multivariados, especialmente nas análises de cluster, por revelarem, de modo consistente, aspectos subjacentes à estrutura intelectual de uma área científica, não identificáveis por análises de cocitação baseadas em frequências absolutas, ao eliminar a dependência do volume de literatura científica referenciada, que varia

entre as diversas áreas (VANZ; STUMPF, 2010, LEYDERSDOFF; VAUGHAM, 2006, GRÁCIO; OLIVEIRA, 2013).

Os índices para medir a similaridade entre os autores cocitados apoiam-se em dois tipos diferentes de procedimentos: na similaridade direta entre dois autores, focando-se somente na frequência de cocitação entre os mesmos, considerada uma medida local, também denominada bilateral; ou na forma com que os autores estão relacionados com os outros autores do conjunto de dados analisados, isto é, com os outros autores cocitados, considerada esta última uma medida global ou multilateral. Vários estudiosos têm apontado que esses métodos podem levar a resultados totalmente diferentes (AHLGREN; JARNEVING; ROUSSEAU, 2003).

Entre os índices relativos de medida local, destaca-se o Cosseno de Salton, que mede a proximidade de dois autores cocitados, independente do volume da literatura referenciada; por outro lado, entre os índices relativos de medida global, destaca-se o Coeficiente de Correlação de Pearson (r), que identifica e evidencia a similaridade do perfil de cocitação dos dois autores cocitados em relação aos outros autores cocitados.

O Cosseno de Salton (CS) é definido como a razão entre a frequência de cocitação de dois autores (X e Y) e a raiz quadrada da multiplicação dos números de citações recebidas individualmente pelos autores X e Y . A expressão matemática é apresentada por:

$$\text{CS} = \frac{\text{cocit}(X,Y)}{\sqrt{\text{cit}(X) \text{cit}(Y)}}$$

em que:

$\text{cocit}(X,Y)$ = número de pesquisas em que os autores X e Y foram cocitados;

$\text{cit}(X)$ = número de pesquisas em que o autor X foi citado;

$\text{cit}(Y)$ = número de pesquisas em que o autor Y foi citado.

Esse índice padroniza os valores de cocitação nas diferentes áreas do conhecimento ao apresentar números variando entre zero e um: quanto maior e mais próximo de um, mais intensa a similaridade temática pesquisada pelos dois autores; quanto

mais próximo de zero, menos intensa é a associação entre eles, na perspectiva do conjunto de autores citantes. Valor zero para CS significa ausência de cocitação entre os dois autores; valor igual a um indica que eles foram citados exatamente nas mesmas pesquisas, o que pode ser decorrente de os autores serem coautores ou por terem proximidade teórica e/ou metodológica.

Para se relativizar as frequências absolutas de cocitação, Small e Sweeney (1985) apontam que o Cosseno de Salton se apresenta mais adequado que o Índice de Jaccard, por tratar, de modo mais efetivo, com as ligações entre artigos com alta citação e baixa citação. Hamers et al. (1989) apontam que o Cosseno de Salton produz frequentemente um valor para o índice, que é duas vezes o número obtido para o Índice de Jaccard. Outros autores – entre eles, Ahlgren, Harneving e Rousseau (2003) e Eck e Waltman (2008) – posicionam-se a favor do uso do Cosseno de Salton. Por outro lado, há autores, como Leydesdorff (2008), que se posicionam pelo Índice de Jaccard como medida mais representativa de similaridade para normalização nos estudos de ACA. Assim, considerando que se observa um maior grupo de pesquisadores que defendem o uso do CS, este tem sido aplicado mais frequentemente na análise de cocitação de autores, o que justifica a escolha do CS para a comparação com o Coeficiente de Correlação de Pearson.

White e Griffith (1981) utilizaram o Coeficiente de Correlação de Pearson (r) na ACA, inicialmente, ao elaborar o mapa dos cientistas da informação.

O Coeficiente de Correlação de Pearson (r) mede a tendência de associação linear de duas variáveis quantitativas. Na Análise de Cocitação de Autores, o Coeficiente de Correlação mede a associação de pares de autores cocitados, dois a dois, considerando o conjunto de valores de cocitações de cada um deles com os demais em estudo, como as variáveis analisadas. O valor de r varia entre -1 e 1, portanto pode ser positiva ou negativa. Uma correlação positiva entre dois autores cocitados indica a mesma tendência de crescimento e decréscimo de valores de cocitação dos dois autores com os demais autores analisados. Por outro lado, uma correlação negativa significa tendências opostas de crescimento e decréscimo de valores de cocitação dos dois autores com os demais analisados, isto é, à medida que um deles tende a

ser cocitado com os demais, o outro tende a não ser cocitado com aqueles. Independente do sentido da correlação (positiva ou negativa), as correlações podem variar também quanto à sua força: desde ausência (igual a zero) até uma correlação muito forte ou até perfeita (valor de r igual a -1 ou 1).

Considerando x_i a frequência de cocitação de um autor X com os demais autores e y_i a frequência de cocitação de um autor Y com os demais autores, para i variando de 1 a n, com n igual a quantidade de autores no estudo, o Coeficiente de Correlação de Pearson (r) entre X e Y é definido pela fórmula:

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

Antes da conversão da matriz de frequências absolutas de cocitação para a matriz de similaridade com os valores de correlação, é necessário decidir como serão tratados os valores da diagonal, ou seja, os valores de cocitação do autor consigo mesmo. Para esta questão, a literatura da área propõe algumas alternativas, a saber: utilizar a maior frequência de cocitação do autor com os demais; adotar a média de cocitação de cada autor como valor da diagonal; usar o número de vezes que o autor foi cocitado consigo mesmo (número de pesquisas na literatura analisada que cita pelo menos dois trabalhos diferentes do autor); utilizar zeros na diagonal; ou considerar a diagonal como um conjunto de valores perdidos (*missing*). Dentre essas alternativas, a última apresenta maior aceitação e uso na comunidade dos pesquisadores, uma vez ser mais operacional e com menos *viés* conceitual, motivo pelo qual nesta pesquisa se utiliza o *missing* para diagonal.

O r de Pearson foi a medida padrão em ACA antes do artigo de Ahlgren, Jarneving e Rousseau (2003), que criticaram seu uso, mostrando que o mesmo não satisfaz como medida de similaridade ou proximidade.

Apresenta-se, na Figura 1, um exemplo hipotético de matrizes de frequência de cocitação absoluta, em que $C_{A,B}$ denota o valor de cocitação entre os autores A e B, supondo-se que o total

de citações recebidas pelo autor A foi igual a 15 e que o autor B recebeu 19 citações, a partir das quais são calculadas as Correlações de Pearson e o Cosseno de Salton, com o objetivo de simular algumas possibilidades de resultados, comparando esses índices relativos.

Figura 1: Matrizes hipotéticas com frequências de cocitação entre autores A, B, P1 a P5.

Matriz 1:

Autores cocitados (frequências)	A	B	P1	P2	P3	P4	P5
A	-	$C_{A,B}$	3	9	5	11	1
B	$C_{A,B}$	-	4	10	7	8	2

Matriz 1: r = 0,89

Matriz 2:

Autores cocitados (frequências)	A	B	P1	P2	P3	P4	P5
A	-	$C_{A,B}$	3	9	5	11	1
B	$C_{A,B}$	-	14	2	6	3	10

Matriz 2: r = -0,84

Matriz 3:

Autores cocitados (frequências)	A	B	P1	P2	P3	P4	P5
A	-	$C_{A,B}$	3	9	5	11	1
B	$C_{A,B}$	-	8	7	7	5	4

Matriz 2: r = 0,04

Nas matrizes presentes na Figura 1, destaca-se que o valor de $C_{A,B}$ não entra no cálculo do Coeficiente de Correlação (r), uma vez que ele está sempre pareado com o valor da diagonal (*missing*). Desse modo, a proximidade de cocitação entre os próprios dois autores não interfere no resultado de r, ou seja, o valor de r independe de os autores serem altamente cocitados ou não.

Na Matriz 1, obteve-se uma alta correlação positiva (r = 0,89) entre os autores cocitados A e B, indicando igual tendência de crescimento dos valores de cocitação de cada um dos autores com os demais (P1 a P5).

Na Matriz 2, obteve-se uma alta correlação negativa (r = -0,84) entre os autores cocitados A e B, indicando tendências opostas de crescimento

ou decrescimento de valores de cocitação dos dois autores com os demais, isto é, à medida que A tende a ser cocitado com os demais (P1 a P5), o autor B tende a não ser cocitado e vice-versa.

Na Matriz 3, obteve-se a quase ausência de correlação ($r = 0,04$) entre os autores citados, não havendo tendência de associação de crescimento ou decrescimento de valores.

Todavia, ressalta-se que, nas três matrizes analisadas, o Coeficiente de Correlação (r) avalia somente a tendência de crescimento dos valores: se os valores de cocitação do autor B fossem duplicados ou multiplicados por uma mesma constante, o valor de r continuaria o mesmo, ou seja, a alta intensidade de r (quer negativa ou positiva) ou sua ausência independe da proximidade de valor dos pares de cocitação entre A e B.

Para se obter o valor do índice relativo de Cosseno de Salton entre os autores A e B, levam-se em consideração somente o valor de cocitação entre A e B ($C_{A,B}$), não necessário para o cálculo da Correlação de Pearson, e os totais de citação de cada autor. Para efeito de simulação do comportamento deste índice, se $C_{A,B}$ for igual a 12, obtém-se $CS = 0,71$, indicando uma significativa similaridade temática entre os dois autores cocitados, advinda de alta coocorrência de citação entre esses autores ou de coautoria. Por outro lado, se $C_{A,B}$ for igual a 3, obtém-se $CS = 0,18$, indicando baixa proximidade temática entre os autores.

Desse modo, complementa-se ainda que, na Matriz 1, caso $C_{A,B}$ fosse igual a três, se teria uma igual tendência de crescimento dos valores de cocitação de A e B com os demais autores (alta correlação positiva) e baixa proximidade temática entre os autores A e B (CS igual a 0,18). Por outro lado, na Matriz 2, raciocínio análogo e inverso pode ser feito para $C_{A,B}$ igual a 12, em que se teria uma tendência oposta de crescimento dos valores de cocitação de A e B com os demais autores (alta correlação negativa) e significativa proximidade temática entre os autores A e B (CS igual a 0,71).

Destaca-se que as informações utilizadas no cálculo do Coeficiente de Correlação de Pearson são distintas daquelas utilizadas no cálculo do Cosseno de Salton, uma vez que, no último, se consideraram apenas dados relativos aos próprios autores, a saber, total de citações de cada um e cocitações

entre eles, sendo assim uma medida local de similaridade. Por outro lado, no cálculo da Correlação de Pearson são considerados os dados de cocitação dos autores com todos os demais na análise (P1 a P5), portanto uma medida global de similaridade.

Isso posto, aponta-se a independência dos dois índices, que se comportam como medidas complementares.

No entanto, não se pode deixar de mencionar os debates no entorno destes dois índices de similares utilizados na ACA. Entre os autores que defendem o uso do Coeficiente de Correlação de Pearson, destacam-se White (2003) e Besnam (2004), enquanto Ahlgren, Jarneving e Rousseau (2003) e Eck e Waltman (2008) se posicionam favoravelmente ao Cosseno de Salton.

3 PROCEDIMENTOS METODOLÓGICOS

Inicialmente, recuperaram-se as matrizes quadradas e simétricas de frequência absoluta de cocitação entre os 38 autores mais citados e de valores de cocitação relativizados pelo Cosseno de Salton, apresentadas nos resultados de Gracio e Oliveira (2013) relativos à temática “Estudos Métricos”, nos artigos indexados na base BRAPCI.

A partir da matriz quadrada e simétrica (38×38) com as frequências absolutas de cocitação entre os 38 autores mais citados, adotando-se *missing* como valores para a diagonal, foi construída a matriz, também quadrada e simétrica 38×38 , com os valores de Correlação de Pearson entre os autores cocitados, utilizando-se o *software SPSS*. Analisaram-se os índices em destaque nas matrizes com os índices normalizados por Cosseno de Salton e por Correlação de Pearson.

A seguir, a partir das duas matrizes quadradas e simétricas 38×38 com índices de Cosseno de Salton e r de Pearson entre os autores cocitados, calculou-se a correlação entre os 703 pares¹ dos índices Cosseno de Salton e r de Pearson a fim de se avaliar a tendência de associação e força de crescimento

¹ Os 703 pares de valores foram calculados a partir da multiplicação do número de linhas e colunas das matrizes 38×38 , excluindo os 38 valores da diagonal principal e dividindo o resultado por 2, dada a simetria da matriz.

Tabela 2: Índices de Correlação (r) de Pearson entre os 38 autores cocitados.

Índice de correlação de P	MEADOWS, J.A.	VANTIN	PRICE, J.de S.	BLANZEL, W.	MUELLER, S. P. M.	SPINAK, E.	SANTOS, R. M.S.	ROUSSEAU, R.	NORONHA, D. P.	PINHEIRO, L. V. R.	ARAÚJO, C. A. A.	BUFREM, L. S.	CAREGNATO, S. E.	VANZ, S.	VELHO, L.	MUGNANI, R.	ALVARADO, R.U.	KOBASHI, N. Y.	CHAPULA, C.M.	STUMPF, I. R.	OLIVEIRA, M.	GARFIELD, E.	PRATES, Y.	POBLACION, D.	KATZ, J. S.	MARTIN, B. R.	ANEGON, F.m.	QUONIAM, L.	FARIA, L. I. L.	SCHUBERT, A.	SOLANA, V.H.	LE COADIG, Y. F.	WASSERMAN, S.	BORSCHIVER, S.	PINTO, A. L.	LETA, J.	GUEDES, V.	JANNUZZI, P.M.
MEADOWS, J. A.	1,00	0,34	0,21	0,42	0,32	0,07	0,51	0,27	0,37	0,48	0,18	0,58	0,53	0,29	-0,03	0,32	0,14	0,18	0,52	0,54	0,37	0,20	0,35	0,28	0,28	0,01	0,01	0,36	0,26	0,01	0,48	0,08	-0,05	0,15	0,24	-0,03	0,01	
VANTI, N.	1,00	0,01	-0,39	0,13	0,39	0,66	0,30	0,28	0,40	0,74	0,56	0,26	0,29	-0,29	0,26	0,41	0,65	0,64	0,41	0,26	0,18	0,62	0,27	-0,46	-0,46	0,01	0,17	-0,28	-0,53	0,21	0,43	-0,35	0,36	0,11	-0,37	0,35	0,24	
PRICE, J. de S.	1,00	0,40	0,29	0,16	-0,29	0,18	-0,01	0,14	0,23	-0,11	0,12	0,23	0,34	0,01	0,27	-0,15	-0,13	0,19	0,32	0,45	-0,10	0,12	0,36	0,40	0,04	-0,03	0,32	0,48	-0,02	0,46	0,35	-0,22	-0,07	0,40	-0,21	0,05		
GLÄNZEL, W.	1,00	0,23	-0,15	-0,53	0,12	-0,04	-0,29	-0,07	-0,49	0,21	0,30	0,69	-0,20	-0,22	-0,52	-0,49	0,18	0,06	0,30	-0,48	0,02	0,70	0,70	0,00	-0,08	0,60	0,85	-0,15	0,15	0,63	-0,47	-0,06	0,69	-0,49	-0,12			
MUELLER, S. P. M.	1,00	0,26	-0,01	0,14	0,65	0,46	0,32	0,12	0,53	0,59	0,44	0,16	0,41	0,08	0,14	0,62	0,70	0,10	0,16	0,53	0,35	0,35	0,21	0,26	0,49	0,18	0,31	0,08	-0,19	0,43	0,50	-0,19	0,19					
SPINAK, E.	1,00	0,35	0,33	0,31	0,17	0,53	0,40	0,28	0,26	0,00	0,24	0,32	0,41	0,45	0,37	0,37	0,29	0,40	0,37	0,03	0,00	0,05	0,17	0,09	0,11	0,18	0,28	0,11	0,05	0,07	0,11	0,10	0,15					
SANTOS, R. M.S.	1,00	0,18	0,21	0,33	0,42	0,73	0,03	0,02	-0,36	0,34	0,34	0,86	0,79	0,24	0,23	0,07	0,72	0,20	-0,55	-0,53	-0,07	0,35	-0,32	-0,52	0,12	0,06	-0,42	0,36	0,02	-0,50	0,38	0,33						
ROUSSEAU, R.	1,00	0,11	0,12	0,55	0,17	0,27	0,22	0,01	0,13	0,29	0,23	0,24	0,50	0,18	0,62	0,13	0,08	-0,02	-0,03	0,20	0,16	0,01	0,16	0,10	0,40	0,09	0,12	-0,05	0,00	0,11	0,14							
NORONHA, D. P.	1,00	0,50	0,32	0,37	0,49	0,46	0,28	0,26	0,41	0,32	0,23	0,52	0,65	-0,08	0,31	0,81	0,08	0,10	0,32	0,24	0,32	-0,05	0,32	0,32	-0,02	-0,15	0,59	0,25	-0,16	0,26								
PINHEIRO, L. V. R.	1,00	0,32	0,56	0,29	0,27	0,00	0,19	0,59	0,38	0,47	0,39	0,68	-0,01	0,48	0,62	-0,17	-0,17	0,21	0,17	-0,04	-0,38	0,28	0,26	-0,33	0,14	0,47	-0,10	0,16	0,16									
ARAÚJO, C. A. A.	1,00	0,44	0,53	0,58	0,00	0,06	0,30	0,38	0,45	0,67	0,45	0,27	0,36	0,31	-0,02	-0,02	-0,10	-0,07	0,00	-0,15	0,11	0,63	0,00	0,37	-0,07	-0,06	0,36	0,01										
BUFREM, L. S.	1,00	0,10	0,16	-0,27	0,35	0,43	0,71	0,78	0,25	0,37	-0,01	0,93	0,43	-0,42	-0,43	0,04	0,29	-0,24	-0,48	0,23	0,22	-0,38	0,32	0,17	-0,37	0,30	0,32											
CAREGNATO, S. E.	1,00	0,85	0,40	-0,03	0,00	0,11	0,16	0,81	0,45	0,08	0,11	0,39	0,39	0,02	-0,05	0,43	0,20	0,00	0,31	0,07	-0,07	0,13	0,40	-0,02	-0,06													
VANZ, S.	1,00	0,46	-0,09	0,12	0,10	0,10	0,78	0,47	-0,02	0,07	0,41	0,40	0,41	0,00	-0,17	0,46	0,18	-0,04	0,44	0,18	-0,01	0,16	0,35	-0,05	-0,11													
VELHO, L.	1,00	-0,13	-0,11	-0,35	-0,32	0,29	0,33	0,10	-0,26	0,27	0,70	0,74	0,03	-0,09	0,90	0,60	-0,19	0,11	0,40	-0,34	0,13	0,80	-0,39	-0,01														
MUGNANI, R.	1,00	0,30	0,46	0,35	0,15	0,14	0,28	0,32	0,31	-0,30	-0,32	0,30	0,92	0,02	-0,07	0,40	0,03	-0,30	-0,08	0,34	-0,01	-0,13	0,94															
ALVARADO, R.U.	1,00	0,36	0,35	0,24	0,60	0,39	0,42	0,37	-0,20	-0,23	0,25	0,32	-0,19	-0,22	0,40	0,44	-0,17	0,05	0,34	-0,08	0,06	0,28																
KOBASHI, N. Y.	1,00	0,77	0,28	0,22	0,00	0,74	0,43	-0,48	-0,51	0,02	0,41	-0,17	-0,49	0,26	0,17	-0,41	0,37	0,20	-0,38	0,32	0,46																	
CHAPULA, C.M.	1,00	0,31	0,36	-0,03	0,76	0,45	-0,52	-0,53	0,09	0,27	-0,26	-0,51	0,28	0,18	-0,34	0,40	0,10	-0,47	0,37	0,31																		
STUMPF, I. R.	1,00	0,45	0,24	0,18	0,50	0,20	0,20	0,14	0,13	0,30	0,11	0,20	0,47	0,10	-0,01	0,20	0,29	-0,03	0,17																			
OLIVEIRA, M.	1,00	0,12	0,40	0,59	0,22	0,20	0,25	0,13	0,29	-0,02	0,23	0,46	-0,04	-0,01	0,40	0,21	-0,01	0,11																				
GARFIELD, E.	1,00	-0,01	-0,02	0,12	0,09	0,05	0,35	0,00	0,28	0,02	0,44	0,08	-0,01	-0,16	0,24	-0,01	0,35																					
PRATES, Y.	1,00	0,34	-0,41	-0,42	0,02	0,29	-0,24	-0,42	0,19	0,15	-0,23	0,33	0,21	-0,41	0,31	0,31																						
POBLACION, D.	1,00	0,01	0,01	0,44	0,38	0,30	-0,06	0,51	0,24	-0,04	-0,12	0,69	0,24	-0,13	0,40																							
KATZ, J. S.	1,00	0,09	-0,07	-0,26	0,66	0,71	-0,30	0,11	0,49	-0,32	-0,05	0,77	-0,32	-0,27																								
MARTIN, B. R.	1,00	-0,06	-0,27	0,65	0,71	-0,30	0,10	0,50	-0,34	-0,04	0,77	-0,33	-0,28																									
ANEGON, F.m.	1,00	0,29	0,05	0,10	0,83	0,15	0,14	-0,36	0,74	0,17	-0,34	0,29																										
QUONIAM, L.	1,00	0,08	-0,04	0,40	-0,08	-0,29	-0,11	0,39	0,09	-0,15	0,97																											
FARIA, L. I. L.	1,00	0,60	-0,13	-0,01	0,31	-0,32	0,22	0,78	-0,30	0,10																												
SCHUBERT, A.	1,00	-0,13	-0,04	0,59	-0,39	-0,02	0,78	-0,39	-0,01																													
SOLANA, V.H.	1,00	0,06	0,06	-0,21	0,71	0,02	-0,20	0,41																														
LE COADIG, Y. F.	1,00	0,12	0,05	-0,04	0,07	0,06	0,04																															
WASSERMAN, S.	1,00	-0,34	-0,02	0,51	-0,34	-0,26																																
BORSCHIVER, S.	1,00	-0,26	-0,42	0,96	-0,13																																	
PINTO, A. L.	1,00	0,24	-0,26	0,38																																		
LETA, J.	1,00	-0,42	0,11																																			
GUEDES, V.	1,00	-0,09																																				
JANNUZZI, P.M.	1,00																																					

Fonte: Grácio e Oliveira (2013)

Apresenta-se a Tabela 2 com os índices de Correlação (r) de Pearson entre os 38 autores cocitados, onde se destacam, com grifo em amarelo, os valores de correlação acima de 0,75, considerados pela literatura científica aqueles com alta intensidade de tendência de associação, significando neste estudo alta intensidade global de cocitação entre os pares de autores citados.

Destacam-se os maiores valores – acima de 0,9 – de índices de correlação entre: Mugnaini, Quoniam e Jannuzzi; Katz e Martin; Guedes e Borschiver; e Bufrem e Prates. Todos esses índices são decorrentes de artigos em coautoria entre eles, observando-se que, em geral, ocorrendo a citação conjunta de um dos autores com outros, o segundo autor também será cocitado com estes. Assim, coautores apresentaram alto índice de cocitação local, expresso pelo Cosseno de Salton, e alto índice de cocitação global, expresso pela Correlação de Pearson.

Entre os altos índices de cocitação por Correlação de Pearson – entre 0,75 e 0,90 – encontram-se: Bufrem e Chapula; Santos e Chapula; Leta e Martin; Leta e Velho; e Kobashi e Chapula, decorrentes de cocitação no conceito próprio do termo.

A Figura 2 apresenta o gráfico de dispersão dos valores dos dois índices analisados – Cosseno de Salton e Correlação de Pearson - ,

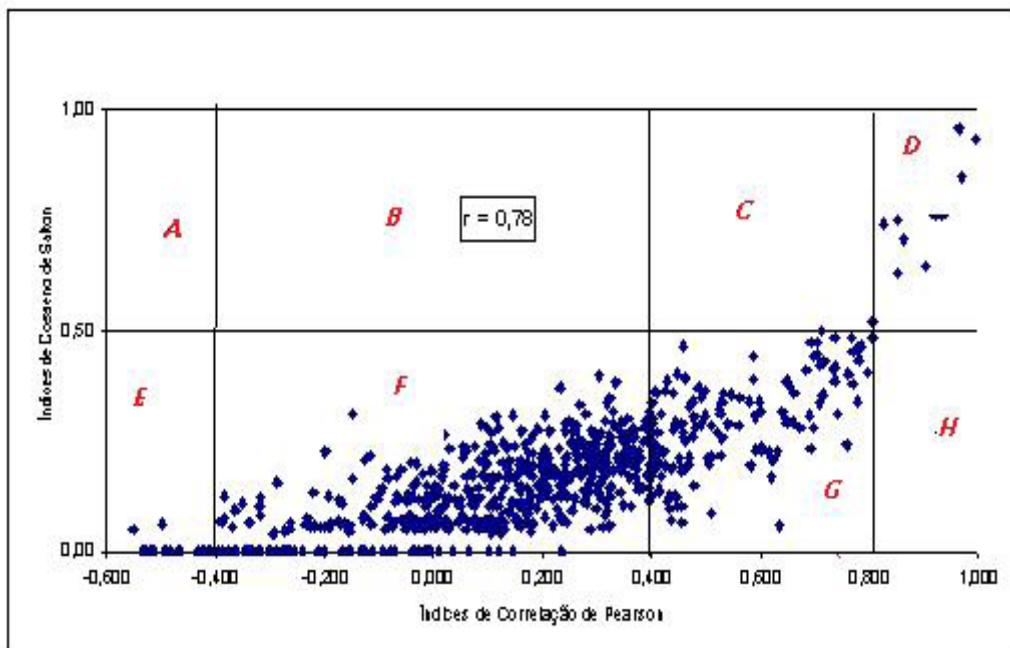
por par ordenado de autores cocitados, e o valor da Correlação de Pearson (r) entre os dois índices normalizados.

Para análise mais elucidativa da Figura 2, dividiu-se o plano do gráfico em oito quadrantes, onde os quadrantes A, B, C e D representam altos valores de Cosseno Salton (acima de 0,50). Para Correlação de Pearson, os quadrantes D e H representam altos valores para este índice. Assim, o quadrante D apresenta altos valores para o Cosseno de Salton e Correlação de Pearson, totalizando 13 pares de índices, todos eles advindos de proximidade, em geral, resultantes de coautoria.

Os quadrantes E, F e G apresentam baixos valores para o Cosseno de Salton e para a Correlação de Pearson, significando que, embora o primeiro meça a proximidade local e o segundo meça a proximidade global, estes índices apresentaram as mesmas tendências de proximidade, validados pela significativa Correlação de Pearson (0,78) entre eles.

Destaca-se que, com exceção de 13 pares de valores de índice (CS e r) contemplados no quadrante D, todos os demais se encontram nos quadrantes E, F e G, com grande concentração no quadrante F, em que o Cosseno de Salton está abaixo de 0,5 e a Correlação de Pearson varia entre -0,4 e 0,4.

Figura 2: Gráfico de dispersão entre os índices de Correlação de Pearson e Cosseno de Salton



Fonte: elaborado pelos próprios autores.

Apresentam-se, na Figura 3, os dendogramas a e b, construídos a partir da análise de *clusters*, em função dos índices de similaridades, de Cosseno de Salton e Correlação de Pearson.

Na Figura 3a, segundo similaridades do Cosseno de Salton, consideram-se seis agrupamentos, a saber:

- G1: composto por oito pesquisadores, é o grupo que, proporcionalmente, apresenta a mais forte presença de estrangeiros, considerando que somente três deles são brasileiros. Destaque-se que a similaridade de cocitação é decorrente em parte de coautorias, a exemplo de Katz e Martin. As temáticas que aproximam esse grupo de pesquisadores se relacionam à colaboração científica, indicadores e avaliação da ciência brasileira.
- G2: este grupo é composto somente por três pesquisadores, sendo um deles estrangeiro. A temática de destaque é o estudo de indicadores, e as cocitações decorrem também de pesquisas em parcerias.
- G3: formado por nove pesquisadores, sendo dois deles espanhóis, este grupo caracteriza-se por apresentar similaridades de cocitações advindas, principalmente, do conceito próprio de cocitação e esporádicas similaridades oriundas de coautorias. Em relação às temáticas, apresentam-se as mais frequentes: epistemologia, comunicação científica, análise de domínio, estruturas científicas – especialmente dos dois pesquisadores espanhóis – e indicadores de produção científica.
- G4: maior grupo, constituído por dez pesquisadores, sendo cinco deles estrangeiros, configura-se por proximidades temáticas e advém, assim como G3, de cocitações no sentido original do termo, considerando que não se registraram coautorias entre seus componentes, exceto entre as três primeiras constituintes do grupo, a saber: Caregnato, Vanz e Stumpf. Caracteriza-se por oferecer referentes teóricos para o tema em estudo, em especial relativos à comunicação científica e seus fundamentos, sociologia da ciência, citação, assim como os fundamentos e desenvolvimento teórico-metodológico dos estudos bibliométricos.
- G5: composto por apenas dois pesquisadores que apresentam similaridade decorrente exclusivamente de coautoria. Apresenta uma revisão dos conceitos bibliométricos, suas leis clássicas sob o contexto da área de gestão.
- G6: grupo formado por seis pesquisadores, sendo dois estrangeiros, com destaque para as similaridades de cocitações advindas de coautorias entre os brasileiros. As temáticas trabalhadas referem-se a questões epistemológicas e aplicadas, sempre sob o olhar do paradigma social.

Figura 3: Dendograma em função dos índices de similaridades

Figura 3a: Dendograma em função dos índices de similaridades de Cosseno de Salton.

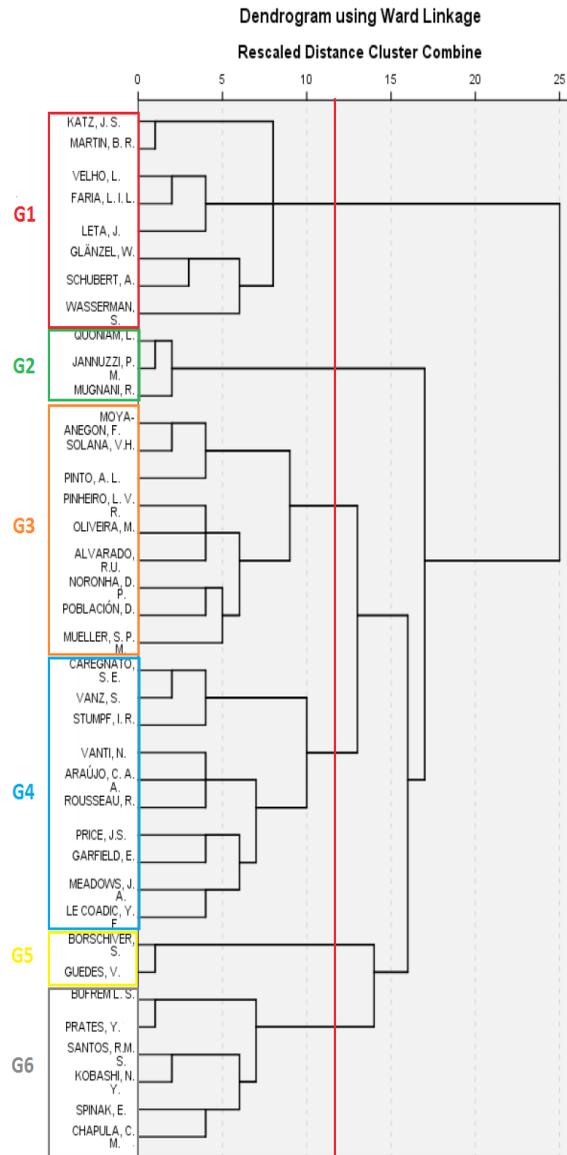
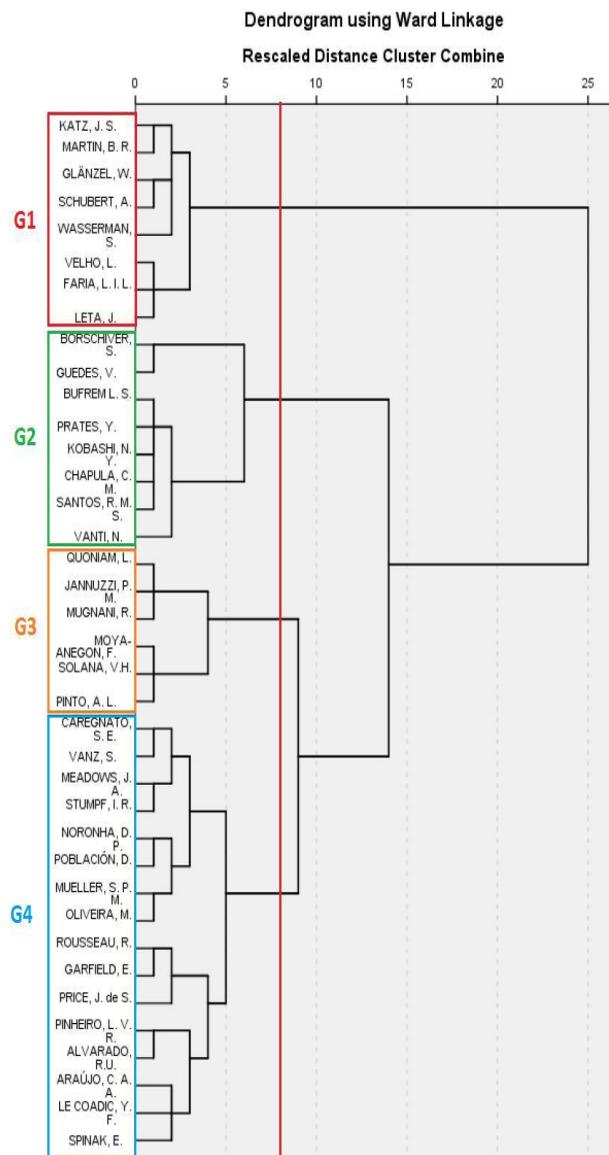


Figura 3b: Dendograma em função dos índices de similaridades de Correlação de Pearson.



Fonte: elaborado pelos próprios autores

Ainda, na Figura 3b, segundo as similaridades em função da Correlação de Pearson, destacaram-se quatro grupos, a saber:

- G1: composto pelos mesmos (oito) pesquisadores presentes no G1 resultante

do agrupamento segundo as similaridades por Cosseno de Salton.

- G2: formado por oito pesquisadores: um estrangeiro, correspondendo a dois pesquisadores do G5, e cinco dos seis

pesquisadores do G6, com exceção do pesquisador Spinak, do dendograma da Figura 3a. Soma-se, ainda, a pesquisadora Vanti, membro do G4 na Figura 3a. Neste grupo, os trabalhos citados são, principalmente, de natureza teórico-conceitual e epistemológica e pesquisas aplicadas, com enfoque nas questões sociais e contextuais.

- G3: constituído por seis pesquisadores: três deles correspondendo a todos os pesquisadores do grupo G2, e os outros três pesquisadores, dos nove membros do G3, relativo ao dendograma da Figura 3a. Referem-se às temáticas relativas à Análise de Domínio e indicadores da produção científica.
- G4: maior grupo, formado por 16 pesquisadores: sete dos nove pesquisadores do G3, oito dos 10 pesquisadores do G4, e um pesquisador dos seis presentes no G6, do dendograma gerado a partir das similaridades por Cosseno de Salton, constituindo-se assim o agrupamento que mais mescla pesquisadores de distintos grupos da Figura 3a. As principais temáticas que agregam os pesquisadores deste grupo são relativas à comunicação científica e seus canais de divulgação, avaliação de periódicos, assim como trabalhos teóricos e aplicados sobre análise de citação e colaboração científica.

Em síntese, comparando-se os dois dendogramas, resultantes das similaridades advindas do Cosseno de Salton e da Correlação linear, destaca-se que as proximidades resultantes de coautoria se mantiveram explícitas nos dois agrupamentos.

5 CONSIDERAÇÕES FINAIS

Este estudo se propôs a apresentar, analisar e comparar os indicadores de proximidade de Coeficiente de Correlação de Pearson (r) e Cosseno de Salton (CS), para a visualização das proximidades nas análises de cocitação de autores, além de agrupar os pesquisadores cocitados a partir dos dois índices. Utilizaram-

se como universo os artigos publicados sobre a temática “Estudos Métricos”, na base de dados BRAPCI.

Com base nas duas matrizes com índices de Cosseno de Salton e r de Pearson entre os 38 autores cocitados, analisaram-se os índices em destaque nas duas matrizes, a partir dos quais observaram-se: cocitações advindas de coautorias, que apresentaram altos índices de cocitação local, expresso pelo Cosseno de Salton; e altos índices de cocitação global, expressos pela Correlação de Pearson, como previsto. Altos índices de cocitação global advindos do conceito próprio do termo foram observados para pares de autores nacionais e estrangeiros, sugerindo a proximidade entre esses referentes, quer seja teórica ou metodológica, como foram os pares de autores Bufrem e Chapula; Santos e Chapula; Leta e Martin; Leta e Velho; e Kobashi e Chapula.

Ratificando a afirmação anterior, no gráfico de dispersão dos valores dos dois índices analisados – Cosseno de Salton e Correlação de Pearson –, constatou-se que altos valores para o Cosseno de Salton e Correlação de Pearson são advindos de proximidade, em geral, resultantes de coautoria. Por outro lado, tendências semelhantes de baixa proximidade local (Cosseno de Salton) e proximidade global (Correlação de Pearson) ocorreram, em geral, para os demais índices.

Comparando os agrupamentos apresentados a partir dos dois dendogramas, resultantes das similaridades advindas do Cosseno de Salton e da Correlação linear, observou-se que as proximidades resultantes de coautoria se mantiveram explícitas nos dois agrupamentos, assim como os grupos constituídos por pesquisadores próximos por cocitação oriunda do conceito próprio do termo.

Finalizando, considera-se que os resultados obtidos contribuíram para dirimir questões relativas à natureza dos indicadores de proximidade e do tipo de análise para os quais são apropriados, ao apresentar, neste estudo, dados nacionais que permitem melhor visualização das similaridades entre os autores cocitados e a estrutura subjacente a uma literatura científica analisada.

PROXIMITY INDICATORS IN AUTHOR CO-CITATION ANALYSIS: a comparative study between the Pearson's correlation coefficient and Salton's cosine

ABSTRACT This study aims to present and analyze Pearson Correlation Coefficient (r) and Salton's Cosine (CS) as proximity indicators to Author Co-citation Analysis and compare the contribution of these two indicators to visualize proximity among co-cited authors, evaluating the correlation between the indicators and identifying clusters of co-cited researchers formed from the two indices, using published articles on the topic "Metric Studies" from BRAPCI database. From a previous research, we retrieved the matrices with absolute frequencies of co-citation of the 38 most cited authors, normalized by Salton's cosine. Then the matrix was yielded with the values of the Pearson's correlation among the co-cited authors, using the SPSS software. The indices were analyzed, especially Salton's cosine and the Pearson's correlation. The correlation among the 703 pairs from both indices were calculated, represented by a scatterplot of values. From each matrix- Salton's Cosine and Pearson's Correlation -, two analyzes of clusters were performed, generating two dendograms, which were used to the comparative analysis of the clusters. In conclusion, we highlight that coauthors have high indices of CS and r . Moreover, we observed a trend toward low values for CS paired with low values of r . We concluded, therefore, that although CS measures local proximity and r measures global proximity, these indices show the same trends in relation to proximity.

Keywords: Author Co-citation Analysis. Normalized Proximity Indicators. Pearson's Correlation Coefficient. Salton's Cosine.

Artigo recebido em 16/12/2015 e aceito para publicação em 22/05/2015

REFERÊNCIAS

- AHLGREN, P.; JARNEVING, B.; ROUSSEAU, R. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. **Journal of the American Society for Information Science and Technology**, v. 54, n.6, p.550-560, 2003.
- Bensman, S.J. Pearson's r and author cocitation analysis: a commentary on the controversy. **Journal of the American Society of Information Science & Technology**, New York, v.55, n.10, p.935-936, 2004.
- ECK, N. J. V.; WALTMAN, L. Appropriate similarity measures for author co-citation analysis. **Journal of the American Society for Information Science and Technology**, v. 59, n. 10, p.1653-1661, 2008.
- GRACIO, M. C. C.; OLIVEIRA, E. F. T. Análise de citação de autores: um estudo teórico-metodológico dos indicadores de proximidade, aplicados ao GT7 da Ancib. **LIINC em Revista**, v. 9, p. 196-213, 2013.
- HAMERS, L. et al. Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. **Information Processing & Management**, v. 25, n. 3, p. 315-318, 1989.
- LEYDESDORFF, L. On the normalization and visualization of author co-citation data: Salton's cosine versus the Jaccard index. **Journal of the American Society for Information Science and Technology**, v. 59, n. 1, p.77-85, 2008.
- 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**, v. 57, n. 12, p.1616-1628, 2006.
- LUUKKONEN, T. et al. The measurement of international scientific collaboration. **Scientometrics**, Amsterdam, v. 28, n. 1, p. 15-36, 1993.
- SMALL, H.; SWEENEY, E. Clustering the Science Citation Index using co-citations: 1. A comparison of methods. **Scientometrics**, v.7, n. 3-6, p.391-409, 1985.
- VANZ, S. A.S.; STUMPF, I.R.C. Procedimentos e ferramentas aplicados aos estudos bibliométricos. **Informação & Sociedade**, v.20, n.2, p. 67-75, 2010.
- WHITE, H.D.; GRIFFITH, B. Author co-citation: a literature measure of intellectual structure. **Journal of the American Society for Information Science & Technology**, 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**, v. 49, n. 4, p. 327-355, 1998.