XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017

GT-7 – Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação

ANÁLISE DE COCITAÇÃO DE AUTORES – ACA: ESTUDO EXPLORATÓRIO COMPARANDO PROXIMIDADE NAS REFERÊNCIAS, SEÇÃO DO ARTIGO E PARÁGRAFO

Rodrigo Aquino de Carvalho (UFRGS)

Sônia Elisa Caregnato (UFRGS)

AUTHOR CO-CITATION ANALYSIS – ACA: EXPLORATORY STUDY COMPARING PROXIMITY IN REFERENCES, SECTION OF ARTICLE AND PARAGRAPH

Modalidade da Apresentação: Pôster

Resumo: O presente trabalho tem como tema a Análise de cocitação de autores – ACA. Para tanto, os objetivos são: 1) determinar a proximidade por seção do artigo e por parágrafo de um par de autores cocitados de uma ACA tradicional; e 2) identificar indícios de uma subtipologia de cocitação de autores na proximidade por parágrafo. O trabalho é de caráter exploratório e metodológico. Um par de autores (cocitação = 23; índice de similaridade = 0,88) foi destacado do principal fator de uma ACA tradicional. Os principais resultados indicam que na proximidade por seção, considerando apenas a ocorrência, há cocitação em 20 artigos citantes (86,96%), enquanto que no parágrafo há cocitação em 10 artigos citantes (43,48%). Os pares a partir das menções foram 100 na proximidade por seção e 20 na proximidade por parágrafo. A distância entre os autores na cocitação por parágrafo oferece um indício de subtipologia, pois quando não há distância pode-se afirmar que a cocitação não apresenta diferença temática. Isso ocorreu apenas em dois artigos citantes (dois dos 20 pares) e mostra que a cocitação foi utilizada para estabelecer uma relação efetiva. Os dados mostram que a cocitação circunstanciada a lista de referência é válida e que outras abordagens podem enriquecer as análises.

Palavras-Chave: Análise de citação; Análise de cocitação; Análise de cocitação de autores; Cocitação por proximidade.

Abstract: The focus of this paper is on author co-citation analysis (ACA). Its objectives are: 1) to determine the proximity of a pair of authors co-cited in the same section or in the same paragraph of an article; and, 2) to identify types of author co-citation by paragraph. The research work is exploratory and methodological. A pair of authors (co-citation = 23; similarity index = 0.88) was highlighted from the main factor of a traditional ACA. The main results indicate that when applying section proximity, considering only the occurrence, there is co-citation in 20 citing articles (86.96%), while in the proximity by paragraph there is co-citation in 10 citation articles (43.48%). Results show 100 pairs of authors co-cited in the section and 20 in the analysis by paragraph. The distance between authors in a paragraph can indicate the existence of categories of co-citation, since no distance means there is no thematic difference. This occurred only in two citing articles (two of the 20 pairs) and shows that co-

citation was used to establish an effective relationship between the two documents. Data show that the co-citation based on the reference list can be enriched by other approaches.

Key-words: citation analysis; Cocitation analysis; Author Co-citation Analysis – ACA; Cocitation proximity.

1 INTRODUÇÃO

A cocitação é um indicador de relacionamento derivado da citação que estabelece a ligação entre duas entidades (autores, documentos ou periódicos) a partir de um documento citante. Estudos nessa natureza são geralmente utilizados para visualizar domínios ou subdomínios do conhecimento (SCHNEIDER; LARSEN; INGWERSEN, 2009) produzido e publicado em um determinado contexto: bases de dados, teses e dissertações, periódicos etc.

O presente trabalho, de caráter metodológico, tem como foco a análise de cocitação de autores (ACA), na área da organização do conhecimento (KO – *Knowledge organization*), de artigos publicados em periódicos indexados na base de dados *Web of Science* (WoS), entre os anos de 2011 e 2015. Para tanto, os objetivo são: 1) determinar a proximidade por seção do artigo e por parágrafo de um par de autores cocitados de uma ACA tradicional, ou seja, feita a parti de dados retirados das referências (proximidade por artigo); e 2) identificar indícios de uma subtipologia de cocitação de autores a partir da proximidade por parágrafo, sendo que a proximidade por frase não será uma categoria separada, mas contida no parágrafo.

Essa abordagem é teste exploratório que visa objetivar a fase de interpretação e validação de uma ACA tradicional, isto é, que seja utilizada no(s) agrupamento(s) principal(is) formado(s) pela análise multivariada (quinta fase de uma ACA tradicional), para ampliar a qualidade de visualização do relacionamento entre os autores, indo além do nível circunstanciado à lista de referência.

2 ASPECTOS CONCEITUAIS E TRABALHOS RELACIONADOS

A ACA é um tipo de análise de cocitação que serve principalmente à visualização de domínios do conhecimento (SCHNEIDER; LARSEN; INGWERSEN, 2009; GRÁCIO; OLIVEIRA, 2013), assim como o acoplamento bibliográfico, a cocitação de documentos e a cocitação de periódicos.

A técnica consiste basicamente em seis passos (McCAIN, 1990) que apresentam diversas questões metodológicas que pode determinar resultados distintos: 1) seleção dos autores; 2) recuperação das frequências de cocitação; 3) compilação da matriz simétrica com os valores

absolutos; 4) normalização da matriz através de um índice de similaridade; 5) análise multivariada da matriz normalizada; e 6) interpretação e validação dos dados.

A seleção dos autores, por exemplo, implica em definir se serão considerados todos os autores das referências ou apenas os primeiros (PERSON, 2011; SCHNEIDER; LARSEN; INGWERSEN, 2009; CARVALHO; CAREGNATO, 2016), além de considerar o número de referências e o número de artigos citantes para a inclusão (GRÁCIO; OLIVEIRA, 2013).

A recuperação da frequência de cocitação implica considerar a proximidade da relação, se será na referência (artigo), na seção do artigo, no parágrafo ou nas frases do documento (LIU; CHEN, 2012) e isso indica a necessidade de observar em como os documentos são mencionadas no texto (DING et al., 2013; CARVALHO; CAREGNATO, 2015), já que a lista de referências iguala documentos mais mencionados com os mencionados apenas uma vez ou mesmo nenhuma (CARVALHO; CAREGNATO, 2015).

Há outras questões importantes que merecem destaques, como definir o ponto de corte do número de artigos citantes para a inclusão nos autores (GRÁCIO; OLIVEIRA, 2013), escolher o índice de similaridade para normalização da matriz (LEYDESDORFF, 2005; GRÁCIO; OLIVEIRA, 2013), qual técnica multivariada utilizar, como interpretar os agrupamentos etc.

O presente trabalho foca na relação entre dados tradicionais de uma ACA e a verificação dessa relação na proximidade por seção e parágrafo, portanto na etapa da formação dos pares. Nesse contexto três trabalhos merecem destaque: Jeong, Song e Ding (2014) propuseram ACA baseada no conteúdo das sentenças citantes, e em comparação com uma ACA tradicional, concluíram que a nova abordagem ofereceu mais detalhes de subdomínios; Bu et al (2017) propõem um modelo que não considera o conteúdo como o anterior, mas utiliza metadados extraídos do texto completo, como número de menções ao citado e número de palavras do contexto da frase citante, além do ano de publicação da referência e concluem que os agrupamentos ficam melhor visualizados e com mais detalhe; por fim, Liu e Chen (2012) apresentam um estudo comparando dados empíricos dos quatro níveis de proximidade de cocitação, artigo (referência), seção de artigo, parágrafo e sentença (frase) e concluem que as cocitações no nível na sentença refletem os dados apresentados por cocitações no nível do artigo e isso parece ser uma evidência interessante que valida os estudos tradicionais, mas não escondem suas deficiências.

Vale salientar que a proximidade por frase estará, no contexto desse trabalho, contida na proximidade por parágrafo, principalmente por lidar com um par de autores e por ser o

parágrafo uma unidade do discurso que possui, por definição, um sentido completo (CUNHA, 2010).

Lidar com a forma como os autores constroem seus discursos a partir de outros é um dos desafios de estudos de citações e cocitação no contexto.

3 PROCEDIMENTOS METODOLÓGICOS

O trabalho é cientométrico, exploratório e de natureza metodológica. O *corpus* de análise é formando por 151 artigos de pesquisa, recuperados na base *Web Of Science* no mês de maio de 2017 e publicados entre os anos de 2011 e 2015. A busca foi realizada pelo termo "*Knowledge organization*" no campo "TÓPICO" da base e refinado pela categoria "INFORMATION SCIENCE LIBRARY SCIENCE", pois a intenção é era ter um grupo de artigos minimamente heterogêneo para testar a ideias propostas.

O corpus de análise gerou três bases de dados de interesse para esse trabalho, que foram manipulas com o uso dos *softwares Microsoft Excel* e *SPSS*[1]. A primeira base de dados é composta pelos dados de caracterização dos 151 artigos, que foram publicados em 23 diferentes periódicos, como demonstrado na tabela abaixo.

Tabela 1. Estatísticas descritivas dos indicadores de caracterização do corpus (N=151).

Medidas	Nº de referências	Nº de referências com autoria pessoal	Nº de referências de autoria não pessoa	Nº de autorias	Nº de autores dos artigos
Total	5771	5445	318	9311	289
Média	38,22	36,06	2,11	61,66	1,91
Mediana	34	32	01	49	02
Máximo	120	111	55	213	06
Mínimo	05	04	00	06	01
Desvio padrão	22,50	21,81	5,29	43,00	1,10

Fonte: dados da pesquisa, 2017.

A segunda base de dados consiste em um *ranking* de 5336 autores citados a partir das 9311 autorias, com as seguintes variáveis: nº de citação; nº de artigos citantes, nº de citações de documento de autoria única do autor; nº de citações de documentos produzidos em coautoria indicando a primeira posição e a secundária (segunda posição ou outra).

A terceira base de dados consiste em uma matriz simétrica de cocitação com 50 autores (nove ou mais artigos citantes), que foi normalizada pelo Cosseno de Salton (GRÁCIO; OLIVEIRA, 2015), com valor da diagonal igual a zero. A técnica de contagem dos pares considerou coautoria como cocitação, como Schneider, Larsen e Ingwersen (2009), principalmente por não interferir nos objetivos da pesquisa. Foi realizada uma análise fatorial na matriz normalizada, utilizando a

[1] *Software* estatístico, cuja sigla significa: "Statistical Package for the Social Sciences".

técnica de componentes principais, analisando a matriz de correlação, sem definir número fixo de fatores, rotação *varimax* e suprimindo coeficientes menores que 0,5. A análise gerou cinco fatores apresentados no Quadro 1.

Foi destacado o par de autores Dahlberg e Hjørland para atender os objetivos do trabalho. Esse par de autores tem a maior frequência de cocitação absoluta (23) e o maior índice de similaridade da matriz de cocitação normalizada (0,88), além de estarem no mesmo agrupamento (motivo para os dados passarem pela análise multivariada). Foram identificados os 23 artigos citantes e um quadro foi criado com os dados de citação em cada artigo citante, nº de seções do artigo, pares de cocitação por seção (ocorrência e menção) e parágrafo (menção).

A proximidade por seção como menção leva em consideração o número de vezes que cada autor aparece na seção (entrada mais data), por exemplo, se o autor *A* é mencionado duas vezes e o autor *B* é mencionado três, a contagem de pares é igual a seis, enquanto que na ocorrência conta-se apenas um par e esse indicador terá como valor máximo o número de seções de cada documento.

4 PRINCIPAIS RESULTADOS E DISCUSSÃO

Os resultados iniciais partem dos fatores formados a partir da matriz simétrica (autores *vs* autores), como demonstrado no Quadro 1. A análise demonstra que a variância explicada é alta e as cargas fatoriais parecem significativas, pois apenas dois autores estão com valores abaixo de 0,6.

Quadro 1. Distribuição dos fatores e cargas fatoriais da ACA.

1º Fator (46,40%*)		2º Fator (22,75%*)		3º Fator (13,12%*)		4º Fator (6,96%*)		5º Fator (3,55%*)	
Autores	CF	Autores	CF	Autores	CF	Autores	CF	Autores	CF
López-Huertas, MJ	,953	Ranganathan, SR	,976	Berners-Lee, T	,960	Bowker, GC	,874	Bawden, D	,876
Albrechtsen, H	,919	La Barre, K	,955	Greenberg, J	,935	Star, SL	,862	Chan, LM	,702
Thellefsen, TL	,907	Gnoli, C	,954	Lassila, O	,925	Berman, S	,762		
Thellefsen, MM	,891	Spiteri, LF	,947	Taylor, A	,924	Beghtol, C	,696		
Frohmann, B	,870	Vickery, BC	,941	Hendler, J	,923	Olson, HA	,691		
Guimarães, JAC	,867	Mazzocchi, F	,899	McGuinness, DL	,914	Tennis, JT	,657		
Buckland, MK	,858	Broughton, V	,823	Noy, NF	,910	Foucault, M	,628		
Hansson, J	,846	Soergel, D	,822	Gruber, TR	,906	Furner, J	,581		
Andersen, J	,843	Szostak, R	,800	Smith, B	,714				
Friedman, A	,778	Jacob, EK	,771	Zeng, ML	,637				
Mai, J-E	,754	Kuhn, TS	,760						
DAHLBERG, I	**,715**	Lancaster, FW	,709						
Bates, MJ	,712	Green, R	,590						
Smiraglia, RP	,678								
Hodge, G	,675								
HJØRLAND, B	**,664**								
Svenonius, E	,624								

* Variância explicada de cada fator (Total da variância explicada 92,78%).
CF – Carga fatorial.

Fonte: Dados de pesquisa, 2017.

Os autores Dahlberg e Hjørland estão presentes (com destaque) no primeiro fator, com outros 15 autores. O quadro 2 apresenta alguns dados detalhados dessa relação de cocitação e atende o primeiro objetivo do texto.

A proximidade por seção dos autores ocorre em 20 artigos citantes, enquanto que a proximidade por parágrafo ocorre em 10. Apenas em três artigos citantes a cocitação acontece em mais de uma seção e o total de cocitação na proximidade por seção fica 23.

Quadro 2. Indicadores de proximidade dos autores Dahlberg e Hjørland.

Artigos citantes (título, ano de publicação e número de seções entre parênteses)	Citações		Cocitações		
			Seção*		Parágrafo
	Dahlberg	Hjørland	Ocorrência	Menção	
Concept theory and semiotics in knowledge organization, 2011 (07)	06	07	02	10	01
Semantics, Classifications and Evidence in a Model for Global Catastrophic Risks, 2011 (09)	01	01	01	04	00
Is classification necessary after Google?, 2012 (07)	01	10	01	02	00
Male Homosexuality in Brazilian Indexing Languages, 2012 (04)	01	04	01	04	00
Metadata About What?, 2012 (07)	01	02	00	00	00
The Methodological Influence of Peirce's Pragmatism on KO, 2012 (03)	02	02	00	00	00
Classifications and concepts, 2013 (05)	01	02	01	02	02
Formal Ontology and the Foundation of Knowledge Organization, 2013 (06)	01	03	01	06	01
Information Sciences Methodological Aspects Applied to Ontology Reuse Tools, 2013 (06)	05	05	01	40	00
Nodes and arcs: concept map, semiotics, and KO, 2013 (04)	01	02	01	01	00
Peircean Semiotics and Subject Indexing, 2013 (04)	01	02	01	05	00
Saaty's Analytic Hierarchies Method for KO in Decision Making, 2013 (04)	01	03	01	03	00
The Blind Men and the Elephant, 2013 (05)	02	05	02	08	00
User-based and Cognitive Approaches to Knowledge Organization, 2013 (12)	01	03	00	00	00
Bibliometrics Contribution to the Metatheoretical and Domain Analysis Studies, 2014 (03)	01	03	01	01	01
Retos y oportunidades en organización del conocimiento en la intersección con las tecnologías de la información, 2014 (08)	01	07	01	01	00
Semantic Relations in Knowledge Organization Systems, 2014 (05)	02	02	01	01	01
Classical Databases and Knowledge Organization, 2015 (07)	01	08	01	01	00
Domain Analysis of Domain Analysis for Knowledge Organization, 2015 (04)	02	05	02	07	05
Evaluating the Practical Applicability of Thesaurus-Based Keyphrase Extraction in the Agricultural Domain, 2015 (06)	01	04	01	00	04
KO in the Intersection with Information Technologies, 2015 (07)	01	10	01	01	02
The Ethics of KO and Representation from a Bakhtinian Perspective, 2015 (06)	02	03	01	02	02
Theories are Knowledge Organizing Systems (KOS), 2015 (06)	01	09	01	01	01
Total	37	102	23	100	20
Total = zero (sem cocitação na proximidade, considerando as 23 cocitações das referências)			03	04	13
Total ≠zero (cocitação válida na proximidade, considerando as 23 cocitações das referências)			20	19	10

Fonte: dados da pesquisa, 2017.

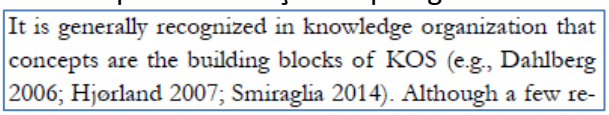
Os dados parecem significativos e demonstram que a cocitação na proximidade por artigo (LIU; CHEN, 2012) ou referência representam a realidade da relação entre os autores. Porém, o número de citação de cada autor nos artigos citantes, para esse caso, indica também que muita informação é perdida e se levar em conta as menções ainda mais, como salientam BU et al (2017) sobre não ser suficientemente informativa a ACA tradicional. O autor Hjørland, por exemplo, é citado 102 vezes nos documentos citantes.

A discussão sobre a relação entre ocorrência e menção é necessária, pois os níveis de proximidade que Liu e Chen (2012) apresentam focam na ocorrência da cocitação e se

apresentam como unidades para estabelecer essa relação. Por exemplo, quando se diz que os autores Dahlberg e Hjørland possuem valor de cocitação igual a 23 na proximidade por seção (quadro 2), diz-se que eles se relacionam nesse número de seções, mas quando esse número é igual a 100 por menção, a unidade de relacionamento se perde. Ou seja, há ainda uma discussão conceitual necessária, mas o nível de proximidade pode ser desconsiderado, como no trabalho de Jeong, Song e Ding (2014).

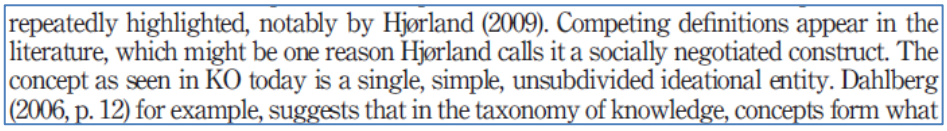
Para atender o segundo objetivo do trabalho foi considerado observar a "distância" entre os autores na proximidade por parágrafo. As figuras abaixo, com dados dos artigos citantes desse trabalho, exemplificam essa diferença.

Figura 1. Exemplo de cocitação no parágrafo sem distância.


It is generally recognized in knowledge organization that concepts are the building blocks of KOS (e.g., Dahlberg 2006; Hjørland 2007; Smiraglia 2014). Although a few re-

Fonte: dados da pesquisa, 2017.

Figura 2. Exemplo de cocitação no parágrafo com distância.


repeatedly highlighted, notably by Hjørland (2009). Competing definitions appear in the literature, which might be one reason Hjørland calls it a socially negotiated construct. The concept as seen in KO today is a single, simple, unsubdivided ideational entity. Dahlberg (2006, p. 12) for example, suggests that in the taxonomy of knowledge, concepts form what

Fonte: dados da pesquisa, 2017.

Na figura 1 é creditada aos três autores a afirmação que antecede as citações (sem analisar o conteúdo), ou seja, podemos supor, nesse contexto, que as citações possuem o mesmo peso, pois não há distância entre as menções. Na figura 2 os autores estão "distantes" e isso indica que o(s) autor(es) citantes elaboraram uma unidade de discurso onde os autores cocitados se contradizem ou se complementam, por exemplo. Das 20 cocitações no nível do parágrafo apresentados no quadro 2, apenas em dois casos não há "distância" entre os autores.

5 CONSIDERAÇÕES FINAIS

Os dados mostram que a cocitação de autores considerando a proximidade por seção e proximidade por parágrafo diminui em relação à proximidade por artigo, se for considerada apenas a ocorrência. Mas considerando as menções o número de pares aumenta e essa abordagem será aplicada nos demais autores do primeiro fator da ACA apresentada nesse texto.

Os estudos de cocitação de autores apresentam diversos desafios e talvez o principal seja entender como os autores citantes determinam essa relação, que parece não ser intencional, como a própria citação.

REFERÊNCIAS

BU, Yu et al. MFTACA: an author co-citation analysis method combined with metadata in full text. In: INTERNATIONAL CONFERENCE ON SCIENTOMETRICS AND INFORMETRICS – ISSI, 16., 2017, Wuhan, China. **Proceedings...** Wuhan: International Society of Scientometrics and Informetrics – ISSI, 2017.

CARVALHO, R. A.; CAREGNATO, S. E. Análise de citação: relação entre referências e menções em artigos. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa – PB. **Anais...** João Pessoa: ANCIB, 2015.

CARVALHO, R. A.; CAREGNATO, S. E. Primeiro *vs* todos os demais autores citados: estudo empírico em artigos na base BRAPCI In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 5., 2016, São Paulo. **Anais...** São Paulo: USP, 2016.

CUNHA, A. G. **Dicionário etimológico da língua portuguesa**. 4. ed. rev. e atual. Rio de Janeiro: Lexikon, 2010. 712p.

DING, Y. et al. The distribution of references across texts: some implications for citation analysis. **Journal of Informetrics**, n. 7, p. 583-592, 2013.

GRÁCIO, M. C. C.; OLIVEIRA, E. F. T. de. Estudos de análise de cocitação de autores: uma abordagem teórico-metodológica para a compreensão de um domínio. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Florianópolis – SC. **Anais...** Florianópolis: ANCIB, 2013.

GRÁCIO, M. C. C.; OLIVEIRA, E. F. T. de. 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, v.25, n.2, p. 105-116, maio/ago. 2015.

JEONG, Y.; SONG, M.; DING, Y. Content-based author co-citation analysis. **Journal of Informetrics**, n.8, p.197-211, 2014.

LEYDESDORFF, L. Similarity measures, author cocitation analysis, and information theory. **Journal of the American Society for Information Science & Technology**, v.56, n.7. 2005.

LIU, S.; CHEN, C. The proximity of co-citation. **Scientometrics**, n.91, p.495-511. 2012.

McCAIN, K. Mapping author intellectual space: a technical overview. **Journal o the American Society for Information Science**. v.41, n.66, p.433-443. 1990.

PERSSON, O. All author citations versus first author citations. **Scientometrics**, v.50, n.2, 2001.

SCHNEIDER, J. W.; LARSEN, B.; INGWERSEN, P. A comparative study of first and all author co-citation counting, and two different matrix generation approaches applied for author co-citation analyses. **Scientometrics**, v. 80, n. 1, p. 105–132, 2009.