8º EBBC
Encontro Brasileiro de Bibliometria e Cientometria
MACEIÓ2022
20, 21 E 22 · JULHO

PROCESSAMENTO DE LINGUAGEM NATURAL E ACOPLAMENTO BIBLIOGRÁFICO:
uma análise da proximidade entre os artigos mais acessados do periódico Scientometrics

Rafael Gutierres Castanha[1]
Bianca Savegnago de Mira[1]

Resumo: Esta pesquisa compara os métodos de Processamento de Linguagem Natural e Acoplamento Bibliográfico normalizados via cosseno de Salton aplicados aos 10 artigos mais acessados de 2020 do periódico Scientometrics. Para tanto, calcula as correlações de Pearson e Spearman, o teste não paramétrico de Wilcoxon e representa os valores normalizados em boxplot. Conclui forte correlação entre textos completos e resumos, e, entre os métodos de Acoplamento Bibliográfico. Entretanto, guarda distinção significativa entre os valores calculados.

Palavras-Chave: Acoplamento bibliográfico. Índice de Similaridade. Processamento de linguagem natural.

1 INTRODUÇÃO

No âmbito dos estudos bibliométricos a proximidade entre dois artigos pode ser aferida por meio da intensidade de acoplamento bibliográfico (AB). Cunhado por Kesller (1963), o método de AB visa agrupar documentos por meio de referências citadas em comum por estes documentos denominadas unidades de acoplamento de modo que, quanto mais unidades de acoplamento em comum entre dois documentos, maior a proximidade (intesidade de acoplamento) entre eles. Kesller (1963) sugere dois critérios de agrupamento (A e B), em que o primeiro analisa quantos documentos se acoplam a um determinado documento e o segundo prevê identificar um grupo de documentos em que todos se acoplam entre si.

Se dois artigos estão acoplados bibliograficamente, ambos possuem ao menos uma referência em comum (citam uma mesma referência). A ideia básica de acoplamento bibliográfico, do ponto de vista matemático, nada mais é do que uma intersecção entre listas de referências de documentos. Porém, do ponto de vista relacional de citações, dois documentos acoplados

[1] Universidade Estadual Paulista (UNESP)

remetem a alguma proximidade científica teórica e/ou metodológica que estes documentos partilham.

Contudo, métodos baseados em processamento de linguagem natural (PLN) podem ser viáveis para este tipo de análise visto que estão ancorados no conhecimento estatístico e possibilitam verificar a presença ou não de uma intersecção entre o conteúdo textual dos documentos. O Processamento de Linguagem Natural (PLN) é definido como um arcabouço técnico computacional com direção teórica orientada para análise e representação de textos cujo objetivo é atingir um processamento de linguagem próximo ao humano aplicado a múltiplas funções ou aplicativos (LIDDY, 2010).

O surgimento do PLN data da década de 1950, o método consistia em uma intersecção entre a inteligência artificial e a linguística. A técnica passou por uma reorientação na década de 1980 com maior aporte estatístico e a utilização do *Machine-learning*. Ambos possibilitam o desenvolvimento de algoritmos que permitem que um programa seja capaz de inferir padrões sobre dados. Com isso as análises tornaram-se mais simples, robustas e rigorosas. (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

Diante do exposto, esta pesquisa compara o desempenho entre o processamento de linguagem natural e o acoplamento bibliográfico, normalizados via cosseno de Salton, para avaliar a proximidade entre o conjunto dos *top* 10 artigos mais acessados do periódico Scientometrics no ano de 2020. O periódico situa-se entre os principais da área de Ciência da Informação. Fundada em 1978 pelo renomeado pesquisador Tibor Braun com foco em análises cientométricas, a Scientometrics reúne mais de 127 volumes. Atualmente é editorada por Wolfgang Glänzel, possui fator de impacto igual a 3,238 (2020) e em 2020 registrou mais de 1 milhão de *downloads* de seus documentos.

2 METODOLOGIA

A fim de comparar a proximidade entre os 10 artigos mais baixados da Scientometrics para avaliar o comportanto de métodos de PLN e no acoplamento bibliográfico e identificar possível independência ou correlação entre eles utilizou-se três métodos baseados PLN e dois baseados no acoplamento bibliográfico. Os métodos de PLN calcularam a proximidade entre os documentos tomando como unidades: i) texto completo (apenas formas ativas; exclusa formas suplementares) ii) resumos (apenas formas ativas; exclusa formas suplementares); iii) palavras-chave. Equanto os métodos baseados no acoplamento bibliográfico utilizaram como

unidades de acoplamento: i) documentos citados (AB documentos); ii) autores citados (AB autores). O primeiro caso refere-se a citar exatamente os mesmo documentos e o segundo, considera-se a obra do autor citado como única e se computou quantos autores em comum os documentos citaram.

As cinco perspectivas foram cálculadas da mesma maneira: computou-se a quantidade de unidades de cada documento, a quantidade de elementos (palavras ativas ou unidades de acoplamento) em comum entre cada par de documentos e em seguida, normalizou-se este valor via Coseno de Salton. Dessa maneira, tem-se, 45 cálculos para cada método. Os textos foram recuperados no portal eletrônico da Scientometrics e submetidos ao processamento no *software* IRAMUTEQ (versão 3.5.1). O IRAMUTEQ é um *software* livre integrado ao ambiente estatístico do *software* R especializado em dados textuais (CAMARGO; JUSTO, 2013). Os textos de cada um dos 10 artigos foram processados individualmente utilizando o dicionário padrão de língua inglesa disponível no *software* e as definições padrão já existentes para a lematização e parametragem. Na lematização as palavras são reduzidas às suas raízes, sendo exclusas características como o tempo verbal e o plural. Os textos são transformados em segmentos textos e as frequências contabilizadas. A inserção dos parâmetros é feita a partir das classes (adjetivos, advérbios, substantivos, etc.) em que se define quais palavras serão consideradas ativas, suplementares ou eliminadas. De acordo com a definição padrão permaneceram como formas ativas as classes dos adjetivos, advérbios, formas não reconhecidas, substantivos comuns e verbos. As formas suplementares, que foram exclusas da análise, são as classes dos: adjetivos dos tipos indefinido, numérico, possessivo e suplementar; advérbios suplementares; artigos; auxiliares; caracteres numéricos; conjunções; onomatopéias; pronomes; substantivos e verbos do tipo suplementar. Para extração da listas de referências (documentos e autores) dos 10 artigos, recuperou-se cada um deles na base de dados Scopus e utilizou-se a ferramenta de extração de autores citados do *software VosViewer.*

Para os cálculos, foi utilizado o código em linguagem R para cálculo de acoplamento bibliográfico de Castanha (2021). O código compara a quantidade de itens em comum, par a par, e fornece a normalização via Cosseno de Salton (CS), em que:

$$CS = \frac{Doc_i \cap Doc_j}{\sqrt{d_i \times d_j}}$$

Em que Doc_i e Doc_j representam os documentos i e j a serem comparados e, d_i e d_j a quantidade de unidades em cada documento a serem analisadas. Para os métodos de PLN, $Doc_i \cap Doc_j$ indicam a quantidade de palavras em comum entre dois documentos e, d_i e d_j denotam a quantidade de palavras totais em ambos documentos (i e j). Com relação aos cálculos de AB, a intersecção $Doc_i \cap Doc_j$ representa a intensidade de acoplamento bibliográfico em que d_i e d_j são unidades de acoplamento. No primeiro caso (AB documentos), a quantidade de documentos citados em por i e j e no segundo caso (AB autores), a quantidade de autores citados pelos documentos i e j.

Posteriormente aos cálculos de proximidade normalizados, submeteu-se os métodos a correlações, de Pearson (r) e Spearman (ρ) (Tabela 1), seguido do teste de Wilcoxon para amostras pareadas (a nível 5%) a fim de de identificar possível associação e diferenças significativas (ou não) entre as medidas. Este procedimento estatístico é inspirado em Lariviere e Gingras (2011) em que os autores comparam diferentes métodos de normalização de citação por campo. Assim, quanto mais forte a correlação entre dois métodos, maior a capacidade de os métodos captarem de maneira similar as proximidades analisadas, e, em caso de diferenças significativas entre os métodos ($p\text{-}valor<0,05$), o teste de Wilcoxon apontará que existe diferença significativa entre os valores calculados. É importante apontar que correlação fraca (não significativa) não implica necessariamente em diferenças significativas entre os métodos, e vice-versa. Para execução dos calculos de correlação e do teste de Wilcoxon utilizou-se o *software* Jamovi.

3 RESULTADOS

Após as extrações supracitadas foram calculadas as correlações entre os cinco métodos propostos como apresenta o Tabela 1. Dos 10 artigos recuperados dois não possuem palavras-chave, sendo que um deles também não possuia resumo. Contudo, optou-se por mantê-los na amostra por serem passíveis de processamento via texto completo e dos dois tipos de acoplamento, por possuirem referências citadas. Ainda, não houve nenhuma palavra-chave em comum na comparação entre 10 artigos, resultando em 45 cálculos normalizados iguais a zero, com isso, o cálculo de proximidades, por assumir valor constante (igual a zero), foi excluído do cálculo de correlação junto aos demais métodos.

Tabela 1 – Matriz de correlação entre as medidas

	Texto Completo	Resumos	AB (documentos)	AB (autores)
Texto completo	1	0,77*	0,295	0,431*
Resumos	0,799*	1	0,327*	0,219
AB (documentos)	0,234	0,326*	1	0,536*
AB (autores)	0,365*	0,265	0,75*	1

Fonte: elaboração própria. *signficativo a nível de 5%.

A Tabela 1 apresenta, na parte triangular superior os cálculos da correlação do ρ de Spearman e na parte triangular inferior, os cálculos do *r* de Pearson. É possível observar forte correlação somente entre texto completo e Resumos, ambos métodos oriundos do PLN. Tal fato pode estar associado a similaridade entre os conteúdos, uma vez que os resumos são elaborados a partir do texto completo. Os resumos condensam as ideias expressas no texto completo e são uma forma de apresentar o artigo aos leitores. Assim, sua construção já está condicionada ao conteúdo do texto completo, essa condição pré-existente suscita a forte correlação obtida.

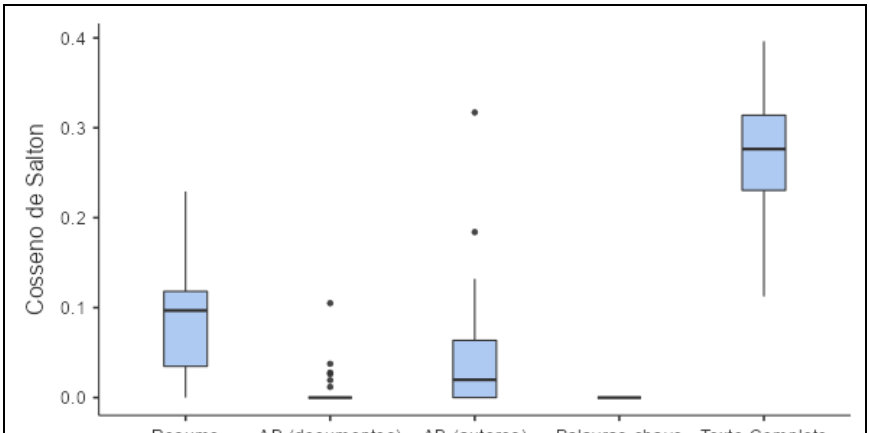
Ainda, é possível observar correlação moderada e forte (Pearson e Spearman) entre os cálculos de acoplamento ao considerar como unidades de acoplamento documentos e autores. Fato é que, se dois artigos estão acoplados por determinado documento, estes artigos estão obrigatoriamente acoplados pelos autores (autores acopladores) destes documentos acopladores. Dessa maneira, cálculos de acoplamento bibliográfico que admitem a obra do autor citado como única enquanto unidade de acoplamento, obrigatoriamente assumirão valores maiores, ou no mínimo iguais, a cálculos que admitem artigos enquanto unidade de acoplamento. Com isso, é possível inferir que $AB\ (autores) \geq AB\ (documentos)$ e assumir a condição lógica de implicação: dado um conjuto de listas de referências (*R)* a serem acopladas, então, $R: \nexists AB\ (documentos) \rightarrow \nexists AB\ (autores)$. Neste caso, a recíproca não é válida, pois, se dois documentos se acoplam por determinados autores, não necessariamente se acoplam por um mesmo documento. O desempenho dos valores de AB (autores) maiores que os de AB (documentos) é visto na representação de *boxplot* da Figura 1 em que os valores dos quartis expressos nas caixas do AB (autores) são maiores que os do AB (documentos).

A Tabela 1 apresenta ainda correlações significantivas a nível de 5% entre texto completo e AB (autores), e, Resumos e AB (documentos). Contudo, é notável que estas correlações não são fortes. Tal falto pode estar relacionado aos valores normalizados de intensidade dos métodos, visto que, AB (documentos) apresenta valores menores que AB (autores) e Resumos apresenta valores menores que texto completo, justamente os métodos que atingem menores

valores, apresentam correlações signficativas. Os valores calculados, e normalizados, a partir de cada método são apresentados no gráfico de *boxplot* na Figura 1.

A Figura 1 apresenta, por meio do *boxplot*, a distribuição segundo os quartis de cada método analisado juntamente com os valores médios e medianos de cada conjunto de dados. Constata-se que os valores médios e medianos dos cálculos de proximidade utilizando o texto completo e os Resumos são maiores que os demais.

Figura 1 – *Boxplot* dos cinco métodos análisados normalizados via cosseno de Salton



	Resumo	AB (documentos)	AB (autores)	Palavras-chave	Texto Completo
Média	0,0797	0,0051	0,0435	0	0,2684
Mediana	0,0968	0	0,0198	0	0,2764

Fonte: elaboração própria.

O maior desempenho de medidas oriundas do PLN possivelmente está relacionado ao fato de que os textos, sejam eles resumos ou completos, são provenientes de um mesmo domínio. A revista Scientometrics possui um escopo bem delimitado com conteúdo voltado à cientometria. Além disso nota-se que alguns artigos abordam as mesmas temáticas. Por exemplo, dentre os dez, dois analisam a produção científica sobre COVID-19 e outros dois versam sobre questões relacionadas à autocitação.

Com relação aos cálculos de acoplamento, é possível ilustrar o superior desempenho ao da obra do autor citado como única enquanto unidade de acoplamento, em relação ao AB (documentos). Equanto AB (autores) assume valores médios e medianos de 0,0435 e 0,0198 (4,35% e 1,98%) de proximidade entre os documentos, AB (documentos) possui 0,005 (0,5%) de proximidade média entre os documentos.

A fim de observar se há diferenças significativas entre os valores de proximidade calculados segundo os diferentes métodos, aplicou-se o teste de Wilcoxon para amostras pareadas entre

os métodos dois a dois, resultando em 8 cálculos. O teste aferiu diferenças significativas (*p-valor<0,05*) em todas oito comparações. Dessa forma, é possível apontar que, mesmo que haja correlaçõs significativas entre métodos, seja baseado em PLN, seja no acoplamento bibliográfico, nenhum método concordou entre si justamente por assumirem valores (normalizados) estatisticamente diferentes entre si. Tal fato suscita que os métodos guardam especificidades entre si, a pesar de possível correlação, em que texto completo desempenha maiores valores, AB (documentos) menores (mas não totalmente nulos) e palavras-chave, completamente nulos.

4 CONSIDERAÇÕES FINAIS

Ao comparar os métodos baseados no PLN e no AB observou-se que as especificidades de cada método influenciaram significativamente na obtenção de correlação forte entre as medidas advindas deles. Os cálculos de acoplamento correlacionaram-se de maneira significativa devido a implicação entre AB (documentos) e AB (autores) em que para cada valor de AB (documentos) há necessariamente um valor de AB (autores). Com relação aos cálculos baseados em PLN foi constatada forte correlação entre textos completos e resumos, visto que há uma dependência de conteúdo entre ambos, pois, o resumo é elaborado a partir do texto.

Por fim, ao comparar o desempenho das medidas normalizadas é possível apontar que, em relação aos valores médios e medianos, os métodos baseados em PLN assumem valores superiores ao calculados via AB com: textos completos > Resumos > AB (autores) > AB (documentos) > Palavras-chaves.

REFERÊNCIAS

CASTANHA, R. G. **rafaelcastanha/The-Bibliographic-Coupler**: The Bibliographic Coupler v1.1.1. trans. by . 24 Jan. 2022. Disponível em: https://doi.org/10.5281/zenodo.5899142. Acesso em: 22 jan. 2022.

CAMARGO, B. V.; JUSTO, A. M. IRAMUTEQ: um software gratuito para análise de dados textuais. **Temas em psicologia**, [*S.l.*], v. 21, n. 2, p. 513-518, 2013.

LARIVIÈRE, V.; GINGRAS, Y. Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. **Journal of informetrics**, Amsterdam, v. 5, n. 3, p. 392-399, 2011.

LIDDY, E. D. Natural Language Processing for Information Retrieval. *In*: BATES, M. J.; MAACK, M. N. (ed.). **Encyclopedia of Library and Information Sciences**. Boca Raton: CRC Press, 2010. Disponível em: https://doi.org/10.1081/E-ELIS3. Acesso em: 31 jan. 2022.

NADKARNI, Prakash M.; OHNO-MACHADO, Lucila; CHAPMAN, Wendy W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, Oxford, v. 18, n. 5, p. 544-551, 2011.

KESSLER, Maxwell Mirton. Bibliographic coupling between scientific papers. **American Documentation**, Hoboken, v. 14, n. 1, p. 10-25, 1963.

SCIENTOMETRICS: an international journal for all quantitative aspects of the science of science, communication in science and science policy. **Top 10 articles 2020 by full-textdownloads!** 2020. Disponível em: https://www.springer.com/journal/11192/updates/18879904. Acesso em: 20 jan. 2022.