DOI: 10.5433/1981-8920.2022v27n3p262

PROCESSAMENTO DE LINGUAGEM NATURAL E ACOPLAMENTO BIBLIOGRÁFICO: UMA ANÁLISE DA PROXIMIDADE ENTRE OS ARTIGOS MAIS ACESSADOS DO PERIÓDICO *SCIENTOMETRICS*

NATURAL LANGUAGE PROCESSING AND BIBLIOGRAPHIC COUPLING: AN ANALYSIS OF THE PROXIMITY BETWEEN THE MOST ACCESSED ARTICLES OF THE SCIENTOMETRICS JOURNAL

Bianca Savegnago de Mira[a]
Rafael Gutierres Castanha[b]

RESUMO

Objetivo: compara os métodos de Processamento de Linguagem Natural e Acoplamento Bibliográfico normalizados via Cosseno de Salton aplicados aos dez artigos mais acessados de 2020 do periódico Scientometrics. **Metodologia:** Calcula a similaridade entre todos os artigos segundo cinco perspectivas, sendo elas: similaridades entre formas ativas do texto completo, formas ativas dos resumos, palavravas-chaves em comum, acoplamento bibliográfico entre documentos e acoplamento bibliográfico de autores. Ademais, calcula as correlações de Pearson e Spearman, aplica o teste não [S. lparamétrico de Wilcoxon a nível de 5% de significância e representa os valores normalizados em boxplot. **Resultados:** Constata que as especificidades de cada método influenciam significativamente na obtenção de correlação significativa entre as medidas em que os dois cálculos de acoplamento se correlacionariam de maneira mais forte entre si, assim como dois cálculos baseados no processamento de linguagem natural. Observa que os cálculos de acoplamento, correlacionaram-se de maneira significativo, pois, para cada valor de acoplamento de documentos há necessariamente, ao menos um valor de acoplamento de autores. Com relação aos cálculos baseados no processamento de linguagem natural, verifica forte correlação entre textos completos e resumos, visto que há uma dependência de conteúdo entre ambos. O teste de Wilcoxon, aferiu diferenças significativas entre todos os pares de medidas comparadas. **Conclusão:** Conclui forte correlação entre textos completos e resumos, e, entre os métodos de acoplamento bibliográfico. Entretanto, guarda distinção significativa entre os valores calculados.

Descritores: Acoplamento bibliográfico. Índice de Similaridade. Processamento de

[a] Doutoranda em Ciência da Informação pelo Programa de Pós-graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho (PPGCI/Unesp), Marília, Brasil. E-mail: bianca.mira@unesp.br

[b] Doutor em Ciência da Informação pela Universidade Estadual Paulista (Unesp). Docente da Universidade de Marília (UNIMAR), Marília, Brasil. E-mail: r.castanha@gmail.com

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 Keslller (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 (intensidade de acoplamento) entre eles.

Keslller (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, em que a cardinalidade desta intersecção pode ser descrita como frequência (ou força) de acoplamento bibliográfico, e as referências responsáveis por conectar dois ou mais documentos, unidades de acoplamento.

Décadas após da idealização do acoplamento bibliográfico, Zhao e Strotmann (2008), propuseram o acoplamento bibliográfico entre autores (ABA), de modo que, ao invés de acoplar documentos, fossem acoplados pesquisadores, tomando como unidade de acoplamento documentos citados em comum por dois pesquisadores, ou autores citados em comum, por dois pesquisadores.

O surgimento do ABA resolvera o problema da estaticidade da medida do AB. Isto é, a frequência de acoplamento bibliográfico entre dois documentos, tomando como unidades de acoplamento as referências citadas por ambos, é necessariamente estática e imutável, visto que, uma vez que dois artigos são publicados, estes artigos não citam novas referências. Logo, a frequência de AB entre dois artigos, será sempre a mesma. Diferente do ABA, em que, basta que

os pesquisadores continuem a publicar, que suas respectivas listas de referentes serão incrementadas, e assim, a frequência do ABA configura-se como uma medida dinâmica.

Do ponto de vista relacional de citações, dois documentos ou pesquisadores acoplados, independentemente da unidade de acoplamento, remetem a alguma proximidade científica teórica e/ou metodológica que estes documentos partilham. Mesmo que os dois marcos teóricos dos estudos voltados ao acoplamento bibliográfico prevejam cálculos de frequência entre documentos ou pesquisadores, tomando como unidades de acoplamento referências ou autores, matematicamente, é possível estender este conceito, enquanto análise de similaridade para quaisquer tipos de unidades, como por exemplo, acoplar periódicos ou instituições, ou utilizar-se de palavras-chaves como unidade de acoplamento.

Nesse sentido, 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. Segundo Liddy (2010) o 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.

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 possibilitaram 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).

A combinação dos métodos de PLN e análises bibliométricas têm se tornado mais frequentes nos últimos anos. É possível categorizar os estudos que realizam essa combinação em dois grupos, um em quem as ferramentas de PLN funcionam como meio de aumentar o desempenho de estudos bibliométricos; e

outro em que se analisa artigos de PLN utilizando métodos bibliométricos (TASKIN; AL, 2019).

A fim de investigar se os métodos de AB e PLN haviam sido combinados em outros trabalhos realizou-se uma pesquisa nas bases *Scopus* e *Web of Science* (*WoS*) com o seguinte *query*: *"bibliographic coupling" AND "natural language processing"*. A busca foi feita no dia 27 de dezembro de 2022 e foram exibidos 5 resultados, 1 obtido pela *WoS* e 4 pela *Scopus*. O artigo exibido pela *WoS* também estava indexado na *Scopus*, assim, somam-se 4 documentos distintos pela busca.

O primeiro resultado da busca, ordenada a partir da data de publicação, é o artigo de Yun, Ahn e Lee (2020) publicado pelo *Journal of Informetrics*. Esse artigo está indexado tanto na *Scopus* quanto na *WoS*. Nele os autores propõem um método alternativo para identificação de frentes de pesquisa que se baseia em análises relacionais e compara os agrupamentos obtidos com o novo método e agrupamentos realizados com classificação baseada no PLN.

Publicado pela *Springer Handbooks*, o capítulo de livro de Thijs (2019) é o segundo resultado obtido. O foco do estudo são as diferentes técnicas e metodologias utilizadas em bibliometria para mapear a estrutura cognitiva da ciência e o uso de técnicas de agrupamento para detecção de tópicos e tópicos emergentes. Tradicionalmente os mapeamentos são construídos por meio de links com citações ou abordagens textuais, mas sob a perspectiva metodológica são alicerçados nos desenvolvimentos da ciência de rede.

O artigo de Zhang *et al.* (2016) é o terceiro resultado, também publicado pelo *Journal of Informetrics*, nele os autores abordam indicadores bibliométricos que são utilizados para determinar medidas de similaridade, um desses indicadores é o AB. Seu objetivo é a construção de um método híbrido de medida de similaridade baseado em múltiplos indicadores para analisar portfólios de patentes. Para o método são propostos dois modelos de similaridade, um categórico, a partir de classificações em patentes e outro semântico que utiliza elementos textuais.

O quarto e último resultado é um capítulo da *CEUR Workshop Proceedings* escrito por Thijs, Glänzel e Meyer (2015) e trata da proposição de

um método baseado na extração de frases nominais usando PLN para melhorar a medição do componente lexical. Os autores realizaram comparações com o método de AB e a detecção de comunidades.

Os quatro estudos demonstram diferentes abordagens metodológicas que combinam ou comparam os métodos de AB e PLN. A quantidade reduzida de trabalhos encontrados que aliam as duas temáticas demonstra a necessidade de expandir estudos que combinem ou comparem ambos.

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 3,801 (2021) e em 2020 registrou mais de 1 milhão de downloads de seus documentos.

2 SIMILARIDADE, ACOPLAMENTO BIBLIOGRÁFICO E PROCESSAMENTO DE LINGUAGEM NATURAL

Analisar a similaridade existente entre conjuntos de dados nada mais é do que detectar aquilo que há em comum, ou seja, se observa a sobreposição de conjuntos com a intenção de quantificar e identificar o que há em comum. Matematicamente esta sobreposição representa a intersecção entre conjuntos.

A sobreposição de conjuntos, sob a perspectiva dos estudos métricos da informação, refere-se diretamente às análises relacionais ou de ligação. Indicadores de ligação, também chamados de indicadores de relação, são medidas (ou frequências) de ocorrência e/ou coocorrência examinadas em autorias, citações, palavras, entre outras formas de relação exibidas na produção científica. Esses indicadores são utilizados na identificação, visualização e análise das ligações verificadas na geração do conhecimento (PRADO; CASTANHA, 2020).

Os indicadores de ligação são aplicáveis nos diferentes níveis de agregação nos estudos sobre coautoria e coocorrência de palavras-chaves, além das análises relacionais de citação, cocitação e acoplamento bibliográfico (AB).

O AB, em especial, é um método concebido para a identificação de agrupamentos de artigos. A utilização do método de AB está direcionada para a detecção de similaridades entre as referências citadas em comum de dois ou mais trabalhos. As referências, por sua vez, são reflexos das citações feitas durante o texto. Basta um único item de referência comum entre dois artigos para que sejam considerados acoplados entre si, de modo que item de referência compartilhada entre artigos é definido como uma unidade de acoplamento (KESSLER, 1963).

A força de acoplamento é calculada a partir da quantidade de documentos iguais na lista de referências. Reforça-se que um aspecto importante do AB entre dois artigos é que ele não é passível de alteração com o passar do tempo (MARSHAKOVA, 1981).

Entretanto, Zhao e Strotmann (2008), ao introduzirem o conceito de acoplamento bibliográfico entre autores (ABA), pontuam que a diferença entre o AB aplicado a artigos e ABA e o AB proposto por Kessler (1963), é que o AB não é passível de alteração, visto que a lista de referência de artigos permanece a mesma, enquanto o ABA é passível de alteração, pois, se os pesquisadores ao serem acoplados continuarem a publicar, novos referentes serão inseridos em suas obras, tanto artigos quanto autores. Segundo Zhao e Strotmann (2021) os autores são a representação das escolas de pensamento, enquanto os artigos que são redigidos por eles representam evidências ou descobertas individuais sobre conceitos, teorias e/ou métodos

O método de acoplamento bibliográfico, seja ele estabelecido entre documentos ou autores, exprime a relação entre elementos citantes, isto é, conecta duas unidades citantes, e é baseado na análise relacional de citação, em que as análises relacionais aparecem como o método mais utilizado para identificar a estrutura de conhecimento e a evolução dinâmica de uma especialidade (HOU; YANG; CHEN, 2018).

A análise relacional de citação proporciona maior conhecimento a respeito da proximidade, vizinhança, associação e interlocução entre documentos e pesquisadores. São esses os aspectos que compõem as relações de conectividade teórico-metodológica de um domínio, uma vez que estas relações são reconhecidas pela comunidade científica (GRÁCIO, 2016).

A identificação e análise das relações que baseiam a estrutura teórico-metodológica de um campo científico torna-se possível por meio das análises relacionais de citação. Estas análises apoiam-se na observação simultânea de unidades de análise (autores, documentos, periódicos e outros). Dessa maneira, também é possível verificar as proximidades, a vizinhança, as associações e interlocuções feitas entre as publicações científicas e seus autores (GRÁCIO, 2020).

Assim, as análises relacionais geram, do ponto de vista bibliométrico, dois resultados importantes: a quantificação das frequências (de citação, acoplamento ou cocitação) e a identificação dos itens citados, citados em comum ou citados concomitantemente (cocitados). Sendo que esta identificação é de extrema utilidade para avaliar a estrutura intelectual do domínio analisado.

Para Hjørland (2013) a suposição de que existem relações de assunto e relações semânticas entre documentos citantes e citados é o que sustenta a funcionalidade do uso de referências bibliográficas e índices de citação para a recuperação de um determinado assunto. Não há como determinar essas relações apenas com base nos títulos ou pela medida de similaridade por meio de coocorrências de palavras entre artigos citantes e citados, é necessário que a medida seja validada a partir de outros métodos.

Sob a perspectiva epistemológica, as unidades de análise não são apenas semelhantes, existem semelhanças e diferenças a depender das particularidades de cada documento. As relações, sejam elas semânticas, de citação ou de assunto configuram tipos diferentes. Especialmente as relações de citação são indicações indiretas de relacionamentos semântico e de assunto (HJØRLAND, 2013).

Nesse sentido, para além das análises relacionais de citação, tem-se os métodos baseados em processamento de linguagem natural (PLN) para a

detecção de similaridades. No entanto, enquanto as análises relacionais de citação subsidiam-se em similaridades e agrupamentos a partir de referentes teóricos (documentos, autores, entre outros), o segundo evidencia similaridades em conteúdos textuais.

O PLN compreende um conjunto de técnicas computacionais que derivam de teorias para a análise automática e representação da linguagem humana, a partir disso é possível executar em computadores tarefas relacionadas à linguagem natural em todos os níveis. A evolução da pesquisa em PNL avançou de cartões no formato físico cujo processamento levava cerca de 7 minutos por frase para o formato digital no qual grandes quantidades de dados podem ser processadas em questão de segundos (YOUNG *et al*., 2018).

Na década de 1990 o PLN passou por um processo intensivo de transformação que havia iniciado na década anterior. Este processo beneficiava-se de modelos que foram construídos a partir de grandes quantidades de dados empíricos da linguagem. Um dos primeiros avanços notáveis do uso de *big data*, antes mesmo do *machine learning* tornar-se amplamente conhecido, foi o aprofundamento estatístico do PLN (HIRSCHBERG; MANNING, 2015).

Inúmeras ferramentas e técnicas foram desenvolvidas com o objetivo de que os sistemas de computador sejam capazes de entender e manipular as linguagens naturais. O PLN está fundamentado em diversas disciplinas, a saber, ciência da computação, ciência da informação, linguística, matemática, engenharia elétrica, engenharia eletrônica, inteligência artificial, robótica e psicologia (CHOWDHURY, 2003).

Em Falcão, Lopes e Souza (2022), os autores realizaram uma revisão sistemática da literatura em PLN em artigos científicos na base de periódicos da CAPES no período de 2010 a 2020 com enfoque em redes neurais. Entre os 115 artigos analisados apenas dois correspondiam à área da Ciência da Informação (CI). Outro ponto observado foi o uso de termos próprios da CI pela Ciência da Computação para desenvolver e melhorar recursos computacionais e que autores da Ciência da Computação têm publicado frequentemente em periódicos voltados à CI (CHANG, 2018 *apud* FALCÃO; LOPES; SOUZA, 2022). Ainda de acordo com Falcão, Lopes e Souza (2022) o uso do PLN pode aperfeiçoar

atividades de CI e evitar a geração de ilhas isoladas de informação.

Outro estudo que aborda o uso do PLN na CI é o de Puerta-Díaz *et al.* (2021). Neste caso realizou-se uma análise dos artigos científicos da base *Web of Science* no período de 2000 a 2019 para identificar quais estariam alinhados aos estudos métricos de informação (EMI). Dos 552 artigos de PNL recuperados dentro da categoria *Information Science Library Science*, 31 correspondiam aos EMI. Os autores concluem que o PLN aplicado aos EMI é utilizado como forma de melhorar o desempenho das pesquisas na disciplina, o método surge com maior frequência a partir da década de 2010 e apresentou crescimento nos últimos 3 anos analisados.

A partir dos estudos mencionados nota-se que o PLN se utiliza de conhecimentos forjados na CI e que inclusive têm atraído pesquisadores de outras áreas para seus periódicos; no entanto, a área da CI, apesar de possuir grande potencial para desenvolver pesquisas com o PLN, apresenta uma quantidade pequena, mas crescente de estudos ano a ano.

3 METODOLOGIA

A fim de comparar a proximidade entre os 10 artigos mais baixados da *Scientometrics* para avaliar o comportamento de métodos de PLN e de AB e identificar possível independência ou correlação entre eles utilizou-se três métodos baseados PLN e dois baseados no acoplamento bibliográfico.

Os 10 artigos, na ordem apresentada pela *Scientometrics*, são: d_1*: Characteristics of scientific articles on COVID-19 published during the initial 3 months of the pandemic* de Girolamo e Reynders (2020) em que são avaliados os tipos de artigos que foram publicados nos 3 primeiros meses da pandemia de COVID-19. Na segunda posição da lista está d_2: *Researchers publishing monographs are more productive and more local-oriented* de Kulczycki e Korytjowski (2020), nele os autores fazem uma análise de monografias de diferentes áreas do conhecimento produzidas por autores poloneses orientada por gênero e senioridade.

O terceiro artigo, d_3, tem como título: *Measuring originality in Science* e

foi escrito pelos autores Shibayama e Wang (2020) que propõem uma medida de originalidade para os artigos baseada na análise de rede de citações. De autoria de Kwiek (2020) o quarto artigo (d_4) chama-se *Internationalists and locals: international research collaboration in a resource-poor system* trabalha com análise de colaboração científica internacional a partir de dados coletados via questionário com pesquisadores poloneses. O quinto artigo, d_5, *Retracted COVID-19 articles: a side-effect of the hot race to publication* de Soltani e Patini (2020), analisa retratações associadas a artigos sobre COVID-19.

Em *Tracking self-citations in academic publishing* de Kacem, Flatt e Mayr (2020), sexto artigo da lista (d_6), os autores analisam mais de 3 milhões de citações extraídas da WoS para avaliar autocitações de acordo com a área temática. O sétimo artigo (d_7) é de autoria de Rogers, Szomszor e Adams (2020) *intitulado de Sample size in bibliometric analysis*, avalia o tamanho de amostras utilizadas em análises bibiométricas a partir de dados da WoS. Na oitava posição, (d_8) está uma carta ao editor sobre a aplicação do Teorema de Thomas à avaliação de pesquisas de rankings universitários, publicada por Bornmann e Marx (2020) chama-se *Thomas theorem in research evaluation*.

O nono artigo, d_9, tem como título *How much is too much? The difference between research influence and self-citation excess* e foi escrito por Szomszor, Pendlebury e Adams (2020), os autores analisam as autocitações de autores com publicações de alto impacto coletadas na WoS. O décimo e último artigo (d_{10}) é de autoria de Fages (2020) e é intitulado *Write better, publish better*, nele o autor analisa artigos da área de Economia sob a perspectiva da qualidade da escrita e da expressividade dos periódicos em que são publicados.

Apresentados os dez artigos selecionados para as análises, os métodos de PLN calcularam a proximidade entre estes documentos tomando como unidades: i) texto completo (apenas formas ativas; exclusas formas suplementares) ii) resumos (apenas formas ativas; exclusas formas suplementares); iii) palavras-chave. Enquanto 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 mesmos 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 calculadas 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 Cosseno 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).

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 parametrização. 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; onomatopeias; pronomes; substantivos e verbos do tipo suplementar. Para extração das listas de referências (documentos e autores) dos 10 artigos, recuperou-se cada um deles na base de dados *Scopus* e utilizou-se as opções *export selected authors* e *export selected cited* references para extração de autores e referencias citadas do *software* VosViewer.

Para os cálculos, foi utilizado o *web-app* para cálculos de acoplamento bibliográfico *The Coupler* de Castanha (2022). A ferramenta compara a quantidade de itens em comum, par a par, e fornece a normalização via Cosseno de Salton (CS), em que:

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

Neste caso, Doc_i e Doc_j representam os documentos *i* e *j* a serem comparados, di 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, di e d_j denotam a quantidade de palavras totais em ambos os 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 por *i* e *j* e no segundo caso (AB autores), a quantidade de autores citados pelos documentos *i* e *j*. Todo corpus de análise, é descrito na Tabela 1.

Tabela 1 – Descrição do corpus de análise

d_i	Texto completo	Resumos	Palavras-Chaves	AB (documentos)	AB (autores)
d_1	644	85	7	16	49
d_2	439	48	5	23	41
d_3	664	56	4	59	108
d_4	1398	79	7	98	126
d_5	255	0	0	28	117
d_6	418	80	0	12	37
d_7	681	87	5	22	34
d_8	238	39	0	12	18
d_9	983	79	6	121	196
d_{10}	354	32	5	14	33

Fonte: elaboração própria.

A Tabela 1 indica a quantidade total de itens a serem submetidos aos cálculos de similaridade, seja via PLN ou acoplamento bibliográfico. Com isso, a segunda e terceira coluna da Tabela 1 representam as quantidades de palavras em sua forma ativa por artigo analisado (d_1 à d_{10}) presentes no texto completo e nos resumos, respectivamente. Já a quarta coluna, representa a quantidade de palavras-chave por artigo. As duas últimas colunas (quinta e sexta) representam, respectivamente, o tamanho da lista de referências de cada artigo e quantidade de autores citados por artigo.

Posteriormente aos cálculos de proximidade normalizados, submeteu-se os métodos a correlações, de Pearson (*r*) e Spearman (ρ) (Tabela 2), seguido do teste de Wilcoxon para amostras pareadas (a nível 5%) a fim 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-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 cálculos de correlação e do teste de Wilcoxon utilizou-se o software Jamovi.

4 RESULTADOS

Após as extrações supracitadas foram computados todos cálculos de similaridade (Apêndice 1) e posteriormente as correlações entre os cinco métodos propostos como apresenta o Tabela 2. Dos 10 artigos recuperados 3 não possuem palavras-chave, sendo que um deles também não possuía resumo. Contudo, optou-se por mantê-los na amostra por serem passíveis de processamento via texto completo e dos dois tipos de acoplamento, por possuírem 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.

A Tabela 2 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 à 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.

Tabela 2 – 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. *significativo a nível de 5%.

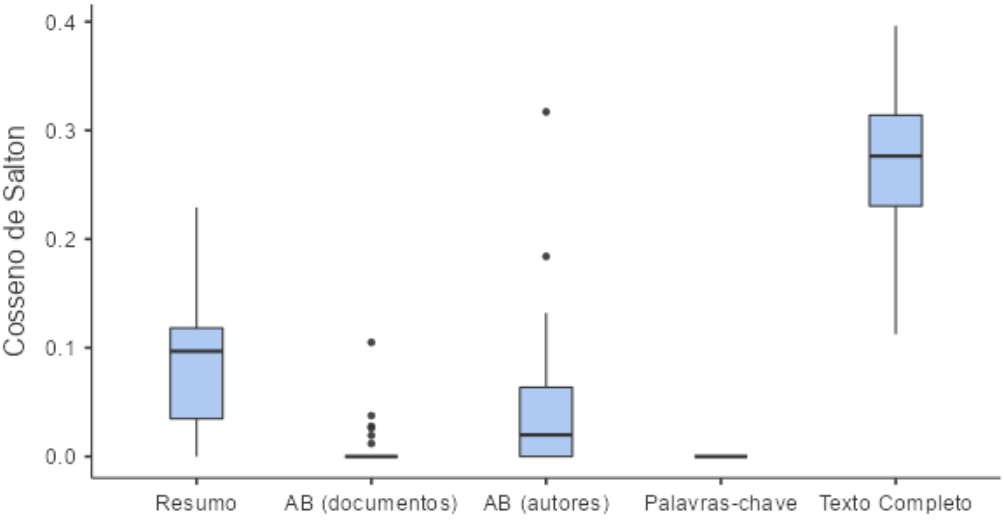
Ainda, é possível observar correlação moderada e forte (Pearson e Spearman) entre os cálculos de acoplamento ao considerar como unidades de acoplamento documentos e autores. Fato é que, se dois artigos estão acoplados por determinado documento, estes artigos estão obrigatoriamente acoplados pelos autores (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) ≥ AB (documentos) e assumir a condição lógica de implicação: dado um conjunto de listas de referências (R) a serem acopladas, então, R: ∃AB (documentos)→∃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 significativas 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 significativas. Os valores calculados, e normalizados, a partir de cada método são apresentados no gráfico de *boxplot* na Figura 1.

Figura 1 – *Boxplot* dos cinco métodos analisados normalizados via Cosseno de Salton



Média	0,0797	0,0051	0,0435	0	0,2684
Mediana	0,0968	0	0,0198	0	0,2764

Fonte: elaboração via software jamovi

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.

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, são eles d_1 e d_5. Neste caso, houve proximidade com relação ao texto completo, mas não entre as demais medidas. Já os artigos d_6 e d_9 versam sobre questões relacionadas à

autocitação e apresentam-se muito próximos apresentarem similares em termos de texto completo, resumos, AB (documentos) e AB (autores), este último, o acoplamento de autores mais intenso de todo corpus analisado.

Ademais, o artigo: d_7 e o artigo d_9 possuem dois autores em comum. O primeiro foi escrito por Rogers, Szomszor e Adams (2020) e o segundo por Szomszor, Pendlebury e Adams (2020), ou seja, tanto Szomszor quanto Adams estão em coautoria em ambos, o que tornou os textos ainda mais próximos visto que possuem a escrita de autores em comum. Estes artigos apresentam-se como os mais similares em termos de resumos e figura-se como a segunda díade mais similar sob a perspectiva de texto completo.

Outro ponto em comum interessante que pode ter aproximado ainda mais o corpus textual pela similaridade são os artigos d_2 e d_4 ambos, por diferentes perspectivas, analisam a produção de pesquisadores poloneses. Esta relação apresentou a maior similaridade em termos de texto completo dentre todos pares analisados.

Com relação aos cálculos de acoplamento, é possível ilustrar o superior desempenho da obra do autor citado como única enquanto unidade de acoplamento, em relação ao AB (documentos). Enquanto AB (autores) assume valores médios e medianos de 0,0435 e 0,0198 (4,35% e 1,98%) de proximidade entre os documentos, o AB (documentos) possui 0,005 (0,5%) de proximidade média entre os documentos e uma mediana igual a zero. É interessante ressaltar que, em ambas perspectivas de acoplamento, a maior intensidade de acoplamento ocorreu entre d_6 e d_9.

Este resultado é esperado, justamente pelo AB (autores) ser uma medida obrigatoriamente maior ou igual que AB (documentos), como supracitado. Com isso, é possível retomar Zhao e Strotmann (2021) e aferir que uma maior similaridade, em termos de autores citados em comum, pode representar uma proximidade teórica mais ampla, indicando influências de escolas de pensamento dos quais os autores acopladores (unidades de acoplamento), responsáveis por acoplarem dois ou mais artigos, advém.

Diferentemente do AB (documentos), em que a proximidade entre os artigos, via documentos citados em comum, representa aspectos teóricos-

metodológicos mais específicos e uteis à obra citante.

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 10 cálculos, como apresenta a Tabela 3.

Tabela 3 – Aplicação do teste de Wilcoxon par a par

		p-valor*
Formas Ativas	Resumos	< 0,001
	Palavras-Chaves	< 0,001
	AB (autores)	< 0,001
	AB (documentos)	< 0,001
Resumos	Palavras-Chaves	< 0,001
	AB (autores)	0,003
	AB (documentos)	< 0,001
Palavras-Chaves	AB (autores)	< 0.001
	AB (documentos)	0,036
AB (autores)	AB (documentos)	< 0,001

Fonte: Elaboração própria. *Cálculos via software jamovi

O teste aferiu diferenças significativas (*p-valor<0,05*) em todas as dez comparações. Dessa forma, é possível apontar que, mesmo que haja correlações 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, apesar de possível correlação, em que texto completo desempenha maiores valores, AB (documentos) menores (mas não totalmente nulos) e palavras-chave, completamente nulos.

5 CONSIDERAÇÕES FINAIS

Esta pesquisa apresentou um estudo comparativo entre cálculos de similaridade entre medidas baseadas no acoplamento bibliográfico (AB) e em métodos de processamento de linguagem natural (PLN). Para isso, utilizou os top 10 artigos com maior número de *downloads* do periódico *Scientometrics* em 2020.

De maneira geral, o estudo tomou como unidade de análise os 10 artigos em questão e calculou a similaridade entre eles utilizando diferentes elementos em cinco vias, sendo três baseadas no PLN e duas no AB: i) número de palavras em sua forma ativa, considerando o texto completo, em comum entre os artigos (PLN); ii) número de palavras em sua forma ativa, considerando somente o resumo de cada artigo, entre os top 10 artigos (PLN); iii) quantidade de palavras-chaves em comum entre os artigos (PLN); iv) AB entre documentos tomando como unidades de acoplamento as referências citada; v) AB entre documentos tomando como unidades de acoplamento os autores citados em cada documento.

A fim de estabelecer comparações entre as cinco medidas, normalizou-se todos os cálculos por meio do índice Cosseno de Salton. Ao normalizar o cômputo das similaridades, foi possível não só correlacionar as variáveis entre si, mas também, comparar os valores numéricos entre eles, uma vez que todos os valores estão sob uma mesma escala (entre 0 e 1).

Assim, esta pesquisa correlacionou todas medidas entre si e comparou-as, ou seja, foi observado se há correlação e possível concordância (ou não) entre os métodos. Utilizou-se as correlações de Pearson e Spearman para as correlações e o teste de Wilconxon para comparação entre os valores numéricos das medidas.

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. Isto é, os dois cálculos de acoplamento se correlacionariam de maneira mais forte entre si, assim como os dois cálculos baseados no PLN (texto completo e resumos).

Desta maneira, verificou-se que 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, ao menos um valor de AB (autores). Este resultado, infere que métodos de acoplamento que levam em conta autores como unidades de acoplamento serão obrigatoriamente maiores ou iguais, a cálculos que tomar por base documentos citados.

Com relação aos dois 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. Ainda, ressalta-se que não houve palavras-chaves em comum entre os 10 artigos e três deles não apresentaram palavras-chave. Pelo fato de não existir intersecção entre os artigos com relação as palavras-chaves, não foi possível correlacionar os valores de acoplamento, uma vez que não houve variabilidade em termos de cálculos de similaridade, pois, todos foram iguais a zero. Tal fato, tornou-se uma limitação deste estudo, visto que em outros domínios é plausível que sejam encontradas palavras-chaves em comum entre artigos de um mesmo periódico.

No que tange os cálculos de correlação entre medidas do PLN com medidas de AB, foi constado que a maior correlação se dá entre texto completo e AB (autores). Nota-se que os cálculos de similaridade tomando o texto completo atinge os maiores valores numéricos dentre as medidas de PLN, assim como AB (autores) é maior que AB (documentos).

Tanto a similaridade entre os textos completos quanto AB (autores) guardam uma distinção perante suas medidas semelhantes, de que: se um documento é citado, necessariamente, os autores são citados; se um existe similaridade entre textos completos, possivelmente, existe similaridade entre os resumos.

Ainda, verificou-se que os maiores destaques, segundo os valores calculados e normalizados, foram os pares: d_2 e d_4 (maior proximidade utilizando texto completo), d_7 e d_9 (maior proximidade entre resumos) e d_6 e d_9 (maior frequência de acoplamento bibliográfico utilizando autores e documentos).

Para compreender a distinção entre cálculos de similaridade, sejam baseados no PLN, seja no AB, foi aplicado o teste de Wilcoxon a nível de 5% de significância entre todos os pares de medidas, contemplando 10 cálculos, em que todos apontaram diferenças significativas entre si. Este resultado aponta que, mesmo existindo correlações significativas entre as medidas normalizadas, todas se diferenciam-se entre si. Tal fato indica que cada forma de computo guarda características próprias e estas características refletem em seu valor normalizado.

Ao estabelecer a relação entre os desempenhos 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.

Como trabalhos futuros, espera-se aumentar o corpus analisados, além de introduzir novas medidas de similaridade como Jaccard ou Dice. Ademais, estabelecer similaridade entre seções de um artigo (introdução, metodologia, resultados e conclusões) via acoplamento bibliográfico entre autores e/ou conteúdo textual para verificar como artigos se conectam.

Por fim salienta-se a necessidade de mais estudos que combinem e realizem novos testes com os métodos. A quantidade de trabalhos que exploram as similaridades entre o acoplamento bibliográfico e o processamento de linguagem natural ainda é incipiente, como apresentado pela busca nas bases *Scopus* e *Web of Science*, mas apresenta grande potencial de desenvolvimento para a área da Ciência da Informação.

REFERÊNCIAS

BORNMANN, L.; MARX, W. Thomas theorem in research evaluation. **Scientometrics**, [*S. l.*], v. 123, n. 1, p. 553-555, 2020. DOI: 10.1007/S11192-020-03389-6

CASTANHA, R. G. The Coupler: uma nova ferramenta bibliométrica para análises relacionais de citação, acoplamento bibliográfico e cocitação. **RDBCI:** Revista Digital de Biblioteconomia e Ciência da Informação, São Paulo, v. 20, 2022. DOI: 10.20396/rdbci.v20i00.8671208

CHOWDHURY, G. Natural language processing. Annual Review of Information Science and Technology. **Asist&T**, [*S. l.*], v. 37, n. 1, p. 51-89, 2003. DOI: 10.1002/aris.1440370103

GIROLAMO, N. D.; REYNDERS, R. M. Characteristics of scientific articles on COVID-19 published during the initial 3 months of the pandemic. **Scientometrics**, [*S. l.*], v. 125, n. 1, p. 795-812, 2020. DOI: 10.1007/S11192-020-03632-0

FALCÃO, L. C. J.; LOPES, B.; SOUZA, R. R. Absorção das tarefas de processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI. **Em**

Questão, Porto Alegre, v. 28, n. 1, p. 13-34, 2021. DOI: 10.19132/1808-5245281.13-34

GRÁCIO, M. C. C. Acoplamento bibliográfico e análise de cocitação: revisão teórico-conceitual. **Encontros Bibli:** Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 21, n. 47, p. 82-99, 2016. DOI: 10.5007/1518-2924.2016v21n47p82

GRÁCIO, M. C. C. **Análises relacionais de citação para a identificação de domínios científicos: uma aplicação no campo dos Estudos Métricos da Informação no Brasil**. Editora UNESP, 2020.

HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. **Science**, [*S. l.*], v. 349, n. 6245, p. 261-266, 2015. DOI: https://www.science.org/doi/10.1126/science.aaa8685

HJØRLAND, B. Citation analysis: A social and dynamic approach to knowledge organization. **Information Processing & Management**, [*S. l.*], v. 49, n. 6, p. 1313-1325, 2013. DOI: 10.1016/j.ipm.2013.07.001

HOU, J.; YANG, X.; CHEN, C. Emerging trends and new developments in information science: A document co-citation analysis (2009-2016). **Scientometrics**, [*S. l.*], v. 115, n. 2, p. 869-892, 2018. DOI: 10.1007/s11192-018-2695-9

KACEM, A.; FLATT, J. W.; MAYR, P. Tracking self-citations in academic publishing. **Scientometrics**, [*S. l.*], v. 123, n. 2, p. 1157-1165, 2020. DOI: 10.1007/S11192-020-03413-9

KESSLER, M. M. Bibliographic coupling between scientific papers. **American documentation**, [*S. l.*], v. 14, n. 1, p. 10-25, 1963. DOI: 10.1002/asi.5090140103

KULCZYCKI, E.; KORYTKOWSKI, P. Researchers publishing monographs are more productive and more local-oriented. **Scientometrics**, [*S. l.*], v. 125, n. 2, p. 1371-1387, 2020. DOI: 10.1007/S11192-020-03376-X

KWIEK, M. Internationalists and locals: international research collaboration in a resource-poor system. **Scientometrics**, [*S. l.*], v. 124, n. 1, p. 57-105, 2020. DOI: 10.1007/S11192-020-03460-2

LARIVIÈRE, V.; GINGRAS, Y. Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. **Journal of informetrics**, [*S. l.*], v. 5, n. 3, p. 392-399, 2011. DOI 10.1016/j.joi.2011.02.001

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. DOI: 10.1081/E-ELIS3

FAGES, D. M. Write better, publish better. **Scientometrics**, [*S. l.*], v. 122, n. 3, p. 1671-1681, 2020. DOI: 10.1007/S11192-019-03332-4

MARSHAKOVA, I. Citation networks in information science. **Scientometrics**, [*S. l.*], v. 3, n. 1, p. 13-25, 1981. DOI: 10.1007/BF02021861

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, [*S. l.*], v. 18, n. 5, p. 544-551, 2011. DOI: 10.1136/amiajnl-2011-000464

PUERTA-DíAZ, M.; MIRA, B. S.; OVALLE-PERANDONES, M.; GRÁCIO, M. C. C.; MARTÍNEZ-ÁVILA, D. O processamento de linguagem natural na área dos estudos métricos da informação: um estudo no período de 2000 a 2019. **Encontros Bibli:** Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 26, 2021. DOI: 10.5007/1518-2924.2021.e76886

PRADO, M. A. R; CASTANHA, R. C. G. Indicadores: conceitos fundamentais e importância em CT&I. *In*: GRÁCIO, M. Cl. C.; MARTÍNEZ-ÁVILA, D.; OLIVEIRA, E. F. T. de; ROSAS, F. S. (org.). **Tópicos da bibliometria para bibliotecas universitárias**. São Paulo: Cultura Acadêmica, 2020. p. 50-70.

ROGERS, G.; SZOMSZOR, M.; ADAMS, J. Sample size in bibliometric analysis. **Scientometrics**, [*S. l.*], v. 125, n. 1, p. 777-794, 2020. DOI: 10.1007/S11192-020-03647-7

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: 27 dez. 2022.

SHIBAYAMA, S.; WANG, J. Measuring originality in science. **Scientometrics**, [*S. l.*], v. 122, n. 1, p. 409-427, 2020. DOI: 10.1007/S11192-019-03263-0

SOLTANI, P.; PATINI, R. Retracted COVID-19 articles: a side-effect of the hot race to publication. **Scientometrics**, [*S. l.*], v. 125, n. 1, p. 819-822, 2020. DOI: 10.1007/S11192-020-03661-9

SZOMSZOR, M.; PENDLEBURY, D. A.; ADAMS, J. How much is too much? The difference between research influence and self-citation excess. **Scientometrics**, [*S. l.*], v. 123, n. 2, p. 1119-1147, 2020. DOI: 10.1007/S11192-020-03417-5

TASKIN, Z.; AL, U. Natural language processing applications in library and information science. **Online Information Review**, [*S. l.*], v. 43, n. 4, p. 676-690, 2019. DOI: 10.1108/OIR-07-2018-0217

THIJS, B. Science mapping and the identification of topics: Theoretical and methodological considerations. *In*: GLÄNZEL, W.; MOED, H. F.; SCHMOCH, U.; THELWALL, M. (ed.). **Springer handbook of science and technology indicators**. Springer, Cham, 2019. p. 213-233. DOI: 10.1007/978-3-030-02511-3_9

THIJS, B.; GLÄNZEL, W.; MEYER, M. S. Using noun phrases extraction for the improvement of hybrid clustering with text-and citation-based components. The example of “Information Systems Research”. *In*: SALAH, A. A.; TONTA, Y.; SALAH, A. A. A.; SUGIMOTO, C.; AL, U. (ed.). **Proceedings of ISSI 2015 Istanbul:** 15th International Society of Scientometrics and Informetrics Conference. Istanbul, Turkey: Bogaziçi University Printhouse, 2015. p. 28-33. Disponível em: http://ceur-ws.org/Vol-1384/paper4.pdf. Acesso em: 12 abr. 2023.

YOUNG, T.; HAZARIKA, D.; PORIA, S.; CAMBRIA, E. Recent trends in deep learning based natural language processing. **IEEE -** Computational intelligenCe magazine, [*S. l.*], v. 13, n. 3, p. 55-75, 2018. DOI: 10.1109/MCI.2018.2840738

YUN, J.; AHN, S.; LEE, J. Y. Return to basics: Clustering of scientific literature using structural information. **Journal of Informetrics**, [*S. l.*], v. 14, n. 4, p. 101099, 2020. DOI: 10.1016/j.joi.2020.101099

ZHANG, Y.; SHANG, L.; HUANG, L.; PORTER, A. L.; ZHANG, G.; LU, J.; ZHU, D. A hybrid similarity measure method for patent portfolio analysis. **Journal of Informetrics**, [*S. l.*], v. 10, n. 4, p. 1108-1130, 2016. DOI: 10.1016/j.joi.2016.09.006

ZHAO, D.; STROTMANN, A. Evolution of Research Activities and Intellectual Influences in Information Science 1996-2005: Introducing Author Bibliographic-Coupling Analysis. **Journal of the American Society for Information Science and Tecnhology**, [*S. l.*], v. 59, n. 13, p. 2070-2086, 2008. DOI: 10.1002/asi.20910

ZHAO, D.; STROTMANN, A. Mapping knowledge domains on Wikipedia: an author bibliographic coupling analysis of traditional Chinese medicine. **Journal of Documentation**, [*S. l.*], v. 78, n. 2, 2021. DOI: 10.1108/JD-02-2021-0039

NATURAL LANGUAGE PROCESSING AND BIBLIOGRAPHIC COUPLING: AN ANALYSIS OF THE PROXIMITY BETWEEN THE MOST ACCESSED ARTICLES OF THE SCIENTOMETRICS JOURNAL

ABSTRACT

Objective: to compare the methods of Natural Language Processing and Bibliographic

Coupling normalized via Salton Cosine applied to the ten most accessed articles of 2020 in the Scientometrics journal. **Methodology:** It calculates the similarity between all articles according to five perspectives, namely: similarities between active forms of the full text, active forms of abstracts, keywords in common, bibliographic coupling between documents and bibliographic coupling of authors. Furthermore, it calculates the Pearson and Spearman correlations, applies the Wilcoxon non-parametric test at a 5% significance level, and represents the normalized values in a boxplot. **Results:** It finds that the specificities of each method significantly influence the achievement of a significant correlation between the measures in which the two coupling calculations would correlate more strongly with each other, as well as two calculations based on natural language processing. Note that the coupling calculations correlated significantly, as for each document coupling value there is necessarily at least one author coupling value. About calculations based on natural language processing, there is a strong correlation between full texts and abstracts, as there is a content dependence between both. The Wilcoxon test measured significant differences between all pairs of compared measurements. **Conclusions:** It concludes a strong correlation between full texts and abstracts, and between bibliographic coupling methods. However, there is a significant difference between the calculated values.

Descriptors: Bibliographic coupling. Similarity Index. Natural language processing.

PROCESAMIENTO DEL LENGUAJE NATURAL Y ENLACE BIBLIOGRÁFICO: UN ANÁLISIS DE LA PROXIMIDAD ENTRE LOS ARTÍCULOS MÁS ACCESO DE LA REVISTA SCIENTOMETRICS

RESUMEN

Objetivo: compara los métodos de Procesamiento del Lenguaje Natural y Acoplamiento Bibliográfico normalizados a través del Coseno de Salton aplicados a los diez artículos más consultados de 2020 de la revista Scientometrics. **Metodología:** Calcular una similitud entre todos los artigos segundo cinco perspectivas, sendo elas: similitudes entre formas activas do texto completo, formas activas dos resumos, palavravas-chaves em comum, acoplamento bibliográfico entre documentos e acoplamento bibliográfico de autores. Además, calcula las correlaciones de Pearson y Spearman, aplica la prueba no paramétrica de Wilcoxon a un nivel de 5% de significancia y representa los valores normalizados en el diagrama de caja. **Resultados:** Encuentra que las especificidades de cada método influyen significativamente en el logro de una correlación significativa entre las medidas en las que los dos cálculos de acoplamiento se correlacionarían más fuertemente entre sí, así como dos cálculos basados en el procesamiento del lenguaje natural. Tenga en cuenta que los cálculos de acoplamiento se correlacionaron significativamente, ya que para cada valor de acoplamiento de documento hay necesariamente al menos un valor de acoplamiento de autor. En cuanto a los cálculos basados en el procesamiento del lenguaje natural, existe una fuerte correlación entre los textos completos y los resúmenes, ya que existe una dependencia de contenido entre ambos. La prueba de Wilcoxon midió diferencias significativas entre todos los pares de medidas comparadas. **Conclusiones:** Concluye una fuerte correlación entre textos completos y resúmenes, y entre métodos de acoplamiento bibliográfico. Sin embargo, existe una diferencia significativa entre los valores calculados.

Descriptores: Acoplamiento bibliográfico. Índice de similitud. Procesamiento natural del lenguaje.

Apêndice 1 – Lista dos cálculos de similaridade via Cosseno de Salton

Acoplamentos		Formas ativas	Resumos	Palavras-Chave	AB (autores)	AB (documentos)
d_1	d_{10}	0,2456	0,0575	0	0,0249	0,0000
d_1	d_8	0,1955	0,0347	0	0,0000	0,0000
d_1	d_5	0,2313	0,0000	0	0,0000	0,0000
d_1	d_6	0,2587	0,0970	0	0,0000	0,0000
d_1	d_2	0,2951	0,1879	0	0,0000	0,0000
d_1	d_3	0,2942	0,1160	0	0,0000	0,0000
d_1	d_4	0,3197	0,1098	0	0,0000	0,0000
d_1	d_7	0,2722	0,1279	0	0,0000	0,0000
d_1	d_9	0,2770	0,0732	0	0,0000	0,0000
d_2	d_5	0,2012	0,0000	0	0,0000	0,0000
d_2	d_6	0,2855	0,0968	0	0,0770	0,0000
d_2	d_3	0,3140	0,0579	0	0,0451	0,0000
d_2	d_9	0,3159	0,1299	0	0,0558	0,0000
d_2	d_{10}	0,2913	0,1021	0	0,0000	0,0000
d_2	d_7	0,3416	0,1083	0	0,0536	0,0000
d_2	d_8	0,2615	0,0462	0	0,1841	0,0000
d_2	d_4	0,3962	0,1949	0	0,0974	0,0000
d_3	d_8	0,2204	0,0000	0	0,1134	0,0000
d_3	d_6	0,2917	0,1046	0	0,0791	0,0376
d_3	d_{10}	0,2764	0,1181	0	0,0335	0,0000
d_3	d_5	0,1562	0,0000	0	0,0000	0,0000
d_3	d_9	0,3320	0,1203	0	0,1237	0,0118
d_3	d_7	0,3155	0,1146	0	0,1320	0,0278
d_3	d_4	0,3535	0,0752	0	0,0343	0,0000
d_4	d_8	0,2420	0,0360	0	0,0630	0,0000
d_4	d_9	0,3653	0,1013	0	0,0636	0,0000
d_4	d_7	0,3666	0,1086	0	0,0153	0,0000
d_4	d_{10}	0,2823	0,0398	0	0,0155	0,0000
d_4	d_5	0,1775	0,0000	0	0,0165	0,0000
d_4	d_6	0,2900	0,1132	0	0,0146	0,0000
d_5	d_{10}	0,1802	0,0000	0	0,0000	0,0000
d_5	d_8	0,1125	0,0000	0	0,0000	0,0000
d_5	d_7	0,1819	0,0000	0	0,0000	0,0000
d_5	d_6	0,1785	0,0000	0	0,0000	0,0000
d_5	d_9	0,1570	0,0000	0	0,0198	0,0000
d_5	d_8	0,2404	0,0000	0	0,1162	0,0000
d_5	d_7	0,3114	0,1199	0	0,0282	0,0000
d_5	d_{10}	0,2360	0,1186	0	0,0000	0,0000
d_5	d_9	0,3403	0,1761	0	0,3171	0,1050
d_7	d_9	0,3796	0,2292	0	0,0857	0,0194
d_7	d_{10}	0,2889	0,1516	0	0,0000	0,0000
d_7	d_8	0,2730	0,0515	0	0,0404	0,0000
d_8	d_{10}	0,2305	0,0566	0	0,0000	0,0000
d_8	d_9	0,2288	0,0721	0	0,0842	0,0262

d_9	d_{10}	0,2724	0,1392	0	0,0249	0,0000

Fonte: elaboração própria.

Recebido em: 28.12.2022
Aceito em: 22.03.2023