



Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação

ISSN: 1518-2924

Programa de Pós-graduação em Ciência da Informação -
Universidade Federal de Santa Catarina

PUERTA-DÍAZ, Mirelys; MARTÍNEZ-ÁVILA, Daniel; MIRA, Bianca Savegnago
de; OVALLE-PERANDONES, María-Antonia; Grácio, Maria Cláudia Cabrini
O PROCESSAMENTO DE LINGUAGEM NATURAL NOS ESTUDOS MÉTRICOS DA INFORMAÇÃO:
UMA ANÁLISE DOS ARTIGOS INDEXADOS PELA WEB OF SCIENCE (2000- 2019).

Encontros Bibli: revista eletrônica de biblioteconomia e
ciência da informação, vol. 26, e76886, 2021, Janeiro-Abril
Programa de Pós-graduação em Ciência da Informação - Universidade Federal de Santa Catarina

DOI: <https://doi.org/10.5007/1518-2924.2021.e76886>

Disponível em: <https://www.redalyc.org/articulo.oa?id=14768130010>

- Como citar este artigo
- Número completo
- Mais informações do artigo
- Site da revista em redalyc.org

UNEM redalyc.org

Sistema de Informação Científica Redalyc

Rede de Revistas Científicas da América Latina e do Caribe, Espanha e Portugal


Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa
acesso aberto

Encontros Bibli


O PROCESSAMENTO DE LINGUAGEM NATURAL NOS ESTUDOS MÉTRICOS DA INFORMAÇÃO: UMA ANÁLISE DOS ARTIGOS INDEXADOS PELA WEB OF SCIENCE (2000- 2019).

Natural Language Processing in Information Metric Studies: an analysis of the articles indexed by the Web of Science (2000-2019).


Mirelys PUERTA-DÍAZ

Doutoranda do Programa de Pós-graduação em Ciência da Informação
Universidade Estadual Paulista, Departamento Ciência da Informação, Marília, Brazil
mirelys.puerta@unesp.br
<https://orcid.org/0000-0002-2312-2540> 


Daniel MARTÍNEZ-ÁVILA
Doutor e Professor.

Universidad de Carlos III de Madrid, Departamento de Biblioteconomía y Documentación, Madrid, Espanha
daniel.martinez@uc3m.es
<https://orcid.org/0000-0003-2236-553X> 

Bianca Savegnago de MIRA

Bacharel em Biblioteconomia.
Universidade Estadual Paulista, Departamento Ciência da Informação, Marília, Brazil
bianca.mira@unesp.br
<https://orcid.org/0000-0001-7913-4084> 

María-Antonia OVALLE-PERANDONES
Doutora e Professora

Universidad Complutense de Madrid, Departamento de Biblioteconomía y Documentación, Madrid, Espanha
maovalle@ucm.es
<https://orcid.org/0000-0002-6149-4724> 

Maria Cláudia Cabrini Grácio
Doutora e Professora

Universidade Estadual Paulista, Departamento Ciência da Informação, Marília, Brazil
cabrini.gracio@unesp.br
<https://orcid.org/0000-0002-8003-0386> 

A lista completa com informações dos autores está no final do artigo 

RESUMO

Objetivo: Identificar a estrutura científica internacional das pesquisas que vinculam o uso do Processamento de linguagem natural no campo dos estudos métricos da informação.

Método: A pesquisa é baseada em uma perspectiva qualitativa própria dos estudos métricos da informação no domínio da organização do conhecimento. A coleta de dados foi realizada em 02/02/2020 no recurso *Web of Science Core Collection* com a expressão "natural language processing", na categoria artigos e revisão, refinada pelas Categorias da *Web of Science Information Science Library Science* e limitada à janela temporal dos últimos 20 anos completos (período de 2000 a 2019). A Análise de Redes Sociais é utilizada como método de pesquisa para examinar e visualizar a rede de colaboração científica, de cocitação e de coocorrência de palavras-chave.

Resultados: Dos 552 documentos recuperados, após a análise dos resumos, observou-se que 31 estavam inseridos no campo dos estudos métricos. A literatura científica mostra um crescente aumento das publicações nos últimos três anos, com 2018 sendo o ano mais produtivo.

Conclusões: Considerando que o conjunto de técnicas de PLN (ex. *bag of words*, *tokenization*, *word stemming*, *part-of-speech tagging* e SVM) vem permitindo ao pesquisador ir além da análise de citação tradicional, para uma análise mais voltada ao conteúdo e contexto da citação, a literatura científica internacional sobre a aplicação do PLN nos estudos métricos da informação tem se mostrado emergente. A revista *Scientometrics* configura o meio de disseminação dos trabalhos que alcançaram maior impacto. Finalizando, a análise de cocitação k-core mostra a existência de um importante núcleo teórico, frequentemente citado na comunidade acadêmica internacional.

PALAVRAS-CHAVE: Processamento da Linguagem Natural. Estudos Métricos da Informação. Análise de Redes Sociais. Pesquisa Científica. Mapeamento da Ciência

ABSTRACT

Objective: To identify the international scientific structure of the research on the use of natural language processing in the information metric studies area.

Methods: It follows qualitative and quantitative approaches of the information metric studies and the knowledge organization domain. The data was retrieved on 02/02/2020 from the *Web of Science Core Collection* using the expression "natural language processing", limited to the document types articles and reviews, the category *Information Science Library*

Science, and the timespan of the last 20 complete years (from 2000 to 2019). A Social Networks Analysis was conducted for the visualization of the scientific collaboration, co-citation, and keywords co-occurrence networks.

Results: Out of the 552 documents retrieved, 31 papers were identified in the information metric studies area. Bibliometric indicators of production, relationship, and impact were considered in the study and showed an increase of publications in the last three years, being 2018 the most productive year.

Conclusions: The international scientific literature on the application of NLP in information metric studies is emerging. Scientometrics was identified as the source that achieved a greatest impact. Finally, the k-core of the co-citation analysis shows the existence of an important theoretical core, often cited in the international academic community. The set of NLP techniques (e.g., bag of words, tokenization, word stemming, part-of-speech tagging, and SVM) allows the researcher to go beyond the traditional citation analysis and focus on content and context of the citations.

KEYWORDS: Natural Language Processing. Information Metric Studies. Social Network Analysis. Scientific Research. Mapping of science.

1 INTRODUÇÃO

A pesquisa em processamento de linguagem natural vem se intensificando há vários anos, desde o final da década de 1940. Nos anos 50, os estudos no campo do Processamento de Linguagem Natural (PLN) procuraram aliar a Inteligência Artificial (IA) à Linguística. Posteriormente, com o avanço das pesquisas científicas interdisciplinares, os estudos em PLN se aproximaram do campo da Recuperação da Informação e, desde a década de 1960, o PLN passou a ser utilizado como técnica de indexação e pesquisa em grandes volumes de texto e como fornecedor de dados estatísticos (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011), revelando também sua relevância para a área da Ciência da Informação. Na década de 1970, desenvolveram-se os modelos matemáticos aplicados à indexação e à recuperação de documentos, baseados na teoria das probabilidades, estendidos nas décadas subsequentes (SMEATON, 1999).

Estudos nos quais os métodos de PLN e as análises bibliométricas são realizados em conjunto são mais comuns nos últimos anos. Segundo Taşkin et al. (2019), essas pesquisas podem ser categorizadas em dois grupos: um no qual o PLN é método dos estudos bibliométricos aplicados; e outro nos quais os artigos que abordam o campo do PLN são o objeto de pesquisa dos estudos metateóricos sustentados no método bibliométrico. Nesse contexto em que é método para os estudos bibliométricos, a análise das citações de qualquer conjunto de artigos em grande escala nos seus quatro níveis (sentença, parágrafo, seção e artigo) de proximidade textual tem constituído uma tarefa complexa e desafiante, para a qual o PLN traz contribuições significativas (LIU; CHEN, 2011).

Apoiando-se em componentes da análise de texto, Glänzel, Heeffer e Thijs (2017) combinam técnicas baseadas em enlaces para agrupar espaços de documentos e detectar tópicos de pesquisa emergentes em larga escala. Aspectos estatísticos, distribuições geográficas e relações de colaboração da pesquisa em computação móvel com PLN foram

abordados por Chen et al. (2018), um aspecto que também foi mostrado por Li e Lei (2019), objetivando aprimorar o processo e os métodos de avaliação da qualidade e produtividade da pesquisa do campo do PLN.

Embora exista um conjunto de estudos na literatura científica internacional com foco na Ciência da Informação e, especificamente, a Bibliometria usando métodos de PLN, são incipientes as análises bibliométricas sobre esse domínio de conhecimento. No contexto brasileiro, observam-se estudos próximos à literatura como o trabalho intitulado “Processamento de Linguagem Natural: em busca de evidências temáticas nas publicações nacionais e contemporâneas” das autoras Ladeira e Alvarenga (2009), o qual aplica a análise de conteúdo para analisar a produção científica nacional e contemporânea na área registrada no Lattes; assim como o posterior estudo de Ferreira e Corrêa (2018) que aplica o software Iramuteq para realizar um estudo métrico temático sobre biblioteca digital no Brasil.

A presente pesquisa é uma versão preliminar e reduzida de um estudo de Puerta-Díaz et al. (2020) apresentado no 7º Encontro Brasileiro de Bibliometria e Cientometria (EBBC) onde panoramicamente descreve-se o contexto da produção científica em termos de distribuição anual, impacto e procedência, as relações de cocitação e os tópicos mais frequentes e emergentes na área. Sendo este uma primeira abordagem a estrutura científica em âmbito internacional das pesquisas que vinculam o uso do PLN no campo dos estudos métricos. Isto posto, em continuidade, o presente estudo objetiva aprofundar a análise das características da literatura científica internacional que aborda o uso do PLN no campo dos Estudos Métricos da Informação (EMI), seguindo o paradigma de análise de domínio da organização do conhecimento. De forma mais específica, descreve-se o contexto da produção científica em termos de distribuição anual, impacto e procedência, assim como as relações de cocitação, identifica a frente de pesquisa e os tópicos emergentes e aqueles que marcam tendências neste interdomínio. Ademais adiciona a aplicação do indicador de impacto, a relação dos artigos que mais se destacam em citações recebidas e os periódicos mais citados pelos autores. Com base nos mesmos dados coletados amplia ainda a caracterização da produção científica que aborda a temática PLN nos Estudos Métricos da Informação ao examinar os fatores associados à presença dos pesquisadores com atuação mais consolidada na temática. No que refere à rede de coocorrência das palavras chave, aprofunda as análises do cluster com maior aderência ao tópico Infometria, Cientometria e Bibliometria e a partir de uma perspectiva temporal identifica os tópicos emergentes.

Considera-se, assim, que a presente pesquisa é relevante para a área já que trata uma temática pouco estudada e descreve a estrutura científica do um interdomínio de conhecimento em construção que precisa ser socializado e examinado com profundidade. Ela contribui para ampliar a compreensão da evolução do desenvolvimento científico do interdomínio do PLN e dos Estudos Métricos da Informação.

2 APROXIMAÇÕES TEÓRICAS AO PROCESSAMENTO DE LINGUAGEM NATURAL NOS ESTUDOS MÉTRICOS DA INFORMAÇÃO

O PLN tem sido definido por vários autores, principalmente da área Ciência da Computação e da Linguística, entre outras. Da perspectiva da Ciência da Informação, Liddy (2001, p.2) a define como:

“um conjunto de técnicas computacionais motivadas teoricamente para analisar e representar textos que ocorrem naturalmente em um ou mais níveis de análise linguística, com o objetivo de obter processamento de linguagem semelhante ao humano para uma variedade de tarefas ou aplicativos”.

Seu objetivo é realizar o processamento de linguagem semelhante ao feito pelos humanos. Sua base teórico-metodológica é interdisciplinar, pois a integram a ciências da computação e da informação, linguística, matemática, engenharia elétrica e eletrônica, inteligência artificial e robótica e psicologia (LIDDY, 2001). Desse modo, PLN é uma área de pesquisa e aplicação que explora como os computadores podem ser usados para entender e manipular texto ou fala em linguagem natural para fazer coisas úteis. Suas aplicações incluem vários campos de estudo; entre eles: tradução automática; processamento e resumo de texto em idioma natural; interfaces de usuário; recuperação da informação entre linguagens e multilíngue (CLIR, siglas em inglês); reconhecimento da linguagem; inteligência artificial; e sistemas especializados (CHOWDHURY, 2005).

No cenário da Ciência da Informação (CI), as publicações sobre a temática aparecem nos anos 80 em periódicos como *Proceedings of the American Society for Information Science*, *Journal of the American Society for Information Science* e no *Annual Review of Information Science and Technology*. Na última década do século XX, a literatura científica sobre PLN na CI amplia-se, inserindo-se em periódicos como *Journal of the American Medical Informatics Association*, *Information Processing & Management*, *Journal of the*

*American Society for Information Science and Technology*¹ e *Journal of Information Science*.

O processamento de linguagem natural divide-se em sete níveis principais (LIDDY, 2001): a) fonológico (interpretação de sons da fala), b) morfológico (busca interpretar a natureza componencial das palavras que são compostas de morfemas), c) léxico (interpretar o significado das palavras individuais), d) sintático (descoberta das estruturas gramaticais das sentenças), e) semântico (determinar os significados das frases, concentrando-se nos significados no nível das palavras), f) discursivo (se enfoca nas propriedades dos textos como um todo e faz conexões entre as frases) e o nível pragmático (compreensão do uso intencional do idioma em situações). A sequência dos níveis segue o critério de complexidade, do nível mais simples (fonológico) ao mais complexo (pragmático), de realização das tarefas do PLN nos estudos, facilitando, assim, as análises de grandes quantidades de textos.

Os sistemas atuais de PLN tendem a programar módulos para atingir principalmente os níveis mais baixos de processamento. Atualmente, existem poucos sistemas de trabalho que incorporam os níveis mais altos (LIDDY, 2010; TAŞKIN; AI, 2019). As análises semântica e sintática são, geralmente, preferidas para as análises de citações baseadas em conteúdo, as quais podem substituir a contagem tradicional de citações e contribuir para o desenvolvimento de uma nova geração de indicadores de citação. Segundo Liddy (2010), em relação aos níveis mais altos, particularmente o nível do discurso, a sua aplicação objetiva reduzir textos maiores a uma representação narrativa abreviada que, embora mais curta, seja mais rica de informação, em relação aos documentos originais. Embora os algoritmos das atuais abordagens do PLN apresentem limitações, estão baseados na representação sintática (também chamada de estrutura sintática) do texto, ou seja, fazem uma contagem das frequências de coocorrência de palavras que aparecem no corpus (CHOWDHARY, 2020), sendo esta uma das principais aplicações na área dos Estudos Métricos da Informação.

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa sustenta-se em uma perspectiva metodológica quali-quantitativa própria dos Estudos Métricos da Informação (EMI), seguindo o paradigma de Análise de domínio

¹ Atual *Journal of the Association for Information Science and Technology (JASIST)*

na Organização do Conhecimento (HJØRLAND, 2002, 2017; SMIRAGLIA, 2015). A coleta de dados foi realizada, em 02/02/2020, na Web of Science - Coleção Principal (SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-EXPANDED, IC). A busca e recuperação dos registros bibliográficos com a expressão "natural language processing" foi utilizada no campo Tópico (título, resumo e palavras-chave), limitada à janela temporal dos últimos 20 anos completos (período de 2000 a 2019), para o tipo de documento artigo e revisão e refinada para a categoria de conhecimento *Information Science Library Science*. Recuperaram-se 552 documentos, dos quais, após análise, 31 estavam inseridos no campo dos EMI, ou seja, os estudos métricos representam 5,6% dos documentos recuperados a partir da busca por "natural language processing" na base Web of Science (WoS) entre os anos de 2000 e 2019 na categoria de conhecimento *Information Science Library Science*.

Foram utilizados indicadores bibliométricos de produção (total de publicações, número de artigos por país, produtividade dos autores), citação (total de citações recebidas, análise de cocitações) ligação (análise de coocorrência de palavras, rede de cocitação, análise da colaboração científica) e de impacto científico (número de citações), a fim de identificar e descrever as tendências de pesquisa presentes nas pesquisas que vinculam o processamento de linguagem natural aos estudos métricos da informação. A seguir, a partir das referências presentes nos 31 registros, construiu-se a matriz de cocitação, considerando apenas o primeiro autor das referências. O modelo adotado é a análise de cocitação ao nível de documentos visando revelar informações mais específicas a partir das referências citadas (CHEN; IBEKWE-SANJUAN; HOU, 2010), como o conjunto de documentos que conformam a frente de pesquisa, assim como a elite de cocitação.

O critério adotado de restrição da análise ao primeiro autor sustentou-se no pressuposto de o primeiro autor ser reconhecido como o responsável pela obra, incluindo o que nela é citado (WHITE, 2001). Ademais, em publicações com mais de três autores, em que as normas estabelecem o uso do termo "et al.", este critério permite a igualdade de tratamento entre os artigos analisados. Teve como critério para a seleção dos pares de cocitação e seu posterior mapeamento os trabalhos com frequência de citação maior que 1. Esta opção decorreu do fato de a identidade de citação de um autor ser constituída, segundo White (2001) pelas obras por ele citadas em mais de um artigo. O processo de extração das informações e criação da matriz foi realizado no software Bibexcel versão 2017. O software Pajek versão 5.10 foi usado para a análise e visualização da rede de cocitação e o algoritmo utilizado para a visualização da rede foi Kamada-Kawai (1991)

(componentes separados), com a perspectiva tradicional de análise de cocitação em nível de documentos. O software VOSviewer 1.6.14 foi usado para criar a rede de temáticas a partir das palavras-chave atribuídas pelo autor extraídas dos 31 artigos.

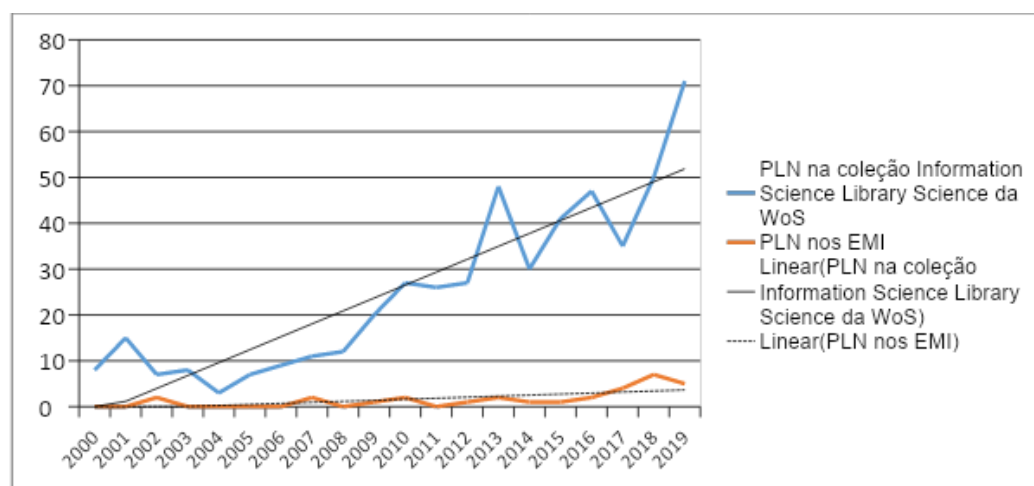
4 RESULTADOS

Observou-se, inicialmente, que quase metade (15) dos artigos foi publicada no periódico *Scientometrics*. A primeira publicação sobre o tema corresponde ao trabalho de Galvez e Moya-Anegón (2007), tratando da padronização dos formatos de dados de origem corporativos. Outras fontes com disseminação mais significativa do tema foram *International Journal on Digital Libraries* e *Journal of the Association for Information Science*, com três publicações cada, *Knowledge Organization* e *World Patent Information*, com duas publicações e as outras seis fontes publicaram apenas um artigo cada.

O Gráfico 1 apresenta a evolução temporal das publicações recuperadas na base WoS. Tanto a escala linear (linha de cor azul) como a escala logarítmica (linha preta contínua) mostra o aumento ininterrupto das pesquisas que aplicam o PLN na presente década. Com base na escala linear o ano de 2018 aparece com o maior total de publicações indexadas, com aderência aos estudos métricos na WoS. Além disso, a escala logarítmica (linha preta descontínua) permitiu observar com maior clareza que o crescente interesse da comunidade científica internacional, embora discreto em relação aos totais de publicações observadas na CI em geral, no uso das técnicas e dos algoritmos de PLN aplicados aos estudos métricos da informação (linha de cor vermelha) ocorre a partir do ano 2011. No ano 2018 em particular, aborda-se o problema da imprecisão dos textos de citação e a possibilidade de se identificar com maior exatidão o contexto na análise de citação, mediante o uso de algoritmos de PLN.

A análise da presença dos pesquisadores com atuação mais consolidada na temática evidenciou que somente oito autores foram responsáveis por mais de um artigo publicado, tendo a seguinte distribuição: com quatro trabalhos Yoon e Kim; com duas publicações Song, Yan, Thijs, Park, Li e Glänzel.

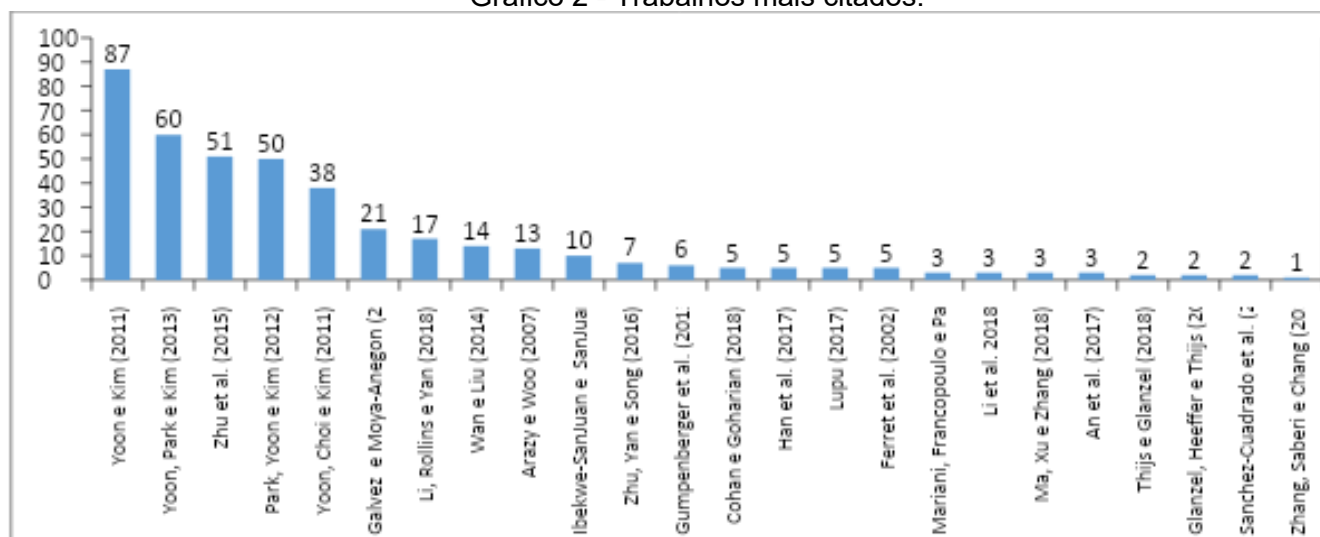
Gráfico 1 - Distribuição das publicações científicas sobre PLN (2000 - 2019).



Fonte: Elaborado pelos autores (2020)

Dos 31 trabalhos, 77% (24 artigos) foram citados (Gráfico 2), destacando-se 5 artigos que receberam a maior quantidade de citações: Yoon e Kim em 2011 no *Scientometrics* (87); Yoon, Park e Kim em 2013 no *Scientometrics* (60); Zhu et al. em 2015 no *JASIST* (51); Park, Yoon e Kim em 2012 no *Scientometrics* (50); Yoon, Choi e Kim em 2011 no *Scientometrics* (38). Dos cinco autores mais produtivos e com as maiores quantidades de citações, encontram-se Yoon, Kim e Park. Ademais, observa-se que o periódico *Scientometrics* aparece como fonte em quatro dos trabalhos que alcançaram maior impacto. Esse comportamento se mantém na totalidade das citações recebidas.

Gráfico 2 - Trabalhos mais citados.



Fonte: Elaborado pelos autores (2020).

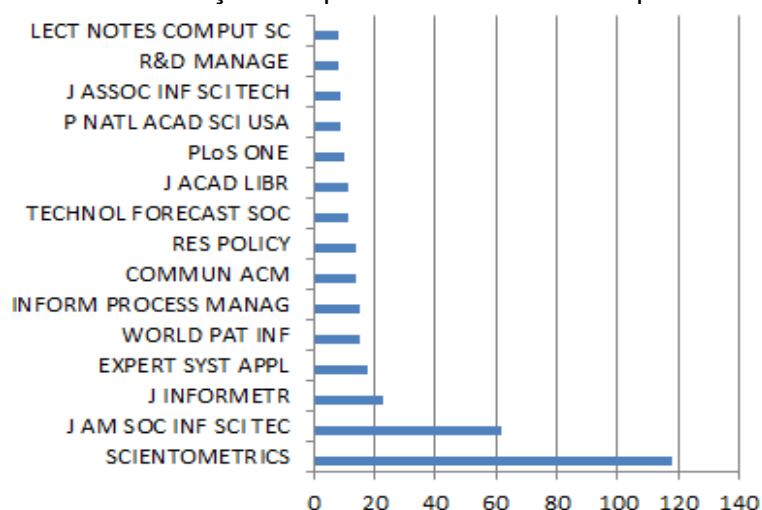
Segundo Szomszor; Pendlebury e Adams (2020) as citações podem ser um indicador da importância da publicação, utilidade ou impacto de curto prazo. Porém, é necessário avaliar se as citações recebidas refletem genuinamente a influência da

publicação ou são consequência da excessiva autocitação dos seus autores. Seguindo esse princípio, destaca-se que após uma análise diacrônica de autocitação (GHIASI; LARIVIÈRE; SUGIMOTO, 2016) nos cinco trabalhos mais citados, os valores totais de autocitações diacrônicas foram calculados após se desconsiderar possíveis duplicações, tanto das citações ao artigo feitas pelo primeiro autor como as citações por qualquer dos autores em outras pesquisas. Foi possível identificar neste corte que nenhum dos trabalhos supera o 25% do total (IOANNIDIS et al., 2019), porém não passaram o limite aceitável de autocitação que evidencie uma má conduta ética. As autocitações são inevitáveis na comunicação acadêmica (GHIASI; LARIVIÈRE; SUGIMOTO, 2016) e embora essa prática pode contribuir para o aumento da contagem de citações dos trabalhos em análise e impor um impacto nas inferências feitas pelos pesquisadores, elas são consideradas normais e demonstrativa da evolução e comprometimento dos autores com a temática.

Ainda sobre os cinco artigos que mais se destacam em citações recebidas, notam-se alguns aspectos interessantes na distribuição por países e por periódicos dessas citações. Três dos trabalhos em que Yoon aparece como autor têm Coreia do Sul, China e Estados Unidos como os três principais países citantes dos artigos. Outro trabalho de Park, Yoon e Kim do ano de 2012 apresenta uma distribuição similar, com a Coreia do Sul e China como os principais países citantes. A Alemanha surge em terceiro lugar, seguida dos Estados Unidos. Destaca-se que o Brasil aparece como o sexto principal país citante do artigo de Yoon, Park e Kim de 2013. Este é o único artigo em que o Brasil aparece entre os citantes, sendo a autoria citante filiada ao Centro Tecnologia Informática e Matemática da Universidade Federal São Carlos. Uma distribuição geográfica diferente das citações é observada no artigo de Zhu et al. do ano de 2015, com a maior parte proveniente da China, seguida da Inglaterra e Estados Unidos. Diferente dos demais, a Coreia do Sul ocupa o 10º segundo lugar entre os citantes.

A respeito das citações realizadas pelos periódicos, *Scientometrics* ocupa o primeiro lugar no número de citações realizadas em quatro dos cinco artigos analisados, sendo a menor presença em Yoon, Park e Kim de 2013 com 15% e a maior em Zhu, Turney, Lemire e Vellino do ano de 2015 com 25%. Desse modo, não só é o periódico mais citado pelos autores (Gráfico 3), mas também um dos que mais os cita.

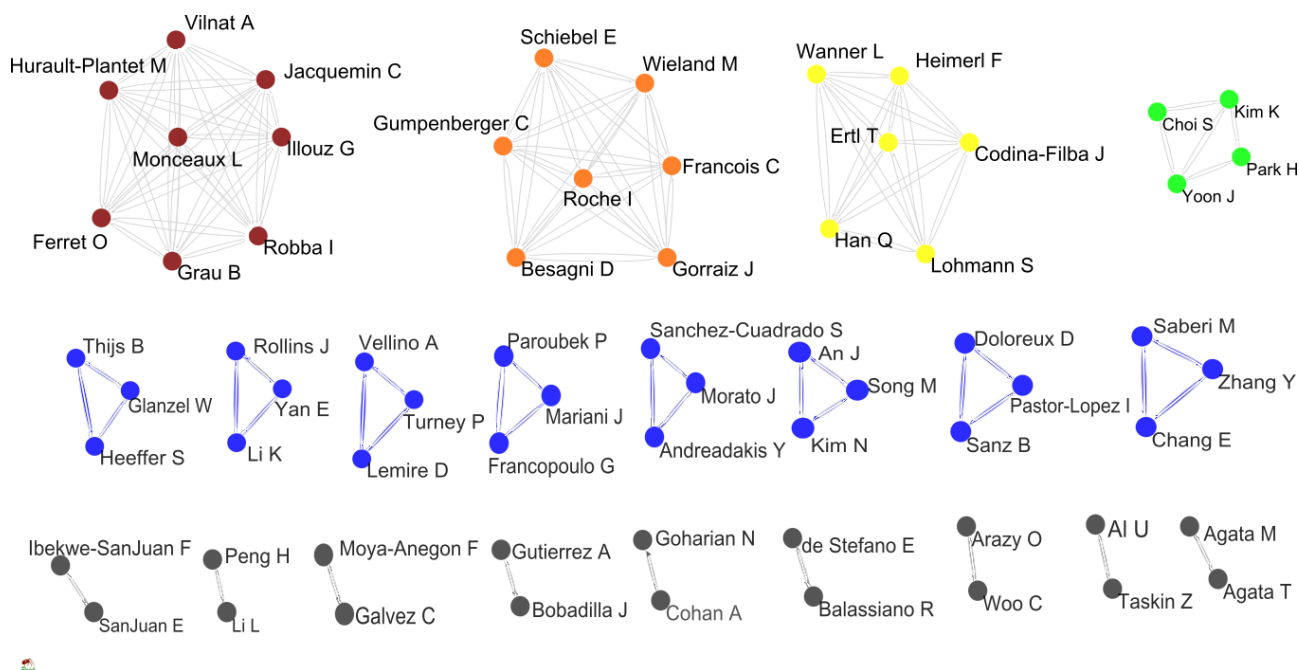
Gráfico 3 - Distribuição dos periódicos mais citados pelos autores



Fonte: Elaborado pelos autores.

A coautoria é um aspecto essencial na análise das relações de colaboração entre autores, a qual revela uma atividade social que visa unir esforços e compartilhar o conhecimento com o objetivo de gerar ciência, assim como a compatibilidade da fundamentação teórico-metodológica entre os coautores. Desde a perspectiva da Análise de Redes Sociais (ARS), a Figura 1 mapeia entre os 91 autores identificados na coleta uma rede de colaboração de 67 autores com pelo menos uma relação de coautoria.

Figura 1 - Rede de coautoria na temática PLN no escopo dos EMI.



Fonte: Elaborado pelos autores em Pajec editado o svg em Inkscape.

Dos 31 artigos extraídos da fonte de dados, somente dois foram publicados em autoria individual. A Figura 1 evidencia que a rede de coautoria é pouco conectada,

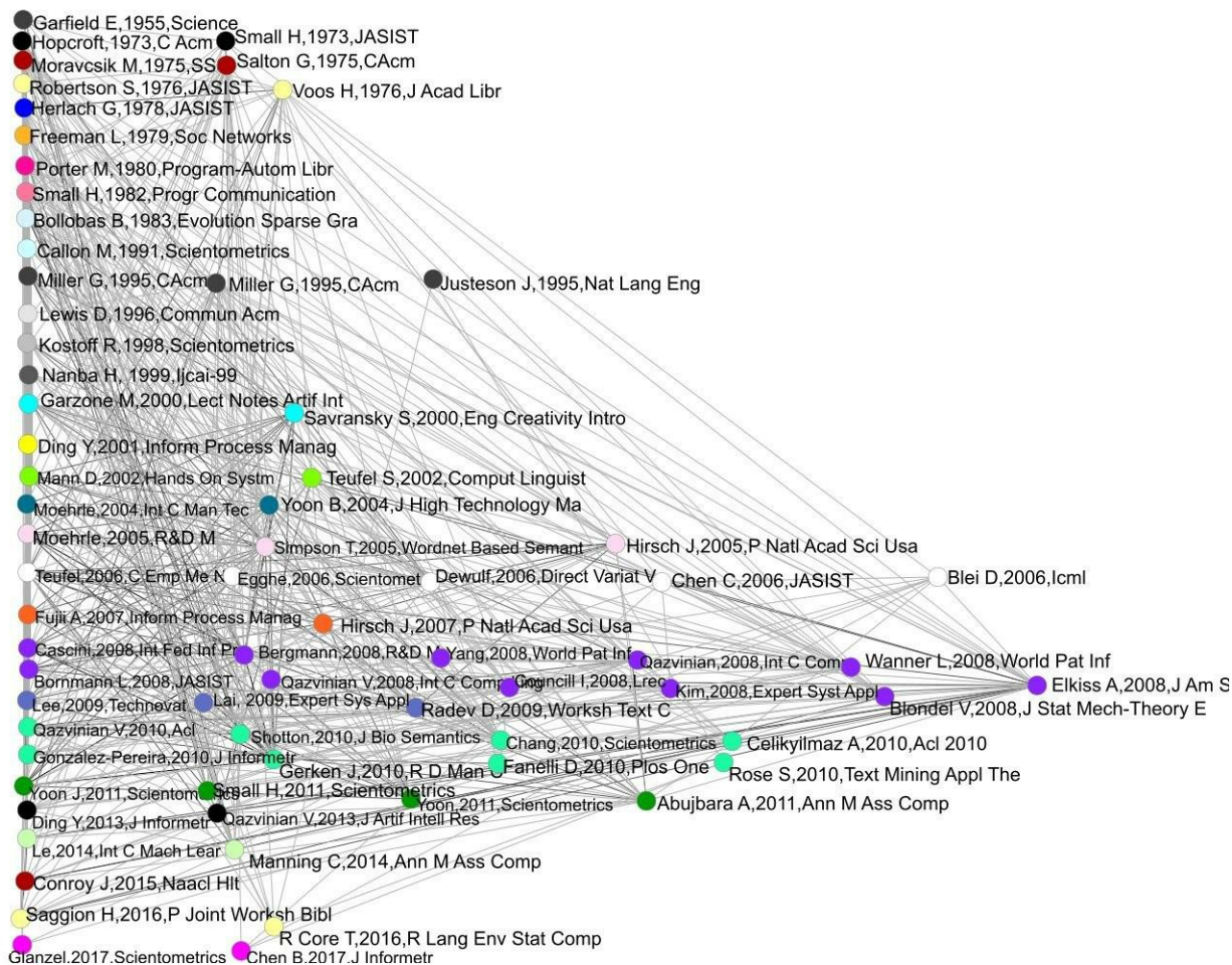
constituída por 21 sub-redes e composta por 67 autores (nós) e 102 relações de coautoria (enlaces). A baixa conexão entre as sub-redes formadas pelos autores está associada ao fato de prevalecer a colaboração intramuros na literatura analisada, com cerca de 66% delas formadas por colaborações intrainstitucionais, ou seja, 14 das 21 sub-redes visualizadas na Figura 1 são formadas por autores vinculados à mesma instituição. Das três maiores estruturas de sub-redes apresentadas na Figura 1, a primeira é constituída por colaboração intrainstitucional e as duas seguintes apresentam colaborações interinstitucionais e internacionais; estas colaborações são a minoria, presentes em apenas quatro das 21 sub-redes. As demais (3) são compostas por colaborações interinstitucionais domésticas (dentro do mesmo país).

A Figura 2 apresenta a visualização da rede de artigos (nós) com frequência de cocitação maior e igual a dois baseados no modelo de linha de tempo, o que facilita a leitura cronológica da cocitação entre eles. A espessura dos segmentos da reta proporcional à frequência de cocitação entre os dois artigos enlaçados. Os clusters de cocitação encontram-se organizados seguindo a data de publicação, com as cores dos círculos representando os anos de publicação. A premissa desta escolha é baseada no fato de os agrupamentos de cocitação revelarem as estruturas intelectuais subjacentes em um campo do conhecimento científico (CHEN; IBEKWE-SANJUAN; HOU, 2010). Esta estratégia de visualização da rede de cocitação além de delimitar a janela temporal na qual se encaixam os trabalhos cocitados ilustra a linha progressiva de relações entre os artigos, o que pode explicar melhor a difusão das ideias científicas e como elas foram se consolidando no domínio científico dos Estudos Métricos.

Seguindo o critério do ano de publicação, a linha cronológica mostra que o trabalho mais antigo citado no conjunto de artigos analisados data de 1955, correspondendo ao trabalho de E. Garfield publicado no periódico Science. Sua localização na rede confirma o fato de que constitui um dos estudos seminais e de alto impacto na área. O artigo de W. Glanzel, Heeffer e Thijs (2017), publicado na Scientometrics, corresponde ao ano mais recente presente na linha de tempo, sobre a análise lexical de publicações científicas para cientometria em nano nível. Do total de 501 pares de cocitação identificados, somente 71 são visualizados na Figura 2. Destaca-se, ainda, que a rede de cocitação está caracterizada por 619 enlaces (cocitação) em sua estrutura, o que proporciona uma densidade igual a 24%, similar aos valores usuais em outras redes de cocitação (YUE, 2010). Seu grau médio é de 17,4 e o grau nodal distribuiu-se nos grupos (k-core) presentes na Tabela 1.

Outra vantagem deste modelo de visualização é a facilidade na identificação da frente de pesquisa, definida como o conjunto de trabalhos publicados recentes (últimos cinco anos) (BOYAK; KLAVANS, 2010). A partir da rede de cocitação construída, identificaram-se, assim, os trabalhos que constituem a frente de pesquisa no campo em estudo, a saber: Manning et al.(2014); Conroy e Davis (2015); Saggion, AbuRaed e Ronzano (2016); R Core Team (2016); Glanzel (2017) e Chen et al. (2017).

Figura 2 - Linha do tempo da Rede de cocitação do PLN.



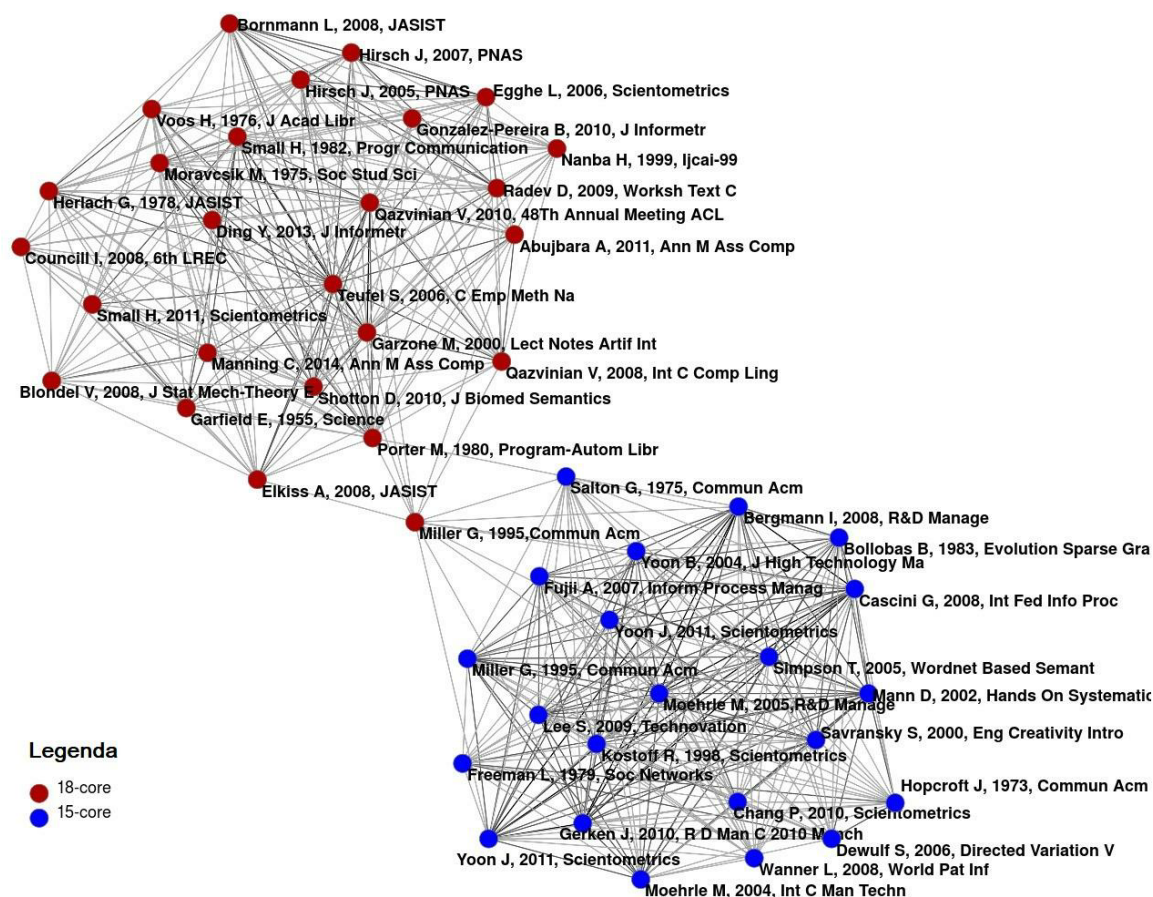
Fonte: Elaborado pelos autores.

A distribuição de grau dos nós (artigos) da rede de cocitação revela concentrações de relação em torno dos trabalhos, como grupos coesos. Todavia, não revela o comportamento dos nós com maior grau de densidade e se estão unidos ou separados na rede. Nesse sentido, recorre-se à análise do k-core, definido como o subgrupo em que os nós estão conectados a pelo menos outros k nós do subgrupo. Assim, a quantidade e a composição dos k-core em uma rede depende do valor k . Observou-se que 32,4% dos nós da rede têm um grau entre 2 e 12, ao passo que 67,6% dos nós apresentaram graus entre

15 e 18. São esses dos subgrupos coesivos, os formados por nós (artigos) com relações mais fortes, diretas, cercadas, frequentes e positivas (SEIDMAN, 1983).

Na Figura 3, representam-se os subgrafos de 18-core (cor vermelha) e 15-core (cor azul) (juntos 67,6%). Ambos representam o núcleo intelectual da cocitação no interdomínio.

Figura 3 - Rede do núcleo intelectual de cocitação no domínio.



Fonte: Elaborado pelos autores em Pajek e SVG, (2020)

A estrutura centro periferia descreve a estrutura das relações entre os documentos de uma rede, a qual consiste num fenômeno social onde se identifica a elite, neste caso de cocitação. Quanto ao indicador de densidade, destaca-se com maior grau de cocitação o par Cascini e Zini (2008) e Moehrle et al. (2005). Os trabalhos com maiores valores de cocitação localizam-se na base da rede da Figura 2. O Quadro 1 apresenta a distribuição dos cinco principais pares de cocitação.

Quadro 1 - Distribuição dos cinco pares de cocitação principais da rede.

Rank	Linha	Valor	Pares de cocitação
1	1-2	3.00000	Cascini e Zini (2008); Moehrle et al. (2005)
2	3-4	3.00000	Bergmann et al.(2008); Yoon, Choi e Kim (2011)
3	5-6	3.00000	Garzone e Mercer (2000); Qazvinian e Radev (2010)
4	1-7	3.00000	Cascini e Zini (2008); Gerken, Moehrle e Walter (2010)
5	2-3	3.00000	Moehrle et al. (2005); Bergmann (2008)

Fonte: elaborado pelos autores (2020)

Considerando a força nas relações entre esses documentos cocitados, pode-se inferir a natureza similar no nível temático ou metodológico existente entre eles, a partir da perspectiva do trabalho citante.

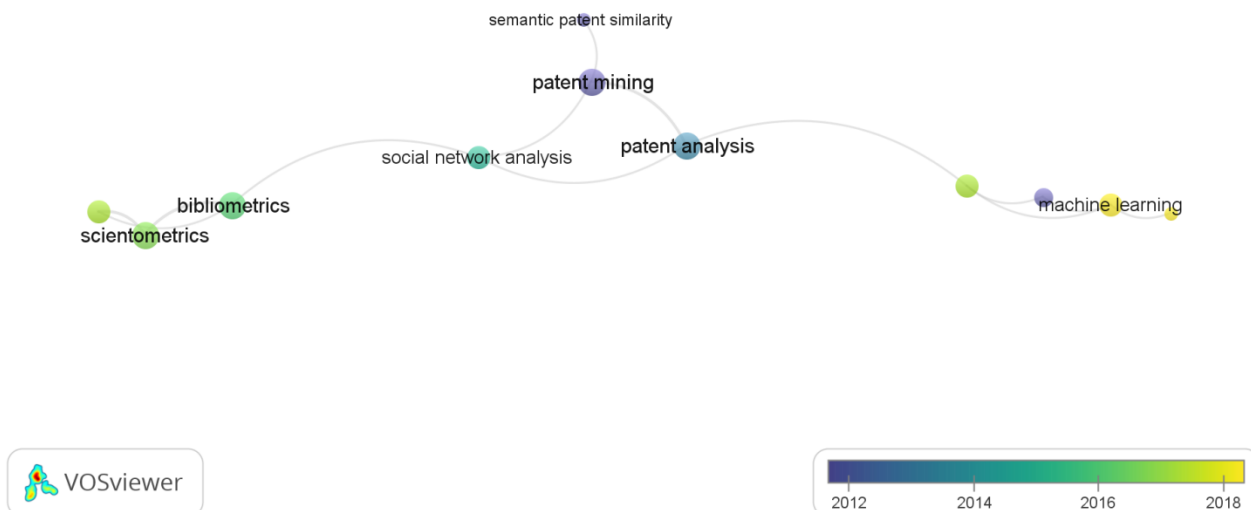
Em relação á análise das palavras-chave e Técnicas de PLN aplicadas nos EMI, a análise das palavras-chave feita no VOSviewer identificou, aplicando o método de contagem completo (VAN ECK; WALTMAN, 2020), um total de 106 termos (palavras chaves atribuídas pelo autor). No mapeamento foram selecionadas as 12 palavras chaves com frequência de duas ou mais ocorrências, as quais encontram-se agrupadas em três *clusters*. O *cluster 1* é composto por 4 palavras chaves (*information retrieval, machine learning, patente information, research topics*), o *cluster 2* é composto por 4 palavras chaves (*patent analysis, patent mining, semantic patent similarity e social network analysis*) e no *cluster 3* encontram-se as pesquisas com maior frequência dos termos *scientometric, bibliometris e informetrics*.

A Figura 4 apresenta a rede de coocorrência das 11 palavras-chave do autor com valor de ocorrência maior que dois, em que quanto maior a frequência de ocorrência de uma temática maior seu destaque (tamanho da palavra). A palavra-chave processamento de linguagem natural foi excluída para que não afetasse a análise da coocorrência das palavras a ela associadas.

O mapa segue o critério de ocorrência ao longo do tempo (*overlay visualization*), destacando a emergência das temáticas na janela temporal em estudo, as quais aparecem representadas em nós com diferentes cores. Por padrão, as cores variam de azul (pontuação mais baixa do ano médio da publicação) a verde e amarelo (pontuação mais

alta do ano médio mais recente). O tamanho dos nós é proporcional à força total do enlace² no conjunto dos textos analisados, sendo 16 o valor obtido neste ponto. Foram mantidos os valores preestabelecidos pelo software para a visualização dos clusters.

Figura 4 - Rede de coocorrência³ das palavras chave (autoria e atribuição pela WoS) nas publicações



Fonte: Elaborado pelos autores em VOSviewer, técnica de Visualização ‘*Overlay Visualization*’.

O mapa compreende a visualização das temáticas de pesquisa com maior frequência que marcam o período de análise (2000-2019). As pesquisas na primeira janela de análise (2000-2005), não visíveis no mapa pela frequência de ocorrência baixa, abordam temas mais relacionados aos sistemas de recuperação da informação ampliada (bases de dados, variações terminológicas e linguísticas, padrões de extração). Neste ponto, os autores exploram o uso de um modelo de recuperação de informação ampliada, um baseado em citações. Em meados da primeira década do século vinte e início de 2010, prevalecem estudos voltados ao debate das análises de indicadores métricos (ranking institucional e de periódicos científicos). No *cluster* 3 do tópico infometria, cientometria e bibliometria, cujas publicações se encaixam entre os anos 2014-2018, destacam estudos que aplicam ferramentas da linguagem natural no processamento do discurso científico com fins de detectar o plágio.

No centro da rede de coocorrência das palavras chave localizam-se os estudos do *cluster* 2, os que estão mais voltados às patentes como unidade de análise, destacando-se a aplicação do PLN para a análise de citações de patentes. Os autores Yoon, Choi e Kim

² Total link strength

³ Palavras com coocorrência maior o igual a 1.

(2011) reconhecem que um fator importante para identificar a possibilidade de violação de patentes é propor um algoritmo de *clustering* para sugerir automaticamente possíveis casos de violação do seu uso. Entre as técnicas de PLN empregadas na limpeza e processamento do corpus e para a extração das unidades de análise, que são as entradas dos posteriores análises lexicais e semânticas, encontram-se as seguintes: *stopwords*, *tokenization* e *word stemming*, seguidas da aplicação das técnicas de *bag of words*, *part-of-speech (POS) tagging* e o processo de *syntactic parsing*.

Adicionalmente, Yoon, Choi e Kim (2011) e Park, Yoon e Kim (2012) fazem uso do WordNet, uma base de dados lexical estruturada de forma hierárquica com relações semânticas entre as palavras (PRINCETON UNIVERSITY, 2010). O programa de PLN Knowledgist (TSOURIKOV et al., 2000) é empregado para extrair e analisar as estruturas gramaticais sujeito-ação-objeto (SAO) das patentes. Resumindo, esses estudos encontram-se no nível semântico de aplicação do PLN, substituindo a tradicional análise das palavras chave como unidades isoladas no corpus da citação de patentes.

Interessante destacar o aumento das pesquisas nos últimos três anos que focam nos diálogos do aprendizado de máquina (*machine learning*), presentes no *cluster* 1, com a mineração de dados científicos (representadas com a cor amarela na Figura 4), visando a descoberta e classificação de tópicos emergentes num domínio em estudo. Especificamente o estudo de Lupu (2017) vincula a recuperação da informação, aprendizado de máquina e o PLN para obter informações sobre propriedade intelectual. Zhu et al (2015) aplicam o pacote LIBSVM *support vector machine* (SVM), um algoritmo de aprendizado de máquina supervisionado, a fim de identificar automaticamente o subconjunto de referências na bibliografia que tem uma influência acadêmica central no artigo citante. Estudos emergentes exploram novas formas de vínculo, classificação e sumarização no contexto das citações no discurso científico, entre os quais aparecem Li et al. (2018) e Cohan e Goharian (2018). Esses autores desenvolvem uma análise de citações e sumarização automáticas, adotando o modelo de tópicos hLDA (*hierarchical Latent Dirichlet Allocation*).

Aplicam-se também diversas técnicas de PLN na menção de citações (DOLOREAU et al., 2019). Nesse estudo, os autores argumentam que há maior precisão quando se considera a frequência com que um artigo é citado dentro do mesmo texto, a fim de representar as reais relações de citação entre os artigos, em lugar de tratar todas as citações com o mesmo peso e ignorar a variedade de funções que executam. Com base

nessa premissa, revisou-se o índice-h e os autores propõem um novo índice bibliométrico, o índice WL, para avaliar o impacto científico de um indivíduo.

Nos últimos anos, adotam-se modelos e técnicas de PLN que buscam aprimorar a qualidade e a precisão dos métodos que os precedem, revolucionando as técnicas de aprendizado de máquina, os chamados métodos de aprendizado profundo (do inglês *deep learning*) (IQBAL et al, 2020). Entre os estudos pioneiros desse modelo, encontra-se o trabalho de Hassan et al. (2018), que aborda o problema da classificação das citações comparando dois modelos tradicionais de ML, SVM e RF, com o modelo LSTM, a fim de distinguir as citações influentes das incidentais. Por outro lado, existem estudos de análise de citação de patentes, que se direcionam à classificação e à análise do valor das patentes e à análise das tendências no desenvolvimento de tecnologias.

5 CONCLUSÕES

O PLN aparece na literatura da Ciência da Informação aplicado aos Estudos Métricos da Informação para melhorar o desempenho das pesquisas na disciplina e com maior presença a partir da década de 2010. Entre os principais achados no presente estudo se conclui que a literatura científica internacional sobre a aplicação do PLN nos Estudos Métricos da Informação é emergente e mostra um discreto aumento nos últimos três anos, comparado com o crescimento na Ciência da Informação.

Considerando as relações de coautoria dos 31 artigos inseridos na temática, observou-se que não compõem uma rede muito conectada, uma vez que estão fragmentados em sub-redes menores. A maioria das sub-redes limita-se à composição de dois pesquisadores, sendo baixo o grau de colaboração entre eles. Esse comportamento pode ser explicado a partir do próprio caráter emergente da inserção da temática nos EMI, ainda não consolidada. No que diz respeito às citações, os trabalhos de maior impacto provêm da Coreia do Sul, China e os Estados Unidos.

Sob a perspectiva das fontes de publicação, quatro dos cinco principais artigos de maior impacto foram publicados no periódico *Scientometrics*. Ademais, este periódico não é só o mais citado pelos autores como também é o maior citante, denotando uma alta aderência na temática PLN vinculada aos Estudos Métricos da Informação; uma característica do periódico que incide nesse resultado é a frequência de publicação de artigos no ano que se considera alta na área dos EMI. Além disso, observou-se a existência de um importante núcleo teórico e metodológico de trabalhos na rede do núcleo intelectual

de cocitação no domínio, o qual está formado por 71 pesquisas que vinculam o tema PLN nos estudos métricos da informação e frequentemente citado na comunidade acadêmica internacional.

Finalizando, após a análise do conjunto de publicações recuperadas, identificou-se, ainda, um conjunto de modelos estatísticos não supervisionados e técnicas de PLN (*bag of words*, *tokenization*, *word stemming*, *part-of-speech (POS) tagging* e *syntactic parsing*) para a aglomeração dos artigos, o modelado de tópicos nos domínios de conhecimento em estudo e a sumarização no contexto das citações no discurso científico. O LIBSVM *support vector machine* (SVM) se aplica como um algoritmo de aprendizado de máquina supervisionado. Com relação ao modelo sistemas de recuperação de informação baseados em citações, destaca-se o uso do algoritmo hLDA. No que diz respeito à mineração de patentes com fins de detecção de plágio, se faz uso da base de dados lexical Network em conjunto com o Knowledgist. Ainda, o programa de PLN torna possível medir a similaridade ou relação semântica entre pares de conceitos presentes no texto da citação de patentes.

As técnicas e modelos de PLN permitem desenvolver tarefas de processamento e análise de alta complexidade para o pesquisador no campo dos Estudos Métricos da Informação e ir além da análise de citação tradicional, para uma análise mais voltada ao conteúdo e contexto da citação. Os objetivos seguidos incluem desde determinar a função específica da citação nos estudos, como revelar as relações de citação entre os artigos, a detecção do plágio, incluindo a violação de patentes, analisar a polaridade da citação, gerar resumos automáticos baseados nas citações de um conjunto de trabalhos e construir sistemas de recuperação de informação baseados em citações.

Como limitação, pode-se indicar que devido ao crescente aumento da sua utilidade na área, seriam necessários estudos subsequentes com foco em outras perspectivas do campo estudado, como a visualização de subdomínios no interdomínio estudado, definidos em função das similaridades das identidades de citação dos artigos, via análise de acoplamento bibliográfico de autores, as lideranças científicas, via análise de atribuição do autor correspondente, entre outros, assim como análises transdisciplinares, principalmente na aplicação dos níveis mais complexos de PLN como método nas pesquisas da área.

AGRADECIMENTOS

Este trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, para bolsa de doutorado Proex/Capes, no. Processo 88887.504100/2020-00.

REFERÊNCIAS

- BERGMANN, I.; BUTZKE, D.; WALTER, L.; FUERSTE, J. P.; MOEHRLE, M. G.; ERDMANN, V. A. Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips: Evaluating the risk of patent infringement. **R&D Management**, v. 38, n. 5, p. 550–562, 2008. Disponível em: <https://doi.org/10.1111/j.1467-9310.2008.00533.x> Acesso em: 24 out. 2020.
- BOYACK, K. W.; KLAUVANS, R. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? **Journal of the American Society for information Science and Technology**, v. 61, n.12, p 2389-2404, 2010. Disponível em: <https://doi.org/10.1002/asi.21419> Acesso em: 24 out. 2020.
- CASCINI, G.; ZINI, M. Measuring patent similarity by comparing inventions functional trees. **Computer-Aided Innovation (CAI)**, v.277, p. 31–42, 2008.
- CHEN, Ch.; IBEKWE-SANJUAN, F.; HOU, J. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. **Journal of the American Society for Information Science and Technology**, v. 61, 7, p. 1386-1409, 2010. Disponível em: <https://doi.org/10.1002/asi.21309> Acesso em: 24 out. 2020.
- CHEN, B.; TSUTSUI, S.; DING, Y.; MA, F. Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics*, vol. 11, n. 4, p. 1175–1189, 2017. Disponível em: <https://doi.org/10.1016/j.joi.2017.10.00> Acesso: 24 out. 2020.
- CHEN, X., DING, R., XU, K., WANG, S., HAO, T., & ZHOU, Y. A bibliometric review of natural language processing empowered mobile computing. **Wireless Communications and Mobile Computing**, v. 2018. Disponível em: <https://doi.org/10.1155/2018/1827074>
- CHOWDHARY, K. R. Natural Language Processing. Em: CHOWDHARY, K. R. **Fundamentals of Artificial Intelligence**. New Delhi: Springer India, p. 603–649, 2020. Disponível em: http://doi.org/10.1007/978-81-322-3972-7_19 Acesso em: 02 fev. 2020.
- CHOWDHURY, G. G. Natural language processing. **Annual Review of Information Science and Technology**, v. 37, n. 1, p. 51–89, 31 Jan. 2005. Disponível em: <https://doi.org/10.1002/aris.1440370103> Acesso em: 02 fev. 2020.
- COHAN, A.; GOHARIAN, N. Scientific document summarization via citation contextualization and scientific discourse. **International Journal on Digital Libraries**, v. 19, n. 2–3, p. 287–303, Sep. 2018. Disponível em: <https://doi.org/10.1007/s00799-017-0216-8>. Acesso em: 02 fev. 2020.
- CONROY, J.M.; DAVIS, S.T. Vector space and language models for scientific document summarization. Em: **Proceedings of NAACL-HLT**, p. 186–191, 2015.
- DOLOREUX, D.; GAVIRIA DE LA PUERTA, J.; PASTOR-LÓPEZ, I.; PORTO GÓMEZ, I.; SANZ, B.; ZABALA-ITURRIAGAGOITIA, J. M. Territorial innovation models: to be or not to

be, that's the question. **Scientometrics**, v. 120, n. 3, p. 1163–1191, Sep. 2019. Disponível em: <https://doi.org/10.1007/s11192-019-03181-1>. Acesso em: 24 jun 2020.

FERREIRA, M. H. W.; CORRÊA, R. F. Estudo métrico temático sobre biblioteca digital no brasil: uma aplicação do software iramuteq. Encontro Brasileiro de Bibliometria e Cienotmetria, v. 6, p. 6º Encontro Brasileiro de Bibliometria e Cienotmetria, 2018. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/117376>. Acesso em: 24 out. 2020.

GALVEZ C; MOYA-ANEGON, F. Standardizing formats of corporate source data. **Scientometrics**, v. 70 n.1, p. 3-26, 2007. Disponível em: [10.1007/s11192-007-0101-0](https://doi.org/10.1007/s11192-007-0101-0) . Acesso em: 24 jun. 2020.

GARZONE, M.; MERCER, R. E. Towards an automated citation classifier. Em: **Advances in Artificial Intelligence**. p. 337-346, 2000.

GERKEN, J.; MOEHRLE, M.; WALTER L. Patents as an information source for product forecasting: Insights from a longitudinal study in the automotive industry. Em: **The R&D management conference**, v. 3, 2010. Disponível em: <https://jmgerken.com/publication/gerken-2010-patents/> Acesso em: 24 out. 2020.

GHIASI, G.; LARIVIÈRE, V; SUGIMOTO, C. Gender differences in synchronous and diachronous self-citations. Em: **21st International Conference on Science and Technology Indicators-STI 2016**. Book of Proceedings. 2016. Disponível em <http://ocs.editorial.upv.es/index.php/STI2016/STI2016/paper/viewFile/4543/2327> Acesso em: 03 nov. 2020.

GLÄNZEL, W.; HEEFFER, S.; THIJS, B. Lexical analysis of scientific publications for nano-level scientometrics. **Scientometrics**, v. 111, n. 3, p. 1897–1906, Jun. 2017. Disponível em: <https://doi.org/10.1007/s11192-017-2336-8>. Acesso em: 02 fev. 2020.

HASSAN SU; IMRAN, M; IQBAL, S; ALJOHANI, NR; NAWAZ, R. Deep context of citations using machine-learning models in scholarly full-text articles. **Scientometrics**, v. 117, n.3, p.1645-62, 2018.

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

HJØRLAND, B. Domain analysis. **Knowledge Organization**, v.44, n. 6, p.436-464, 2017.

IQBAL, S.; HASSAN, S. U.; ALJOHANI, N. R.; ALELYANI, S.; NAWAZ, R.; BORNMANN, L. A Decade of In-text Citation Analysis based on Natural Language Processing and Machine Learning Techniques: An overview of empirical studies. 2020. arXiv preprint Disponível em: <https://arxiv.org/abs/2008.13020>. Acesso em: 02 nov. 2020.

IOANNIDIS, J. P. A.; BAAS, J.; KLAVANS, J.; BOYACK, K. W. A standardized citation metrics author database annotated for the scientific field. **PLOS Biology**, v. 17, n. 8, e. 3000384, ago. 2019. Disponível em: <https://doi.org/10.1371/journal.pbio.3000384> Acesso em: 06 nov. 2020.

- KAMADA, T.; KAWAI, S. A general framework for visualizing abstract objects and relations. *ACM Transactions on Graphics, Connecticut*, v. 10, p. 1-39, 1991.
- LADEIRA, A. P.; ALVARENGA, L. Processamento de linguagem natural: em busca de evidências temáticas nas publicações nacionais contemporâneas. In: **Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação**, 10, 2009, João Pessoa. Anais... João Pessoa: Ancib, 2009.
- LI, L.; MAO, L.; ZHANG, Y.; CHI, J.; HUANG, T.; CONG, X.; PENG, H. Computational linguistics literature and citations oriented citation linkage, classification and summarization. *International Journal on Digital Libraries*, v. 19, n. 2–3, p. 173–190, Sep. 2018. Disponível em: <https://doi.org/10.1007/s00799-017-0219-5>. Acesso em: 02 fev. 2020.
- LI, X.; LEI, L. A bibliometric analysis of topic modelling studies (2000–2017). *Journal of Information Science*, p. 0165551519877049, 2019.
- LIDDY, E. D. **Natural language processing**. p.1-15, 2001. Disponível em: <https://surface.syr.edu/cgi/viewcontent.cgi?article=1019&context=cnlp> Acesso em: 26 Jul. 2020.
- LIDDY, E. D. Natural Language Processing for Information Retrieval. Em: BATES, M. J.; MAACK, M. N. (Eds.). **Encyclopedia of Library and Information Sciences**. CRC Press, 2010. Disponível em: <https://doi.org/10.1081/E-ELIS3>. Acesso em: 26 Jul. 2020.
- LIU, Sh.; CHEN, Ch. The effects of co-citation proximity on co-citation analysis. Em: **Proceedings of ISSI**, p. 474-484. 2011.
- LUPU, M. Information retrieval, machine learning, and Natural Language Processing for intellectual property information. *World Patent Information*, v. 49, p. A1–A3, 2017. Disponível em: <https://doi.org/10.1016/j.wpi.2017.06.002> Acesso: 26 Jul. 2020.
- MANNING, C. D., SURDEANU, M., BAUER, J., FINKEL, J., BETHARD, S. J., & MCCLOSKEY, D. The Stanford CoreNLP natural language processing toolkit. Em: **Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, p. 55-60, 2014. Disponível em: <https://www.aclweb.org/anthology/N15-3.pdf> Acesso em: 26 Jul. 2020.
- MOEHRLE, M. G; WALTER, L; GERITZ, A; MULLER, S. Patent-based inventor profiles as a basis for human resource decisions in research and development. **R and D Management**, v. 35, n. 5, p. 513–524, 2005. <https://doi.org/10.1111/j.1467-9310.2005.00408.x>. Acesso em: 26 Jul. 2020.
- NADKARNI, P. M.; OHNO-MACHADO, L; CHAPMAN, W. W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, v. 18, n. 5, p. 544-551, 2011.
- PARK, H.; YOON, J; KIM, K. Identifying patent infringement using SAO based semantic technological similarities. **Scientometrics**, v.90, n.2, p. 515-529, 2012. Disponível em: <https://doi.org/10.1007/s11192-011-0522-7> Acesso em: 2 dez. 2020.

- PRINCETON UNIVERSITY. About WordNet. WordNet. Princeton University. 2010. Disponível em <https://wordnet.princeton.edu/>. Acesso em: 26 oct. 2020.
- PUERTA-DIAZ, 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. **Anais do 7º Encontro Brasileiro de Bibliometria e Cientometria**. Salvador: EDUFBA, 2020. p. 145-152. Disponível em: <http://repositorio.ufba.br/ri/handle/ri/32385>. Acesso em: 2 dez. 2020.
- QAZVINIAN, V.; RADEV, D. R. Identifying non-explicit citing sentences for citation-based summarization. Em: **Proceedings of the 48th annual meeting of the association for computational linguistics**, p. 555-564, 2010.
- R CORE TEAM. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2016. Disponível em: <https://www.R-project.org/> Acesso: 24 out. 2020.
- SAGGION, H.; ABURAED, A.; RONZANO, F. Trainable citation-enhanced summarization of scientific articles. Em: CABANAC, G; CHANDRASEKARAN, MK; FROMMHOLZ, I; JAIDKA, K; KAN, M; MAYR, P; WOLFRAM, D.(eds). **Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)**; 2016 Jun 23; Newark, United States.CEUR Workshop Proceedings, p. 175-86, 2016.
- SEIDMAN, S. B. Network structure and minimum degree. **Social networks**, v.5 n.3, p. 269-287, 1983.
- SZOMSZOR M; PENDLEBURY DA; ADAMS J. How much is too much? The difference between research influence and self-citation excess. **Scientometrics**, v.123, n.2, p. 1119-1147, 2020.
- SMEATON, A. F. Using NLP or NLP Resources for Information Retrieval Tasks. In: STRZALKOWSKI, T. (ed.). **Natural Language Information Retrieval**. Dordrecht: Springer Netherlands, 1999. v. 7, p. 99–111. Disponível em: http://link.springer.com/10.1007/978-94-017-2388-6_4. Acesso em: 26 Jul. 2020.
- SMIRAGLIA, R. **Domain analysis for knowledge organization: tools for ontology extraction**. Chandos Publishing, p. 116, 2015.
- TASKIN, Z.; AL, U. Natural language processing applications in library and information science. **Online Information Review**, v. 43, n. 4, p. 676–690, 12 Aug. 2019. Disponível em: <https://doi.org/10.1108/OIR-07-2018-0217>. Acesso em: 26 Jul. 2020.
- TSOURIKOV, V. M.; BATCHILO, L. S.; SOVPEL, I. V. Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (SAO) structures. United States Patent No. 6167370. 2000
- VAN ECK, N. J.; WALTMAN, L. **VOSviewer manual**. Leiden: Univeriteit Leiden, v. 1, n. 1, p. 1-53, 2020.

WHITE, H. D. Authors as Citers over Time. **Journal of the American Society for Information Science and Technology**, v. 52, n. 2, p. 87–108, 2001.

YOON, J.; CHOI, S.; KIM, K. Invention property-function network analysis of patents: a case of silicon-based thin film solar cells. **Scientometrics**, v. 86, n. 3, p. 687–703, 2011. Disponível em: <https://doi.org/10.1007/s11192-010-0303-8>. Acesso em: 26 Jul. 2020.

YOON J.; KIM K. Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. **Scientometrics**, v.88 n.1, p.213-28, 2011. Acesso em: 26 Jul. 2020.

YOON J; PARK H; KIM K. Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. **Scientometrics**, v.94, n.1, p.313-31, 2013. Disponível em: <http://doi.org/10.1007/s11192-012-0830-6> Acesso em: 26 Jul. 2020.

YUE, H. Core and visualization analysis based on network of co-citation. Em: 2010 2nd **IEEE International Conference on Information Management and Engineering**. IEEE, p. 266-269, 2010. Disponível em: <http://doi.org/10.1109/ICIME.2010.5478291>. Acesso em: 26 Jul. 2020.

ZHU XD; TURNEY P; LEMIRE D; VELLINO A. Measuring Academic Influence: Not All Citations Are Equal. **Journal of the Association for Information Science and Technology**, v.66, n.2, p.408-27, 2015. Disponível em: <http://doi.org/10.1002/asi.23179> Acesso em: 26 Jul. 2020.

Notas

AGRADECIMENTOS

Não se aplica

CONTRIBUIÇÃO DE AUTORIA

Concepção e elaboração do manuscrito: M. Puerta-Díaz, B. S. Mira

Coleta de dados: M. Puerta-Díaz, B. S. Mira

Análise de dados: M. Puerta-Díaz, B. S. Mira, Maria Cláudia Cabrini Grácio, M-A Ovalle-Perandones

Discussão dos resultados: M. Puerta-Díaz, B. S. Mira, D. Martínez-Ávila, M-A Ovalle-Perandones, Maria Cláudia Cabrini Grácio

Revisão e aprovação: D. Martínez-Ávila, Maria Cláudia Cabrini Grácio, M-A Ovalle-Perandones

CONJUNTO DE DADOS DE PESQUISA

Todo o conjunto de dados que dá suporte aos resultados deste estudo foi publicado no próprio artigo.

FINANCIAMENTO

Este trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

CONSENTIMENTO DE USO DE IMAGEM

Não se aplica

APROVAÇÃO DE COMITÊ DE ÉTICA EM PESQUISA

Não se aplica.

CONFLITO DE INTERESSES

Não se aplica.

LICENÇA DE USO

Os autores cedem à **Encontros Bibli** os direitos exclusivos de primeira publicação, com o trabalho simultaneamente licenciado sob a [Licença Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) (CC BY) 4.0 International. Esta licença permite que **terceiros**



remixem, adaptem e criem a partir do trabalho publicado, atribuindo o devido crédito de autoria e publicação inicial neste periódico. Os **autores** têm autorização para assumir contratos adicionais separadamente, para distribuição não exclusiva da versão do trabalho publicada neste periódico (ex.: publicar em repositório institucional, em site pessoal, publicar uma tradução, ou como capítulo de livro), com reconhecimento de autoria e publicação inicial neste periódico.

PUBLISHER

Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no [Portal de Periódicos UFSC](#). As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES

Enrique Muriel-Torrado, Edgar Bisset Alvarez, Camila Barros.

HISTÓRICO

Recebido em: 01/09/2020 – Aprovado em: 29/12/2020 – Publicado em: 20/02/2021