

# O PROCESSAMENTO DE LINGUAGEM NATURAL NA ÁREA DOS ESTUDOS MÉTRICOS DA INFORMAÇÃO: UM ESTUDO NO PERÍODO DE 2000 A 2019

---

*Mirelys Puerta-Díaz*

Mestra

Programa de Pós-Graduação em Ciência da Informação

Universidade Estadual Paulista (Unesp)

E-mail: mirelys.puerta@unesp.br

*Bianca Savegnago de Mira*

Bacharel

Programa de Pós-Graduação em Ciência da Informação

Universidade Estadual Paulista (Unesp)

E-mail: bianca.mira@unesp.br

*María-Antonia Ovalle-Perandones*

Doutora

Universidad Complutense de Madrid

E-mail: maovalle@ucm.es

*Maria Cláudia Cabrini Grácio*

Doutora

Universidade Estadual Paulista (Unesp)

E-mail: cabrini.gracio@unesp.br

*Daniel Martínez-Ávila*

Doutor

Universidad de Carlos III de Madrid

E-mail: daniel.martinez@uc3m.es

## INTRODUÇÃO

Nos anos 1950, iniciaram-se os estudos sobre Processamento de Linguagem Natural (PLN), procurando aliar a Inteligência Artificial (IA) à Linguística. Posteriormente, com o avanço das pesquisas científicas interdisciplinares, a PLN e a Recuperação da Informação se aproximaram e 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.

No cenário bibliométrico, uma atividade desafiante constitui analisar as citações do conjunto de artigos em grande escala nos seus quatro níveis de proximidade (sentença, parágrafo, seção e artigo). (LIU; CHEN, 2011) Nesse contexto, apoiando-se em componentes da análise de texto, Glänzel e Thijs (2017) combinam técnicas baseadas em enlaces para agrupar espaços de documentos e detectar tópicos de pesquisa emergentes em larga escala.

Além disso, aspectos estatísticos, distribuições geográficas e relações de colaboração da pesquisa em computação móvel com PLN foram abordados por Chen e demais autores (2018), um aspecto que também foi mostrado por Li e Lei (2019), objetivando aprimorar o processo e métodos de avaliação da qualidade e produtividade da pesquisa do campo do PLN.

Isto posto, o presente estudo objetiva identificar a estrutura científica internacional das pesquisas que vinculam o uso do PLN no campo dos estudos métricos. Portanto, pretende-se mapear o contexto da produção científica em termos de distribuição anual, impacto e procedência, as relações de cocitação e os tópicos mais frequentes e emergentes.

## **PROCEDIMENTOS METODOLÓGICOS**

A coleta de dados foi realizada na base Web of Science em 02 de fevereiro de 2020. A busca e recuperação dos registros bibliográficos com a expressão “natural language processing” foi utilizada nos campos de título, resumo e palavras-chave, limitada ao período de 2000 a 2019, para o tipo documentário artigos e revisão e refinada pelas categorias da Web of Science (Information Science Library Science). Recuperaram-se 502 documentos, dos quais, após a análise dos resumos, palavras-chaves e fontes de publicação, observou-se que apenas 31 estavam inseridos no campo dos estudos métricos, ou seja, os estudos métricos representam cerca de 6,2% dos documentos recuperados a partir da busca por “Natural Language Processing” na base Web of Science (WoS) entre os anos de 2000 e 2019.

Foram analisados indicadores bibliométricos de produção e de citação, por meio da análise das referências e das citações recebidas, a fim de caracterizar as tendências presentes nas pesquisas que vinculam o PLN 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. Teve como critério para a seleção dos pares de cocitação e seu posterior mapeamento os trabalhos com frequência de citação maior que 1. O processo de extração das informações e criação da matriz foi realizado no software Bibexcel versão 2017. O software Pakej foi usado para a análise e visualização da rede de cocitação e o algoritmo utilizado para a visualização da rede foi Kamada-Kawai (componentes separados), com a perspectiva tradicional de análise

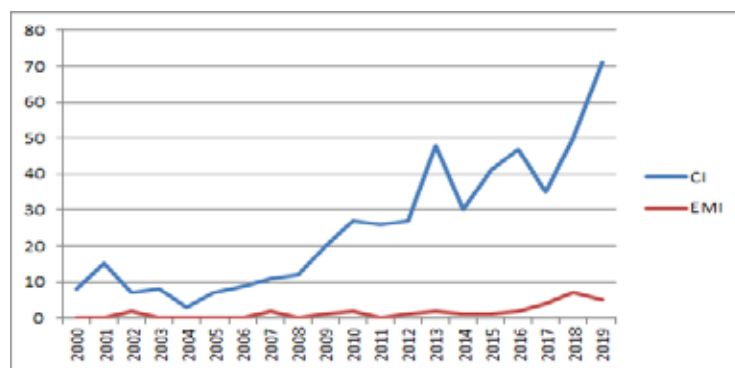
de cocitação em nível de documentos. O Vosviewer 1.6.14 foi usado para criar a rede de temáticas a partir das palavras-chave extraídas dos 31 documentos.

## RESULTADOS E DISCUSSÃO

Observou-se, inicialmente, que quase metade (15) dos documentos foram publicados no periódico *Scientometrics*. O primeiro artigo sobre o tema foi publicado em 2007, em coautoria de C. Galvez e Félix de Moya-Anegón, 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 demais fontes (6) publicaram apenas uma pesquisa cada.

A evolução temporal das publicações (Figura 1) mostra o crescente interesse da comunidade científica internacional no uso de técnicas e algoritmos PLN aplicados nos estudos métricos da informação e na Ciência da Informação (CI). Destaca-se 2018 como o ano mais produtivo das pesquisas em PLN com aderência nos Estudos Métricos em Informação (EMI), em que se aborda o problema da imprecisão dos textos de citação e a possibilidade de se identificar com maior precisão o contexto na análise de citações mediante o uso de algoritmos de PLN.

Figura 1 - Distribuição das publicações científicas sobre PLN (2000 - 2019) na WoS



Fonte: elaborado pelos autores.

A análise do *corpus* de pesquisa revelou oito pesquisadores com atuação mais consolidada na temática, a saber: Yoon e Kim, com quatro trabalhos cada, e Song, Yan Thijs, Park, Li e Glänzel, com duas publicações cada. Em termos de produtividade por país destacam-se Coreia do Sul, presente em 19,3% dos trabalhos, Espanha e Estados Unidos em 16%, França em 12,9%, Canadá e China, ambos em 9,7%; o Brasil aparece em um dos 31 trabalhos.

Dos 31 trabalhos, 77% foram citados. Destacam-se cinco estudos com as maiores quantidade de citações recebidas: Yoon e Kim (2011) *Scientometrics* (87); Yoon, Park e Kim (2013) *Scientometrics* (60), Zhu e demais autores (2015) *JASIST* (51), Park, Yoon e Kim (2012) *Scientometrics* (50); Yoon, Choi e Kim (2011) *Scientometrics* (38). Na interseção dos autores mais produtivos e aqueles com as maiores quantidades de citações, encontram-se Yoon, Kim e Park. Ainda, o *Scientometrics* aparece como a principal fonte dos trabalhos de maior impacto. Esse comportamento se mantém na totalidade das citações recebidas.

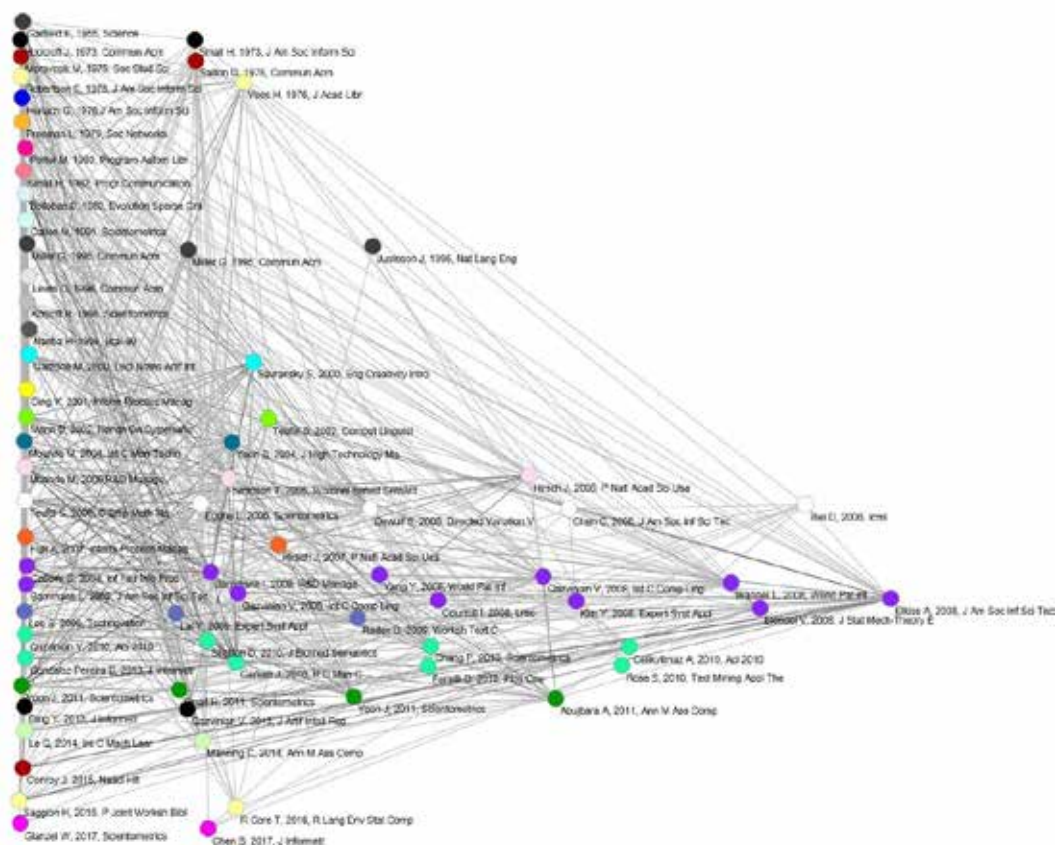
Destacaram-se Coreia do Sul, China e Estados Unidos como os principais países citantes de três (dos cinco) trabalhos em que Yoon aparece como autor. Outro trabalho (PARK; YOON; KIM, 2012) apresenta uma distribuição similar, com a Coreia do Sul e China os principais países citantes e a Alemanha em terceiro lugar, seguida dos Estados Unidos. Destaca-se que o Brasil aparece como o sexto principal país citante do artigo de Yoon, Park e Kim (2013), sendo este o único artigo em que o país aparece entre os citantes, de autoria dos citantes filiada ao Centro Tecnologia Informática e Matemática da UFSCar.

A distribuição geográfica diferente para as citações recebidas é observada para o artigo de Zhu e demais autores (2015), com a maior parte proveniente da China, seguida da Inglaterra e Estados Unidos. Diferente dos demais, a Coreia do Sul ocupa o 12º lugar entre os citantes.

Quanto à fonte citante dos cinco artigos, destaca-se o periódico *Scientometrics*, ao ocupar o primeiro lugar no número de citações realizadas em quatro dos cinco artigos, sendo a maior em Zhu, Turney, Lemire e Vellino (2015). Desse modo, este periódico é simultaneamente o mais citado pelos cinco artigos e também sua principal fonte citante.

A Figura 2 apresenta a rede de cocitação, em que os nós representam os trabalhos com frequência de cocitação maior ou igual a dois, a espessura das ligações é proporcional à frequência de cocitação entre eles. As cores dos nós representam os anos de publicação, partindo da premissa de que a linha temporal para os *clusters* de cocitação facilita a visualização da estrutura intelectual subjacente de forma diacrônica. (CHEN; IBEKWE-SANJUAN; HOU, 2010)

Figura 2 - Linha do tempo da Rede de cocitação dos trabalhos que vinculam PLN nos EMI

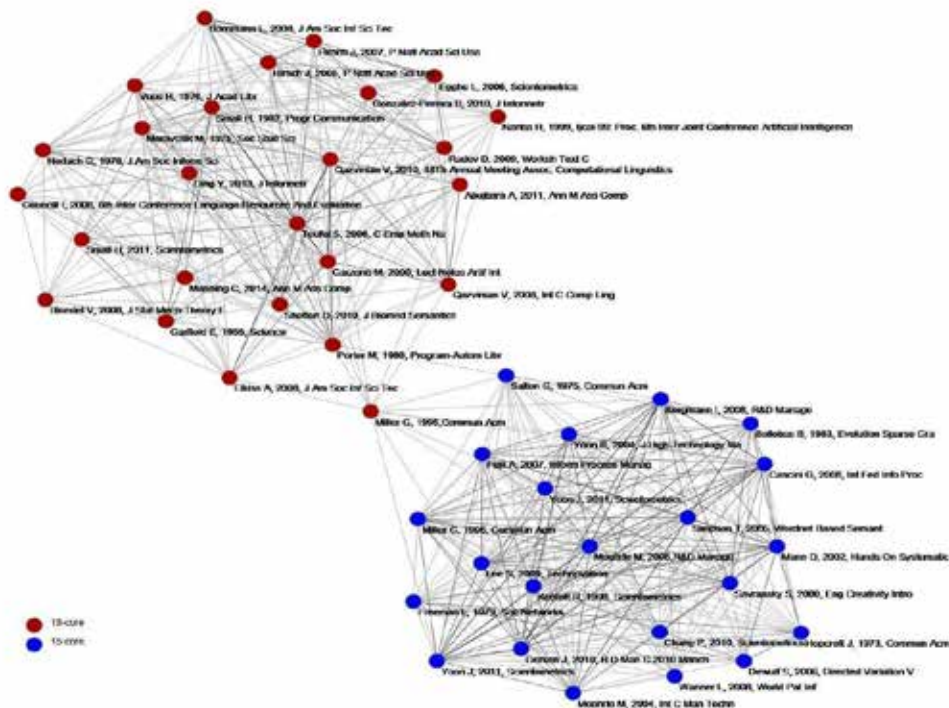


Fonte: elaborado pelos autores.

A Figura 2 evidencia 619 ligações em sua estrutura, o que proporciona uma densidade igual a 24%, similar aos valores usuais em outras redes de cocitação. (YUE, 2010) O extremo superior visualiza o nó mais antigo, enquanto o extremo inferior o mais atual.

A Figura 3 apresenta dois subgrafos da Figura 2: 18-core (cor vermelha) e 15-core (cor azul) (juntos 67,6%), que representam o núcleo intelectual da cocitação no campo estudado. A estrutura centro periferia descreve a estrutura das relações entre os nós da rede, a qual consiste num fenômeno social onde se identificam os autores com maior número de publicações, neste caso de cocitação.

Figura 3 - Rede do núcleo intelectual de cocitação no domínio



Fonte: elaborado pelos autores.

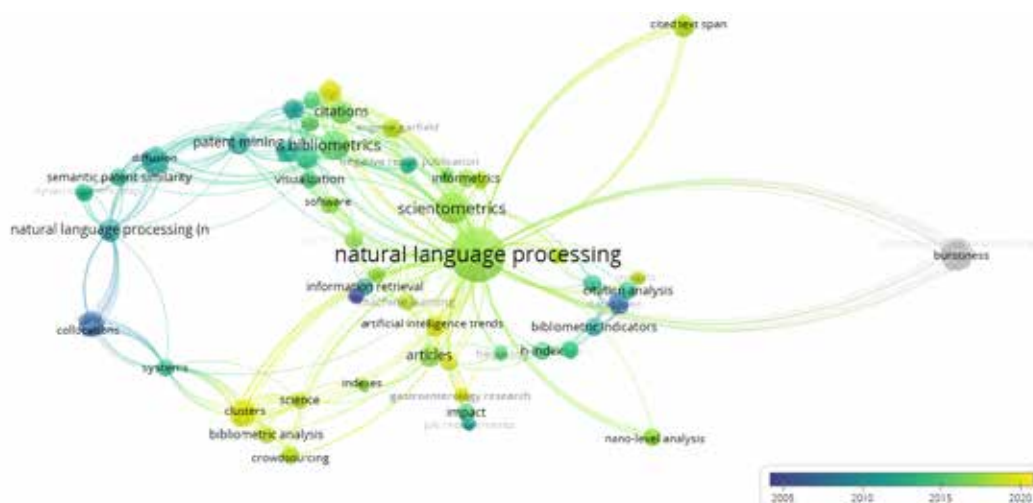
A linha cronológica mostra 1955 como a menor data associada a um nó na rede, correspondente ao trabalho de Garfield E, publicado no periódico *Science*. Sua localização na rede confirma o fato de constituir um dos estudos seminais e de alto impacto na área. O nó do artigo de Glanzel W, 2017, *Scientometrics* corresponde ao último ano na linha de tempo, sobre análise lexical de publicações científicas para Cientometria em *nano nível*.<sup>1</sup> Quanto ao indicador de densidade, destaca-se o par de cocitação Cascini G, 2008, *Int Fed Info Proc*- Moehrl M, 2005, *R&D Manage*.

A análise das palavras-chave identificou 169 termos com frequência entre 1 e 31 ocorrências, agrupados em 14 *clusters*. A Figura 4 apresenta a rede de coocorrência das palavras-chave, em que quanto maior a frequência de ocorrência de uma temática maior seu destaque (tamanho da palavra). O mapa segue o critério de ocorrência ao longo do tempo, destacando a emergência das temáticas na janela temporal em estudo, as quais aparecem representadas em nós com diferentes cores. O tamanho dos nós é proporcional à frequência de ocorrência no conjunto dos textos analisados. As pesquisas que marcam a primeira janela de análise (2000-2005) abordam temas relacionados aos

<sup>1</sup> Considera-se o nano nível como o micro nível na Bibliometria mas numa escala menor. Vários autores como Cronin tem usado o termo nano-bibliometria para a avaliação de autores individuais. (GLÄNZEL; THIJJS, 2017)

sistemas de recuperação da informação (bases de dados, variações terminológicas e linguísticas, padrões de extração). Nos meados da primeira década do século XX e início dos anos 2010, prevalecem estudos voltados ao debate das análises de indicadores métricos (ranking institucional e de periódicos científicos). Destaca-se o aumento das pesquisas nos últimos três anos que focam nos diálogos do *machine learning* na mineração de dados científicos (representadas com a cor amarela), visando a descoberta e classificação de tópicos emergentes num domínio em estudo.

Figura 4 - Rede de coocorrência das palavras chave nas publicações analisadas



Fonte: elaborado pelos autores.

Uma temática no âmbito da cientometria que se destaca é a aplicação de ferramentas da linguagem natural no processamento do discurso científico com fins de detectar o plágio. Nessa linha, destaca-se a aplicação do PLN para a análise de citações nas patentes. Os autores Yoon, Choi e Kim (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.

Observa-se, ainda, um *cluster* com baixa frequência de ocorrência, embora suas palavras-chave descrevam uma das temáticas emergentes relacionadas ao PLN no escopo dos Estudos Métricos. Li e demais autores (2018) e Cohan e Goharian (2018) exploram novas formas de vínculo, classificação e sumarização no contexto das citações no discurso científico.

Aplicam-se também diversas técnicas de PLN na menção de citações, 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.

## CONCLUSÕES

A literatura científica internacional sobre a aplicação do PLN nos estudos métricos da informação é emergente e mostra um crescente aumento nos últimos três anos. O ano de 2018 se destaca como o mais produtivo em relação à temática. Sob a perspectiva da fonte das publicações, quatro dos cinco principais trabalhos de maior impacto correspondem ao periódico *Scientometrics*. O periódico é também o principal citante, tendo assim uma alta aderência na temática PLN vinculada aos estudos métricos da informação. As citações aos trabalhos de maior impacto provêm da Coreia do Sul, China e Estados Unidos.

Finalizando, observou-se a existência de um importante núcleo teórico que vincula o tema PLN nos estudos métricos da informação, frequentemente citado na comunidade acadêmica internacional.

## REFERÊNCIAS

- CHEN, X.; DING, R.; XU, K. et al. A bibliometric review of natural language processing empowered mobile computing. *Wireless Communications and Mobile Computing*, Chichester, v. 2018, p. 1-21, 2018.
- CHEN, C.; IBEKWE-SANJUAN, F.; HOU, J. The structure and dynamics of cocitation clusters: a multiple-perspective co-citation analysis. *Journal of the American Society for Information Science and Technology*, New York, v. 61, n. 7, p. 1386-1409, 2010.
- GLÄNZEL, W.; THIJS. Lexical analysis of scientific publications for nano-level scientometrics. *Scientometrics*, Dordrecht, v. 111, n. 3, p. 1897-1906, 2017.
- LI, X.; LEI, L. A bibliometric analysis of topic modelling studies (2000-2017). *Journal of Information Science*, Amsterdam, p. 1-15, 2019.
- LIU, S.; CHEN, C. The effects of co-citation proximity on co-citation analysis. In: CONFERENCE OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS, 13., 2011, Durban. *Anais [...]*. Durban: ISSI, 2011. p. 474-484.
- NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, Philadelphia, v. 18, n. 5, p. 544-551, 2011.
- YUE, H. Core and visualization analysis based on network of co-citation. In: INTERNATIONAL CONFERENCE ON INFORMATION MANAGEMENT AND ENGINEERING, 2., 2010, Chengdu. *Anais [...]*. Chengdu: IEEE, 2010. p. 266-269.