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

PUBLICAÇÕES EM ACESSO ABERTO:
identificando a representatividade dos principais tópicos de pesquisa

Patrícia Mascarenhas Dias[1]
Thiago Magela Rodrigues Dias[1]
Gray Farias Moita[1]

Resumo: A publicação de artigos em periódicos de acesso aberto surge como um interessante mecanismo para a divulgação de pesquisas científicas, já que facilita e viabiliza o acesso a elas, tendo em vista que não existem barreiras, em especial financeiras, para acessar os conteúdos desse tipo de publicação. Logo, este trabalho visa realizar um estudo sobre a frequência dos principais tópicos de pesquisa utilizados no conjunto de publicações emperiódicos de acesso aberto por pesquisadores do Brasil. Como resultados das análises foi possível verificar como alguns tópicos são frequentemente utilizados pelos pesquisadores brasileiros na descrição de seus estudos.

Palavras-Chave: Tópicos de Pesquisa. Produção Científica. Mineração de Texto.

1 INTRODUÇÃO

A publicação científica em acesso aberto faz parte de um cenário mais amplo em prol da abertura do conhecimento em geral (acesso aberto, dados abertos, recursos educacionais abertos, software livre, licenças abertas) e constitui essencialmente um movimento em direção à concepção da informação e do conhecimento como bens públicos (FURNIVAL; SILVA-JEREZ, 2017).

Tendo em vista que grande parte das pesquisas científicas no país é financiada com recursos públicos, geralmente em instituições de ensino ou centros de pesquisa públicos, é de se esperar que os resultados de tais estudos sejam divulgados sem nenhum tipo de barreira, principalmente financeira. Nesse contexto, aliados às vantagens que as publicações em acesso aberto apresentam, como disponibilidade, visibilidade e acessibilidade, diversos esforços estão sendo empregados para que cada vez mais artigos científicos sejam publicados em periódicos de acesso aberto.

Diante disso, compreender quais os principais tópicos de pesquisa estão sendo investigados nos atigos publicados em periódicos de acesso aberto possibilita identificar um panorama das

[1] Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)

principais temáticas estudadas. Permite, ainda, verificar a representatividade de determinados tópicos presentes nos artigos analisados.

2 DESENVOLVIMENTO

A bibliometria tem como objetivo desenvolver padrões e modelar matematicamente os processos para as medições e, a partir dos resultados, traçar previsões e tomar as possíveis decisões. Por meio de suas técnicas, a bibliometria procura estudar os aspectos quantitativos da ciência e da produção científica como uma atividade que envolve caracterísitcas sociais, econômicas e políticas. Ela fornece um instrumental para estudos que visam mapear o conhecimento científico e extrair informações, bem como a compreensão de como a produção científica tem sido realizada (HAYASHI, 2012).

Dentre as principais leis bibliométricas, tem se a Lei de Zipf. A Lei de Zipf está relacionada à frequência de ocorrência de palavras em um dado texto. Essa lei desenvolveu e estendeu uma lei empírica observada por Estoup em 1916, a qual estabelece uma relação entre a posição de uma palavra e a frequência de seu aparecimento em um texto longo. A Lei de Zipf é assim formulada: r .f = c, sendo que "r" é a posição da palavra, "f" é a frequência e "c" é uma constante. Zipf extraiu sua lei de um princípio geral do "esforço mínimo", segundo o qual uma palavra cujo custo de utilização seja pequeno ou cuja transmissão demande um esforço mínimo é frequentemente usada em um texto grande (KLEINUBING, 2010).

No contexto deste trabalho, no intuito de melhor compreender os principais tópicos de pesquisa, investigados pelos pesquisadores brasileiros em periódicos de acesso aberto, foi utilizado o repositório de dados currículares da Plataforma Lattes.

Para a definição do conjunto de dados a ser analisados neste trabalho, optou-se por extrair as palavras dos títulos dos artigos publicados em periódicos de aceso aberto (a saber 2.090.015 artigos). Para a identificação dos artigos a serem considerados foi realizado um cruzamento com os periódicos de cada artigo com a relação de periódicos de acesso aberto obtida no DOAJ. A escolha da extração das palavras dos títulos dos artigos em detrimento das palavras-chave vinculadas aos artigos, se deve ao fato de que aproximadamente, apenas 17% dos artigos analisados possuíam palavras-chave vinculados a eles. Além disso, diversos trabalhos têm utilizados as palavras dos títulos das publicações como objeto de análise (CUNHA *et al.*, 2013; VINKERS *et al.*, 2015; MRYGLOD *et al.*, 2016; RONDA-PUPO, 2016).

Além disso, tendo em vista que o cadastramento das palavras-chave de um artigo científico em seus currículos é de inteira responsabilidade dos respectivos pesquisadores, e isso é feito livremente por eles, significa que podem ser inseridos quaisquer conjuntos de caracteres como uma palavra-chave. A partir disto, geralmente tem-se uma coleção muito grande de palavras-chave e sem nenhum padrão (GOMES, 2018).

Logo, para as análises aqui realizadas, os títulos de todas as publicações do conjunto identificado foram considerados. Os títulos passaram por um processo de tratamento de dados que visou identificar as palavras que que posteriormente serão objeto de análise. Todas as etapas do processo de tratamento podem ser visualizadas na Tabela 1.

Tabela 1 – Etapas do Processo de Tratamento dos Dados

ETAPA DA DO ALGORITMO	RESULTADO
Recebimento do Título	UMA ESTRATÉGIA PARA IDENTIFICAÇÃO DE ARTIGOS EM PERIÓDICOS DE ACESSO ABERTO NA PLATAFORMA LATTES.
LowerCase	uma estratégia para identificação de artigos em periódicos de acesso aberto na plataforma lattes
StopWords_PT	estratégia identificação artigos periódicos acesso aberto plataforma lattes
StopWords_EN	estratégia identificação artigos periódicos acesso aberto plataforma lattes
Identificação de Termos	estratégia
	identificação
	artigos
	periódicos
	acesso
	aberto
	plataforma
	lattes

Fonte: Elaborado pelos autores.

Como pode ser observado, para cada artigo seu título é recuperado e dessa forma, o processo de tratamento dos dados é inicializado. Na etapa de LowerCase, todas as palavras são convertidas para minúsculo com a proposta de padronizar o conjunto, bem como, evitar que palavras sejam mapeadas em tópicos distintos por algumas possuírem letras em maiúsculo e outras não. Já no processo de remoção de stopWords (StopWords_PT e StopWords_EN), são removidos todos os termos que não possuem valores semânticos significativos para caracterizar um tópico de pesquisa e, com isso, diminuir o volume de palavras a serem processadas e analisadas. Foram removidos os stopWords inicialmente em português e

posteriormente em inglês, tendo em vista que são os idiomais mais utilizados, conforme já apresentado.

Como o objeto inicial de análise é o título dos artigos, em que, se existe uma preocupação com a descrição geral do estudo a ser apresentado, a quantidade de stopWords é significativa, diferentemente das palavras-chave, justificando a remoção para as análises a serem realizadas. Após, na última etapa Identificação de Termos as palavras são separadas em tópicos, que irão compor um dicionário de termos para a contagem das frequências.

3 RESULTADOS

Inicialmente foram identificados 28.636.958 tópicos, considerando todas as palavras dos títulos dos artigos. Após a remoção das duplicatas o conjunto foi reduzido a 423.364 palavras únicas. Posteriormente, com a remoção das stopWords, o conjunto passou a ter um total de 393.896 palavras que se tornaram objeto de análise.

Considerando a característica dos títulos das publicações que em geral necessitam da utilização de stopWords na sua composição, todos os primeiros 15 termos identificados são stopWords em português ou inglês.

Aproximadamente 64% das publicações em periódicos de acesso aberto analisadas neste trabalho são em português, logo se justifica uma quantidade considerável de termos neste idioma. A Tabela 2 apresenta o resultado da extração e ordenação pela frequência das palavras dos títulos de cada artigo analisado, após todo o tratamento dos dados.

Tabela 2 – Distribuição das palavras por posição (x) e suas frequências (y)

Posição (x)	Frequência(y)	Palavras
1	88.485	brazil
2	85.470	brasil
3	72.100	estudo
4	71.618	avaliação
5	69.314	análise
6	58.977	saúde
7	46.863	educação
8	44.131	study
9	43.411	brazilian
10	42.365	rio
11	41.411	patients
12	38.754	diferentes
13	37.788	produção
14	37.586	ensino

15	37.395	estado
⋮	⋮	⋮
393.894	1	zzaa
393.895	1	zzgam
393.896	1	zzgamma

Fonte: Elaborado pelos autores.

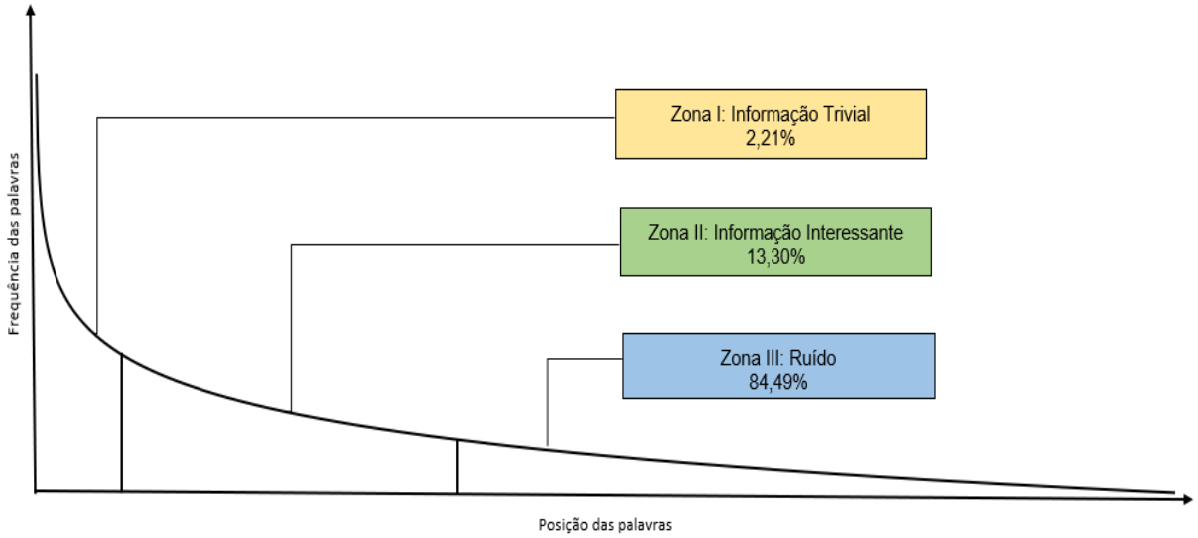
Como pode ser observado, mesmo após a remoção das stopWords é possível verificar que dentre as palavras mais frequentes, a maioria está em português, com algumas destas palavras em sua versão em inglês, como por exemplo, as duas palavras mais frequentes. Já nas últimas posições, se encontram palavras com uma frequência muito baixa. Percebe-se que tais palavras não possuem conteúdo semântico, sendo uma hipótese para a existência de tais palavras, erros de digitação no momento de cadastro do título da publicação em um determinado currículo. Percebe-se ainda, que dentre as palavras mais frequentes, se encontram tópicos que geralmente fazem parte dos títulos das publicações, já que são importantes para indicar métodos, técnicas, objetos ou localidades.

No intuito de avaliar o conjunto de palavras que estão vinculadas as publicações de artigos em periódicos de acesso aberto, utilizou-se a Lei de Zipf. No trabalho de Quoniam (1992), o autor descreve a curva de Zipf, em que a mesma é dividida em três zonas de distribuição:

- Zona I - Informação trivial ou básica: define os temas centrais da análise bibliométrica;
- Zona II - Informação interessante: localiza-se entre as Zonas I e III e mostra os temas periféricos, uma informação potencialmente inovadora. É aí que as transferências de tecnologia relacionadas aos novos temas devem ser consideradas;
- Zona III - Ruído: tem como característica possuir conceitos ainda não emergentes, onde é impossível afirmar se eles serão emergentes ou se são apenas ruído estatístico.

Neste contexto, o conjunto de palavras identificadas nesta tese, após todo o tratamento de dados já apresentado, foi dividido em três zonas textuais de distribuição (Figura 1).

Figura 1 – Divisão das Palavras Identificadas nas Três Zonas Textuais.



Fonte: Elaborado pelos autores.

A primeira zona identificada (Zona I), possui 2,21% das palavras analisadas, tais palavras que são as mais frequentes, descrevem quais são os temas centrais do conjunto analisado. Apesar de contemplar um baixo percentual de palavras, a frequência delas, corresponde a 47,93% de todo o conjunto, comprovando a sua representatividade. Já na Zona II, que possui 13,3% das palavras, engloba um conjunto de tópicos que ocorrem em menor frequência que os da Zona I, e por não serem palavras utilizadas com tanta fregquência são caracterizadas como temas emergentes, já que se caracterizam como informação potencialmente inovadora. Por fim, na Zona III que possui a grande maioria das palavras (84,49%), se caracteriza por agregar tópicos com baixa frequência, considerados ruídos. Aqui, cabe destacar como já apontado, os problemas originados do livre cadastramento dos dados das publicações por parte dos indivíduos em seus currículos, em que é real a inserção de dados incorretos, seja por erros de digitação ou até mesmo de codificação no momento de copiar um texto de documentos em diversos formatos. Um total de 149.891 palavras possuem apenas uma ocorrência.

Como pode ser observado, existe uma grande disparidade entre os conjuntos de palavras que compõem cada uma das Zonas identificadas. No intuito de melhor compreender cada uma destas Zonas, diversas técnicas de análise e visualização de dados podem ser aplicadas. A Figura 2, apresenta uma nuvem de palavras da Zona 1.

Figura 2 – Zona I: Informação Trivial ou Básica

Fonte: Elaborado pelos autores.

Observa-se que diante das palavras mais frequentes no conjunto analisado, as palavras “brazil” e “brasil” se destacam significamente. Como uma hipótese, pode-se inferir que tais palavras são amplamente utilizadas por indicar localidades de aplicação das pesquisas, principalmente por considerar praticamente em sua totalidade artigos publicados por brasileiros, nos idiomas inglês e português. Além disso, desta-se ainda as palavras “estudo”, “avaliação”, “análise”, “study”, e “analysis”, que geralmente indicam métodos utilizados para a realização das pesquisas. Importante também, destacar as palavras “saúde” e “educação” com 58.977 e 46.863 ocorrências respectivamente, apresentando-se também como tópicos muito representativos nas pesquisas realizadas. Ressalta-se aqui, a ocorrência de outras palavras como “rio”, “paulo”, “caso” e “meio” que também possuem frequência significativa, mas que podem ter sofrido influência do método utilizado para identificar as palavras dos títulos, tendo em vista que são palavras que também podem ter sido derivadas de palavras compostas.

4 CONCLUSÕES

Resslata-se que no estudo aqui apresentado, foi adotada a Lei de Zipf no intuito de identificar os principais tópicos de pesquisa dos pesquisadores brasileiros em publicações em periódicos de acesso aberto. Para tanto, foi inicialmente no repositório de dados curriculares da Plataforma Lattes, todo sos artigos publicados neste meio de divulgação. Considerando que “Brazil e Brasil” indicam localidades, e que os termos “Estudo, Avaliação e Análise” indicam métodos utilizados, destacam-se neste contexto, os termos “Saúde e Educação” como os mais representativos no conjunto analisado. Além disso, também é importante destacar que não foi

possível realizar a unificação das palavras no singular e plural, bem como, a utilização de palavras compostas, tendo em vista que seriam necessárias a adoção de técnicas como por exemplo de Processamento de Linguagem Natural, como radicalização e n-gramas que vão além do escopo deste trabalho.

REFERÊNCIAS

CUNHA, M. V. *et al.* Redes de títulos de artigos científicos variáveis no tempo. *In*: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2., 2013, Maceió. **Anais** […]. Maceió: SBC, 2013. p. 194-205.

FURNIVAL, A. C. M.; SILVA-JEREZ, N. S. Percepções de pesquisadores brasileiros sobre o acesso aberto à literatura científica. **Informação & Sociedade: Estudos**, Campina Grande, v. 27, n. 2, 2017.

GOMES, J. O. **Uma análise temporal dos principais tópicos de pesquisa da ciência brasileira a partir das palavras-chave de publicações científicas**. 2018. 127 f. Tese (Doutorado em Modelagem Matemática e Computacional) — Centro Federal de Educação Tecnológica de Minas Gerais, Dezembro 2018.

HAYASHI, M. C. P. I. Sociologia da Ciência, Bibliometria e Cientometria: Contribuições para a Análise da Produção Científica. *In*: SEMINÁRIO DE EPISTEMOLOGIA E TEORIAS DA EDUCAÇÃO, 4., 2012, São Paulo. **Anais** [...]. São Paulo: Episted, 2012.

KLEINUBING, L. S. Análise bibliométrica da produção científica em gestão da informação na base de dados LISA. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 8, n. 2, p.1-11, 2010.

MRYGLOD, O. *et al.* Quantifying the evolution of a scientific topic: reaction of the academic community to the Chornobyl disaster. **Scientometrics**, Budapest, v. 106, n. 3, p. 1151-1166, 2016.

QUONIAM, L. Bibliométrie sur des référence bibliographiques: methodologie. *In*: DESVALS H.; DOU, H. (org.). **La veille technologique**. Paris: [*s.n.*], 1992. p. 244-262.

RONDA-PUPO, G. A. Knowledge map of Latin American research on management: Trends and future advancement. **Social Science Information**, London, v. 55, n. 1, p. 3-27, 2016.

VINKERS, C. H. *et al.* Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. **BMJ**, London, v. 351, 2015.