

REDES DE PALAVRAS-CHAVE COMO MECANISMO PARA A IDENTIFICAÇÃO DOS TÓPICOS DE INTERESSES DOS PESQUISADORES BRASILEIROS

Jether Oliveira Gomes

Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, MG, Brasil
jethergomes@yahoo.com.br

Thiago Magela Rodrigues Dias

Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, MG, Brasil
thiagomagela@gmail.com

Gray Farias Moita

Centro Federal de Educação Tecnológica de Minas Gerais, MG, Brasil
gray@dppg.cefetmg

1 INTRODUÇÃO

O grande aumento do número de publicações científicas disponíveis na internet tem atraído interesses de pesquisadores de todos os domínios, que visam compreender como o conhecimento científico tem evoluído. São encontrados na literatura diversos trabalhos com finalidades distintas, como: saúde pública no Brasil (PEREIRA et. al., 2007); identificação de especialistas (CHAGAS et. al., 2015); e; redes sociais acadêmicas (MENA-CHALCO et. al., 2012). Esses estudos podem servir à diversos propósitos, tais como fornecer base para a construção de políticas de incentivo a pesquisa visando novos avanços na ciência.

Também vem se destacando a análise das palavras-chave das publicações científicas, pois as mesmas foram selecionadas cuidadosamente por seus autores com o foco de evidenciar os principais assuntos que permeiam o trabalho de forma clara e objetiva (MCCLOSKEY, 1988).

Neste contexto, o repositório da Plataforma Lattes é tido como um diferencial (LANE, 2010). Esse repositório é composto de dados re-



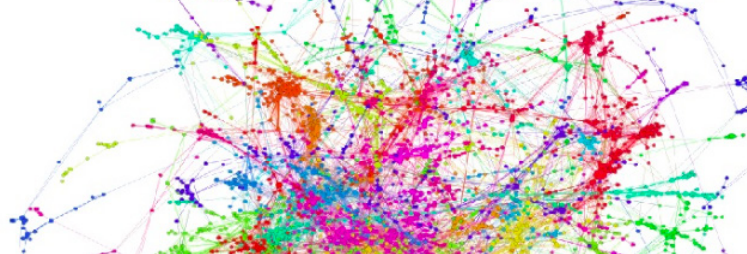
ferentes a grupos de pesquisas, instituições e de currículos de mais de cinco milhões de indivíduos (DIAS, 2016). O grande volume de dados existente na base curricular da Plataforma Lattes encontra-se disponível livremente na internet, e, ainda não foi amplamente analisado (DIGIAMPIETRI, 2015).

Sendo assim, este trabalho tem o objetivo de mapear o conhecimento científico através da rede de palavras-chave formada a partir dos dados de 55 anos de trajetória da ciência brasileira de pesquisadores doutores que possuem currículos cadastrados na Plataforma Lattes. Para em seguida, identificar os principais tópicos de pesquisas estudados de acordo com o grau de cada vértice, e, concomitantemente, possibilitar uma análise comparativa com as palavras-chave ranqueadas pela frequência, medida de importância amplamente utilizada na literatura.

2 DESENVOLVIMENTO

Apesar da base curricular da Plataforma Lattes contar com mais de cinco milhões currículos cadastrados, neste trabalho, foi considerado apenas os currículos cujos indivíduos possuem o nível de capacitação doutorado concluído. Entretanto, vale ressaltar que o número de doutores representa cerca de 5% da quantidade total de indivíduos com currículos cadastrados na Plataforma Lattes, contudo, esses são responsáveis por cerca de 70% de todos os artigos cadastrados nos currículos; que, aliado a diversidade e contemporaneidade de atualização dessas informações, pondera a validade dos dados para auxiliar na compreensão sobre a produção científica brasileira (DIAS, 2016).

Os currículos dos doutores na versão XML (*eXtensible Markup Language*) da Plataforma Lattes, foram coletados com a utilização do *LattesDataXplorer* (DIAS, 2016). Em seguida, foram selecionados os currículos dos doutores, e posteriormente, foram processados para extrair as informações de todos os artigos de anais de congressos e periódicos, armazenando-as à parte num arquivo de publicações científicas, definindo o conjunto de dados centrais a serem estudados. As informações incluem: identifica-



dor do currículo; grande área da publicação; ano de publicação; tipo de publicação; título e palavras-chave.

Para a caracterização das redes, fez-se necessário uma adaptação no método *ISCooll*, proposto por Dias e Moita (2015) para identificar colaborações científicas em grande volume de dados. Enquanto o *ISCooll* original utiliza um dicionário para vincular os artigos a seus autores, o *ISCooll* adaptado adota um dicionário para vincular os artigos ao conjunto união das palavras-chave.

As redes de palavras-chave são baseadas no processo de adição de cliques por justaposição, onde cada artigo possui um conjunto de palavras-chave que formam uma clique. Ou seja, para cada artigo analisado uma clique é gerada, na qual palavras-chave se caracterizam como vértices e os vínculos entre eles são estabelecidos quando duas ou mais palavras-chave são utilizadas no mesmo artigo. Esse processo se repete para todos os artigos analisados, e, ao final, uma rede é gerada.

3 RESULTADOS

Os dados foram coletados em abril de 2017, totalizando 265.170 currículos de indivíduos com doutorado concluído. Para as análises, foram considerados os artigos publicados em anais de congressos e em periódicos entre 1962 e 2016, equivalendo a 10.040.664 artigos e 24.256.312 palavras-chave. A rede foi construída com base em todo conjunto de palavras-chave que os doutores associaram aos seus artigos cadastrados nos currículos da Plataforma Lattes, consistindo em 2.088.220 vértices e 20.093.922 arestas, onde sua componente gigante é responsável 99% de abrangência dessa rede.

Em continuação, para conhecer os tópicos que se constituem como os principais assuntos de pesquisa contidos na rede, todos os vértices (palavras-chave processadas) foram ranqueados de acordo com seu número de grau, e, posteriormente, todas as palavras-chave processadas tiveram suas frequências calculadas. Para facilitar a comparação entre as palavras-chave classificadas pelas duas medidas de importâncias utilizadas, a Tabela 1 apresenta as principais palavras-chave ranqueadas pela medida de frequ-



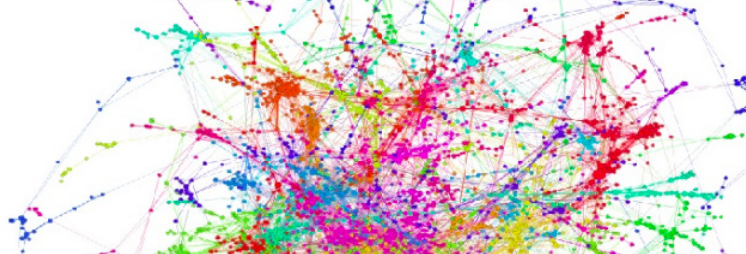
ência e as respectivas ordens de ranqueamento considerando o grau do vértice referente ao conjunto histórico dos dados.

TABELA 1 - COMPARATIVO ENTRE O RANQUEAMENTO DAS PALAVRAS-CHAVE ENTRE 1962 E 2016

1962-2016	Frequência	Grau	R. Frequência	R. Grau
Educação	75.855	30.951	1	1
Formação do Professor	46.032	15.923	2	13
Enfermagem	40.228	13.710	3	25
Epidemiologia	40.158	25.398	4	3
Crianças	29.048	18.047	5	9
Políticas Públicas	28.100	15.929	6	12
Amazônia	27.626	23.800	7	4
Bovinos	27.236	16.883	8	10
Idoso	26.291	11.589	9	40
Ensino	25.760	14.268	10	18
Ratos	25.480	20.270	11	7
Brasil	24.981	28.527	12	2
Diagnóstico	24.520	22.172	13	5
Educação Ambiental	23.194	10.103	14	61
Cultura	22.123	15.026	15	15

Fonte: Elaborado pelos autores, 2018.

Como pode ser observado, apesar de existir um alto percentual de palavras-chave iguais entre as principais ranqueadas pela frequência e grau, praticamente a ordem de destaque dessas não é a mesma, exceto para “Educação” e “Cultura”. Diante disto, para melhor compreender a relação entre os resultados encontrados, uma análise de correlação foi realizada. Os resultados apresentam um coeficiente de correlação positivo de “0,9572” para frequência e grau. Pelo valor obtido, destaca-se a comprovação de uma relação direta entre a frequência das palavras-chave e medidas baseada em grau nas redes de palavras-chave, indicando



que a medida de grau é uma alternativa interessante para identificação dos principais tópicos investigados.

4 CONSIDERAÇÕES FINAIS

Este estudo contribuiu para o mapeamento do conhecimento científico brasileiro a partir de redes de palavras-chave, e, concomitantemente, para a identificação dos principais tópicos de estudados por parte dos doutores brasileiros. Ao comparar as palavras-chave ranqueadas pelas medidas de frequência e grau, foi possível constatar uma alta correlação considerando todo o conjunto de análise, indicando que o grau pode ser uma alternativa interessante para calcular a popularidade de um tópico de pesquisa. Por fim, como trabalhos futuros, espera-se que sejam realizadas análises temporais nas redes para destacar a evolução dos tópicos estudados, e ainda, que sejam consideradas características qualitativas apropriadas ao contexto da Plataforma Lattes, como o fator de impacto dos periódicos.

REFERÊNCIAS

- CHAGAS, F. M. et. al. Algoritmo de classificação de especialistas em áreas na base de currículos Lattes. **Em Questão**, v. 21, n. 2, p. 119-139, 2015. ISSN 1808-5245.
- DIAS, T. M. R.; MOITA, G. F. A method for the identification of collaboration in large scientific databases. **Em Questão**, v. 21, n. 2, p. 140-161, 2015. ISSN 1808-5245.
- DIAS, T. M. R. **Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes**. 2016. 200 f. Tese (Doutorado) - Programa de Pós-Graduação em Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016.
- DIGIAMPIETRI, L. A. **Análise da rede social acadêmica brasileira**. 2015. 160 f. Tese (Livre Docência) - Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2015.
- LANE, J. Let's make science metrics more scientific. **Nature**, v. 464, n. 7288, p. 488-489, 2010.



MCCLOSKEY, D. N. **The rhetoric of economics**. Madison, Wisconsin: University of Wisconsin Press, 1988.

MENA-CHALCO, J. P. et. al. Caracterizando as redes de coautoria de currículos Lattes. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2012, Curitiba. **Anais...** Curitiba: BraSNAM, 2012.

PEREIRA, J.C.R. et. al. Who's who and what's what in Brazilian Public Health Sciences. **Scientometrics**, v. 73, n. 1, p. 37- 52, 2007.