Análises de Palavras-chave como Fonte de Dados para Obtenção de Conhecimento Sobre a Evolução da Ciência

Jether Gomes de Oliveira[1], Thiago Magela Rodrigues Dias[2], Gray Farias Moita[3], Adilson Luiz Pinto[4]

[1] 0000-0001-3448-6823, Doutor, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, MG, Brasil. jethergoliveira@gmail.com
[2] 0000-0001-5057-9936, Doutor, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, MG, Brasil. thiagomagela@cefetmg.br
[3] 0000-0002-6510-1019, Doutor, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, MG, Brasil. gray@dppg.cefetmg.br
[4] 0000-0002-4142-2061, Doutor, Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil. adilson@cin.ufsc.br

Tipo de trabalho: comunicação

Palavras-chave: Plataforma Lattes, Bibliometria, Palavras-chave, Tópicos de Pesquisa.

1 Introdução

O grande número de informações disponibilizadas pela internet e a sociabilização da atividade científica por parte de redes de pesquisadores são os fatores essenciais para o atual desenvolvimento da ciência (BRITO et. al., 2016). Para Dias (2016), serviços como bibliotecas digitais e sítios para registro individual de produção científica são alguns exemplos de como a internet tem contribuído na quantidade de trabalhos publicados, permitindo que usuários não apenas acessem conteúdo disponível, mas também possam registrar a sua produção científica.

Neste contexto, pesquisadores de todos os domínios têm dedicado esforços com intuito de analisar a produção científica sobre diferentes perspectivas, como: análises de citações, indicadores de produtividade, colaboração científica e análise de tópicos baseados em termos extraídos de resumos, títulos e palavras-chave de artigos científicos. Nesse último caso, um tópico pode ser entendido como termo que representa um dos assuntos associados a um determinado documento (BORGES et. al., 2015).

Os trabalhos que analisam tópicos também funcionam como uma boa revisão da literatura, o que permite, por exemplo, a verificação por parte do setor industrial se o que está sendo desenvolvido pela ciência contempla as necessidades da indústria (KHAN e WOOD, 2015). Para tanto, tais trabalhos geralmente exploram repositórios de artigos científicos, analisando seus títulos, resumos ou até todo o texto para extrair tópicos de pesquisas e analisá-los através de análises bibliométricas ou técnicas de análises de redes sociais.

Diversos outros trabalhos têm utilizado análises bibliométricas e técnicas de redes sociais, para extração de conhecimento acerca do que tem sido desenvolvido nas mais variadas áreas de pesquisas com propósitos distintos, a saber: taxas de carbono (ZHANG et al., 2016), linhas de produtos de software (HERADIO et. al. 2016), criatividade (ZHANG et. al. 2015); educação matemática (FADIGAS et. al., 2009); biomedicina (MADLOCK-BROWN,2014); epidemiologia na Alemanha (PETER et. al., 2016); e; saúde no Brasil (PEREIRA et. al., 2007).

Dentre os trabalhos encontrados na literatura, geralmente as análises realizadas utilizam repositórios internacionais, que são específicas de uma determinada área ou periódico. No entanto, por tratarem de análises específicas e utilizarem de repositórios internacionais, não podem representar de forma abrangente o que é produzido no Brasil. Com isso, analisar fonte de dados que englobe diversos tipos de publicações, principalmente em veículos nacionais de diversas áreas, passa a ser uma tarefa relevante para a compreensão da ciência brasileira.

De acordo com Brito et. al. (2016), estudos desta natureza são considerados urgentes no Brasil e podem retratar o que é desenvolvido e publicado em ciência, tecnologia e inovação, possibilitando gerar parâmetros que podem nortear esforços e investimentos para impulsionar resultados de pesquisa. Em Trucolo (2016), o autor destaca que muitas vezes os investimentos focam em áreas de pesquisas já consolidadas e populares, nas quais se acredita que haverá retorno, ou ainda, identificadas como tendências globais. Entretanto, num país com dimensões continentais como o Brasil, uma estratégia interessante seria investir nas áreas e tópicos de pesquisas com os maiores potenciais de crescimento, ampliando as chances do retorno da investigação científica e canalizando os recursos, que na maioria das vezes são reduzidos.

Para análises sobre o patamar científico brasileiro, o repositório de dados da Plataforma Lattes é tido como um diferencial (LANE, 2010). Esse repositório é composto por dados de grupos de pesquisas, instituições e currículos de mais de cinco milhões de indivíduos (DIAS, 2016). Esses currículos concentram dados sobre formação acadêmica, áreas de atuação, trabalhos em anais de congressos e em periódicos, entre outros. A Plataforma Lattes também é fonte de informação para órgãos que avaliam o Sistema Nacional de Pós-Graduação do Brasil, e agências de fomento que financiam pesquisas e ofertam bolsas de estudos. Contudo, mesmo estando disponível livremente na internet, esses dados ainda não foram amplamente analisados (DIGIAMPIETRI, 2015). Dos trabalhos que exploram estes dados, poucos são os que analisam as palavras-chave das publicações científicas. Geralmente utilizam termos extraídos dos títulos dos artigos de um conjunto restrito de currículos, na tentativa de destacar os principais assuntos abordados.

Assim sendo, é proposto neste trabalho uma análise temporal das principais palavras-chave existentes nos artigos científicos de todos os indivíduos com doutorado concluído que possuem os currículos cadastrados na Plataforma Lattes. Para tanto, inicialmente, as palavras-chave dos artigos publicados em anais de congressos e periódicos são extraídas e processadas, para posteriormente serem analisadas através de análises bibliométricas e técnicas de análises de redes sociais, para destacar os principais tópicos de interesses dos pesquisadores brasileiros de todas as grandes áreas do conhecimento ao longo de 55 anos de pesquisas registrados na base curricular da Plataforma Lattes.

2 Metodologia

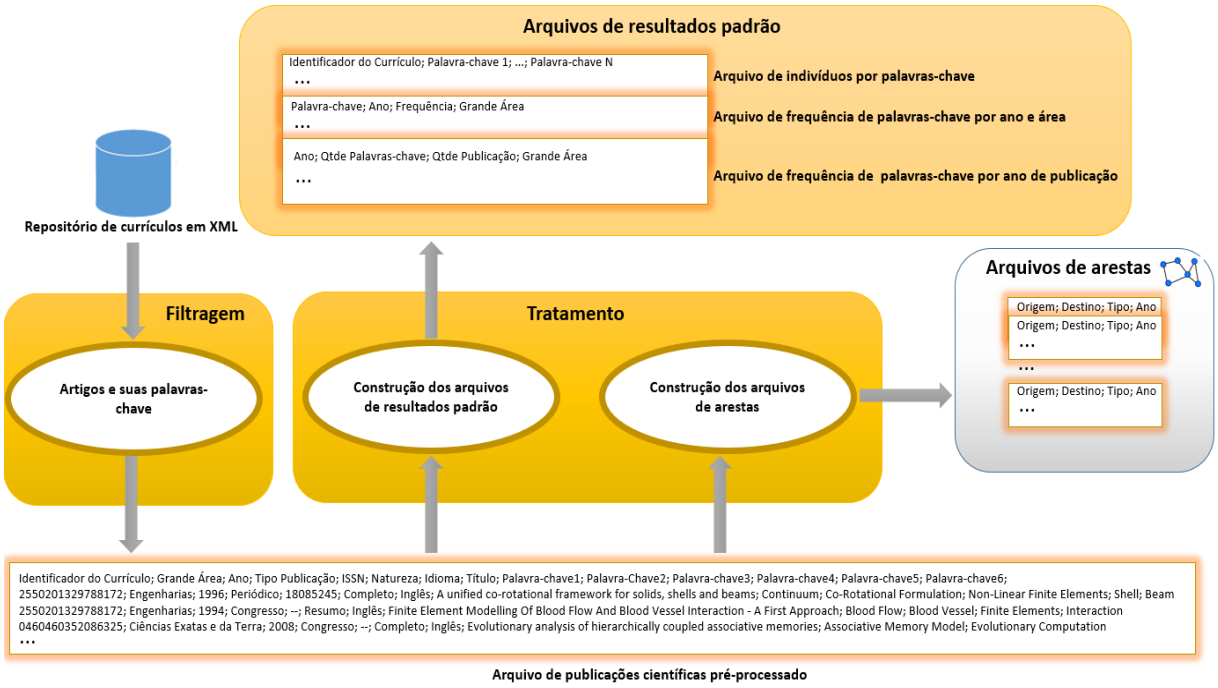
A motivação da escolha da Plataforma Lattes como fonte de informação está relacionada basicamente a quatro fatores: (1) aos dados estarem disponíveis na internet e não terem sido amplamente analisados (DIGIAMPIETRI, 2015); (2) tratar da integração de dados de produções científicas de todas as áreas de C&T existentes na ciência brasileira ao longo de toda trajetória; (3) por não negligenciar os artigos publicados em periódicos nacionais que muitas vezes não são indexados, e também os artigos de anais de congresso (DIAS, 2016); e, (4) por ser uma poderosa fonte para fornecimento de dados de alta qualidade para medir e avaliar o desempenho acadêmico nacional (LANE, 2010).

Neste ponto, vale ressaltar que, apesar dos dados disponibilizados, estes são apenas visualizados através de uma interface de consulta que apresenta apenas os dados de um único currículo. Logo, para uma análise mais detalhada de grupos, instituições ou até toda a nação de cientistas brasileiros, técnicas e ferramentas para análises dos dados se fazem necessárias.

A aquisição dos currículos dos doutores na versão XML (*eXtensible Markup Language*) foi realizada através da utilização do LattesDataXplorer, desenvolvido por Dias (2016) para coletar os dados científicos contido nos currículos cadastrados na Plataforma Lattes.

Logo, a Figura 1 apresenta uma visão geral do arcabouço de componentes desenvolvidos que suporta as análises desejadas. Nela, os componentes "filtragem dos dados" e "tratamento dos dados" são responsáveis por todo o processo de seleção, tratamento e modelagem das informações dos currículos que realmente necessitam ser processadas para atingir os objetivos propostos, e, concomitantemente, diminuir o tempo de processamento computacional.

Figura 1 – Processo de filtragem e tratamento dos dados.



O componente de "filtragem" realiza a etapa de mineração nos currículos para extrair as informações dos artigos, armazenando-as à parte num arquivo de publicações científicas, com isso,

definindo o conjunto de dados centrais a serem estudados. As informações dos artigos incluem: identificador do currículo; grande área da publicação; ano de publicação; tipo de publicação; issn do periódico; idioma da publicação; título e palavras-chave.

Por outro lado, o componente de "tratamento", processa os dados do arquivo de publicações científicas para tratá-los e caracterizá-los, e, a partir destes, constrói um conjunto dos arquivos de resultados padrão e arquivos de arestas para facilitar as análises. Esse componente realiza basicamente três etapas, a saber: limpeza dos dados, identificação de colaboração científica, e construção dos arquivos.

A limpeza dos dados serve para realizar o processamento das palavras-chave de tal forma a excluir os termos que não representam tópicos de estudos e agrupar as palavras que possuem mesmo valor semântico. Para isso, inicialmente, o método desenvolvido obtém as palavras-chave de cada artigo analisado. Em seguida, cada uma das palavras-chave são associadas ao idioma cadastrado para o artigo, servindo de referência no processo de radicalização. O processo de lowercase converte todas as palavras para minúsculo no intuito de padronizar o conjunto. No processo de stopWords são removidos os termos que não possuem valores semânticos. Posteriormente, no processo de normalização todos os acentos e pontuações são retirados das palavras-chave. Finalmente, o processo de radicalização é responsável pela redução da palavra-chave a seu radical. Contudo, em caso de palavras-chave compostas, este processo é executado em cada termo individualmente, e, concatenado formando uma única palavra.

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 únicos (colaboração) publicados em anais de congressos e em periódicos referentes ao período de 1962 até 2016, totalizando 10.040.664 artigos e 24.256.312 palavras-chave. Adicionalmente, vale destacar que a grande maioria dos currículos foram atualizados recentemente, onde 49,6% (131.660) possuem data de última atualização em 2017 e 73,3% (194.626) atualizados nos últimos dois anos. Neste ponto, vale ressaltar que o número de doutores representa cerca de 5% da quantidade total de indivíduos da Plataforma Lattes, porém, estes são responsáveis cerca de 70% de todos os artigos cadastrados nos currículos; que, aliado a diversidade e contemporaneidade de atualização destas informações, pondera a validade dos dados para auxiliar na compreensão sobre a evolução da produção científica brasileira (DIAS, 2016).

Inicialmente, 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 quinze principais palavras-chave ranqueadas pela medida de frequência e as respectivas ordem 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

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 destas não são a mesma, exceto para "Educação" e "Cultura". Diante disto, para melhor compreender a relação entre os resultados encontrados, uma análise de correlação foi realizada. Para tal análise, foi calculado o coeficiente de Spearman considerando todo o conjunto de palavras-chave entre frequência e grau. Os resultados apresentam um coeficiente de correlação positivo de "0,9572" para frequência e grau.

Para conhecer os tópicos que se constituem como os principais assuntos de pesquisa em cada época, todas as 2.088.220 palavras-chave únicas foram ranqueadas de acordo com a medida de importância de popularidade baseada na frequência por quinquênios entre 1962 e 2016, para facilitar o entendimento sobre a evolução temporal dos principais temas abordados. Incialmente, na tentativa de mapear a evolução dos principais interesses científicos por parte dos doutores brasileiros ao longo do tempo, a Tabela 2 apresenta o coeficiente de similaridade das 15 palavras-chave mais populares entre cada par de quinquênios do período analisado.

Tabela 2: Coeficiente de similaridade das palavras-chave mais frequentes entre os quinquênios.

	19621966	19671971	19721976	19771981	19821986	19871991	19921996	19972001	20022006	20072011	20122016
1962-1966	X	46,66%	26,66%	20,00%	20,00%	20,00%	20,00%	13,33%	13,33%	**0%**	6,66%
1967-1971		X	**33,33%**	20,00%	26,66%	20,00%	20,00%	13,33%	13,33%	**0%**	6,66%
1972-1976			X	73,33%	66,66%	53,33%	53,33%	40,00%	20,00%	**0%**	6,66%
1977-1981				X	73,33%	66,66%	66,66%	53,33%	33,33%	13,33%	20,00%
1982-1986					X	86,66%	80,00%	66,66%	46,66%	20,00%	26,66%
1987-1991						X	**93,33%**	73,33%	53,33%	26,66%	33,33%
1992-1996							X	80,00%	60,00%	33,33%	40,00%
1997-2001								X	80,00%	46,66%	53,33%
2002-2006									X	66,66%	73,33%
2007-2011										X	86,66%
2012-2016											X

Como pode ser notado, não houve ao longo do período analisado 100% de similaridade dos tópicos centrais estudados entre pares de quinquênios. Isso mostra que a cada 5 anos os principais interesses científicos tem mudado por algum motivo. No entanto, ao comparar os pares de quinquênios sequentes, nota-se que em 100% das comparações realizadas revelam que parte dos principais tópicos de pesquisas estudados no quinquênio atual foram considerados relevantes no

quinquênio anterior, destacando os períodos de 1987-1991 e 1992-1996, onde o percentual de similaridade foi de 93,33%, divergindo apenas quanto as palavras-chave “Peixes” e “Enfermagem”. Nesse caso, a palavra-chave “Peixes” que foi frequentemente utilizada nos períodos de 1982-1986 e 1987-1991, deu lugar a palavra-chave “Enfermagem” no quinquênio de 1992-1996. Uma hipótese para tal acontecimento, é o processo de regulamentação do exercício profissional que reconheceu as categorias de enfermeiro durante a década de 1980.

4 Considerações Finais

Este estudo contribuiu para a identificação dos tópicos de pesquisas que obtiveram maior interesses em cada época por parte dos doutores brasileiros que possuem currículos cadastrados na Plataforma Lattes. Ao realizar uma análise temporal entre as principais palavras-chave de cada período, foi possível verificar que as preferências centrais dos doutores sobre os tópicos de pesquisas foram alterando gradualmente ao longo do tempo, possivelmente devido a uma demanda externa da sociedade ou por determinado assunto ter atingido maturidade. Adicionalmente, ao comparar as principais palavras-chave pela medida de frequência com a medida de grau, foi possível constatar alta correlação considerando todo o conjunto de análise. Diante disso, constata-se que a utilização do grau como medida de importância se mostra uma alternativa interessante à frequência.

Agradecimentos

Os autores agradecem ao CNPq e CEFET-MG pelo auxílio na pesquisa.

Referências

BORGES, V. A. et al. Uma análise exploratória de tópicos de pesquisa emergentes em Informática na Educação. Revista Brasileira de Informática na Educação, Porto Alegre, v. 23, n. 1, p. 1-13, mar. 2015.

BRITO, A. G. C. et al. Exploração da Plataforma Lattes por assunto: proposta de metodologia. Transinformação, Campinas, v. 28, n.1, p. 77-86, jan./abr. 2016.

DIAS, T. M. R. Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes. 2016. Tese (Doutorado em Modelagem Matemática e Computacional) - Programa de Pós-Graduação em Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais, Minas Gerais, 2016.

DIAS, T. M. R.; MOITA, G. F. A method for the identification of collaboration in large scientific databases. Em Questão, Porto Alegre, v. 21, n. 2, p. 140-161, maio/ago. 2015.

DIGIAMPIETRI, L. A. Análise da rede social acadêmica brasileira. 2015. Tese (Livre Docência) Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2015.

FADIGAS, I. et al. Análise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática. Educação Matemática Pesquisa, v.11, n.1, p. 67-193, 2009.

HERADIO, R. et al. A bibliometric analysis of 20 years of research on software product lines. Information and Software Technology, Amsterdam, v. 72, p. 1- 15, Apr. 2016.

LANE, J. Let's make science metrics more scientific. Nature, London, v. 464, n. 7288, p. 488-489, Mar. 2010.

MADLOCK-BROWN, C. R. A framework for emerging topic detection in biomedicine. 2014. Tese (Doutorado em Filosofia em Informática na Saúde) - Graduate College, University of Iowa, Iowa City, 2014

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

PETER, R. S. et al. Epidemiologic research topics in Germany: a keyword network analysis of 2014 DGEpi conference presentations. European Journal of Epidemiology, London, v. 31, n. 6, p. 635- 638, June 2016.

TRUCOLO, C. C. Análise de tendências em redes sociais acadêmicas. 2016. 65 f. Dissertação (Mestrado em Ciências) - Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2016.

ZHANG, K. et al. A bibliometric analysis of research on carbon tax from 1989 to 2014. Renewable and Sustainable Energy Reviews, Amsterdam, v. 58, p. 297-310, May 2016.