

ESTUDO DOS PRINCIPAIS TÓPICOS DE PESQUISA NO BRASIL COM ADOÇÃO DE TÉCNICAS DE ANÁLISES DE REDES SOCIAIS

Jether Oliveira Gomes

Doutor em Modelagem Matemática Computacional
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET)
E-mail: jetherog@gmail.com

Thiago Magela Rodrigues Dias

Doutor em Modelagem Matemática Computacional
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET)
E-mail: thiagomagela@cefetmg.br

Gray Farias Moita

Doutor em Aeronáutica
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET)
E-mail: gray@dppg.cefetmg.br

Patrícia Mascarenhas Dias

Doutoranda em Modelagem Matemática Computacional
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET)
E-mail: patriciamdias@gmail.com

INTRODUÇÃO

Pesquisadores de todos os domínios têm concentrado esforços com o intuito de analisar a produção científica sob 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 – ou descritor – que representa um dos assuntos associados a um determinado documento. (BORGES; NOGUEIRA, BARBOSA, 2015) Assim, esforços para analisar tópicos de pesquisas constituem formas de melhorar a compreensão do comportamento de indivíduos ou grupos de pesquisadores, bem como de evidenciar o que tem sido produzido acerca da ciência. Adicionalmente, várias outras informações podem ser descobertas pela comunidade científica, a saber: (1) identificação dos principais tópicos em discussão; (2) identificação de novos tópicos; (3) identificação de tópicos isolados; (4) co-ocorrência dos tópicos de pesquisas; (5) descoberta de tópicos

sendo abordados por áreas diferentes; (6) indicação de possíveis tendências de pesquisas e (7) identificação de especialistas em determinadas áreas ou assuntos. A partir dessas descobertas, ainda é possível realizar análises baseadas em agrupamento, ranqueamento e recomendação.

Dentre os diversos trabalhos encontrados na literatura, principalmente devido ao crescimento exponencial da produção científica registrada nas últimas décadas (VINKERS; TIJDINK; OTTE 2015), nota-se um interesse global de todas as áreas de pesquisas em conhecer o que se tem feito acerca da ciência. Contudo, no contexto nacional, as análises geralmente utilizam repositórios de dados internacionais, que são específicos de uma determinada área ou periódico, e, com isso, realizam estudos sobre um determinado assunto ou área de pesquisa. No entanto, justamente por se tratarem de análises específicas e utilizarem repositórios de dados internacionais, geralmente, tais trabalhos não conseguem representar em totalidade o que é produzido no Brasil.

Assim sendo, este trabalho apresenta uma análise baseada em métricas de redes sociais das principais palavras-chave existentes nos artigos científicos de todos os indivíduos com doutorado concluído e, que possuem currículos cadastrados na Plataforma Lattes. Para tanto, inicialmente, as palavras-chave de cada publicação são extraídas e processadas, para posteriormente serem analisadas por meio de análises de redes sociais com o objetivo de destacar os principais tópicos de interesse dos pesquisadores brasileiros de todas as grandes áreas do conhecimento, bem como a relação entre eles, ao longo dos últimos 55 anos de pesquisas registradas nos currículos da Plataforma Lattes. Assim, é possível realizar um mergulho na história da ciência brasileira a fim de descobrir os tópicos mais importantes em determinadas épocas e utilizá-los como apoio para diversos tipos de tomadas de decisão.

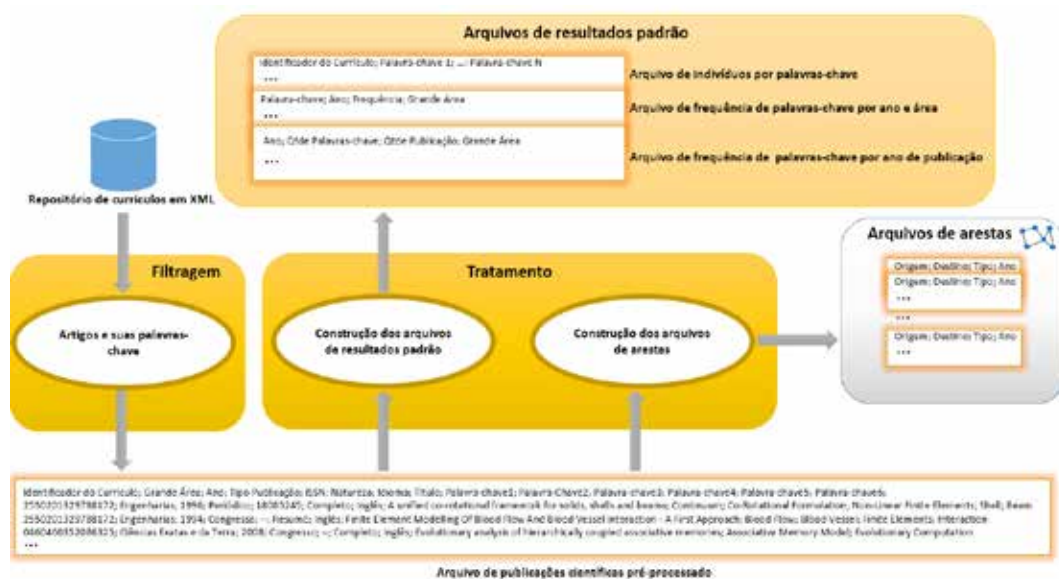
MATERIAIS E MÉTODOS

A motivação para a escolha da Plataforma Lattes como fonte de informação para estudo está relacionada a quatro fatores: (1) as informações estarem disponíveis livremente na internet e não terem sido amplamente analisadas; (2) tratar da integração de um enorme volume de dados de todas as áreas de Ciência e Tecnologia (C&T) existentes na ciência brasileira, com registros de dados acadêmicos e as produções científicas dos pesquisadores e instituições ao longo de toda trajetória; (3) por não negligenciar dados de artigos publicados em periódicos nacionais que muitas vezes não são indexados, e também os artigos publicados em anais de congresso que não são facilmente recuperados tendo em vista que cada evento possui seu repositório, diferentemente de repositórios internacionais (DIAS, 2016); e, (4) por ser reconhecida como uma poderosa fonte para fornecimento de dados de alta qualidade para medir e avaliar o desempenho acadêmico nacional, como mostrado por Lane (2010).

A aquisição desses currículos é realizada pelo LattesDataXplorer, proposto por Dias (2016) para coletar e tratar os dados contidos nos currículos cadastrados na Plataforma Lattes. Após a aquisição de todos os currículos cadastrados na Plataforma Lattes, foi utilizado o módulo de seleção do LattesDataXplorer para selecionar, dentre estes, os currículos que possuem o nível de capacitação doutorado concluído, definindo assim o conjunto principal de arquivos XML a serem processados para a realização dos estudos apresentados neste trabalho.

Logo, a Figura 1 apresenta uma visão geral do arcabouço de componentes que suporta as análises das palavras-chave dos artigos científicos cadastrados nos currículos da Plataforma Lattes. Nela, os componentes de “filtragem dos dados” e “tratamento dos dados”, desenvolvidos na atual pesquisa, são responsáveis por todo o processo de seleção, tratamento e caracterização das informações que 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



Fonte: arquivo de publicações científicas pré-processado.

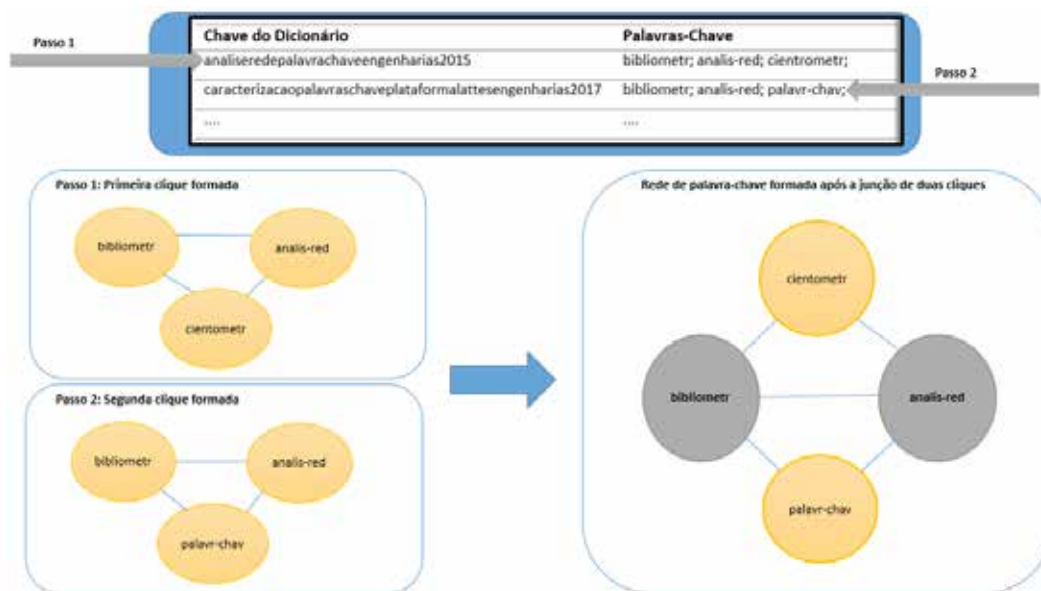
De posse dos currículos a serem analisados, o componente de “filtragem dos dados” realiza a etapa de filtragem em todos os arquivos XML com intuito de extrair as informações referentes aos artigos publicados em anais de congressos e em periódicos, armazenando-as num arquivo de publicações científicas, e com isso, definindo o conjunto de informações centrais a serem processadas durante todo o desenvolvimento da atual pesquisa. As informações dos artigos incluem: identificador do currículo; grande área da

publicação; ano de publicação; tipo de publicação; ISSN do periódico; natureza – no caso de anais de congresso, completo ou resumo –; idioma em que a publicação foi escrita; título do artigo e suas respectivas palavras-chave. Já quanto ao componente de “tratamento dos dados”, este processo tem o intuito de pré-processar os dados contidos no arquivo de publicações científicas para tratá-los e caracterizá-los, e a partir destes construir um conjunto de arquivos para facilitar as análises a serem realizadas. Esse componente é dividido em dois processos: (1) tratamento das palavras-chave para a construção dos arquivos de resultado padrão e (2) tratamento das palavras-chave para construção dos arquivos de arestas. Ambos os processos realizam basicamente três etapas, a saber: (1) limpeza e agrupamento dos dados, (2) normalização dos dados e (3) construção dos arquivos.

O conjunto de arquivos de arestas foi criado para permitir as análises de redes de palavras-chave. As redes de palavras-chave são baseadas no processo de adição de cliques por justaposição e sobreposição, em que cada artigo possui um conjunto de palavras-chave que formam uma clique. Ou seja, para cada artigo analisado um clique é gerada, na qual palavras-chave se caracterizam como vértices e os vínculos entre estes são estabelecidos quando duas ou mais palavras-chave são utilizadas no mesmo artigo.

Este processo se repete para todos os artigos analisados e, ao final, uma rede é gerada a partir da inserção de cliques. A Figura 2 exemplifica a construção da rede de palavras-chave a partir de dois artigos que foram processados pelas etapas descritas.

Figura 2 - Processo de construção de redes de palavras-chave considerando dois artigos



Fonte: arquivo de publicações científicas pré-processado.

As redes de palavras-chave são modeladas a partir dos arquivos de arestas, que foram construídos no seguinte formato: origem (vértice de origem), destino (vértice de destino) e tipo (neste caso, não direcionado). Apesar de não serem consideradas neste trabalho, vale ressaltar que existem outras possibilidades para configuração de redes de palavra-chave, como, por exemplo, a construção de redes de pesquisadores, onde os vértices são os identificadores dos currículos que representam um determinado indivíduo e as arestas são formadas por palavras-chave que pesquisadores tenham utilizado em comum. Diante disso, é possível modelar uma rede onde os vínculos entre pesquisadores são o vocabulário em comum utilizado para descrever seus trabalhos.

RESULTADOS

Os dados utilizados nesta caracterização foram coletados em abril de 2017, totalizando 265.170 currículos de indivíduos com doutorado concluído das diversas áreas do conhecimento científico. Para esta caracterização foram considerados os artigos de anais de congressos e periódicos referentes ao período de 1962 até 2016.

Inicialmente, foi construída uma rede com base em todo o conjunto de palavras-chave que os doutores associaram a seus artigos cadastrados nos currículos da Plataforma Lattes, totalizando 2.088.220 vértices e 20.093.922 arestas, em que sua componente gigante é responsável por 1.903.657 (91%) vértices e 20.046.249 (99%) arestas. Considerando que 98,6% das palavras-chave únicas possuem frequência de utilização inferior a 100, tomou-se a decisão de reconfigurar o conjunto sem estas palavras-chave para facilitar os cálculos e visualizações. Com isso, a nova rede é composta por 112.812 vértices e 8.779.131 arestas e sua componente gigante possui 112.799 vértices e 8.779.131 arestas.

Alguns dos trabalhos que analisam dados com métricas de redes têm analisado medidas de centralidade para destacar os vértices (no caso, palavras-chave) mais importantes para a rede analisada. Neste contexto, o grau de um vértice se habilita como uma alternativa interessante à medida de importância de tópicos, mais especificamente, a medida baseada em frequência. Logo, os vértices de maior grau foram ranqueados a partir das redes de quinquênios entre 1962 e 2016 e da rede referente a todo histórico de análise (Tabela 1).

Tabela 1 - As palavras-chave que apresentam maior grau entre 1962 e 2016

Ranque	Palavras-chave (1962-2016)	Grau
1	Educação	30.951
2	Brasil	28.527
3	Epidemiologia	25.398
4	Amazônia	23.800
5	Diagnóstico	22.172

Ranque	Palavras-chave (1962-2016)	Grau
6	Enfermagem	21.940
7	Ratos	20.270
8	Avaliação	18.751
9	Crianças	18.047
10	Bovinos	16.883
11	Morfologia	16.504
12	Políticas Públicas	15.929
13	Formação do Professor	15.923
14	Tratamento	15.550
15	Cultura	15.026

Fonte: elaborado pelos autores.

Como mostrado nessa tabela, a palavra-chave “Educação” possui o maior número de grau dentre todas do conjunto, ou seja, é a mais abrangente quanto a conexões com palavras-chave distintas. À vista disso, “Educação” se consolida como a palavra-chave de maior interesse dos doutores entre 1962 e 2016. Em relação às principais palavras-chave ranqueadas pelas medidas de frequência em períodos de cinco anos, percebe-se, em média, 66% de compatibilidade entre as medidas de grau e frequência. Outra observação é que as palavras-chave “Brasil”, “Morfologia” e “Tratamento”, ambas destacadas como importantes pelo grau e não encontradas entre as principais pela medida de frequência, representam temas de investigação genéricos, ou seja, normalmente aplicados em trabalhos de diferentes áreas.

Ressalta-se que os termos Brasil e Amazônia se caracterizam como localidades e possivelmente sofrem influência de estudos, relacionados a outros temas e que podem ter sido realizados nestas localidades.

Outra informação interessante que pode ser extraída das redes é a medida de co-ocorrência dos vértices – palavras-chave –, para, em seguida, ser possível verificar quais pares de palavras-chave são mais utilizadas conjuntamente. Isso pode, por exemplo, auxiliar no entendimento sobre o conteúdo dos artigos publicados. Sendo assim, a Tabela 2 apresenta os pares das palavras-chave que mais co-ocorrem, ou seja, as arestas mais densas da rede de palavras-chave, que foi construída a partir de todo o conjunto histórico dos dados.

Tabela 2 - As principais co-ocorrências de pares de palavras-chave entre 1962 e 2016

Ranque	Pares de Palavra-chave	Co-ocorrência
1	Germinação – Sementes	3.191
2	Educação – Formação do Professor	3.068
3	Aids – Hiv	2.975
4	Adolescentes – Crianças	2.792

Ranque	Pares de Palavra-chave	Co-ocorrência
5	Educação – Ensino	2.487
6	Germinação – Vigor	2.249
7	Geoprocessamento – Sensoriamento Remoto	2.077
8	Educação – Cultura	1.873
9	Aprendizagem – Ensino	1.869
10	Envelhecimento – Idoso	1.861
11	Memória – História	1.845
12	Educação – História	1.841
13	Currículo – Formação do Professor	1.807
14	Currículo – Educação	1.684
15	Enfermagem – Cuidados	1.675

Fonte: elaborado pelos autores.

Notadamente, o destaque da palavra-chave “Educação” quanto à sua frequência de utilização e o alto valor do grau, possivelmente, tem influência quanto sua presença em 33,3% dos principais pares apresentados. Apesar disto, destacam-se pares de palavras-chave que não apareceram entre as principais encontradas pelas medidas de importância utilizadas, como, por exemplo: (1) “Germinação – Sementes”, aresta mais densa da rede, (2) “Aids – Hiv”, (3) “Germinação – Vigor”, (4) “Geoprocessamento – Sensoriamento Remoto”, (5) “Aprendizagem – Ensino” e (6) “Memória – História”.

CONSIDERAÇÕES

Ao analisar as redes de palavras-chave, percebe-se claramente a evolução destas ao longo do tempo, principalmente quanto à quantidade de arestas e vértices, que possuem crescimento em todos os períodos, com exceção dos dois últimos quinquênios. Essa evolução segue o comportamento da produção científica brasileira. Também pôde ser notado o impacto da retirada/inserção de vértices e arestas – no caso, palavras-chave e suas relações – perante os valores das métricas aplicadas. Por exemplo, a adição de novas palavras-chave (vértices) na componente gigante impacta diretamente o número de componentes isolados, enquanto, a inserção de arestas torna a componente gigante mais densa, e, com isso, há impacto direto nos valores de grau médio, diâmetro e caminho mínimo da rede.

REFERÊNCIAS

BORGES, V. A.; NOGUEIRA, B. M.; BARBOSA, E. F. 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, 2015.

DIAS, T. M. R. *Um Estudo Sobre a Produção Científica Brasileira a Partir de Dados da Plataforma Lattes*. 2016. Tese (Doutorado em Modelagem Matemática e Computacional) – Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 206.

LANE, J. Let's make science metrics more scientific. *Nature*, New York, v. 464, p. 488-489, 2010.

VINKERS, C. H.; TIJDINK, J. K.; OTTE, W. M. J. Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. *BMJ*, London, p. 1-8, 2015.