

Uma análise dos principais tópicos de pesquisas investigados pelos pesquisadores doutores brasileiros

Jether Oliveira Gomes

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

Thiago Magela Rodrigues Dias

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

Gray Farias Moita

Doutor; Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, MG, Brasil;
gray@dppg.cefetmg.br

Resumo: A produção e publicação de trabalhos científicos apresentou um crescimento impressionante nas últimas décadas, sendo a internet o principal fator de acesso e difusão desses para a comunidade. Diante disto, nota-se um interesse global de todas áreas do conhecimento quanto a estudos sobre dados de produções científicas, a fim de conhecer o que se tem feito acerca da ciência. Sendo que, o entendimento sobre como as pesquisas têm evoluído, pode servir de base para construção de políticas científicas, visando novos avanços na ciência, ou impulsionar grupos de pesquisas a se tornarem mais produtivos. Neste contexto, o objetivo do estudo é analisar os tópicos de pesquisa publicados ao longo dos últimos 55 anos de trajetória da ciência brasileira de pesquisadores doutores, no intuito de mapear o conhecimento científico e identificar temas em destaques. Para isso, são realizadas a caracterização geral e uma análise dos dados referentes às palavras-chave dos artigos publicados em anais de congressos e em periódicos, pelos doutores que têm currículos cadastrados na Plataforma Lattes. Os resultados mostraram uma visão geral sobre as palavras-chave utilizadas pelos doutores, e, assim, identificação e análise dos principais tópicos de pesquisas desenvolvidos por eles. Como considerações finais, constatou-se a validade dos resultados do ponto de vista quantitativo, para auxiliar no entendimento sobre o desenvolvimento da ciência brasileira.

Palavras-chave: Palavras-chave. Tópicos de pesquisas. Análise bibliométrica. Plataforma Lattes. Ciência brasileira.

1 Introdução

O grande número de informações disponibilizadas pela internet e a sociabilização da atividade científica, por parte de redes de pesquisadores, são os principais fatores para o atual desenvolvimento da ciência (BRITO; QUONIAM; MENA-CHALCO, 2016). Para Dias (2016), serviços como bibliotecas digitais e sítios para registros de produção científica são alguns exemplos de como a internet tem contribuído no avanço da quantidade de trabalhos publicados, permitindo que usuários não apenas acessem conteúdo disponível, mas também possam registrar sua produção científica a partir de sua interação com esse meio. Desta forma, trabalhos publicados podem ser acessados instantaneamente, contribuindo para a expansão do conhecimento.

Segundo Carneiro (2003), o conhecimento é o principal elemento para a geração do desenvolvimento. Além de a divulgação científica contribuir para popularização do conhecimento, ela aproxima o cidadão dos benefícios que tem direito de requerer para a melhoria do seu bem-estar social, oferecendo-lhe uma visão mais clara sobre os problemas que enfrenta.

Em uma sociedade competitiva em escala mundial, implementar o conhecimento técnico e científico é tarefa primordial, a fim de alavancar o desenvolvimento econômico e social (CAVACINI, 2016), principalmente para os países em desenvolvimento que são consumidores desse conhecimento. Contudo, muitas vezes, o plano de implementação é condicionado à existência de recursos limitados e uma imposição, cada vez maior, de racionalidade e objetividade na aplicação dos poucos recursos disponíveis.

Tão importante quanto ter investimentos é ter habilidade para controlar, entender e medir o patamar científico das nações e/ou grupos individualizados, negócios e fundações que devem decidir suas prioridades científicas. Portanto, torna-se imprescindível estudar a produção científica para a implementação desse conhecimento, a superação das dificuldades e o alcance dos pressupostos de racionalidade e objetividade (SAES, 2005).

Neste contexto, pesquisadores de todos os domínios têm dedicado esforços para analisar a produção científica sobre diferentes perspectivas como, por exemplo, por meio de análise de tópicos de pesquisas. Nesse 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 aprimorar a compreensão do que tem sido produzido acerca da ciência. Adicionalmente, outras informações podem ser descobertas: (1) principais tópicos em discussão; (2) coocorrência dos tópicos de pesquisas e; (3) tendências de pesquisas. A partir destas descobertas, ainda é possível realizar análises baseadas em agrupamento, ranqueamento e recomendação.

Aliado a isso, análises considerando o índice temporal, associado ao uso dos tópicos, podem permitir a identificação de assuntos mais importantes em determinada época e, também, possibilitar o estudo da evolução dos tópicos desenvolvidos. Entender essa evolução pode ajudar na compreensão do histórico dos interesses em pesquisas de uma determinada instituição ou área de pesquisa.

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

Entretanto, títulos e resumos de trabalhos podem não representar os assuntos abordados, devido à necessidade de se preocupar com a semântica e a estrutura dos termos. Logo, uma abordagem interessante é a análise das palavras-chave de publicações científicas, visto que são inseridas, cuidadosamente, por seus respectivos autores para descrever os assuntos principais que permeiam o trabalho de forma clara e objetiva (MCCLOSKEY, 1998; YI; CHOI, 2012). Pei-Chun, Hsin-Ning e Te-Yi (2010) destacam que uma palavra-chave é o portador fundamental mais básico do conhecimento. Com

isso, no artigo, palavras-chave de publicações científicas são igualmente referenciadas como tópicos de pesquisas.

Dentre os trabalhos encontrados, nota-se um interesse global de todas as áreas de pesquisas em conhecer o que se tem feito acerca da ciência (VINKERS; TIJDINK; OTTE, 2015). Contudo, no contexto nacional, geralmente, as análises realizadas utilizam repositórios de dados internacionais que são específicos de uma determinada área ou periódico. No entanto, justamente por se tratarem de análises específicas e utilizarem repositórios internacionais, não podem representar em totalidade o que é produzido pelas diversas áreas do conhecimento.

Para Brito, Quoniam e Mena-Chalco (2016), estudos desta natureza são considerados urgentes no Brasil e podem retratar o que é desenvolvido e publicado em ciência, possibilitando gerar parâmetros para orientar esforços e investimentos em pesquisa. Trucolo (2016) destaca que, geralmente, os investimentos focam em áreas de pesquisas já consolidadas ou tendências globais, nas quais se acredita que haverá retorno. Entretanto, uma estratégia interessante seria investir em assuntos com maiores potenciais de crescimento, ampliando as chances do retorno da investigação científica e canalizando os recursos.

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 utilizada por órgãos que avaliam o Sistema Nacional de Pós-Graduação do Brasil e agências de fomento que financiam pesquisas e ofertam bolsas de estudos.

Os dados da Plataforma Lattes estão disponíveis livremente na internet. No entanto, estes ainda não foram amplamente analisados, apesar de ser utilizado para avaliar e verificar dados de pesquisadores e/ou grupos desses

(DIGIAMPIETRI, 2015). Dos trabalhos que exploram dados da Plataforma Lattes para análises bibliográficas, poucos são os que analisam conteúdo. Estes utilizam termos extraídos dos títulos dos artigos de um conjunto restrito de currículos como, por exemplo, dados de indivíduos de um determinado programa de pós-graduação ou área do conhecimento específica, na tentativa de destacar os assuntos abordados por esses grupos restritos.

Assim sendo, este trabalho realiza a caracterização geral e uma análise inédita sobre os dados referentes às palavras-chave dos artigos publicados em anais de congressos e em periódicos pelos doutores que têm currículos na Plataforma Lattes. Para tanto, inicialmente, as palavras-chave são extraídas, para, posteriormente, serem analisadas a partir de análises bibliométricas, no intuito de destacar os principais tópicos de interesses dos doutores brasileiros ao longo dos últimos 55 anos de pesquisas registrados em seus currículos.

2 Trabalhos correlatos

Esforços para identificar e analisar tópicos de pesquisa constituem uma forma de melhorar a compreensão do que se tem produzido acerca da ciência e, a partir disto, auxiliar nos mais variados cenários quanto à tomada de decisão. Segundo Khan e Wood (2015), estes estudos, geralmente, baseiam-se em popularidade (analisa a frequência de utilização de termos contidos nos títulos ou palavras-chave de artigos científicos) ou redes de citações (analisa as relações entre citações de artigos e autores). Adicionalmente, Choi, Yi e Lee (2011) destacam a importância em analisar as redes de palavras-chave para evidenciar os tópicos mais importantes baseado em medidas de centralidade.

O trabalho de Zhu e outros (2013) analisou uma rede de 111.444 palavras-chave de artigos da Ciência da Informação extraídos do repositório da Scopus em 2008. Os autores identificaram, pelo grau de centralidade, que algumas palavras-chave são mais importantes que outras. Diante de tal constatação, compararam a identificação por frequência com medidas de centralidade, justificando que a análise baseada em grau tende a ser mais eficiente.

Khan e Wood (2015) analisaram as redes de palavras-chave e de termos extraídos dos títulos de 893 artigos publicados entre 1995 e 2014, pela área de Gestão da Tecnologia da Informação, contidos na Web of Science. Os autores relatam que poucas palavras são frequentemente utilizadas; as redes de palavras-chave são mais apropriadas para representar os temas de pesquisas e; foi possível realizar um mapeamento dos assuntos discutidos pela área.

Pollack e Adler (2015) aplicaram análises bibliométricas para descobrir as tendências de pesquisas da área de Gerenciamento de Projetos a partir de 94.472 artigos das bibliotecas Scopus e Web of Science publicados entre 1962 e 2012. Eles realizaram comparações entre as palavras-chave e os termos de resumos dos artigos quanto à frequência, sua variação de utilização ao longo do tempo e a coocorrência para identificar os tópicos emergentes.

Em Borges, Nogueira e Barbosa (2015) realizaram uma análise dos tópicos de pesquisas na área de Informática na Educação, a fim de identificar os principais assuntos. Como apoio à análise, eles construíram hierarquias de tópicos com o algoritmo K-means a partir de 4.053 artigos publicados, no período de 2011 a 2014, nas principais conferências da área.

O trabalho de Mryglod e outros (2016) estudou a produção científica relacionada ao acidente de Chernobyl. Para isto, utilizaram análises bibliométricas e técnicas de redes sociais em 9.500 publicações extraídas da Scopus e Ukrainika naukova. Os autores mediram a distribuição dos trabalhos publicados, as taxas de crescimento e, além disso, analisaram os títulos e resumos das publicações para detecção dos termos mais importantes utilizados.

Zhu e outros (2015) analisaram as palavras-chave de 363.458 teses publicadas, entre 1986 e 2014, referentes à indústria de petróleo e gás da China National Knowledge Internet, para entender a evolução da área e evidenciar os assuntos atuais. Inicialmente, realizaram a mineração dos dados e, em seguida, analisaram as frequências, coocorrências e as redes das palavras-chave.

Ronda-Pupo (2016) utilizou uma combinação de técnicas de coocorrência e de redes sociais nos termos extraídos dos títulos e resumos de 2.264 artigos da biblioteca Web of Science, para determinar os tópicos que

comporam a disciplina de Gestão na América Latina e no Caribe nos últimos 25 anos.

Diversos outros trabalhos têm utilizado análises bibliométricas, técnicas de redes sociais e análises de tendências, para entendimento sobre o desenvolvimento da ciência nas mais variadas áreas do conhecimento e diferentes propósitos como, taxas de carbono (ZHANG et al., 2016), linhas de produtos de software (HERADIO et al. 2016), mapeamento científico e tendências (HONG et al., 2016; SILVA et al., 2016); epidemiologia na Alemanha (PETER et al., 2016); biomedicina (MADLOCK-BROWN, 2014); saúde no Brasil (SAES, 2005; PEREIRA et al., 2007) e; pesquisas em universidades (CHEN et al., 2015).

Neste ponto, é importante ressaltar que os trabalhos citados até o momento utilizam dados de repositórios internacionais (determinada área ou periódico) e com uma quantidade limitada de registros. Por outro lado, a seguir, são apresentados os poucos e recentes trabalhos encontrados que analisam conteúdo das publicações e que utilizam fontes de dados nacionais (mais especificamente dados da Plataforma Lattes).

Assim, Medeiros e Mena-Chalco (2013) analisaram 633.508 currículos e 4.946.990 publicações para estudar a rede social dos indivíduos de áreas distintas. Adicionalmente, calcularam as frequências dos termos dos títulos para identificar quais são mais utilizados em cada época. Em continuação, utilizaram mapas de termos e nuvem de palavras para as 200 palavras mais frequentes de cada área por período de tempo.

O trabalho de Digiampietri, Peres e Silva(2014) apresentou uma análise sobre a rede social formada a partir 2.002 currículos de indivíduos que atuam no Brasil, em Inteligência Artificial ou Inteligência Computacional. Além disso, analisaram a frequência relativa dos termos dos artigos publicados, entre 1993 a 2012, para ilustrar a evolução dos temas abordados ao longo de cada ano.

Em Trucolo e Digiampietri (2014a), os autores aplicaram regressões nas medidas de importância TF-IDF dos termos extraídos de 34.289 títulos de artigos publicados pelos doutores da Ciência da Informação entre 1991 e 2012,

para identificar tendências de pesquisas. Assim, um termo era considerado tendência se apresentasse uma alta previsão de TF-IDF. Trucolo e Digiampietri (2014b) aplicaram as mesmas regressões nos termos de 57.501 títulos de artigos, publicados entre 1911 e 2011, pelos professores de Ciência da Computação. Em continuação, Trucolo (2016) considerou a fonte geradora da informação e, com isso, utilizou, também, resultados de métricas de redes sociais para análises de tendências nos mesmos dados estudados em Trucolo e Digiampietri (2014b).

Dentre os trabalhos apresentados, verificou-se que analisar termos dos títulos dos artigos é o foco principal de grande parte dos estudos (no caso da Plataforma Lattes, são todos). Embora, esta estratégia tenha validade, ela tem suas limitações, uma vez que nem sempre os títulos conseguem expressar todo o conteúdo abordado pelo trabalho. Com isso, a estratégia de analisar palavras-chave dos artigos vem ganhando força na literatura, visto que tem o foco principal de evidenciar os tópicos centrais que permeiam o trabalho de forma objetiva.

Consequentemente, o presente trabalho apresenta-se como a primeira análise textual abrangente, ao considerar as palavras-chave dos artigos publicados pelos doutores brasileiros ao longo dos últimos 55 anos registrados em seus currículos cadastrados na Plataforma Lattes.

3 Metodologia

As subseções (3.1) Obtenção dos dados e (3.2) Filtragem e tratamento dos dados apresentam o detalhamento das atividades que suportam o trabalho.

3.1 Obtenção dos dados

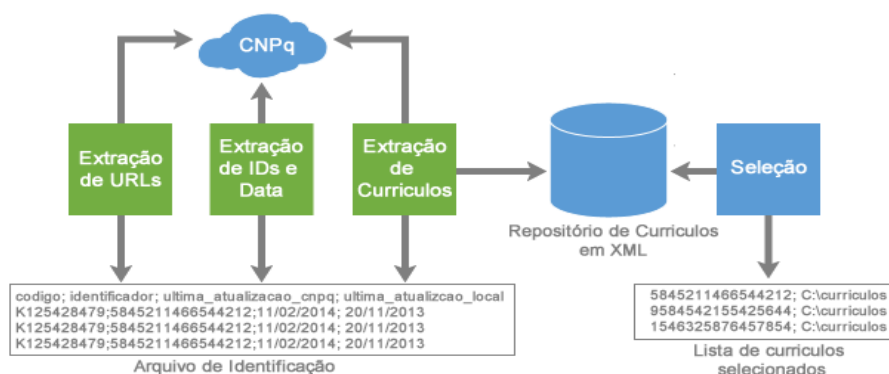
A escolha da Plataforma Lattes como fonte de dados está relacionada: (1) aos dados estarem disponíveis na internet e não terem sido amplamente analisados (DIGIAMPIETRI, 2015); (2) tratar da integração de dados de produções científicas de todas as áreas de C&T existentes na ciência brasileira; (3) por 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).

Apesar dos dados disponíveis, estes são apenas visualizados por uma interface de consulta que apresenta cada currículo individualmente. Assim, para uma análise mais detalhada de grupos de pesquisadores ou instituições, técnicas e ferramentas para análises dos dados se fazem necessárias (DIAS, 2016).

A obtenção dos dados ocorreu em abril de 2017, totalizando 265.170 currículos de doutores. Para tanto, foi utilizado o *LattesDataXplorer*, proposto por Dias (2016), para coletar os dados existentes no repositório da Plataforma Lattes, como ilustrado na Figura 1.

Figura 1 - Processo de extração de dados *LattesDataXplorer*



Fonte: Dias (2016).

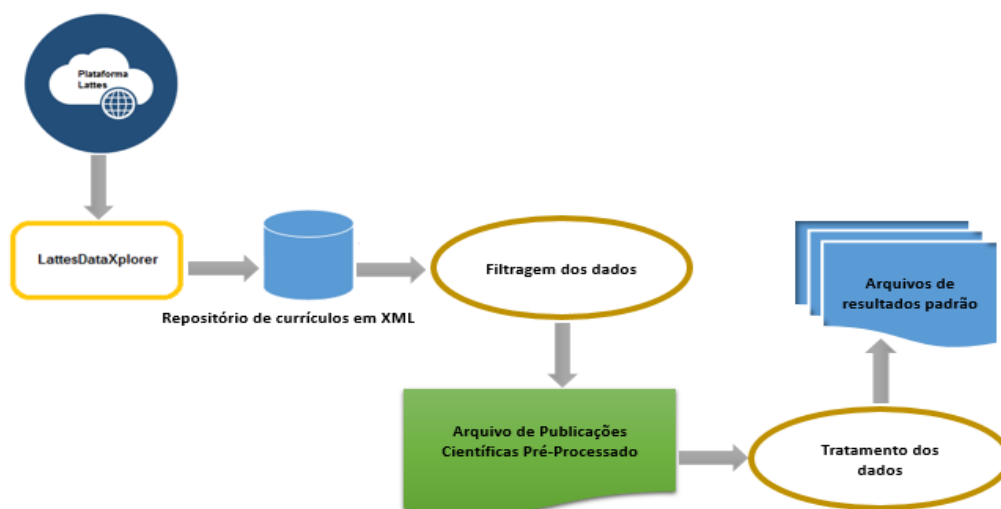
O processo de extração do *LattesDataXplorer* para obtenção dos currículos foi realizado em três etapas: (1) extração de URLs, responsável por adquirir as referências únicas para os currículos cadastrados e, assim, possibilitar o acesso individual a cada currículo; (2) extração de IDs, para possibilitar o acesso a cada currículo e extrair seu identificador e; (3) extração dos currículos, responsável pelo *download* e armazenamento em disco dos arquivos no formato XML (*eXtensible Markup Language*) (DIAS, 2016).

3.2 Filtragem e tratamento dos dados

Na tentativa de mapear o desenvolvimento científico dos doutores brasileiros foram utilizadas as informações de artigos científicos publicados em anais de congressos e em periódicos extraídos dos currículos dos indivíduos com doutorado concluído da Plataforma Lattes. A respectiva justificativa é que artigos são considerados o principal caminho de disseminação de novos conhecimentos na maioria das disciplinas científicas (RONDA-PUPO, 2016).

A Figura 2 apresenta uma visão geral do conjunto de componentes desenvolvidos para suportar as análises desejadas. Nela, “filtragem dos dados” e “tratamento dos dados” são os responsáveis pelo processo de seleção e tratamento dos dados dos currículos a serem analisados e, concomitantemente, diminuem o tempo de processamento computacional.

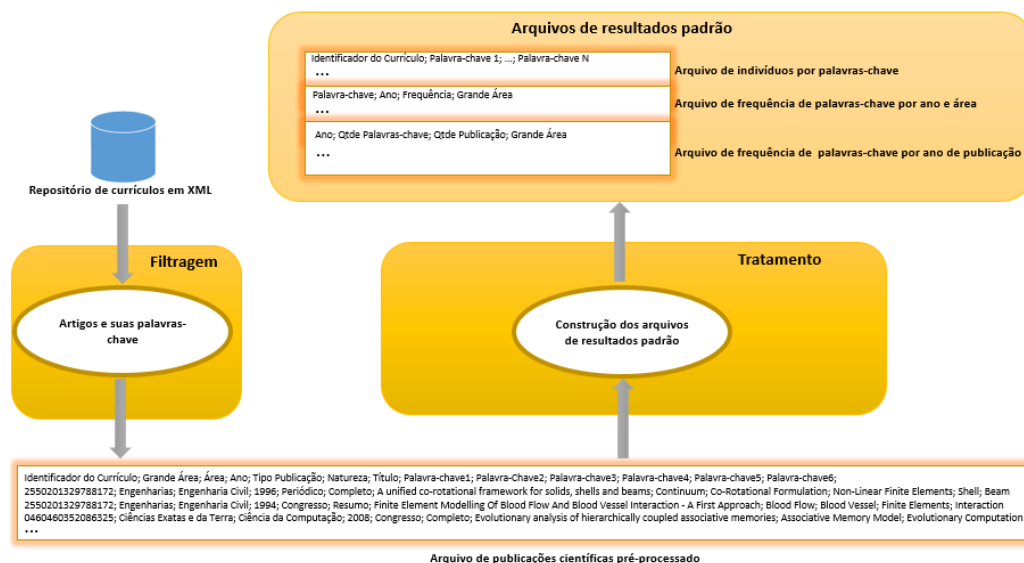
Figura 2 - Visão geral do conjunto de componentes



Fonte: Elaborado pelos autores.

Os componentes de “filtragem dos dados” e “tratamento dos dados” são ilustrados com maiores detalhes na Figura 3.

Figura 3 - Processo de filtragem e tratamento dos dados



Fonte: Elaborado pelos autores.

Após a aquisição dos currículos, o componente de “filtragem dos dados” realiza a etapa de mineração nos arquivos XML para extrair as informações dos artigos, armazenando-as em um arquivo de publicações científicas e, com isso, definindo o conjunto de dados centrais a serem processados. As informações dos artigos incluem: identificador do currículo; grande área da publicação; ano de publicação; tipo de publicação; título e palavras-chave.

O componente de “tratamento dos dados” tem o intuito de processar os dados do arquivo de publicações científicas para tratá-los e caracterizá-los e, partir destes, construir um conjunto de arquivos para facilitar as análises. Esse componente realiza o processo necessário para o tratamento das palavras-chave dos artigos para construção dos arquivos de resultado padrão. Esse processo realiza, basicamente, três etapas: (3.2.1) limpeza e agrupamento dos dados, (3.2.2) normalização dos dados e (3.2.3) construção dos arquivos.

3.2.1 Limpeza e agrupamento dos dados

O cadastramento das palavras-chave dos artigos nos currículos é de responsabilidade dos pesquisadores e, isso é feito livremente por eles. Assim, geralmente, tem-se uma coleção muito grande de palavras-chave. Na tentativa de contornar este problema, foi desenvolvido um método que processa as

palavras-chave para excluir possíveis termos irrelevantes para a pesquisa. Além disso, o método agrupa em um dicionário, as palavras-chave que foram escritas de formas distintas, mas que têm mesmo valor semântico.

O método inicia-se obtendo as palavras-chave extraídas de cada artigo. Diante disso, cada uma das palavras-chave é associada ao idioma do artigo que foi cadastrado pelo próprio pesquisador, para servir de referência no processo de radicalização (*stemming*). Em continuação, no processo de *lowercase*, todas as palavras são convertidas para minúsculo com a proposta de padronizar o conjunto.

No processo de *stopWords*, são removidos os termos que não apresentam valores semânticos significativos para caracterizar um tópico de pesquisa e, com isso, diminuir o volume de dados a serem processados e analisados. Em seguida, cada uma das palavras-chave passa por um processo de normalização para extrair as letras acentuadas e substituí-las pelo seu equivalente sem acentuação.

E, por último, o processo de *stemming* que consiste na redução das palavras-chave a seu radical. O processo de *stemming* é importante para evitar a inclusão de palavras-chave com o mesmo significado de formas distintas, possibilitando a redução do conjunto de forma significativa. No caso de palavras-chave compostas, este processo é executado em cada termo individualmente e, em seguida, são concatenados formando uma única palavra. A Tabela 1 apresenta um exemplo de transformação de uma palavra-chave de artigo científico, após a execução da etapa de limpeza e agrupamento.

Tabela 1 - Exemplo de transformação de uma palavra extraída de um artigo científico

Etapa Algoritmo	Resultado
Recebimento da Palavra-Chave	Gerência de Dados
Deteção de idioma	Gerência de Dados, Português
LowerCase	gerência de dados, Português
StopWords	gerência dados, Português
Normalização	gerencia dados, Português

Radicalização

gerenc dad, Português

gerenc-dad

Fonte: Elaborado pelos autores.

3.2.2 Normalização dos dados

No caso dos currículos cadastrados na Plataforma Lattes, não é uma situação incomum dois pesquisadores cadastrarem um artigo científico que publicaram juntos, utilizando diferentes informações como, por exemplo, as respectivas palavras-chave. Em contrapartida, em determinadas análises, o preenchimento igualitário das informações pode gerar duplicidade de dados quando considerado todo o repositório. Por exemplo, a Tabela 2 exemplifica uma situação que um mesmo artigo aparece três vezes, isso ocorre, pois, tal artigo foi publicado em coautoria. Neste caso, a frequência de cada palavra-chave que aparece igualmente deveria ser um, pois se trata do mesmo artigo publicado. No entanto, para as palavras “Bibliometria” e “Redes Sociais” a frequência é 3.

Tabela 2 - Exemplo de um artigo publicado em coautoria

Autor	G. Área	Título da Publicação	Ano	Palavras-Chave
....				
ID 1	Engenharias	Análise de Rede de Palavra-Chave	2010	Bibliometria; Redes Sociais; Cientrometria
ID 2	Engenharias	Análise de Rede de Palavra-Chave	2010	Bibliometria; Redes Sociais; Análise de Tópicos
ID 3	Engenharias	Análise de Rede de Palavra-Chave	2010	Bibliometria; Redes Sociais; Mineração de Texto
....				

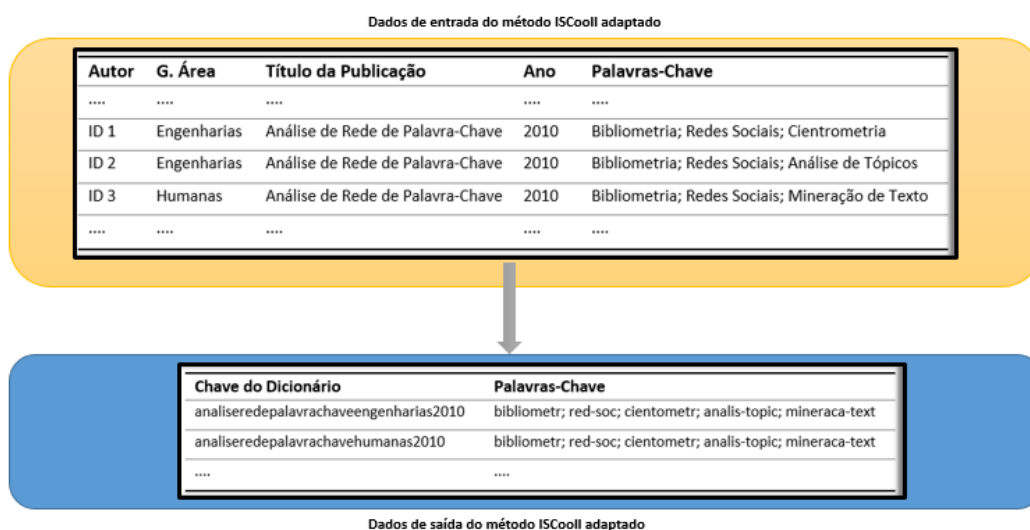
Fonte: Elaborado pelos autores.

Sendo assim, a etapa de normalização dos dados tem o objetivo de analisar os artigos para eliminar as inconsistências citadas. Para tanto, fez-se necessário uma adaptação no método ISCooll proposto por Dias e Moita (2015), a fim de identificar colaborações científicas automaticamente em grande volume

de dados. Enquanto o método ISCooll original utiliza um dicionário para vincular os artigos (chaves de um dicionário) a seus autores (identificadores dos currículos), o método ISCooll adaptado adota um dicionário para vincular os artigos ao conjunto união das respectivas palavras-chave. Outra diferença é que, enquanto o método ISCooll original utiliza o título do artigo concatenado com seu ano de publicação, o método ISCooll adaptado adiciona também a grande área da publicação, ou em caso de inexistência, a primeira grande área de atuação informada no currículo.

Assim, a inserção da grande área para a formação da chave do dicionário no método ISCooll adaptado dá-se pela decisão de considerar a contribuição do artigo publicado em coautoria realizada por diferentes grandes áreas. A Figura 4 apresenta um exemplo de resultado do método ISCooll adaptado, aplicado a um conjunto de artigos publicados em coautoria de diferentes grandes áreas.

Figura 4 - Exemplo de resultado do método *ISCooll* adaptado



Fonte: Elaborado pelos autores.

Portanto, ao término do processamento do conjunto de artigos pelo método ISCooll adaptado, o dicionário contemplará: (1) em casos de publicações em coautoria, mas quais os autores informaram a mesma grande área, será considerado um único artigo acompanhado com conjunto união das palavras-chave associadas e (2) em casos de coautoria, nas quais os autores

informaram grandes áreas distintas, o número de artigos serão de acordo com a quantidade de diferentes grandes áreas informadas pelos seus coautores.

3.2.3 Construção dos arquivos

A estratégia da construção dos arquivos tem como intuito facilitar as análises e contribuir na diminuição dos dados a serem processados. Nesta etapa, foi construído um conjunto de arquivos de resultados padrão para cada tipo de publicação. Para análises mais abrangentes, os arquivos de resultados padrão foram divididos em: arquivos que consideram a coautoria (processados pelas etapas de limpeza e normalização) e artigos que contém todo o conjunto de artigos (processados apenas pela etapa de limpeza).

O conjunto de arquivos de resultados padrão forma o repositório principal para manipulação e análises estatísticas sobre os tópicos de pesquisas. Inicialmente, foram construídos três arquivos de seguintes formatos:

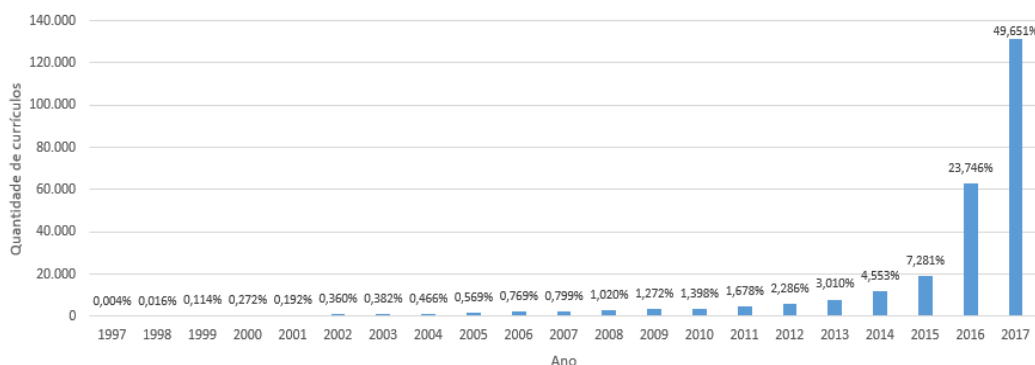
- a) arquivo de indivíduos por palavras-chave (identificador do currículo, palavra-chave1, ... palavra-chaveN);
- b) arquivo de palavras-chave por ano e grande área (palavra-chave, ano, frequência e grande área);
- c) arquivo de frequência de palavras-chave por ano e quantidade de publicação (ano, frequência palavra-chave, quantidade de publicação e grande área).

Por fim, vale destacar que a construção dos componentes deu-se na linguagem de programação *Python*, com a utilização das funcionalidades disponíveis na biblioteca *Natural Language ToolKit (NLTK)*. Neste caso, a *NLTK* contribuiu com os termos referentes às *stopWords* e aos algoritmos *Porter Stemmer* para a radicalização de língua inglesa, e *Snowball Stemmer* para a língua portuguesa. Para a geração dos resultados apresentados na próxima seção, foram manipulados e processados os arquivos de resultados padrão, utilizando *Python* e suas bibliotecas *Pandas* (para manipulação e análise de dados) e *NetworkX* (para construir e analisar redes e grafos).

4 Análise dos resultados

Os dados foram coletados em abril de 2017, totalizando 265.170 currículos de doutores. Para as análises, foram considerados os artigos publicados em anais de congressos e em periódicos referentes ao período de 1962 até 2016. Diante disto, espera-se que os artigos publicados em 2016 já estejam registrados nos currículos dos doutores, uma vez que tais artigos são cadastrados por eles após a data de publicação. A Figura 5 apresenta a distribuição dos currículos dos doutores com base na última data de atualização.

Figura 5 - Distribuição dos currículos dos doutores pela data da última atualização



Fonte: Elaborado pelos autores.

Apesar de existirem 10 currículos com data da última atualização em 1997, essa quantidade é irrelevante perante o conjunto total. De acordo com Dias (2016), a não atualização desses currículos pode ter motivos variados e de difícil reconhecimento. Contudo, a grande maioria foi atualizada recentemente, na qual 49,651% dos currículos (131.660) apresentam data de última atualização em 2017, e 73,397% (194.626) foram atualizados nos últimos dois anos.

Cada indivíduo tem um currículo na Plataforma Lattes, não existindo a possibilidade de duplicação da produção deste. No entanto, se considerado todo o repositório, um artigo produzido em colaboração pode aparecer várias vezes. Assim, para tornar as análises coesas ao contexto, foi necessária a consideração de duas visões sobre os dados extraídos: a **visão geral** que contabiliza todos os artigos e palavras-chave extraídos dos currículos e a **visão colaboração** que

considera apenas um único artigo em coautoria de mesma grande área do conhecimento associado ao conjunto união de suas respectivas palavras-chave. A Tabela 3 apresenta o quantitativo de artigos e palavras-chave extraídos dos currículos dos doutores, considerando as visões gerais e de colaboração.

Tabela 3 - Quantitativo de artigos e suas palavras-chave dos currículos de doutores

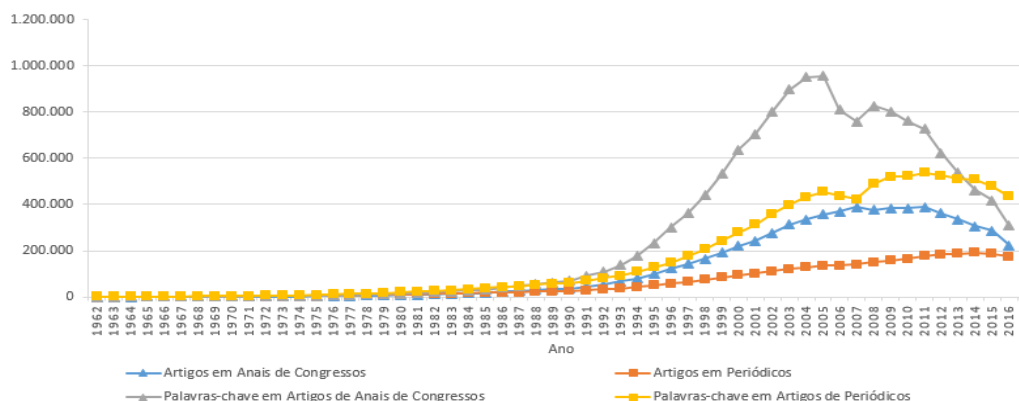
Tipo de Publicação		Artigos	Palavras-chave vinculadas aos artigos
Geral	Anais de Congresso	8.881.237 (65,4%)	14.284.910 (61,6%)
	Periódicos	4.685.011 (34,6%)	8.889.527 (38,4%)
	Total	13.566.248 (100%)	23.174.437 (100%)
Colaboração	Anais de Congresso	6.776.671 (67,5%)	14.839.228 (61,2%)
	Periódicos	3.263.993 (32,5%)	9.417.084 (38,8%)
	Total	10.040.664 (100%)	24.256.312 (100%)

Fonte: Elaborado pelos autores.

É possível destacar a preferência dos doutores sobre a publicação de artigos em congressos. Como esperado, o total de artigos, na visão colaboração, é cerca de 25% menor que da visão geral, devido ao fato de terem artigos registrados em coautoria. Entretanto, o número total de palavras-chave da visão colaboração é maior que da visão geral, fato este, devido aos autores cadastrarem diferentes palavras-chave e/ou grandes áreas para o mesmo artigo realizado em coautoria.

Outra situação esperada é o número de palavras-chave serem maior que o número de artigos, visto que cada artigo pode ter até seis palavras-chave associadas. Contudo, não há garantias que isso aconteça para todo conjunto de currículos analisados, uma vez que não existe obrigatoriedade quanto à inserção dessas palavras durante o cadastro na Plataforma Lattes. Isso pode ser mais bem analisado por meio da Figura 6, construída com base na visão colaboração, para ilustrar o número de artigos e palavras-chave ao longo dos anos.

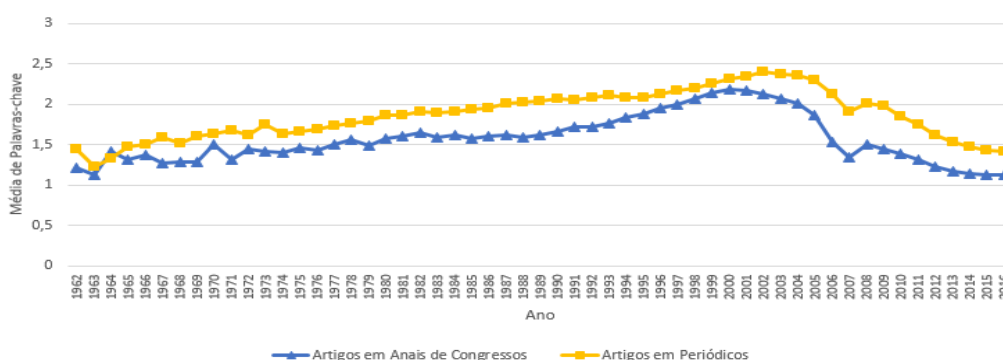
Figura 6 - Quantitativo de artigos e suas palavras-chave por ano



Fonte: Elaborado pelos autores.

Como pode ser observado, para a maioria dos anos, as curvas das palavras-chave, em anais de congressos e em periódicos, apresentam o mesmo comportamento dos artigos. Outra situação interessante é que, o maior número de palavras-chave não ocorreu no mesmo ano do maior número de artigos. Percebe-se, ainda, que as palavras-chave tiveram um crescimento considerável entre 1995 e 2005: em congressos cresceram 443% e em periódicos, 197%. No entanto, a partir de 2005, houve uma queda de 67% em congressos e 18,9% em periódicos e, consequentemente, após 2014, o número de palavras-chave em periódicos tornou-se maior que em congressos. Em continuação, a Figura 7, construída com base na visão geral, apresenta a média das palavras-chave por artigos ao longo dos anos.

Figura 7 - Média de palavras-chave por artigo publicado

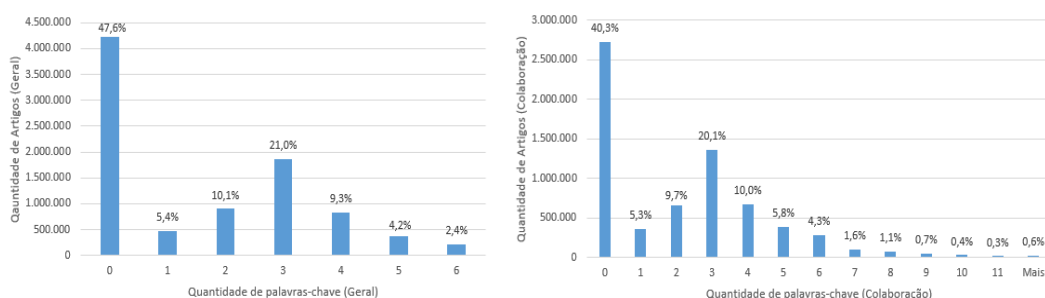


Fonte: Elaborado pelos autores.

É possível verificar que a média de palavras-chave por artigo de periódico é, geralmente, superior a de congressos. Contudo, o baixo valor de

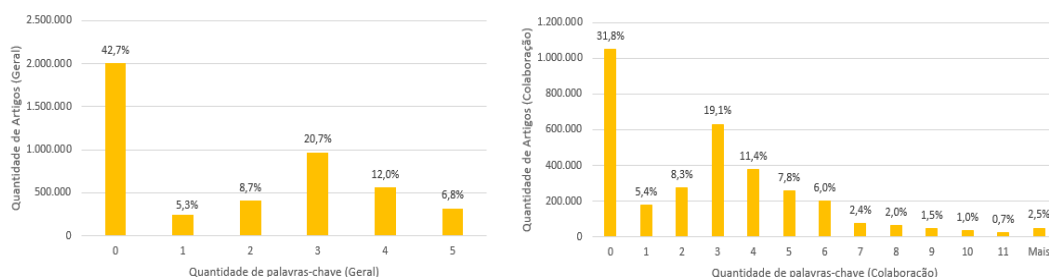
média encontrada é devido à influência dos artigos que apresentam poucas (ou nenhuma) palavra(s)-chave associada(s). Para facilitar essa visualização, os histogramas referentes à distribuição do número de artigos pela quantidade de palavras-chave foram gerados, a partir das visões geral e colaboração (Figuras 8 e 9).

Figura 8 - Distribuição dos artigos de congressos pelo número de palavras-chave



Fonte: Elaborado pelos autores.

Figura 9 - Distribuição dos artigos de periódicos pelo número de palavras-chave



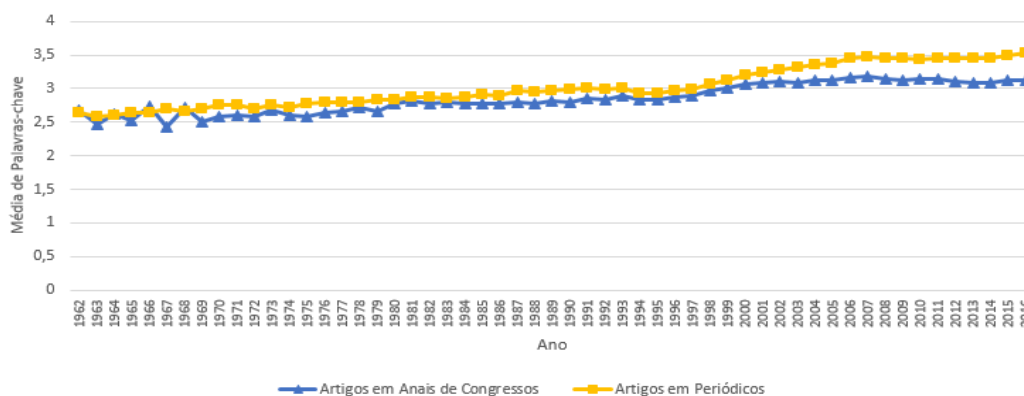
Fonte: Elaborado pelos autores.

Logo, fica explícito que grande parte dos artigos (em média 45% visão geral e 36% colaboração) não apresenta palavras-chave. Além disso, cerca de 15% têm entre uma e duas palavras-chave. Adicionalmente, por meio da visão colaboração, é possível comprovar que os doutores que realizaram trabalhos em coautoria, inseriram diferentes palavras-chave para o mesmo artigo, pois, os números de palavras são superiores a capacidade aceita pela Plataforma Lattes.

Diante disso, fica evidente o quanto os artigos sem palavras-chave têm impactos na média de palavras-chave por artigo. Com base nisso, optou-se por excluir todos os artigos sem palavras-chave associadas. Assim, as médias de

palavras-chave por artigo alteram significativamente, conforme mostrado na Figura 10, construída com base na visão geral.

Figura 10 - Média de palavras-chave por artigo publicado



Fonte: Elaborado pelos autores.

Igualmente ao analisado na Figura 8, a média de palavras-chave por artigo em periódicos é superior à média em congressos para, praticamente, todo o período e com valores médios superiores. Outra informação interessante é que, a partir de 1999, a média de palavras-chave sempre foi superior a três para artigos em congressos e periódicos, destacando-se as médias de periódicos que, para o mesmo período, cresceu cerca de 8% em relação a congressos.

A grande quantidade de palavras-chave cadastradas na Plataforma Lattes pode servir de base para verificar o desenvolvimento da ciência brasileira, uma vez que têm a função de destacar os principais assuntos que permeiam o artigo. Apesar disto, a estratégia de analisar termos extraídos de títulos dos artigos é o foco principal da maioria dos estudos encontrados e, no caso da Plataforma Lattes, é o foco de todos. Embora, esta estratégia tenha utilidade, notadamente, ela apresenta suas limitações, já que nem sempre os títulos conseguem expressar todo o conteúdo de um trabalho.

Com base nisso e de forma inédita, a Tabela 4 e a Figura 11, construídas a partir da visão geral e utilizando apenas o conjunto de artigos que tem palavras-chave, possibilitam verificar o percentual de palavras-chave que estão associadas aos respectivos títulos dos artigos.

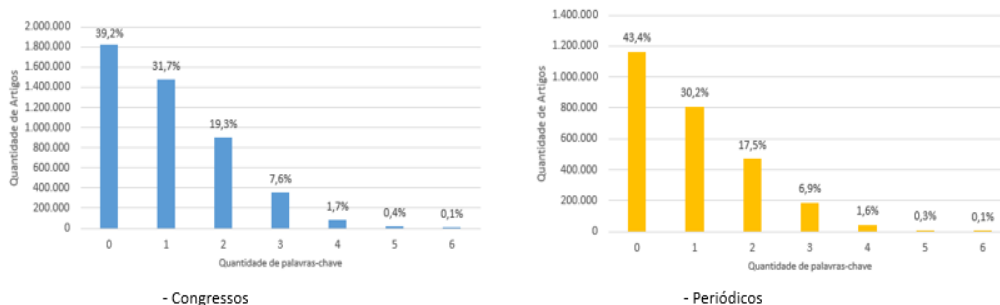
Tabela 4 - Número de palavras-chave contidas nos títulos das publicações

Artigos Publicados	Artigos Publicados que contém palavras-chave	Total de Palavras-chave	Palavras-chave contidas nos títulos dos artigos
Anais de Congresso	4.650.037	14.284.910	4.756.539 (33,3%)
Periódicos	2.682.718	8.889.527	2.536.825 (28,6%)

Fonte: Elaborado pelos autores.

Como pode ser observado, os resultados apresentados pelas palavras-chave são bem diferentes dos que podem ser mostrados pelos títulos, pois, no caso dos doutores que têm currículos cadastrados na Plataforma Lattes, apenas 33,3% das palavras-chave de artigos em anais de congressos e 28,6% em periódicos estão contidas nos títulos de seus artigos publicados. Para maiores detalhes desta análise, a Figura 11 apresenta a distribuição dos artigos pelo número de palavras-chave existentes nos títulos.

Figura 11 - Distribuição dos artigos pelo número de palavras-chave existentes nos títulos



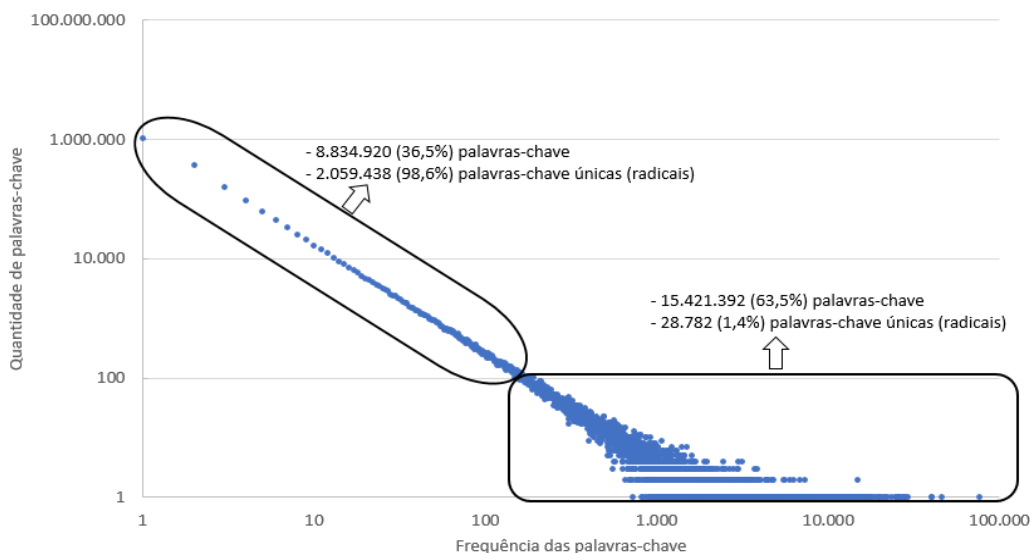
Fonte: Elaborado pelos autores.

Dessa maneira, é possível verificar que à medida que o número de palavras-chave nos títulos aumenta, o percentual de artigos diminui aceleradamente. Dentre estes, destaca-se o percentual elevado (cerca de 40%) dos artigos que não apresentam palavras-chave no título. Isso fundamenta, ainda mais, a originalidade dos estudos sobre palavras-chave da Plataforma Lattes.

Uma informação importante, para compreender o conjunto das palavras-chave dos artigos, diz respeito a sua quantidade e frequência. Com base na visão

colaboração, a Figura 12 apresenta a distribuição de quantidade e frequência de todas as palavras-chave. Além disso, mostra o quantitativo das palavras-chave únicas, ou seja, aquelas que foram processadas pela etapa de limpeza e agrupamento dos dados, conforme descrito na Seção “Desenvolvimento”.

Figura 12 - Distribuição de quantidade de palavras-chave por sua frequência



Fonte: Elaborado pelos autores.

Assim, observa-se que 24.256.312 palavras-chave (conjunto global), após serem processadas, reduziram-se a 2.088.220 (palavras-chave únicas), comprovando a importância das etapas de tratamento dos dados para diminuição das informações a serem manipuladas. Também, nota-se que 63,5% do conjunto global das palavras-chave equivalem a 1,4% do total de palavras-chave únicas de frequência superior a 100, enquanto os 36,5% restantes equivale a 98,6% das palavras-chave únicas que aparecem menos de 100 vezes. Isto confirma uma das leis clássicas da Bibliometria, Lei de Zipf ou Lei do Mínimo Esforço, na qual se descreve que um pequeno número de palavras é usado muito mais frequentemente.

Em continuação, para conhecer os tópicos que se constituem como os principais assuntos de pesquisa, todas as palavras-chave únicas foram ranqueadas de acordo com a medida de importância baseada na frequência. A Tabela 5 apresenta a classificação das principais palavras-chave e, dentre estas,

o ranque do quantitativo de doutores envolvidos em cada palavra. A coluna “Flutuação” apresenta o desvio de posição no ranque do número de doutores envolvidos em cada palavra-chave, se comparado ao ranque da frequência.

Tabela 5 - As principais palavras-chave e o número doutores entre 1962 e 2016

Ranque	Palavras-chave		Doutor		Flutuação
1	Educação	75.855	Educação	20.126	0
2	Formação do Professor	46.032	Ensino	10.080	8
3	Enfermagem	40.228	Políticas Públicas	9.104	3
4	Epidemiologia	40.158	Epidemiologia	8.512	0
5	Crianças	29.048	Formação do Professor	7.961	-3
6	Políticas Públicas	28.100	Cultura	7.903	9
7	Amazônia	27.626	Brasil	7.516	5
8	Bovinos	27.236	Crianças	7.376	-3
9	Idoso	26.291	Diagnóstico	6.958	4
10	Ensino	25.760	Educação Ambiental	5.823	4
11	Ratos	25.480	Idoso	5.737	-2
12	Brasil	24.981	Amazônia	4.797	-5
13	Diagnóstico	24.520	Enfermagem	3.975	-10
14	Educação Ambiental	23.194	Bovinos	3.793	-6
15	Cultura	22.123	Ratos	3.665	-4

Fonte: Elaborado pelos autores.

Notadamente, a palavra-chave “Educação” tem destaque na pesquisa nacional entre os doutores, visto seu alto valor de popularidade. Nesse caso, “Educação” foi também a palavra que mais despertou o interesse do maior número (20.126) de doutores ao longo da história. As palavras-chave “Formação do Professor” e “Ensino” corroboram as pesquisas científicas ligadas à educação.

Saúde é outro assunto que recebeu grande atenção, destacam-se as palavras-chave “Enfermagem”, “Epidemiologia”, “Crianças”, “Idoso”, “Ratos” e “Diagnósticos”. Também, se pode observar temas relacionados à pecuária, problemas ambientais e humanas, como “Bovinos”, “Amazônia” e “Educação Ambiental”, “Cultura” e “Políticas Públicas”. E, por último, não menos importante, destaca-se a palavra-chave “Brasil”, tema genérico de investigação que, frequentemente, é aplicado a trabalhos em áreas distintas.

5 Considerações finais

Este estudo apresentou as primeiras análises bibliométricas sobre as palavras-chave de artigos publicados em anais de congressos e em periódicos pelos doutores que têm currículos na Plataforma Lattes. A partir dos resultados encontrados, foi possível: (1) apresentar a caracterização geral das palavras-chave utilizadas pelos doutores; (2) verificar a diferença dos resultados encontrados pelas palavras-chave e termos extraídos dos títulos e; (3) destacar os principais tópicos desenvolvidos pelos doutores nos últimos 55 anos.

Tais análises caracterizam-se como um importante mecanismo, pois podem identificar quais tópicos mais impactantes são desenvolvidos no Brasil e, assim, servir como apoio a diversos tipos de tomada de decisão. Ao analisar as palavras-chave da Plataforma Lattes, é possível considerar artigos publicados em congressos, o que não seria factível verificar em outras fontes de dados internacionais. Com isso, é possível obter uma visão mais precisa dos tópicos mais investigados pela ciência brasileira.

A sistemática apresentada para a filtragem, tratamento e manipulação das palavras-chave contidas nos artigos de currículos da Plataforma Lattes forneceu uma metodologia nova para estudos sobre a identificação e a análises de tópicos de pesquisas. O mesmo procedimento pode ser replicado em outros conjuntos de palavras-chave de artigos científicos.

Por fim, constatou-se a validade dos resultados do ponto de vista quantitativo, uma vez que foi possível auxiliar no entendimento sobre o desenvolvimento da ciência brasileira; por outro lado, no caso da Plataforma

Lattes, que possui dados de periódicos e conferências das mais variadas áreas de pesquisas, com diferentes níveis de qualidade, análises apenas quantitativas para destacar tópicos podem induzir a interpretações equivocadas. Com isso, espera-se que com trabalhos futuros sejam incorporadas análises que considerem fatores temporais, para determinar a relevância de um tópico em determinada época, características qualitativas apropriadas ao contexto em análise e, ainda, análises baseadas em métricas de redes sociais para determinar as palavras-chave mais centrais e, consequentemente, com maior grau de importância.

Referências

- BRITO, A. G. C.; QUONIAM, L.; MENA-CHALCO, J. P. Exploração da Plataforma Lattes por assunto: proposta de metodologia. **Transinformação**, Campinas, v. 28, n.1, p. 77-86, jan./abr. 2016.
- 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, mar. 2015.
- CARNEIRO, D. L. C. M. C&T em prol da cidadania. **Minas Faz Ciência**, Belo Horizonte, v. 1, n.15, p. 5-5, 2003.
- CAVACINI, A. Recent trends in Middle Eastern scientific production. **Scientometrics**, Dordrecht, v. 109, n. 1, p. 423-432, Oct. 2016.
- CHEN, G. et al. Identifying the research focus of Library and Information Science institutions in China with institutions-specific keywords. **Scientometrics**, Dordrecht, v. 103, n. 2, p. 707-724, May 2015.
- CHOI, J.; Yi, S.; Lee, K. C. Analysis of keyword networks in MIS research and implications for predicting knowledge evolution. **Information & Management**, Amsterdam, v. 48, n. 8, p. 371-381, Dec. 2011.
- 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. ISSN 1808-5245.

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.

DIGIAMPIETRI, L. A.; PERES, S. M.; SILVA, L. A. Rede de Relacionamentos Brasileira de Inteligência Artificial e Computacional. In: ENCONTRO DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL, 11., 2014, São Carlos. **Anais...** São Carlos: SBC, 2014. p. 1-6.

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.

HONG, Y. et al. Knowledge structure and theme trends analysis on general practitioner research: a co-word perspective. **BioMed Central Family Practice**, London, v. 17, n. 10, p. 1-15, Jan. 2016.

KHAN, G. F.; WOOD, J. Information technology management domain: emerging themes and keyword analysis. **Scientometrics**, Dordrecht, v. 105, n. 2, p. 959-972, Nov. 2015.

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

PEI-CHUN, L.; HSIN-NING, S.; TE-YI, C. Assessment of ontology-based knowledge network formation by Vector-Space Model. **Scientometrics**, Dordrecht, v. 85, n. 3, p. 689-703, Dec. 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.

MCCLOSKEY, D. N. **The rhetoric of economics**. Wisconsin: University of Wisconsin Press, 1998.

MEDEIROS, C.; MENA-CHALCO, J. P. The dynamics of multidisciplinary research networks: mining a public repository of scientists CVs. In: WORLD SOCIAL SCIENCE FORUM, 2013, Montreal. **Anais...** Montreal: UNESCO, 2013. p. 1-17.

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

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.

POLLACK, J.; ADLER, D. Emergent trends and passing fads in project management research: a scientometric analysis of changes in the field. **International Journal of Project Management**, Amsterdam, v. 33, n. 1, p. 236-248, 2015.

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

SAES, S. G. **Aplicação de métodos bibliométricos e da co-word analysis na avaliação da literatura científica brasileira em ciências da saúde de 1990 a 2002**. 2005. Tese (Doutorado em Saúde Pública) - Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2005.

SILVA, F. N. et al. Using network science and text analytics to produce surveys in a scientific topic. **Journal of Informetrics**, Amsterdam, v. 10, n. 2, p. 487-502, May 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.

TRUCOLO, C. C.; DIGIAMPIETRI, L. A. Análise de tendências da produção científica nacional na área de Ciência da Informação: estudo exploratório de mineração de textos. **AtoZ: novas práticas em informação e conhecimento**, Curitiba, v. 3, n. 2, p. 87-94, jul./dez. 2014a.

TRUCOLO, C. C.; DIGIAMPIETRI, L. A. Trend Analysis of the Brazilian Scientific Production in Computer Science. **Revista de Sistemas de Informação da FSMA**, Macaé, n. 14, p. 2-10, jul./dez. 2014b.

VINKERS, C. H.; TIJDINK, J. K.; OTTE, W. M. Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. **British Medical Journal**, London, v. 351, Dec. 2015.

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.

ZHU, D. et al. Small-world phenomenon of keywords network based on complex network. **Scientometrics**, Dordrecht, v. 97, n. 2, p. 435-442, Nov. 2013.

ZHU, L. et al. Keywords co-occurrence mapping knowledge domain research base on the theory of Big Data in oil and gas industry. **Scientometrics**, Dordrecht, v. 105, n. 1, p. 249-260, Oct. 2015.

YI, S., CHOI, J. The organization of scientific knowledge: the structural characteristics of keyword networks. **Scientometrics**, Dordrecht, v. 90, n. 3, p. 1016-1026, Mar. 2012.

An analysis of the main research topics investigated by Brazilian researchers with doctoral degrees

Abstract: The production and publication of scientific papers presented an impressive growth in the last decades, being the Internet the main factor of access and diffusion of these to the community. Given this context, there is a global interest in all areas of knowledge regarding studies of scientific production data, in order to know what has been done about science. The understanding of how the researches have evolved may serve as a basis to build scientific policies that accelerate the progress of science or to impel research groups to become more productive. In this context, the aim of this study is to analyze the research topics published during the last 55 years in the Brazilian Science trajectory of researchers with a doctoral degree, in order to map scientific knowledge and identify hot topics. For this, the general characterization and an analysis of data related to the keywords of the articles published in congresses annals and in periodicals are carried out, written by the researchers who have curricula registered in the Lattes Platform. The results showed an overview of the keywords used by the PhDs, and identification and analysis of the main research topics developed by them. Also the validity of the results from the quantitative point of view was verified, to help the understanding about the development of the Brazilian science.

Keywords: Keywords. Research topics. Bibliometric analysis. Lattes Platform. Brazilian science.

Recebido: 14/7/2017

Aceito: 3/10/2017

