CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

UMA ANÁLISE TEMPORAL DOS PRINCIPAIS TÓPICOS DE PESQUISA DA CIÊNCIA BRASILEIRA A PARTIR DAS PALAVRAS-CHAVE DE PUBLICAÇÕES CIENTÍFICAS

JETHER OLIVEIRA GOMES

Orientador: Prof. Dr. Gray Farias Moita
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof. Dr. Thiago Magela Rodrigues Dias
Centro Federal de Educação Tecnológica de Minas Gerais

Belo Horizonte
Março de 2019

JETHER OLIVEIRA GOMES

UMA ANÁLISE TEMPORAL DOS PRINCIPAIS TÓPICOS DE PESQUISA DA CIÊNCIA BRASILEIRA A PARTIR DAS PALAVRAS-CHAVE DE PUBLICAÇÕES CIENTÍFICAS

Tese apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Doutor em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional
Linha de pesquisa: Sistemas Inteligentes

Orientador: Prof. Dr. Gray Farias Moita

Coorientador: Prof. Dr. Thiago Magela Rodrigues Dias

Belo Horizonte
Março de 2019

Gomes, Jether Oliveira
G633u Uma análise temporal dos principais tópicos de pesquisa da ciência brasileira a partir das palavras-chave de publicações científicas. / Jether Oliveira Gomes. – – Belo Horizonte, 2019.
124 f. : il.

Tese (Doutorado) – Centro Federal de Educação Tecnológica de Minas Gerais, Programa de Pós-Graduação em Modelagem Matemática e Computacional, 2018.
Orientador: Prof. Dr. Gray Farias Moita
Coorientador: Prof. Dr. Thiago Magela Rodrigues Dias

Bibliografia

1. Modelos Matemáticos. 2. Ciência – Terminologia - Pesquisa. 3. Bibliometria. I. Moita, Gray Farias. II. Centro Federal de Educação Tecnológica de Minas Gerais. III. Título

CDD 511.8

Elaboração da ficha catalográfica pela Bibliotecária Elisângela Gonçalves Barbieri CRB-6: 2624 / CEFET-MG

Esta folha deverá ser substituída pela cópia digitalizada da folha de provação fornecida pelo Programa de ós-graduação.

Dedico esta tese a minha esposa Layse e aos meus pais José Francisco e Edvirgens.

AGRADECIMENTOS

Agradeço a Deus por tudo. Por me amparar nos momentos difíceis, me capacitar para superar as dificuldades e me suprir em todas as minhas necessidades. Agradeço a Ele todas as conquistas alcançadas durante a minha vida. Agradeço a minha esposa Layse, que é o maior presente que Deus poderia ter me dado nesta vida. Por toda felicidade, carinho, compreensão, apoio, incentivo e dedicação.

Em especial aos meus grandes amigos, professores, orientadores e ídolos, Dr. Gray Farias Moita e Dr. Thiago Magela Rodrigues Dias, aos quais aprendi a admirar e respeitar por serem profissionais exemplares, pela dedicação, comprometimento, paciência, amizade, confiança e seus conhecimentos repassados. Tenho muito orgulho de citá-los como os principais responsáveis pela minha formação acadêmica e pela realização do sonho que é alcançar o mais alto grau acadêmico. A vocês, meu eterno muito obrigado!

À instituição CEFET-MG, que sempre me proporcionou uma ótima estrutura para que eu pudesse exercer as minhas atividades como aluno desde a época de mestrado. Em especial, agradeço a todos os professores do PPGMMC (Programa de Pós-Graduação em Modelagem Matemática e Computacional) que contribuíram significativamente em minha formação, principalmente Henrique Elias Borges, Paulo Eduardo Maciel de Almeida, Marcone Jamilson Freitas Souza e Rodrigo Tomas Nogueira Cardoso.

Também registro meu agradecimento aos professores Adilson Luiz Pinto (UFSC) e Henrique Elias Borges, pelas importantes considerações, análises e sugestões que incentivaram a continuidade e o aprimoramento dos estudos descritos nesta tese.

Aos queridos amigos do grupo de pesquisa ABC do CEFET-MG, principalmente, Tales Moreira e Patrícia Dias pelo grande apoio durante o desenvolvimento desta tese.

Aos queridos amigos da Vale/VLI, em especial, Daniel Dutra, Fernando Alcantara, Leandro Fulgêncio, José Campolina, Kesley Julianelli, Nilton Cordeiro e Vanessa Ajeje, pelos importantíssimos direcionamentos de carreira e/ou apoio incondicional durante o processo de doutoramento. Tenho muito orgulho de citá-los como as minhas principais referências no ambiente corporativo. Vocês impactaram muito minha vida profissional. Meu eterno muito obrigado!

Aos queridos amigos do meio acadêmico, principalmente Bruno Pimentel, Frederico Miranda, Guilherme Tavares e Vinícius Morais, pelo grande apoio durante o processo de doutoramento.

Por fim, agradeço a todos os amigos, colegas e familiares que contribuíram direta e indiretamente para a realização desta tese.

“O conhecimento serve para encantar as pessoas, não para humilhá-las.”
(Mario Sergio Cortella)

RESUMO

A produção de trabalhos científicos apresentou um crescimento significativo nas últimas décadas, sendo a internet o principal meio para seu acesso e difusão. Consequentemente, nota-se um esforço global de todas as áreas do conhecimento na confecção de estudos sobre dados da produção científica, a fim de conhecer o que tem sido pesquisado. Tais estudos podem servir a diversos propósitos, como fornecer embasamento para a construção de políticas de incentivo à pesquisa visando novos avanços na ciência. No entanto, a maioria dos trabalhos encontrados analisam conjuntos restritos de dados, que são originados de repositórios internacionais, geralmente específicos de uma determinada área. Por outro lado, poucos e recentes trabalhos utilizam fontes de dados nacionais, manipulando, no entanto, uma quantidade restrita dos dados disponíveis em análises preliminares. Apesar de tais trabalhos apresentarem resultados significantes, não foram encontrados estudos prévios que englobem, de forma abrangente, o que é produzido pela comunidade científica brasileira. Além disso, a estratégia principal de alguns desses estudos, principalmente os que utilizam repositórios de dados nacionais, é analisar os títulos das publicações. Em outra perspectiva, uma estratégia que vem se destacando é a análise das palavras-chave das publicações científicas, tendo em vista que 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. Nesse contexto, esta tese apresenta-se como uma proposta de análise textual sobre o desenvolvimento científico brasileiro registrado ao longo da história na base curricular da Plataforma Lattes. Concomitantemente, é também a primeira análise sobre todo o conjunto de palavras-chave das publicações dos indivíduos com doutorado concluído que possuem currículos cadastrados na Plataforma Lattes. Para tanto, foi desenvolvido um arcabouço de componentes com a finalidade de filtrar e tratar os dados dos currículos para padronização e definição das informações essenciais a serem analisadas, e, ao mesmo tempo, diminuir o processamento computacional para geração dos resultados desejados. Os resultados iniciais são apresentados a partir da aplicação de análises bibliométricas e técnicas baseadas em análises de redes sociais sobre as palavras-chave de artigos publicados em anais de congressos e em periódicos do conjunto selecionado. Assim, quantitativamente, foi possível apresentar uma caracterização geral das palavras-chave utilizadas pelos doutores, e, com isso, destacar os principais tópicos de pesquisa desenvolvidos por eles ao longo dos últimos anos. No entanto, a análise apenas quantitativa dos tópicos pode não explorar por completo as características existentes nos dados contidos na Plataforma Lattes. Por isso, foi desenvolvida nesta tese a medida de importância de tópicos TF.FI, que leva em consideração tanto as características quantitativas das palavras-chave dos artigos quanto as características qualititativas (Fator de Impacto) dos periódicos em que tais artigos foram publicados. Como resultado final, as palavras-chave dos artigos publicados em periódicos entre 1997 e 2016 foram ranqueadas utilizando a medida TF.FI e analisadas em suas respectivas grandes áreas do conhecimento. Os resultados apresentados servirão de base para diversos outros estudos que visam entender o desenvolvimento da ciência brasileira nas diversas áreas do conhecimento.

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

ABSTRACT

The last few decades have shown a significant increase in the production of scientific papers and the internet has been its main vehicle of distribution and access. Consequently, a global effort has been made in all areas of knowledge through studies about scientific production data, aiming to gather information on what has been studied. Such studies can serve different purposes, one of them being to provide the basis for research incentive policies seeking new scientific development. However, most of the published studies analyze restricted groups of data originated in international data repositories that are generally specific to a certain area. On the other hand, a rare few recent studies have been utilizing national data sources. They manipulate, however, only a restrict amount of the data available in previous analysis. Even though these studies show significant results, it is still difficult to find earlier studies that encompass the production of the Brazilian scientific community as a whole. Besides, the strategy used by most studies is to analyze the titles of publications, especially by those using national data repositories. From another point of view, the strategy of analyzing the keywords in scientific papers stands out. Keywords are carefully chosen by their authors so as to point out the main topics in the paper in a clear and direct way. In this context, this thesis proposes a textual analysis of scientific development in Brazil as registered throughout history on the Lattes Platform basis of curricula vitae. It is also the first study encompassing the whole set of keywords in the studies made by individuals with a PhD whose curricula are registered on Lattes Platform. A group of components was developed in order to filter and process the data in the curricula with the purpose of standardizing and defining the essential information for analysis and, simultaneously, decrease the amount of computer processing which is necessary for generating the desired results. The initial results were obtained by applying bibliometric and technical analysis. These techniques were applied to the keywords found in papers published in congress proceedings and journals by the group of individuals considered in this research. It was possible then to present a quantitative study on the keywords used by PhDs and, thus, to highlight the main research topics developed through the last few years. Nevertheless, the mere quantitative analysis of topics may not be able to fully exploit the existing features in the data available through the Lattes Platform. Therefore, in this thesis, the importance measure of TF.FI topics was developed, which takes into account both the quantitative characteristics of the article keywords and the qualitative (Impact Factor) of the journals in which such articles were published. As a final result, the keywords of articles published in journals between 1997-2016 were ranked using the TF.FI measure and analyzed in their respective broad areas of expertise. The results presented will serve as the basis for several other studies aimed at understanding the development of Brazilian science in the various areas of knowledge.

KEYWORDS: Keywords, Research topics, Bibliometrics, Keyword network, Lattes Platform

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AI	*Activity Index*
ALC	Gestão na América Latina e no Caribe
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNKI	*China National Knowledge Internet*
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
C&T	Ciência e Tecnologia
ETD	*Emerging Trending Detection*
FI	Fator de Impacto
GTI	Gestão da Tecnologia da Informação
HTML	*HyperText Markup Language*
IA	Inteligência Artificial
KAI	*Keyword Activity Index*
NLTK	*Natural Language ToolKit*
P&D	Pesquisa e Desenvolvimento
SBBD	Simpósio Brasileiro de Banco de Dados
TDT	*Topic Trending Detection*
TF	Frequência de Termos
TF-IDF	Frequência Inversa do Documento
TF-FI	Frequência de Termos * Fator de Impacto
TVG	*Time Varying Graph*
URL	*Uniform Resource Locator*
WoS	*Web of Science*
XML	*eXtensible Markup Language*

SUMÁRIO

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, redes de relacionamentos e sítios para registro individual de produção científica são alguns exemplos de como a internet tem contribuído consideravelmente na quantidade de trabalhos publicados, permitindo que usuários não apenas acessem conteúdo disponível, mas também possam registrar a sua produção técnica e científica a partir de sua interação com esse meio. Dessa maneira, trabalhos publicados e disponibilizados 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 da divulgação científica contribuir para a popularização do conhecimento, aproxima o cidadão comum dos benefícios que tem direito de reivindicar para a melhoria do seu bem-estar social, dando-lhe uma visão mais clara sobre as causas e efeitos dos problemas que enfrenta no dia a dia.

Numa sociedade competitiva em escala mundial, implementar o conhecimento técnico e científico é tarefa primordial para alavancar o desenvolvimento econômico e social (CAVACINI, 2016), principalmente para os países em desenvolvimento que são consumidores deste conhecimento, de modo a identificar suas necessidades e peculiaridades (SAES, 2005). Porém, muitas vezes seu plano de implementação é condicionado pela existência de recursos limitados e por 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, superação das dificuldades e alcance dos pressupostos de racionalidade e objetividade (SAES, 2005).

Nesse contexto, 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 et al., 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.

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

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 por meio de análises bibliométricas ou técnicas de análises de redes sociais ou análises de tendências. Entretanto, títulos e resumos de trabalhos podem não representar exatamente os assuntos abordados, devido à necessidade de se preocuparem com a semântica e a estrutura dos termos. Logo, uma abordagem interessante é a análise das palavras-chave de um conjunto de publicações científicas, pois estas são inseridas cuidadosamente por seus respectivos autores e fornecem uma possibilidade de descrever os assuntos principais que permeiam o trabalho de forma clara e objetiva sem se preocupar com questões semânticas (MCCLOSKEY, 1998; YI e CHOI, 2012; KHAN e WOOD, 2015). Lee et al. (2010) destacam o pressuposto de que uma palavra-chave é o portador fundamental mais básico do conhecimento. Por isso, nesta tese, palavras-chave de publicações científicas são igualmente referenciadas como tópicos de pesquisas.

Dentre os diversos trabalhos encontrados na literatura, principalmente devido ao crescimento exponencial da produção científica registrada nas últimas décadas (VINKERS et al., 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, não podem representar em totalidade o que é produzido no Brasil.

De acordo com Brito et al. (2016), estudos dessa natureza são considerados urgentes no Brasil e podem retratar o que é desenvolvido e publicado em ciência, tecnologia e inovação, possibilitando gerar parâmetros que possam nortear esforços e investimentos para impulsionar resultados de pesquisa. Trucolo (2016), por sua vez, destaca que muitas vezes os investimentos são focados 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 – tanto em extensão geográfica quanto em diversidade cultural –, 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 a realização de análises sobre o patamar científico brasileiro, o importante e abrangente repositório de dados da Plataforma Lattes, desenvolvida e mantida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), é tido como um diferencial. Esse repositório é composto por dados referentes a grupos de pesquisas, instituições e currículos de mais de 8 milhões de indivíduos, número este que tem crescido linearmente a cada ano. Esses currículos têm suas informações geridas por seus próprios possuidores (TRUCOLO, 2016) e congregam dados sobre produções bibliográficas; formação acadêmica; áreas de atuação; trabalhos em anais de congressos e em periódicos; orientações; participação em eventos, entre outros. A Plataforma Lattes também é utilizada como fonte de informação para órgãos que avaliam o Sistema Nacional de Pós-Graduação do Brasil e para as agências de fomento que financiam pesquisas e ofertam bolsas de estudos.

O grande volume de dados existente nos currículos da Plataforma Lattes, disponível livremente na internet, representa o instrumento de maior importância para estudos sobre a produção científica brasileira (FERRAZ et al., 2014). No entanto, esse repositório ainda não foi amplamente analisado, apesar de ser utilizado para avaliar ou verificar dados de pesquisadores e/ou grupos destes (DIGIAMPIETRI, 2015). Dos trabalhos que exploram os dados dos currículos cadastrados na Plataforma Lattes como principal fonte para análises bibliográficas, poucos são os que analisam os tópicos investigados. Os que o fazem utilizam termos extraídos dos títulos de publicações científicas 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, esta tese apresenta uma análise temporal das principais palavras-chave existentes nos artigos científicos de todos os indivíduos com doutorado concluído 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 bibliométricas e técnicas 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.

1.1 Descrição do Problema

A ciência apresentou um crescimento impressionante nas últimas décadas quanto ao número de registros de publicações de artigos científicos (RIDKER e RIFAI, 2013). Tal crescimento é verificado por resultados de esforços coletivos de pesquisadores de todos os domínios em todas as regiões do mundo. Embora essa sistematização social tenha existido em épocas anteriores, o novo paradigma da tecnologia da informação, ou seja, a internet e seus serviços, é quem fornece a base fundamental para sua ampliação em toda a estrutura social (CASTELLS, 2010).

Para Ferreira (2010), a comunidade científica tem a necessidade e o compromisso de tornar público todo o conhecimento produzido durante o desenvolvimento de suas pesquisas. A divulgação deste conhecimento tem ajudado a compreender melhor o mundo, o que permite o desenvolvimento da sociedade e dos padrões da vida humana.

Num âmbito mais específico, entender o avanço da produção científica é de suma importância para aumentar a competitividade de determinados negócios ou auxiliar na criação de novas políticas públicas (SAES, 2005; CHEN et al. 2015; CAVACINI, 2016). Concomitantemente, vale destacar que geralmente há uma quantidade limitada de recursos para fomentar a pesquisa e um grande número de pesquisadores ou instituições interessadas nesses recursos. Logo, quanto mais amplo e preciso for esse entendimento, maior será a possibilidade de se determinar os recursos de maneira correta. Contudo, esse tipo de avaliação é uma tarefa extremamente complexa, pois envolve a análise de diferentes características tanto quantitativas como qualitativas. Além disso, não há um consenso sobre quais medidas ou características devem ser consideradas para a avaliação da produtividade científica (DIGIAMPIETRI, 2015).

Em continuação, com o crescimento exponencial do volume de publicações científicas em todas as áreas, e, consequentemente, das mais variadas fontes de informações na internet (BRITO et al., 2016), ficou ainda mais difícil para cientistas e decisores políticos medirem e avaliarem as complexas inter-relações do novo conhecimento que vem sendo gerado (SAES, 2005; OHNIWA et al., 2010; ZHU et al. 2015; SILVA et al. 2016). Nesse sentido, a presença de um grande volume de dados disponível em diferentes formatos e repositórios dificulta a realização de análises por parte dos usuários, que muitas vezes necessitam de uma determinada visão dos dados que não é explorada pelas interfaces de consultas frequentemente disponíveis, e a partir disso, cria-se a necessidade de mecanismos que facilitem as buscas e análises dos dados disponíveis.

Nesse sentido, tem sido revelado um especial interesse por partes dos pesquisadores de todos os domínios de pesquisas quanto à extração de conhecimentos de repositórios de dados científicos (MADLOCK-BROWN, 2014). Essa extração de conhecimento ocorre sob diversas perspectivas, nas quais se inclui, mesmo que de forma menos frequente, a identificação e a análise da evolução de tópicos de pesquisas. Dessa forma, é possível identificar o comportamento dos tópicos principais que os pesquisadores têm produzido ao longo do tempo. Esses trabalhos têm foco nos estudos de um determinado assunto ou área de conhecimento específica, e geralmente se utilizam de repositórios de dados estrangeiros como fonte principal para suas análises (BRITO et al., 2016). Assim, a literatura não apresenta análises de tópicos de pesquisas de maneira abrangente, que envolvam dados científicos capazes de representar em totalidade o histórico da produção dos pesquisadores brasileiros.

Vale ressaltar que, para a análise da produção científica de um determinado país, uma das maiores dificuldades pode estar relacionada à aquisição de um conjunto abrangente de informações relevantes, normalmente presentes em vários repositórios de dados nos mais diferentes formatos. Entretanto, no caso do Brasil, essa atividade pode ser facilitada pela utilização do repositório curricular da Plataforma Lattes.

Os currículos cadastrados na Plataforma Lattes possuem registros acadêmicos e, por isso, tornam-se uma fonte singular de informação, que pode servir como um instrumento da maior importância para o estudo do que foi produzido pela ciência brasileira. Atualmente, o repositório possui cadastros de diferentes tipos de publicações de mais de 55 anos de pesquisas dos cientistas brasileiros, sendo quase 14 milhões de trabalhos publicados em anais de congressos e em periódicos referentes a todas as áreas de conhecimento. Portanto, ao realizar análises da produção científica brasileira, é importante ressaltar que o Brasil é o quinto maior e quinto país mais populoso do mundo, que combinado à grande diversidade cultural, geográfica e social, torna a análise da comunidade científica brasileira ainda mais interessante e desafiadora (TRUCOLO, 2016).

Para Ferraz et al. (2014), não existe no mundo um repositório curricular nacional único semelhante à Plataforma Lattes. Mesmo com tal relevância e com os dados livremente na internet, o repositório de dados da plataforma não foi amplamente explorado. Dos trabalhos que utilizam o repositório da Plataforma Lattes como principal fonte de dados, poucos são os que identificam e analisam tópicos de pesquisas, e quando o fazem, realizam análises preliminares em um conjunto restrito de dados, como, por exemplo, de um determinado assunto ou área específica.

Neste ponto, tendo em vista a dificuldade na coleta, tratamento, processamento e análise de grande volume de dados científicos, é importante destacar que, geralmente, tantos os trabalhos que utilizam repositório de dados estrangeiros quanto os que utilizam dados da Plataforma Lattes, analisam um conjunto específico de dados a partir de termos existentes nos títulos e/ou resumos de publicações cientificas. Entretanto, os termos extraídos de títulos podem não representar, na totalidade, os assuntos que são tratados no documento científico, visto que possuem uma preocupação dos autores quanto à compreensão semântica. Diferentemente das palavras-chave, que são frequentemente associadas a trabalhos publicados e cujo objetivo é descrever os assuntos principais que permeiam a pesquisa de forma objetiva. Porém, a análise de palavras-chave de artigos cadastrados na Plataforma Lattes não é uma tarefa trivial, tendo em vista que a escolha das palavras-chave não segue um padrão pré-definido, e, dessa forma, pode ser feita sem a necessidade de seguir critérios específicos (FERRAZ et al., 2014). Além disso, Medeiros e Mena-chalco (2013) destacam que, dado o enorme volume de dados, não é uma tarefa trivial visualizar as redes de termos extraídos dos títulos de publicações dos indivíduos que possuem currículos cadastrados na Plataforma Lattes e que declaram atuar nas grandes áreas das Ciências Humanas, Ciências Sociais Aplicadas ou Linguísticas, Letras e Artes.

Sendo assim, o problema principal abordado nesta tese é identificar a evolução de todo o histórico do que tem sido desenvolvido acerca da comunidade científica pelos pesquisadores brasileiros nas diversas grandes áreas do conhecimento. Para tratar deste assunto, a presente tese realiza a extração, processamento e caracterização de todas as palavras-chave contidas em artigos publicados em anais de congressos e em periódicos por todos os indivíduos com doutorado concluído que possuem currículos cadastrados na Plataforma Lattes, para em seguida serem estudadas por meio de análises bibliométricas e técnicas de análises de redes sociais.

1.2 Objetivos e Contribuições

Buscando um entendimento mais abrangente sobre o que tem sido desenvolvido pela comunidade científica brasileira ao longo de toda sua trajetória, esta tese tem como objetivo geral identificar e analisar temporalmente o comportamento dos principais tópicos de pesquisas de interesse dos pesquisadores brasileiros, utilizando para isso as palavras-chave de quase 14 milhões de artigos científicos referentes a todas as grandes áreas do conhecimento, baseado em análises bibliométricas e aplicação de técnicas de redes sociais.

Para tanto, esta tese se propôs os seguintes objetivos específicos:

- Realização de uma revisão na literatura acerca dos seguintes temas: utilização de palavras-chave de artigos científicos; análise de tópicos de pesquisas e medidas de importância para classificação destes; mineração de textos; técnicas de análises de redes sociais e análises bibliométricas.
- Extração, tratamento e caracterização dos tópicos de pesquisas desenvolvidos pelos pesquisadores brasileiros ao longo dos últimos 55 anos de história cadastrados no repositório de dados da Plataforma Lattes.
- Análise bibliométrica para determinação de indicadores e compreensão sobre o desenvolvimento científico ao longo do tempo.
- Construção de redes de palavras-chave e, em seguida, aplicação de técnicas de análises de redes sociais para extração de conhecimento acerca delas.
- Adoção de medidas de importância para a classificação dos tópicos de interesses dos pesquisadores brasileiros ao longo dos anos.
- Proposta de uma nova medida de importância que combina tanto características quantitativas quanto qualitativas, capaz de levar em consideração a heterogeneidade dos dados científicos da Plataforma Lattes durante a classificação dos tópicos de interesses dos pesquisadores brasileiros ao longo do tempo.
- Realização de uma análise comparativa entre as medidas de importância de tópicos quantitativas e qualitativas para avaliar as respostas encontradas em um grande repositório de dados científicos heterogêneo e, com isso, selecionar a medida que melhor se adequa ao contexto dos dados utilizados para os estudos.
- A partir da medida de importância mais apropriada ao contexto, realização de análises temporais sobre a produção científica brasileira num âmbito geral e por grande área do conhecimento, no intuito de extrair conhecimento acerca do comportamento dos principais tópicos de pesquisas de maior interesse.

O propósito é explorar quantitativamente e qualitativamente todo o conjunto de palavras-chave de todos os artigos científicos publicados em anais de congressos e em periódicos por todos os doutores que possuem currículos cadastrados na Plataforma Lattes, não só estatisticamente, mas também com base em técnicas de análises de redes sociais, para verificar os esforços dedicados pelos pesquisadores brasileiros, e, concomitantemente, destacar os principais tópicos de pesquisas de cada uma das grandes áreas do conhecimento.

Quanto às contribuições, esta tese traz:

- Caracterização geral dos dados referentes às palavras-chave extraídas de todos os artigos publicados em anais de congresso e em periódicos existentes no repositório curricular da Plataforma Lattes. Com isso, é possível avaliar a qualidade do conjunto de palavras-chave, o que poderá servir de insumo para a construção de um *corpus*, e, consequentemente gerar uma taxonomia padrão de termos para a classificação das publicações científicas. Assim, será possível evitar que as inserções das palavras-chave pelos pesquisadores brasileiros em seus currículos se deem por campos textos sem nenhum tipo de validação.
- Análise quantitativa e qualitativa dos principais tópicos de pesquisas desenvolvidos por pesquisadores brasileiros em suas grandes áreas do conhecimento ao longo do tempo, como auxílio ao entendimento do que se tem produzido acerca da ciência brasileira. Assim, tópicos evidenciados como importantes em períodos passados podem ser comparados com os resultados esperados por políticas científicas passadas. Já os tópicos identificados como importantes em períodos mais recentes podem servir de (1) insumo para construção de políticas científicas futuras; (2) atração de novos cientistas, expandindo laços sociais; (3) orientação de professores quanto à aplicação de trabalhos em salas de aula, e (4) compreensão da mudança nos interesses das grandes áreas do conhecimento com o passar do tempo.
- A verificação de que a simples contagem das palavras-chave utilizadas por indivíduos em seus artigos pode possibilitar a identificação de especialistas de domínios, que podem ser convidados para diversos fins, como premiações por mérito acadêmico, revisão de artigos, formação de bancas para defesas de pesquisas, palestras, participação em novos grupos de pesquisas para impulsionar resultados, entre outros.
- Proposta de uma nova medida de importância de tópico de pesquisa (TF.FI) que é capaz de levar em consideração tanto as características quantitativas das palavras-chave dos artigos científicos quanto as características qualitativas dos periódicos em que tais artigos foram publicados. A medida TF.FI tem como foco principal o ranqueamento de tópicos de pesquisas advindos de fontes heterogêneas de dados, como no caso da Plataforma Lattes, que congrega dados de periódicos e conferências das mais variadas áreas de pesquisas com diferentes níveis de qualidade.
- Os métodos desenvolvidos e empregados nesta tese para extração de conhecimento de um enorme volume de dados científicos de características específicas possibilitam a construção de uma ferramenta para visualização e análise de tópicos de interesses dos pesquisadores brasileiros a partir do poderoso repositório curricular da Plataforma Lattes. Pesquisadores de todos os domínios, instituições de ensino, órgãos de pesquisas e decisores políticos poderiam utilizar a ferramenta para auxiliar na compreensão da evolução dos tópicos de pesquisas de interesses desenvolvidos pela ciência brasileira e na identificação de especialistas em determinados assuntos para auxiliar em diversos tipos de tomadas de decisão. Essa ferramenta tornaria isso possível sem a necessidade de os interessados terem conhecimentos técnicos acerca de análises bibliométricas, análises de tendências e técnicas de redes sociais, além de não se preocuparem com a coleta, o tratamento e a manipulação dos conjuntos de dados.

1.3 Estrutura da Tese

Esta tese foi organizada em sete capítulos. Após a Introdução, a Figura 1.1 apresenta graficamente a organização dessa pesquisa.

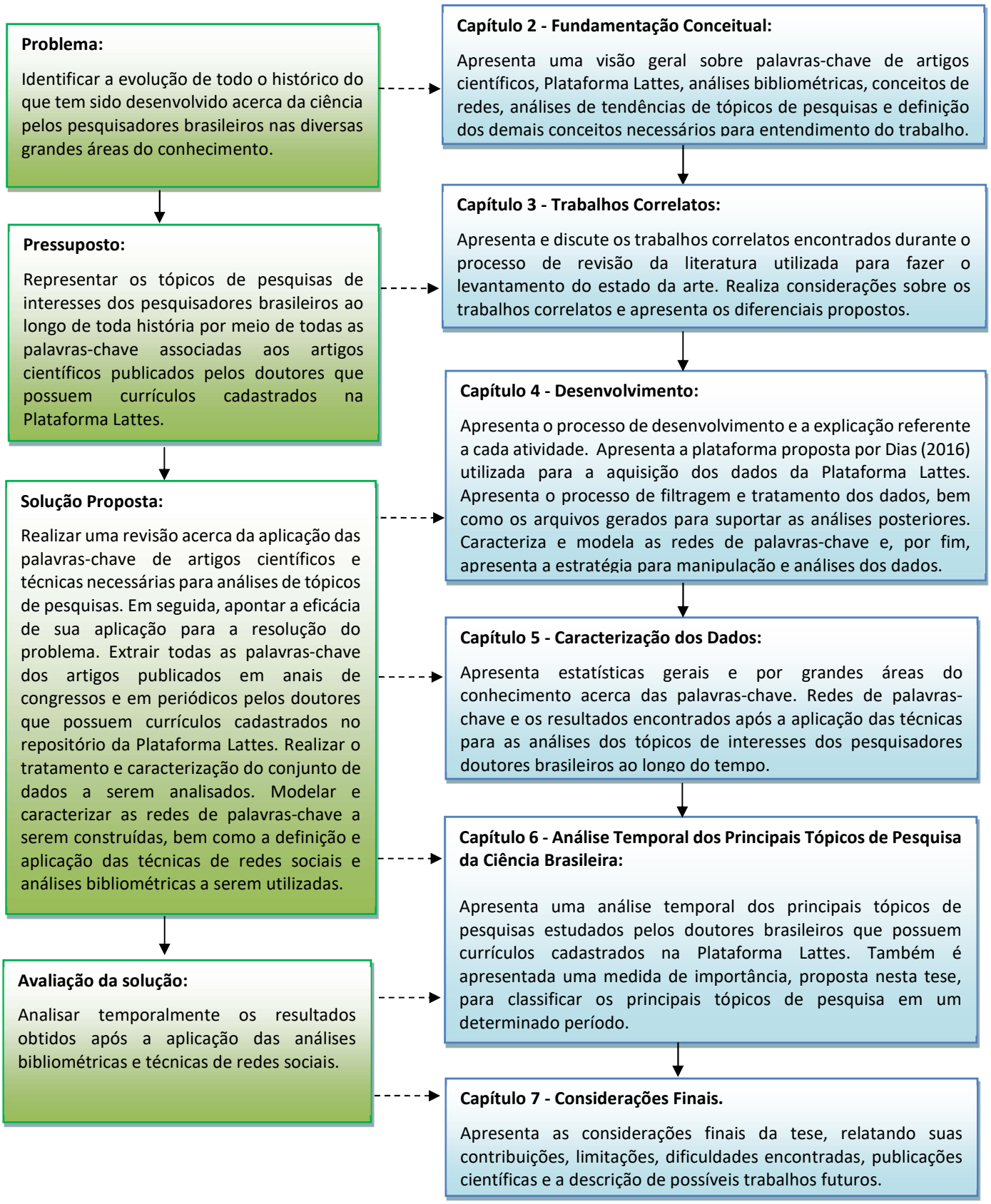


Figura 1.1 - Apresentação gráfica da organização da pesquisa em capítulos.

FUNDAMENTAÇÃO CONCEITUAL

Este capítulo modela o referencial teórico que é importante para o entendimento desta tese. Inicia-se com uma breve descrição acerca de palavras-chave de artigos científicos e Plataforma Lattes nas seções (2.1) e (2.2). Em seguida, nas seções (2.3), (2.4) e (2.5) são apresentadas definições, conceitos e principais técnicas adotadas para a realização de análises bibliométricas, análises de redes sociais e análise de tendências. Por fim, na Seção (2.6) são apresentadas as principais medidas de importância utilizadas para ranquear tópicos em documentos científicos.

2.1 Palavras-chave de Artigos Científicos

O pesquisador investiga, experimenta e produz conhecimentos em determinada área de estudo, colaborando com crescimento do conhecimento científico. Geralmente, os artigos científicos se mostram eficazes e rápidos para a divulgação dos resultados de pesquisa. São veiculados impressos ou eletrônicos em publicações especializadas como revistas, jornais científicos e em periódicos. Durante a confecção e submissão de um artigo científico, o pesquisador necessita elaborar as palavras-chave, objeto deste estudo, com intuito de melhor identificar e caracterizar o trabalho a ser publicado.

No trabalho de Ventura e Silva (2013), os autores definem palavras-chave como termos que descrevem o conteúdo semântico dos documentos. Segundo eles, estes termos são úteis para o processo de indexação e recuperação de documentos e para auxiliar os pesquisadores a identificar o tópico principal de um determinado texto com agilidade durante o processo de revisão na literatura. Já em Aquino e Aquino (2013), os autores ressaltam que palavras-chave constituem a menor seção da escrita em artigos científicos, contudo, de extrema importância, pois normalmente são utilizadas para permitir que o artigo seja posteriormente encontrado em sistemas eletrônicos de pesquisa. Sua escrita é composta de três ou quatro termos relevantes que melhor evidenciam as ideias centrais do texto, podendo ser termos simples e compostos, ou expressões características (PEREIRA, 2011).

De acordo com Miguéis et al. (2013), a investigação sobre a importância e as caraterísticas das palavras-chave de artigos científicos tem incidido sobre vários aspectos, como o da (1) eficiência na recuperação da informação através de algoritmos; (2) etiquetagem de documentos; (3) o impacto das palavras-chave em resumos de artigos na predição de citações; e (4) comparação com os termos extraídos de títulos, resumos e textos integrais para identificar principais assuntos abordados em documentos. Para maiores detalhes sobre conceitos e aplicações de palavras-chave, os seguintes trabalhos podem ser consultados: Sohrabi e Iraj (2017), Campos e Gomes (2008), Kobashi (2007), Maculan (2014) e Meireles et al. (2016).

No entanto, em Gonçalves (2008), o autor relata que geralmente não existe uma uniformização quanto à forma de escrita das palavras-chave entre os veículos de publicação. Tal fato deve-se possivelmente à pulverização de meios de publicação nas diversas áreas do conhecimento e a particularidade na definição das palavras-chave por parte dos pesquisadores. Assim, dado o enorme volume de publicações existentes na atualidade e dada sua relevância, a análise das palavras-chave inseridas por autores em suas produções científicas podem fornecer indícios para verificar fenômenos e tendências acerca do desenvolvimento científico (LEE et al., 2010; SU e LEE, 2010; YI e CHOI, 2012; KHAN e WOOD, 2015) como, por exemplo, evidenciar tópicos de interesses de determinado grupo de pesquisadores ou instituições de pesquisas, bem como evolução destes por meio de uma análise temporal.

2.2 A Plataforma Lattes

A Plataforma Lattes é a fonte principal de informação para estudos desta tese. De acordo com Lane (2010), em artigo publicado na revista Nature, a Plataforma Lattes é um poderoso exemplo de boas práticas para fornecimento de dados de alta qualidade. A autora relata também que órgãos federais, instituições e órgãos financiadores são incentivadores assíduos desta plataforma e que ela é uma das fontes de dados de pesquisadores mais confiáveis existente.

Essa plataforma foi construída em 1997 pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) em parceria com o Grupo Stela, da Universidade Federal de Santa Catarina. Posteriormente, tornou-se grande conhecida da comunidade científica e tecnológica no Brasil, países latino-americanos e europeus (DIAS et al., 2014). Ela é constituída por um único sistema de informação e seu importante repositório de dados é composto pela integração de informações dos currículos, grupos de pesquisa e de instituições. A Plataforma Lattes é utilizada pelo CNPq, agências e fundações de apoio à ciência e tecnologia, para auxiliar ações de planejamento, gestão e operacionalização de fomento, bem como para a formulação das políticas dos Ministério de Ciência e Tecnologia (CNPq, 2016). Além disso, é utilizada também para órgãos que avaliam o Sistema Nacional de Pós-Graduação do Brasil (BRITO et al. 2016).

O modelo de currículo fornecido pela Plataforma Lattes é tido como um padrão nacional para registro de experiências dos estudantes e pesquisadores do país, adotado pela maioria das instituições de fomento, universidades e institutos de pesquisa do Brasil. Por sua riqueza de dados e sua crescente confiabilidade e abrangência, se tornou elemento indispensável para análise de mérito e competência dos pleitos de financiamentos na área de ciência e tecnologia (DIGIAMPIETRI, 2015). A partir disto, o repositório de dados da Plataforma Lattes passou a ser uma importante fonte para vários estudos em diversas áreas do conhecimento, a fim de analisar fenômenos e tendências acerca da ciência brasileira com o intuito de produzir conhecimento original.

Atualmente, são mais de 8 milhões de currículos cadastrados na Plataforma Lattes. Esses currículos contemplam dados históricos de pesquisadores de todas as grandes áreas do conhecimento. As grandes áreas do conhecimento cadastradas na Plataforma Lattes estão distribuídas da seguinte forma: Ciências Agrárias, Ciências Biológicas, Ciências da Saúde, Ciências Exatas e da Terra, Ciências Humanas, Ciências Sociais Aplicadas, Engenharias e Linguísticas Letras e Artes. Cada currículo está disponível em dois formatos (HTML e XML) e dividido em oito seções principais, a saber: dados gerais, formação, atuação, projetos, produções, eventos, orientações e bancas.

A seguir, são apresentadas as informações mais relevantes de cada seção citada anteriormente:

- **Dados gerais**: contém o nome do autor do currículo, sua biografia, as formas de citação bibliográficas deste, seu endereço profissional e a lista de prêmios e títulos.
- **Formação acadêmica / titulação**: contém os dados referentes às formações e titulações, como, ensino fundamental, médio, profissional, curso de aperfeiçoamento, graduação, especialização, mestrado, doutorado, pós-doutorado, livre-docência e formações complementares.
- **Projetos**: são divididos em quatro grupos principais: pesquisa, desenvolvimento tecnológico, extensão e outros tipos.
- **Produções**: são distribuídas em produção bibliográfica, técnica e artística/cultural. Para cada tipo de produção, é permitido o cadastro de um conjunto de palavras-chave que melhor definem cada produção (é partir das palavras-chave cadastradas nas produções bibliográficas, mais precisamente de anais de congressos e periódicos, que são realizados os estudos para o desenvolvimento da atual pesquisa).
- **Orientações**: compreende informações sobre os tipos de orientação: dissertação de mestrado, tese de doutorado, monografia de conclusão de curso de graduação, aperfeiçoamento ou especialização, iniciação científica, supervisão de pós-doutorado, e orientação de outra natureza.
- **Eventos**: contém as informações sobre participação em eventos, congressos, exposições e feiras.
- **Bancas**: inclui informações tanto de participação em bancas de trabalhos de conclusão quanto da participação em bancas de comissões julgadoras.

Contudo, é importante destacar que as informações contidas nos currículos cadastrados no repositório da Plataforma Lattes são preenchidas manualmente pelos respectivos indivíduos responsáveis. Entretanto, existem apenas alguns tipos simples de controle das informações (por exemplo, impedindo o cadastramento de informações com o preenchimento de campos obrigatórios ou exigindo-se que algumas informações sejam obtidas de uma lista fornecida pela Plataforma). Porém, não há nenhum tipo mais sofisticado de verificação da completude ou corretude da informação (DIGIAMPIETRI, 2015). No caso do cadastramento de palavras-chave das publicações bibliográficas, nem o preenchimento obrigatório é exigido. Com isso, pode ser questionável o grau de confiabilidade das informações cadastradas pelos indivíduos em seus próprios currículos, pois, como não há supervisão direta, eles podem fornecer dados incertos (BRITO et al., 2016).

Todos os dados que constam na plataforma estão disponíveis para visualização pública na internet, com o objetivo de dar maior transparência e mais confiabilidade às atividades de fomento do CNPq e das agências que a utilizam; fortalecem o intercâmbio entre pesquisadores e instituições; e é fonte inesgotável de informações para estudos e pesquisas (MUGNAINI et al., 2011).

Embora os dados estejam disponibilizados, estes são apenas visualizados por meio de uma interface de consulta que apresenta cada currículo individualmente, e com isso, não permite uma análise mais abrangente. Diante disto, para uma análise mais detalhada de grupos, instituições ou até de toda a nação de cientistas brasileiros, se fazem necessárias técnicas e ferramentas para a tratamento e análises desses dados científicos.

2.3 Bibliometria

Existem diversas técnicas relacionadas à análise quantitativa da produção científica, algumas das mais usadas são: bibliometria, cienciometria, informetria e webometria. Todas têm funções semelhantes, mas, ao mesmo tempo, cada uma delas propõe medir a difusão do conhecimento científico e o fluxo da informação sob enfoques diversos (VANTI, 2002). Contudo, a tese descrita no presente texto enfoca em estudos acerca da bibliometria.

Pritchard (1969) foi quem popularizou o uso do termo "bibliometria", quando sugeriu em seu trabalho e que deveria substituir o termo "bibliografia estatística", até então utilizado. A diferença essencial entre essas abordagens se dá pelo fato da bibliometria utilizar mais métodos quantitativos do que discursivos. Para ele, a bibliometria pode ser entendida como um conjunto de técnicas e métodos quantitativos para a gestão de instituições envolvidas com o tratamento de informação científica.

Já para Tague-Sutcliffe (1992), a bibliometria é definida como o estudo dos aspectos quantitativos da produção, disseminação e uso da informação. Ele afirma que, na bibliometria, são desenvolvidos modelos matemáticos para medir esses processos, usando seus resultados para previsões e apoio à tomada de decisões. No trabalho de Araújo (2006), o autor ressalta que a utilização de técnicas estatísticas e matemáticas na busca por uma avaliação objetiva da produção científica é o ponto central da bibliometria.

Portanto, informações de avaliação e monitoramento da produção científica de indivíduos em particular, grupos de pesquisas, instituições, regiões geográficas, periódicos ou eventos, podem ser medidos pela utilização de análises bibliométricas em dados de trabalhos científicos. Porém, é importante ressaltar que não se trata de uma tarefa trivial, uma vez que a gestão e a avaliação da atividade científica exigem a formulação de técnicas e a formação de recursos humanos habilitados para compreender os fenômenos da construção do conhecimento e de como transformá-lo em resultados econômicos ou estratégicos (SANTOS e KOBASHI, 2009).

Em Vanti (2002), a autora elenca várias aplicações concretas que podem utilizar técnicas bibliométricas para descoberta de conhecimento científico. Algumas destas podem ser visualizadas a seguir:

- Identificar as tendências e o crescimento do conhecimento em uma área;
- Identificar as revistas do núcleo de uma disciplina;
- Mensurar a cobertura das revistas secundárias;
- Identificar os usuários de uma disciplina;
- Prever as tendências de publicação;
- Estudar a dispersão e a obsolescência da literatura científica;
- Prever a produtividade de autores individuais, organizações e países;
- Medir o grau e padrões de colaboração entre autores;
- Analisar os processos de citação e co-citação;
- Determinar o desempenho dos sistemas de recuperação da informação;
- Avaliar os aspectos estatísticos da linguagem, dos termos e das frases;
- Avaliar a circulação e uso de documentos em um centro de documentação;
- Medir o crescimento de determinadas áreas e o surgimento de novos tópicos.

Devido à importância, diversos trabalhos na literatura preocupam-se em ilustrar e atualizar os conceitos acerca de indicadores bibliométricos e suas aplicações. Neste contexto, podem ser citados os trabalhos de Guedes e Borschiver (2005), Araújo (2006), Santos e Kobashi (2009) e Machado JR et al. (2014). De um modo geral, todos os trabalhos citados apresentam as leis consideradas o cerne da disciplina e demais princípios bibliométricos. As três Leis Clássicas da Bibliometria são: *Lei de Lotka, Lei de Bradford e Leis de Zipf.*

A *Lei de Lotka* ou *Lei do Quadrado Inverso* surgiu em 1926, a partir de um estudo sobre a produtividade de cientistas mediante a contagem de pesquisadores presentes no *Chemical Abstracts*. Lotka constatou que um número menor de pesquisadores, supostamente de maior prestígio em uma determinada área do conhecimento, produz grande parte da literatura científica, e muitos pesquisadores, supostamente de menor prestígio, produzem pouco. A partir daí, formulou a lei dos quadrados inversos: $a_n = a_1/n^c$, onde a_n é a quantidade de pesquisadores que publicaram determinado número (n) de trabalhos, a_1 é o número de pesquisadores que publicaram um trabalho e c um valor constante para cada determinada área do conhecimento ($c \sim 2$). Em seguida, Price afirmou que $1/3$ da literatura é produzida por menos de $1/10$ dos pesquisadores mais produtivos, levando a uma média de 3,5 trabalhos por pesquisador e 60% dos pesquisadores produzindo um único trabalho (PELEIAS et al., 2010). Desde que estabelecida esta Lei, muitos estudos têm sido conduzidos para investigar a produtividade dos pesquisadores em distintas áreas do conhecimento. Porém, vale destacar que estudos realizados por Voos (1974), Bogaert et al. (2000) e Rao (1986) propõem variações e/ou questionamentos sobre a definição inicial de Lotka.

A *Lei de Bradford* ou *Lei da Dispersão*, que incide sobre conjunto de periódicos, surgiu de pesquisas conduzidas por Samuel C. Bradford em 1934. A proposição do estudo era descobrir a extensão na qual artigos científicos de um assunto específico apareciam em periódicos destinados a outros assuntos. Diante disso, ele observou que, em uma coleção de periódicos

de uma área do conhecimento, há vários núcleos de periódicos, e o número de periódicos por zona aumenta, enquanto a produtividade diminui. Assim, Samuel C. Bradford concluiu que considerando-se os periódicos dispostos em ordem decrescente de produtividade de artigos por dado assunto, pode-se definir três zonas, cada uma contendo $1/3$ dos artigos, onde a primeira possui poucos periódicos altamente produtivos, a segunda contém um conjunto maior de periódicos menos produtivos e, por fim, uma terceira com uma quantidade ainda maior de periódicos, no entanto, com produtividade ainda mais baixa que a segunda (ARAUJO, 2006). Esta lei é um instrumento útil para o desenvolvimento de políticas de aquisição e de descarte de periódicos, em nível de gestão de sistemas de recuperação da informação, gestão da informação e do conhecimento científico e tecnológico (GUEDES e BORSCHIVER, 2005). Porém, vale destacar que Araújo (2006) traz diversos autores que reformularam e aperfeiçoaram a definição inicial desta lei.

A *Lei de Zipf ou Lei do Menor Esforço*, formulada em 1949, descreve a relação entre termos num determinado texto suficientemente grande e sua ordem de série. Após análises num texto longo, Zipf observou uma correlação entre o número de termos diferentes e a frequência de seu uso. Em seguida, concluiu que existe uma regularidade fundamental na seleção e no uso dos termos onde um pequeno número de termos é usado muito mais frequentemente, indicando supostamente o assunto do documento. Sua proposta é de que, se listarmos os termos que ocorrem num texto em ordem decrescente de frequência, a ordem de série (posição) de um termo na lista multiplicada por sua frequência é igual a uma constante. A equação para esse relacionamento é: $k = r.f$, onde r é a posição do termo, f é a sua frequência e k é a constante. Por meio desse procedimento, foi possível verificar que um termo é utilizado muitas vezes em detrimento de outros. Igualmente, nas leis anteriores citadas, o método foi sendo aperfeiçoado, principalmente com estudos de frequência e co-ocorrência de descritores (ARAUJO, 2006).

Como meio de buscar conhecimento sobre a produção científica, diversos estudos pautam-se nas Leis Clássicas da Bibliometria que podem ser utilizadas individualmente ou ainda combinadas. A Figura 2.1 apresenta a bibliometria e as suas principais leis.

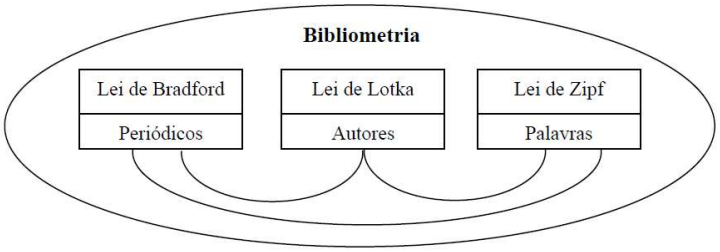


Figura 2.1 - As principais leis da bibliometria (Fonte: Machado JR et al., 2014).

Como complemento às técnicas de análise bibliométrica clássica, existem diversas outras teorias, como, por exemplo, a teoria epidêmica da transmissão de ideias desenvolvida por Goffman e Newill (1967). No trabalho, os autores relatam similaridades entre a propagação de ideias dentro de uma comunidade com a transmissão das doenças infecciosas, após realizarem um estudo por comparação do ciclo da esquistossomose e da informação.

Outra importante abordagem explorada na bibliometria para avaliar a dinâmica e o crescimento da literatura é análise de citações. Citação pode ser considerada a inserção de referências de obras consultadas para construção de conhecimento num determinado trabalho. No campo de estudo de citações, diversos princípios bibliométricos podem ser aplicados. Maiores detalhes sobre cada princípio podem ser encontrados no trabalho de Guedes e Borschiver (2005).

Em Araújo (2006), o autor destaca que por meio de análises de citações e seus princípios bibliométricos, informações importantíssimas podem ser reveladas, a saber: (1) evidência de elos entre indivíduos, instituições e áreas de pesquisas; (2) procedência geográfica das pesquisas; (3) autores mais influentes numa determinada área; (4) origem de pesquisas; (5) periódicos mais citados; (6) tipo de documento mais utilizado, dentre inúmeras outras. Mais informações acerca de análise de citações podem ser encontradas no trabalho de Vanz e Caregnato (2003), onde as autoras realizaram uma revisão dos estudos de citação na literatura com intuito de entender os processos de comunicação científica entre os pesquisadores.

Finalmente, vale destacar a Lei dos 80/20, que consiste na hipótese de que, em um fenômeno, 80% da demanda de informação se satisfaz com 20% do conjunto de fontes de informação. Essa lei é muita aplicada na redução de acervos (GUEDES e BORSCHIVER, 2005). A Tabela 2.1 relaciona as principais leis e princípios bibliométricos, seus focos de estudo e suas principais aplicações na gestão da informação, em sistemas de informação e em comunicação científica e tecnológica.

Tabela 2.1: Leis e princípios bibliométricos, seus focos de estudo, principais aplicações e áreas de interesse (Fonte adaptado: GUEDES e BORSCHIVER, 2005).

Bibliometria		
Leis e Princípios	**Focos de Estudo**	**Principais Aplicações**
Lei de Bradford	Periódicos	Estimar grau de relevância de periódicos em dada área do conhecimento
Lei de Lotka	Autores	Estimar grau de relevância de autores em dada área do conhecimento
Lei de Zipf	Palavras	Indexação automática de artigos científicos e tecnológicos
Ponto de Transição (T) de Golfam	Palavras	Indexação automática de artigos científicos e tecnológicos
Colégios Invisíveis	Citações	Identificação da elite de pesquisadores em dada área do conhecimento
Fator de imediatismo	Citações	Estimar o grau de relevância de artigos, cientistas e periódicos científicos em determinada área do conhecimento
Acoplamento Bibliográfico	Citações	Estimar o grau de ligação de dois ou mais artigos
Co-citação	Citações	Estimar o grau de ligação de dois ou mais artigos
Obsolescência da Literatura	Citações	Estimar o declínio da literatura de determinada área do conhecimento
Vida-média	Citações	Estimar a vida-média de uma unidade de literatura de dada área do conhecimento
Teoria Epidêmica de Golfmam	Citações	Estimar a razão de crescimento e declínio de determinada área do conhecimento
Lei do Elitismo	Citações	Estimar a razão de crescimento e declínio de determinada área do conhecimento
Frente de Pesquisa	Citações	Identificação de um padrão de relação múltipla entre autores que se citam
Lei dos 80/20	Demanda de Informação	Composição, ampliação e redução de acervos

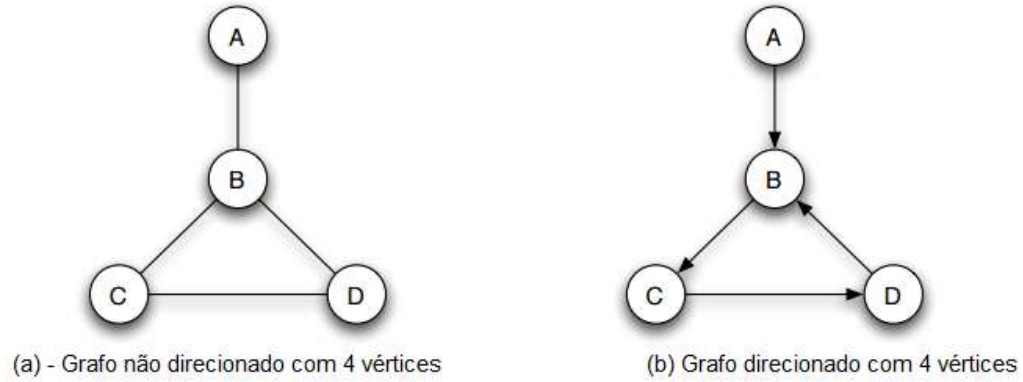
Diante do relatado na seção corrente, é possível verificar a relevância das análises bibliométricas como um dos elementos para apoiar pesquisadores na construção de indicadores que representam conhecimento acerca da dinâmica e evolução da produção científica e tecnológica. Neste ponto, é importante ressaltar que, dado o fato da bibliometria ter abrangência interdisciplinar, pode ser aplicada em qualquer área do conhecimento na busca de informações acerca do desenvolvimento científico.

2.4 Conceitos de Redes Sociais

O estudo de redes tem sido utilizado como ferramenta para entender vários estudos de sistemas naturais e sociais que contêm elementos que se relacionam entre si. No contexto desta tese, o conhecimento científico é considerado um sistema complexo com muitos conceitos associados que se adaptam ao mundo real ao longo do tempo (YI e CHOI, 2012), e, portanto, pode ser modelado por conceito de redes. O tipo de uma rede é definido pelo contexto em que está inserida e o seu relacionamento. Assim, diferentes tipos de relacionamentos produzem diferentes tipos de redes (MENEZES, 2012). Em Yi e Choi (2012), Medeiros e Mena-Chalco (2013), Cunha (2013), Cunha et al. (2013), Zhu et al. (2013), Souza et al. (2014), Easley e Kleinberg (2010), Khan e Wood (2015), Pollack e Adler (2015), Fraga (2016) e Brito et al. (2016), os autores apresentam diversos sistemas reais representados por diferentes tipos de redes em diversas áreas, como, por exemplo, as redes sociais.

De acordo com Newman (2001), uma rede social é um conjunto de pessoas ou grupos que têm algum tipo de relação entre si. No domínio científico, uma possibilidade de rede social é a de colaboração científica, em que vértices correspondem aos autores de publicações e as arestas à relação de coautoria. Redes Sociais são frequentemente representadas como grafos. Logo, um grafo $G = (V, E)$ é composto por um conjunto de vértices ou nós V e um conjunto de arestas E, sendo que cada aresta representa uma conexão entre dois elementos do conjunto V. Os grafos podem ter características esparsas ou densas. Um grafo esparso é aquele cujo número de arestas ($|E|$) é muito menor do que o número de vértices ao quadrado ($|V|^2$). Enquanto um grafo denso é aquele cujo número de arestas é próximo do quadrado do número de vértices (CORMEN et al., 2001).

De acordo com Diestel (2010), um grafo pode ser do tipo direcionado ou não direcionado, dependendo do direcionamento das relações entre seus vértices (Figura 2.2). Num grafo não direcionado (Figura 2.2(a)), uma aresta é representada pelo par não ordenado de vértices u, v pertencentes ao conjunto V e indica que u e v possuem uma relação. Por exemplo, se o grafo representa uma rede social de amizades, como o *Facebook*, u é amigo de v e v é amigo de u. Já em um grafo direcionado (Figura 2.2(b)) uma aresta é representada por um par ordenado de vértices u, v pertencentes ao conjunto V e indica uma relação direcional de u para v. Por exemplo, em uma rede social de seguidores de geradores de conteúdo, como o *Instagram*, uma aresta u, v indica que o usuário u segue as fotos do usuário v. Neste contexto, essa aresta não indica nenhuma relação entre o usuário v e o usuário u (DIGIAMPIETRI, 2015).



(a) - Grafo não direcionado com 4 vértices (b) Grafo direcionado com 4 vértices

Figura 2.2 - Exemplo dos tipos de grafos (Fonte: EASLEY e KLEINBERG, 2010)

Arestas que conectam vértices a si mesmo são chamadas de laço. Continuamente, é possível haver múltiplas arestas entre o mesmo par de vértices. Os grafos que possuem esta característica são chamados de multigrafos, enquanto os grafos que não possuem laços e nem arestas múltiplas são chamados de grafos simples (DIESTEL, 2010). Um grafo simples pode ser considerado grafo completo se existirem arestas ligando cada um dos pares de vértices deste grafo. Assim, um grafo completo não direcionado possui $(|V|.(V-1))/2$ arestas. Em continuação, os grafos podem ou não ter ponderações. Em grafos ponderados, cada aresta possui um peso associado. Este peso pode indicar, por exemplo, a intensidade da relação entre os dois vértices ou uma medida de distância ou custo entre eles.

Na maioria das vezes, em análise de redes, existe o foco no estudo de regiões específicas. Neste caso, vale destacar o conceito de subgrafos. Um subgrafo (Figura 2.3) é um grafo menor que pode ser obtido retirando arestas ou vértices de um grafo $G = (V, E)$. Outro conceito importante é o de clique. Um clique de um grafo é um subgrafo completo deste grafo. Sendo assim, é importante destacar que, uma vez definido o conceito de um grafo e sabendo que o mesmo é apropriado para o conceito de redes, a utilização do termo redes ou grafos se refere à mesma definição para o restante da tese.

Figura 2.3 - Exemplo de um grafo G e seus subgrafos G' e G'' (Fonte: DIESTEL, 2010)

Para entender melhor o comportamento de uma rede social, tanto do ponto de vista quantitativo quanto qualitativo, geralmente são utilizadas técnicas consolidadas baseadas na teoria dos grafos. Com a aplicação das técnicas, é possível caracterizar a rede, entender o fluxo de informação e a dinâmica dos vértices e das relações entre eles. Tendo em vista melhorar a compreensão e a padronização da nomenclatura que permeia a tese descrita, alguns conceitos fundamentais para as análises de redes são descritos a seguir, em conformidade com os trabalhos de Easley e Kleinberg (2010), Benevenuto (2010), Diestel (2010), Cunha (2013), Digiampietri (2015) e Dias (2016):

- **Grau e Grau Médio -** O grau de um nó corresponde ao número de arestas conectadas a ele, o grau médio é a média desses valores. A conexão dos vértices em uma rede é devidamente classificada a partir da distribuição das frequências dos graus.

- **Caminho mínimo ou Distância Geodésica -** O caminho mínimo entre dois vértices em um grafo não ponderado é o caminho que possui o menor tamanho (número de arestas) entre todos os caminhos que conectam os dois vértices. O tamanho desse caminho, para grafos não ponderados, é também chamado de distância geodésica.

- **Componente Gigante -** Um componente conexo de um grafo é um subgrafo no qual todos os nós estão conectados uns aos outros por caminhos. O componente conexo com maior quantidade de nós de um grafo é muitas vezes chamado de componente gigante. A porcentagem de nós no componente gigante é uma medida bastante utilizada na análise de redes. Pertencer ao componente gigante costuma ser associado a fazer parte do principal fluxo de conhecimento (ou informação) da rede. Assim, um valor elevado para esta medida costuma ser considerada uma característica positiva da rede. A Figura 2.4 exemplifica um grafo, destacando sua componente gigante.

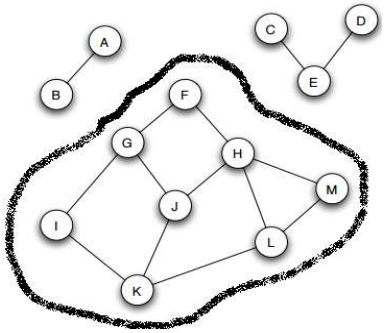


Figura 2.4 - Exemplo de um grafo G com três componentes. A componente em destaque representa a componente gigante. (Fonte: adaptado de EASLEY e KLEINBERG, 2010)

- **Diâmetro -** O diâmetro é definido como a maior distância geodésica contida no grafo. O diâmetro de um grafo pode variar de um mínimo de 1, caso o grafo seja completo, a um máximo de $V - 1$, onde V é o número de vértices no grafo. Em situações onde o grafo é não conectado, seu diâmetro refere-se ao maior entre os diâmetros dos componentes conexos.

- **Centralidade -** As medidas de centralidade visam identificar o quão central um vértice é na rede de acordo com alguma característica. Existem diversas medidas de centralidade, as mais utilizadas são: centralidade de grau, baseada no grau do vértice; centralidade de proximidade (*closeness*), baseada na distância entre o vértice analisado e os demais da rede; centralidade de intermediação (*betweeness*), baseada na frequência em que um dado vértice aparece entre todos os caminhos mínimos da rede; e, centralidade de autovalor (*eigenvalue*), baseada na importância dos vértices que estão ligados ao vértice em análise. As medidas de centralidade da rede são baseadas nas medidas de centralidade dos vértices da rede e servem para indicar o quão importante o vértice mais central de cada rede é para a sua rede.

- **Densidade -** A densidade corresponde à razão entre o número de arestas do grafo e o número máximo possível de arestas para o mesmo grafo. Em um grafo simples não-direcionado, a densidade é calculada como $2 \,.\, |E| \,/(|V| \,.\, (|V-1|))$.

- **Coeficiente de aglomeração ou Coeficiente de clusterização -** O coeficiente de clusterização de um vértice $u, cc(u)$ é a razão do número de arestas existentes entre os vizinhos de um vértice e o total de arestas possíveis entre os vizinhos de u. Como exemplo, a Figura 2.5 mostra o valor do coeficiente de clusterização para o vértice escuro em três cenários diferentes. No primeiro, todos os vizinhos do vértice estão conectados entre si e, consequentemente, o cc do vértice é um. No segundo cenário, existe apenas uma aresta entre os vizinhos do vértice dentre as três possíveis, deixando o vértice com $cc = 1/3$. No último cenário, não há nenhuma aresta entre os vizinhos do vértice escuro e, portanto, o cc do vértice é zero. Com isso, pode ser notado que o coeficiente de clusterização funciona como uma medida da densidade de arestas estabelecidas entre os vizinhos de um vértice. O coeficiente de clusterização de uma rede, cc, é a média do coeficiente de clusterização de todos os vértices.

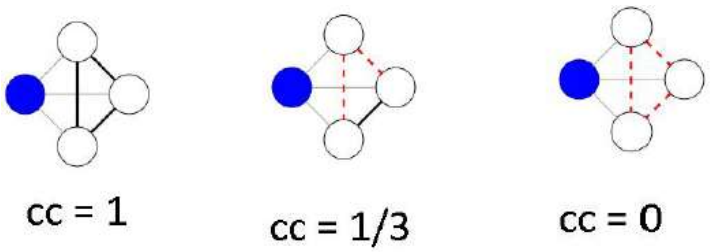


Figura 2.5 - Exemplo do cálculo do coeficiente de clusterização de um nodo em três cenários diferentes (BENEVENUTO, 2010)

Apesar de vários trabalhos terem sidos verificados acerca do contexto de análises de redes no passado, foi com o avanço dos computadores que essa área de pesquisa teve um relevante crescimento, já que tornou-se possível o estudo de redes referentes a grande volume de dados. Tais redes podem englobar milhares de nós com milhões de arestas, o que era impossível de se modelar alguns anos atrás, uma vez que os computadores possuíam recursos limitados se comparados com os atuais (DIAS, 2016). Entretanto, analisar grandes redes não é uma tarefa trivial. De acordo com Cunha (2013), é necessário entender todo o relacionamento entre as partes do sistema, caso contrário o pesquisador poderá colher resultados incompletos acerca do estudo. Logo, para uma análise holística mais confiável é necessário saber como as redes se formam e como seus elementos relacionam entre si.

Por fim, nesse sentido, diferentes modelos de redes são frequentemente apresentados na literatura, como, por exemplo: redes aleatórias (ERDOS e RENY, 1960), redes de mundo pequeno ou *Small-Word* (MILGRAN, 1967; WU e WATTS, 2002; BENEVENUTO, 2010), redes livres de escala ou *Scale Free / power-law* (BARABASI e ALBERTI, 1999; FALOUTSOS, et al., 1999), e as redes baseadas em cliques (CALDEIRA, 2005; CUNHA, 2013; FADIGAS e PEREIRA, 2013). No entanto, como o desenvolvimento desta tese concentra esforços exclusivamente em modelos de redes construídas a partir de cliques, apenas este é detalhado a seguir:

- **Redes de Cliques.** As redes de cliques são formadas por um conjunto de cliques unidas, parcialmente unidas ou não unidas. O elemento básico em uma rede desta natureza não é o vértice, mas sim um conjunto deles mutualmente conectados. Isso implica que os valores dos índices de densidade, coeficiente de aglomeração, caminho mínimo médio e diâmetro para uma clique terão valor igual a 1 (CUNHA, 2013). Portanto, ao unir cliques, constrói-se o que denomina-se rede de cliques. Segundo Fadigas e Pereira (2013), existem duas maneiras de se conectar duas cliques: justaposição ou sobreposição. O primeiro processo se dá pela união de duas cliques a partir de um único vértice comum. O segundo, pela junção de duas cliques com dois ou mais vértices em comum (Figura 2.6). Este modelo de rede possui diversas aplicações, como, por exemplo: redes de coautoria, redes de atores de filmes, redes de discursos orais, redes de títulos, entre outros (PEREIRA et al., 2011; CALDEIRA, 2005; MENEZES, 2012; CUNHA, 2013; SOUZA, 2015; TUESTA et al., 2015).

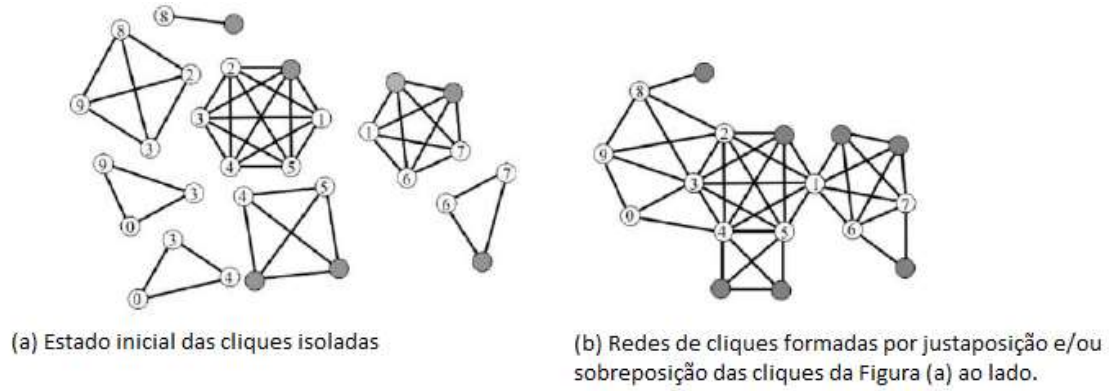


Figura 2.6 - Exemplo do estado inicial de cliques isoladas e uma configuração para redes de cliques. (FADIGAS e PEREIRA, 2013)

2.5 Análise de Tendências de Pesquisas

O termo análise de tendências é comumente utilizado por diferentes áreas do conhecimento em diferentes pontos de vista. Entretanto, não existe uma definição formal e amplamente aceita por essas diferentes áreas. A partir disto, para descrever esta seção, foram utilizados os conceitos explanados nos trabalhos de Trucolo e Digiampietri (2014a), Digiampietri (2015) e Trucolo (2016).

A análise de tendências em Sociologia geralmente observa mudança nos comportamentos de indivíduos ou grupos destes ao longo do tempo. Logo, tendências são padrões de comportamento social e estilo de vida que são identificados em determinados intervalos de tempo. Já para a área de exatas, a análise de tendências consiste em modelar séries temporais a partir de fonte de dados com intuito de entender o comportamento desses dados e, então, fornecer indicativos sobre valores futuros das variáveis dessas séries.

Apesar de serem pontos de vistas de áreas diferentes, a análise de tendência consiste em reunir dados para tentar determinar um padrão de comportamento de algum grupo de informações ao longo do tempo. Os grupos de informações para estudo que não possuem modificação relacionada a um fator de tempo não necessitam de utilização de técnicas de análises de tendências. Em contrapartida, existem diversos tipos de conjuntos de dados que possuem essa variação temporal, como, por exemplo, os artigos científicos produzidos por pesquisadores. Os tópicos de pesquisas abordados por esses documentos textuais científicos variam ao longo dos anos e podem representar a evolução do conhecimento construído pela comunidade científica.

Assim, no contexto de documentos textuais temporais, Kontostathis et al. (2004) definem uma tendência como um tópico que cresce de interesse ou utilidade ao longo do tempo. Nesse sentido, *Emerging Trending Detection (ETD)* é o modelo de aplicação de detecção de tendências que identifica tópicos novos ou que estão aumentando em importância ao longo do tempo dentro do corpo de documentos. Para este tipo de modelo, geralmente utiliza-se como fatores índices de importância baseados em frequência dos tópicos e o tempo. No entanto, a velocidade de aumento na popularidade de um determinado tópico pode definir se é uma tendência do tipo *bursty* ou emergente. De acordo com Decker et al. (2007), tendências do tipo *bursty* estão associadas a termos que tiveram aumento repentino num curto intervalo de tempo, já as tendências emergentes são associadas a termos que possuem crescimento, independentemente da taxa de popularidade nos últimos anos.

Contudo, muitas vezes a geração dos termos importantes a serem considerados durante a identificação de tendências num determinado conjunto de documentos necessita de esforços humanos, ou seja, de especialistas de domínio. Com o intuito de reduzir a intervenção humana dentro dos processos de identificação de tendências por ETD, são frequentemente utilizadas técnicas de *Topic Trending Detection (TDT)*, em que algoritmos buscam agrupar informações similares de documentos em tópicos para, em seguida, servir como entrada para as análises. Soluções TDT têm contribuído para implementação de modelos de ETD totalmente automáticos (KONTOSTATHIS et al. 2004). Neste ponto, no caso desta tese, vale destacar que o esforço dedicado para a identificação dos principais tópicos de pesquisa acerca de um conjunto de documentos científicos foi bastante reduzido, uma vez que são utilizadas as palavras-chave inseridas em artigos científicos como fonte para representação de tais tópicos em estudo.

Identificar tendências pode auxiliar na melhoria da qualidade de processos ou aumentar competitividade em negócios. Diversos trabalhos inseridos nessa perspectiva de análise de conteúdo identificam tendências para muitos tipos de aplicações, como, por exemplo: consultas realizadas na máquina de busca do *Google*, *twitter*, mercado de ações, notícias de um modo geral, e-mails, e, no ambiente acadêmico. Neste último, os resultados de análises de tendências sobre conhecimento científico podem fornecer insumos importantíssimos para decisores políticos elaborarem políticas científicas mais eficazes, servir de base para as agências de fomentos identificar o contexto da pesquisa proposta em um projeto, auxiliar pesquisadores a verificar o comportamento recente de tópico para definir suas pesquisas futuras, identificar autoridades em assuntos de determinada área de pesquisas para diferentes propósitos, dentre vários outros.

2.6 Índices para Medir Importância de Tópicos de Pesquisa

Uma vez identificados os tópicos que representam uma coleção de documentos textuais, geralmente uma medida de importância é calculada, associando pesos para estes tópicos. Logo, uma medida de importância mensura o poder discriminatório de um tópico num determinado documento ou em conjunto destes (SALTON e BUCKLEY, 1988; BORGES et al., 2015). Os pesos associados aos tópicos são utilizados como base tanto em processos de identificação de tendências quanto em ranqueamento de tópicos mais populares (ou "quentes" ou importantes).

De acordo com Abe e Tsumoto (2009), uma das medidas de importância mais utilizadas para aferir pesos de tópicos em coleções de documentos textuais é a frequência inversa do documento (TF-IDF). O fator TF_{ij} se refere a frequência simples de um tópico i e é calculada pela soma das ocorrências do referido tópico num determinado documento j. Neste ponto, vale destacar que frequência simples de um tópico, por si só, representa outra medida de importância de popularidade de um tópico amplamente utilizada em trabalhos de análise de tópicos científicos, dentre estes: Decker et al. (2007), Choi e Lee (2011), Zhu et al. (2013), Medeiros e Mena-Chalco (2013), Vinkers et al. (2015), Zhu et al. (2015), Ohniwa et al. (2010), Liu et al. (2012), Chen et al. (2015) e Chen e Xiao (2016).

Já o fator IDF do tópico i, frequência inversa do documento, é calculado como sendo o logaritmo do quociente do número de todos os documentos analisados (N) pelo número de documentos que contêm o respectivo tópico i (DF_i) , ou seja, $IDF_i = \log(N/DF_i)$. Esse fator é responsável por distinguir a homogeneidade ou heterogeneidade dos documentos por meio de seus tópicos. Com isso, o $TF - IDF_{ij} = \ TF_{ij} \ . \log(N/DF_i)$. Assim, um alto valor de TF-IDF significa que este tópico tem uma alta frequência em poucos documentos, mas raramente aparece em todos os documentos. Enquanto um valor baixo de TF-IDF significa que um tópico aparece em muitos documentos, mas com uma frequência similar em cada documento. Quanto maior o valor do TF-IDF do tópico, mais importante ele é para a coleção de documentos (SALTON e BUCKLEY, 1988). Como exemplo de utilização da medida TF-IDF no contexto científico, os trabalhos Trucolo e Digiampietri (2014b) e Trucolo e Digiampietri (2014c) podem ser considerados.

Além de tais medidas de importância serem aplicadas em análises de documentos científicos para ranqueamento de tópicos, estas são consolidadas em outros tipos de pesquisas, como recuperação de informação e agrupamento de documentos. Como a seleção do texto completo dos artigos científicos é uma tarefa complexa, as análises bibliométricas em tópicos de pesquisas são geralmente realizadas por meio de estudos de dados dos artigos científicos, a saber: palavras-chave, títulos ou resumos, que, se comparados com outros tipos de documentos (notícias, blogs, etc.), possuem características exclusivas. Além disso, durante o ranqueamento dos tópicos de pesquisas, os dados extraídos dos artigos são tratados igualmente sem ponderação apropriada (CHEN et al., 2015). Por fim, os conceitos descritos nesta seção e em seções anteriores deste capítulo, são suficientes para auxiliar no entendimento do desenvolvimento desta tese. No próximo capítulo, são apresentadas as principais referências que abordam os assuntos que permeiam esta tese.

TRABALHOS CORRELATOS

Este capítulo apresenta o resumo do conjunto de trabalhos relacionados a esta tese. Primeiramente, nas seções (3.1) e (3.2) são apresentados trabalhos que analisam conteúdo de textos com o intuito de compreender como os pesquisadores têm realizado suas pesquisas. Porém, a única diferença é que os trabalhos discutidos na Seção (3.1) não utilizam dados da Plataforma Lattes em seus estudos, diferentemente da Seção (3.2), onde todos os trabalhos apresentados têm como fonte principal de suas análises os dados dos currículos cadastrados na Plataforma Lattes. Por fim, na Seção (3.3), são feitas considerações acerca dos trabalhos encontrados, bem como os diferenciais propostos.

3.1 Trabalhos envolvendo dados de Publicações Científicas

Nesta seção, são apresentados trabalhos que de alguma forma têm analisado termos a partir de um conjunto de dados textuais (geralmente publicações científicas) para compreender como tem ocorrido a evolução científica ao longo do tempo, destacar os tópicos de interesse num determinado período e evidenciar quais são de investigação atual.

Em sua pesquisa de doutorado, Saes (2005) analisou a produção científica brasileira no campo da saúde. A pesquisa incluiu 3.066 revistas, correspondendo a um total de 38.349 artigos publicados entre 1990 e 2002 extraídos da *Web of Knowledge.* Os métodos bibliométricos e a co-ocorrência de termos foram os instrumentos utilizados para medir a atividade científica. Os descritores e identificadores dos artigos foram relacionados com outras variáveis, como, autores, ano de publicação e filiação da publicação. A partir daí, foi possível identificar redes de cooperação entre pesquisadores e instituições, bem como os descritores do campo do conhecimento estudado.

Ohniwa et al. (2010) propuseram um método simples para identificar tópicos emergentes a partir de termos utilizados para indexar artigos científicos no campo de ciências da vida. Para eles, um tópico emergente é definido por um grupo formado pela co-ocorrência de termos emergentes. Já os termos emergentes são termos que tiveram maior aumento na taxa de aparecimento. Para validação do método, utilizaram os termos *Mesh* associados aos 13.954.189 artigos publicados entre 1971 e 2008 do *PubMed*. Primeiramente, realizaram o cálculo da taxa de aparecimento de todos os termos *Mesh*. Para tanto, considerando o ano A de um respectivo termo *Mesh*, o cálculo da taxa de aparecimento deste termo é a soma das frequências de termos *Mesh* nos anos $A + 1$ *e* $A + 2$ dividido pela soma de sua frequência nos anos $A - 1, A$, $A + 1$ *e* $A + 2$. Assim, os autores definem os termos emergentes, como termos *Mesh,* que possuem o valor da taxa de aparecimento entre os 5% de todos os termos calculados num determinado ano. Em seguida, formaram os grupos dos termos emergentes por janelas de tempo e definiram os tópicos emergentes alinhado com os conceitos referentes aos termos emergentes de cada grupo. Por fim, demonstraram uma detalhada revisão acerca dos tópicos emergentes ao longo de 30 anos de pesquisas no campo de ciências da vida.

No trabalho de Liu et al. (2012), os autores analisaram 2.647 artigos e suas respectivas 9.538 palavras-chave das principais revistas cadastradas no repositório chinês *Full Text,* com intuito de mapear a estrutura intelectual acerca do campo de pesquisa "biblioteca digital", desenvolvido na China durante o período de 2002 a 2011. Inicialmente, foram elencadas as palavras-chave mais frequentes, onde pôde ser visto que as somas de 66 palavras-chave representavam 55% da soma total do conjunto analisado. Em seguida, os autores construíram uma matriz de co-ocorrência com base nas 66 palavras-chave mais frequentes, que serviu de insumo para a construção do agrupamento hierárquico e a rede de palavras-chaves. Por fim, após a aplicação de técnicas estatísticas e de redes sociais, foi possível elencar os principais tópicos estudados atualmente dentro do campo de pesquisa biblioteca digital na China.

Em Yi e Choi (2012), os autores ressaltaram que a identificação de padrões em redes de citação pode ajudar na compreensão da organização do desenvolvimento científico, no entanto, possui limitações (LEE et al. 2010; SU e LEE 2010). Diante disto, os autores propuseram uma alternativa por meio de análises de redes de palavras-chave de artigos científicos. Para tanto, utilizaram dados de publicações de 3 revistas científicas da área de Administração entre 1980 a 2009. Na construção da rede, duas palavras-chave estão vinculadas quando aparecem em uma mesma publicação em um artigo. Devido à grande quantidade de palavras-chave que podem tornar as redes a serem analisadas muito extensas e de difícil processamento, foram selecionadas apenas três áreas específicas para análise. Consequentemente, foram aplicadas métricas de análise de redes e características topológicas foram identificadas e comparadas com as redes de citações para o mesmo conjunto de dados. Concluíram que, ao contrário das redes de citações, as redes com as palavras-chave não possuem característica de mundo pequeno, seguem o conceito de Lei de Potência, ou seja, poucas palavras-chave possuem muitos vínculos e uma grande maioria possui uma quantidade baixa de vínculos. Os autores justificaram que os resultados são consistentes tendo em vista a viabilidade de representar o conhecimento científico de uma pesquisa com o elemento palavra-chave. A partir disto, sugeriram novos trabalhos, principalmente tendo como característica explorar a estrutura hierárquica que as redes de palavras-chave possuem.

Cunha et al. (2013) investigaram a evolução temporal de redes semânticas formadas a partir de termos extraídos dos títulos de artigos científicos publicados na Nature de 1999 a 2008, com intuito de verificar a existência de padrões de relacionamento em uma comunidade científica. Embasados pela Nature ter sua publicação semanal, os títulos foram agrupados por semana em 507 arquivos de texto para representar o período de estudo. Posteriormente, para a construção das janelas temporais do *Time Varying Graph (TVG),* método proposto por Casteigts et al. (2011) para estudar redes dinâmicas, estes arquivos foram agrupados em grupos de 8. Portanto, a primeira janela do TVG (t = 1) foi formada pelas publicações nos meses de janeiro e fevereiro de 1999. Isso se repete até a última janela, t = 507. Para cada janela do TVG, foram construídas as redes de cliques dos títulos. Em seguida, elas foram analisadas utilizando métricas clássicas de redes e métricas baseadas em cliques, propostos por Fadigas e Pereira (2013) e Santoro et al. (2011). Os resultados permitiram visualizar a dinâmica do adensamento das redes em épocas diferentes e constataram que apresentam o fenômeno de mundo pequeno. Os autores indicam que estudos futuros sobre frequência desses termos podem revelar algo sobre a capacidade de eles interagirem com grupos de temáticas diferentes, que podem indicar áreas do conhecimento humano.

Em Zhu et al. (2013), os autores modelaram uma rede de palavras-chave baseado num conjunto de 111.444 palavras-chave de artigos científicos do campo de Ciência da Informação, extraídos do repositório da *Scopus* no ano de 2008. Antes da construção da rede, todas as palavras-chave passaram por um processo de tratamento com o intuito de remover os caracteres que não possuem valores semânticos. Posteriormente à construção da rede, técnicas baseadas em análises de redes sociais foram aplicadas e identificado um efeito de mundo pequeno, mostrando que as palavras-chave estão próximas umas das outras. Além disto, foi identificado por meio de medidas de centralidade de grau, que algumas palavras possuem um número alto de vínculos com outras e que isto demonstra a sua importância na rede. Diante de tal constatação, os autores realizaram um estudo preliminar sobre como detectar os tópicos mais relevantes de uma linha de pesquisa. Este método também foi comparado com a estratégia de identificação por frequência de termos, justificando que a análise baseada em grau de centralidade tende a ser mais eficiente.

O trabalho de Trucolo e Digiampietri (2014a) realizou uma revisão sistemática com o intuito de investigar o estado da arte das técnicas de identificação e análise de tendências em aplicações de interação social em âmbito tecnológico. Eles destacaram várias técnicas utilizadas com objetivos diferentes, por isso, categorizaram-nas. As técnicas de identificação de tópicos buscam agrupar termos por algum tipo de similaridade; as técnicas de detecção de tendências são baseadas em frequência e conseguem identificar aumentos repentinos no aparecimento de tópicos em determinados períodos; e as técnicas de reconhecimento de padrões procuram aprender o comportamento dos tópicos e auxiliar na predição de tais comportamentos. Em continuação, os autores relataram que o domínio das aplicações influi na utilização de técnicas diferentes e que não existe um padrão de avaliação que facilite a comparação de resultados entre trabalhos, além da dificuldade na replicação dos experimentos devido à indisponibilidade do conjunto dos dados analisados. Ainda destacaram que poucos trabalhos utilizam a estrutura das fontes de informação como variável nos experimentos e, por isso, há um grande campo a ser explorado sobre a utilização da teoria de redes sociais na detecção e na análise de tendências.

Vinkers et al. (2015) estudaram os resumos e títulos de artigos científicos extraídos da PubMed publicados entre 1974 a 2014, com o objetivo de investigar as tendências de utilização de classes de termos positivos, negativos e neutros ao longo dos anos. Segundo eles, entender o comportamento da frequência de utilização desses termos pode gerar insumos para auxiliar na interpretação sobre o desenvolvimento da ciência. Para tanto, definiram conjuntos de termos, sendo: 25 selecionados criteriosamente por especialistas e 100 aleatoriamente para cada classe. Os termos foram quantificados individualmente por período de tempo e comparados com os livros publicados entre 1975 e 2009 cadastrados no *Google Books Ngram Viewer*. Os resultados mostraram que os resumos e os títulos estão cada vez mais sendo escritos com termos positivos, o que levou os autores a entenderem que o uso da linguagem positiva mais evidente está provavelmente relacionado com o surgimento de resultados positivos dominantes na literatura científica. Contudo, os autores listaram diversos fatores que poderiam estar contribuindo para o exagero dos pesquisadores na utilização de uma linguagem positiva perante seus resultados.

Em Pollack e Adler (2015), os autores realizaram análises bibliométricas com o objetivo de descobrir as tendências de pesquisas da área de Gerenciamento de Projetos a partir de 94.472 artigos cadastrados nos repositórios da *Scopus* e *Web of Science* publicados entre 1962 e 2012. Eles analisaram as palavras-chave e os termos extraídos dos resumos dos artigos científicos quanto à frequência, a variação quanto à utilização ao longo do tempo e a co-ocorrência. A partir disto, realizaram comparações entre a frequência dos termos e períodos onde existiu um rápido aumento de utilização para identificar os tópicos emergentes. Por fim, os autores concluíram que, dentre outras descobertas, os resultados apresentam indicativos sobre uma mudança de aplicação de técnicas de engenharia para uma perspectiva mais organizacional ao longo do tempo.

Khan e Wood (2015) analisaram as palavras-chave e os títulos de artigos referentes à área de Gestão da Tecnologia da Informação (GTI), com a proposta de mapear o conhecimento científico por meio de construção de redes e destacar os tópicos emergentes. As redes de termos extraídos dos títulos e palavras-chave foram construídas com base em aproximadamente 2000 termos dos 893 artigos publicados entre 1995 e 2014 referentes a 40 revistas, 64 países, 914 instituições e 1914 pesquisadores extraídos da *Web of Science*. A comparação entre as redes se deu com o resultado das métricas de número de componentes, diâmetro, densidade, coeficiente de clusterização, grau médio, intermediação e medidas de centralidade. Já para a identificação dos tópicos emergentes, foi utilizado o algoritmo *Burst Detection* (KLEINBERG, 2002), inicialmente definido para identificar tendências *bursty* num conjunto de notícias. Após as análises realizadas, os autores conseguiram avaliar o que tem sido produzido pela comunidade acadêmica quanto à área de GTI, evidenciaram os tópicos de pesquisas emergentes e os que estão sendo menos utilizados. Concluíram também que: (1) ambas as redes apresentam características de distribuição de lei de potência, ou seja, poucas palavras são frequentemente utilizadas e (2) redes de palavras-chave são mais apropriadas para representar os temas de pesquisas que redes de termos extraídos dos títulos. Contudo, apesar da validade dos resultados, sugerem limitações no estudo devido os dados analisados não serem suficientes para generalizar as conclusões acerca da área da GTI.

Zhu et al. (2015) analisaram as palavras-chave de 363.458 teses, publicadas entre 1986 e 2014, referente a indústria de petróleo e gás cadastradas no repositório da *China National Knowledge Internet* (CNKI), para entender o processo de desenvolvimento da área e evidenciar os assuntos atuais. Inicialmente, utilizaram a teoria de *Big Data* para extração dos dados e, em seguida, analisaram as frequências das palavras-chave individuais e co-ocorrências, bem como as redes construídas. Em continuação, os autores concluíram que foi possível identificar que aplicações de tecnologias em P&D (pesquisa e desenvolvimento) predominam na indústria de petróleo e gás, apresentando um grande número de pesquisadores interessados nos mais variados propósitos e um número menor de pesquisadores concentrados em propósitos similares.

Borges et al. (2015) realizaram uma análise exploratória dos tópicos de pesquisas emergentes na área de Informática na Educação com intuito de identificar os principais assuntos estudados. Como apoio à análise, os autores construíram a hierarquia de tópicos a partir de 4.053 artigos científicos publicados no período de 2011 a 2014 nas 10 principais conferências (5 nacionais e 5 internacionais) da área. O algoritmo de agrupamento hierárquico utilizado foi

o *K-means* (MACQUEEN, 1967), dado sua simplicidade computacional e os bons resultados presentes na literatura. Com isso, por meio de análises dos grupos hierárquicos foi possível verificar que a área de Informática na Educação possui forte vertente de pesquisas direcionadas por estudos de caso ou provas de conceitos. Além disso, os autores destacaram que a comunidade científica cada vez mais discute a inclusão e atualização de disciplinas para considerar a informática como apoio ao processo educacional, como, por exemplo, jogos educacionais.

Já no trabalho de Chen et al. (2015), os autores analisaram palavras-chave de artigos com o propósito de evidenciar o foco de pesquisa das universidades da China na área de Ciência da Informação. Para tal, propuseram e utilizaram o método KAI *(Keyword Activity Index)* para identificar os principais tópicos das 8 melhores universidades quanto a toda produção da área, que foi representada por um total de 277.721 palavras-chave referentes a 65.653 artigos publicados no período de 2000 a 2013 registrados no repositório da *Full Text.* O KAI foi baseado no método AI (*Activity Index*) proposto por Frame (1977), frequentemente utilizado na economia para verificar se um país possui vantagem em determinado campo de pesquisa se comparado com outros, por número de publicações. Em seguida, pela técnica de co-ocorrência e com o apoio de especialistas, foram construídos e comparados os clusters dos tópicos das publicações de cada instituição com maiores valores de KAI, calculados a partir das 200 palavras-chave mais frequentes de cada instituição. Porém, apesar da relevância da pesquisa, os autores sugerem novos estudos para definir critérios para seleção de palavras-chave a serem analisadas, pois é um problema comum na bibliometria. A partir disto, Chen e Xiao (2016) propuseram o método TF-KAI, baseado no método KAI citado, e uma variação do TF-IDF para a seleção de palavras-chave de artigos científicos. Esses métodos levam em consideração a frequência e a discriminação das palavras-chave dentro e fora do campo de estudo. O desempenho dos métodos propostos foi comparado com o método de frequência de termos (TF). Para tanto, especialistas analisaram listas de palavras-chave encontradas por cada método, no intuito de destacar as palavras-chave mais importantes para representar o campo Biblioteca Digital na China. Os resultados mostraram que o método TF-KAI foi o que obteve maior quantidade de palavras-chave correspondentes às escolhas dos especialistas, e que apresentaram de forma mais específica os temas estudados.

O trabalho de Mryglod et al. (2016) estudou a produção científica relacionada ao maior acidente nuclear da história, que ocorreu em 1986 na usina de Chernobil na Ucrânia, para apresentar a repercussão da comunidade científica nacional e internacional acerca do tema. Para isto, utilizaram análises bibliométricas, técnicas de análises de textos e de redes sociais nos 9.500 artigos científicos extraídos dos repositórios da *Socups* e *Ukrainika naukova*. Os autores mediram a distribuição dos trabalhos publicados ao longo do tempo, as taxas de crescimento e as propriedades de redes de co-autoria em diferentes campos de pesquisa e países. Além disso, analisaram os títulos e resumos das publicações para detecção dos termos mais importantes utilizados. Em seguida, concluíram que foi possível realizar a análise comparativa nacional, internacional e interdisciplinar do que se tem publicado acerca do acidente de Chernobil e também demonstrar que o respectivo assunto atrai o interesse de diversos pesquisadores, atualmente, em diferentes áreas do conhecimento com diferentes objetivos.

Numa pesquisa de doutorado, Fraga (2016) aplicou o método não supervisionado, proposto por Shibata et al. (2008), para detectar linhas de pesquisas emergentes em redes de citações de publicações científicas no campo de Bioenergia e de Empreendedorismo. Nestas redes, os vértices são representados pelos artigos e as arestas são as referências entre eles. As 93.250 publicações (64.009 de Bioenergia e 29.241 de Empreendedorismo) utilizadas para a construção das redes foram extraídas da *Web of Science* referentes ao período de 1980 a 2014. Durante as análises, foram encontrados 46 sub-grupos emergentes na linha de Bioenergia, enquanto que, na linha de Empreendedorismo, foi verificado que este não é um campo de pesquisa emergente, mas sim um campo em crescimento, onde os novos estudos são contribuições incrementais baseadas nos trabalhos pioneiros. Em seguida, os termos mais relevantes de cada grupo foram analisados utilizando o TF-IDF a partir dos resumos das publicações. Para finalizar, os resultados obtidos foram validados a partir de análise de especialistas, ou seja, conhecedores dos campos estudados.

Ronda-Pupo (2016) utilizou uma combinação de técnicas de co-ocorrência de termos e de redes sociais nos termos extraídos dos títulos e resumos de 2.264 artigos científicos cadastrados no repositório da *Web of Science,* com o objetivo de determinar os tópicos de pesquisas que compuseram a disciplina de Gestão na América Latina e no Caribe (ALC) nos últimos 25 anos. No intuito de estudar a evolução dos tópicos, o autor realizou análises por países em dois períodos de tempo, 1988 a 2000 e 2001 a 2013. Os resultados encontrados mostraram que as linhas de pesquisas Gestão Estratégica e Gestão de Inovação e Tecnologia formaram o desenvolvimento da disciplina de Gestão na ALC. As linhas necessárias para o desenvolvimento econômico e social da região, como Empreendedorismo, Estratégia Cooperativa e Gestão do Setor Público, aparecem como subdesenvolvidas. O autor também apresentou possíveis estratégias para o desenvolvimento futuro da disciplina de Gestão na ALC.

Em Silva et al. (2016), os autores propuseram uma metodologia baseada em técnicas de redes sociais e análises de textos para selecionar artigos mais relevantes de cada área, mapear a organização dessas áreas, suas comunidades e conexões. Primeiramente, os artigos mais relevantes foram destacados por meio da utilização de redes de citações, onde, cada artigo é tratado como um nó da rede e cada citação de um artigo por outro é considerada uma conexão. Em seguida, foram realizadas análises dos termos extraídos dos títulos e resumos das publicações com intuito de destacar os termos mais importantes para rotular as comunidades existentes. Para testar o método, foram utilizados artigos das áreas de Redes Complexas e Cristais Fotônicos cadastrados na *Web of Science* e os resultados foram avaliados por especialistas do domínio. Assim, foram identificadas duas comunidades na área de Cristais Fotônicos, sendo uma de engenheiros, voltados para telecomunicações, e outra de físicos e químicos, que fabricam os materiais. Também foi constatado que as comunidades são pouco conectadas entre si, o que significa que o conhecimento existente na área pode não estar sendo utilizado por pesquisadores da própria área, pelo fato de uma comunidade quase não saber o que se passa na outra. Os autores concluíram que as ferramentas automáticas são úteis para fornecer uma visão geral das áreas científicas e para auxiliar os cientistas na realização de pesquisas sistemáticas sobre um tópico específico.

Hong et al. (2016) realizaram a combinação de análises bibliométricas, co-ocorrência de termos e técnicas de análises de redes sociais, para identificar a estrutura do conhecimento e a evolução dos tópicos de pesquisas relacionados a médicos que atuaram como clínicos gerais no período de 1999 a 2014. Os dados para estudos foram coletados no repositório da *PubMed*, sendo 10.704 publicações e seus 42.695 termos *Mesh* associados. Estes dados foram agrupados e analisados de acordo com as respectivas faixas de tempo: 1999 a 2004, 2004 a 2008 e 2009 a 2014, para facilitar a verificação da evolução dos tópicos de pesquisas. Além dos resultados encontrados fornecerem uma visão geral clara do desenvolvimento de pesquisa acerca do assunto, também apontaram que temas relacionados a idosos serão amplamente discutidos no futuro, devido ao crescimento da população idosa. Entretanto, os autores citam duas limitações da pesquisa: (1) os dados extraídos do *PubMed* não podem representar em totalidade o desenvolvimento da pesquisa na área de médicos que atuam como clínicos gerais e (2) apenas os termos mais frequentes foram analisados, e assim, é possível que fique sem mostrar algum possível termo emergente.

Diversos outros trabalhos têm utilizado análises bibliométricas, co-ocorrência de termos, técnicas de redes sociais e análises de tendências, para extração de conhecimento acerca do que tem sido desenvolvido nas mais variadas áreas de pesquisas com propósitos distintos, como: taxas de carbono (ZHANG et al., 2016), produtos de *software* (HERADIO et al. 2016), criatividade (ZHANG et al., 2015), estrutura intelectual (RAVIKUMAR et al., 2015; CAVACINI, 2016); epidemiologia na Alemanha (PETER et al., 2016); saúde pública no Brasil (PEREIRA et al., 2007); biomedicina (MADLOCK-BROWN, 2014); palavras-chave mais representativas num *corpus* (LUCAS e LARA, 2014); educação em matemática (FADIGAS et al., 2009); atendimento eletrônico (SOUZA et al., 2014); sistemas embarcados (SOUZA et al., 2015); banco de dados (KAUER e MOREIRA, 2013). Contudo, apesar dos trabalhos apresentados nesta seção realizarem análises de conteúdo textuais, as primeiras abordagens encontradas na literatura para avaliar os interesses dos pesquisadores analisavam apenas dados quantitativos, como, número de pesquisadores, quantidade de citações e o número de artigos publicados (OHNIWA et al., 2010). Callon et al. (1983), Callon et al. (1986) e Leydesdorff (1995) foram uns dos autores que iniciaram estudos qualitativos (por exemplo, conteúdo dos artigos) associados aos dados quantitativos mencionados. Sendo assim, é possível compreender e classificar os principais tópicos de pesquisas num determinado período de tempo, dado a importância de tal compreensão para auxiliar nas tomadas de decisões estratégicas.

3.2 Trabalhos envolvendo Repositório de Dados da Plataforma Lattes

A Plataforma Lattes é um dos repositórios de dados de pesquisadores mais confiáveis existentes no mundo (LANE, 2010). Além disso, se constitui como uma fonte inesgotável de informações sobre a ciência brasileira (MUGNAINI et al. 2011). Como esta tese tem o objetivo principal de realizar um estudo acerca do que os pesquisadores brasileiros têm produzido em ciência nas grandes áreas do conhecimento ao longo do tempo, nesta seção são explanados e discutidos os trabalhos que mais se relacionam com a pesquisa em estudo, pelo simples fato de utilizarem a mesma fonte de dados, para, de certo modo, realizarem análises de conteúdo das publicações.

No trabalho de Granada et al. (2006), os autores apresentaram um método para identificar a similaridade entre pessoas que trabalham com o mesmo tema. O Coeficiente de Jaccard foi aplicado como medida de similaridade por comparações de análises entre as palavras-chave, títulos e textos completos das publicações científicas referentes a um conjunto de 10 currículos de professores da mesma escola de informática. Os resultados de similaridades obtidos foram: 100% palavras-chave, 90% títulos dos artigos e 60% textos dos artigos. Para os autores, uma possível explicação é o fato de que as palavras-chave são utilizadas para descrever a área de atuação ou de interesse. Os títulos e o texto dos artigos variam dependendo do enfoque a ser dado para o trabalho e tema em que foi publicado.

Medeiros e Mena-Chalco (2013) analisaram aproximadamente 670 mil currículos e mais de 5 milhões de publicações a fim de estudar a rede social dos indivíduos que atuam nas grandes-áreas: Ciências Humanas, Ciências Sociais Aplicadas ou Linguística, Letras e Artes. Os autores mediram em períodos de duração de 5 anos, de 1991 a 2010, o comportamento de alguns aspectos bibliométricos e métricas de redes sociais para o conjunto de dados, em que encontraram algumas características da dinâmica das redes formadas. Adicionalmente, realizaram uma análise de frequência dos termos extraídos dos títulos das publicações para identificar quais estão sendo mais utilizados por período de tempo, e que, possivelmente, são as mais importantes para estas áreas. Neste caso, os autores ressaltaram que a utilização de uma rede de termos para a visualização das interações referentes a grandes grupos é inviável devido ao grande número de nós e conexões que aparecem. Portanto, como apoio a esta análise, foram utilizados os conceitos de mapas de termos e conceitos de árvores de mapas para as 200 palavras mais frequentes de cada área por períodos de tempo estudados.

Em Digiampietri et al. (2014b), foi apresentada uma análise sobre a rede social formada a partir dos relacionamentos de indivíduos que atuam no Brasil em Inteligência Artificial (IA) ou Inteligência Computacional (IC). Para tanto, estudaram dados referentes a 2002 currículos que possuíam os termos “Inteligência Artificial” ou “Inteligência Computacional” no campo “Áreas de Atuação” e um endereço no Brasil no campo “Endereço Profissional”. Inicialmente, construíram 28 redes de colaboração científica, sendo uma rede nacional que contempla todos os pesquisadores e 27 redes compostas apenas por pesquisadores de cada um dos estados. Em seguida, realizaram a contagem dos termos mais usados nos títulos de artigos publicados em inglês. Na etapa de análises, três tipos de métricas foram aplicadas. O primeiro tipo é referente à análise de redes sociais e tem como intuito explicar características das redes construídas. O segundo tipo de métrica é derivado da área de mineração de textos, com a proposta de identificar os termos mais frequentes nos títulos dos artigos publicados e quantificar as respectivas frequências relativa considerando os estados e território nacional. Já o terceiro tipo analisa a frequência relativa dos termos nos artigos publicados entre 1993 e 2012 para ilustrar a evolução dos temas abordados ao longo de cada ano. De acordo com os autores, os resultados obtidos constituem-se como produção de conhecimento sobre a pesquisa nacional e têm potencial para apoiar tomadas de decisão por parte de pesquisadores e grupos de pesquisa, e para suportar o estabelecimento de diretrizes nacionais para a área. Como um dos trabalhos futuros, pretende-se realizar um estudo mais detalhado sobre a dinâmica temporal dos valores das métricas de análise de redes sociais a fim de explicar, com mais robustez, como a área se desenvolveu no decorrer dos anos e quais são as tendências observadas se o contexto de política nacional de pesquisa permanecer o mesmo.

Trucolo e Digiampietri (2014b) desenvolveram e aplicaram uma metodologia em 34.289 títulos de publicações científicas referentes ao período de 1991 e 2012 dos doutores que atuam no campo da Ciência da Informação, com intuito de identificar tendências de assuntos para curto, médio e longo prazo. A metodologia foi desenvolvida em três etapas: (1) obtenção dos dados dos currículos, (2) extração automática dos termos mais importantes inseridos nos títulos das publicações e (3) realização de regressões lineares e não lineares dos índices de importância TF-IDF dos termos extraídos. Assim, um termo foi considerado tendência se possuísse uma alta previsão da medida TF-IDF em relação aos demais. Já em Trucolo e Digiampietri (2014c), a mesma metodologia foi aplicada em 57.501 títulos dos artigos publicados entre 1911 e 2011 pelos professores permanentes da área de Ciência da Computação avaliados no triênio 2007-2009. Em ambos os trabalhos, os autores destacaram a relevância dos resultados, porém tais resultados consistem em um passo inicial considerando-se todo o potencial de análise de tendência da produção científica nacional. Em continuação, Trucolo (2016), em complemento à análise dos trabalhos anteriores, considerou a fonte geradora da informação e, com isso, utilizou também resultados de métricas de redes sociais para detecção de tendências nos mesmos dados estudados em Trucolo e Digiampietri (2014c). Diferentes técnicas de inteligência artificial foram utilizadas para realizar as previsões de tendências dos termos. Enfim, o autor observou que a inclusão de medidas oriundas da análise de redes sociais nos dados de treinamento permitiu uma predição com um erro bem menor do que o obtido utilizando-se apenas a análise temporal dos documentos.

Apesar de terem sido encontrados poucos trabalhos que analisam termos a partir das publicações cadastradas na base curricular da Plataforma Lattes, vários outros trabalhos com diferentes propósitos podem ser citados: (1) Identificação de especialistas em determinadas áreas (BRITO et al. (2016); CHAGAS et al., 2015); (2) Identificação de áreas de pesquisas (MIYATA et al., 2013); (3) Genealogia Científica (MOREIRA et al., 2016; DIGIAMPIETRI et al., 2014a); (4) Análises bibliométricas (CARDOSO e MACHADO, 2007; MUGNAINI et al., 2011; MUGNAINI et al., 2014; LIMA et al. 2014; DIAS, 2016); (5) Prospecção e organização de dados científicos (FERNANDES et al., 2011; MENA-CHALCO e CESAR-JUNIOR, 2009; MENA-CHALCO e CESAR-JUNIOR, 2013); (6) Caracterização e análises de redes sociais acadêmicas (MENA-CHALCO et al., 2012a; MENA-CHALCO et al., 2012b; PERES-CERVANTES et al., 2013; BOAVENTURA et al., 2014; MENA-CHALCO et al., 2014; ARAUJO et al., 2014; DIAS et al., 2013a; DIAS et al., 2013b; DIAS et al., 2014; DIAS et al., 2015; DIAS e MOITA, 2016).

Por fim, mesmo não utilizando análises de termos em documentos científicos, vale destacar o recente trabalho de Dias (2016), pois, diferentemente dos trabalhos encontrados na literatura, apresenta um estudo abrangente sobre a produção científica brasileira. Para isso, o autor desenvolveu o *LattesDataXplorer,* um arcabouço responsável pela extração de todo o conjunto de dados dos currículos. Neste arcabouço, também foram implementadas técnicas para análises bibliométricas e métricas baseadas em análises de redes sociais para estudo das redes de colaboração científica. Assim, foi possível caracterizar todo o repositório de currículos da Plataforma Lattes e, ainda, apresentar uma visão detalhada sobre a produção científica brasileira. Segundo o autor, os resultados encontrados possuem caráter inédito tendo em vista a abrangência das análises realizadas e pela grande quantidade de indivíduos considerados. Como um dos trabalhos futuros, sugere o estudo do conteúdo das publicações para entender quais tópicos foram e são estudados pela ciência brasileira.

3.3 Considerações sobre Trabalhos Correlatos e Diferenciais Propostos

Na literatura revisada, verificou-se um grande esforço por parte dos pesquisadores de todas as áreas e nações em compreender o conhecimento científico e como ele tem evoluído por meio de estudos sobre dados das publicações. Dentre estes, a maioria dos trabalhos visam realizar análises sobre conjuntos restritos de dados, que são originados de repositórios internacionais específicos (determinada área ou periódico) e com uma quantidade limitada de registros. Por outro lado, dos trabalhos que utilizam fontes de dados nacionais, mais especificamente dados da Plataforma Lattes, pouquíssimos e recentes foram os casos que analisaram conteúdo das publicações e, quando o fizeram, também utilizaram uma quantidade restrita dos dados disponíveis em análises preliminares. Portanto, em se tratando de pesquisas que de algum modo analisam dados científicos referentes ao Brasil, tanto extraídos de repositórios internacionais quanto nacionais, não é possível englobar, em totalidade, o que é produzido no país.

Além disso, do ponto de vista do conteúdo analisado das publicações, foi verificado que a estratégia de analisar termos extraídos de títulos e/ou resumos dos artigos é o foco principal de grande parte dos estudos encontrados na literatura (no caso da Plataforma Lattes, são praticamente todos). Embora esta estratégia contribua na compreensão da evolução científica, claramente possui suas limitações, uma vez que nem sempre os títulos conseguem expressar todo o conteúdo abordado por um determinado trabalho, tendo em vista a necessidade do autor da pesquisa se preocupar com a estrutura semântica em sua descrição. Já quanto aos resumos dos artigos, a maior dificuldade está na extração automática dos termos que realmente representam o assunto central do trabalho. Com isso, outra estratégia que vem ganhando força na literatura é a análise das palavras-chave que os autores inserem em suas publicações, pois as mesmas foram selecionadas cuidadosamente por eles (MCCLOSKEY, 1988) com o foco de evidenciar os tópicos centrais que permeiam o trabalho (YI e CHOI, 2012) sem se preocupar com questões semânticas.

Em continuação, os trabalhos geralmente têm o foco principal em detectar tópicos de pesquisas populares ou tendências (classificadas em emergente ou *bursty*), a partir de análises de frequência de termos, análises de taxas de crescimento de utilização dos termos, análises de termos utilizando técnicas de clusterização e análise de mapas do conhecimento por redes de termos. Nesses estudos, existem duas abordagens diferentes que dependem do objetivo da pesquisa: (1) usam uma quantidade mais abrangente de termos para explorar as características estruturais do campo de estudo em um nível mais amplo e (2) usam alguns termos “mais importantes” com o intuito de analisar os principais tópicos de investigação e suas relações com maiores detalhes (CHEN e XIAO, 2016). Contudo, menos atenção tem sido dada à medida de importância para ranqueamento desses termos a serem analisados e interpretados, e geralmente são considerados os termos mais populares com base em sua frequência ou medidas de redes baseadas em centralidade, ambas comprovadamente correlatas, de acordo com o estudo de Choi et al. (2011). No entanto, a importância de um termo no contexto científico não deveria ser igualmente tratada sem ponderação apropriada, devido às características particulares dos dados analisados das publicações (CHEN e XIAO, 2016). Isso é mais importante ainda em fontes heterogêneas de dados científicos, como no

caso da Plataforma Lattes, que congrega dados de periódicos e conferências das mais variadas áreas de pesquisas com diferentes níveis de qualidade. Assim, medidas de importância dos termos deveriam levar em consideração características adicionais apropriadas ao contexto em análise, como, por exemplo, impacto da fonte de publicação dos artigos que contêm o termo.

A grande maioria dos trabalhos que identificam tópicos populares e analisam tendências têm como foco final extrair conhecimento de um conjunto de dados textuais para apoiar algum tipo de tomada de decisão, utilizando técnicas de análises de documentos gerais, ao invés de propor novas técnicas que contemplem as características específicas dos documentos científicos. As poucas pesquisas que propõem novas técnicas não realizam validações que facilitam a comparação de resultados entre os trabalhos (TSENG et al., 2009; TRUCOLO e DIGIAMPIETRI, 2014a), principalmente, pela inexistência de métricas globais que permitam tal avaliação (MADLOCK-BROWN, 2014; CHEN e XIAO, 2016), pela dificuldade na replicação dos experimentos devido à indisponibilidade dos dados utilizados em pesquisas anteriores e pelos diferentes contextos de aplicações (TRUCOLO e DIGIAMPIETRI, 2014a; TSENG et al. 2009). Alternativamente, alguns trabalhos utilizam especialistas de domínio para avaliar os resultados encontrados por suas técnicas e, dependendo do retorno destes, consideram-nas como relevantes (TSENG et al., 2009; CHEN e XIAO, 2016).

Apesar dos trabalhos apresentados neste Capítulo utilizarem abordagens baseadas em técnicas computacionais e estatísticas, estudos baseados em análises de especialistas ainda são amplamente utilizados, principalmente para auxílio na construção de novas políticas científicas. Porém, métodos baseados na análise de especialistas são caros, lentos e tendenciosos, além de não conseguir levar em consideração todas as características existentes em grande volume de dados (MADLOCK-BROWN, 2014; FRAGA, 2016). As técnicas computacionais apresentam vantagem da escalabilidade de análise, adequação ao volume de dados e demandam menos esforços para conseguir resultados objetivos (FRAGA, 2016). Ainda assim, tais trabalhos explanados neste Capítulo apresentam resultados significantes para a literatura científica. Apesar de tal relevância, não foram encontrados estudos prévios sobre uma análise temporal dos tópicos desenvolvidos por pesquisadores brasileiros ao longo da história da ciência. Consequentemente, o presente estudo apresenta-se como a primeira análise textual abrangente sobre o desenvolvimento científico brasileiro ao longo dos 55 anos de história registrada nos currículos dos doutores cadastrados na Plataforma Lattes. Além disso, é também a primeira análise sobre as palavras-chave que os pesquisadores inseriram em seus artigos cadastrados em seus currículos. O conteúdo desta tese está inserido no contexto de processamento, caracterização, modelagem e análise de um grande volume de dados científicos, utilizando métodos computacionais e estatísticos. Para isso, é apresentado a seguir, no Capítulo *Desenvolvimento,* um arcabouço de componentes que foi desenvolvido com a responsabilidade de filtrar e tratar os dados, que em capítulos posteriores são analisados por meio de análises bibliométricas e técnicas de redes sociais.

MATERIAIS E MÉTODOS

Neste capítulo, é descrito o arcabouço de componentes que suporta o desenvolvimento desta tese. Na Seção (4.1) é apresentado o procedimento necessário para a aquisição de todos os currículos XML cadastrados na Plataforma Lattes; a Seção (4.2) descreve o processo realizado para a seleção, filtragem e tratamento dos dados a serem analisados; e, para encerrar o capítulo, a Seção (4.3) apresenta a estratégia utilizada para manipular e processar os dados que são apresentados e discutidos no capítulo a seguir.

4.1 Aquisição dos Dados

De forma resumida ao que foi detalhado em capítulos anteriores, 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 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).

Como citado na Seção (2.2), os currículos cadastrados na Plataforma Lattes são disponibilizados em dois formatos: HTML (*HyperText Markup Language*) e XML (*eXtensible Markup Language*), sendo a versão XML a mais adequada para o processamento automático, pois possui todas as seções e campos bem delimitados. Além disso, a versão XML possui os dados referentes às palavras-chave cadastradas em publicações científicas, principal objeto de estudo desta pesquisa, que não são disponibilizadas na versão HTML. Portanto, para compor o repositório de dados a ser analisado, apenas a versão XML de todos os currículos cadastrados na Plataforma Lattes foi utilizada. A aquisição destes currículos é realizada pelo LattesDataXplorer, proposto por Dias (2016) para coletar e tratar os dados contidos nos currículos cadastrados na Plataforma Lattes. A seguir, na Figura 4.1, é apresentada uma visão geral do processo de extração de todos os currículos XML cadastrados na Plataforma Lattes pela ferramenta LattesDataXplorer.

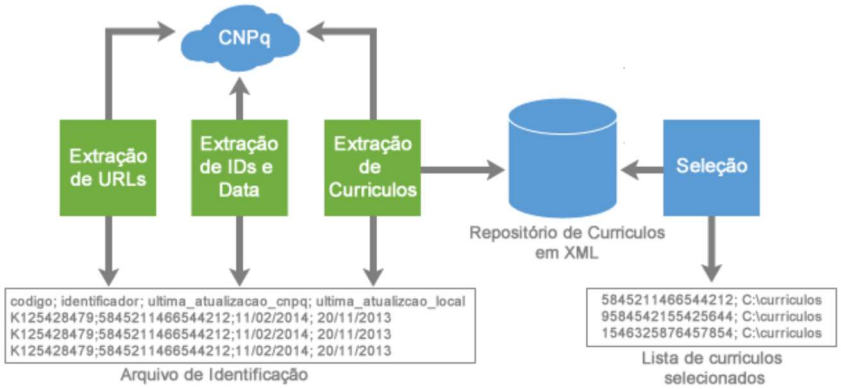


Figura 4.1 - Processo de extração de dados LattesDataXplorer (Fonte: Dias, 2016).

O processo de extração do LattesDataXplorer para coleta dos currículos da Plataforma Lattes foi realizado em três etapas: (1) extração de URLs, responsável por adquirir as referências únicas para todos os currículos cadastrados e, assim, possibilitar o acesso individual a cada currículo, (2) extração de Ids e Data, para possibilitar o acesso a cada currículo e extrair seu identificador, bem como a data de última utilização, e (3) extração dos currículos, responsável por extrair e armazenar os currículos cuja data de atualização na Plataforma Lattes seja divergente da data de atualização do currículo armazenado localmente (DIAS, 2016).

4.2 Filtragem e Tratamento dos Dados

Conforme explanado na Seção (2.2), os currículos XML da Plataforma Lattes permitem o cadastramento de vários tipos de informações referentes às contribuições dos pesquisadores para o desenvolvimento da ciência. Apesar disto, nesta tese, optou-se por utilizar apenas informações de artigos científicos publicados em anais de congressos e em periódicos, com o intuito de assim representar a produção científica brasileira. A justificativa para estudo dos artigos científicos em detrimento de outras informações disponíveis nos currículos se dá pelo fato dos artigos serem considerados o principal caminho de disseminação de novos conhecimentos na maioria das disciplinas científicas (RONDA-PUPO, 2015).

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 nesta tese.

Logo, a Figura 4.2 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 os responsáveis por todo o processo de seleção, tratamento e caracterização das informações dos currículos que realmente necessitam ser processadas para atingir os objetivos propostos, e, concomitantemente, diminuir o tempo de processamento computacional.

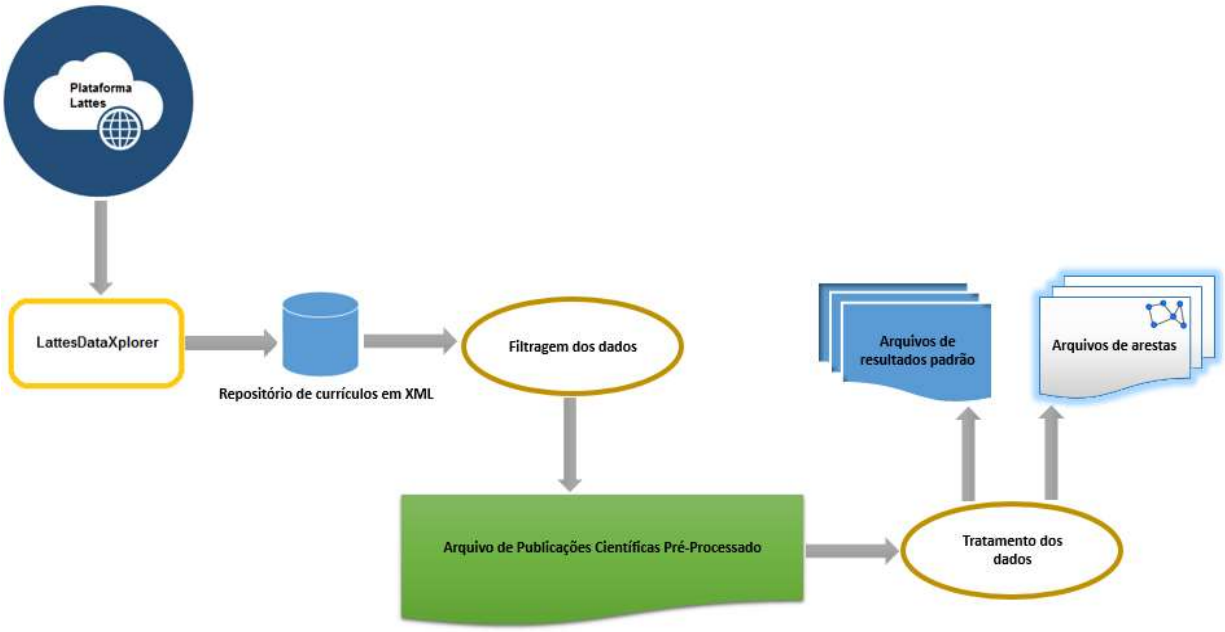


Figura 4.2 - Visão geral do arcabouço de componentes que suporta as análises de palavras-chave dos artigos.

Os componentes de "filtragem dos dados" e "tratamento dos dados" são ilustrados com maiores detalhes na Figura 4.3.

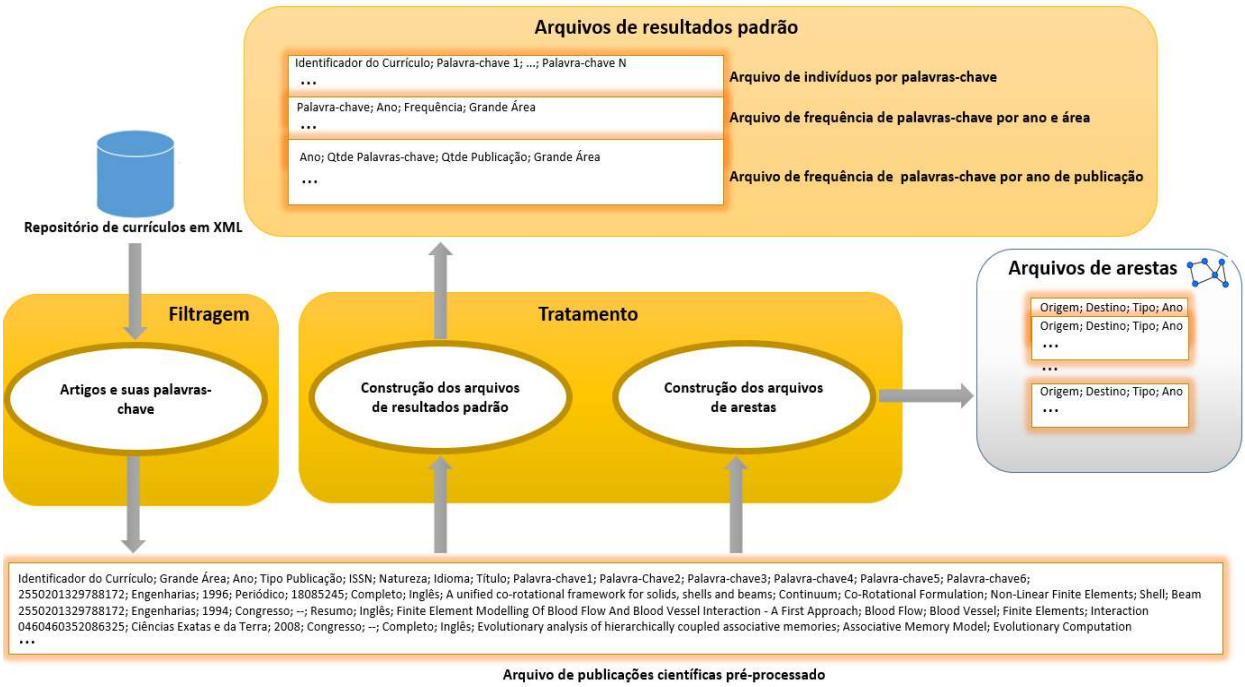


Figura 4.3 - Processo de filtragem e tratamento dos dados.

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.

Limpeza e Agrupamento dos dados

Tendo em vista que o cadastramento das palavras-chave de um artigo científico em seus currículos é de inteira responsabilidade dos respectivos pesquisadores, e isso é feito livremente por eles, significa que podem ser inseridos quaisquer conjuntos de caracteres como uma palavra-chave. A partir disto, geralmente tem-se uma coleção muito grande de palavras-chave e sem nenhum padrão. Na tentativa de contornar este problema, foi desenvolvido um método que realiza o processamento das palavras-chave com o objetivo de excluir possíveis termos com ruídos ou que não representam um tópico de pesquisa. Além disso, a aplicação do método faz com que seja possível agrupar, em um dicionário, as palavras-chave que foram escritas de formas distintas, mas que possuem mesmo valor semântico.

O método inicia-se obtendo a quantidade de palavras-chave extraídas de cada artigo. Diante disso, cada uma das palavras-chave passa por um processo de detecção de idioma, que serve de referência durante o processo de radicalização (*stemming*). Neste caso, o processo de detecção de idioma associa à palavra-chave o idioma que foi inserido no respectivo artigo cadastrado pelo próprio pesquisador em seu currículo. Em continuação, no processo de *lowercase*, todas as palavras são convertidas para minúsculo com a proposta de padronizar o conjunto.

Já no processo de remoção de s*topWords,* são removidos todos os termos que não possuem valores semânticos significativos para caracterizar um tópico de pesquisa e, com isso, diminuir o volume de palavras-chave a serem processadas e analisadas. Em seguida, cada uma das palavras-chave passa por um processo de normalização que visa extrair as letras acentuadas e substituí-las pelo seu equivalente sem acentuação.

E, por último, existe o processo de radicalização, que consiste na redução das palavras-chave a seu radical com base no idioma do artigo recuperado durante o processo de detecção de idioma. O processo de radicalização é importante para evitar a inclusão de palavras-chave com o mesmo significado de formas distintas, possibilitando que o conjunto seja reduzido 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 4.1 apresenta um exemplo de transformação de uma palavra-chave de artigo científico após a execução da etapa de limpeza dos dados.

Tabela 4.1: Exemplo de transformação de uma palavra extraída de um artigo científico.

Etapa Algoritmo	Resultado
Recebimento da Palavra-Chave	Gerência de Dados
Detecção 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

As atividades desta etapa são utilizadas durante a execução das etapas de normalização e na construção dos arquivos, para compor o processamento necessário durante a geração dos arquivos de resultados padrão e arquivos de arestas, objetivo principal deste capítulo.

Normalização dos dados

Boaventura et al. (2014) destacam que o maior desafio no tratamento dos dados de currículos da Plataforma Lattes resulta do fato de que as informações são preenchidas manualmente por seus respectivos donos. Assim, não é uma situação incomum dois pesquisadores cadastrarem um artigo científico que publicaram juntos utilizando diferentes informações, como título e palavras-chave. Em contrapartida, em determinadas análises, o preenchimento igualitário das informações gera duplicidade de dados de artigos quando considerado todo o repositório de publicações, uma vez que o cálculo de popularidade com base na frequência de uma palavra-chave poderia ser falseado pelo número de co-autores do trabalho, impactando diretamente nos resultados finais. Por exemplo, a Tabela 4.2 exemplifica uma situação que um mesmo artigo aparece três vezes quando são considerados todas as publicações de um conjunto de currículos. Isso ocorre pois tal artigo foi publicado em coautoria. Neste caso, a popularidade de cada palavra-chave que aparece igualmente em todo o conjunto destas palavras deveria ser 1, pois está se tratando do mesmo artigo publicado. No entanto, para as palavras “Bibliometria” e “Redes Sociais” a popularidade é 3.

Tabela 4.2: Exemplo de um artigo publicado em co-autoria com informações iguais em seus currículos.

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
....	....	....	....	....

Sendo assim, a etapa de normalização dos dados tem o objetivo de analisar todos os artigos considerados nesta tese com o intuito de eliminar as inconsistências citadas anteriormente. Para tanto, fez-se necessário uma adaptação no método ISCooll que foi proposto por Dias e Moita (2015) para identificar colaborações científicas automaticamente em um grande volume de dados com custo linear, tendo em vista os diversos problemas envolvidos, como duplicatas nos nomes dos autores e divergência em títulos referentes a uma mesma publicação quando cadastradas por coautores. Logo, 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 processadas a partir da etapa de limpeza e agrupamento dos dados. Outra diferença está na formação da chave do dicionário; enquanto o método ISColl original utiliza somente o título do artigo concatenado com seu ano de publicação, o método ISCooll adaptado adiciona também a grande área da publicação (ou caso haja inexistência desta, adiciona a primeira grande área de atuação informada pelo pesquisador em seu currículo). Assim, a inserção da grande área para a formação da chave do dicionário no método ISCooll adaptado se dá pela decisão de considerar a contribuição do artigo publicado em coautoria, em casos que os pesquisadores inseriram diferentes grandes áreas para a mesma publicação. A Figura 4.4 apresenta um exemplo de resultado dos métodos ISCooll original e adaptado aplicado a um conjunto de artigos publicados em coautoria por pesquisadores que inseriram diferentes grande área nos mesmos.

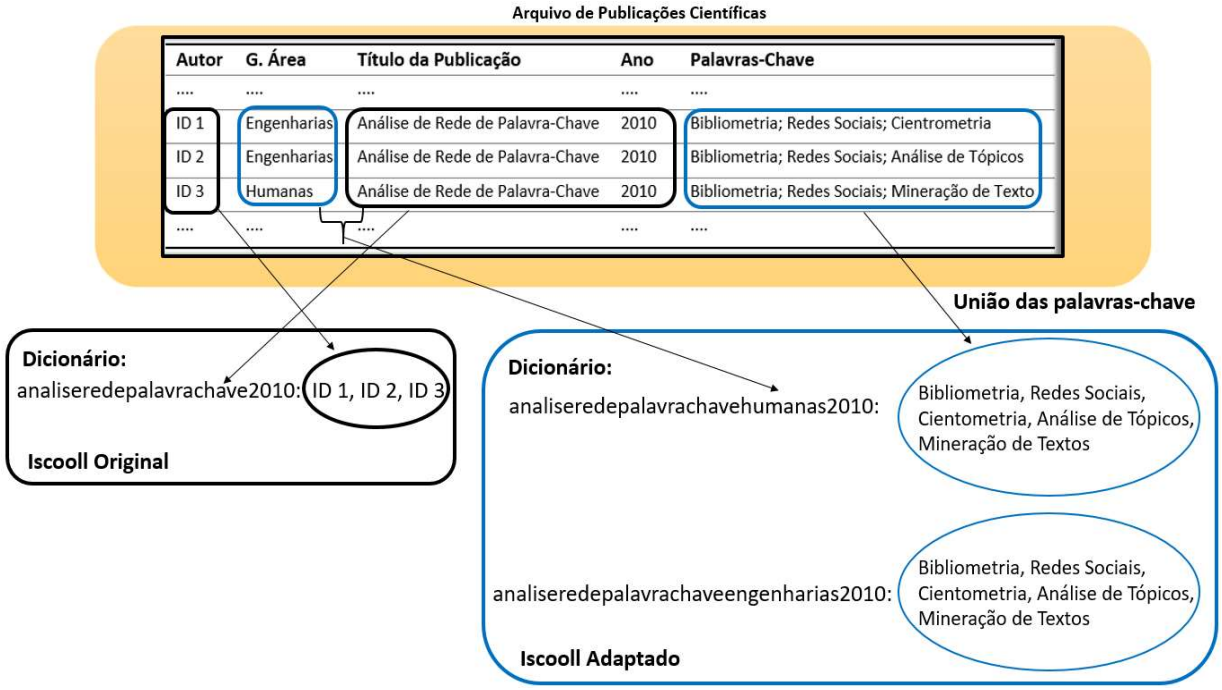


Figura 4.4 - Exemplo de processamento dos métodos *ISCooll* original e adaptado.

Portanto, ao término do processamento de todos os artigos publicados em anais de congressos e em periódicos cadastrados na base curricular da Plataforma Lattes pelo método ISCooll adaptado, o dicionário contemplará: (1) em casos de publicações em coautoria, onde os autores informaram a mesma grande área do conhecimento, será considerado um único artigo acompanhado com conjunto união das palavras-chave associadas por eles, e (2) em casos de coautoria onde os autores informaram grandes áreas do conhecimento distintas, um número de artigos será de acordo com a quantidade de diferentes grandes áreas do conhecimento informadas pelos seus coautores. A partir disto, é possível também executar a próxima etapa, construção de arquivos, utilizando os dados de publicações realizados em coautoria (processados por esta etapa), ou seja, sem considerar um mesmo artigo duplicado se considerado todo o conjunto.

Construção dos arquivos

A estratégia da construção dos arquivos tem como intuito facilitar a extração de conhecimento durante as análises e contribuir com a diminuição do volume de dados a serem processados. Nesta etapa, foram construídos dois conjuntos de arquivos para cada tipo de publicação, a saber: arquivos de resultados padrão e arquivos de arestas. Para uma análise mais abrangente, os arquivos de resultados padrão foram divididos em: arquivos que consideram a coautoria dos artigos (processados pelas etapas de limpeza e normalização) e os que contém todo o conjunto de artigos (processados apenas pela etapa de limpeza).

O conjunto de arquivos de resultados padrão forma o repositório principal que possibilita a manipulação e as análises estatísticas sobre os tópicos de pesquisas desenvolvidos por pesquisadores brasileiros em suas grandes áreas do conhecimento. Inicialmente, foram construídos três arquivos de seguintes formatos: (1) **Arquivo de indivíduos por palavras-chave** (Identificador do currículo, palavra-chave1, ... palavra-chaveN); (2) **Arquivo de palavras-chave por ano e grande área** (Palavra-chave, ano, frequência e grande área) e (3) **Arquivo de frequência de palavras-chave por ano e quantidade de publicação** (Ano, frequência palavra-chave, quantidade de publicação e grande área).

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, conforme descrito na Seção (2.5.2), em que 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 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 4.5 exemplifica a construção da rede de palavras-chave a partir de dois artigos que foram processados pelas etapas de limpeza e normalização dos dados.

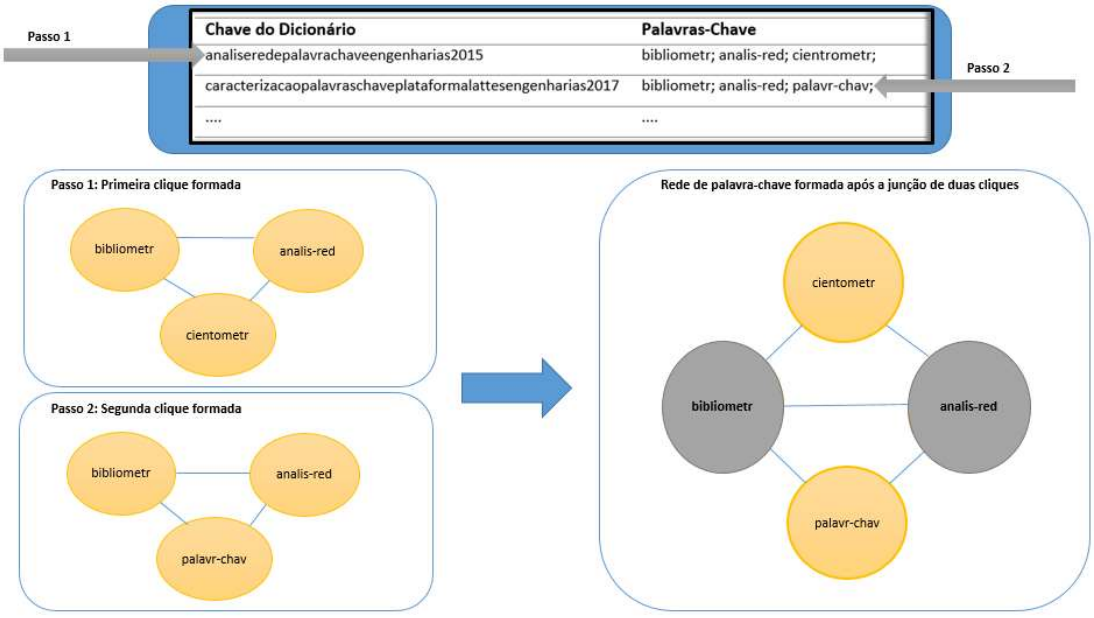


Figura 4.5 - Exemplo do processo de construção de redes de palavras-chave considerando dois artigos.

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 nesta tese, 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. Assim, redes de pesquisadores vinculados por palavras utilizadas em comum podem indicar proximidade entre pesquisadores, permitindo agrupá-los segundo os temas com que trabalham.

Por fim, vale destacar que a construção dos componentes se deu na linguagem de programação *Python* com a utilização das funcionalidades disponíveis na biblioteca *Natural Language ToolKit (NLTK),* criada pelo departamento de Ciência da Computação e Ciência da Informação da Universidade da Pennsylvania para processamento de linguagem natural (BIRD et al., 2009). Neste caso, a *NLTK* contribuiu com o conjunto de termos referentes às *stopWords* e a implementação do algoritmo *Porter Stemmer* para a radicalização de língua inglesa e *Snowball Stemmer* para a língua portuguesa.

4.3 Manipulação e Análises dos Dados

Como descrito na Seção (2.1), as palavras-chave de um artigo científico possuem o objetivo de melhor descrever o conteúdo de um trabalho a ser publicado e seu estudo pode ser aplicado a diversos propósitos. No contexto desta tese, entender quais são as palavras-chave em destaque de uma grande coleção de artigos científicos cadastrados na base curricular da Plataforma Lattes pode representar o conhecimento científico brasileiro, ou seja, indicar influência de determinados tópicos de pesquisa que são mais relevantes, identificação de tópicos isolados, descoberta de tópicos abordados por áreas distintas, auxílio em processos de revisão bibliográfica, definição do corpo do conhecimento em determinada área ou assunto, verificação de similaridades de interesses entre pesquisadores, indicação de possíveis tendências de pesquisas, dentre outras.

Assim, após uma revisão bibliográfica sobre o assunto, foram selecionadas análises bibliométricas, técnicas para análises de redes sociais e análises de tendências a serem testadas e combinadas para tal estudo, visando um entendimento abrangente sobre a evolução da ciência brasileira nas diferentes grandes áreas do conhecimento. No próximo capítulo são apresentados os resultados de tais análises e técnicas utilizadas por meio de uma caracterização geral de todas as palavras-chave de artigos científicos publicados em anais de congressos e em periódicos cadastrados na base curricular da Plataforma Lattes. Vale destacar que estes resultados foram gerados a partir da manipulação e processamento dos arquivos de resultados padrão e arquivos de arestas (Seção 4.2) utilizando a linguagem de programação Python e suas bibliotecas *Pandas (para manipulação e análise de dados)* e *NetworkX (para analisar redes e grafos)*.

CARACTERIZAÇÃO DOS DADOS

Este Capítulo apresenta uma caracterização geral dos dados científicos dos doutores que possuem currículos cadastrados na Plataforma Lattes. Assim, são apresentadas estatísticas gerais dos dados extraídos (Seção 5.1), estatísticas por grandes áreas do conhecimento (Seção 5.2) e as redes de palavras-chave (Seção 5.3). Os resultados exibidos neste Capítulo podem impulsionar diversos outros trabalhos nas mais diferentes áreas de pesquisa.

5.1 Estatísticas Gerais

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. Esses dados foram coletados, filtrados, tratados e modelados conforme descrito no Capítulo de "Desenvolvimento". Para esta caracterização foram considerados os artigos de anais de congressos e periódicos referentes ao período de 1962 até 2016, representando 55 anos de registros de dados acadêmicos da ciência brasileira na base curricular da Plataforma Lattes. Diante disto, espera-se que os artigos publicados em 2016 já estejam registrados nos currículos dos doutores, uma vez que tais artigos são normalmente cadastrados por eles após sua data de publicação. Entretanto, frequentemente são encontrados currículos que estão há algum tempo sem atualização. A Figura 5.1 apresenta a distribuição dos currículos dos doutores com base na última data de atualização.

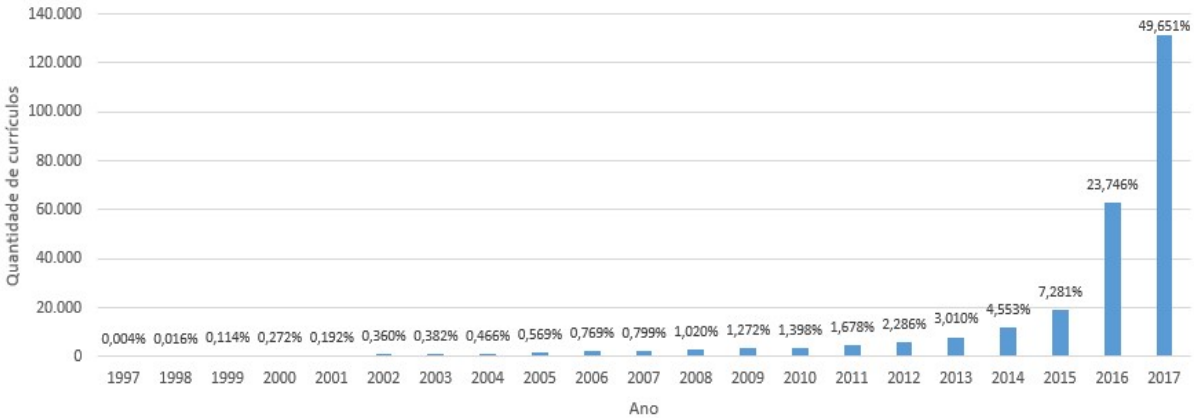


Figura 5.1 - Distribuição dos currículos dos doutores pela data da última atualização.

Diante disso, observa-se que apesar de existirem 10 currículos com data da última atualização em 1997, essa quantidade é irrelevante quando comparada ao conjunto total de currículos atualizados em anos posteriores. De acordo com Dias (2016), a não atualização desses currículos pode ter motivos variados e de difícil reconhecimento. Contudo, a grande maioria dos currículos foram atualizados recentemente, em que 49,813% (132.090) dos currículos possuem data de última atualização em 2017 e 73,559% (195.056) foram atualizados nos últimos dois anos.

É importante destacar a diversidade de produções científicas que estão registradas neste conjunto de currículos. Neste contexto, a Tabela 5.1 apresenta o quantitativo de palavras-chave dos principais tipos de produção científica entre 1962 e 2016 registrados pelos doutores brasileiros em seus currículos.

Tabela 5.1: Quantitativo dos dados por produção científica e suas palavras-chave entre 1962 e 2016.

Tipo de Produção Científica	Produção científica	Palavras-chave vinculadas
Anais de Congresso	**8.881.237 (46,88%)**	**14.288.895 (45,64%)**
Periódicos	**4.685.011 (24,73%)**	**8.892.490 (28,40%)**
Livros	394.521 (2,08%)	762.364 (2,43%)
Capítulos de Livros	1.046.780 (5,53%)	1.849.985 (5,91%)
Textos em Jornais ou Revistas	854.199 (4,51%)	1.002.335 (3,20%)
Trabalhos Técnicos	1.765.493 (9,32%)	1.742.449 (5,57%)
Patentes	36.987 (0,20%)	50.015 (0,16%)
Orientações de Mestrado	942.717 (4,98%)	1.979.823 (6,32%)
Orientações de Doutorado	297.406 (1,57%)	670.466 (2,14%)
Orientações de Pós-Doutorado	40.487 (0,21%)	71.024 (0,23%)
Total	**18.944.838 (100%)**	**31.309.846 (100%)**

Como é possível observar, a soma dos artigos publicados em anais de congressos e em periódicos representam mais de 70% do número total de produção científica e palavras-chave, reforçando assim, a justificativa para esta tese utilizar somente os dados dos artigos para representar a produção científica brasileira, conforme descrito na Seção (4.2).

Cada indivíduo possui um único currículo cadastrado no repositório da Plataforma Lattes, não existindo a possibilidade de dados duplicados sobre a produção deste. Por exemplo, se considerado todo o repositório curricular da Plataforma Lattes, os dados de um artigo produzido em colaboração podem aparecer várias vezes, tendo em vista que ele pode ser registrado por cada um de seus autores. Neste ponto, com o propósito de customizar as análises para torná-las mais abrangentes e coesas ao contexto, foi necessária a consideração de duas visões sobre os dados extraídos, a saber: (1) a **visão geral**, que contabiliza todos os artigos e palavras-chave extraídos dos currículos e, (2) a **visão colaboração**, como apresentado na etapa de normalização dos dados na Seção (4.2), que considera apenas um único artigo publicado em coautoria de mesma grande área do conhecimento associado ao conjunto união de suas respectivas palavras-chave. Sendo assim, a Tabela 5.2 apresenta o quantitativo de todos os artigos e palavras-chave extraídos dos currículos dos doutores, considerando as visões geral e de colaboração.

Tabela 5.2: Quantitativo de artigos científicos e suas palavras-chave dos currículos dos 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%)**

Os dados apresentados correspondem a toda produção de artigos publicados em anais de congressos e em periódicos existentes nos currículos extraídos. A quantidade de artigos registrados nos currículos dos doutores representa cerca de 70% do número total de artigos contidos no repositório da Plataforma Lattes (DIAS, 2016), que, aliado à diversidade e à contemporaneidade de atualização destes, pondera a validade do conjunto para os estudos iniciais desta pesquisa. Portanto, é possível destacar uma considerável preferência dos doutores sobre a publicação de artigos em congressos tanto na visão geral quanto em colaboração, e consequentemente, o mesmo vale quanto ao número de palavras-chave. Como esperado, o número total de artigos na visão de colaboração é cerca de 25% menor que o total da visão geral, devido ao fato de os artigos terem sidos registrados em coautoria. Entretanto, o número total de palavras-chave vinculadas aos artigos publicados em anais de congressos e em periódicos da visão colaboração é maior que o número de palavras-chave dos respectivos tipos de artigos da visão geral, fato este, que se justifica devido aos autores cadastrarem diferentes palavras-chave e/ou grandes áreas do conhecimento para um mesmo artigo realizado em coautoria.

Outra situação também esperada é quanto ao número de palavras-chave ser maior que o número de artigos, visto que cada artigo pode ter até 6 palavras-chave associadas durante seu cadastramento na Plataforma Lattes. Porém, no caso da Plataforma Lattes, não há garantias que, para todo conjunto analisado, a relação entre o número de artigos e o número de palavras-chave seja bem definida, pois não existe obrigatoriedade quanto à inserção destas palavras, fazendo com que a responsabilidade dos dados inseridos seja dos próprios possuidores dos currículos. Isso pode ser melhor analisado pela Figura 5.2, construída a partir dos dados da visão de colaboração, para ilustrar os quantitativos dos artigos e suas palavras-chave ao longo dos anos.

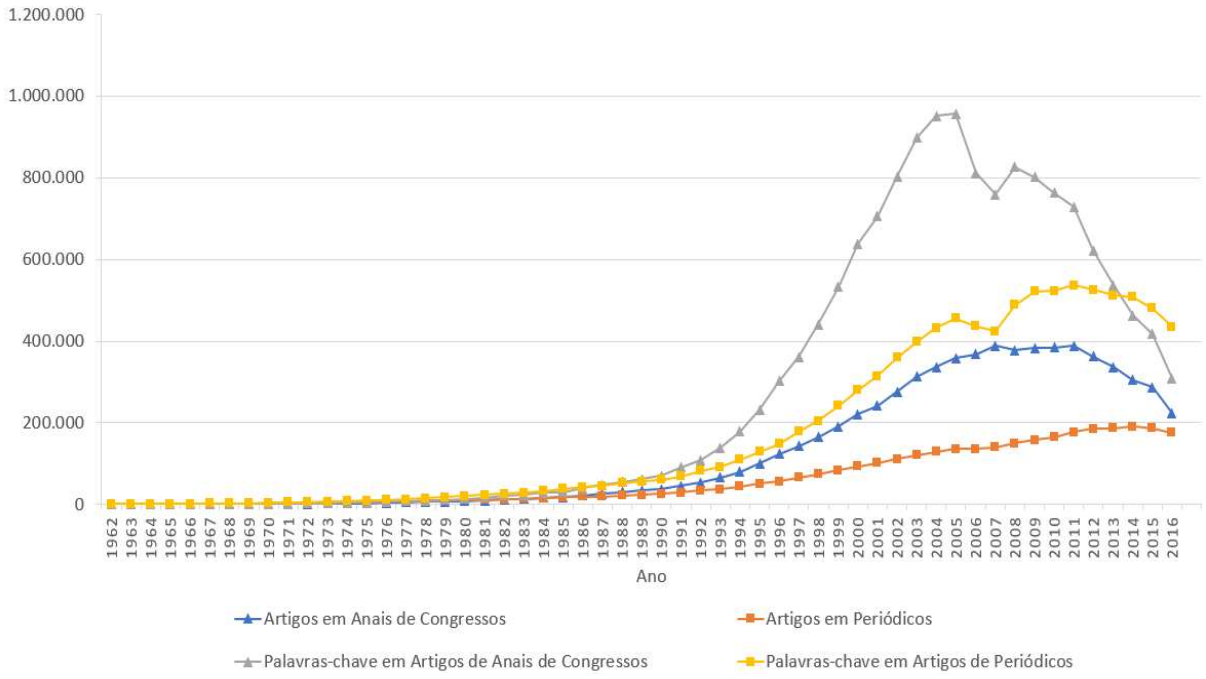


Figura 5.2 - Quantitativo de artigos e suas palavras-chave por ano.

Por meio dessa figura é possível perceber que, para a maioria dos anos, as curvas das palavras-chave de anais de congressos e periódicos apresentam o mesmo comportamento das respectivas curvas dos artigos, tanto em crescimento quanto em queda. As exceções de comportamentos em anais de congressos se deram: (1) no ano de 2008, os doutores registraram uma queda no número de artigos e tiveram um aumento de palavras-chave e; (2) nos anos de 1963, 2006, 2007, 2009, 2010 e 2011, o número de palavras-chave teve queda enquanto o número de artigos aumentou. Em relação aos dados de periódicos, os anos de 2007, 2012, 2013 e 2014 apresentaram crescimento no número de artigos e queda nos números palavras-chave. Outra situação interessante é que o maior número de palavras-chave cadastrado não ocorreu no mesmo ano do maior número de artigos publicados. Em artigos de anais de congressos, o maior número de palavras-chave foi registrado em 2005 (957.122) e o maior número de artigos em 2011 (388.369). No caso de artigos em periódicos, foi registrado em 2011 (537.159) o maior número de palavras-chave e em 2014 (191.045) seu maior número de artigos. Percebe-se ainda que o número de palavras-chave teve um crescimento considerável entre 1995 e 2005, quando palavras-chave em artigos de anais de congressos cresceram 443% e em artigos de periódicos 197%. Apesar disso, a partir de 2005 o número de palavras-chave em artigos de anais de congressos teve queda de aproximadamente 67%, enquanto a queda em artigos de periódicos foi de 18,9% em relação a 2011. Essa queda acentuada do número de palavras-chave em artigos de anais de congressos nos últimos anos possivelmente demonstra um maior interesse dos doutores quanto a periódicos, visto que a partir de 2014 o número de palavras-chave de artigos em periódicos ultrapassou o número de palavras-chave de artigos em anais de congressos.

A Figura 5.3, construída com base na visão geral, apresenta a média das palavras-chave por artigos em anais de congressos e em periódicos, para possibilitar uma análise mais detalhada acerca da relação entre a quantidade de palavras-chave associadas aos artigos publicados ao longo dos anos.

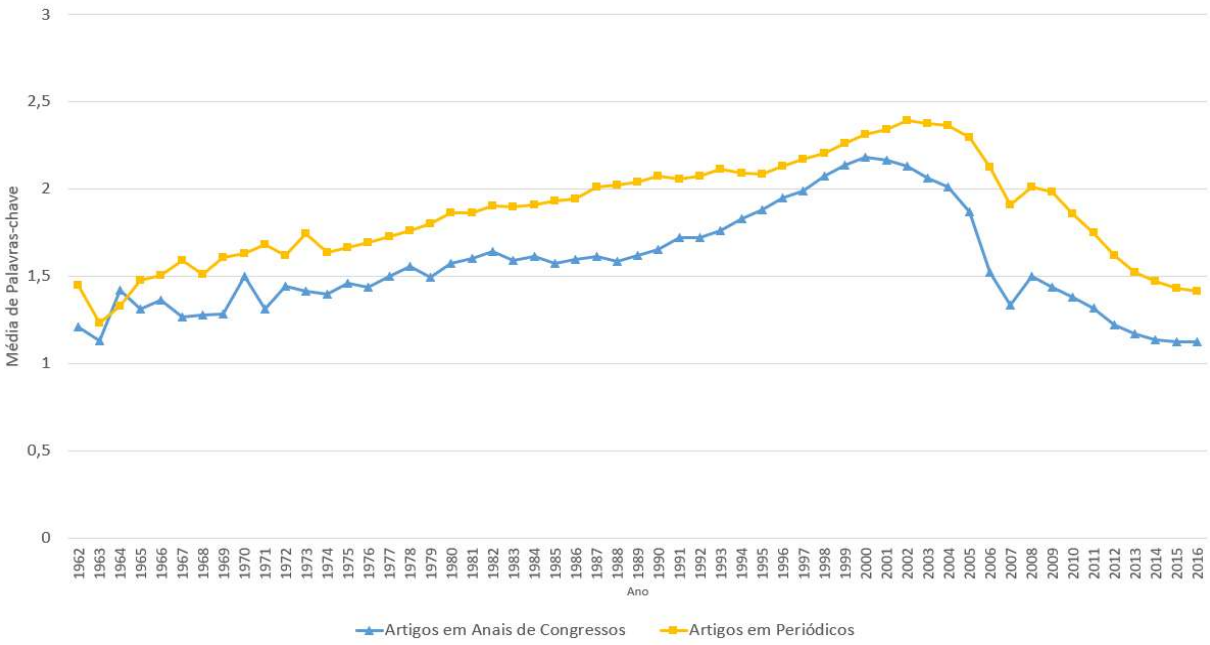


Figura 5.3 - Média de palavras-chave por artigo publicado.

Assim, é possível perceber que a queda de palavras-chave no período entre 2004 e 2009 foi possivelmente impulsionada pela queda no número de publicações. Além disso, em praticamente todo o período analisado, a média de palavras-chave por artigo publicado em periódico é superior à média em anais de congressos. Neste ponto, destaca-se a queda de palavras-chave no período entre 2004 e 2009 possivelmente impulsionada pela queda no número de publicações. Contudo, 2,5 palavras-chave por artigo em periódico é a maior média encontrada. Se comparada à capacidade máxima de palavras-chave por artigo cadastrado na Plataforma Lattes, esse valor é baixo. Neste caso, os baixos valores médios se devem pela influência significativa por parte dos artigos que possuem poucas (ou nenhuma) palavra(s)-chave associada(s). Para facilitar a visualização desta informação, foram gerados os histogramas referentes à distribuição do número de palavras-chave por quantidade de artigos a partir das visões geral e colaboração respectivamente (ver Figura 5.4 e 5.5).

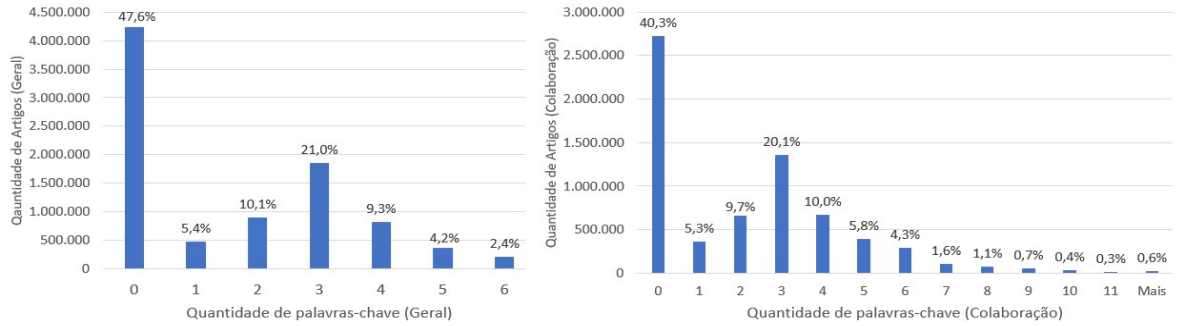


Figura 5.4 - Distribuição dos artigos em anais de congresso pelo número de palavras-chave.

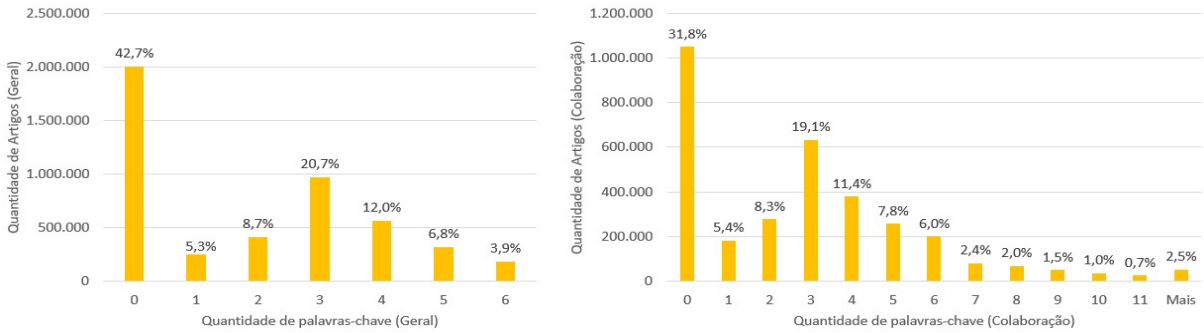


Figura 5.5 - Distribuição dos artigos em periódicos pelo número de palavras-chave.

Logo, fica explícito que grande parte dos artigos (em média 45% visão geral e 36% colaboração) em anais de congressos e em periódicos não possuem sequer uma palavra-chave associada. Além disso, cerca de 15% dos artigos têm entre 1 e 2 palavras-chave. Adicionalmente, pelos histogramas baseados na visão colaboração, é possível comprovar que os doutores que realizaram trabalhos em coautoria inseriram diferentes palavras-chave para este mesmo artigo, sendo que aparecem números de palavras-chave maiores que 6 (capacidade limite de palavras-chave por artigo aceita pela Plataforma Lattes). Diante disso, fica evidente o quanto os artigos sem palavras-chave que os doutores inserem em seus currículos têm impacto neste tipo de análise. Com base nisso, dado o foco deste trabalho, para as análises a partir deste ponto, optou-se por desconsiderar todos os artigos sem palavras-chave associadas. Assim, conforme apresentado na Figura 5.6, construída com base na visão geral, as médias de palavras-chave por artigo se alteram significativamente.

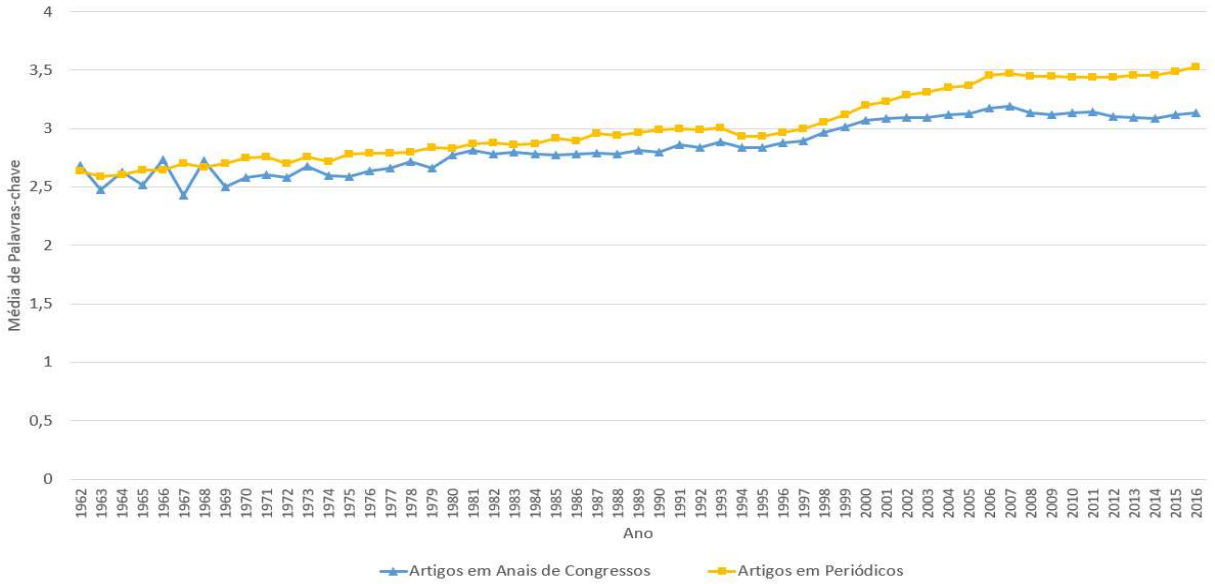


Figura 5.6 - Média de palavras-chave por artigo publicado.

Igualmente ao analisado na Figura 5.3, a média de palavras-chave por artigo em periódico é superior à média de palavras-chave por artigo em anais de congressos para praticamente todo o período analisado, o que sugere que os doutores dão mais atenção ao cadastramento de artigos em periódicos. Contudo, sem a influência dos artigos que não contêm palavras-chave, as médias de palavras-chave por artigo são superiores e giram entre 2,4 e 3,5. Outra informação interessante é que, a partir de 1999, a média de palavras-chave por artigo sempre foi maior que 3 em anais de congressos e em periódicos, destacando-se as médias em periódicos, que para o mesmo período, cresceram cerca de 8% em relação a anais de congresso. Uma possível justificativa para estes valores pode estar relacionada às exigências dos veículos de publicação, que geralmente solicitam o cadastramento de 3 palavras-chave por artigo.

A grande quantidade de palavras-chave cadastradas na base curricular da Plataforma Lattes pode, por exemplo, servir de base para verificar o desenvolvimento da ciência brasileira, uma vez que palavras-chave têm a função de destacar os principais assuntos que permeiam o artigo e são escolhidas pelos próprios autores. Apesar disto, conforme apresentado no Capítulo de "Trabalhos Correlatos", a estratégia de analisar termos extraídos de títulos dos artigos é o foco principal da maioria dos estudos encontrados, principalmente os trabalhos que utilizam o repositório de dados da Plataforma Lattes (CUNHA et al., 2013; DIGIAMPIETRI et al., 2014b; TRUCOLO e DIGIAMPIETRI, 2014b; TRUCOLO e DIGIAMPIETRI, 2014c; VINKERS et al., 2015; SOUZA et al., 2015; MRYGLOD et al., 2016; RONDA-PUPO, 2016; SILVA et al., 2016). Embora esta estratégia tenha utilidade, notadamente possui suas limitações, pois nem sempre os títulos conseguem expressar todo o conteúdo de um trabalho, tendo em vista a necessidade do autor da pesquisa se preocupar com a estrutura semântica em sua descrição, conforme já discutido anteriormente.

Com base nisso e de forma inédita, a Tabela 5.3, construída a partir da visão geral e utilizando apenas o conjunto de artigos que possuem palavras-chave, possibilita verificar o percentual das palavras-chave associadas aos artigos que estão contidas em seus respectivos títulos.

Tabela 5.3: Número de palavras-chave contidas nos títulos das publicações.

Artigos Publicados	Artigos 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%)

Como pode ser observado, os resultados apresentados pela estratégia de analisar palavras-chave são bem diferentes dos que podem ser mostrados por meio das análises de títulos dos artigos, pois, no caso dos doutores que possuem currículos cadastrados na Plataforma Lattes, apenas 33,3% das palavras-chave de artigos em anais de congressos e 28,6% em periódicos estão contidos nos títulos de seus respectivos artigos publicados. Para maiores detalhes desta análise, as Figuras 5.7 e 5.8 apresentam a distribuição dos artigos pelo número de palavras-chave existentes em seus títulos.

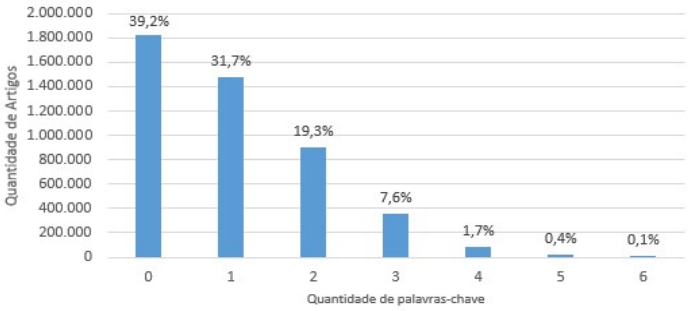


Figura 5.7 - Distribuição dos artigos em congressos pelo número de palavras-chave existentes nos títulos.

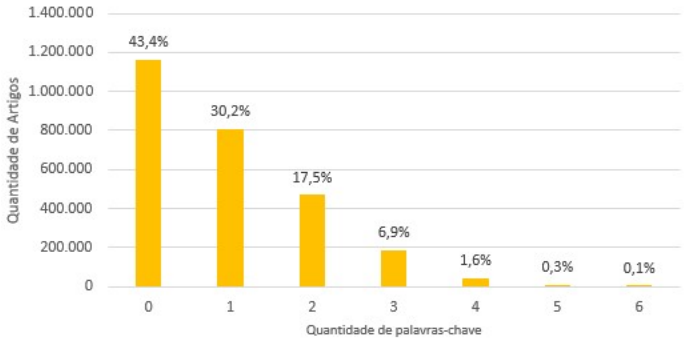


Figura 5.8 - Distribuição dos artigos de periódicos pelo número de palavras-chave existentes nos títulos.

É possível verificar um padrão de valores em ambas as figuras (5.7 e 5.8). A medida que a quantidade de palavras-chave aumenta, o percentual de artigos que possuem suas palavras-chave presentes nos títulos diminui aceleradamente. Dentre estes, destaca-se o percentual elevado (cerca de 40%) dos artigos que não possuem nenhuma de suas palavras-chave contidas em seus títulos. Esta informação é de extrema importância, pois, diversos outros estudos que utilizam a estratégia de analisar os títulos das publicações podem não considerar de forma efetiva o conteúdo dos trabalhos em análise. Com base nisso e, aliado às importantes características das palavras-chave dos artigos científicos, fundamenta-se ainda mais o desenvolvimento da atual pesquisa quanto à sua originalidade.

Uma informação muito importante para a compreensão da organização do conjunto das palavras-chave em análise diz respeito ao quantitativo de palavras por frequência de utilização. Assim, com base na visão colaboração, a Figura 5.9 apresenta a distribuição da quantidade de palavras-chave por frequência. Além disso, mostra também o quantitativo e o percentual de palavras-chave únicas, ou seja, palavras que foram processadas pela etapa de limpeza e agrupamento dos dados (Seção 4.2).

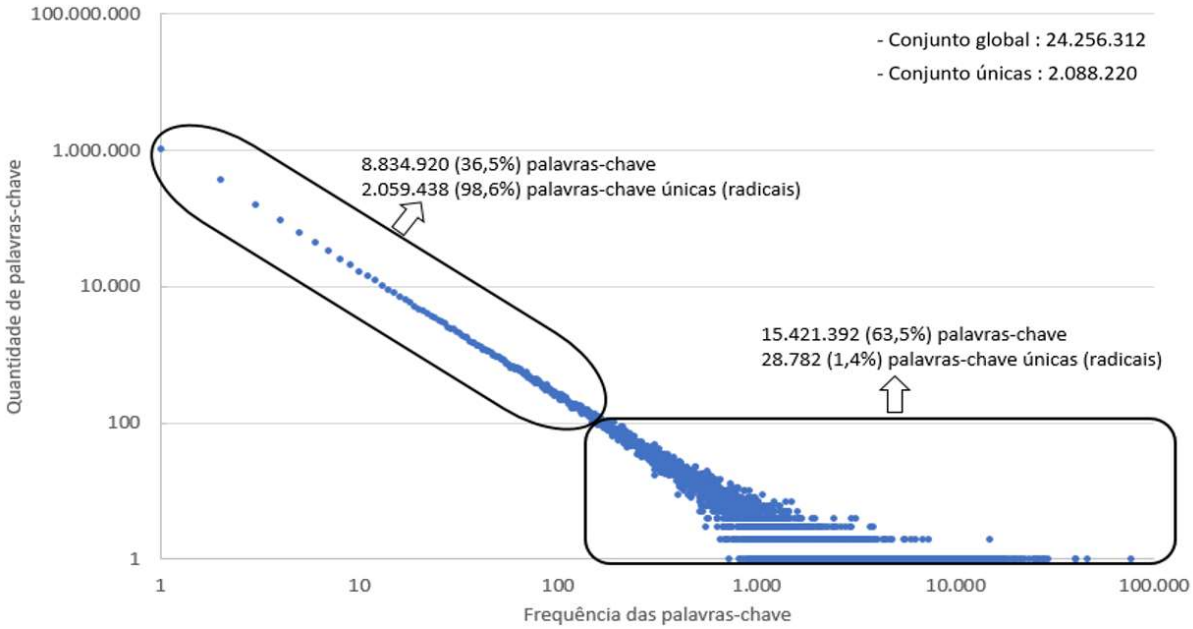


Figura 5.9 - Distribuição de quantidade de palavras-chave por sua frequência de utilização.

Observa-se que 24.256.312 palavras-chave (conjunto global), após processadas, se reduziram a 2.088.220 palavras-chave únicas, comprovando a importância do tratamento dos dados para diminuição de informações a serem analisadas. Também pode ser notado que 63,5% do conjunto global das palavras-chave equivale a 1,4% do total de palavras-chave únicas que possuem frequência superior a 100, enquanto que, os 36,5% restantes equivale a 98,6% das palavras-chave únicas que foram utilizadas menos de 100 vezes. Dentre estes, destaca-se que uma única palavra-chave foi utilizada 75.855 vezes e 1.059.962 diferentes palavras aparecem apenas uma vez cada. Essa informação mostra que os doutores brasileiros têm concentrado mais esforços acerca de um conjunto específico de temas ao longo do tempo. Confirmando uma das leis clássicas da Bibliometria, *Lei de Zipf* (Seção 2.3), que sugere que um pequeno número de palavras é usado muito mais frequentemente.

Além disso, é possível classificar os pesquisadores pela quantidade de palavras-chave utilizadas em cada conjunto (global e únicas), e consequentemente, destacar os doutores que utilizam o maior número de palavras-chave; identificar os doutores que trabalham com o maior número de tópicos distintos e, por meio da operação de divisão do número de palavras-chave únicas pelo número de palavras-chave do conjunto global, identificar a intensidade com que os doutores se dedicam a seus temas. Com base na visão colaboração, a Tabela 5.4 apresenta os doutores que mais utilizaram palavras-chave em cada conjunto.

Tabela 5.4: Os doutores que utilizaram maior número de palavras-chave em cada conjunto.

	Referência Global			Referência Únicas			Referência Razão GU		
Doutor	**Global**	**Únicas**	**Razão GU**	**Global**	**Únicas**	**Razão GU**	**Global**	**Únicas**	**Razão GU**
1º	**6.428**	933	6,88	3.075	**1.727**	1,78	1.380	10	**138**
2º	**5.248**	418	12,55	4.524	**1.688**	2,68	1.290	11	**117,27**
3 º	**5.004**	418	11,97	3.902	**1.432**	2,72	100	1	**100**
4 º	**4.625**	1.321	3,50	3.540	**1.393**	2,54	198	2	**99**
5 º	**4.524**	1.688	2,68	3.470	**1.359**	2,55	332	5	**66,4**
6 º	**4.474**	1.010	4,42	3.522	**1.326**	2,65	927	14	**66,21**
7 º	**4.343**	897	4,84	4.625	**1.321**	3,50	666	11	**60,54**
8 º	**4.170**	241	17,30	2.233	**1.304**	1,71	2.253	39	**57,76**
9 º	**3.999**	1.123	3,56	2.436	**1.284**	1,89	632	11	**57,45**
10 º	**3.988**	307	12,99	3.676	**1.241**	2,96	57	1	**57**
11 º	**3.902**	1.432	2,72	2.041	**1.177**	1,73	279	5	**55,8**
12 º	**3.760**	727	5,17	2.919	**1.163**	2,50	55	1	**55**
13 º	**3.724**	768	4,84	3.999	**1.123**	3,56	53	1	**53**
14 º	**3.676**	1.241	2,96	3.175	**1.110**	2,86	53	1	**53**
15 º	**3.637**	540	6,73	1.595	**1.096**	1,455	1189	23	**51,69**

As colunas “Referência Global”, “Referência Únicas” e “Referência Razão GU” (onde Razão GU é razão entre global e únicas), norteiam o ranqueamento dos doutores perante utilização das palavras-chave, uma vez que optou-se por apresentar os respectivos quantitativos nos conjuntos analisados. Por exemplo, para identificar o doutor que inseriu o maior número de palavras-chave em seus artigos, basta consultar a coluna “Referência Global”, e verificar que o 1º possui 6.428 palavras-chave no global, que se reduz a 933 únicas e 6,88 de razão entre únicas e global. Pela coluna “Referências Únicas”, identifica-se que o 1º doutor trabalhou com 1.727 palavras-chave únicas, representando o pesquisador que mais dedicou esforços a diferentes tópicos de pesquisas ao longo de sua trajetória. Já na coluna “Referências Razão GU”, o 1º doutor possui em média 138 de intensidade de dedicação por tópico de pesquisa, fornecendo fortes indícios sobre ser especialista em algum dos assuntos que atua. Como complemento a esta análise, pode ser utilizado o arquivo “Arquivo de Indivíduos por palavras-chave” (Seção 4.2) para: (1) destacar os pesquisadores que mais utilizaram determinadas palavras-chave, e com isso, possibilitar a identificação de especialistas de domínio; (2) identificar o número de pesquisadores que utilizam determinada palavra-chave, para verificar a quantidade de indivíduos envolvidos por tópico de pesquisa, e; (3) identificar as palavras-chave utilizadas por determinado pesquisador.

Em continuação, para conhecer os tópicos que se constituem como os principais assuntos de pesquisa, todas as 2.088.220 palavras-chave únicas foram ranqueadas de acordo com a medida de importância de popularidade baseada na frequência. Esse ranqueamento foi realizado, (1) por quinquênios entre 1962 e 2016, para facilitar o entendimento sobre a evolução temporal dos principais temas abordados (ver Apêndice Tabela A.1) e (2) levando em consideração todo o histórico do conjunto, para destacar os principais tópicos abordados nos 55 anos de registros. Incialmente, na tentativa de mapear a evolução dos principais interesses científicos por parte dos doutores brasileiros ao longo do tempo, a Tabela 5.5 apresenta o coeficiente de similaridade das 15 palavras-chave mais populares entre cada par de quinquênios do período analisado. Neste contexto, coeficiente de similaridade mede o quão dois conjuntos de palavras-chave são iguais (a taxa da interseção entre eles), independentemente da ordem do ranqueamento destas palavra-chave em cada conjunto.

Tabela 5.5: Coeficiente de similaridade das palavras-chave mais frequentes entre os quinquênios analisados.

	1962-1966	1967-1971	1972-1976	1977-1981	1982-1986	1987-1991	1992-1996	1997-2001	2002-2006	2007-2011	2012-2016
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 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. Outra possibilidade é a análise do coeficiente de similaridade dos tópicos de pesquisas entre os quinquênios não sequentes, pois, é permitido verificar se tópicos que foram destaques num passado mais distante se mantiveram, desapareceram ou retornaram à evidência em períodos mais recentes. Neste contexto, destaca-se que nenhum dos tópicos de pesquisas evidenciados no quinquênio de 2007-2011 apareceram entre os principais tópicos estudados nos três primeiros quinquênios. Entretanto, a palavra-chave "Brasil", tema de investigação genérico e presente nos principais tópicos de pesquisas dos três primeiros quinquênios analisados, ressurge novamente como importante no período de 2012-2016.

Para complementar a análise quanto à sumarização dos tópicos de pesquisas mais relevantes desenvolvidos pelos doutores brasileiros ao longo do tempo, a Tabela 5.6 apresenta o ranqueamento das principais palavras-chave, considerando o período completo de análise, e, adicionalmente, realiza a classificação destas pelo quantitativo de doutores distintos que as utilizaram. Assim, a coluna "Palavra-chave (1962-2016)" ordena as palavras-chave com base na medida de popularidade de frequência, a coluna "Doutor (1962-2016)" apresenta a ordenação das palavras-chave pelo número de doutores que as utilizaram e, 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 5.6: As palavras-chave mais frequentes e o quantitativo de doutores entre 1962 e 2016.

Ranque	Palavras-chave (1962-2016)		Doutor (1962-2016)		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

Notadamente, a palavra-chave “Educação” tem destaque na pesquisa nacional, visto seu alto valor de popularidade. Contudo, foi apenas no quinquênio 1977-1981 que figurou-se entre as principais pela primeira vez, para não mais ficar fora desse ranque posteriormente. No período de 1997-2001 registrou um aumento de 254% se comparado ao quinquênio anterior, alcançando o topo em popularidade, de onde não saiu até o último período analisado (2012-2016). As palavras-chave “Formação do Professor” e “Ensino”, ambas ranqueadas entre as primeiras no período de 1997-2001 até 2012-2016, corroboram à dedicação dos doutores acerca de pesquisas ligadas a educação. Nesse caso, “Educação” foi também a mais abrangente quanto ao número de doutores (20.126), mantendo seu índice de flutuação igual “0”. Apesar disto, vale ressaltar que o número de doutores envolvidos com determinada palavra-chave não garante que esta será a mais frequente. Por exemplo, “Ensino” é a segunda em número de doutores e décima quanto a frequência, enquanto, “Enfermagem” é décima terceira em doutores e terceira mais frequente. Este fato deve-se a determinados doutores serem mais produtivos que seus pares e utilizarem as mesmas palavras-chave.

Saúde é outro assunto que recebe grande atenção, como pode ser notado, as palavras-chave “Enfermagem”, “Epidemiologia”, “Crianças”, “Idoso”, “Ratos” e “Diagnósticos” estão entre as mais utilizadas considerando todo o período. Dentre estas, destaca-se “Ratos”, (possivelmente utilizada em artigos que envolvam experimentos) que apareceu a primeira vez entre as mais importantes em 1977-1981, perdendo evidência nos dois últimos quinquênios analisados. Nesse ponto, vale destacar que diversos outros tópicos relacionados à saúde foram ranqueados como principais em todos os períodos analisados. Por exemplo, “Aids” teve destaque nos períodos de 1987-1991 e 1992-1996, reflexo do surgimento da doença na década de 1980; já “Doença de Chagas” e o protozoário causador (Trypanosoma cruzi) tiveram maior atenção nos períodos de 1967-1971 até 1992-1996. Também se pode observar temas relacionados à agricultura e pecuária, como “Bovinos”, “Milho”, “Soja”, “Sementes”, “Adubação” e “Suínos”. Dentre estes, no período entre 1962 e 2016, destaca-se “Bovinos”, presente entre as principais palavras-chave de 1962-1966 até 2002-2006, perdendo atenção nos quinquênios de 2007-2011 e 2012-2016 (períodos em que não houve nenhum tópico, dentre os principais, referentes a assuntos que envolvem a agricultura e/ou pecuária).

As palavras-chave “Amazônia” e “Educação Ambiental”, possivelmente relacionados a temas que envolvem problemas ambientais, contaram também com atenção especial por parte dos doutores, principalmente a palavra “Amazônia”, utilizada frequentemente de 1977-1981 até 2012-2016. Em contrapartida, as palavras-chave “Cultura” e “Políticas Públicas” ganharam relevância em quantidade de utilização recentemente, ou seja, nos dois últimos quinquênios do período analisado. E por último, não menos importante, destaca-se a palavra-chave “Brasil”, que exceto o período de 2007-2011, figurou em popularidade em todos os quinquênios analisados.

5.2 Estatísticas por Grande Área do Conhecimento

A grande área do conhecimento é uma das informações requeridas durante o cadastramento dos artigos publicados nos currículos da Plataforma Lattes. De acordo com Dias (2016), a lista das grandes áreas disponíveis na Plataforma Lattes segue, de modo geral, mesma utilizada pela CAPES, salvo pequenas distinções. Por exemplo, na Plataforma Lattes existe a opção da grande área “Outra” em detrimento da Multidisciplinar usada pela CAPES. Assim, torna-se factível realizar agrupamentos por grandes áreas do conhecimento na tentativa de entender quais são as contribuições de cada uma para a ciência brasileira e, com isso, verificar se existem grandes áreas distintas trabalhando no mesmo tema de pesquisa. Neste ponto, conforme o Capítulo de “Desenvolvimento”, é importante destacar duas situações contempladas pela atual pesquisa: (1) em casos do não preenchimento da grande área, foi associado a cada artigo, a principal grande área de atuação do respectivo autor do currículo, que, em caso de inexistência, a grande área “Outra” foi atribuída e, (2) em casos de artigos realizados em coautoria, quando o mesmo artigo foi preenchido com grandes áreas do conhecimento diferentes por seus respectivos autores em seus currículos, tomou-se a decisão de replicar esse artigo em cada uma das grandes áreas informadas. A seguir, as Tabelas 5.7 e 5.8, construídas com base nas visões geral e colaboração (ver Seção 5.1), apresentam os quantitativos de artigos e suas palavras-chave por grande área do conhecimento.

Tabela 5.7: Artigos em anais de congresso e periódicos e suas palavras-chave, visão geral.

Grande Área	Anais de Congresso		Periódicos	
	Artigos	Palavras-chave	Artigos	Palavras-chave
Ciências Agrárias	1.545.824 (17,4%)	**2.675.313 (18,72%)**	654.319 (13,96%)	1.489.671 (16,75%)
Ciências Biológicas	1.374.645 (15,47%)	2.104.980 (14,73%)	745.803 (15,91%)	1.427.793 (16,06%)
Ciências da Saúde	**1.929.251 (21,72%)**	2.436.750 (17,05%)	**1.254.289 (26,77%)**	**2.074.407 (23,33%)**
Ciências Exatas e da Terra	1.168.606 (13,15%)	1.925.422 (13,47%)	645.463 (13,77%)	1.150.055 (12,93%)
Ciências Humanas	995.985 (11,21%)	1.862.199 (13,03%)	483.708 (10,32%)	1.071.795 (12,05%)
Ciências Sociais Aplicadas	505.072 (5,68%)	905.858 (6,34%)	343.270 (7,32%)	659.054 (7,41%)
Engenharias	930.790 (10,48%)	1.821.212 (12,74%)	304.097 (6,49%)	644.829 (7,25%)
Linguística, Letras e Artes	253.215 (2,85%)	475.996 (3,33%)	144.007 (3,07%)	317.396 (3,57%)
Outra	177.849 (2,00%)	77.180 (0,54%)	110.055 (2,34%)	54.527 (0,61%)
Total	**8.881.237 (100%)**	**14.284.910 (100%)**	**4.685.011 (100%)**	**8.889.527 (100%)**

Tabela 5.8: Artigos em anais de congresso e periódicos e suas palavras-chave, visão colaboração.

Grande Área	Anais de Congresso		Periódicos	
	Artigos	Palavras-chave	Artigos	Palavras-chave
Ciências Agrárias	1.055.864 (15,58%)	**2.690.973 (18,13%)**	391.365 (11,99%)	1.463.971 (15,45%)
Ciências Biológicas	1.052.861 (15,53%)	2.318.147 (15,62%)	487.447 (14,93%)	1.658.380 (17,61%)
Ciências da Saúde	**1.459.741 (21,54%)**	2.495.641 (16,81%)	**813.826 (24,93%)**	**2.143.681 (22,76%)**
Ciências Exatas e da Terra	899.138 (13,26%)	2.093.349 (14,10%)	439.256 (13,45%)	1.291.121 (13,71%)
Ciências Humanas	855.866 (12,62%)	1.845.203 (12,43%)	432.052 (13,23%)	1.106.906 (11,75%)
Ciências Sociais Aplicadas	422.929 (6,24%)	928.006 (6,25%)	291.937 (8,94%)	665.511 (7,06%)
Engenharias	696.512 (10,27%)	1.916.713 (12,91%)	210.263 (6,44%)	709.994 (7,53%)
Linguística, Letras e Artes	228.704 (3,37%)	465.312 (3,13%)	136.480 (4,18%)	320.626 (3,40%)
Outra	105.056 (1,55%)	85.884 (0,57%)	61.367 (1,88%)	56.894 (0,60%)
Total	**6.776.671 (100%)**	**14.839.228 (100%)**	**3.263.993 (100%)**	**9.414.084 (100%)**

Ao analisar o percentual das palavras-chave de artigos em anais de congressos e em periódicos das grandes áreas, percebe-se que os valores são bem próximos quando comparado as visões geral e colaboração. Com isso, para facilitar esta análise são considerados os valores médios entre anais de congressos e periódicos das visões geral e colaboração. Logo, em relação as palavras-chave de artigos em anais de congressos, destaca-se a "Ciências Agrárias", pois, representam mais de 18% do conjunto total de palavras-chave, contudo, não é a responsável pelo maior número de artigos publicado. Isto aconteceu devido os indivíduos que atuam na grande área Ciências da Saúde (responsável pelo maior número de artigos publicados) não terem associados palavras-chave em parte dos artigos cadastrados em seus currículos. A respeito dos periódicos, cerca de 23% das palavras-chave de seus artigos referem-se à produção da grande área de "Ciências da Saúde", que nesse caso, também foi responsável pelo maior número de artigos publicados. Com base nos baixos valores encontrados da grande área "Outra", optou-se pela desconsideração das informações referentes a esta grande área para o restante das análises. Outra informação interessante pode ser percebida através da comparação entre os valores médios de palavras-chave por artigo referentes às grandes áreas, em que em anais de congressos, as "Engenharias" lideram com uma média de 2,75 palavras-chave por trabalho produzido, enquanto em periódicos, a "Ciências Agrárias" é a principal, e, assim, contribui em média com 3 palavras-chave a cada novo artigo publicado.

Conforme discutido em capítulos anteriores, entender em quais tópicos de pesquisas a comunidade científica brasileira (grupos de pesquisadores ou áreas específicas) tem empregado seus esforços pode contribuir no auxílio a diversos tipos de tomadas de decisão, como, por exemplo, impulsionar resultados de pesquisas e direcionar escolhas de temas para novas pesquisas. Assim, para conhecer os principais tópicos de pesquisas de cada grande área do conhecimento, todo o conjunto de palavras-chave únicas, resultado do processamento dos dados com base na visão colaboração, foram agrupadas de acordo com a grande área associada ao artigo, para, posteriormente, serem ranqueadas de acordo com a medida de importância baseada em frequência. Assim, a Tabela 5.9 apresenta o ranqueamento das principais palavras-chave utilizadas nos artigos científicos publicados entre 1962 e 2016 por cada grande área.

Tabela 5.9: As palavras-chave mais frequentes em cada grande área entre 1962 e 2016.

Ranque	Ciências Agrárias	Frequência	Ciências Biológicas	Frequência	Ciências da Saúde	Frequência
1	Bovinos	22.570	Taxonomia	15.110	Enfermagem	37.640
2	Cães	18.890	Ratos	11.770	Epidemiologia	26.527
3	Sementes	13.177	Trypanosoma Cruzi	10.912	Idoso	20.819
4	Soja	13.049	Amazônia	9.788	Crianças	20.470
5	Milho	12.984	Ecologia	9.105	Diagnóstico	13.365
6	Ovinos	12.417	Cerrado	8.977	Adolescentes	13.098
7	Controle Biológico	11.532	Morfologia	8.504	Qualidade de Vida	11.403
8	Qualidade	11.167	Biodiversidade	8.255	Obesidade	11.290
9	Equinos	10.508	Mata Atlântica	7.588	Ratos	10.818
10	Irrigação	10.342	Peixes	7.511	Educação Física	10.661
11	Suínos	9.633	Conservação	7.201	Tratamento	10.567
12	Nutrição	9.566	Citogenética	6.608	Aids	9.908
13	Cultivo	9.443	Anatomia	6.329	Educação e Saúde	9.703
14	Germinação	9.429	Plantas Medicinais	6.291	Câncer	8.639
15	Adubação	8.810	Epidemiologia	5.647	Educação	8.494
Ranque	**Ciências Exatas e da Terra**	**Frequência**	**Ciências Humanas**	**Frequência**	**Ciências Sociais Aplicadas**	**Frequência**
1	Sensoriamento Remoto	7.885	Educação	50.333	Comunicação	9.974
2	Geoprocessamento	5.378	Formação do Professor	37.174	Políticas Públicas	7.825
3	Biodiesel	4.426	Políticas Públicas	14.583	Estratégia	6.597
4	Ensino de Química	4.230	Gênero	12.700	Sustentabilidade	6.522
5	Amazônia	4.136	História	12.606	Educação	5.965
6	Geoquímica	4.015	Currículo	11.595	Inovação	5.660
7	Catalisador	3.956	Educação Ambiental	11.418	Jornalismo	5.308
8	Sedimentos	3.685	Cultura	11.001	Turismo	5.216
9	Adsorção	3.557	Memória	10.386	Marketing	5.073
10	Síntese	3.475	Trabalho	10.265	Design	5.033
11	Educação e Matemática	3.397	Psicanálise	9.192	Brasil	4.800
12	Ensino de Física	3.384	Identidade	9.098	Desenvolvimento Sustentável	4.556
13	Educação	3.200	Educação Infantil	8.987	Desenvolvimento Regional	4.415
14	Espectroscopia	2.999	História da Educação	8.777	Meio Ambiente	4.297
15	Flavonoides	2.938	Educação Especial	8.620	Administração	4.279
Ranque	**Engenharias**	**Frequência**	**Linguística, Letras e Artes**	**Frequência**		
1	Compósitos	6.088	Literatura	6.350		
2	Otimização	5.535	Leitura	6.024		
3	Concreto	5.300	Análise Discursiva	5.893		
4	Simulação	4.994	Discurso	5.650		
5	Corrosão	4.937	Literatura Brasileira	5.612		
6	Reciclagem	4.388	Identidade	4.587		
7	Biodiesel	4.360	Formação do Professor	4.075		
8	Modelagem	4.155	Ensino	4.019		
9	Ergonomia	3.977	Memória	3.864		
10	Biomateriais	3.958	Artes	3.663		
11	Cerâmicas	3.852	Poesia	3.592		
12	Propriedades Mecânicas	3.755	Tradução	3.237		
13	Elementos Finitos	3.715	Literatura Comparada	3.128		
14	Resíduos Sólidos	3.611	Língua Portuguesa	3.094		
15	Adsorção	3.579	Linguagem	3.083		

Como pode ser observado, destaca-se a palavra-chave “Educação”, utilizada 50.333 vezes pelos doutores que publicaram na grande área “Ciências Humanas”; quinta maior grande área em quantitativo de palavras-chave utilizadas; e concomitantemente, a palavra-chave mais frequente dentre todas as grandes áreas, tendo aparecido 12.693 a mais que “Enfermagem”, palavra-chave mais utilizada em publicações da grande área “Ciências da Saúde”, e, a segunda mais frequente. Isso mostra que a grande área “Ciências Humanas” contribuiu cerca de 66% para tornar a palavra-chave “Educação” a mais utilizada pelos doutores brasileiros em toda história, uma vez que foi utilizada 75.855 vezes no total (Tabela 5.5). Além disso, é a palavra-chave que mais vezes apareceu entre as principais de todas as grandes áreas, ou seja, também é tema de pesquisa de interesses nas “Ciências da Saúde”, “Ciências Exatas e da Terra” e “Ciências Sociais Aplicadas”. Outras palavras-chave são utilizadas por diferentes grandes áreas. Assim, para facilitar visualização desta relação, a Tabela 5.10 apresenta o coeficiente de similaridade das principais palavras-chave entre cada par de grande área do conhecimento.

Tabela 5.10: Coeficiente de similaridade das principais palavras-chave por grandes áreas entre 1962 e 2016.

	Ciências Agrárias	Ciências Biológicas	Ciências da Saúde	Ciências Exatas e da Terra	Ciências Humanas	Ciências Sociais Aplicadas	Engenharias	Linguística, Letras e Artes
Ciências Agrárias	X	0%	0%	0%	0%	0%	0%	0%
Ciências Biológicas		X	**13,33%**	6,66%	0%	0%	0%	0%
Ciências da Saúde			X	6,66%	6,66%	6,66%	0%	0%
Ciências Exatas e da Terra				X	6,66%	6,66%	**13,33%**	0%
Ciências Humanas					X	**13,33%**	0%	**20,00%**
Ciências Sociais Aplicadas						X	0%	0%
Engenharias							X	0%
Linguística, Letras e Artes								X

Pelos resultados da tabela, verifica-se que apenas “Ciências Agrárias” não possui relação de similaridade dentre seus respectivos principais tópicos com os assuntos centrais das outras grandes áreas. Em contrapartida, ao menos um par das grandes áreas restantes se relacionam entre si quanto ao interesse em pelo menos um tópico de pesquisa, dentre estas, destaca-se a “Ciências Exatas e da Terra”, que possui o maior número de relações: “Ciências Biológicas” (“Amazônia”); “Ciências da Saúde” (“Educação”); “Ciências Humanas” (“Educação”); “Ciências Sociais Aplicadas” (“Educação”) e; “Engenharias” (“Adsorção”, ”Biodiesel”). No entanto, quando observada a intensidade do relacionamento, evidenciam-se os pares das grandes áreas a seguir: (1) “Ciências Humanas” e “Linguística, Letras e Artes” possuem a maior semelhança (20%) entre os principais tópicos de pesquisas que trabalham, são eles: “Identidade”, “Formação do Professor” e, “Memória”; (2) “Ciências Humanas” e “Ciências Sociais Aplicadas” que se vinculam por meio dos temas “Educação” e “Políticas Públicas”; (3) “Ciências Biológicas” e ”Ciências da Saúde” discutem amplamente os temas “Ratos” e “Epidemiologia” e; (4) “Engenharias” e “Ciências Exatas e da Terra” possuem interesses em “Adsorção” e ”Biodiesel”.

Por fim, outra análise interessante foi realizada pela comparação das principais palavras-chave de cada grande área, com as principais palavras-chave de todo o conjunto. Assim, é permitido visualizar as grandes áreas que possuem forte influência sobre a utilização das principais palavras-chave do conjunto total analisado. Logo, a Tabela 5.11 apresenta a relação das palavras-chave mais frequentes (retiradas da Tabela 5.5) associadas as grandes áreas que tiveram estas palavras entre as mais utilizadas no período de 1962 até 2016. À visto disso, percebe-se que todas as palavras-chave frequentemente utilizadas em todo o conjunto estão claramente entre os interesses principais de ao menos uma das grandes áreas, destacando-se a palavra-chave "Educação", como já citado, é assunto discutido nas "Ciências da Saúde", "Ciências Humanas", "Ciências Exatas e da Terra" e, "Ciências Sociais Aplicadas".

Tabela 5.11: As principais palavras-chave de todo o conjunto associadas as grandes áreas entre 1962 e 2016.

Palavras-chave	Grandes Áreas que possuem interesses de destaque nas palavra-chave
Educação	Ciências da Saúde, Ciências Humanas, Ciências Exatas e da Terra e Ciências Sociais Aplicadas
Formação do Professor	Ciências Humanas e Linguística, Letras e Artes
Enfermagem	Ciências da Saúde
Epidemiologia	Ciências da Saúde e Ciências Biológicas
Crianças	Ciências da Saúde
Políticas Públicas	Ciências Humanas e Ciências Sociais Aplicadas
Amazônia	Ciências Exatas e da Terra
Bovinos	Ciências Agrárias
Idoso	Ciências da Saúde
Ensino	Linguística, Letras e Artes
Ratos	Ciências da Saúde e Ciências Biológicas
Brasil	Ciências Sociais Aplicadas
Diagnóstico	Ciências da Saúde
Educação Ambiental	Ciências Humanas
Cultura	Ciências Humanas

5.3 Redes de Palavras-Chave

As redes de palavras-chave podem se constituir como uma importante estratégia para se estudar sobre o desenvolvimento da ciência. Essas redes são formadas a partir das palavras-chave que são associadas aos artigos científicos e estudadas por meio de métricas específicas de redes sociais. Como os artigos cadastrados nos currículos da Plataforma Lattes são associados a um ano de publicação, torna-se possível a construção de redes de palavras-chave a partir de conjuntos de dados referentes a intervalos de tempos específicos, e com isso, a partir de análises de tais redes, destacar as principais palavras-chave em uma determinada época, verificar a evolução entre seus relacionamentos ao longo do tempo e, ainda, estudar a estrutura da(s) rede(s) formada(s). Diante disto, as métricas discutidas na Seção (2.4) foram selecionadas para analisar as redes de palavras-chave caracterizadas e construídas com base no processo descrito na Seção (4.3). Essas métricas também são aplicadas em diversos trabalhos na literatura, conforme apresentado no Capítulo "Trabalhos Correlatos".

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. Nesse caso, o cálculo de algumas métricas, como, por exemplo, “diâmetro da rede” e “caminho mínimo médio” necessita do dispêndio de um alto tempo computacional, que, dado sua limitação para entrega desta etapa da pesquisa, tornou-se inviável. Assim sendo, e, amparado pela análise apresentada através da Figura 5.9, onde 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 referentes aos resultados apresentados pelas Figuras (5.10), (5.11) e (5.12) e Tabela (5.12). 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, ou seja, 99% de abrangência da rede completa.

Em continuação, como a componente gigante de uma rede contém a maior quantidade de vértices conectados, entender sua evolução pode revelar informações interessantes. Para tanto, fez-se necessário a construção das redes referentes a cada ano entre o intervalo de 1962 a 2016 para medir a evolução temporal da componente gigante a partir do número de vértices, diâmetro da rede e caminho mínimo médio ao longo dos anos, conforme apresentado nas Figuras (5.10), (5.11) e (5.12) respectivamente.

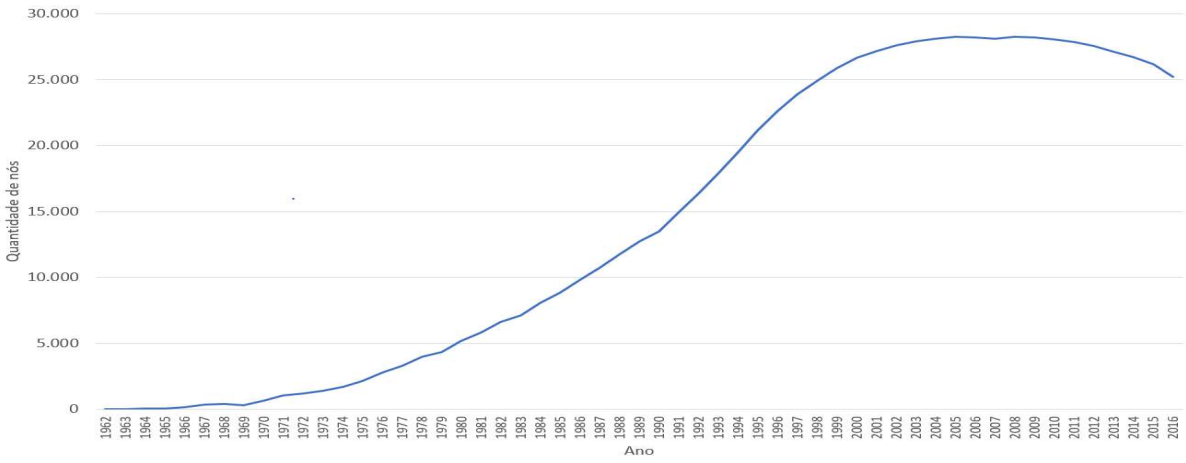


Figura 5.10 - Evolução temporal da quantidade de vértices.

Pela da Figura 5.10, percebe-se um crescimento acentuado do componente a partir do início da década de 80 até o fim da década de 90, totalizando 178.085 vértices a mais se comparado às décadas anteriores. Apesar de alguns anos da década de 2000 apresentar crescimento em relação a década de 1990 e os anos de 2009 a 2016, são registradas quedas consecutivas no número de vértices, e, por isso, a tendência é continuar diminuindo em anos posteriores. O crescimento do componente gigante ao longo dos anos é devido a outros componentes irem se conectando a ele, o que pode significar mais proximidades entre os tópicos de pesquisas e/ou o aumento no interesse dos pesquisadores por novas pesquisas. Já a redução do componente pode apresentar uma mudança de interesses quanto aos tópicos estudados, fortalecer a concentração de esforços em tópicos específicos ou até refletir uma queda na produção científica.

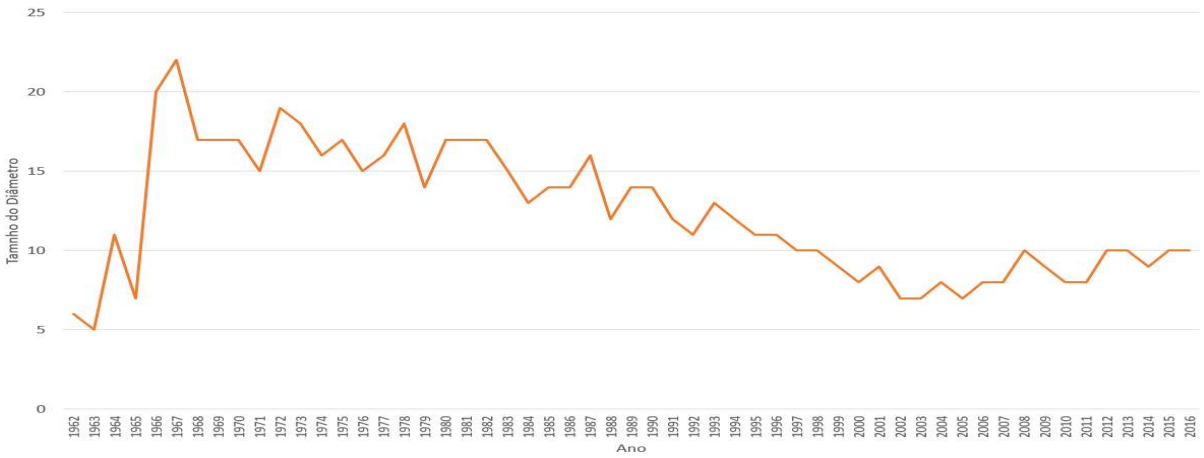


Figura 5.11 - Evolução do diâmetro da rede.

Diante da Figura 5.11, é possível observar a evolução temporal do diâmetro do componente gigante. Assim, percebe-se que, dentre os anos iniciais de análise, destacam-se os maiores valores de diâmetro em 1966 (20) e 1967 (22), resultado do componente gigante com baixa densidade. Porém, a medida que o componente gigante vai se tornando mais conectado, o valor de diâmetro tende a diminuir e a se estabilizar, como aconteceu a partir de 1997, atingindo 10 para não ser maior em anos posteriores.

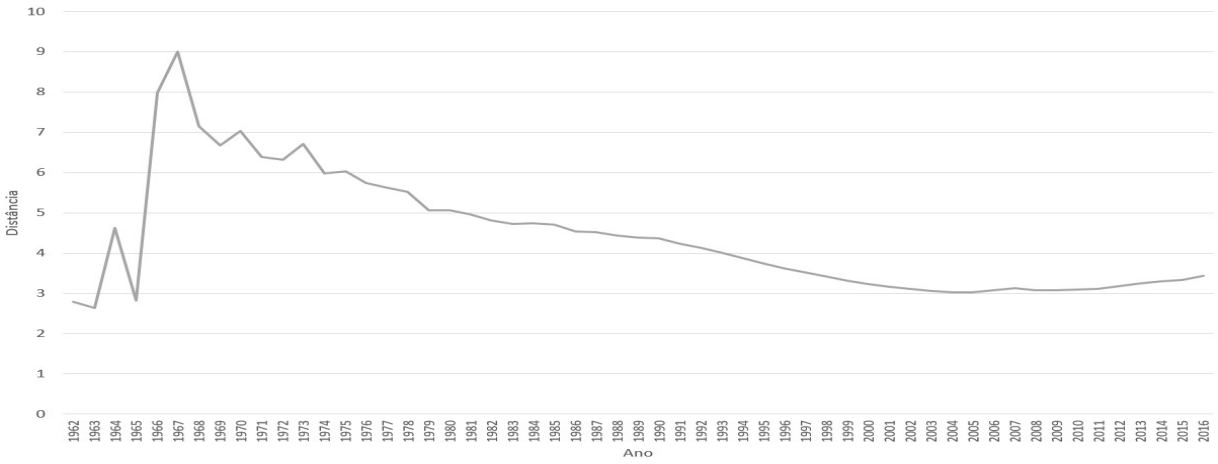


Figura 5.12 - Evolução do caminho mínimo médio.

Na Figura 5.12, é apresentada a evolução temporal do caminho mínimo médio da componente gigante. Como na medida do diâmetro, os valores de caminho mínimo médio também se alteram de acordo com o aumento da conectividade do componente gigante. Diante disso, observa-se que esta medida segue um padrão semelhante ao gráfico do diâmetro da rede, onde os maiores caminhos mínimo médio ocorreram nos anos de 1966 (7,9) e 1967 (9,0) e estabilizaram a partir de 1997 (3,5), devido ao componente ter-se tornado mais conectado.

A Tabela 5.12 apresenta a sumarização das redes de palavras-chave construídas a cada 5 anos entre o intervalo de 1962 a 2016, bem como os resultados das métricas aplicadas em cada uma destas, para, em seguida, ser possível analisar como estão estruturadas.

Tabela 5.12: Resultados das métricas adotas.

	Métricas									
Período	Total de Vértices	Total de Arestas	Grau médio dos Vértices	Total de Vértices no Componente Gigante	% de Vértices no Componente Gigante	Total de arestas no Componente Gigante	Densidade da Rede	Diâmetro da Rede	Caminho mínimo médio	Total de Componentes Isolados
1962-1966	1.918	2.430	2,53	924	0,48%	1.855	0,00132179	17	6,337	568
1967-1971	4.318	7.555	3,49	2.960	0,68%	7.045	0,00081058	15	5,358	937
1972-1976	8.067	19.773	4,90	6.484	0,80%	19.396	0,00060776	15	4,785	1.265
1977-1981	12.915	50.936	7,88	11.320	0,87%	50.657	0,00061080	13	4,190	1.348
1982-1986	17.636	107.216	12,15	16.276	0,92%	107.090	0,00068946	10	3,835	1.243
1987-1991	22.136	202.808	18,32	21.180	0,95%	202.750	0,00082782	10	3,544	904
1992-1996	26.347	480.786	36,49	26.010	0,98%	480.783	0,00138527	10	3,117	335
1997-2001	28.243	1.158.287	82,02	28.182	**0,99%**	1.158.287	0,00290428	6	2,735	62
2002-2006	28.687	**1.867.607**	**130,20**	28.678	**0,99%**	**1.867.607**	**0,00453900**	**5**	**2,547**	10
2007-2011	**28.763**	1.811.113	125,93	**28.761**	**0,99%**	1.811.113	0,00437846	**5**	2,570	**3**
2012-2016	28.635	1.365.603	95,37	28.606	**0,99%**	1.365.603	0,00333100	6	2,699	30

Notadamente, é possível observar uma evolução considerável entre as redes de 1962-1966 e 1992-1996 em praticamente todas as medidas realizadas. Dentre estas, destacam-se: (1) aumento do número de vértices em 1.273%; (2) aumento do número de arestas em 19.685%; (3) aumento do grau médio em cerca de 1.342% e; (4) a abrangência da componente gigante em 98% sobre a rede completa. Ainda é possível verificar um aumento quanto à densidade da rede, e, com isso, a redução dos componentes isolados em 69% e do caminho mínimo médio em cerca de 103%. A evolução das redes nesse período relaciona-se com o início do avanço na produção científica brasileira. Também pode-se observar que as redes do período de 1997 a 2016 apresentam um baixo número de componentes que não se encontram nas respectivas componentes gigantes, isso provavelmente aconteceu devido às palavras-chave isoladas apresentarem ruídos (por exemplo, erros de digitação) ou termos específicos, como fórmulas matemáticas.

Entretanto, foram as redes de 1997-2001 e 2002-2006 que apresentaram um enorme crescimento quanto ao número de arestas se comparadas as redes anteriores, especialmente a rede de 2002-2006 que atingiu 1.867.607 de arestas, o maior número dentre todas as redes. Isso impactou diretamente no maior valor de grau médio dos vértices (130,20) e nos melhores valores encontrados para densidade das redes (0,00453900) e caminho mínimo médio (2,547). Este fato coincidiu com o maior período de crescimento da produção científica brasileira. Apesar do número de vértices da rede de 2002-2006 ter sido maior se comparado com as redes anteriores, foi por meio da rede de 2007-2011 o registro do maior número de vértices entre todas as redes analisadas, fato este que teve impacto diretamente quanto ao menor número de componentes isolados dentre todas as redes. Curiosamente, isso ocorreu na rede que registrou a primeira queda do número de arestas, que, com isso, registrou queda

no número de grau médio, densidade da rede, e, aumento no caminho médio. Tal acontecimento pode estar associado a mudanças de interesses em tópicos de estudos, alcance da maturidade de determinados tópicos de pesquisas ou com o início da queda da produção científica brasileira. Por fim, a rede de 2012-2016, se comparada a de 2007-2011, apresenta uma queda de 25% no número de arestas (segunda queda consecutiva em arestas), que reflete na alteração de todos os valores das medidas realizadas.

Conforme apresentado no Capítulo de "Trabalhos Correlatos", recentemente, um número crescente de trabalhos encontrados têm dedicado esforços nos estudos de redes de palavras-chave, como apoio ao entendimento sobre o desenvolvimento da ciência. Alguns destes trabalhos 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 apresentada na Seção (5.1), mais especificamente, a medida baseada em frequência (ver Tabela 5.5). Para facilitar a comparação entre os resultados encontrados pela medida de importância citada, os vértices de maior grau foram ranqueados a partir das redes de quinquênios entre 1962 e 2016 (ver Apêndice Tabela A.2) e da rede referente a todo histórico de análise (Tabela 5.13).

Tabela 5.13: 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
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

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 visto disso, "Educação" se consolida como a palavra-chave de maior interesse dos doutores entre 1962 e 2016 sob a ótica das duas medidas de importâncias de tópicos utilizadas até o momento. Em complemento disso, no que tange à semelhança entre as principais palavras-chave elencadas pelas medidas de importância levando em consideração todo o histórico, cerca de 80% das principais palavras-chave encontradas pela medida de frequência também foram encontradas através do cálculo do grau. Em relação as principais palavras-chave ranqueadas pelas medidas de importâncias em períodos de 5 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 de pesquisas.

Outra informação interessante que pode ser extraída das redes é a medida de coocorrência 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 5.14 apresenta os pares das palavras-chave que mais coocorrem, ou seja, as arestas mais densas da rede de palavras-chave, que foi construída a partir de todo o conjunto histórico dos dados.

Tabela 5.14: 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
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

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”.

5.4 Análise Comparativa entre as medidas de Grau e Frequência

Para facilitar a comparação entre as palavras-chave classificadas pelas medidas de importâncias utilizadas, a Tabela 5.15 apresenta as principais palavras-chave ranqueadas pela frequência e a respectiva ordem de ranqueamento por grau entre o período de 1962 e 2016.

Tabela 5.15: 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	6
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

Os resultados apresentados nessa tabela confirmam o nível de similaridade entre as principais palavras-chave encontradas pelas medidas de importâncias de tópicos utilizadas, conforme discussão sobre a Tabela (5.13). 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 é a mesma, exceto para as palavras-chave “Educação” que apareceu em primeiro no ranque para ambas medidas de importância, e, “Cultura” a décima quinta para frequência e grau. Diante disto, para melhor compreender a relação entre os resultados das medidas de importâncias aplicadas, uma análise de correlação foi realizada. Para tal análise, como apresentado na Figura (5.13), foi calculado o coeficiente de Spearman considerando todo o histórico de palavras-chave entre frequência e grau, visando verificar a existência de uma relação direta entre os resultados das medidas de importância.

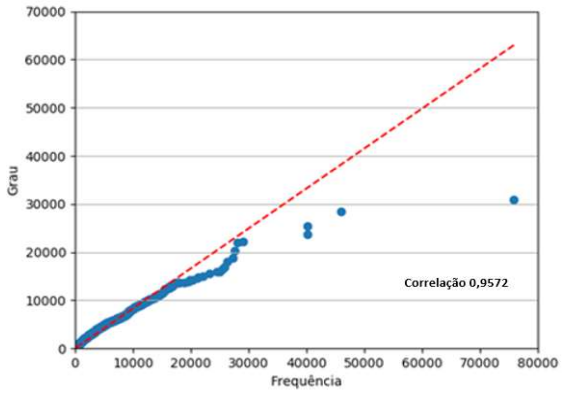


Figura 5.13 - Correlação entre as medidas de importância frequência e grau.

Como pode ser observado, os resultados apresentam um coeficiente de correlação positivo de “0,9572” entre frequência e grau. Pelo valor obtido, destaca-se a comprovação de uma relação direta entre a frequência das palavras-chave e medida baseada em grau das redes de palavras-chave, indicando que a medida de grau é uma alternativa interessante para identificação das principais palavras-chave estudadas.

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, conforme explanado na Seção (3.3), a Plataforma Lattes possui dados de periódicos e conferências das mais variadas áreas de pesquisas com diferentes níveis de qualidade, na qual medidas de importância apenas quantitativas para destacar tópicos de pesquisas podem induzir a interpretações equivocadas. Com isso, no próximo capítulo é proposta uma nova medida de importância de tópicos que consideram características qualitativas das publicações (no caso, o fator de impacto dos periódicos), e em seguida, será feita uma análise comparativa com as medidas de importância de frequência e grau utilizando somente o conjunto de dados referentes aos artigos publicados em periódicos pelos doutores brasileiros.

ANÁLISE TEMPORAL DOS PRINCIPAIS TÓPICOS DE PESQUISA DA CIÊNCIA BRASILEIRA

Neste capítulo, é apresentada uma análise temporal dos principais tópicos de pesquisas estudados pelos doutores brasileiros que possuem currículos cadastrados na Plataforma Lattes. Assim, a Seção (6.1) apresenta um estudo que culmina na seleção da medida de importância mais adequada ao contexto para destacar os principais tópicos de pesquisas. Nas Seções (6.2) e (6.3), os principais tópicos são analisados em sua evolução temporal, respectivamente num âmbito geral e por grande área do conhecimento.

6.1 Medida de Importância de Tópico Selecionada

Inicialmente, são apresentados os dados de publicações dos doutores brasileiros em periódicos, adicionados à sua respectiva característica qualitativa FI (*Fator de Impacto*), no intuito de fornecer um conjunto de informações quantitativas e qualitativas para testar o desempenho das medidas de importância de tópicos (Subseção 6.1.1). Posteriormente, é descrita a medida de importância *TF.FI*, proposta nesta tese com o intuito de realizar o ranqueamento das palavras-chave considerando aspectos quantitativos e qualitativos dos dados referentes aos artigos em periódicos (Subseção 6.1.2). Por fim, é apresentada uma análise comparativa entre as medidas de frequência, grau e *TF.FI,* com a finalidade de verificar qual a melhor medida para se considerar as características existentes nos dados das produções científicas cadastradas na base curricular da Plataforma Lattes (Subseção 6.1.3).

6.1.1 Definição do Conjunto de Dados

O periódico é o principal meio de comunicação formal utilizado pelos pesquisadores para disseminação dos resultados de suas pesquisas na maioria das áreas do conhecimento (FREIRE, 2012). Com base nisso, foram selecionados os dados dos artigos publicados em periódicos pelos doutores brasileiros contidos no arquivo de publicações científicas referente à visão colaboração (Seção 4.2) para representar o conjunto das informações a serem exploradas até a finalização desta tese. Essa exploração será feita sob duas perspectivas, a saber: (1) o teste do desempenho das medidas de tópicos e (2) o embasamento para a realização de uma análise temporal dos principais assuntos estudados em cada grande área do conhecimento da ciência brasileira. A Figura 6.1 apresenta o diagrama de Venn para facilitar a compreensão da relação entre número total de palavras-chave únicas associadas

aos artigos em anais de congressos e em periódicos pelos doutores em seus currículos cadastrados na Plataforma Lattes.

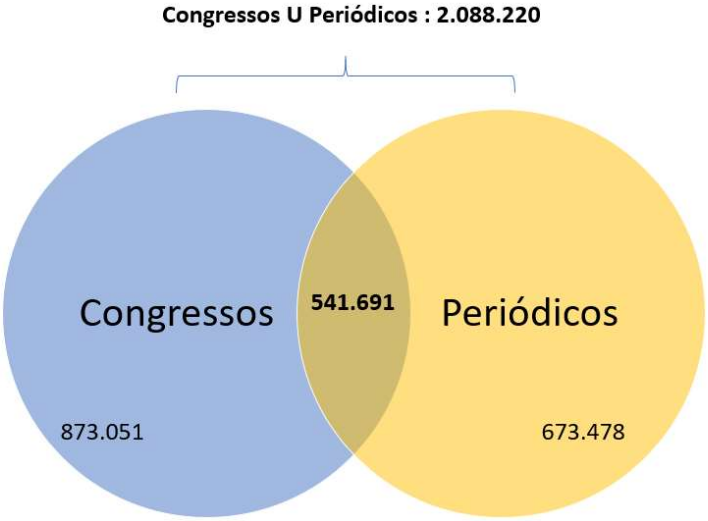


Figura 6.1 - Número de palavras-chave em artigos publicados em anais de congressos e em periódicos.

Como pode ser observado, somente cerca de 26% das palavras-chave únicas são utilizadas concomitantemente em artigos de anais de congressos e em periódicos pelos doutores. Isso possivelmente sugere que os assuntos principais publicados em anais de congressos são relativamente diferentes dos abordados em periódicos. Outra informação importante é que o conjunto dos artigos publicados em periódicos totaliza 1.215.169 palavras-chave únicas, ou seja, representa 58% da abrangência do conjunto união das palavras-chave utilizadas em anais de congressos e em periódicos entre 1962 e 2016.

Como citado em capítulos anteriores, a Plataforma Lattes congrega dados de periódicos das mais variadas áreas do conhecimento com diferentes níveis de qualidade. Sabe-se que a publicação de artigos em periódicos de prestígio possibilita ao pesquisador reconhecimento junto a seus pares, citações por outros autores e visibilidade no cenário nacional e/ou internacional. Por isso é importante considerar a qualidade e a confiabilidade de cada periódico para tornar as análises mais coesas em relação ao contexto. Assim sendo, os dados referentes aos artigos dos doutores brasileiros contidos no arquivo de publicações científicas (Seção 4.2) foram complementados com o valor do Fator de Impacto (FI) de cada periódico. Em seguida, o componente de "tratamento dos dados" (Seção 4.2) foi ajustado para adicionar a somatória dos valores anuais de FI dos periódicos às respectivas palavras-chave dos artigos contidas nos arquivos de resultado padrão. Dessa maneira, as palavras-chave de artigos publicados em periódicos de relevância têm valor maior de FI em relação às palavras-chave de artigos publicados em periódicos de menor impacto. Isso foi possível por meio da utilização do conjunto de dados do JCR (*Journal Citation Reports*), que permite, a partir do ISSN do periódico, a obtenção de informações referentes a seu impacto e influência na comunidade científica ano a ano desde 1997. A Figura 6.2 ilustra a adaptação necessária no arcabouço de componentes (Seção 4.2) para a inserção do valor de FI dos periódicos no conjunto dos dados contidos no arquivo de publicações científicas e nos arquivos de resultado padrão.

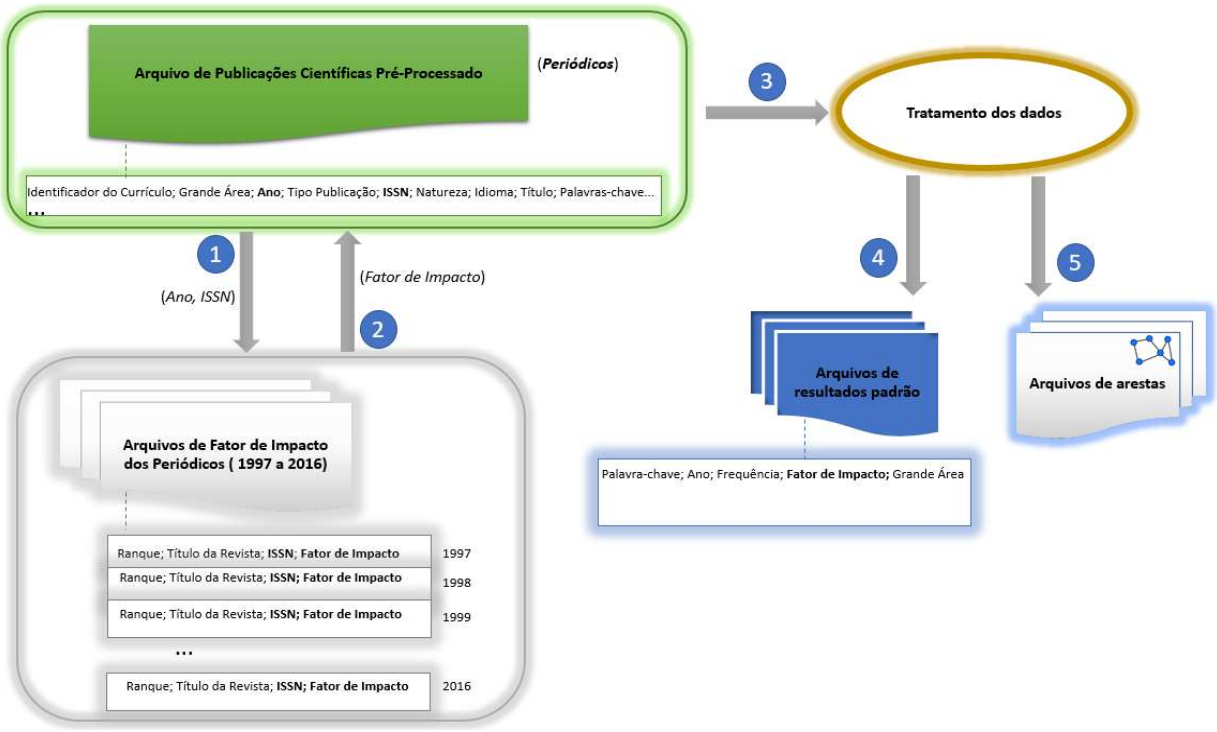


Figura 6.2 - Processo para adição de FI dos periódicos ao conjunto dos dados científicos dos doutores.

Dessa forma, o conjunto de dados de publicações em periódicos selecionado para estudo passa a contemplar o Fator de Impacto destes e refere-se ao período de 1997 a 2016, ou seja, da data inicial de disponibilização do FI ao último ano considerado nas análises. Com isso, o número de palavras-chave do conjunto global foi reduzido de 9.417.084 para 8.255.113, e as palavras-chave únicas a serem analisadas foram reduzidas de 1.215.169 para 1.068.806, ou seja, para 88% do conjunto de palavras-chave únicas inseridas em artigos publicados em periódicos pelos doutores brasileiros durante todo o histórico de análise.

6.1.2 Medida de Importância de Tópico TF.FI

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 disso, auxiliar nos mais variados cenários de tomada de decisão. Esses estudos podem ser viabilizados pela utilização das medidas de importância de tópicos baseadas na frequência das palavras-chave de artigos científicos ou em medidas de centralidade de redes de palavras-chave, conforme resultados preliminares apresentados no capítulo anterior. Apesar da validade desses resultados, no caso da Plataforma Lattes, que possui palavras-chave de artigos publicados em periódicos de diferentes níveis de impacto, análises que contemplam apenas atributos quantitativos não conseguem explorar em completude as características qualitativas existentes nesses dados. À vista disso, nesta tese foi desenvolvida a medida de importância de tópico TF.FI, que é capaz de levar em consideração tanto as características quantitativas das palavras-chave dos artigos quanto as características qualitativas dos periódicos em que tais artigos foram publicados.

Para tanto, a TF.FI pondera a frequência das palavras-chave de um conjunto de artigos publicados em periódicos com a soma dos valores de FI dos periódicos em que foram publicados os artigos que contêm as respectivas palavras-chave. Essa medida parte do pressuposto de que as palavras-chave de um artigo publicado em um periódico de alta relevância são mais importantes do que as mesmas palavras-chave de um outro artigo publicado em periódico de menor relevância. Por isso, as palavras-chave de artigos publicados em periódicos de destaque recebem maior valor de FI como parcela de ponderação. Quanto a isso, é importante destacar que: (1) o período temporal para o cálculo da medida TF.FI de determinado tópico é considerado de acordo com as decisões tomadas para cada análise; (2) os valores de FI de um mesmo periódico podem se alterar ao longo dos anos, visto que esses valores são recalculados anualmente; e (3) quando são levadas em conta as palavras-chave de um artigo que foi publicado em um periódico que não possui valor de FI, seu valor no somatório de FI para as respectivas palavras-chave é considerado 0, mas sua utilização é contabilizada no cálculo da quantidade de vezes que determinadas palavras apareceram no somatório da frequência.

A Figura 6.3 apresenta a equação para o cálculo da medida TF.FI de um determinado tópico num intervalo de tempo, para que posteriormente esses tópicos sejam ranqueados com o objetivo de destacar os mais importantes do conjunto de dados em análise.

$$TF.FI_{K\,(i,m)} = \sum_{n=i}^{m} TF_{K\,(n)} \cdot \sum_{n=i}^{m} FI_{K\,(n)}$$

Em que:

$TF_{k(n)}$ = Quantidade da palavra-chave k inserida nos artigos publicados em determinado ano

$FI_{k(n)}$ = Soma dos valores de FI dos periódicos onde foram publicados os artigos que contêm a palavra-chave k em determinado ano

i = Ano inicial do período de análise

m = Ano final do período de análise

Figura 6.3 - Medida de importância de tópicos TF.FI.

Para melhor entendimento, a Figura 6.4 apresenta um exemplo de utilização da medida de TF.FI a partir de um conjunto de dados de publicações científicas.

Autor	G. Área	Título da Publicação	Ano	Palavras-chave	Periódico	FI
...	...	...	...	...	...	...
ID 1	Engenharias	Titulo 1	2008	Cientometria; Redes Sociais; Tópicos de Pesquisa	0000-0001	10,00
ID 2	Engenharias	Titulo 2	2009	Bibliometria; Tópicos de Pesquisa	0000-0002	15,00
ID 3	Engenharias	Titulo 3	2012	Bibliometria; Redes Sociais; Tópicos de Pesquisa	0000-0002	20,00
ID 4	Engenharias	Titulo 4	2013	Bibliometria; Redes Sociais; Mineração de Texto	0000-0002	2,00
...	...	...	...	...	...	...

Palavras-chave	TF	FI	TF.FI
Tópicos de Pesquisa	3	45	135
Bibliometria	3	37	111
Redes Sociais	3	32	96
Cientometria	1	10	10
Mineração de Texto	1	2	2

Figura 6.4 - Exemplo de aplicação da medida de importância TF.FI em um conjunto de dados científicos.

Como pode ser observado, “Tópicos de Pesquisa” tem popularidade maior que “Bibliometria” e “Redes Sociais”. Embora essas palavras-chave possuam o mesmo valor de índice de frequência, “Tópicos de Pesquisa” refere-se a artigos de periódicos de maior fator de impacto.

6.1.3 Análise Comparativa entre Medidas de Importância de Tópicos

Aqui são comparados os resultados da aplicação das medidas de frequência (TF), grau e *TF.FI* sobre as palavras-chave dos artigos publicados em periódicos pelos doutores brasileiros no período de 1997 a 2016, para verificar qual medida se adequa melhor ao contexto dos dados contidos na base curricular da Plataforma Lattes. Primeiramente, foi realizada a interseção entre os conjuntos das 100 primeiras palavras-chave ranqueadas por cada medida de importância, no intuito de avaliar a similaridade entre os resultados encontrados pelas diferentes medidas. A escolha do tamanho 100 para os conjuntos justifica-se pelo fato de nenhum dos estudos empíricos anteriores sobre bibliometria de palavras-chave terem analisado uma quantidade maior de palavras (CHEN e XIAU, 2016) e por considerarmos esse valor significativo. A Figura 6.5 apresenta o diagrama de Venn referente à sobreposição das 100 primeiras palavras-chave ranqueadas pelas medidas de frequência (TF), grau e TF.FI.

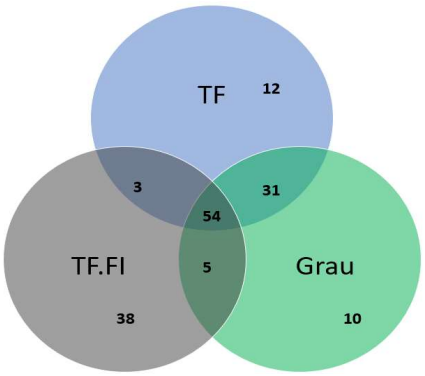


Figura 6.5 - Interseção entre 100 primeiras palavras-chave ranqueadas pelas medidas de TF, Grau e TF.FI.

Notadamente, existem 54 palavras-chave comuns aos três conjuntos ranqueados pelas diferentes medidas de importância, que consequentemente representam 54% de similaridade entre as três medidas utilizadas. O conjunto ranqueado pela medida de TF.FI possui 38 palavras-chave exclusivas, enquanto o conjunto referente à medida de grau possui 10, representando, respectivamente, o maior e o menor número de palavras-chave distintas. Outra informação relevante é que a medida de grau tem apenas 15 palavras-chave diferentes do conjunto ranqueado pela TF, ou seja, essas duas medidas de importância apresentaram 85% de similaridade entre suas respostas. Já a medida de TF.FI gerou um conjunto mais diversificado em relação às demais medidas, pois tem 43 palavras-chave diferentes da TF e 41 da medida de grau. Adicionalmente, para melhor compreender a relação entre os resultados das medidas de importância utilizadas, uma análise de correlação foi realizada. Para tal análise, como apresentado na Figura 6.6, foi calculado o coeficiente de Spearman considerando todas as palavras-chave contidas em artigos publicados em periódicos de 1997 a 2016, entre TF e grau, TF.FI e TF, e TF.FI e grau, visando verificar o nível de relação entre os resultados das medidas de importância.

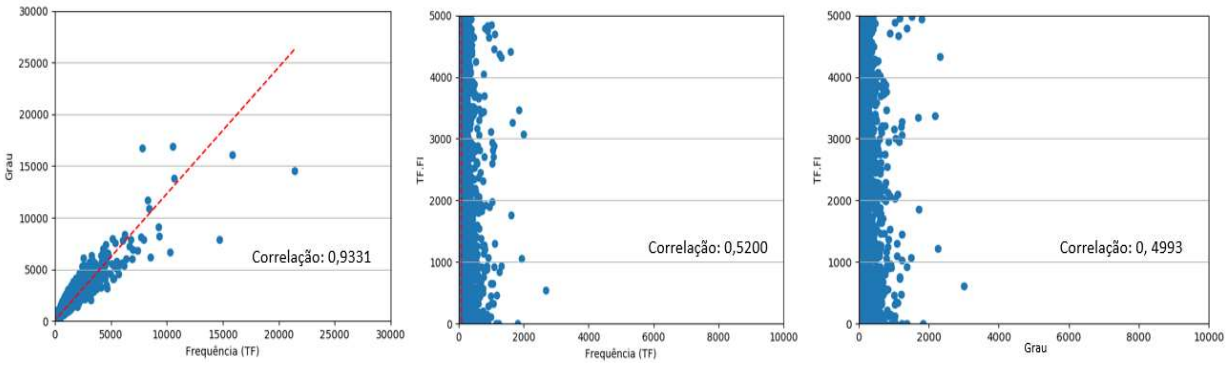


Figura 6.6 - Correlação entre as medidas de importância frequência e grau; e frequência e TF.FI.

Conforme apresentado, os resultados apontam um coeficiente de correlação positivo de "0,9331" entre TF e grau, "0,5200" entre TF e TF.FI e "0,4993" entre grau e TF.FI. Pelos valores obtidos, destaca-se: (1) uma relação muito forte entre as medidas de TF e grau, possivelmente por essas duas medidas considerarem apenas aspectos quantitativos dos dados; e (2) uma relação levemente moderada entre TF e TF.FI e grau e TF.FI, possivelmente devido ao fato de a medida TF.FI considerar tanto as características quantitativas das palavras-chave quanto as qualitativas dos diferentes periódicos. Por fim, apesar da ampla utilização na literatura das medidas de TF e grau para o ranqueamento das palavras-chave a partir de um conjunto de dados de publicações científicas, no caso da Plataforma Lattes a medida TF.FI se apresenta como uma interessante alternativa para esse ranqueamento, visto que leva em consideração as características intrínsecas nos dados referentes ao histórico de publicações em periódicos pelos doutores brasileiros.

6.2 Análise Temporal dos Principais Tópicos de Pesquisas

Para conhecer os tópicos que se constituem como os principais assuntos de pesquisa discutidos por pesquisadores doutores brasileiros, o novo conjunto de análise representado por 1.068.806 palavras-chave únicas, contidas nos artigos publicados em periódicos entre 1997 e 2016, foi reanqueado de acordo com a medida de importância de popularidade proposta nesta tese, a TF.FI (Seção 6.1.2). Esse ranqueamento foi realizado (1) por biênios entre 1997 e 2016, para facilitar o entendimento sobre o comportamento da evolução temporal dos principais temas abordados (ver Apêndice, Tabela A.3), e (2) levando em consideração todo o histórico do conjunto dos dados a serem analisados, para destacar os principais tópicos abordados nos últimos 20 anos de pesquisa da ciência brasileira de acordo com seu registro pelos doutores brasileiros que possuem currículos cadastrados na Plataforma Lattes. Inicialmente, na tentativa de mapear a evolução dos principais interesses científicos por parte dos doutores, a Tabela 6.1 apresenta o coeficiente de similaridade (taxa de interseção entre os conjuntos) das 15 palavras-chave mais populares entre cada par de biênios do período analisado. Neste ponto vale ressaltar que, para melhorar a visualização dos valores de TF.FI apresentados nas tabelas desta e da próxima Seção, os mesmos foram truncados.

Tabela 6.1: Coeficiente de similaridade das palavras-chave ranqueadas por TF.FI entre os biênios analisados.

	1997-1998	1999-2000	2001-2002	2003-2004	2005-2006	2007-2008	2009-2010	2011-2012	2013-2014	2015-2016
1997-1998	X	**66,66%**	53,33%	53,33%	60,00%	53,33%	53,33%	53,33%	53,33%	33,33%
1999-2000		X	80,00%	73,33%	73,33%	66,66%	66,66%	53,33%	53,33%	33,33%
2001-2002			X	80,00%	80,00%	80,00%	80,00%	66,66%	66,66%	46,66%
2003-2004				X	73,33%	73,33%	73,33%	60,00%	60,00%	46,66%
2005-2006					X	86,66%	86,66%	73,33%	73,33%	53,33%
2007-2008						X	**93,33%**	80,00%	80,00%	60,00%
2009-2010							X	86,66%	86,66%	66,66%
2011-2012								X	**93,33%**	80,00%
2013-2014									X	80,00%
2015-2016										X

Como notado, não houve ao longo do período 100% de similaridade nos tópicos centrais estudados entre pares de biênios. Isso mostra que a cada 2 anos os principais interesses científicos dos doutores brasileiros têm mudado por algum motivo, como, por exemplo, por uma nova demanda da sociedade ou a maturação de um determinado tema. No entanto, ao comparar os pares de biênios sequentes, nota-se que 100% das comparações realizadas revelam que ao menos 66,66% dos principais tópicos de pesquisas estudados no biênio atual foram considerados relevantes no biênio anterior, destacando os períodos de 2007-2008 a 2009-2010 e 2011-2012 a 2013-2014, em que o percentual de similaridade foi de 93,33%, havendo divergência apenas quanto às palavras-chave “Taxonomia” e “Biodiesel”, no primeiro período, e “Citocina” e “Nanopartícula”, no segundo. Como exemplo disso, observa-se que a palavra-chave “Taxonomia”, que teve destaque de popularidade em 2005-2006 e 2007-2008, deu lugar à palavra-chave “Biodiesel” no biênio de 2009-2010, que se manteve com popularidade destacada até o biênio de 2015-2016. Uma hipótese para tal acontecimento é a constante preocupação global com a busca por melhorias nas questões climáticas, situação na qual o biodiesel se destaca como uma importante fonte renovável de energia. Soma-se a isso o fato de o Brasil ser um dos maiores produtores mundiais de biocombustível (INFOMONEY, 2018).

Outra possibilidade é a análise do coeficiente de similaridade dos tópicos de pesquisas entre os biênios não sequentes, pois assim é possível verificar se tópicos que foram destaques num passado mais distante mantiveram, desapareceram ou retornaram à evidência em períodos mais recentes. Nesse contexto, percebe-se que 33,33% dos tópicos de pesquisas (5 palavras-chave) evidenciados como os principais no biênio de 2015-2016 também apareceram no ranque nos dois primeiros biênios analisados e que, à medida que os principais tópicos do último biênio são comparados com os dos biênios mais recentes, o percentual de similaridade aumenta ou se mantém. Isso mostra que, apesar dos principais interesses dos doutores terem mudado a cada 2 anos, boa parte desses tópicos principais transcenderam mais de um biênio em evidência, como é o caso das palavras-chave “Epidemiologia”, “Ratos”, “Amazônia”, “Câncer” e “Brasil”, que receberam destaque em popularidade nos vinte anos em análise.

Para complementar a análise quanto à sumarização dos tópicos de pesquisas mais relevantes desenvolvidos pelos doutores brasileiros e publicados em periódicos, a Tabela 6.2 apresenta o ranqueamento das principais palavras-chave considerando todo o período analisado.

Tabela 6.2: As principais palavras-chave ranqueadas por TF.FI entre 1997 e 2016.

Ranque	Palavras-chave (1997-2016)	TF.FI
1	Epidemiologia	197.941.520
2	Amazônia	97.458.265
3	Brasil	69.948.788
4	Ratos	69.872.440
5	Câncer	60.588.976
6	Trypanosoma Cruzi	41.543.320
7	Diagnóstico	37.740.892
8	Estresse Oxidativo	37.103.109
9	Obesidade	32.621.174
10	Diabetes	29.315.396
11	HIV	26.772.660
12	Óxido Nítrico	26.289.914
13	Inflamação	25.895.219
14	Taxonomia	19.505.268
15	Bovinos	19.281.820

Nota-se claramente que temas relacionados à saúde têm destaque na pesquisa nacional, haja vista o alto valor de popularidade das palavras-chave “Epidemiologia”, “Ratos”, “Câncer”, “Trypanossoma Cruzi”, “Diagnóstico”, “Estresse Oxidativo”, “Obesidade”, “Diabetes”, “HIV”, “Óxido Nítrico” e “Inflamação”, que representam 73,33% do conjunto das palavras mais populares. Dentre estas, destaca-se “Epidemiologia”, que esteve no topo em todos os biênios analisados, possivelmente devido a sua contribuição para a política de saúde pública, já que se propõe a estudar fenômenos de saúde/doença e seus fatores condicionantes e determinantes na população humana. Meio ambiente é outro assunto que recebe grande atenção, pois as palavras-chave “Amazônia” e “Taxonomia” possuem popularidade destacada, sendo que a primeira, como citado anteriormente, se destaca em todos os biênios analisados, provavelmente pelo interesse da sociedade em questões como desmatamento, preservação ambiental e aquecimento global. Já a palavra-chave “Brasil” (tema de investigação genérico) indica que os doutores têm dedicado esforços a pesquisas relacionadas ao país. Por fim, “Bovinos”, tema de interesse do agronegócio, teve maior relevância nos períodos de 1999-2000 a 2009-2010.

6.3 Análise Temporal dos Principais Tópicos de Pesquisas por Grande Área

Aqui são apresentadas as principais contribuições dos doutores em periódicos para cada grande área do conhecimento da ciência brasileira. Para isso, os dados utilizados na seção anterior foram agrupados por grande área e, em seguida, receberam a aplicação da medida de importância TF.FI para a realização de um ranqueamento das palavras-chave em cada um desses grupos. Do ponto de vista da apresentação dos resultados, esse ranqueamento foi realizado de acordo com o exposto na seção anterior, ou seja: (1) por biênios entre 1997 e 2016, em que foi calculado o coeficiente de similaridade das 15 palavras-chave mais populares entre cada par de biênios do período analisado; e (2) levando em consideração todo o histórico dos dados a serem analisados.

A grande área Ciências Agrárias (Tabela 6.3, 6.4; Apêndice Tabela A.4) não apresentou durante os 20 anos de pesquisas em análise 100% de similaridade nos principais tópicos estudados entre os pares de biênios. Isso mostra que a cada 2 anos os principais interesses dos doutores brasileiros que atuam nas Ciências Agrárias têm mudado por algum motivo. Entretanto, ao analisar os pares de biênios sequentes, nota-se que essa mudança ocorre paulatinamente ao longo do tempo, pois, pelo menos 53% dos principais tópicos estudados em um biênio corrente também foram estudados no biênio anterior. Nesse caso, vale destacar o par de biênios 2007-2008 e 2009-2010, em que o percentual de similaridade foi de 86,66%, havendo divergência apenas quanto às palavras-chave "Proteínas" e "Suínos", amplamente estudadas no primeiro biênio, e "Sémen" e "Qualidade", destacadas no segundo. Quando analisada a similaridade entre biênios não sequentes, percebe-se que no mínimo 33,33% dos principais tópicos ficaram mais de 2 anos em evidência. Nesse sentido, vale ressaltar as palavras-chave "Bovinos", "Cães", "Equinos", "Epidemiologia" e "Soja", que permaneceram em evidência em todos os biênios analisados. Outra observação é que 73,33% dos principais tópicos em evidência no biênio de 2001-2002 continuaram em destaque pelos próximos 5 biênios consecutivos, mostrando a constância de boa parte dos interesses dos doutores durante 12 anos de pesquisa.

Assuntos possivelmente relacionados à agricultura e à pecuária são de interesses dos doutores brasileiros que atuam na grande área das Ciências Agrárias, na qual "Bovinos", "Equinos", "Ovinos", "Caprinos", "Milho", "Soja", "Cana-de-Açúcar" e "Sémen" possuem alta popularidade considerando todo o período de análise, destacando-se "Bovinos", que é a primeira do ranque e possui popularidade 5 vezes maior que a segunda palavra-chave. Além disso, "Bovinos" só não esteve no topo em popularidade no biênio de 2015-2016, quando apareceu na segunda colocação. Outras palavras-chave como "Epidemiologia", "PCR", "Nutrição" e "Proteína" provavelmente estão relacionadas a estudos que envolvem saúde no contexto dos animais. Já a palavra-chave "Brasil" (tema de investigação genérico) indica que historicamente os doutores têm dedicado esforços a pesquisas de interesses do país. Por fim, 26,66% das principais palavras-chave das Ciências Agrárias no período de 1997-2016 também foram destaques no conjunto completo dos dados (Seção 6.2), sendo que essa grande área contribuiu com cerca de 55% da popularidade total do tópico "Bovinos" e em média 1,8% para os tópicos "Epidemiologia", "Amazônia" e "Brasil".

Tabela 6.3: Coeficiente de similaridade das palavras-chave das Ciências Agrárias ranqueadas por TF.FI.

	1997-1998	1999-2000	2001-2002	2003-2004	2005-2006	2007-2008	2009-2010	2011-2012	2013-2014	2015-2016
1997-1998	X	60,00%	60,00%	53,33%	60,00%	46,66%	53,33%	53,33%	46,66%	53,33%
1999-2000		X	**53,33%**	46,66%	53,33%	46,66%	46,66%	46,66%	**33,33%**	40,00%
2001-2002			X	**73,33%**	**73,33%**	**73,33%**	**73,33%**	**73,33%**	46,66%	53,33%
2003-2004				X	66,66%	53,33%	60,00%	60,00%	53,33%	53,33%
2005-2006					X	80,00%	80,00%	66,66%	46,66%	60,00%
2007-2008						X	**86,66%**	66,66%	46,66%	53,33%
2009-2010							X	73,33%	46,66%	53,33%
2011-2012								X	60,00%	66,66%
2013-2014									X	80,00%
2015-2016										X

Tabela 6.4: As principais palavras-chave das Ciências Agrárias ranqueadas por TF.FI entre 1997 e 2016.

Ranque	Palavras-chave (1997-2016)	TF.FI
1	**Bovinos**	10.594.044
2	**Amazônia**	2.477.482
3	Equinos	2.384.996
4	Cães	2.023.197
5	**Epidemiologia**	1.964.808
6	**Brasil**	1.380.288
7	Ovinos	1.306.800
8	Caprinos	1.001.817
9	PCR	901.340
10	Nutrição	894.504
11	Milho	875.310
12	Soja	835.354
13	Proteína	823.256
14	Cana-de-Açúcar	760.104
15	Sémen	741.810

Em Ciências Biológicas (Tabela 6.5, 6.6; Apêndice Tabela A.5), também não houve 100% de similaridade nos principais tópicos estudados entre os pares de biênios para o período analisado. Ou seja, os principais interesses dos doutores brasileiros que atuam nas Ciências Biológicas mudaram ao longo do tempo. Contudo, o interesse dos doutores não mudou repentinamente, pois por meio da análise de pares de biênios sequentes nota-se que no mínimo 60% dos principais tópicos estudados no biênio corrente também foram estudados no biênio anterior. À vista disso, vale destacar os períodos 2009-2010 a 2011-2012 e 2011-2012 a 2013-2014, em que o percentual de similaridade foi de 93,33%, havendo divergência apenas quanto às palavras-chave “Proteômica” e “Apoptose”, presentes no primeiro período, e “Genoma” e “Citocina”, no segundo. Pela análise dos biênios não sequentes, é possível observar que 26,66% dos tópicos de pesquisas (4 palavras-chave) evidenciados como os principais no biênio de 2015-2016 também tiveram destaque no biênio 1997-1998. Além disso, nota-se que à medida que os principais tópicos de um determinado biênio são comparados com biênios mais recentes, o percentual de similaridade na maioria das vezes aumenta. Entretanto, pode ser notado que 80% dos principais tópicos que estiveram em evidência no biênio de 2005-2006 continuaram em destaque nos próximos 4 biênios, mostrando a regularidade dos interesses dos doutores em 10 anos de pesquisa.

Em especial, temas relacionados à saúde recebem mais esforços dos doutores das Ciências Biológicas, visto o alto valor de popularidade das palavras-chave “Trypanossoma Cruzi”, “Ratos”, “Estresse Oxidativo”, “Óxido Nítrico”, “Inflamação”, “Câncer”, “Genoma”, “Citocina” e “Epidemiologia”, que representam 60% do conjunto das palavras mais populares. Dentre estas, destaca-se “Trypanossoma Cruzi”, protozoário causador da Doença de Chagas, que esteve no topo do ranque em 50% dos biênios analisados e presente nos outros 50% restantes. Assuntos ligados à natureza também receberam atenção, uma vez que as palavras-chave “Amazônia”, “Taxonomia”, “Cerrado” e “Ecologia” possuem popularidade destacada, provavelmente devido ao interesse da sociedade em questões como preservação ambiental e aquecimento global. Já as palavras-chave “Brasil” e “Conservação” indicam temas de investigação genéricos, por isso aplicados em diversos contextos. Por fim, 66,66% das

principais palavras-chave das Ciências Biológicas do período entre 1997 e 2016 também foram destaques no conjunto completo dos dados (Seção 6.2), sendo que essa grande área contribuiu com cerca de 74% da popularidade total do tópico “Taxonomia”, 48% de “Trypanosoma Cruzi”, 32% de “Óxido Nítrico”, 26% de “Estresse Oxidativo”, 25% de “Inflamação” e em média 13% de “Epidemiologia”, “Amazônia”, “Brasil”, “Ratos” e “Câncer”.

Tabela 6.5: Coeficiente de similaridade das palavras-chave das Ciências Biológicas ranqueadas por TF.FI.

	1997-1998	1999-2000	2001-2002	2003-2004	2005-2006	2007-2008	2009-2010	2011-2012	2013-2014	2015-2016
1997-1998	X	60,00%	40,00%	40,00%	33,33%	40,00%	33,33%	33,33%	33,33%	**26,66%**
1999-2000		X	66,66%	60,00%	53,33%	46,66%	53,33%	46,66%	53,33%	46,66%
2001-2002			X	80,00%	73,33%	66,66%	73,33%	66,66%	66,66%	60,00%
2003-2004				X	80,00%	73,33%	66,66%	66,66%	66,66%	60,00%
2005-2006					X	**80,00%**	**80,00%**	**80,00%**	**80,00%**	73,33%
2007-2008						X	80,00%	80,00%	73,33%	66,66%
2009-2010							X	**93,33%**	86,66%	80,00%
2011-2012								X	**93,33%**	80,00%
2013-2014									X	80,00%
2015-2016										X

Tabela 6.6: As principais palavras-chave das Ciências Biológicas ranqueadas por TF.FI entre 1997 e 2016.

Ranque	Palavras-chave (1997-2016)	TF.FI
1	**Trypanosoma Cruzi**	19.851.264
2	**Amazônia**	15.420.223
3	**Ratos**	14.580.026
4	**Taxonomia**	14.472.090
5	**Brasil**	10.183.510
6	**Estresse Oxidativo**	9.702.869
7	**Óxido Nítrico**	841.1054
8	**Inflamação**	6.548.580
9	**Câncer**	6.285.000
10	Genoma	5.452.656
11	Cerrado	5.178.456
12	Conservação	4.860.870
13	Citocina	4.700.072
14	**Epidemiologia**	4.040.556
15	Ecologia	3.887.695

No contexto das Ciências da Saúde (Tabela 6.7, 6.8; Apêndice Tabela A.6) pode ser observado que os principais interesses dos doutores brasileiros que nelas atuam também mudaram durante o período em análise, devido não existir 100% de similaridade entre os pares de biênios contrapostos. Apesar disso, a partir da análise comparativa entre biênios sequentes é possível notar apenas uma mudança sutil dos principais interesses a cada 2 anos, visto que ao menos 66,66% (10 palavras-chave) dos tópicos centrais de um determinado biênio também foram referência no biênio que o antecede. Assim sendo, vale destacar os pares de biênios 2009-2010 e 2011-2012; 2011-2012 e 2013-2014; 2013-2014 e 2015-2016, em que o percentual de similaridade foi de 86,66%, ou seja, divergindo em apenas (“Tratamento”,

“Insuficiência Cardíaca”) e (“Estresse Oxidativo”, “Diabete Mellitus”) estudadas no primeiro par de biênios, (“Aids”, “Diabete Mellitus”) e (“Brasil”, “Exercício”) no segundo par e (“Diagnóstico”, “Ratos”) e (“Envelhecimento”, “Diabete Mellitus”) no terceiro. Por outro lado, quando analisados os biênios não sequentes, pode ser verificado que no mínimo 40% dos principais tópicos (6 palavras-chave) perduraram por mais de dois anos em evidência, e que, à medida que os principais tópicos de um determinado biênio são comparados com biênios mais distantes temporalmente, o percentual de similaridade sempre diminui ou se mantém, mostrando a troca de interesses por parte dos doutores de forma gradativa ao longo do tempo. Entretanto, pode ser notado que 66,66% dos principais tópicos que estiveram em evidência no biênio de 1999-2000 continuaram em destaque nos próximos 4 biênios, mostrando a estabilidade de parte dos principais interesses dos doutores em 10 anos de pesquisa.

Naturalmente, 100% dos principais tópicos ranqueados entre 1997 e 2016 nessa grande área são relacionados ao contexto da saúde, mesmo que não literalmente, como é o caso de “Crianças”, “Idoso”, “Adolescente” e “Brasil”. Dentre os principais tópicos, destacam-se “Epidemiologia”, “Diabetes” e “Diabetes Mellitus”, sendo que “Epidemiologia” é a primeira do ranque e possui popularidade cerca de 5 vezes maior que a segunda palavra-chave. Além disso, “Epidemiologia” esteve no topo em popularidade em todos os biênios analisados ao longo dos 20 anos de pesquisa. Em contrapartida, as palavras-chave “Diabetes” e “Diabetes Mellitus” demonstram a grande atenção dedicada a esse tema pelos doutores que atuam nas Ciências da Saúde. O tópico “Diabetes” obteve destaque em todos os biênios, e “Diabetes Mellitus” apareceu em 4 biênios ao longo do período em análise. Por fim, 53,33% das principais palavras-chave das Ciências da Saúde entre 1997 e 2016 também foram destaques no conjunto completo dos dados (Seção 6.2), sendo que essa grande área contribuiu com cerca de 50% da popularidade total do tópico “Obesidade”, 44% de “Epidemiologia”, 42% de “HIV”, 33% de “Diabetes”, 30% de “Diagnóstico” e em média 17% de “Brasil”, “Ratos” e “Câncer”.

Tabela 6.7: Coeficiente de similaridade das palavras-chave das Ciências da Saúde ranqueadas por TF.FI.

	1997-1998	**1999-2000**	**2001-2002**	**2003-2004**	**2005-2006**	**2007-2008**	**2009-2010**	**2011-2012**	**2013-2014**	**2015-2016**
1997-1998	X	66,66%	73,33%	80,00%	66,66%	60,00%	60,00%	60,00%	46,66%	46,66%
1999-2000		X	**66,66%**	**66,66%**	**66,66%**	**66,66%**	60,00%	53,33%	46,66%	**40,00%**
2001-2002			X	80,00%	66,66%	66,66%	66,66%	60,00%	53,33%	46,66%
2003-2004				X	66,66%	66,66%	66,66%	60,00%	53,33%	46,66%
2005-2006					X	73,33%	66,66%	66,66%	60,00%	53,33%
2007-2008						X	80,00%	80,00%	80,00%	66,66%
2009-2010							X	**86,66%**	80,00%	73,33%
2011-2012								X	**86,66%**	86,66%
2013-2014									X	**86,66%**
2015-2016										X

Tabela 6.8: As principais palavras-chave das Ciências da Saúde ranqueadas por TF.FI entre 1997 e 2016.

Ranque	Palavras-chave (1997-2016)	TF.FI
1	**Epidemiologia**	86.670.856
2	**Obesidade**	16.482.438
3	**Câncer**	16.119.070
4	**Ratos**	11.951.712
5	**Diagnóstico**	11.548.350
6	**HIV**	11.351.275
7	**Diabetes**	9.917.658
8	Crianças	936.7995
9	Aids	896.5494
10	Idoso	719.0760
11	Tratamento	719.0760
12	Adolescente	634.2382
13	Enfermagem	612.5238
14	**Brasil**	601.0400
15	Diabete Mellitus	468.1145

Na grande área Ciências Exatas e da Terra (Tabela 6.9, 6.10; Apêndice Tabela A.7), por sua vez, pode ser observado que os principais interesses dos doutores brasileiros que atuam nessa grande área também mudaram ao longo do tempo. No entanto, a partir da análise comparativa entre biênios sequentes, é possível notar que boa parte dos principais tópicos de pesquisas estudados no biênio atual foram relevantes no biênio anterior, destacando-se: (1) uma mudança de 46,66% para 73,33% de similaridade entre os biênios de 1999-2000 a 2001-2002; (2) a similaridade de 73,33% (11 palavras-chave) entre os pares de biênios 1999-2000 e 2001-2002; 2001-2002 e 2003-2004; 2011-2012 e 2013-2014; (3) com exceção dos dois primeiros biênios, a similaridade dos conjuntos principais varia em apenas um único tópico a cada 2 anos. Pela comparação entre biênios não sequentes, é possível observar que 20% dos tópicos de pesquisas evidenciados como os principais no biênio de 2015-2016 também foram destaque nos 3 primeiros biênios. Nesse caso, as palavras-chave são “Cosmologia”, “Estrutura Cristalina” e “Amazônia”, destacando-se “Estrutura Cristalina”, que, diferentemente das outras duas palavras, que foram destaques em todos os biênios analisados, figurou no topo do ranque nos biênios de 2005-2006 e 2007-2008, e só não foi destaque no biênio de 2013-2014.

Assuntos possivelmente relacionados à quimica e à física são de interesse dos doutores brasileiros que atuam na grande área das Ciências Exatas e da Terra, visto que, com exceção das palavras-chave “Amazônia” e “Cosmologia”, todos os 13 tópicos principais restantes referem-se diretamente a tais áreas. Apesar disso, vale destacar que “Cosmologia” pode referir-se tanto à astronomia quanto à física. Já o destaque da palavra-chave “Amazônia” pode dever-se ao fato de ela representar o contexto dos estudos de parte dos outros temas em evidência. Por fim, apenas 6,66% das principais palavras-chave das Ciências Exatas e da Terra do período de 1997-2016 também foram destaques no conjunto completo dos dados (Seção 6.2), sendo que essa grande área contribuiu com cerca de 3,3% da popularidade total do tópico “Amazônia”.

Tabela 6.9: Coeficiente de similaridade das palavras-chave das Ciências Exatas e da Terra por TF.FI.

	1997-1998	1999-2000	2001-2002	2003-2004	2005-2006	2007-2008	2009-2010	2011-2012	2013-2014	2015-2016
1997-1998	X	46,66%	33,33%	33,33%	40,00%	26,66%	33,33%	26,66%	20,00%	**20,00%**
1999-2000		X	**73,33%**	73,33%	53,33%	53,33%	40,00%	20,00%	13,33%	**20,00%**
2001-2002			X	**73,33%**	46,66%	60,00%	46,66%	26,66%	13,33%	**20,00%**
2003-2004				X	66,66%	66,66%	60,00%	40,00%	20,00%	26,66%
2005-2006					X	66,66%	66,66%	53,33%	33,33%	33,33%
2007-2008						X	66,66%	46,66%	33,33%	46,66%
2009-2010							X	66,66%	46,66%	60,00%
2011-2012								X	**73,33%**	53,33%
2013-2014									X	66,66%
2015-2016										X

Tabela 6.10: As principais palavras-chave das Ciências Exatas e da Terra por TF.FI entre 1997 e 2016.

Ranque	Palavras-chave (1997-2016)	TF.FI
1	**Amazônia**	3.263.260
2	Cosmologia	2.400.428
3	Biodiesel	2.374.000
4	Estrutura Cristalina	2.090.760
5	Nanopartícula	1.928.646
6	Magnetismo	1.809.632
7	Adsorção	1.563.840
8	Filme Fino	1.466.150
9	Flavonoides	1.298.088
10	Fotoluminescência	1.207.152
11	Sol-Gel	1.156.730
12	Catalisador	1.119.410
13	Óleo Essencial	1.079.670
14	Espectroscopia	986.592
15	Mercúrio	959.438

A grande área Ciências Humanas (Tabela 6.11, 6.12; Apêndice Tabela A.8) também não apresentou durante os 20 anos de pesquisas em análise 100% de similaridade nos principais tópicos estudados entre os pares de biênios. Isso mostra que, a cada 2 anos, os principais interesses dos doutores brasileiros que atuam nessa grande têm mudado por algum motivo. Entretanto, ao analisar os pares de biênios sequentes, nota-se que, em média, 57% dos principais tópicos estudados em um biênio corrente também foram estudados no biênio anterior. Nesse caso, vale destacar os pares de biênios 2009-2010 a 2011-2012 e 2011-2012 a 2013-2014, em que o maior percentual de similaridade foi de 66,66%, havendo divergências apenas entre ("Trabalho", "Políticas Públicas", "Educação a Distância", "Ansiedade", "Psicanálise") e ("Ciência", "Violência", "Crianças", "Saúde", "Autismo"), estudadas no primeiro par de biênios, e ("Ciência", "Gênero", "Crianças", "Autismo", "Cognição") e ("Idoso", "Infância", "Ansiedade", "América Latina", "Formação do Professor"), no segundo par. Pela análise comparativa entre biênios não sequentes, é possível observar que 26,66% dos principais tópicos (4 palavras-chave) que estiveram em evidência, no biênio de 1997-1998,

continuaram em evidência nos 5 últimos biênios, destacando-se as palavras-chave “Cultura”, “Memória” e “Brasil”, que figuraram entre as principais durante o período ressaltado.

No contexto das Ciências Humanas percebe-se uma atenção especial dos doutores brasileiros a temas relacionados à educação, à saúde e ao bem-estar do ser humano, possivelmente representada pelo estudo das palavras-chave “Educação”, “Psicologia”, “Amazônia”, “Violência”, “Epidemiologia”, “Ansiedade”, “Trabalho”, “Gênero”, “Adolescente”, “Ciência” e “Crianças”. Além disso, também se destacam temas referentes à cultura e à história, como no caso das palavras-chave “Cultura”, “Arqueologia” e “Memória”. Por fim, 20% das principais palavras-chave das Ciências Humanas do período 1997-2016 também foram destaques no conjunto completo dos dados (Seção 6.2), sendo que essa grande área contribuiu em média com 0,5% da popularidade total dos tópicos “Epidemiologia”, “Amazônia” e “Brasil”.

Tabela 6.11: Coeficiente de similaridade das palavras-chave das Ciências Humanas ranqueadas por TF.FI.

	1997-1998	**1999-2000**	**2001-2002**	**2003-2004**	**2005-2006**	**2007-2008**	**2009-2010**	**2011-2012**	**2013-2014**	**2015-2016**
1997-1998	X	46,66%	33,33%	46,66%	46,66%	**26,66%**	**26,66%**	**26,66%**	**26,66%**	**26,66%**
1999-2000		X	46,66%	46,66%	53,33%	46,66%	46,66%	40,00%	40,00%	40,00%
2001-2002			X	53,33%	53,33%	46,66%	46,66%	40,00%	40,00%	46,66%
2003-2004				X	60,00%	40,00%	60,00%	46,66%	33,33%	40,00%
2005-2006					X	60,00%	46,66%	46,66%	46,66%	46,66%
2007-2008						X	53,33%	53,33%	66,66%	46,66%
2009-2010							X	**66,66%**	60,00%	60,00%
2011-2012								X	**66,66%**	46,66%
2013-2014									X	60,00%
015-2016										X

Tabela 6.12: As principais palavras-chave das Ciências Humanas ranqueadas por TF.FI entre 1997 e 2016.

Ranque	**Palavras-chave (1997-2016)**	**TF.FI**
1	Educação	676.848
2	Memória	670.145
3	**Amazônia**	547.800
4	**Brasil**	533.994
5	Cultura	242.292
6	Gênero	222.365
7	Ansiedade	185.808
8	Psicologia	168.480
9	Trabalho	166.968
10	Adolescente	157.353
11	Violência	156.604
12	Arqueologia	153.272
13	**Epidemiologia**	140.140
14	Ciência	136.500
15	Crianças	119.880

A grande área Ciências Sociais Aplicadas (Tabela 6.13, 6.14; Apêndice Tabela A.9) também não apresentou 100% de similaridade nos principais tópicos estudados entre os pares de biênios. Isso mostra que a cada 2 anos os principais interesses dos doutores brasileiros que atuam nessa grande área têm mudado por algum motivo. Contudo, ao analisar os pares de biênios sequentes, nota-se que parte dos principais tópicos de pesquisas estudados no biênio atual foram relevantes no biênio anterior, destacando-se: (1) uma mudança brusca de interesse por parte dos doutores nos dois primeiros biênios, em que 14 dos 15 principais tópicos estudados no biênio de 1999-2000 não tiveram destaque em 1997-1998; (2) os períodos de 2009-2010 e 2011-2012, em que o percentual de similaridade foi de 60,00%, havendo divergência apenas quanto às palavras-chave "Biodiesel", "Comunicação", "Governo", "Agronegócio", "Estratégia" e "Sustentabilidade", que tiveram popularidade destacada em 2011-2012; enquanto que, "Ciência da Informação", "Bibliometria", "Turismo", "Internet", "Educação" e "Saúde", que foram destaque em 2009-2010. Ao analisar o coeficiente de similaridade dos tópicos de pesquisas entre os biênios não sequentes, foi possível verificar que somente a palavra-chave "Amazônia" permaneceu em evidência em todos os biênios.

Notadamente, a palavra-chave "Brasil" (tema de investigação genérico) tem destaque ao longo do tempo na grande área Ciências Sociais Aplicadas, haja vista seu alto valor de popularidade e devido ter figurado entre as principais palavras-chave em praticamente todos os biênios, exceto no período de 1999-2000. Isso indica que historicamente os doutores que atuam na grande área Ciências Sociais Aplicadas têm dedicado seus esforços a pesquisas de interesse do país. As palavras-chave "Amazônia", "Sustentabilidade" e "Consumismo", possivelmente relacionadas a temas que envolvem problemas ambientais, contaram também com uma atenção especial por partes desses doutores. Os assuntos que abrangem temas relacionados a Comunicação, *Marketing* e Vendas também recebem grande atenção, como pode ser notado pela presença das palavras-chave "Marketing", "Comunicação", "Inovação", "Estratégia" e "Design", que estão entre as mais populares considerando todo o período de análise dos dados. A utilização dessas palavras-chave pode estar relacionada com a modernização progressiva do Marketing desde a década de 90, relacionada à evolução e à popularização da internet, das redes sociais e dos dispositivos móveis. Também vale destacar temas que podem estar relacionados a mais de uma linha de estudo, como é o caso da palavra-chave "Gênero", que pode referir-se a estudos que tentam compreender as relações de gênero na cultura e sociedade humanas ou gêneros gramaticais.

As palavras-chave "Saúde", "Educação" e "Políticas Públicas" são consideradas temas centrais para o desenvolvimento de qualquer nação, e por isso também receberam grande dedicação dos doutores que atuam na grande área Ciências Sociais Aplicadas. Em contrapartida, a popularidade das palavras-chave "Bibliometria" e "Gestão do Conhecimento" deve-se certamente ao fato de a Ciência da Informação ser, ou ter sido até pouco tempo atrás, uma das áreas da grande área Ciências Sociais Aplicadas. Por fim, 13,33% das principais palavras-chave do período 1997-2016 das Ciências Sociais Aplicadas também foram destaques no conjunto completo dos dados (Seção 6.2), sendo que essa grande área contribuiu em média com 2% da popularidade total das palavras-chave "Amazônia" e "Brasil".

Tabela 6.13: Coeficiente de similaridade das palavras-chave das Ciências Sociais Aplicadas por TF.FI.

	1997-1998	1999-2000	2001-2002	2003-2004	2005-2006	2007-2008	2009-2010	2011-2012	2013-2014	2015-2016
1997-1998	X	**6,66%**	20,00%	13,33%	13,33%	26,66%	40,00%	26,66%	33,33%	26,66%
1999-2000		X	**6,66%**	6,66%	20,00%	6,66%	13,33%	13,33%	6,66%	**6,66%**
2001-2002			X	26,66%	20,00%	33,33%	20,00%	33,33%	26,66%	26,66%
2003-2004				X	26,66%	40,00%	20,00%	20,00%	26,66%	20,00%
2005-2006					X	33,33%	26,66%	26,66%	20,00%	40,00%
2007-2008						X	46,66%	53,33%	53,33%	46,66%
2009-2010							X	**60,00%**	40,00%	46,66%
2011-2012								X	46,66%	46,66%
2013-2014									X	53,33%
2015-2016										X

Tabela 6.14: As principais palavras-chave das Ciências Sociais Aplicadas por TF.FI entre 1997 e 2016.

Ranque	Palavras-chave (1997-2016)	TF.FI
1	**Brasil**	176.349
2	**Amazônia**	98.464
3	Marketing	84.182
4	Inovação	79.648
5	Comunicação	54.950
6	Sustentabilidade	49.140
7	Bibliometria	39.900
8	Educação	35.460
9	Estratégia	33.840
10	Saúde	31.450
11	Design	29.512
12	Políticas Públicas	28.960
13	Gênero	27.864
14	Gestão do Conhecimento	26.841
15	Consumismo	21.917

As Engenharias (Tabela 6.15, 6.16; Apêndice Tabela A.10) também não apresentaram 100% de similaridade nos principais tópicos estudados entre os pares de biênios. Ou seja, a cada 2 anos os principais interesses dos doutores brasileiros que atuam nessa grande área mudaram. Contudo, ao analisar os pares de biênios sequentes, nota-se que em média 71% dos principais tópicos de pesquisas estudados no biênio atual foram relevantes no biênio anterior, destacando os pares de biênios 2005-2006 a 2007-2008; 2007-2008 a 2009-2010; 2011-2012 a 2013-2014 e 2013-2014 a 2015-2016, em que o percentual de similaridade foi de 86,66% (13 palavras-chave), havendo divergência, portanto, em apenas 2 dos 15 principais tópicos a cada 4 anos de estudo. Ao analisar os biênios não sequentes, é possível observar que 26,66% dos tópicos de pesquisas (4 palavras-chave) evidenciados como os principais no biênio de 2015-2016 também tiveram destaque no biênio 1997-1998 e, à medida que os principais tópicos de um determinado biênio são comparados com biênios mais recentes, o percentual de similaridade geralmente aumenta, mostrando que a alteração dos principais interesses dos doutores alternaram linearmente ao longo do tempo.

Dentre os principais tópicos elencados nas Engenharias entre 1997 e 2016, cerca de 86,66% (13 palavras-chave) dizem respeito a temas relacionados a engenharia dos materiais, sendo que 30% desses tópicos se referem a assuntos voltados para o desenvolvimento sustentável: “Biomassa”, “Biodiesel”, “Lipase” e “Biomateriais”. Nesse ponto, destaca-se “Biodiesel”, que apareceu pela primeira vez entre as principais palavras-chave no biênio 2007-2008 e atingiu o topo em popularidade em todos os biênios seguintes. Ainda é possível observar as palavras-chave “Modelagem” e “Otimização”, que possivelmente se relacionam à melhoria dos processos em um determinado domínio. Por fim, vale destacar que nenhuma das principais palavras-chave das Engenharias do período de 1997-2016 foram destaques no conjunto completo dos dados apresentados na Seção (6.2).

Tabela 6.15: Coeficiente de similaridade das palavras-chave das Engenharias ranqueadas por TF.FI.

	1997-1998	1999-2000	2001-2002	2003-2004	2005-2006	2007-2008	2009-2010	2011-2012	2013-2014	2015-2016
1997-1998	X	46,66%	46,66%	40,00%	40,00%	40,00%	40,00%	26,66%	33,33%	**26,66%**
1999-2000		X	66,66%	66,66%	60,00%	53,33%	53,33%	33,33%	46,66%	40,00%
2001-2002			X	46,66%	53,33%	46,66%	40,00%	26,66%	40,00%	33,33%
2003-2004				X	66,66%	66,66%	66,66%	46,66%	60,00%	53,33%
2005-2006					X	**86,66%**	73,33%	53,33%	66,66%	53,33%
2007-2008						X	**86,66%**	66,66%	80,00%	66,66%
2009-2010							X	66,66%	80,00%	66,66%
2011-2012								X	**86,66%**	73,33%
2013-2014									X	**86,66%**
2015-2016										X

Tabela 6.16: As principais palavras-chave das Engenharias ranqueadas por TF.FI entre 1997 e 2016.

Ranque	Palavras-chave (1997-2016)	TF.FI
1	Biodiesel	1.891.960
2	Compósitos	1.396.200
3	Biomateriais	1.057.614
4	Cerâmicas	798.950
5	Microestrutura	723.095
6	Nanocompósitos	693.666
7	Modelagem	663.884
8	Propriedade Mecânica	634.848
9	Otimização	591.384
10	Adsorção	556.250
11	Filme fino	492.600
12	Lipase	420.486
13	Biomassa	407.037
14	Nanopartícula	353.363
15	Corrosão	347.422

A grande área Linguística, Letras e Artes (Tabela 6.17, 6.18; Apêndice Tabela A.11) apresentou um baixo percentual de similaridade entre os principais tópicos estudados entre os pares de biênios, sendo 33,33% o maior índice. Além disso, pode ser notado que, em vários pares de biênios, os principais interesses dos doutores brasileiros que atuam nessa grande área

mudaram em sua totalidade (ou seja, 0% de similaridade), principalmente ao analisar pares de biênios não sequentes. Nesse sentido, vale destacar que todos os principais tópicos estudados no primeiro biênio perderam evidência a partir de 2007. Por outro lado, quando analisados os biênios sequentes, nota-se que em média 16% dos principais tópicos de pesquisas estudados no biênio atual foram relevantes no biênio anterior, destacando-se (1) 0% de similaridade entre os pares dos 3 primeiros biênios e (2) 33,33% de similaridade entre os pares de biênios 2003-2004 a 2005-2006 e 2011-2012 a 2013-2014.

As palavras-chave evidenciadas como as mais importantes entre 1997 e 2016 mostram que os doutores brasileiros que atuam na grande área Linguística, Letras e Artes dedicaram esforços em apenas um único tópico referente a Artes, no caso "Música". Esta palavra-chave apareceu entre as principais pela primeira vez no biênio 2011-2012 e permaneceu em evidência somente até o biênio seguinte. Por fim, vale destacar que nenhuma das principais palavras-chave do período 1997-2016 da grande área Línguistica, Letras e Artes foi destaque no conjunto completo dos dados apresentados na Seção (6.2).

Tabela 6.17: Coeficiente de similaridade das palavras-chave da Linguística, Letras e Artes por TF.FI.

	1997-1998	1999-2000	2001-2002	2003-2004	2005-2006	2007-2008	2009-2010	2011-2012	2013-2014	2015-2016
1997-1998	X	**0,00%**	13,33%	20,00%	13,33%	**0,00%**	**0,00%**	**0,00%**	**0,00%**	**0,00%**
1999-2000		X	**0,00%**	0,00%	0,00%	0,00%	6,66%	6,66%	6,66%	6,66%
2001-2002			X	6,66%	6,66%	0,00%	6,66%	20,00%	13,33%	**0,00%**
2003-2004				X	**33,33%**	6,66%	0,00%	0,00%	6,66%	**0,00%**
2005-2006					X	13,33%	6,66%	13,33%	13,33%	6,66%
2007-2008						X	20,00%	13,33%	6,66%	6,66%
2009-2010							X	26,66%	6,66%	6,66%
2011-2012								X	**33,33%**	6,66%
2013-2014									X	13,33%
2015-2016										X

Tabela 6.18: As principais palavras-chave das Linguística, Letras e Artes por TF.FI entre 1997 e 2016.

Ranque	Palavras-chave (1997-2016)	TF.FI
1	Memória	15.000
2	Cultura	11.407
3	Discurso	7.368
4	Tradução	6.670
5	Identidade	6.076
6	Linguagem	5.530
7	Bilinguismo	5.348
8	Literatura	5.254
9	Música	5.202
10	Psicolinguística	4.321
11	Poesia	3.298
12	Fonologia	2.688
13	Narrativa	2.640
14	Violência	2.493
15	Escrita	2.133

Por fim, existem assuntos centrais que são tratados com destaque por diferentes grandes áreas do conhecimento. Com isso em vista, para facilitar a visualização dos tópicos que possuem popularidade destacada em mais de uma grande área, a Tabela 6.19 apresenta o coeficiente de similaridade das principais palavras-chave dos artigos publicados pelos doutores brasileiros em periódicos no período de 1997 a 2016 entre os pares de grandes áreas.

Tabela 6.19: Coeficiente de similaridade das principais palavras-chave por grandes áreas entre 1997 e 2016.

	Ciências Agrárias	**Ciências Biológicas**	**Ciências da Saúde**	**Ciências Exatas e da Terra**	**Ciências Humanas**	**Ciências Sociais Aplicadas**	**Engenharias**	**Linguística, Letras e Artes**
Ciências Agrárias	X	20,00%	13,33%	6,66%	20,00%	13,33%	0%	0%
Ciências Biológicas		X	**26,66%**	6,66%	20,00%	13,33%	0%	0%
Ciências da Saúde			X	0%	**26,66%**	6,66%	0%	0%
Ciências Exatas e da Terra				X	6,66%	6,66%	**20,00%**	0%
Ciências Humanas					X	**26,66%**	0%	**20,00%**
Ciências Sociais Aplicadas						X	0%	0%
Engenharias							X	0%
Linguística, Letras e Artes								X

Pelos resultados da tabela, verifica-se que cada uma das grandes áreas “Engenharias” e “Linguística, Letras e Artes” possui similaridade com apenas uma única outra grande área: “Ciências Exatas e da Terra” e “Ciências Humanas”, respectivamente. Em contrapartida, ao menos um par das outras grandes áreas restantes se relacionam entre si quanto ao interesse em pelo menos um tópico de pesquisa. Dentre estas, destaca-se a grande área “Ciências Humanas”, que possui o maior número de relações: “Ciências Agrárias” (“Amazônia”, “Epidemiologia”, “Brasil”); “Ciências Biológicas” (“Amazônia”, “Epidemiologia”, “Brasil”); “Ciências da Saúde” (“Crianças”, “Adolescentes”, “Epidemiologia”, “Brasil”); “Ciências Exatas e da Terra” (“Amazônia”); “Ciências Sociais Aplicadas” (“Amazônia”, “Gênero”, “Brasil”, “Educação”), e “Linguística, Letras e Artes” (“Memória”, “Violência”, “Cultura”). Nesse sentido, vale destacar que, as grandes áreas “Ciências Biológicas”, “Ciências Humanas” e “Ciências Agrárias” dedicam grandes esforços aos tópicos “Amazônia”, “Epidemiologia” e “Brasil”. No entanto, quando observada a intensidade do relacionamento, evidenciam-se os seguintes pares de grandes áreas, com 26,66% de semelhança entre si: (1) “Ciências Biológicas” e “Ciências da Saúde”, que têm em comum os tópicos “Câncer”, “Epidemiologia”, “Ratos” e “Brasil”; (2) “Ciências da Saúde” e “Ciências Humanas”, que se vinculam pelos temas “Crianças”, “Adolescentes”, “Epidemiologia” e “Brasil”; (3) “Ciências Humanas” e ”Ciências Sociais Aplicadas”, que discutem amplamente os temas “Amazônia”, “Brasil”, “Educação” e “Gênero”.

CONSIDERAÇÕES FINAIS

Este Capítulo apresenta as considerações finais desta tese (Seção 7.1), as publicações e submissões resultantes do trabalho realizado (Seção 7.2) e, para finalizar, indicações referentes a alguns dos possíveis trabalhos futuros (Seção 7.3).

7.1 Considerações Finais

Foi verificado que o grande aumento no número de publicações científicas disponíveis na web tem tornado a bibliometria um objeto de estudo para um crescente número de pesquisadores de todas as áreas do conhecimento, interessados em compreender o que a comunidade científica tem produzido. Tal entendimento serve principalmente como auxílio para a tomada de decisão em diversos níveis, como: (1) em um nível operacional, o pesquisador que deseja iniciar uma nova pesquisa pode escolher assuntos que estão em foco ou que tendem a ganhar repercussão num curto espaço de tempo, e (2) em um nível estratégico, essa compreensão é primordial para os governos, negócios e fundações, que devem decidir sobre investimentos para a construção de políticas científicas e ter competências e insumos para analisar o patamar científico das nações e/ou grupos individualizados, seja para impulsionar resultados sobre determinada pesquisa, seja para iniciar estudos sobre novos temas.

Nesse contexto, esta tese apresenta o desenvolvimento de um estudo inédito para identificar e analisar temporalmente o comportamento dos principais tópicos de pesquisas de interesse da comunidade científica brasileira. Para tanto, utiliza as palavras-chave de quase 14 milhões de artigos científicos publicados pelos doutores em anais de congressos e periódicos cadastrados na base curricular da Plataforma Lattes. Esses artigos possuem datas de publicação que vão de 1962 a 2016, representando os últimos 55 anos da história do desenvolvimento da ciência brasileira. As principais motivações para o desenvolvimento desta tese foram: (1) a relevância do repositório de dados da Plataforma Lattes para o estudo do desenvolvimento da ciência no Brasil; (2) o potencial das palavras-chave dos artigos para a descrição dos assuntos principais que os permeiam; (3) a inexistência de estudos amplos sobre as palavras-chave dos artigos cadastrados nos currículos da Plataforma Lattes, e (4) a inexistência de um estudo abrangente com foco na compreensão de quais os tópicos desenvolvidos pelos pesquisadores brasileiros.

Os resultados apresentados foram alcançados a partir de análises bibliométricas e baseados em técnicas para análises de redes sociais, aplicadas sobre as palavras-chave de artigos publicados em anais de congressos e em periódicos pelos doutores que possuem currículos cadastrados na Plataforma Lattes. Nesse ponto, vale destacar que o número de doutores representa cerca de 5% da quantidade total de indivíduos da Plataforma Lattes, porém são eles os responsáveis por mais de 70% de todos os artigos cadastrados no total de currículos, fornecendo, assim, um considerável conjunto de dados para as análises desta pesquisa. Para tanto, após aquisição dos currículos dos doutores pelo *LattesDataXplorer*, fez-se necessário a

construção dos componentes responsáveis pela filtragem e tratamento das palavras-chave dos artigos, para a padronização e definição do conjunto de arquivos que contêm os dados essenciais para o estudo e concomitantemente, diminuir o processamento computacional para a geração dos resultados desejados. Nesse sentido, com o propósito de possibilitar a customização das análises e assim torná-las mais abrangentes e coesas ao contexto, foi necessária a consideração de duas visões sobre os dados extraídos: (1) a **visão geral**, que contabiliza todos os artigos e palavras-chave extraídos dos currículos e, (2) a **visão colaboração**, que considera apenas um único artigo publicado em coautoria de mesma grande área do conhecimento associado ao conjunto união de suas respectivas palavras-chave.

Pelos resultados encontrados no Capítulo 5, foi possível apresentar uma caracterização geral das palavras-chave utilizadas pelos doutores brasileiros, o que possibilitou a verificação de diversas informações. Como, por exemplo, que o maior número de palavras-chave não ocorreu no mesmo ano em que houve o maior número de publicações, implicando que não há garantia de uma relação bem definida entre número de palavras-chave e número de artigos. Foi identificado que tal acontecimento está relacionado ao cadastramento de artigos científicos com baixo número de palavras-chave, em que, com base na visão geral, 47,6% em congressos e 42,7% em periódicos das publicações não possui sequer uma palavra-chave. Já com base na visão colaboração, esse percentual se reduz para 40,3% em congressos e 31,8% em periódicos, devido ao fato de alguns dos autores que realizaram trabalhos em coautoria terem inserido palavras-chave e outros não. Essa informação é muito relevante, pois um percentual considerável de publicações sem palavras-chave vinculadas pode ser um limitador dependendo do tipo de análise a ser realizada. Outra informação destacada pela visão colaboração foi que a quantidade de palavras-chave associadas a determinados artigos é maior que a capacidade permitida pela Plataforma Lattes. Isso se deve ao fato de diferentes autores terem cadastrado palavras-chave diversas para o mesmo artigo publicado em coautoria. Também foi notada uma preferência dos doutores pelo preenchimento dos dados de artigos em periódicos, pois a média de palavras-chave por artigo em periódico é superior à média por artigo em anais de congressos para praticamente todo o período analisado. Destaca-se que nos últimos anos a média de palavras-chave por artigo em periódico tem crescido consecutivamente, atingindo 3,52 palavras-chave (maior valor registrado) em 2016, enquanto a média por artigo em congressos tem oscilado em torno de 3,0.

Além disso, ao comparar as palavras-chave com seus respectivos títulos das publicações, foi verificado que apenas 33,3% das palavras-chave em congressos e 28,6% em periódicos estão contidas nos títulos. Essa informação é muito importante para orientação de pesquisas futuras, pois a análise de termos extraídos dos títulos é uma estratégia amplamente utilizada na literatura para estudar tópicos de pesquisas, e, nesse caso, esses trabalhos podem não estar considerando de forma efetiva o conteúdo dos documentos analisados. Em continuação, percebeu-se a necessidade de um estudo com relação à frequência sobre o conjunto das palavras-chave, visto que 98,6% delas são utilizadas menos de 100 vezes, enquanto as outras poucas palavras-chave são extremamente referenciadas.

Posteriormente ao ranqueamento dos principais tópicos de pesquisas em cada quinquênio, foi possível verificar que grande parte dos tópicos pesquisados pelos doutores em um quinquênio atual também foi de interesse no quinquênio anterior. Assim, a evolução dos

principais interesses dos doutores aconteceu gradualmente ao longo dos anos, o que possivelmente é ocasionado pela demanda ou maturação de um determinado tema de pesquisa. Contudo, foi possível identificar que entre 2007 e 2011 não se estudou nenhum tema que foi destaque entre 1962 e 1972 e que, no quinquênio seguinte (2012-2016), foi registrado entre os principais tópicos o ressurgimento de um tema que foi destaque nos três primeiros quinquênios. Logo, observa-se a possibilidade de um estudo mais detalhado para mapear e evidenciar as trocas de determinados tópicos de pesquisas desenvolvidos pelos doutores ao longo do tempo.

No âmbito das grandes áreas do conhecimento, tem-se "Ciências Agrárias" como a responsável pelo maior número de palavras-chave em congressos (cerca de 18%) e "Ciências da Saúde" em periódicos (cerca de 23%). Em contrapartida, o menor número de palavras-chave está relacionado à grande área "Outra", com apenas 0,6%. Esse baixo percentual, além de contemplar as áreas pré-definidas pela Plataforma Lattes, também contém os dados de artigos que não possuíam grande área atribuída. Em seguida, foi observado que as grandes áreas que possuem mais interesses em comum são "Linguística, Letras e Artes" e "Ciências Humanas", que se assemelham em 20% dos principais tópicos estudados. Por outro lado, a grande área "Ciências Agrárias" não possui relação com nenhuma outra grande área. A partir disso, nota-se a possibilidade de um estudo temporal detalhado para mapear a variação dos principais tópicos de pesquisas de cada grande área ao longo do tempo.

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. Esse processo foi mais intenso na rede de 2002-2006, período de maior crescimento da produção científica brasileira, em que foi registrado o maior número de arestas das redes e, consequentemente, o maior grau médio, o maior valor de densidade, o menor diâmetro e o menor caminho mínimo médio.

Comparando as 15 principais palavras-chave ranqueadas pelas duas medidas de importância de tópicos, verificou-se unanimidade quanto a "Educação" ser a palavra mais relevante do conjunto. No entanto, destaca-se a relação entre as medidas de grau e frequência, as quais foram similares em 73% para o conjunto das 15 principais palavras-chave. Porém, a comprovação da relação direta entre grau e frequência se confirmou através do alto valor de correlação ("0,9572") considerando-se todo o conjunto de análise. Diante disso, constata-se que a utilização do grau como medida de importância é uma alternativa interessante à frequência. Além disso, percebe-se que, ao analisar a co-ocorrência das palavras-chave, foi possível identificar que "Germinação-Sementes" co-ocorreram o maior número de vezes (3.191), e não foram destacadas entre as palavras-chave mais relevantes pelas medidas de importâncias utilizadas. Como resultado desses primeiros estudos, publicações foram

produzidas, submetidas e aceitas, tanto em periódicos quanto em anais de congressos. Tais publicações são detalhadas na próxima seção.

Apesar da constatação da validade dos resultados apresentados no Capítulo 5 para auxiliar o entendimento sobre o desenvolvimento da ciência brasileira, 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 não explorar em completude as características existentes nos dados. À Vista disso, nesta tese foi desenvolvida a medida de importância TF.FI, que é capaz de levar em consideração tanto as características quantitativas das palavras-chave dos artigos quanto as características qualitativas dos periódicos em que tais artigos foram publicados.

Em continuação, foi realizada uma análise comparativa entre os resultados da aplicação das medidas de frequência (TF), grau e TF.FI no conjunto das palavras-chave dos artigos publicados em periódicos entre 1997 e 2016 com a finalidade de verificar qual a melhor medida para considerar as características existentes nos dados da base curricular da Plataforma Lattes. Assim, por meio da interseção entre as 100 primeiras palavras-chave ranqueadas por cada medida de importância, foi possível verificar que 54 palavras-chave são comuns entre as três medidas utilizadas. Além disso, foi observado que: (1) 85% dos conjuntos de TF e grau são similares; e (2) a TF.FI gerou um conjunto mais diversificado em relação às demais medidas. A comprovação da relação direta entre TF e grau se deu pelo seu alto valor de correlação (“0,9331”), contra (“0,5200”) entre TF e TF.FI para todo conjunto. Diante disso, constata-se que a utilização da TF.FI se mostra uma alternativa promissora às medidas de frequência e grau, visto que considera as características intrínsecas dos dados referentes ao histórico de publicações em periódicos de diferentes níveis de qualidade.

Após o ranqueamento dos principais tópicos de pesquisas em cada biênio por meio da aplicação da medida TF.FI no conjunto das palavras-chave dos artigos publicados em periódicos entre 1997 e 2016, foi possível verificar que 73,33% dos principais interesses dos doutores brasileiros estão relacionados à saúde, haja vista que o tópico “Epidemiologia” foi o primeiro do ranque em todos os biênios analisados. Observou-se que apesar dos principais tópicos terem mudado a cada 2 anos, boa parte deles perdurou em evidência por mais de um biênio, comportamento similar ao encontrado nas grandes áreas do conhecimento, exceto em Linguística, Letras e Artes, que apresentou mudança completa dos principais interesses em diversos pares de biênios. Também nota-se que a grande área Ciências Biológicas apresentou o maior número de palavras-chave similares ao conjunto completo dos dados, no total de 10 das 15 palavras-chave. Em contrapartida, as grandes áreas Engenharias e Linguísticas, Letras e Artes apresentaram 0% de similaridade.

Além disso, devido ao grande volume de palavras-chave e à pouca padronização em seu registro por parte dos pesquisadores, espera-se também a utilização de um *thesaurus* para tal finalidade. Um *thesaurus* genérico que contemple todas as grandes áreas do conhecimento facilitará este tipo de análise, visto que poderá reduzir drasticamente a quantidade de palavras-chave de mesmo valor semântico a serem processadas. Por fim, espera-se que diversos outros trabalhos que tratam da questão dos tópicos de pesquisas possam se beneficiar dos resultados aqui apresentados.

7.2 Publicações

A seguir, são apresentados os artigos que foram publicados a partir dos primeiros resultados dos estudos das palavras-chave de publicações científicas de doutores que possuem currículos cadastrados na Plataforma Lattes, no intuito de evidenciar os principais tópicos de pesquisas desenvolvidos pelas Engenharias:

- GOMES, J.O.; DIAS, T. M. R.; MOITA, G. F. **Análise dos tópicos de pesquisa dos pesquisadores brasileiros que atuam nas áreas de engenharias**. In: IBERIAN LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGINEERING (CILAMCE), XXXVII, Brasília, 2016. *(**Obs.:** selecionado para publicação na **Revista RIPE** (2016)).*

- GOMES, J.O.; DIAS, T. M. R.; MOITA, G. F. **Análise dos tópicos de pesquisa dos pesquisadores brasileiros que atuam nas áreas de engenharias**. Revista Interdisciplinar de Pesquisa em Engenharia (RIPE), v. 2, p. 138-151, 2016.

- GOMES, J.O.; et al. **Análise temporal dos principais tópicos de pesquisas em engenharias no Brasil**. In: IBERIAN LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGINEERING (CILAMCE), XXXVIII, Florianopolis, 2017.

- GOMES, J.O.; DIAS, T. M. R.; MOITA, G. F. **Analysis of the Topics of Research of Brazilian Researchers Acting in the Areas of Engineering**. In: 19TH INTERNATIONAL CONFERENCE ON CYBERMETRICS, SCIENTOMETRICS, INFORMETRICS AND BIBLIOMETRICS, Barcelona, 2017. *(**Obs.:** selecionado para publicação no **WASET** (2017)).*

- GOMES, J.O.; DIAS, T. M. R.; MOITA, G. F. **Analysis of the Topics of Research of Brazilian Researchers Acting in the Areas of Engineering**. World Academy of Science (WASET), v.11, p. 1064-1069, 2017.

- GOMES, J.O.; DIAS, T. M. R.; MOITA, G. F. **Identification of Principal Research Topics Studied in Engineering in Brazil**. In: IBERIAN LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGINEERING (CILAMCE), XXXVII, Paris, França, 2018.

Em continuação, também foram publicados artigos que descrevem o arcabouço de componentes responsável pela filtragem e tratamento dos dados contidos nos currículos cadastrados na Plataforma Lattes (Capítulo 4) e os resultados das análises sobre a caracterização num âmbito geral e por grande área do conhecimento de toda a produção dos doutores brasileiros em periódicos e anais de congressos entre 1962 e 2016 (Capítulo 5):

- GOMES, J.O.; DIAS, T. M. R.; MOITA, G. F. **Uma análise dos principais tópicos de pesquisas investigados pelos pesquisadores doutores brasileiros**. Em Questão, v. 24, p. 55-82, 2018.

- GOMES, J.O.; DIAS, T. M. R.; MOITA, G. F. **Uma análise temporal dos principais tópicos de pesquisas da ciência brasileira a partir das palavras-chave de publicações científicas**. Pesquisa Brasileira em Ciência da Informação e Biblioteconomia, v. 13, p. 21-31, 2018. (Pesquisa em andamento)

- GOMES, J.O.; et al. **Visualização analítica das palavras-chaves nos eventos científicos: proposta a partir do Currículo Lattes**. Ciência da Informação, v. 25, p. 216-233, 2018.

- GOMES, J.O.; DIAS, T. M. R.; MOITA, G. F. **Redes de palavras-chave como mecanismo para a identificação dos tópicos de interesses dos pesquisadores brasileiros**. In: VI ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA (EBBC), Rio de Janeiro, 2018.

- GOMES, J.O.; DIAS, T. M. R.; MOITA, G. F. **Uma análise temporal dos principais tópicos de pesquisas em ciências sociais aplicadas no Brasil**. In: XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), Londrina-PR, 2018.

- GOMES, J.O.; DIAS, T. M. R.; MOITA, G. F. **Identificação das palavras-chave mais frequentes registradas nos currículos da Plataforma Lattes**. In: XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), Marília-SP, 2017.

- GOMES, J.O.; DIAS, T. M. R.; MOITA, G. F. **Mapeamento e análise do conhecimento científico brasileiro a partir de redes de palavras-chave**. In: REUNIÓN LATINOAMERICANA DE ANÁLISIS DE REDES SOCIALES (RLARS), Florianópolis, 2017.

Finalmente, vale destacar que outros artigos acerca dos resultados apresentados no Capítulo 6 estão em fase de elaboração para submissão a periódicos de relevância na área, com o intuito de ampliar a compreensão sobre os principais tópicos abordados em periódicos pelos doutores brasileiros ao longo do tempo.

7.3 Trabalhos Futuros

Com a proposta de explorar o potencial dos dados cadastrados na base curricular da Plataforma Lattes para extrair conhecimento acerca do conhecimento científico brasileiro, são sugeridos os seguintes trabalhos futuros:

- Construção de uma ferramenta similar ao *google trends* para facilitar a análise sob diferentes perspectivas dos tópicos de pesquisas estudados na comunidade científica brasileira ao longo do tempo, como: por instituição, localização geográfica, linha de pesquisa, grande área do conhecimento, programas de pesquisa, entre outros;

- Utilização de um *thesaurus* para melhorar a classificação das palavras-chave devido à pouca padronização existente no registro destas por parte dos pesquisadores;

- Identificar tendências de tópicos de pesquisas para auxiliar interessados em diversos níveis de tomadas de decisão no contexto científico, como: no direcionamento para estudos e na construção de políticas científicas;

- Construção e aplicação de uma medida de importância que utilize o *qualis* como fator qualititativo para ranqueamento de palavras-chave;

- Identificação de especialistas de domínio por meio do estudo de popularidade das palavras-chave de artigos científicos.

- Analisar as publicações da base curricular da Plataforma Lattes que não possuem palavras-chave vinculadas com o intuito de verificar se elas são oriundas de um mesmo autor, de uma determinada área do conhecimento e/ou tipo de publicação.

- Analisar o baixo percentual de palavras-chave contidas nos respectivos títulos das publicações. Como parte do estudo, sugere-se uma avaliação das palavras-chave cadastradas nos artigos da base curricular da Plataforma Lattes *versus* palavras-chave contidas em repositórios renomados na literatura. Esse estudo tem relevância destacada, pois pode influenciar nos resultados dos trabalhos futuros de análises de tópicos de pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

ABE, T.; TSUMOTO, S. Evaluating a method to detect temporal trends of phrases in research documents. **IEEE Xplore Digital Library**, Hong kong, China, 2009. ISBN: 978-1-4244-4642-1. Disponível em: http://ieeexplore.ieee.org/document/5250711

AQUINO, I. S.; AQUINO, I.S. Análise sobre a forma da escrita de palavras-chave em artigos científicos na área de ciências agrárias publicados no período de 1999 a 2011. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 18, n. 37, p. 227 - 238, 2013. ISSN 1518-2924. Disponível em: http://dx.doi.org/10.5007/1518-2924.2013v18n37p227

ARAUJO, C. A. A. Bibliometria: evolução histórica e questões atuais. **Em Questão**. v.12, n.1, 2006. ISSN 1807-8893.

ARAÚJO, E.B.; et al. Collaboration networks from a large cv database: dynamics, topology and bonus impact. **PloS One**, v. 9, n. 3, p. 90537, 2014.

BARABASI, A. L.; ALBERT, R. **Emergence of scaling in random networks. Science**, v. 286, n. 5439, p. 509-512, 1999.

BASTIAN, M.; HEYMANN, S. The open graph viz plataform. **Proceedings of XXX Sunbelt Social Networks Conference**, Riva Del Garda, Itália, 2010.

BENEVENUTO, F.D.S. **Uma Análise Empírica de Interações em Redes Sociais**. 2010. 149 (Tese de Doutorado). Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, UFMG, 2010.

BIRD, S.; KLEIN, E.; LOPER, E. (2009). **Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit**. ISBN: 10: 0-596-51649-5.

BOAVENTURA, M.; et al. Caracterização Temporal das Redes de Colaboração Científica nas Universidades Brasileiras: Anos 2000-2013. In: **Brazilian Workshop on Social Network Analysis and Mining**, 2014. Brasília, 2014.

BOGAERT, J.; ROUSSEAU, R; HECKE, P. V. Percolation as a model for informetric distributions: fragment size distribution characterized by Bradford curves. **Scientometrics**, v. 47, n.2, p. 195- 206, 2000. Disponível em: http://dx.doi.org/10.1023/A:1005678707987

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**, v. 23, n. 1, p. 1-13, 2015.
Disponível em: http://dx.doi.org/10.5753/rbie.2015.23.01.85

BRITO, A. G. C.; QUONIAM, L.; MENA-CHALCO, J. P. Exploração da Plataforma Lattes por assunto: proposta de metodologia. **Transinformação**, v.28, n.1, p. 77-86, 2016. Disponível em: http://periodicos.puc-campinas.edu.br/seer/index.php/transinfo/article/view/2519

CALDEIRA, S. **Caracterização da Rede de Signos Linguísticos: Um Modelo baseado no Aparelho Psíquico de Freud**. 2005. (Dissertação de Mestrado). Centro de Pós-Graduação e Pesquisa Visconde de Cairu, Salvador, Brasil.

CALLON, M.; et al. From translations to problematic networks: An introduction to co-word analysis. **Social Science Information,** v. 22, n. 2, p. 191-235, 1983. Disponível em: http://journals.sagepub.com/doi/abs/10.1177/053901883022002003

CALLON, M.; LAW, J.; RIP, A. **Mapping the dynamics of science and technology:** sociology of science in the real world. 1. Ed., Palgrave Macmillan, p. 242, 1986. Disponível em: http://link.springer.com/book/10.1007%2F978-1-349-07408-2

CAMPOS, M. L. A.; GOMES, H. E. Taxonomia e Classificação: o princípio de categorização. **DataGramaZero** - Revista Ciência da Informação, v.9, n.4, p. 1-13, 2008. Disponível em: http://www.enancib.ppgci.ufba.br/artigos/GT2--101.pdf

CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. **Revista de Administração Pública**, v. 42, n. 3, p. 495-528, 2008.

CARNEIRO, D.M. C&T em prol da cidadania. **Fapemig**. 2003. Disponível em: http://lqes.iqm.unicamp.br/canal_cientifico/pontos_vista/pontos_vista_artigos_opiniao27-1.html>

CASTELLS, M. **A sociedade em rede**. São Paulo: Paz e Terra, 2010.

CASTEIGTS, A.; et al. Time-varying graphs and dynamic networks. **Springer - ADHOC-NOW of Lecture Notes in Computer Science,** v. 6811, p. 346-359, 2011.

CAVACINI, A. Recent trends in Middle Eastern scientific production. **Scientometrics**, v. 109, n. 1, p. 423-432, 2016.Disponível em: http://link.springer.com/article/10.1007/s11192-016-1932-3

CHAGAS, F. M.; PEREZ-ALCAZAR, J. J.; DIGIAMPIETRI, L. A. 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. Disponível em: http://seer.ufrgs.br/index.php/EmQuestao/article/view/52652

CHEN, G.; XIAO, L.; HU, C.; ZHAO, X. Identifying the research focus of Library and Information Science institutions in China with institutions-specific keywords. **Scientometrics**, v. 103, n. 2, p. 707-724, 2015. Disponível em: http://link.springer.com/article/10.1007/s11192-015-1545-2

CHEN, G.; XIAO, L. Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods. **Journal of Informetrics**, v. 10, n. 1, p. 212-223, 2016.

CHOI, J.; SANGYOON, Y.; LEE, K.C. Analysis of keyword networks in MIS research and implications for predicting knowledge evolution. **Information & Management**, v. 48, n. 8, p. 371-381, 2011. Disponível em: http://dx.doi.org/10.1016/j.im.2011.09.004

CNPQ. **Conselho Nacional de Desenvolvimento Científico e Tecnológico**. Brasil, 2016. Disponível em: http://lattes.cnpq.br

CORMEN, T. H.; et al. **Introduction to Algorithms.** Cambridge, MA: MIT Press, 2001.

CUNHA, M. V.; et al. Redes de títulos de artigos científicos variáveis no tempo. **II Brazilian Workshop on Social Network Analysis and Mining** (BraSNAM 2013). Maceió - AL: SBC. p. 01-10. 2013.

CUNHA, M.V. **Redes Semânticas baseadas em títulos de artigos científicos**. 2013. 111 (Dissertação de Mestrado). Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial, Salvador, Bahia, 2013.

DECKER, S.J.; et al. Detection of Bursty and Emerging Trends towards Identification of Researchers at the Early Stage of Trends. **CiteSeer**, 2007.

DIAS, T. M. R.; MOITA, G. F. Análise da Rede de Colaboração Científica dos Congressos Ibero Latino Americano de Métodos Computacionais em Engenharia Extraídos da Plataforma Lattes. In: **Iberian Latin American Congress on Computational Methods in Engineering**, 2013. Pirenópolis, 2013.

DIAS, T. M. R.; et al. Modelagem e Caracterização de Redes Científicas: Um Estudo Sobre a Plataforma Lattes. In: **Brazilian Workshop on Social Network Analysis and Mining**, 2013. Maceió, 2013a.

DIAS, T. M. R; DIAS, P. M.; MOITA, G. F. Analysis Collaboration Networks of Scientific Publications. In: **Contecsi – International Conference of Information Systems and Technology Management**, 2013. São Paulo, 2013b.

DIAS, T. M. R.; et al. Identificação e caracterização de redes científicas de dados curriculares. **Revista Brasileira de Sistemas de Informação**, v.7, n.3, p. 5-18, 2014.

DIAS, T. M. R. **Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes**. 2016. 181 (Tese de Doutorado). Programa de Pós-Graduação em Modelagem Matemática e Computacional, PPGMMC/CEFET-MG, 2016.

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. Disponível em: http://seer.ufrgs.br/index.php/EmQuestao/article/view/53259

DIAS, T. M. R.; MOITA, G. F. Adoção da Plataforma Lattes como fonte de dados para caracterização de redes científicas. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, v. 21, n. 47, p. 16-26, 2016.

DIESTEL, R. **Graph Theory**. Graduate Texts in Mathematics. 4. ed. Heidelberg: Springer-Verlag, v. 173, 2010.

DIGIAMPIETRI, L. A. **Análise da rede social acadêmica brasileira**. 2015. 160. Tese (Livre Docência). Escola de Artes, Ciências e Humanidades da Universidade de São Paulo USP, São Paulo, 2015.

DIGIAMPIETRI, L. A.; et al. BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs. **PLoS ONE**, v. 9, n. 4, p. 1-12, 2014a.
Disponível em: http://dx.doi.org/10.1371/journal.pone.0094541

DIGIAMPIETRI, L. A.; PERES, S. M.; SILVA, L. A. Rede de Relacionamentos Brasileira de Inteligência Artificial e Computacional. **XI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2014)**, p. 1-6, 2014b, São Carlos, SP, Brasil.

EASLEY D.; KLEINBERG, J. **Networks, Crowds and Markets: Reasoning About a Highly Connected World.** Cambridge University, 2010.
Disponível em: https://www.cs.cornell.edu/home/kleinber/networks-book

ERDOS, P.; RENYI, A. On the evolution of random graphs. **Mathematical Institute of the Hungarian Academy of Sciences**, v. 5, n.1, p. 17-61, 1960.
Disponível: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.153.5943

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.

FADIGAS, I.; PEREIRA, H. A network approach based on cliques. **Physica A: Statistical Mechanics and its Applications**, v. 392, n. 10, p. 2576-2587, 2013.

FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the Internet topology. **Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)**, v.29, n.4, p. 251–262, 1999.
Disponível em http://dl.acm.org/citation.cfm?doid=316188.316229

FERRAZ, R. R. N.; QUONIAM, L.; MACCARI, E. A. The use of ScriptLattes tool for extraction and on-line availability of academic production from a departament of stricto sensu in management. In: **International Conference on Information Systems and Technology Management**, XI, São Paulo, p. 663 – 679, 2014.

FERREIRA, A. G. C. Bibliometria na avaliação de periódicos científicos. **Data Grama Zero**. Revista de Ciência da Informação**,** v. 11, n. 3, p. 1-9, 2010.

FERNANDES, G. O.; SAMPAIO, J. O.; SOUZA, J. M. XMLattes - A Tool for Importing and Exporting Curricula Data. In: **Worldcomp – World Congress in Computer Science, Computer Engineering and Applied Computing**, 2011. Las Vegas, Nevada, USA, 2011.

FRAGA, R. **Detecção de linhas de pesquisas emergentes em redes de publicações científicas.** 2016. 107 (Tese de Doutorado). Programa de Pós-Graduação em Desenvolvimento Econômico, Universidade Estadual de Campinas, UNICAMP, 2016.

FRAME, J. D. Mainstream research in Latin America and the Caribbean. **Interciencia**, v. 2, n. 3, p. 143-148, 1977.

FREIRE, G. H. A. O principal canal de comunicação para os pesquisadores. **Biblionline**. v. 8, n. esp., p. 1-2, 2012.

GOFFMAN, W.; NEWILL, V. A. Communication and epidemic processes. **Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences**, v. 298, n.1454, p. 316-334, 1967.
Disponível: http://rspa.royalsocietypublishing.org/content/298/1454/316.abstract

GONÇALVES, A. L. Uso de resumos e palavras-chave em Ciências Sociais: uma avaliação. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**. v. 13, n. 26, p. 78-93, 2008.

GRANADA, R.; et al. Formação de Equipes Profissionais através da Avaliação da Similaridade entre Currículos. **III Simpósio Brasileiro de Sistemas de Informação**, 2006, Curitiba, Brasil.

GUEDES, V. L. S.; BORSCHIVER, S. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. **Encontro Nacional de Ciência da Informação**, p.1-18, 2005, Florianópolis, Brasil.

HERADIO, R.; et al. A bibliometric analysis of 20 years of research on software product lines. **Information and Software Technology**, v. 72, n. 1, p. 1-15, 2016.
Disponível em: http://dx.doi.org/10.1016/j.infsof.2015.11.004

HONG, Y.; et al. Knowledge structure and theme trends analysis on general practitioner research: A Co-word perspective. **Bio Med Central**, v. 17, 2016. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4734860

INFOMONEY. **InfoMoney – site especializado em investimentos pessoais**. Brasil, 2018. Disponível em: https://www.infomoney.com.br/mercados/agro/noticia/5846881/brasil-segundo-maior-produtor-consumidor-biodiesel-mundo

KHAN, G. F.; WOOD, J. Information technology management domain: emerging themes and keyword analysis. **Scientometrics**, v. 105, n. 2, p. 959-972, 2015.
Disponível em: http://link.springer.com/article/10.1007/s11192-015-1712-5

KAUER, A.; MOREIRA, V. Evolução dos Temas de Interesse do SBBD ao Longo dos Anos. SBBD - **Simpósio Brasileiro de Banco de Dados**, 2013, Recife, Brasil.

KLEINBERG, J. Bursty and hierarchical structure in streams. **In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 91–101. ACM Press, 2002.

KOBASHI, N. Y. Fundamentos semânticos e pragmáticos da construção de instrumentos de representação de informação. **DataGramaZero**: Revista de Ciência da Informação, v.8, n.6, p. 1-9, 2007. Disponível em: http://periodicos.ufpb.br/index.php/pbcib/article/view/6120

KONTOSTATHIS, A.; et al. A Survey of Emerging Trend Detection in Textual Data Mining. **Chapter - Survey of Text Mining**, p. 185-224, 2004.
Disponível em: http://link.springer.com/chapter/10.1007%2F978-1-4757-4305-0_9

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

LEE, PC.; SU, HN.; CHAN, TY. Assessment of ontology-based knowledge network formation by Vector-Space Model. **Scientometrics**, v. 85, n. 3, p. 689-703, 2010.

LEYDESDORFF, L. **The challenge of scientometrics: the Development, Measurement and Self-organization of Scientific Communications**. Leiden University. 1995.

LIMA, H.; et al. Assessing the profile of top Brazilian computer science researchers. **Scientometrics**, v. 103, p. 879- 896, 2015. Disponível em: http://dx.doi.org/ 10.1007/s11192-015-1569-7

LIU, G. Y.; HU, J. M.; WANG, H. L. A co-word analysis of digital library field in China. **Scientometrics**, v. 91, n. 1, p. 203-217, 2012.

LUCAS, E. O.; LARA, M. L. G. Palavras-chave como elo entre artigos e autores: visualizações possíveis. In: **IV Encontro Brasileiro de Bibliometria e Cientometria**, Recife, 2014.

MACHADO JR., et al. Análise de viabilidade de utilizar as leis da Bibliometria em diferentes bases de pesquisa. In: **XXXVIII Encontro da ANPAD - EnANPAD**, 2014, Rio de Janeiro, Brasil.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. **Proceedings of Berkeley Symposium on Mathematical Statistics and Probability**, v.1: p., 281-297, University of California Press, California, 1967.

MACULAN, B. C. M. S. **Taxonomia facetada e navegacional**: um mecanismo de recuperação. Curitiba: Appris, p.232, 2014.

MADLOCK-BROWN, C. R. **A framework for emerging topic detection in biomedicine**. 2014. 110 (Thesis). University of Iowa, 2014. Disponível em: http://ir.uiowa.edu/etd/1483/

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

MEDEIROS, C.; MENA-CHALCO, J. P. The Dynamics of Multidisciplinary Research Networks - Mining a Public Repository of Scientists CVs. **World Social Science: social transformations and the digital ages**, 2013, Montreal, Canadá.

MEIRELES, M. R. G; CENDON, B. V.; ALMEIDA, P. E. M. Comparação do processo de categorização de documentos utilizando palavras-chave e citações em um domínio de conhecimento restrito. **Transinformação**. v.28, n.1, p. 87-96, 2016. Disponível em: http://dx.doi.org/10.1590/2318-08892016002800007

MENA-CHALCO, J. P.; CESAR-JUNIOR, R. M. ScriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31-39, 2009.

MENA-CHALCO, J. P.; et al. Brazilian bibliometric coauthorship networks. **Journal of the Association for Information Science and Technology**, v. 65, n. 7, p. 1424-1445, 2014. Disponível em: http://dx.doi.org/10.1002/asi.23010

MENA-CHALCO, J. P.; CESAR-JUNIOR, R. M. **Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes.** In: Bibliometria e Cientometria: reflexões teóricas e interfaces, São Carlos, São Paulo, p. 109-128, 2013.

MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A.; CESAR-JR, R. M. Caracterizando as redes de coautoria de currículos Lattes. **Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)**, Curitiba, Brasil, 2012a.

MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A.; OLIVEIRA, L. B. Perfil de produção acadêmica dos programas brasileiros de pós-graduação em Ciência da Computação nos triênios 2004-2006 e 2007-2009. **Revista da Faculdade de Biblioteconomia e Comunicação da UFRGS**. p. 14-18, Porto Alegre, 2012b.

MENEZES, V. S. D. A. **Análise de Redes Sociais Científicas**. 2012. 206. (Tese de Doutorado). Programa de Engenharia de Sistemas e Computação, UFRJ/COPPE, Rio de Janeiro, 2012.

MILGRAN, S. The small world problem. **Psychology today**, v.2, p. 60-67, 1967.

MIGUÉIS, A.; et al. A importância das palavras-chave dos artigos científicos da área das ciências farmacêuticas, depositados no estudo geral: estudo comparativo com os termos atribuídos na Medline. **Revista de Ciência da Informação e Documentação**, v.4, n.2, p. 112-125, 2013. Disponível em: http://dx.doi.org/10.11606/issn.2178-2075.v4i2p112-125

MIYATA, B.; KANO, V.; DIGIAMPIETRI, L. Combinando mineração de textos e análise de redes sociais para a identificação das áreas de atuação de pesquisadores. **II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2013).** SBC. Maceió, Brasil, p.10, 2013.

MOREIRA, L. R. F.; DIAS, T. M. R.; MOITA, G. F. Uma estratégia baseada em árvores genealógicas científicas para visualização da relação orientador-orientado. In: **Encontro Brasileiro de Bibliometria e Cientometria**, 2016. São Paulo: USP, 2016.

MRYGLOD, O.; et al. Quantifying the evolution of a scientific topic: reaction of the academic community to the Chornobyl disaster. **Scientometrics**, v. 106, n. 3, p. 1151- 1166, 2016. Disponível: http://link.springer.com/article/10.1007/s11192-015-1820-2

MUGNAINI, R.; LEITE, P.; LETA, J. Fontes de informação para análise de internacionalização da produção científica brasileira. **PontodeAcesso**, v. 5, n. 3, 2011.

MUGNAINI, R.; DIGIAMPIETRI, L. A.; MENA-CHALCO, J. P. Comunicação científica no brasil (1998-2012): indexação, crescimento, fluxo e dispersão. **Transinformação**, v. 26, n.3, p. 239-252, 2014.

NEWMAN, M. E. J. The structure of scientific collaboration networks. **Proceedings of the National Academy of Sciences,** v. 98, n.2, p. 404-409, 2001.

NLTK. **Natural Language Toolkit: NLTK 3.0 documentation**. 2015. Disponível em: http://www.nltk.org/

OHNIWA, R. L.; HIBINO, A.; TAKEYASU, K. Trends in research foci in life science fields over the last 30 years monitored by emerging trends. **Scientometrics**, v. 85, n. 1, p. 111-127, 2010. Disponível em: http://link.springer.com/article/10.1007/s11192-010-0252-2

PELEIAS, I. R.; et al. Dez anos de pesquisa científica em controladoria no Brasil (1997-2006). **Revista de Administração e Inovação (RAI)**, v. 7, n. 1, p. 193-2017, 2010.

PEREIRA, H. B. B.; et al. Semantic networks based on titles of scientific papers. **Physica A: Statistical Mechanics and its Applications**, v. 390, n. 6, p. 1192-1197, 2011.

PEREIRA, M. G. **Artigos científicos: como redigir, publicar e avaliar**. Rio de Janeiro: Guanabara Koogan, p.384, 2011

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. Disponível em: http://dx.doi.org/ 10.1007/s11192-007-1787-8

PEREZ-CERVANTES, E.; et al. Using link prediction to estimate the collaborative influence of researchers. In: **IEEE 9th International Conference on E-science**. China, Beijing, p. 293-300, 2013.

PETER, R.S.; et al. Epidemiologic research topics in Germany: a keyword network analysis of 2014 DGEpi conference. **European Journal of Epidemiology**, v. 31, n.6, p. 635- 638, 2016. Disponível em: https://link.springer.com/article/10.1007%2Fs10654-016-0141-y

POLLACK, J.; ADLER, D. Emergent trends and passing fads in project management research: A scientometric analysis of changes in the field. **Project Management**, v. 33, n. 1, p. 236-248, 2015. Disponível em: http://dx.doi.org/10.1016/j.ijproman.2014.04.011.

PRITCHARD, A. Statistical bibliography or bibliometrics? **Journal of Documentation**, v. 25, n.4, p. 348-349, 1969.

RAO, I. K. R. Métodos quantitativos em biblioteconomia e ciência da informação. **Associação dos Bibliotecários do Distrito Federal**, 1986.

RAVIKUMAR S., AGRAHARI A., SINGH S.N. Mapping the intellectual structure of scientometrics: a co-word analysis of the journal Scientometrics (2005–2010). **Scientometrics**, v. 102, n. 1, p. 929-955, 2015. Disponível em: http://link.springer.com/article/10.1007/s11192-014-1402-8

RIDKER, P. M.; RIFAI, N. Expanding Options for Scientific Publication. Is More Always Better? **Circulation**, v. 127, n. 1, p.155-156, 2013.

RONDA-PUPO, G. Knowledge map of Latin American research on management: Trends and future advancement. **Social Science Information**, v. 55, n. 1, p. 3 – 27, 2015. Disponível em: http://journals.sagepub.com/doi/abs/10.1177/0539018415610225

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. 183. (Tese de Doutorado). Programa de Pós-Graduação em Saúde Pública, USP, São Paulo, 2005.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing & Management**, v. 24, n. 5, p. 513 – 523, 1988. Disponível em: http://dx.doi.org/10.1016/0306-4573(88)90021-0.

SANTOS, R. N. M.; KOBASHI, N. Y. Bibliometria, cientometria, informetria: conceitos e aplicações. **Ciência da Informação**, Brasília, v. 2, n. 1, p.155-172, 2009. Disponível em: http://basessibi.c3sl.ufpr.br/brapci/v/a/7766

SANTORO, N.; et al. Time-varying graphs and social network analysis: Temporal indicators and metrics**. Computer Science - Social and Information Networks, 2011**. Disponível em http://arxiv.org/abs/1102.0629

SHIBATA, N.; et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. **Technovation**, v. 28, n. 11, p. 758-775, 2008. ISSN 0166-4972. Disponível: http://www.sciencedirect.com/science/article/pii/S0166497208000436

SILVA, F.N.; et al. Using network science and text analytics to produce surveys in a scientific topic. **Journal of Informetrics**, v. 10, n. 2, p. 487 – 502, 2016.
Disponível: http://www.sciencedirect.com/science/article/pii/S1751157715301966

SOHRABI, B.; IRAJ, H. The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts. **Scientometrics**, v. 110, n. 1, p. 243-251, 2017.
Disponível em: https://link.springer.com/article/10.1007/s11192-016-2161-5

SOUZA, J.; et al. Análise de redes de palavras baseada em títulos extraídos de um sistema de atendimento. **III Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2014)**. CSBC. Brasília, Brasil, p. 87-96, 2014.

SOUZA, J.; et al. Análise de rede de termos em Sistemas Embarcados através de análise da rede de termos em títulos de trabalhos científicos. **IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015)**. CSBC. Recife, Brasil, 2015. Disponível em: http://dx.doi.org/10.13140/RG.2.1.3246.3207

SOUZA, J. **Utilizando Títulos de Artigos Científicos na Construção de Redes Semânticas para Caracterizar Áreas de Pesquisa**. 2015. 87 (Dissertação de Mestrado). Programa de Pós-Graduação do Departamento de Informática da Universidade Federal da Paraíba, Universidade Federal da Paraíba, João Pessoa, Paraíba, 2015.

SU, HN.; LEE, PC. Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight. **Scientometrics**, v. 85, n. 1, p. 65-79, 2010. Disponível em: http://link.springer.com/article/10.1007/s11192-010-0259-8

TAGUE-SUTCLIFFE, J. An introduction to informetrics. **Information Processing & Management**, v. 28, n. 1, p. 1-3, 1992.
Disponível: http://www.sciencedirect.com/science/article/pii/030645739290087G

TRUCOLO, C. C.; DIGIAMPIETRI, L. A. Uma revisão sistemática acerca das técnicas de identificação e análise de tendências. In: **Simpósio Brasileiro de Sistemas de Informação (SBSI).** 2014a. p. 639-650.

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, v. 3, n. 2, p. 87-94, 2014b.
Disponível em: http://revistas.ufpr.br/atoz/article/view/41341

TRUCOLO, C. C.; DIGIAMPIETRI, L. A. Trend Analysis of the Brazilian Scientific Production in Computer Science. **Revista de Sistemas de Informação da FSMA**. 2014c. p. 2-10.

TRUCOLO, C.C. **Análise de tendências em redes sociais acadêmicas**. 2016. (Dissertação de Mestrado). Programa de Pós-Graduação em Sistemas de Informação, São Paulo, Brasil.

TSENG, YH.; et al. A comparison of methods for detecting hot topics. **Scientometrics**, v. 81, n. 73, 2009. Disponível em: http://link.springer.com/article/10.1007/s11192-009-1885-x

TUESTA, E. F.; et al. Analysis of an Advisor-Advisee Relationship: An Exploratory Study of the Area of Exact and Earth Sciences in Brazil. **Plos One**, v. 10, p. 1-18, 2015.
VANZ, S. A. S.; CAREGNATO, S. E. Estudos de Citação: uma ferramenta para entender a comunicação científica. **Em Questão**. v.9, n.2, p. 295-307, 2003. ISSN 1807-8893.

VANTI, N. A. P. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. **Ciência da Informação**, v. 31, n. 2, p. 152-162, 2002.

VENTURA, J.; SILVA, J. Automatic extraction of explicit and implicit keywords to build document descriptors. **Progress in Artificial Intelligence**, v. 8154, p. 492-503, 2013. Disponível em: http://dx.doi.org/10.1007/978-3-642-40669-0_42

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**, 2015. Disponível em: http://www.bmj.com/content/351/bmj.h6467

VOOS, H. Lotka and information science. **Journal of the American Society of Information Science**, v. 25, n. 4, p. 270-272, 1974. Disponível: http://dx.doi.org/10.1002/asi.4630250410

ZHANG K.; et al. A bibliometric analysis of research on carbon tax from 1989 to 2014. **Renewable and Sustainable Energy Reviews**, v. 58, p. 297-310, 2016.
Disponível em: http://dx.doi.org/10.1016/j.rser.2015.12.089

ZHANG W.; et al. Knowledge map of creativity research based on keywords network and co-word analysis, 1992–2011. **Quality & Quantity**, v. 49, n. 3, p. 1023-1038, 2015. Disponível em: http://link.springer.com/article/10.1007/s11135-014-0032-9

ZHU D.; et al. Small-world phenomenon of keywords network based on complex network. **Scientometrics**, v. 97, n. 2, p. 435-442, 2013. Disponível em: http://dx.doi.org/10.1007/s11192-013-1019-3

ZHU L.; et al. Keywords co-occurrence mapping knowledge domain research base on the theory of Big Data in oil and gas industry. **Scientometrics**, v. 105, n. 1, p. 249-260, 2015.
Disponível em: http://dx.doi.org/10.1007/s11192-013-1019-3

YI, S.; CHOI, J. The organization of scientific knowledge: the structural characteristics of keyword networks. **Scientometrics**, v. 90, n. 3, p. 1016-1026, 2012.
Disponível em: http://dx.doi.org/ 10.1007/s11192-011-0560-1

WU, J.; WATTS, D. J. Small Worlds: The Dynamics of networks between order and randomness. **ACM SIGMOD Record**, v. 31, n.4, p. 74-75, New York, 2002.
Disponível em http://dl.acm.org/citation.cfm?doid=637411.637426

Apêndices

Tabela A.1 – As palavras-chave mais frequentes por quinquênios entre 1962 e 2016.

Ranque	1962-1966	Frequência	1967-1971	Frequência	1972-1976	Frequência
1	Taxonomia	49	Taxonomia	82	Ratos	154
2	Sistemática	30	Doença de Chagas	61	Taxonomia	149
3	Hymenoptera Apoidea	27	Bovinos	58	Bovinos	120
4	Fungos	25	Brasil	54	Cultivo	120
5	Microbiologia	23	Sistemática	53	Soja	109
6	Micologia	23	Cirurgia	49	Brasil	100
7	Brasil	22	Microscopia Eletrônica	43	Doença de Chagas	98
8	Morfologia	22	Coleóptera	37	Glândulas salivares	91
9	Pólen	21	Sedimentologia	37	Milho	91
10	Microscopia Eletrônica	20	Diptera	36	Sementes	91
11	Coleóptera	19	Morfologia	34	Morfologia	90
12	Sol	19	Anatomia	33	Comportamento	88
13	Bovinos	18	Ultraestrutura	31	Adubação	87
14	Geologia	18	Abelhas	30	Trypanosoma Cruzi	86
15	Rio Grande do Sul	18	Cardiologia	30	Ecologia	79
Total de Palavras-chave do Quinquênio		**7.553**		**22.829**		**60.712**
Ranque	**1977-1981**	**Frequência**	**1982-1986**	**Frequência**	**1987-1991**	**Frequência**
1	Soja	519	Bovinos	754	Epidemiologia	1.126
2	Bovinos	396	Soja	722	Bovinos	1.056
3	Trypanosoma Cruzi	383	Trypanosoma Cruzi	631	Taxonomia	1.016
4	Sementes	328	Epidemiologia	613	Ratos	1.006
5	Brasil	312	Brasil	581	Brasil	942
6	Ratos	307	Sementes	579	Trypanosoma Cruzi	942
7	Taxonomia	258	Ratos	567	Soja	913
8	Milho	255	Taxonomia	506	Educação	895
9	Germinação	238	Milho	505	Amazônia	866
10	Cultivo	219	Doença de Chagas	484	Diagnóstico	854
11	Amazônia	211	Amazônia	480	Sementes	820
12	Ecologia	210	Educação	453	Milho	780
13	Adubação	206	Cultivo	405	Crianças	723
14	Educação	204	Diagnóstico	400	Peixes	716
15	Suínos	203	Peixes	392	Aids	664
Total de Palavras-chave do Quinquênio		**147.645**		**311.951**		**606.292**
Ranque	**1992-1996**	**Frequência**	**1997-2001**	**Frequência**	**2002-2006**	**Frequência**
1	Epidemiologia	3.063	Educação	10.324	Educação	19.680
2	Educação	2.912	Epidemiologia	7.069	Formação do Professor	11.378
3	Bovinos	2.468	Bovinos	5.808	Epidemiologia	10.325
4	Ratos	2.309	Enfermagem	5.799	Crianças	8.290
5	Diagnóstico	2.051	Ratos	5.563	Enfermagem	8.197
6	Brasil	1.923	Formação do Professor	4.950	Bovinos	7.916
7	Amazônia	1.899	Crianças	4.912	Ratos	7.780
8	Aids	1.867	Diagnóstico	4.777	Amazônia	7.232
9	Milho	1.852	Brasil	4.587	Diagnóstico	7.178
10	Sementes	1.814	Milho	4.336	Educação Ambiental	6.911
11	Crianças	1.793	Ensino	4.235	Idoso	6.826
12	Soja	1.710	Amazônia	4.168	Ensino	6.688
13	Trypanosoma Cruzi	1.645	Sementes	3.856	Avaliação	6.140
14	Taxonomia	1.562	Avaliação	3.663	Políticas Públicas	6.133
15	Enfermagem	1.534	Soja	3.594	Brasil	6.001
Total de Palavras-chave do Quinquênio		**1.519.397**		**3.896.270**		**6.502.794**
Ranque	**2007-2011**	**Frequência**	**2012-2016**	**Frequência**		
1	Educação	23.078	Educação	18.205		
2	Formação do Professor	15.850	Formação do Professor	12.427		
3	Enfermagem	12.910	Enfermagem	11.078		
4	Epidemiologia	10.609	Políticas Públicas	8.473		
5	Políticas Públicas	9.885	Sustentabilidade	7.151		
6	Idoso	9.212	Epidemiologia	7.057		
7	Crianças	7.722	Idoso	6.681		
8	Educação Ambiental	7.344	Ensino	5.832		
9	Biodiesel	7.250	Amazônia	5.744		
10	Amazônia	6.924	Gênero	5.341		
11	Cultura	6.442	Educação Ambiental	5.185		
12	Gênero	6.434	Crianças	5.133		
13	Ensino	6.405	Brasil	4.819		
14	Sustentabilidade	6.345	Memória	4.775		
15	Meio Ambiente	6.287	Biodiesel	4.698		
Total de Palavras-chave do Quinquênio		**6.373.163**		**4.807.706**		

Tabela A.2 – As palavras-chave que apresentam maior grau entre 1962 e 2016.

Ranque	1962-1966	Grau	1967-1971	Grau	1972-1976	Grau
1	Taxonomia	65	Bovinos	130	Ratos	261
2	Brasil	49	Taxonomia	118	Taxonomia	234
3	Morfologia	49	Cirurgia	104	Bovinos	221
4	Bovinos	47	Brasil	91	Brasil	206
5	Microscopia Eletrônica	47	Microscopia Eletrônica	89	Cultivo	182
6	Crustáceos	40	Doença de Chagas	75	Soja	175
7	Rio Grande do Sul	38	Sistemática	74	Comportamento	157
8	Pólen	35	Morfologia	73	Morfologia	156
9	Amazônia	33	Dípteras	72	Milho	155
10	Histoquímica	33	Ultraestrutura	69	Produto	152
11	Decapodas	33	Comportamento	64	Produção	147
12	Dípteras	32	Histoquímica	61	Ecologia	145
13	Ratos	32	Técnica Cirúrgica	60	Microscopia Eletrônica	143
14	Pernambuco	31	Ratos	57	Proteína	135
15	Abelhas	29	Cerrado	57	Epidemiologia	133
Total Grau do Quinquênio		**14.746**		**45.972**		**122.092**
Ranque	**1977-1981**	**Grau**	**1982-1986**	**Grau**	**1987-1991**	**Grau**
1	Soja	606	Brasil	1.126	Brasil	1.800
2	Brasil	599	Bovinos	1.029	Ratos	1.562
3	Bovinos	555	Epidemiologia	902	Epidemiologia	1.482
4	Ratos	520	Amazônia	878	Amazônia	1.471
5	Trypanosoma Cruzi	507	Ratos	834	Bovinos	1.456
6	Taxonomia	464	Soja	817	Diagnóstico	1.322
7	Milho	407	Taxonomia	811	Taxonomia	1.321
8	Amazônia	373	Diagnóstico	788	Trypanosoma Cruzi	1.150
9	Sementes	365	Trypanosoma Cruzi	785	Tratamento	1.108
10	Cultivo	361	Milho	722	Crianças	1.097
11	Controle	359	Cultivo	692	Morfologia	1.093
12	Ecologia	328	Cerrado	640	Soja	1.065
13	Proteína	322	Tratamento	622	Educação	1006
14	Comportamento	320	Sementes	610	Ecologia	991
15	Crescimento	315	Comportamento	589	Milho	985
Total Grau do Quinquênio		**301.662**		**640.578**		**1.255.610**
Ranque	**1992-1996**	**Grau**	**1997-2001**	**Grau**	**2002-2006**	**Grau**
1	Epidemiologia	3.496	Brasil	7.551	Educação	12.116
2	Brasil	3.436	Epidemiologia	7.119	Brasil	10.019
3	Ratos	3.313	Educação	6.908	Epidemiologia	10.007
4	Diagnóstico	2.943	Ratos	6.658	Diagnóstico	9.176
5	Amazônia	2.876	Diagnóstico	6.262	Amazônia	8.882
6	Bovinos	2.716	Amazônia	5.704	Ratos	8.598
7	Educação	2.438	Bovinos	5.461	Crianças	7.575
8	Tratamento	2.339	Crianças	4.938	Avaliação	7.488
9	Crianças	2.300	Avaliação	4.909	Bovinos	7.325
10	Morfologia	2.183	Tratamento	4.779	Milho	7.260
11	Taxonomia	2.055	Morfologia	4.676	Morfologia	6.687
12	Ecologia	2.054	Ecologia	4.316	Tratamento	6.480
13	Crescimento	1.969	Qualidade	4.123	Qualidade	6.339
14	Avaliação	1.948	Controle	4.036	Formação do Professor	6.004
15	Milho	1.917	Milho	3.970	Meio Ambiente	5.996
Total Grau do Quinquênio		**3.042.668**		**7.855.920**		**13.256.146**
Ranque	**2007-2011**	**Grau**	**2012-2016**	**Grau**		
1	Educação	12.886	Educação	11.453		
2	Epidemiologia	10.075	Brasil	8.357		
3	Brasil	9.308	Amazônia	8.016		
4	Amazônia	8.806	Epidemiologia	7.735		
5	Diagnóstico	7.773	Políticas Públicas	7.369		
6	Formação do Professor	7.672	Sustentabilidade	6.743		
7	Políticas Públicas	7.484	Formação do Professor	6.551		
8	Crianças	7.334	Enfermagem	6.359		
9	Avaliação	6.853	Memória	5.629		
10	Ratos	6.821	Gênero	5.097		
11	Cultura	6.389	Diagnóstico	5.031		
12	Sustentabilidade	5.902	Avaliação	4.931		
13	Enfermagem	5.898	Cultura	4.879		
14	Bovinos	5.776	Crianças	4.871		
15	Milho	5.729	Ratos	4.835		
Total Grau do Quinquênio		**12.450.738**		**9.701.290**		

Tabela A.3 – As principais palavras-chave ranqueadas por TF.FI em cada biênio.

Ranque	1997-1998	TF.FI	1999-2000	TF.FI	2001-2002	TF.FI
1	Epidemiologia	251.133	Epidemiologia	486.999	Epidemiologia	759.696
2	Ratos	216.384	Trypanosoma Cruzi	359.080	Amazônia	656.019
3	Trypanosoma Cruzi	190.836	Ratos	349.418	Ratos	595.625
4	Amazônia	134.330	Amazônia	211.485	Brasil	352.256
5	Diagnóstico	65.348	Diagnóstico	155.496	Trypanosoma Cruzi	312.892
6	HIV	59.892	Brasil	141.771	Diagnóstico	263.718
7	Aids	55.554	Genoma	126.776	Câncer	258.875
8	Cardiologia	41.208	PCR	122.144	PCR	258.440
9	Doença de Chagas	30.702	Aids	116.676	Óxido Nítrico	205.938
10	Câncer	30.276	HIV	98.208	Genoma	196.688
11	Tratamento	29.181	Crianças	94.105	Bovinos	174.248
12	Brasil	29.050	Sequenciamento	90.240	Proteínas	142.803
13	Crianças	27.280	Câncer	85.550	Diabetes	115.885
14	Memória	26.978	Óxido Nítrico	79.765	Inflamação	105.588
15	Ecologia	25.979	Bovinos	74.226	HIV	94.170
Total de Palavras-chave do Biênio		**382.868**		**521.213**		**674.110**
Ranque	**2003-2004**	**TF.FI**	**2005-2006**	**TF.FI**	**2007-2008**	**TF.FI**
1	Epidemiologia	1.304.217	Epidemiologia	1.539.142	Epidemiologia	2.419.450
2	Amazônia	1.074.897	Ratos	1.111.624	Ratos	995.601
3	Ratos	876.420	Amazônia	860.360	Brasil	945.216
4	Genoma	687.360	Câncer	851.032	Amazônia	765.011
5	Câncer	660.400	Brasil	690.480	Trypanosoma Cruzi	595.782
6	Brasil	453.000	Óxido Nítrico	482.868	Câncer	577.252
7	Trypanosoma Cruzi	429.552	Aids	435.848	Diagnóstico	496.762
8	Diagnóstico	386.595	Diabetes	432.990	Estresse Oxidativo	486.920
9	Óxido Nítrico	337.450	Trypanosoma Cruzi	390.630	Óxido Nítrico	463.690
10	Diabetes	331.296	Diagnóstico	373.214	HIV	414.117
11	Bioinformática	264.421	Estresse Oxidativo	330.863	Diabetes	398.853
12	Bovinos	221.020	HIV	278.720	Bovinos	290.862
13	PCR	209.385	Bovinos	230.520	Obesidade	284.428
14	Tratamento	184.658	Taxonomia	223.539	Inflamação	273.504
15	Estresse Oxidativo	178.839	Inflamação	200.646	Taxonomia	252.340
Total de Palavras-chave do Biênio		**830.327**		**891.994**		**912.361**
Ranque	**2009-2010**	**TF.FI**	**2011-2012**	**TF.FI**	**2013-2014**	**TF.FI**
1	Epidemiologia	4.108.686	Epidemiologia	4.026.439	Epidemiologia	3.742.560
2	Amazônia	1.678.172	Amazônia	1.900.250	Amazônia	1.513.576
3	Ratos	1.215.665	Brasil	1.432.689	Brasil	1.454.421
4	Brasil	1.211.580	Câncer	1.352.239	Câncer	1.122.030
5	Câncer	984.004	Estresse Oxidativo	1.265.891	Obesidade	1.030.140
6	Estresse Oxidativo	749.496	Ratos	1.170.246	Estresse Oxidativo	1.011.738
7	Diagnóstico	670.907	Biodiesel	1.150.092	Biodiesel	972.416
8	Trypanosoma Cruzi	601.028	Obesidade	908.208	Inflamação	726.119
9	Obesidade	558.279	Inflamação	745.130	Diabetes	565.110
10	Diabetes	520.007	HIV	705.280	Antioxidante	547.545
11	Óxido Nítrico	509.850	Diagnóstico	661.713	Diagnóstico	537.000
12	Inflamação	490.311	Trypanosoma Cruzi	628.208	Trypanosoma Cruzi	482.608
13	HIV	483.875	Antioxidante	566.690	Ratos	469.650
14	Biodiesel	443.090	Diabetes	522.244	HIV	451.552
15	Bovinos	429.420	Citocina	492.336	Nanopartícula	441.870
Total de Palavras-chave do Biênio		**1.044.271**		**1.063.348**		**1.019.004**
Ranque	**2015-2016**	**TF.FI**	**1997-2016**	**TF.FI**		
1	Epidemiologia	3.154.392	Epidemiologia	197.941.520		
2	Amazônia	2.065.728	Amazônia	97.458.265		
3	Brasil	1.424.475	Brasil	69.948.788		
4	Obesidade	1.140.099	Ratos	69.872.440		
5	Câncer	1.063.920	Câncer	60.588.976		
6	Estresse Oxidativo	978.336	Trypanosoma Cruzi	41.543.320		
7	Inflamação	954.636	Diagnóstico	37.740.892		
8	Biodiesel	784.836	Estresse Oxidativo	37.103.109		
9	Nanopartícula	455.800	Obesidade	32.621.174		
10	Citocina	424.060	Diabetes	29.315.396		
11	Antioxidante	411.600	HIV	26.772.660		
12	Óleo Essencial	361.335	Óxido Nítrico	26.289.914		
13	Diabetes	350.448	Inflamação	25.895.219		
14	Ratos	329.199	Taxonomia	19.505.268		
15	Biomassa	328.192	Bovinos	19.281.820		
Total de Palavras-chave do Biênio		**915.617**		**8.255.113**		

Tabela A.4 – As principais palavras-chave das Ciências Agrárias ranqueadas por TF.FI em cada biênio.

Ranque	1997-1998	TF.FI	1999-2000	TF.FI	2001-2002	TF.FI
1	Bovinos	10.620	Bovinos	49.542	Bovinos	121.706
2	Equinos	7.020	Equinos	6.946	Equinos	35.908
3	Soja	3.794	Milho	5.760	Amazônia	12.236
4	Amazônia	2.720	Temperatura	4.588	Cães	11.451
5	Milho	2.232	Soja	4.336	Proteína	10.854
6	Epidemiologia	1.456	Xylella Fastidiosa	3.362	Soja	8.775
7	Nitrogênio	1.320	Citrus	3.186	Sémen	8.664
8	Biomassa	1.196	Ovinos	3.080	Epidemiologia	7.830
9	Solos	1.090	Epidemiologia	3.024	Brasil	7.171
10	Caprinos	1.038	Sequenciamento	2.820	Ovinos	6.966
11	Sémen	1.032	Nitrogênio	2.814	Biotecnologia	6.215
12	Resistência	870	Caprinos	2.814	Milho	5.430
13	Cães	870	Genética	2.668	PCR	5.341
14	Fósforo	847	Cães	2.480	Caprinos	4.720
15	Temperatura	801	Sementes	2.472	Genoma	4.577
Total de Palavras-chave do Biênio		**57.803**		**87.258**		**109.798**
Ranque	**2003-2004**	**TF.FI**	**2005-2006**	**TF.FI**	**2007-2008**	**TF.FI**
1	Bovinos	110.240	Bovinos	145.860	Bovinos	161.007
2	Amazônia	41.205	Cães	32.648	Cães	34.903
3	Cães	21.412	Amazônia	30.294	Equinos	25.925
4	Milho	19.850	Equinos	20.140	Caprinos	24.479
5	Ovinos	18.961	Ovinos	16.128	Epidemiologia	24.415
6	Genoma	15.015	Milho	15.609	Ovinos	20.984
7	Epidemiologia	13.630	Nutrição	15.557	Amazônia	16.008
8	Cana-de-Açúcar	11.984	Desempenho	15.435	PCR	15.698
9	Equinos	8.579	Brasil	14.400	Brasil	15.660
10	Frangos de corte	7.375	Epidemiologia	13.596	Desempenho	15.453
11	Brasil	7.144	Soja	13.250	Plantio Direto	15.453
12	Tomate	6.972	Café	11.760	Proteína	14.105
13	Reprodução	6.468	Suínos	10.686	Soja	14.025
14	Sémen	6.300	Sémen	9.747	Suínos	13.616
15	Soja	5.909	Caprinos	9.180	Nutrição	12.464
Total de Palavras-chave do Biênio		**132.793**		**139.059**		**146.514**
Ranque	**2009-2010**	**TF.FI**	**2011-2012**	**TF.FI**	**2013-2014**	**TF.FI**
1	Bovinos	206.412	Bovinos	134.616	Bovinos	77.106
2	Equinos	56.550	Epidemiologia	55.860	Brasil	39.579
3	Cães	47.960	Amazônia	38.225	Epidemiologia	36.424
4	Epidemiologia	44.278	Cães	32.448	Cães	34.713
5	Ovinos	38.880	Equinos	25.680	Equinos	34.320
6	Amazônia	38,001	Ovinos	25.256	Amazônia	29.430
7	PCR	33.756	Brasil	24.325	Biodiesel	23.355
8	Brasil	30.173	Milho	23.287	Controle Biológico	19.710
9	Caprinos	26.900	Antioxidante	22.630	Gado	17.680
10	Soja	21.736	PCR	22.278	Probiótico	17.018
11	Nutrição	21.648	Controle Biológico	21.672	Biomassa	15.936
12	Sémen	21.583	Melhoramento Genético	20.610	Soja	15.686
13	Qualidade	18.073	Caprinos	20.604	Cana-de-Açúcar	14.996
14	Desempenho	17.810	Soja	20.444	Cerrado	14.700
15	Plantio Direto	17.216	Qualidade	16.422	Antioxidante	14.518
Total de Palavras-chave do Biênio		**163.665**		**160.343**		**147.722**
Ranque	**2015-2016**	**TF.FI**	**1997-2016**	**TF.FI**		
1	Amazônia	71.273	Bovinos	10.594.044		
2	Bovinos	44.117	Amazônia	2.477.482		
3	Biomassa	28.536	Equinos	2.384.996		
4	Cães	27.020	Cães	2.023.197		
5	Brasil	24.924	Epidemiologia	1.964.808		
6	Epidemiologia	22.610	Brasil	1.380.288		
7	Milho	20.292	Ovinos	1.306.800		
8	Antioxidante	20.043	Caprinos	1.001.817		
9	Equinos	18.020	PCR	901.340		
10	Biodiesel	17.653	Nutrição	894.504		
11	Controle Biológico	16.625	Milho	875.310		
12	Probiótico	14.259	Soja	835.354		
13	Soja	12.880	Proteína	823.256		
14	Ovelhas	12.864	Cana-de-Açúcar	760.104		
15	Nutrição	12.410	Sémen	741.810		
Total de Palavras-chave do Biênio		**126.113**		**1.271.068**		

Tabela A.5 – As principais palavras-chave das Ciências Biológicas ranqueadas por TF.FI em cada biênio.

Ranque	1997-1998	TF.FI	1999-2000	TF.FI	2001-2002	TF.FI
1	Trypanosoma Cruzi	123.968	Trypanosoma Cruzi	255.304	Trypanosoma Cruzi	182.016
2	Ratos	62.496	Ratos	104.469	Amazônia	136.024
3	Amazônia	32.482	Amazônia	55.440	Ratos	128.640
4	Óxido Nítrico	14.946	Óxido Nítrico	37.779	Óxido Nítrico	89.800
5	Ecologia	13.774	PCR	35.875	Brasil	63.444
6	Taxonomia	13.032	Genoma	32.736	PCR	61.246
7	Memória	11.850	Inflamação	25.190	Taxonomia	57.800
8	Esquistossomo	11.205	Taxonomia	24.924	Genoma	51.700
9	Leishmania	10.494	Brasil	23.276	Inflamação	41.041
10	Metabolismo	9.523	Apoptose	22.860	Estresse Oxidativo	28.556
11	Doença de Chagas	7.936	Malária	19.688	Câncer	25.573
12	Malária	7.722	Doença de Chagas	17.576	Lectina	24.824
13	Lectina	5.680	Sequenciamento	17.232	Citocina	22.005
14	HIV	5.133	Memória	15.984	Apoptose	21.660
15	Macrófago	5.084	Macrófago	15.566	Proteína	19.923
Total de Palavras-chave do Biênio		**72.742**		**96.631**		**120.473**
Ranque	**2003-2004**	**TF.FI**	**2005-2006**	**TF.FI**	**2007-2008**	**TF.FI**
1	Trypanosoma Cruzi	243.530	Ratos	250.800	Trypanosoma Cruzi	267.960
2	Ratos	219.436	Trypanosoma Cruzi	217.930	Ratos	213.156
3	Amazônia	195.468	Óxido Nítrico	163.416	Taxonomia	191.136
4	Genoma	158.608	Taxonomia	160.071	Brasil	149.556
5	Óxido Nítrico	136.203	Amazônia	151.956	Óxido Nítrico	141.000
6	Brasil	83.056	Brasil	119.680	Estresse Oxidativo	138.252
7	Estresse Oxidativo	76.287	Estresse Oxidativo	100.934	Amazônia	132.936
8	Taxonomia	75.468	Câncer	74.681	Cerrado	76.570
9	Câncer	73.780	Inflamação	66.736	Epidemiologia	69.784
10	Bioinformática	54.538	Citocina	66.079	Citocina	69.044
11	Citocina	48.205	Genoma	61.425	Câncer	61.380
12	Inflamação	44.020	Cerrado	54.400	Mata Atlântica	59.940
13	PCR	43.618	Conservação	52.993	Diabetes	56.784
14	Insulina	37.570	Diabetes	50.799	Inflamação	52.020
15	Ecologia	37.506	Bioinformática	47.414	Ecologia	47.867
Total de Palavras-chave do Biênio		**148.316**		**156.620**		**157.801**
Ranque	**2009-2010**	**TF.FI**	**2011-2012**	**TF.FI**	**2013-2014**	**TF.FI**
1	Taxonomia	270.861	Taxonomia	362.604	Taxonomia	238.700
2	Amazônia	244.408	Amazônia	272.616	Estresse Oxidativo	204.930
3	Trypanosoma Cruzi	239.436	Estresse Oxidativo	265.776	Amazônia	176.478
4	Ratos	208.030	Trypanosoma Cruzi	226.784	Trypanosoma Cruzi	160.908
5	Estresse Oxidativo	189.745	Brasil	216.814	Brasil	160.200
6	Brasil	178.422	Ratos	193.556	Câncer	129.800
7	Óxido Nítrico	141.250	Cerrado	160.758	Inflamação	129.360
8	Epidemiologia	139.040	Inflamação	159.750	Epidemiologia	107.240
9	Cerrado	133.152	Câncer	154.330	Conservação	93.324
10	Conservação	115.020	Conservação	143.016	Proteômica	90.954
11	Inflamação	106.596	Biodiversidade	123.000	Biodiversidade	87.606
12	Câncer	98.230	Epidemiologia	116.688	Cerrado	81.983
13	Citocina	97.121	Citocina	109.802	Genoma	79.365
14	Biodiversidade	83.980	Óxido Nítrico	107.212	Ratos	70.959
15	Apoptose	83.592	Proteômica	87.210	Óxido Nítrico	67.150
Total de Palavras-chave do Biênio		**175.077**		**173.483**		**157.762**
Ranque	**2015-2016**	**TF.FI**	**1997-2016**	**TF.FI**		
1	Amazônia	237.468	Trypanosoma Cruzi	19.851.264		
2	Estresse Oxidativo	204.204	Amazônia	15.420.223		
3	Taxonomia	197.106	Ratos	14.580.026		
4	Inflamação	192.156	Taxonomia	14.472.090		
5	Brasil	139.195	Brasil	10.183.510		
6	Câncer	117.160	Estresse Oxidativo	9.702.869		
7	Trypanosoma Cruzi	96.400	Óxido Nítrico	841.1054		
8	Biodiversidade	91.632	Inflamação	6.548.580		
9	Cerrado	77.500	Câncer	6.285.000		
10	Obesidade	75.075	Genoma	5.452.656		
11	Citocina	71.820	Cerrado	5.178.456		
12	Genoma	71.604	Conservação	4.860.870		
13	Conservação	70.070	Citocina	4.700.072		
14	Epidemiologia	67.480	Epidemiologia	4.040.556		
15	Malária	64.824	Ecologia	3.887.695		
Total de Palavras-chave do Biênio		**140.750**		**1.399.655**		

Tabela A.6 – As principais palavras-chave das Ciências da Saúde ranqueadas por TF.FI em cada biênio.

Ranque	1997-1998	TF.FI	1999-2000	TF.FI	2001-2002	TF.FI
1	Epidemiologia	161.670	Epidemiologia	279.916	Epidemiologia	381.300
2	Cardiologia	39.996	Crianças	72.562	Ratos	109.134
3	Diagnóstico	37.168	Diagnóstico	67.298	Diagnóstico	100.368
4	Aids	33.524	Aids	64.798	Câncer	93.636
5	Ratos	31.239	Tratamento	50.076	Resina Composta	66.856
6	HIV	26.961	HIV	48.060	Tratamento	62.073
7	Tratamento	23.805	Ratos	46.512	Obesidade	57.090
8	Crianças	19.965	Câncer	35.119	Diabetes	54.216
9	Infarto Agudo do Miocárdio	14.014	Resina Composta	32.118	Crianças	47.450
10	Câncer	12.903	Endodontia	21.735	HIV	42.276
11	Diabete Mellitus	11.088	Diabetes	21.546	Aids	37.536
12	Hipertensão	10.005	Prevenção	19.832	Hipertensão	36.777
13	Epilepsia	7.980	Epilepsia	19.152	Dentina	36.354
14	Diabetes	7.936	Tuberculose	16.107	Laser	32.368
15	Doença de Chagas	6.540	Helicobacter Pylori	14.934	Cardiologia	31.590
Total de Palavras-chave do Biênio		**88.688**		**121.503**		**165.120**
Ranque	**2003-2004**	**TF.FI**	**2005-2006**	**TF.FI**	**2007-2008**	**TF.FI**
1	Epidemiologia	628.178	Epidemiologia	707.128	Epidemiologia	940.430
2	Câncer	154.830	Câncer	256.700	HIV	172.425
3	Diagnóstico	152.711	Aids	225.144	Câncer	162.000
4	Ratos	142.691	Ratos	172.150	Ratos	161.824
5	Diabetes	128.205	Diabetes	148.400	Diagnóstico	156.156
6	Tratamento	117.465	Diagnóstico	128.960	Obesidade	137.379
7	Obesidade	115.393	HIV	123.520	Diabetes	123.265
8	Cardiologia	91.945	Crianças	98.046	Aids	116.232
9	Crianças	73.304	Obesidade	75.775	Tratamento	109.200
10	Aids	71.556	Prognostico	52.800	Crianças	92.380
11	HIV	61.886	Prevalência	49.704	Idoso	90.984
12	Hipertensão	55.594	Tratamento	49.608	Brasil	85.272
13	Stent	36.050	Tuberculose	48.348	Endodontia	84.043
14	Epilepsia	33.912	Diabete Mellitus	39.600	Estresse Oxidativo	68.557
15	Ressonância Magnética	32.085	Brasil	39.176	Adolescente	61.585
Total de Palavras-chave do Biênio		**212.461**		**221.715**		**213.240**
Ranque	**2009-2010**	**TF.FI**	**2011-2012**	**TF.FI**	**2013-2014**	**TF.FI**
1	Epidemiologia	1.628.795	Epidemiologia	1.747.820	Epidemiologia	1.432.284
2	Obesidade	290.250	Obesidade	420.546	Obesidade	464.135
3	Câncer	280.598	Enfermagem	350.779	Câncer	256.360
4	Ratos	233.415	Câncer	330.440	Diabetes	188.055
5	Idoso	231.830	HIV	284.394	Adolescente	183.414
6	Enfermagem	222.054	Adolescente	237.770	HIV	180.565
7	HIV	213.538	Ratos	221.427	Enfermagem	174.024
8	Diagnóstico	172.072	Crianças	212.715	Idoso	164.673
9	Aids	169.824	Idoso	208.840	Inflamação	140.185
10	Crianças	164.124	Estresse Oxidativo	165.953	Diagnóstico	126.549
11	Diabetes	162.944	Aids	162.342	Estresse Oxidativo	125.316
12	Adolescente	159.032	Diabetes	158.632	Crianças	119.712
13	Tratamento	124.581	Diagnóstico	150.064	Exercício	113.155
14	Insuficiência Cardíaca	116.144	Inflamação	118.350	Ratos	101.304
15	Inflamação	99.880	Diabete Mellitus	105.512	Brasil	86.180
Total de Palavras-chave do Biênio		**247.266**		**238.296**		**213.231**
Ranque	**2015-2016**	**TF.FI**	**1997-2016**	**TF.FI**		
1	Epidemiologia	1.329.654	Epidemiologia	86.670.856		
2	Obesidade	457.250	Obesidade	1.6482.438		
3	Câncer	198.940	Câncer	1.6119.070		
4	Inflamação	170.632	Ratos	1.1951.712		
5	Estresse Oxidativo	135.373	Diagnóstico	1.1548.350		
6	Enfermagem	127.680	HIV	11.351.275		
7	Adolescente	105.789	Diabetes	9.917.658		
8	Diabetes	95.676	Crianças	936.7995		
9	Idoso	89.925	Aids	896.5494		
10	Envelhecimento	82.896	Idoso	719.0760		
11	HIV	82.288	Tratamento	719.0760		
12	Brasil	77.546	Adolescente	634.2382		
13	Crianças	75.486	Enfermagem	612.5238		
14	Diabete Mellitus	74.483	Brasil	601.0400		
15	Exercício	71.994	Diabete Mellitus	468.1145		
Total de Palavras-chave do Biênio		**178.684**		**1.900.204**		

Tabela A.7 – As principais palavras-chave das Ciências Exatas e da Terra ranqueadas por TF.FI em cada biênio.

Ranque	1997-1998	TF.FI	1999-2000	TF.FI	2001-2002	TF.FI
1	Espectroscopia	10.864	Magnetismo	22.842	Magnetismo	37.881
2	Magnetismo	10.738	Semicondutor	12.540	Amazônia	22.785
3	Semicondutor	10.700	Cosmologia	12.312	Filme Fino	20.862
4	Cosmologia	9.310	Filme Fino	12.221	Estrutura Cristalina	16.384
5	Amazônia	6.237	Raman	10.058	Cosmologia	15.394
6	Filme fino	4.819	Mercúrio	8.736	Sol-gel	15.008
7	Mecânica Estatística	4.730	Supercondutividade	7.600	Catalisador	13.407
8	Caos	4.290	Estrutura Cristalina	6.375	Raman	12.257
9	Estrutura Cristalina	3.920	Amazônia	6.080	Semicondutor	11.921
10	Polímero	2.862	Sol-gel	5.396	Supercondutividade	11.718
11	Termodinâmica Irreversível	2.652	Espectroscopia	5.082	NMR	10.560
12	Síntese	2.500	Caos	4.565	Sílica	10.100
13	Flavonoides	2.440	Supersimetria	4.280	EPR	9.555
14	Tokamak	2.254	Catalisador	4.260	Óculos	9.345
15	Luminescência	2.160	NMR	4.221	Fotoluminescência	8.586
Total de Palavras-chave do Biênio		**58.734**		**77.026**		**99.373**
Ranque	**2003-2004**	**TF.FI**	**2005-2006**	**TF.FI**	**2007-2008**	**TF.FI**
1	Filme Fino	50.886	Estrutura Cristalina	45.790	Estrutura Cristalina	43.152
2	Amazônia	38.850	Cosmologia	40.446	Biodiesel	32.770
3	Magnetismo	33.201	Espectrometria	34.578	Filme Fino	11.310
4	Genoma	32.300	Filme Fino	24.476	Cosmologia	29.160
5	Sol-gel	27.030	Magnetismo	26.499	Catalisador	26.529
6	Bioinformática	24.932	Espectroscopia	26.394	Fotoluminescência	26.334
7	Estrutura Cristalina	23.184	Catalisador	24.924	Espectrometria	25.938
8	Fotoluminescência	21.666	Fotoluminescência	24.824	Magnetismo	24.634
9	Cosmologia	21.168	Amazônia	23.449	Nanopartícula	24.360
10	Raman	16.157	Nanopartícula	22.116	Adsorção	24.150
11	Espectroscopia	15.456	Sol-Gel	21.842	Amazônia	20.536
12	NMR	15.052	Mercúrio	21.472	Sol-gel	18.564
13	Supercondutividade	13.189	Flavonoides	21.021	NMR	16.640
14	Catalisador	12.870	Adsorção	19.188	Nanotubo de carbono	16.310
15	Nanopartícula	12.445	Bioinformática	18.643	Cruzamento	15.876
Total de Palavras-chave do Biênio		**114.400**		**127.330**		**119.614**
Ranque	**2009-2010**	**TF.FI**	**2011-2012**	**TF.FI**	**2013-2014**	**TF.FI**
1	Biodiesel	62.400	Biodiesel	207.600	Biodiesel	110.189
2	Estrutura Cristalina	53.116	Física de Partículas	102.486	Nanopartícula	57.514
3	Amazônia	50.600	Amazônia	53.130	Amazônia	56.120
4	Nanopartícula	40.754	Adsorção	44.022	Óleo Essencial	31.706
5	Cosmologia	36.915	Nanopartícula	37.905	Novos Fenômenos	26.112
6	Adsorção	36.432	Cosmologia	31.137	Antioxidante	25.944
7	Flavonoides	30.960	Particle Seaches	28.210	Adsorção	25.807
8	Fotoluminescência	26.250	Estrutura Cristalina	25.803	Particle Seaches	25.772
9	Nanocompósitos	25.823	Flavonoides	24.386	Citotóxico	24.990
10	Magnetismo	23.625	Antioxidante	24.309	Cosmologia	22.078
11	Óleo Essencial	22.330	Novos Fenômenos	23.100	Flavonoides	21.033
12	Nanotubo de carbono	21.844	Óleo Essencial	20.732	DFT	19.694
13	Bioinformática	18.618	Bioinformática	20.416	Biosensor	18.486
14	Catalisador	17.617	Fotoluminescência	20.367	XPS	18.480
15	Raman	18.240	Biosensor	17.625	Graphen	18.360
Total de Palavras-chave do Biênio		**128.131**		**131.058**		**129.531**
Ranque	**2015-2016**	**TF.FI**	**1997-2016**	**TF.FI**		
1	Amazônia	78.720	Amazônia	3.263.260		
2	Biodiesel	66.154	Cosmologia	2.400.428		
3	Nanopartícula	59.211	Biodiesel	2.374.000		
4	Óleo Essencial	34.104	Estrutura Cristalina	2.090.760		
5	Citotóxico	32.065	Nanopartícula	1.928.646		
6	Adsorção	30.256	Magnetismo	1.809.632		
7	Cosmologia	29.300	Adsorção	1.563.840		
8	DFT	24.570	Filme Fino	1.466.150		
9	Nanocompósitos	19.080	Flavonoides	1.298.088		
10	Graphen	18.711	Fotoluminescência	1.207.152		
11	Nanotubo de carbono	15.104	Sol-Gel	1.156.730		
12	Antioxidante	13.440	Catalisador	1.119.410		
13	Câncer	11.323	Óleo Essencial	1.079.670		
14	Sensoriamento Remoto	11.178	Espectroscopia	986.592		
15	Estrutura Cristalina	10.829	Mercúrio	959.438		
Total de Palavras-chave do Biênio		**118.378**		**1.103.575**		

Tabela A.8 – As principais palavras-chave das Ciências Humanas ranqueadas por TF.FI em cada biênio.

Ranque	1997-1998	TF.FI	1999-2000	TF.FI	2001-2002	TF.FI
1	Memória	1.420	Educação	2.664	Amazônia	13.332
2	Educação	1.284	Cultura	2.163	Memória	2.772
3	Política	632	Gênero	1.095	Cidadania	1.808
4	Etnologia Indígena	540	Ansiedade	1.026	Governo	1.219
5	Escola	505	Etnologia Indígena	783	Ansiedade	1.120
6	Democracia	423	Ética	723	Educação	1.105
7	Desenvolvimento	252	Memória	680	Avaliação	1.080
8	Cultura	243	Política	564	Pesquisa	896
9	Aids	242	História	407	Aids	756
10	Aprendizagem	205	Filosofia	380	Cultura	662
11	Globalização	200	Amazônia	356	Gênero	600
12	Antropologia	190	Democracia	276	Psicologia	572
13	Estresse	164	Educação Matemática	249	Ensino	555
14	Brasil	158	Universidade	246	Brasil	488
15	Ensino e Pesquisa	135	Brasil	215	Prevenção	481
Total de Palavras-chave do Biênio		**44.985**		**57.622**		**72.079**
Ranque	**2003-2004**	**TF.FI**	**2005-2006**	**TF.FI**	**2007-2008**	**TF.FI**
1	Memória	9.312	Filosofia	6.656	Educação	17.400
2	Trabalho	8.910	Psicologia	4.980	Arqueologia	12.191
3	Cultura	4.741	Memória	4.676	Memória	8.272
4	Ciência	3.782	Ciência	3.575	Psicologia	5.100
5	Gênero	3.690	Amazônia	3.381	Violência	3.808
6	Ansiedade	2.139	Ansiedade	3.120	História	3.654
7	Amazônia	2.023	Educação	2.578	Infância	3.432
8	Brasil	1.603	Desenvolvimento	1.911	Depressão	3.150
9	Avaliação	1.400	Memória	1.536	Aprendizagem	2.772
10	Psicanalise	1.266	Ratos	1.144	Formação do Professor	2.733
11	Cognição	544	Arqueologia	888	Ansiedade	2.464
12	Desenvolvimento	496	Cultura	866	Adolescente	2.355
13	Escola	483	Brasil	836	Amazônia	2.282
14	Memória	475	Adolescente	826	Cultura	2.230
15	Aprendizagem	408	Aprendizagem	824	Gênero	2.180
Total de Palavras-chave do Biênio		**86.899**		**92.810**		**104.839**
Ranque	**2009-2010**	**TF.FI**	**2011-2012**	**TF.FI**	**2013-2014**	**TF.FI**
1	Memória	12.348	Brasil	40.670	Memória	11.088
2	Educação	10.836	Memória	19.845	Amazônia	10.920
3	Gênero	8.115	Educação	15.376	América Latina	10.500
4	Amazônia	5.425	Adolescente	12.006	Educação	9.455
5	Brasil	4.836	Crianças	8.448	Brasil	8.646
6	Trabalho	4.632	Amazônia	8.352	Formação do Professor	6.230
7	Políticas Públicas	3.150	Saúde	8.008	Violência	5.795
8	Cultura	2.994	Ciência	6.435	Adolescente	4.392
9	Cognição	2.914	Gênero	5.430	Idoso	4.366
10	Educação a Distância	2.568	Epidemiologia	3.906	Ansiedade	3.876
11	Depressão	2.530	Violência	3.744	Infância	3.388
12	Epidemiologia	2.448	Depressão	3.399	Depressão	3.198
13	Adolescente	2.366	Cognição	3.379	Epidemiologia	3.078
14	Ansiedade	2.220	Cultura	3.090	Cultura	2.975
15	Psicanalise	2.220	Autismo	3.010	Saúde	2.492
Total de Palavras-chave do Biênio		**126.878**		**136.053**		**136.732**
Ranque	**2015-2016**	**TF.FI**	**1997-2016**	**TF.FI**		
1	Brasil	9.744	Educação	676.848		
2	Amazônia	8.853	Memória	670.145		
3	Epidemiologia	6.683	Amazônia	547.800		
4	Educação	5.445	Brasil	533.994		
5	Memória	4.191	Cultura	242.292		
6	Cultura	3.751	Gênero	222.365		
7	Políticas Públicas	2.348	Ansiedade	185.808		
8	Arqueologia	2.136	Psicologia	168.480		
9	Sexualidade	2.080	Trabalho	166.968		
10	Ensino	2.052	Adolescente	157.353		
11	Violência	2.000	Violência	156.604		
12	América Latina	1.740	Arqueologia	153.272		
13	Psicanalise	1.544	Epidemiologia	140.140		
14	Saúde Mental	1.460	Ciência	136.500		
15	Ansiedade	1.269	Crianças	119.880		
Total de Palavras-chave do Biênio		**126.010**		**984.907**		

Tabela A.9 – As principais palavras-chave das Ciências Sociais Aplicadas ranqueadas por TF.FI em cada biênio.

Ranque	1997-1998	TF.FI	1999-2000	TF.FI	2001-2002	TF.FI
1	Tecnologia da Informação	124	Sequenciamento	1.440	Brasil	770
2	Inflação	84	Genoma	1.260	Estratégia	273
3	Brasil	70	Bioinformática	231	Desenvolvimento	234
4	Bibliometria	55	Agente patogénico	231	Discurso	128
5	Marketing	38	Xylella Fastidiosa	231	Família	96
6	Amazônia	18	DNA	204	Economia Política	62
7	Sindicato	15	Políticas Públicas	146	Tecnologia da Informação	55
8	Organização Industrial	12	Genética	125	Base de Dados	52
9	Cientometria	9	Laranja	125	Energia	46
10	Microeconometria	9	Qualidade de Vida	110	Qualidade de Vida no Trabalho	44
11	Firma	9	Citrus	102	Indústria	34
12	Infometria	6	CVC	102	Sustentabilidade	31
13	Ciência da Informação	6	Fitopatógeno	102	Pobreza	29
14	Gestão do Conhecimento	4	Planejamento Urbano	56	Amazônia	28
15	Web 2.O	3	Amazônia	36	Mulher	28
Total de Palavras-chave do Biênio		**20.124**		**27.444**		**35.847**
Ranque	**2003-2004**	**TF.FI**	**2005-2006**	**TF.FI**	**2007-2008**	**TF.FI**
1	Amazônia	1.672	Sustentabilidade	735	Gênero	1.960
2	Brasil	708	Comunicação	405	Sustentabilidade	1.810
3	Sustentabilidade	585	Saúde	396	Brasil	930
4	Cultura	132	Brasil	350	Mercado de Trabalho	693
5	Mercado de Trabalho	108	Competitividade	306	Estratégia	600
6	Saúde	92	Qualidade de Vida	300	América Latina	497
7	Economia	75	Amazônia	270	Produção Científica	448
8	Governança	72	Empreendedorismo	232	Design	392
9	Capitalismo	68	Antropologia	208	Amazônia	325
10	Crescimento Econômico	62	Reforma Agrária	192	Pobreza	305
11	Peptídeos	48	Educação	164	Governança	261
12	Desenvolvimento Rural	40	Produção Científica	150	Educação	196
13	Pobreza	37	Financiamento	148	Marketing	187
14	Bioética	36	Planejamento Urbano	138	Inovação	165
15	Comércio Eletrônico	36	Avaliação	132	Gestão do Conhecimento	159
Total de Palavras-chave do Biênio		**46.639**		**56.636**		**63.855**
Ranque	**2009-2010**	**TF.FI**	**2011-2012**	**TF.FI**	**2013-2014**	**TF.FI**
1	Inovação	2.460	Brasil	6.688	Marketing	5.962
2	Brasil	2.430	Amazônia	2.700	Inovação	5.180
3	Saúde	2.176	Inovação	2.604	Brasil	4.914
4	Consumismo	1.581	Design	2.184	Bibliometria	3.325
5	Amazônia	1.224	Sustentabilidade	1.820	Amazônia	3.036
6	Marketing	955	Gestão do Conhecimento	1.323	Sustentabilidade	1.980
7	Gestão do Conhecimento	796	Agronegócio	1.253	Epidemiologia	1.584
8	Educação	786	Marketing	1.225	Agricultura Familiar	1.467
9	Políticas Públicas	748	Comunicação	1.122	Administração	1.160
10	Design	650	Estratégia	1.092	Estratégia	1.146
11	Turismo	645	Biodiesel	1.080	Gestão do Conhecimento	930
12	Internet	588	Políticas Públicas	1.004	Agricultura	910
13	Ciência da Informação	417	Governo	925	Gênero	852
14	Bibliometria	414	Consumismo	903	Cultura	776
15	Energia	396	Energia	888	Gestão Pública	690
Total de Palavras-chave do Biênio		**77.286**		**89.261**		**96.237**
Ranque	**2015-2016**	**TF.FI**	**1997-2016**	**TF.FI**		
1	Brasil	9.100	Brasil	176.349		
2	Comunicação	4.491	Amazônia	98.464		
3	Sustentabilidade	4.096	Marketing	84.182		
4	Bibliometria	3.380	Inovação	79.648		
5	Inovação	3.090	Comunicação	54.950		
6	Universidade	2.720	Sustentabilidade	49.140		
7	Educação	2.601	Bibliometria	39.900		
8	Marketing	2.574	Educação	35.460		
9	Amazônia	2.254	Estratégia	33.840		
10	Interdisciplinar	1.716	Saúde	31.450		
11	Empreendedorismo	1.416	Design	29.512		
12	Estratégia	957	Políticas Públicas	28.960		
13	Agricultura Familiar	912	Gênero	27.864		
14	Internet	798	Gestão do Conhecimento	26.841		
15	Ensino Superior	756	Consumismo	21.917		
Total de Palavras-chave do Biênio		**96.195**		**609.524**		

Tabela A.10 – As principais palavras-chave das Engenharias ranqueadas por TF.FI em cada biênio.

Ranque	1997-1998	TF.FI	1999-2000	TF.FI	2001-2002	TF.FI
1	Filme fino	560	Compósitos	3.434	Compósitos	4.928
2	Elementos Finitos	546	Modelagem	2.808	Filme fino	4.819
3	Modelagem	495	Filme fino	2.652	Cerâmicas	4.000
4	Otimização	450	Biomateriais	2.464	Microestrutura	3.876
5	Zeólito	406	Elementos Finitos	1.980	Modelagem	3.569
6	Alumina	352	Alumina	1.755	Cristalização	3.105
7	Transporte	348	Sol-Gel	1.450	Reologia	2.541
8	Compósitos	343	Cerâmicas	1.340	Biomateriais	2.340
9	Tokamak	342	Corrosão	1.140	Corrosão	2.320
10	Cerâmicas	308	Microestrutura	1.008	Varistor	2.128
11	Sno2	300	Cristalização	925	Permeabilidade	2.025
12	Corrosão	280	Redes Neurais	784	Processamento	1.920
13	Polianilina	266	Propriedade Mecânica	697	Elementos Finitos	1.764
14	Diamante CVD	265	Sinterização	660	Concreto	1.683
15	Varistor	264	Óculos	648	Redes Neurais	1.488
Total de Palavras-chave do Biênio		**24.414**		**35.336**		**47.855**
Ranque	**2003-2004**	**TF.FI**	**2005-2006**	**TF.FI**	**2007-2008**	**TF.FI**
1	Filme fino	13.110	Compósitos	17.751	Compósitos	18.998
2	Cerâmicas	12.441	Microestrutura	12.792	Cerâmicas	16.016
3	Compósitos	9.525	Cerâmicas	12.090	Filme fino	15.456
4	Microestrutura	8.910	Filme fino	6.834	Biomateriais	13.720
5	Biomateriais	5.529	Modelagem	6.264	Propriedade Mecânica	11.286
6	Modelagem	4.100	Propriedade Mecânica	6.188	Fotoluminescência	9.870
7	Nanopartícula	3.952	Otimização	5.967	Otimização	9.840
8	Otimização	3.434	Biomateriais	4.815	Biodiesel	9.118
9	Sinterização	3.174	Sensores	4.067	Nanocompósitos	8.633
10	Propriedade Mecânica	3.008	Cristalografia de Raios X	3.886	Cristalografia de Raios X	7.315
11	Lipase	2.744	Fotoluminescência	3.854	Adsorção	7.008
12	Fluído Magnético	2.610	Lipase	3.717	Microestrutura	6.723
13	Fotoluminescência	2.544	Corrosão	3.483	Modelagem	6.314
14	Cristalização	2.450	Adsorção	3.364	Lipase	6.097
15	Alumina	2.409	Redes Neurais	3.234	Corrosão	4.824
Total de Palavras-chave do Biênio		**60.557**		**67.536**		**71.765**
Ranque	**2009-2010**	**TF.FI**	**2011-2012**	**TF.FI**	**2013-2014**	**TF.FI**
1	Biodiesel	49.368	Biodiesel	130.634	Biodiesel	112.917
2	Nanocompósitos	28.032	Compósitos	28.310	Biomateriais	29.172
3	Compósitos	23.807	Nanocompósitos	25.758	Nanocompósitos	27.789
4	Biomateriais	20.150	Biomateriais	20.881	Biomassa	20.300
5	Otimização	12.998	Adsorção	16.107	Compósitos	20.230
6	Cerâmicas	12.480	Biomassa	15.000	Modelagem	19.685
7	Lipase	11.815	Lipase	12.782	Adsorção	17.696
8	Propriedade Mecânica	11.655	Propriedade Mecânica	12.096	Propriedade Mecânica	15.836
9	Modelagem	11.328	Modelagem	10.176	Microestrutura	12.760
10	Microestrutura	10.890	Quitosana	9.956	Nanopartícula	11.988
11	Adsorção	10.449	Otimização	9.628	Etanol	10.971
12	Fotoluminescência	7.888	Etanol	9.588	Corrosão	10.864
13	Biodegradável	7.705	Corrosão	9.559	Cerâmicas	10.692
14	Corrosão	6.532	Nanopartícula	8.772	Lipase	9.638
15	Alumina	6.384	Cerâmica	8.732	Otimização	7.600
Total de Palavras-chave do Biênio		**80.820**		**86.105**		**90.713**
Ranque	**2015-2016**	**TF.FI**	**1997-2016**	**TF.FI**		
1	Biodiesel	98.400	Biodiesel	1.891.960		
2	Compósitos	31.040	Compósitos	1.396.200		
3	Biomateriais	27.522	Biomateriais	1.057.614		
4	Biomassa	24.244	Cerâmicas	798.950		
5	Adsorção	23.868	Microestrutura	723.095		
6	Microestrutura	20.184	Nanocompósitos	693.666		
7	Nanopartícula	19.190	Modelagem	663.884		
8	Propriedade Mecânica	18.696	Propriedade Mecânica	634.848		
9	Nanocompósitos	18.592	Otimização	591.384		
10	Otimização	15.678	Adsorção	556.250		
11	Modelagem	10.890	Filme fino	492.600		
12	Modelo Matemático	9.559	Lipase	420.486		
13	Cerâmicas	9.545	Biomassa	407.037		
14	Etanol	9.159	Nanopartícula	353.363		
15	Simulação	7.663	Corrosão	347.422		
Total de Palavras-chave do Biênio		**85.026**		**650.127**		

Tabela A.11 – As principais palavras-chave da Linguística, Letras e Artes ranqueadas por TF.FI em cada biênio.

Ranque	1997-1998	TF.FI	1999-2000	TF.FI	2001-2002	TF.FI
1	Astrócito	10	Psicolinguística	264	Cultura	588
2	Análise Discursiva	10	Tradução	228	Fonologia	38
3	Minimalismo	5	Teatro	83	Austin	18
4	Discurso da Mídia	4	Sintaxe	23	Fonética	16
5	Interação Social	3	Linguagem	19	Correlação	12
6	Preconceito	3	Redes Neurais	11	Ritual	12
7	Triflato	3	Neurociência	11	Glutamato	8
8	Hippo Campus	2	Afasiologia	11	Astrócito	8
9	Gfap	2	Hidrocarboneto	10	Difusão	6
10	Carbeto de Silício	2	Carbocátion	10	Campo	4
11	Lítio	2	Shakespeare	9	Hippo Campus	2
12	Gliosis	2	Língua e Cultura	6	S100B	2
13	Gênero Textual	2	Gramática Gerativa	5	Excitotoxicidade	2
14	Poli Metil Silano	2	Pirólise	4	Soro	2
15	Negociação	1	Cerâmica	3	Movimento Facial	2
Total de Palavras-chave do Biênio		**13.963**		**16.863**		**21.306**
Ranque	**2003-2004**	**TF.FI**	**2005-2006**	**TF.FI**	**2007-2008**	**TF.FI**
1	Memória	250	Memória	590	Escrita	158
2	Neurolinguística	26	Cognição	82	Arte	134
3	Apoptose	15	Escrita	68	Educação	106
4	Lítio	8	Neurolinguística	30	Estética	61
5	Minimalismo	8	Aprendizagem de Línguas	26	Psicanalise	37
6	Taxonomia	3	Estresse Crônico	24	Esclerose Múltipla	28
7	Neuroproteção	2	Lítio	24	Neurite Optica	28
8	Hippo Campus	2	Bilinguismo	24	Musicoterapia	28
9	Estresse Crônico	2	Walter Benjamin	18	Política	20
10	Léxico mental	2	Tempo	18	Bilinguismo	19
11	Processamento de Sinais	2	Bakhtin	16	Prosódia	17
12	Neuroinformática	2	Ciência	13	Compostos Fenólicos	10
13	Gênero Gramatical	2	Hippo Campus	12	Matrinxã	10
14	Imunossupressão	1	Depressão	12	Português Brasileiro	7
15	Caninos	1	Antioxidante	8	Neuroproteção	6
Total de Palavras-chave do Biênio		**25.039**		**26.939**		**29.841**
Ranque	**2009-2010**	**TF.FI**	**2011-2012**	**TF.FI**	**2013-2014**	**TF.FI**
1	Discurso	206	Identidade	884	Memória	696
2	Sociolinguística	159	Linguagem	336	Literatura	612
3	Cultura	152	Cultura	320	Bilinguismo	504
4	Fotografias	93	Música	292	Violência	495
5	Humor	54	Fonologia	180	Música	365
6	Vocabulário	30	Cognição	132	Tradução	182
7	Musicoterapia	26	Gênero	120	Poesia	160
8	Infância	23	Fonética	99	Dança	57
9	Bilinguismo	22	Pragmatismo	96	Fonologia	55
10	Resenha	17	Ideologia	86	Mídia	44
11	Ambiguidade	16	Língua Inglesa	68	Preconceito Linguístico	39
12	Esclerose Múltipla	12	Dança	62	Controle Executivo	36
13	Contexto	10	Argumentação	60	Trauma	24
14	Identidade	9	Política	41	Compreensão	24
15	Linguagem	8	Bilinguismo	38	Fonética	23
Total de Palavras-chave do Biênio		**37.353**		**39.646**		**37.114**
Ranque	**2015-2016**	**TF.FI**	**1997-2016**	**TF.FI**		
1	Literatura	315	Memória	15.000		
2	Narrativa	264	Cultura	11.407		
3	Psicolinguística	126	Discurso	7.368		
4	Morfologia	81	Tradução	6.670		
5	Ultrassonografia	72	Identidade	6.076		
6	Opera	66	Linguagem	5.530		
7	Ressonância Magnética	30	Bilinguismo	5.348		
8	Voz	28	Literatura	5.254		
9	Português	24	Música	5.202		
10	Bilinguismo	22	Psicolinguística	4.321		
11	Processamento	18	Poesia	3.298		
12	Corpus	15	Fonologia	2.688		
13	Diabete mellitus	14	Narrativa	2.640		
14	Metodologia de Pesquisa	13	Violência	2.493		
15	Filme Fino	12	Escrita	2.133		
Total de Palavras-chave do Biênio		**34.212**		**282.276**		