Uma Plataforma para a Extração, Integração e Análises de Dados Científicos em Acesso Aberto

Patrícia Mascarenhas Dias[1], Thiago Magela Rodrigues Dias[2]

[1] 0000-0002-8448-6874, Doutoranda, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, MG, Brasil. patriciamdias@gmail.com
[2] 0000-0001-5057-9936, Doutor, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, MG, Brasil. thiagomagela@cefetmg.br

Tipo de trabalho: comunicação

Palavras-chave: Acesso Aberto, Plataforma Lattes, Extração de Dados.

1 Introdução

Uma nova geração de serviços disponíveis principalmente na Web está mudando a forma de divulgar e disponibilizar a produção científica e tecnológica. Existe, atualmente, uma tendência que reforça a troca de informações e a colaboração entre as pessoas. A forte relação entre os domínios científico e socioeconômico tem gerado um interesse crescente pela compreensão dos mecanismos que norteiam as atividades científicas, sendo possível apontar diversos trabalhos que analisam aspectos específicos como as características da linguagem e dos discursos empregados na comunicação científica (HOFFNAGEL, 2009) ou, ainda, a relação de colaboração entre pesquisadores e grupos de pesquisa (DING, 2011; REVOREDO et al., 2012; STROELE; ZIMBRÃO; SOUZA, 2012).

Para Mugnaini et al. (2014), o levantamento da produção científica de um país permite estudar diversos aspectos que podem ser qualificados como resultados mensuráveis de seu respectivo sistema de ciência, tecnologia e inovação. Acompanhar o fluxo de comunicação científica das diversas áreas facilita o processo de avaliação dos resultados de pesquisa, cujas características são tão diversificadas quanto o é a própria ciência. Para os autores, a análise quantitativa da ciência, que se dá a partir de indicadores, é tida como complementar à análise qualitativa, realizada pelos pares em diversas etapas, desde a formação, com a entrada e progressão na carreira de cientista, até a avaliação das pesquisas empreendidas, consubstanciada em manuscritos e títulos.

Com a competição cada vez mais acirrada entre as instituições de pesquisa, torna-se importante para seus integrantes determinar a abrangência de suas publicações e a qualidade de seus trabalhos quando comparados aos de outros grupos, bem como identificar potenciais colaboradores com o intuito de realizar trabalhos em colaboração visando impulsionar a sua produção científica e obter melhores resultados em suas pesquisas. É possível apontar trabalhos recentes nos quais é mostrado que grupos de pesquisa com uma rede social científica bem conectada tendem, geralmente, a ser mais produtivos (LOPES et al., 2011; BRANDÃO et al., 2013). Para Wanderley et al. (2014), evidências apontam que a forma como os pesquisadores colaboram tem forte impacto sobre a sua produtividade.

No entanto, o grande volume de dados sobre produção científica disponível em diferentes formatos e em diferentes repositórios dificulta a realização de estudos e consultas por parte de usuários que necessitam de uma visão unificada desses dados para, por exemplo, possibilitar a identificação de grupos de indivíduos que estejam trabalhando com determinado tema em diferentes instituições ou regiões. Além disso, o crescimento e a evolução da Web deram origem a uma grande quantidade de dados textuais pouco estruturados, heterogêneos e armazenados, sem nenhuma preocupação com padronização em diferentes repositórios. Tais características requerem muitas vezes a utilização de técnicas específicas para integração e conciliação de dados provenientes de diferentes fontes (LIU, 2011; CHRISTEN, 2012).

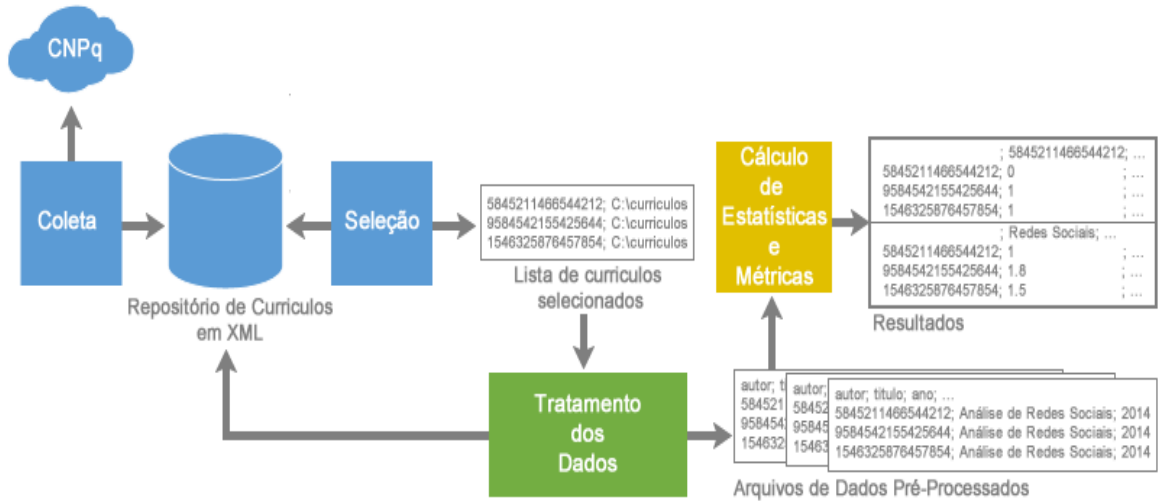
Estudos bibliométricos, principalmente em grandes repositórios bibliográficos, não são tarefas triviais tendo em vista a quantidade de dados a serem analisados e as características dos repositórios, que, em sua maioria, não possuem um padrão definido. Atualmente, grande parte desses estudos tem utilizado como principais fontes de dados resultados de consultas a repositórios internacionais que apresentam dados sobre trabalhos científicos, geralmente publicados em periódicos indexados. Entretanto, muitos desses repositórios negligenciam trabalhos publicados em periódicos nacionais que muitas vezes não são indexados e grande parte dos artigos publicados em anais de congressos, que constituem importante meio de publicação de algumas áreas do conhecimento como, por exemplo, a Ciência da Computação (LAENDER et al., 2008).

Logo, fica evidente a dificuldade que existe para se realizar estudos abrangentes que possam apresentar de forma ampla análises sobre toda a produção científica de um grande conjunto de indivíduos que estejam vinculados a diferentes instituições ou que atuam em áreas distintas, como, por exemplo, o conjunto de todos os pesquisadores de um determinado país. Diante disso, este trabalho apresenta um estudo sobre a produção científica registrada nos currículos cadastrados na Plataforma Lattes, possibilitando, desta forma, que o framework proposto neste trabalho, possa também ser utilizado em fontes distintas de dados em acesso aberto para estudos bibliométricos e baseados em colaborações científicas.

2 Materiais e Métodos

Um dos propósitos deste trabalho é utilizar tecnologias envolvidas no processo de extração de dados da Web para realizar a coleta de todos os currículos da Plataforma Lattes. Com os currículos extraídos, outras consultas são realizadas com o intuito de enriquecer ainda mais os dados obtidos, como, por exemplo, número de citações dos artigos ou informações adicionais como a qualidade dos meios de publicação destes artigos. Para isso, um arcabouço denominado LattesDataXplorer foi desenvolvido (Figura 1).

Figura 1: Visão geral do LattesDataXplorer.



O LattesDataXplorer é responsável por englobar todo o conjunto de técnicas e métodos para a coleta, tratamento e análise dos dados utilizados neste trabalho. Ele é composto por um conjunto de componentes que são responsáveis por todo o processo de coleta e tratamento dos dados. O processo de extração de todos os dados curriculares da Plataforma Lattes é dividido em três componentes que objetivam minimizar o custo computacional: 1) extração de URLs, que é responsável por extrair as referências únicas para todos os currículos cadastrados, e dessa forma, possibilitar o acesso individual a cada currículo, 2) extração de Ids e Data, que visa acessar cada currículo e extrair o seu identificador individual, bem como a data de última atualização, 3) extração de currículos, que é responsável por extrair e armazenar os currículos cuja data de atualização na Plataforma Lattes seja divergente da data de atualização do currículo armazenado localmente.

Todas essas etapas se fazem necessárias, já que o ideal é manter os dados curriculares atualizados com a maior frequência possível, possibilitando a realização de análises com dados atualizados, tendo em vista que, com a estratégia adotada, não se faz necessário coletar todo o repositório de dados a cada nova extração. É importante, ainda, considerar que os currículos atualizados podem ter novos dados inseridos, bem como, a alteração e exclusão de dados já registrados, o que torna o processo de atualização de campos específicos uma tarefa complexa.

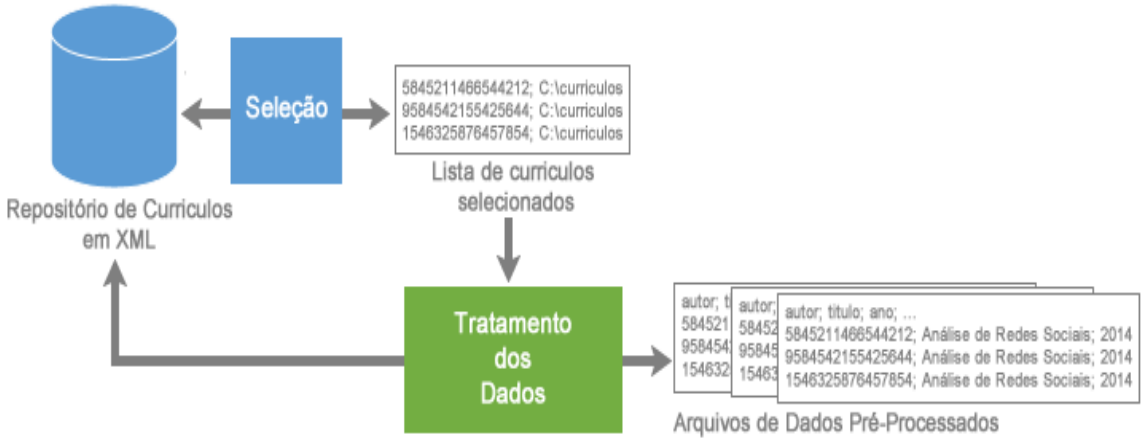
Com todos os currículos armazenados localmente em formato XML, a possibilidade de manipulação dos dados com flexibilidade permite explorar todo o potencial que os dados curriculares da Plataforma Lattes oferecem. Com o intuito de analisar grupos específicos de currículos, como os formados por docentes de um programa de pós-graduação ou de uma instituição em particular, é descrito o componente denominado Seleção, para a composição de subgrupos de currículos baseada em informações presente em seus registros.

Diante disso, todas as etapas posteriores à geração de grupos específicos utilizam a lista de currículos selecionados no momento de identificar quais serão os currículos a serem analisados. As listas ficam armazenadas em um repositório de listas, em que várias listas podem ser geradas e sendo estas referenciadas no momento de análise. Logo, o componente de seleção se apresenta como uma importante facilidade para a seleção de subgrupos dos currículos a serem analisados.

Após a identificação dos subgrupos, estudos que considerem somente um conjunto particular de indivíduos ou de todo o repositório podem ser realizados.

Na etapa de identificação e caracterização das redes de colaboração científica, o componente de Tratamento dos Dados é responsável por analisar os currículos dos grupos selecionados e gerar as redes de colaboração científica. O componente utiliza a lista de currículos selecionados, gerada para um determinado grupo, o que possibilita identificar quais são os currículos que vão ser analisados. A lista possui o identificador e o local de armazenamento do referido currículo e, diante desta indicação, apenas os currículos presentes na lista são considerados. Além disso, o componente também é responsável por analisar cada um dos currículos, de tal forma que sejam produzidos arquivos de dados pré-processados que vão possibilitar a realização de análises bibliométricas da produção científica, como também da colaboração científica dos indivíduos que compõem os grupos em análise (Figura 2).

Figura 2: Processo de tratamento dos dados.



Para a identificação de colaborações, podem ser utilizados vários elementos que permitam relacionar pares de indivíduos, como: publicação de trabalhos em conjunto (ou seja, coautorias), orientações, participação em bancas de avaliação, organização de eventos, participação em projetos de pesquisa, dentre outros. Diante disso, especificamente na identificação de colaboração científica em publicações, um pré-processamento nos títulos dos trabalhos para a identificação é realizado.

Em um documento textual, podem ser encontrados diversos tokens que não possuem valor semântico. Estes tokens são importantes para melhor entendimento do texto de forma geral e também são classificados como stopwords. As stopwords compõem uma lista denominada stoplist, a qual é frequentemente utilizada em sistemas de mineração de textos. Palavras que compõem a stoplist são removidas do documento original, reduzindo o texto a ser analisado, de modo a não provocar problemas no conteúdo semântico do documento. Uma stoplist bem elaborada permite a eliminação de muitos termos irrelevantes, tornando mais eficiente o resultado obtido pelo processo de mineração de textos (CARRILHO-JUNIOR, 2007). A normalização e a padronização consistem em remover caracteres especiais, letras com acentuação e espaços, e deixar a string resultante com todas as letras em minúsculo.

De posse de todos os currículos armazenados e utilizando a lista de currículos previamente selecionados, que especifica o subgrupo de indivíduos a serem analisados, é proposto um método

para a identificação de colaborações científicas. Para isso, todos os títulos dos artigos cadastrados no currículo de cada um dos autores são analisados e, consequentemente, se tornam a base para construção da rede de colaboração. Para auxiliar o processo de caracterização, é utilizado um dicionário que possibilita vincular os artigos (chaves do dicionário) a seus autores (identificadores).

Ao considerar o custo computacional, a adoção de métodos que trabalham com comparação de títulos entre as publicações torna-se inviável quando é necessário o processamento de uma grande quantidade de dados, como é o caso de todos os currículos inseridos na Plataforma Lattes, tendo em vista o tempo de resposta para avaliar redes de grande porte com centenas de milhares de títulos. Já o método utilizado pelo LattesDataXplorer possui um custo computacional para n produções da ordem de $\theta(n)$ comparações, já que a única comparação realizada é a da existência da chave no dicionário, ou seja, a única comparação realizada com cada chave (título) transformada é se ela existe ou não no dicionário de identificação utilizado para a caracterização das redes.

Além da identificação das colaborações, são gerados arquivos de dados pré-processados que contêm todas as informações dos nós que vão compor a rede. Estes arquivos possuem informações como titulação, proficiência, afiliação, áreas de pesquisa, dentre outras. Estes arquivos são importantes, pois, diante das informações que eles agregam, é possível realizar diversas outras análises bibliométricas sem a necessidade de consultar o conjunto de currículos novamente.

Como resultados das etapas de tratamentos dos dados, são produzidos arquivos de dados pré-processados que contêm todas as informações necessárias para o cálculo de diversas métricas bibliométricas e baseadas em análise de redes sociais. O cálculo das estatísticas e métricas utilizando esses arquivos é facilitado, já que eles sumarizam todos os dados contidos nos currículos. Consequentemente, esses arquivos se tornam a base para o componente de Cálculo de Estatísticas e Métricas.

3 Resultados

É importante destacar, ainda, a diversidade dos dados registrados no conjunto de currículos considerado, que referem-se a artigos publicados em anais de congresso e em periódicos, apresentação de trabalhos científicos, participação em eventos, nível de formação acadêmica, orientações realizadas, dentre outros. É importante ressaltar, ainda, que um determinado trabalho pode estar registrado em currículos distintos, já que pode ter sido realizado em colaboração envolvendo mais de um indivíduo. Logo, no repositório da Plataforma Lattes, um trabalho pode aparecer várias vezes, tendo em vista que ele pode ter sido registrado por cada um de seus autores. A Tabela 1 apresenta o quantitativo geral de todos os trabalhos registrados nos currículos dos doutores coletados para a geração da coleção.

Tabela 1: Quantitativo dos dados dos currículos dos doutores em abril de 2018.

Tipo de Trabalho	*Quantidade*
Artigos em Anais de Congresso	10.548.672
Artigos em Periódico	5.458.385
Capítulos de Livro	1.351.632
Livros	522.634
Textos em Jornais e Revistas	1.072.786
Trabalhos Técnicos	1.547.435
Outras Produções Bibliográficas	904.976

A quantidade de dados registrada corrobora a importância da Plataforma Lattes, confirmando a sua condição de um dos principais repositórios de dados científicos atualmente existentes em todo o mundo (LANE, 2010) e caracterizando-se como uma fonte extremamente rica para análise da produção científica brasileira. A partir da Tabela 1, é possível observar a tendência de publicação de artigos em anais de congresso, seguida em menor número pela publicação de artigos em periódicos e de capítulos de livro.

4 Considerações Finais

Considerando o grande interesse de diversos trabalhos recentes que visam analisar dados de publicações científicas, os conjuntos de dados identificados neste trabalho caracterizam-se como importante fonte de informação para diversos novos estudos em diferentes áreas. Por sumarizar dados específicos, como produção científica, formação acadêmica e orientações, os conjuntos de dados possíveis de serem extraídos dos currículos cadastrados na Plataforma Lattes, possibilitam diversos novos estudos.

5 Referências

BRANDÃO, M. A. et al. Using link semantics to recommend collaborations in academic social networks. In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON WORLD WIDE WEB COMPANION, 22, 2013, Anais …, Rio de Janeiro, p. 833-840, 2013.

CARRILHO-JUNIOR, J. R. Desenvolvimento de uma Metodologia para Mineração de Textos. 2007. 113 p. (Mestrado). Departamento de Engenharia Elétrica, Pontífica Universidade Católica do Rio de Janeiro, 2007.

CHRISTEN, P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, Berlin, 2012.

DING, Y. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. Informetrics, v. 5, n. 1, p. 187-203, 2011.

HOFFNAGEL, J. C. A prática de citação em trabalhos acadêmicos. Cadernos de Linguagem e Sociedade, v. 10, n. 1, p. 71, 2009.

LAENDER, A. H. F. et al. Assessing the research and education quality of the top Brazilian Computer Science graduate programs. SIGCSE Bulletin, v. 40, n. 2, p. 135-145, 2008.

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

Liu, B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Second Edition, Springer, Berlin, 2011.

LOPES, G. R.; et al. Ranking Strategy for Graduate Programs Evaluation. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY AND APPLICATIONS - IEEE, 7, Anais… Austrália, p. 59-64, 2011.

MUGNAINI, R. et al. Comunicação científica no Brasil (1998-2012): indexação, crescimento, fluxo e dispersão. Transinformação, v. 26, n. 3, p. 239-252, 2014.

REVOREDO, K. et al. Mining scientific literature for analysis of collaboration in research communities. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 1, Curitiba, Anais… Curitiba, 2012.

STRÖELE, V.; ZIMBRÃO, G.; SOUZA, J. M. Análise de redes sociais científicas: modelagem multi-relacional. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 1, Curitiba, Anais … Curitiba, 2012.

WANDERLEY, A. J. et al. Identificando correlações entre métricas de Análise de Redes Sociais e o h-index de pesquisadores de Ciência da Computação. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS DND MINING, 3, Brasília, Anais… Brasília, 2014.