cadernos BAD

9ª conferência luso-brasileira acessoaberto
2 a 4 de outubro 2018 - ISCTE-IUL - Lisboa - Portugal

Obtenção de dados científicos a partir de repositórios de dados curriculares

Thiago Magela Rodrigues Dias

CEFET-MG

thiagomagela@gmail.com

Elaine Rosangela de Oliveira Lucas

UDESC

lanilucas@gmail.com

Patricia Mascarenhas Dias

CEFET-MG

patriciamdias@gmail.com

Gray Farias Moita

CEFET-MG

gray@dpp.cefetmg.br

Resumo

Os estudos sobre dados científicos têm atraído o interesse de pesquisadores de diversas áreas do conhecimento, tendo em vista seu potencial para melhor compreender como as pesquisas em uma determinada área têm sido realizadas, ou como grupos de pesquisadores têm colaborado no desenvolvimento de seus trabalhos. Assim, este estudo descreve o processo de extração, tratamento e caracterização de uma coleção de dados contendo informação científicas sobre os indivíduos com currículos cadastrados na Plataforma Lattes. O trabalho também apresenta uma descrição quantitativa sobre os dados coletados, bem como uma descrição geral dos conjuntos de dados extraídos.

Palavras-chave: Plataforma Lattes, Dados Curriculares, Dados Científicos

Obtaining scientific data from curricular data repositories

Abstract

Studies on scientific data have attracted the interest of researchers in various fields of knowledge, in view of their potential to better understand how research in a given area has been carried out, or how groups of researchers have collaborated in the development of their work. Thus, this study describes the process of extracting, treating and characterizing a collection of data containing scientific information about individuals with curricula registered in the Lattes Platform. The paper also presents a quantitative description of the data collected, as well as a general description of the extracted data sets.

Key-words: Lattes Platform, Curriculum Data, Scientific Data

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), bem como a relação de colaboração entre pesquisadores e grupos de pesquisa (DING, 2011; REVORENDO et al., 2012; STROELE, ZIMBRÃO E 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 a própria ciência. No entanto, o grande volume de dados sobre produção científica disponível em diferentes formatos e em diferentes repositórios dificulta a realização de estudos, bem como a consulta 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.

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 geralmente 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).

Assim, é evidente a dificuldade existente para se realizar estudos abrangentes que possam apresentar, de forma ampla, análises sobre a produção científica de um grande conjunto de indivíduos que estejam vinculados a diferentes instituições ou que atuem em áreas distintas, como, por exemplo, o conjunto de todos os pesquisadores com um determinado nível de formação ou de uma determinada área de atuação. Diante disso, este trabalho apresenta uma coleção de dados estratificados extraídos dos currículos de todos os indivíduos com doutorado concluído cadastrados na Plataforma Lattes. Essa coleção, inclui dados sobre a formação, orientações concluídas e em andamento, produção científica e colaborações desses indivíduos, possibilitando, desta forma, a realização de diversos estudos sobre esse segmento de pesquisadores brasileiros.

Trabalhos Relacionados

Em um trabalho pioneiro, Mena-Chalco, Digiampietri e Oliveira (2012) analisam os programas de Ciência da Computação brasileiros, identificando o rápido crescimento em termos de produção bibliográfica e formação acadêmica ocorrido na área. Nesse trabalho, os autores apresentam uma descrição do perfil de produção acadêmica dos programas de Ciência da Computação avaliados pela CAPES nos triênios 2004-2006 e 2007-2009, tendo como base as publicações e orientações concluídas listadas nos currículos dos docentes associados a esses programas disponíveis na Plataforma Lattes. A identificação do perfil de produção acadêmica desses docentes envolveu quatro etapas: (1) identificação dos programas de pós-graduação avaliados pela CAPES nos triênios 2004-2006 e 2007-2009, (2) identificação dos docentes permanentes associados a cada um dos programas de pós-graduação considerados, (3) coleta dos currículos Lattes dos docentes associados a cada programa e, por fim, (4) extração das produções acadêmicas de cada programa (produções bibliográficas e orientações concluídas). Os resultados obtidos pela análise mostram que a área de Ciência da Computação no Brasil caracteriza-se por publicar, preferencialmente, trabalhos completos em anais de congressos (54%), aparecendo em segundo lugar a produção de artigos completos em periódicos (14%). Ainda nesse mesmo trabalho, os autores mostram que, no segundo triênio (2007-2009), houve um aumento na produção de artigos em periódicos em relação ao triênio anterior (2004-2006), influenciado provavelmente pela adoção por parte da CAPES de uma nova sistemática para a composição do Qualis.

Boaventura et al. (2014) apresentam uma caracterização da evolução das redes de colaboração científica brasileiras, representadas pelas redes de coautoria das seguintes universidades brasileiras: UFAM, UFMG, UFPE, UFRGS, UFRJ, UNB, UNICAMP e USP. Essa caracterização abrange os anos entre 2000 e 2013, tendo como foco os pesquisadores (docentes) dessas universidades. Métricas tradicionais de redes sociais, tais como densidade e diâmetro, grau de centralidade, grau de proximidade e grau de intermediação, foram

adotadas para análise das redes de coautoria. Os autores apresentam também uma análise da endogamia dessas instituições e a sua correlação com os resultados da avaliação trienal dos Programas de Pós-Graduação realizada pela CAPES. Por fim, os autores identificam a existência de diversos grupos de pesquisa, representados por docentes que frequentemente publicam em conjunto. As principais conclusões da análise apresentada pelos autores são as seguintes: (1) as redes de coautoria das universidades analisadas possuem a mesma característica de densificação das colaborações entre os seus docentes, sendo que os seus diâmetros relativamente pequenos quando comparados às suas cardinalidades indicam que elas possuem características de redes de mundo pequeno ; (2) universidades com maior porcentagem de Programas de Pós-Graduação com conceitos CAPES acima de 5 possuem valores menores de endogamia, ou seja, seus docentes colaboram mais com colegas de outras instituições.

Lima et al. (2015) fazem uma avaliação do desempenho dos principais pesquisadores que atuam na área de Ciência da Computação. Para isso são utilizados dados extraídos da Plataforma Lattes referentes a 406 pesquisadores bolsistas de produtividade em pesquisa do CNPq da área enquadrados nas cinco modalidades da bolsa (1A, 1B, 1C, 1D e 2). A avaliação considerou três dimensões centrais, sendo: tempo de carreira do pesquisador correspondente ao total de anos após a conclusão do doutorado, quantidade de alunos orientados e produtividade científica medida em função do volume de publicações e citações. Com relação ao tempo de carreira dos pesquisadores, os autores observaram que em geral aqueles que possuem os níveis mais elevados de bolsa também possuem maior tempo de carreira. Já com relação às orientações, a proporção de alunos de mestrado orientados durante a carreira dos pesquisadores é de aproximadamente um aluno por ano, para praticamente todos os níveis. No caso da produção científica dos pesquisadores, observou-se que o volume de publicações aumenta com o nível dos pesquisadores, exceto no caso do nível 1A, que se assemelha ao 1C, por incluir pesquisadores mais antigos cujos programas, no início de suas carreiras, ainda não tinham atingido a maturidade. Em resumo, a avaliação realizada pelos autores demonstrou haver coerência entre as dimensões avaliadas e as modalidades das bolsas em que os pesquisadores se enquadram na área de Ciência da Computação.

Materiais e Métodos

Para a geração da coleção de dados apresentada neste trabalho, foram coletados da Plataforma Lattes os currículos de todos os doutores ali registrados. Segundo Lane, em artigo publicado na revista Nature (LANE, 2010), medir e avaliar o desempenho acadêmico de seus pares é um fator crucial para qualquer comunidade científica. A autora descreve esforços empregados para a construção de repositórios confiáveis de dados científicos que poderiam permitir análises com o objetivo de explorar e compreender como a ciência tem evoluído. Embora tais esforços sejam importantes, alguns apresentam problemas que comprometem o sucesso dessas iniciativas. Neste cenário, a Plataforma Lattes é citada como exemplo de boas práticas para o fornecimento de dados de alta qualidade sobre a produção científica de um país e de como a sua utilização tem sido incentivada por órgãos federais, instituições

acadêmicas e agências de fomento a pesquisa. Por fim, a autora destaca que a Plataforma Lattes é uma das fontes de dados sobre pesquisadores mais confiáveis existentes atualmente.

Para Ferraz, Quoniam e Maccari (2014), até o presente momento, não existe no mundo um repositório curricular nacional semelhante à Plataforma Lattes, sendo que somente repositórios de dados referenciais, de onde se podem extrair referências bibliográficas, e fontes de informação secundárias estão disponíveis para livre acesso. Dessa forma, a Plataforma Lattes é um instrumento da maior importância para o estudo da produção científica brasileira.

Tendo em vista a ausência de uma interface de consulta e recuperação de dados dos currículos na Plataforma Lattes, e considerando todo o potencial dos dados armazenados, estratégias para coleta e análises dos dados se fazem necessárias. Para a extração e tratamento do conjunto de dados, utilizou-se um arcabouço denominado LattesDataXplorer (Figura 1) desenvolvido especificamente para a coleta, extração e tratamento de dados da Plataforma Lattes (DIAS, 2016). Esse arcabouço adota técnicas usualmente empregadas na coleta e extração de dados de documentos disponíveis na Web para realizar essas tarefas sobre os currículos da Plataforma Lattes.

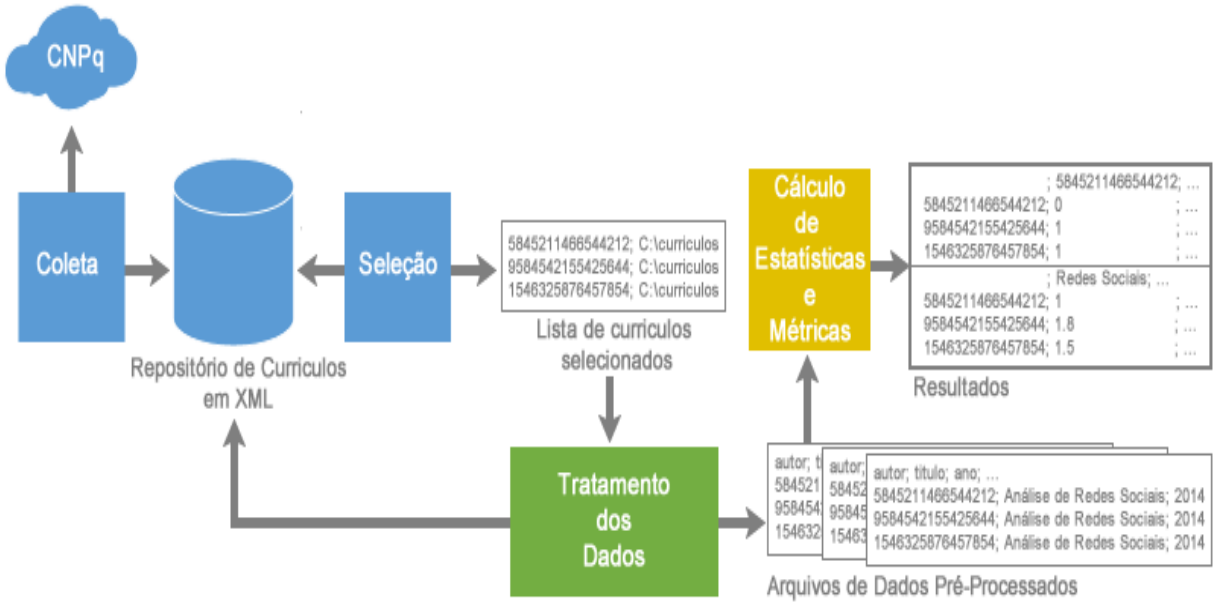

Figura 1

Visão geral do LattesDataXplorer (DIAS, 2016).

O processo de coleta e extração dos dados da Plataforma Lattes envolve três etapas que são realizadas por meio de três componentes específicos que, para minimizar o custo

computacional envolvido, executam respectivamente as seguintes funções: 1) extração de URLs, que visa obter os códigos de identificação de todos os currículos cadastrados na plataforma, possibilitando assim acessar individualmente cada um deles; 2) extração de Ids e Datas de Atualização, que visa extrair de cada currículo o seu identificador individual e a data de sua última atualização; e 3) coleta dos currículos, que visa coletar e armazenar em um repositório local os currículos cuja data de atualização na Plataforma Lattes seja divergente da data de atualização armazenada localmente ou que ainda não tenham sido coletados.

Resultados

Uma grande parte dos editais de financiamento a projetos abertos nos país por agências de amparo à pesquisa, por instituições de ensino e pelo próprio CNPq consideram a análise dos currículos Lattes dos proponentes como um dos principais itens do processo de avaliação das propostas apresentadas. Isto tem sido um grande incentivo para que os pesquisadores mantenham seus currículos atualizadas, tornando a Plataforma Lattes uma fonte de dados extremamente rica para análise da produção científica brasileira. Além disso, estudantes que pretendam dar continuidade aos seus estudos em programas de pós-graduação, no país ou no exterior, financiados por agências governamentais brasileiras, também devem cadastrar seus currículos na Plataforma Lattes, contribuindo de forma significativa para o aumento da quantidade de currículos atualmente disponíveis nessa plataforma.

Conforme ressaltado por Dias (2016), apesar de o conjunto de indivíduos com doutorado concluído representar apenas 5,38% de todos os currículos cadastrados na Plataforma Lattes, esses indivíduos são detentores de 64,67% dos artigos em anais de congressos e 74,51% dos artigos em periódicos registrados em todo o conjunto de currículos contidos na Plataforma Lattes, corroborando, assim, a importância da coleção apresentada neste trabalho.

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.

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**

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

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.

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.

Referências Bibliográficas

BOAVENTURA, M. [et al.] (2014) - Caracterização Temporal das Redes de Colaboração Científica nas Universidades Brasileiras: Anos 2000-2013. In: *BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING*, 3, 2014, Anais ... Brasilia, 2014.

DIAS, T. M. R. (2016) - *Um Estudo da Produção Científica Brasileira a partir de Dados da Plataforma Lattes*. Tese de Doutorado, Programa Pós-Graduação em Modelagem Matemática e Computacional, CEFET-MG.

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

FERRAZ, R. R. N., QUONIAM, L., MACCARI, E. A. (2014) - The Use of Scriptlattes tool for extraction and online availability of academic production from a departament of stricto sensu in management. In *Proceedings of the International Conference on Information Systems and Technology Management*, São Paulo, Brasil.

HOFFNAGEL, J. C. (2009) - A prática de citação em trabalhos acadêmicos. *Cadernos de Linguagem e Sociedade*, Vol. 10, n.º 1, p. 71.

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

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

LIMA, H. [et al.] (2015) - Assessing the profile of top Brazilian computer science researchers. *Scientometrics*, Vol. 103, n.º 3, p. 879-896.

MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A.; OLIVEIRA, L. B. (2012) - 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*, Porto Alegre, Vol. 18.

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

REVOREDO, K. et al. (2012) - Mining scientific literature for analysis of collaboration in research communities. In*: Proceedings of the Brazilian Workshop on Social Network Analysis and Mining,* Curitiba, Brasil.

STRÖELE, V., ZIMBRÃO, G., SOUZA, J. M. (2012) - Análise de redes sociais científicas: modelagem multi-relacional. In: *Proceedings of the Brazilian Workshop on Social Network Analysis and Mining*, Curitiba, Brasil.