DISTRIBUIÇÃO E ANÁLISE DO CONJUNTO DE DOUTORES BRASILEIROS BASEADO EM GÊNERO

Email:
moniqueosantiago@gmail.com
thiagomagela@gmail.com

Monique de Oliveira Santiago, Thiago Magela Rodrigues Dias[1]

RESUMO

Gênero é um termo abrangente que tem gerado muitos estudos e esforços de diversas áreas, a fim de ampliar as abordagens e perspectivas sobre o tema. O estudo aqui apresentado serve a diversos propósitos, como direcionar o fomento à pesquisa e compreender a importância que o domínio gênero demanda sobre o cenário da ciência. Logo, o objetivo deste trabalho é caracterizar e analisar com ênfase em gênero o conjunto de doutores com currículos cadastrados na Plataforma Lattes. Para isso, após a aquisição dos dados curriculares do conjunto de doutores, aplica-se técnica computacional de mineração de texto para identificar o gênero da palavra através do primeiro nome do pesquisador. Esta técnica utiliza-se de uma tabela contendo 608 sufixos distribuídos entre cada letra do alfabeto, que é consultada através da última letra do nome do indivíduo a ser identificado, verificando se existe sufixo correspondente ao inverso do nome sem a última letra. A identificação de gênero do conjunto de doutores possibilitou caracterizar de modo geral, constatando, conforme esperado, um cenário com maioria masculina em 53,07% e feminina com 46,93%. Os resultados apresentados mostram uma caracterização geral do conjunto e podem auxiliar no entendimento sobre o domínio gênero no desenvolvimento da ciência brasileira.

Palavras-Chave: Identificação de gênero, Plataforma Lattes, Doutores brasileiros.

ABSTRACT

Gender is a comprehensive term that has generated many studies and efforts from various areas in order to broaden approaches and perspectives on the subject. The study presented here serves several purposes, such as directing the promotion of research and understanding the importance that gender domain demands on the science scene. Therefore, the objective of this work is to characterize and analyze with emphasis on gender the set of doctors with curricula registered in the Lattes Platform. For this, after the acquisition of the curricular data of the set of doctors, it is applied the computational technique of text mining to identify the gender of the word through the first name of the researcher. This technique uses a table containing 608 suffixes distributed between each letter of the alphabet, which is queried through the last letter of the name of the individual to be identified, checking whether there is suffix corresponding to the inverse of the name without the last letter. The gender identification of the group of doctors made it possible to characterize in a general way, noting, as expected, a scenario with the male majority in 53.07% and female with 46.93%. The results presented show a general characterization of the set and can help in understanding the gender domain in the development of Brazilian science.

Keywords: Identification of gender, Lattes Platform, Brazilian Doctors.

[1] Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Pesquisa apresentada no XIX ENANCIB – GT-7 – Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação
Acesso ao artigo: http://enancib.marilia.unesp.br/index.php/XIXENANCIB/xixenancib/paper/view/1247

1. INTRODUÇÃO

Gênero é um termo amplo e muito explorado na literatura, abrangendo diversas áreas e abordagens. Logo, para este estudo, seguiremos a premissa que gênero é "[...] um elemento constitutivo de relações sociais baseadas nas diferenças percebidas entre os sexos e o gênero é uma forma primária de dar significado às relações de poder." (SCOTT, 1995, p86).

Estudos sobre gênero tem inspirado inúmeras pesquisas e podem servir a diversos propósitos como compreender a produção científica, ampliar as perspectivas sobre o tema, problematizar e discutir as relações assimétricas de poder entre homens e mulheres. Van Arensbergen *et al.* (2012) sugere que as diferenças de desempenho por gênero estão desaparecendo após realizar um estudo sobre as diferenças de gênero na produção científica na área de ciências sociais, utilizando dados dos pedidos de bolsas de pesquisa na Holanda. O estudo de Sousa e Perucchi (2013) identifica a frequência da produção científica de mulheres e homens, e verifica se há hierarquia de gênero com base nas temáticas escolhidas pelos pesquisadores, utilizando a base de apresentações orais do XIII Enancib. Já no estudo de Dias e De Lima (2015) é realizado um levantamento das pesquisas sobre mulheres e relações de gênero publicadas em periódicos da Ciência da Informação, utilizando a base de dados BRAPCI.

No contexto atual são inúmeras as bases de dados disponíveis para estudos, porém para realizar amplas análises e fazer um levantamento da formação acadêmica, o repositório de dados curriculares da Plataforma Lattes[2] é um diferencial (LANE, 2010). A Plataforma Lattes coleta uma grande quantidade de informações, porém não disponibiliza algumas destas para a consulta pública, como exemplo, o campo sexo. Como o nome está diretamente relacionado ao gênero do indivíduo, utilizando de técnica computacional como a mineração de texto é possível identificar se o primeiro nome é referente ao gênero feminino ou gênero masculino.

Nesse sentido, este trabalho visa caracterizar com ênfase em gênero através de técnica de mineração de texto e analisar o conjunto de pesquisadores doutores brasileiros, tendo como fonte de dados os seus currículos cadastrados na Plataforma Lattes. Segundo Naldi *et al.* (2004), é viável utilizar o primeiro nome para produzir indicadores de gênero robustos que podem ser aplicados a qualquer conjunto de dados contendo nomes próprios. Para tanto, inicialmente será identificado o gênero da palavra através do primeiro nome do pesquisador cadastrado na Plataforma Lattes e posteriormente será apresentado a caracterização geral do conjunto.

2. DESENVOLVIMENTO

Trata-se de um estudo quantitativo que utiliza coleta, análise e recursos estatísticos de dados dos doutores cadastrados na Plataforma Lattes. É caracterizado como descritivo, pois visa caracterizar o conjunto e analisar o cenário, gerando resultados e viabilizando novas pesquisas.

2.1 Coleta de dados

O conjunto de dados para análise são provenientes das informações curriculares cadastradas na Plataforma Lattes e foi escolhida para a extração de dados por possuir uma ampla

[2] Plataforma Lattes: http://lattes.cnpq.br/

quantidade de informações, como formação acadêmica, atuação profissional, orientações acadêmicas, produções técnicas e científicas. Os registros coletados em abril de 2018 totalizaram mais de 5.743.000, utilizando o extrator LattesDataXplorer, um arcabouço desenvolvido por Dias (2016) responsável por coletar e tratar os dados científicos. Para este estudo será considerado o conjunto de indivíduos que possui como titulação máxima concluída doutorado, totalizando 290.131 currículos. Essa escolha pode ser justificada pelo fato de que os mesmos são responsáveis por grande parte de publicações em periódicos e anais de congresso, possuir em geral data de atualização de seus currículos recente e também por ser responsável pela formação dos alunos nos programas de pós-graduação *stricto sensu* no Brasil (DIAS, 2016).

Após o tratamento dos currículos pelo LattesDataXplorer é gerado como um dos extratos de dados, um arquivo tabulado separando os valores por um delimitador do tipo vírgula. No Quadro 1 são apresentados os campos do arquivo e suas respectivas descrições.

Campos	**Descrições**
id	Identificador do currículo
label	Nome completo
titulacao	Titulação máxima
grande_area	Grande área de atuação
area	Área de atuação
instituicao	Instituição, endereço profissional
cidade	Cidade de atuação profissional
pais	País de atuação profissional
uf	Estado de atuação profissional

Quadro 1: Estrutura do arquivo

Os registros possuem informações que possibilitam caracterizar o conjunto em análise sob diversas perspectivas, exceto o campo gênero, objeto de estudo deste trabalho.

2.2 Identificação de gênero

Ao inserir um currículo ou atualizar o cadastro, na opção "Dados gerais / Identificação / Dados pessoais", a Plataforma Lattes possui o campo sexo que permite escolher entre masculino ou feminino, porém esta informação não é exibida na consulta pública, e por este motivo não é possível recuperá-la. Nesse sentido, executou-se de técnica de mineração de texto para verificar o gênero de cada pesquisador.

A técnica para identificar o gênero utilizada foi desenvolvida por Carnut[3] no formato de um arquivo de identificação que contém uma tabela com 608 sufixos distribuídos entre cada letra do alfabeto. Para identificar o gênero da palavra, consulta-se a tabela através da última letra do nome e verifica se o sufixo corresponde ao inverso do nome sem a última letra. A tabela de sufixos compreende as vogais "a", "e", "i", "o" e "u" com respectivamente 72, 175, 96, 30 e 4 sufixos cada. Neste grupo, os nomes terminados em "e" normalmente são femininos, no entanto há 175

[3] Arquivo de identificação de Carnut: http://www.postcogito.org/Kiko/PlanilhaMascFemPtBr.html

exceções que os tornam masculinos, sendo este o caso mais difícil de resolver. As letras “g”, “h”, “l”, “m”, “n”, “r”, “s”, “t”, “y” e “z”, contém respectivamente 5, 12, 13, 8, 33, 33, 51, 7, 55 e 6 sufixos cada. As consoantes “b”, “c”, “d”, “k” e “p” contém entre um a dois sufixos cada. As restantes “f”, “j”, “q”, “v”, “w”, “x” não contém sufixos, indicando que, caso o nome terminar com uma destas letras será do gênero masculino.

Como o arquivo de dados extraído da Plataforma Lattes possui 290.131 linhas que representa cada indivíduo e o arquivo de identificação disponibilizado por Carnut é um processo manual de inserir apenas um registro por vez, implementou-se o algoritmo em uma linguagem de programação para ler todas as linhas do arquivo automaticamente. Python[4] foi a linguagem escolhida por ser muito utilizada para análise e modelagem de dados, gratuita, fácil de aprender, multiplataforma, entre outros. Assim, a implementação do algoritmo realiza a leitura de todo o arquivo de dados e para cada nome do indivíduo a ser identificado realiza-se a limpeza do primeiro nome removendo acentos e convertendo em minúsculo, retira-se a última letra e inverte-se o primeiro nome, consulta-se a tabela de sufixos através da última letra do nome e verifica se contém sufixo correspondente ao inverso do primeiro nome, caso contenha retornará “F” para gênero feminino, caso não contenha retornará “M” para gênero masculino.

2.3 Prova de conceito

Para testar a acurácia do algoritmo implementado, o mesmo foi aplicado na base de dados disponibilizada na plataforma Kaggle[5], denominada candidatos_deputados_2014_final, contendo nome completo e gênero de candidatos ao cargo de deputado no ano de 2014, além de outras informações gerais disponibilizadas no arquivo. A base contém 21.124 registros com 56 campos de dados. Os campos que interessam ao estudo são “nome_cadidato” e “descricao_sexo”. Utilizando um filtro por gênero masculino e feminino no campo “descricao_sexo” do arquivo, obteve-se 14.962 masculino e 6.162 feminino. Ao executar o algoritmo implementado neste estudo, obteve-se o resultado de 15.143 pessoas do gênero masculino e 5.981 do gênero feminino que representa um percentual de aproximadamente 97% de precisão e comprova a confiabilidade do algoritmo.

Após provar a acurácia do algoritmo e coletar os dados da Plataforma Lattes, foi realizada uma caracterização com o conjunto de doutores quanto ao gênero masculino e feminino, executando o algoritmo implementado na base de dados curriculares da Plataforma Lattes. Para o total de 290.131 indivíduos, obteve-se 154.001 (53,07%) do gênero masculino e 136.130 (46,93%) do gênero feminino.

3. RESULTADOS

Analisando o gênero e a titulação, obtém-se a caracterização geral apresentada na Figura 1 a porcentagem do gênero masculino que corresponde a 53,07% do total de indivíduos, 111.186 possuem doutorado e 42.815 pós-doutorado, e para os indivíduos do gênero feminino com 46,93% de representatividade, 96.948 possuem doutorado e 39.182 possuem pós-doutorado.

[4] Linguagem de programação utilizada Python: https://www.python.org/
[5] Base de dados Kaggle: https://www.kaggle.com/eliezerfb/candidatos-deputado-federal-e-estadual-2014

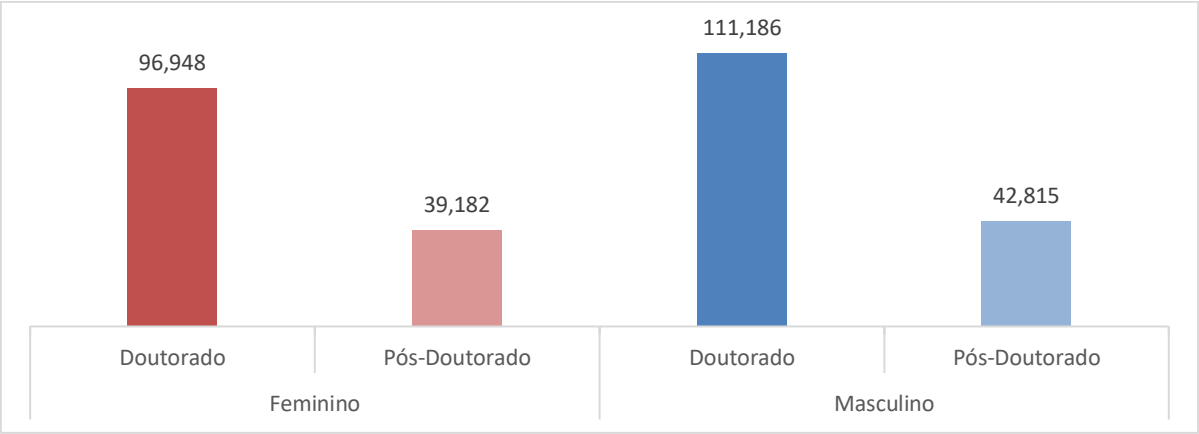

Figura 1: Caracterização Geral do Conjunto Analisado

Também foram analisados o gênero e estado de atuação profissional dos indivíduos. Obteve-se, em todos os estados a quantidade de doutores como maioria masculina, conforme apresentado na Tabela 1. Além disso também foram analisados o gênero e país de atuação profissional dos indivíduos. Além do Brasil, os países mais representativos, somando a quantidade de indivíduos femininos e masculinos superior a 100, são Chile, Colômbia, Estados Unidos e Portugal. Dentre estes, todos apresentaram a quantidade de indivíduos femininos menores que masculinos.

Tabela 1: Distribuição por estado

Estado	Feminino	Masculino	Estado	Feminino	Masculino
AC	196	320	PB	2.197	2.677
AL	756	941	PE	3.404	3.622
AM	1.018	1.321	PI	845	968
AP	161	223	PR	6.552	7.957
BA	3.885	4.116	RJ	12.878	14.316
CE	2.550	3.061	RN	1.736	2.250
DF	2.928	3.923	RO	275	331
ES	1.304	1.776	RR	207	267
GO	2.064	2.499	RS	8.777	9.076
MA	947	1.114	SC	3.687	4.732
MG	9.561	11.839	SE	832	979
MS	1.296	1.436	SP	26.594	31.513
MT	1.235	1.436	TO	429	564
PA	1.841	2.338			

A análise entre gênero e grande área do conhecimento corrobora com o estudo de Olinto (2011), comprovando que os homens predominam nas carreiras exatas e as mulheres tem predominância nas áreas biológicas e saúde. Na Figura 2 verifica-se que as grandes áreas do conhecimento consideradas predominantemente femininas, sendo elas Ciências Biológicas,

Ciências da Saúde, Ciências Humanas e Linguísticas, Letras e Artes possuem maior quantidade de doutores do gênero feminino. As áreas de atuação profissional que apresentaram maior representatividade feminina, totalizando indivíduos femininos e masculinos superior a 3.000 na análise entre gênero e áreas de atuação, foram Artes, Bioquímica, Ciência e tecnologia de alimentos, Comunicação, Educação, Enfermagem, Farmácia, Genética, Letras, Linguística, Medicina veterinária, Microbiologia, Odontologia, Psicologia e Saúde coletiva.

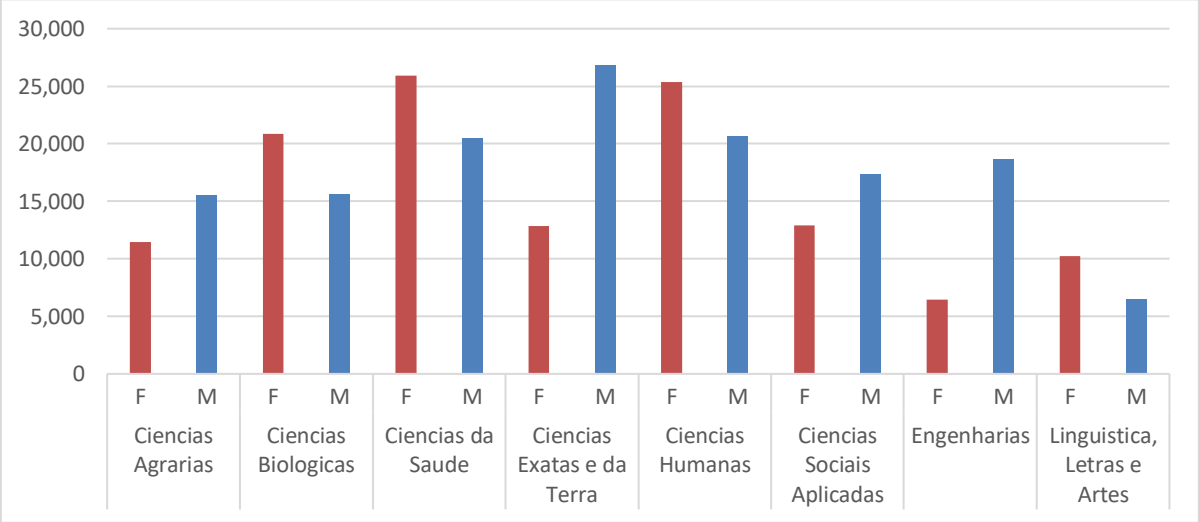

Figura 2: Distribuição por grandes áreas do conhecimento

A próxima análise foi referente ao gênero e instituição atual de atuação profissional do pesquisador. As trinta instituições mais representativas, contendo o total de indivíduos femininos e masculinos superior a 1.500, são apresentadas na Tabela 2. Dentre estas, apenas 20% possuem maior quantidade de indivíduos do gênero feminino.

Tabela 2: Distribuição por instituições de atuação profissional do pesquisador

Instituições	Feminino	Masculino	Total
Universidade de São Paulo	5.656	7.171	12.827
Universidade Estadual Paulista Júlio de Mesquita Filho	2.680	3.328	6.008
Universidade Federal do Rio de Janeiro	2.962	3.027	5.989
Universidade Estadual de Campinas	2.185	2.649	4.834
Universidade Federal de Minas Gerais	2.101	2.277	4.378
Universidade Federal do Rio Grande do Sul	1.963	2.066	4.029
Instituto Federal de Educação	1.549	2.335	3.884
Universidade Federal de Santa Catarina	1.522	1.965	3.487
Universidade de Brasília	1.447	1.768	3.215
Universidade Federal Fluminense	1.452	1.670	3.122
Universidade Federal de Pernambuco	1.452	1.574	3.026
Universidade Federal de São Paulo	**1.493**	1.376	2.869
Universidade Federal do Paraná	1.281	1.574	2.855

Universidade do Estado do Rio de Janeiro	**1.288**	1.225	2.513
Universidade Federal da Bahia	**1.265**	1.184	2.449
Universidade Federal da Paraíba	1.116	1.283	2.399
Universidade Federal do Rio Grande do Norte	1.060	1.315	2.375
Empresa Brasileira de Pesquisa Agropecuária	799	1.557	2.356
Universidade Federal do Ceara	1.000	1.349	2.349
Universidade Federal de Goiás	1.106	1.214	2.320
Fundação Oswaldo Cruz	**1.429**	880	2.309
Universidade Federal de Santa Maria	965	1.149	2.114
Universidade Federal de São Carlos	940	1.086	2.026
Universidade Federal do Para	873	1.142	2.015
Universidade Federal de Uberlândia	851	1.101	1.952
Universidade Tecnológica Federal do Paraná	696	1.149	1.845
Universidade Federal do Espirito Santo	794	972	1.766
Universidade Estadual de Maringá	**835**	807	1.642
Universidade Federal de Viçosa	620	1.020	1.640
Universidade Estadual de Londrina	**807**	738	1.545

Os resultados apresentados mostram uma caracterização geral do conjunto no Brasil e a contribuição deste trabalho pode ser definida por "identificação, registro, categorização que levem à reflexão e síntese sobre a produção científica de uma determinada área" (MOROSINI, FERNANDES, 2014), destacando o novo na produção científica. É importante ressaltar que como alguns registros apresentaram inconsistência em suas informações, como campos vazios ou palavras grafadas incorretamente, estes foram desconsiderados nas análises.

4. CONSIDERAÇÕES FINAIS

Este estudo apresentou a caracterização com ênfase em gênero através de técnicas de mineração de texto e análises sobre o conjunto de pesquisadores doutores brasileiros, tendo como fonte de dados os seus currículos cadastrados na Plataforma Lattes. A partir dos resultados encontrados, foi possível apresentar a caracterização geral do conjunto de doutores por gênero visando compreender a distribuição do cenário feminino e masculino.

Como resultados obteve-se a caracterização geral, correspondendo a maioria de doutores masculinos, conforme era esperado. No âmbito das grandes áreas do conhecimento, 50% correspondeu a maioria feminina. Já as análises por estado obteve-se em todos os estados a maioria masculina. Por fim, o aspecto geral de instituições a que os doutores estão vinculados, apenas 20% tem participação predominantemente feminina. Para trabalhos futuros, espera-se incorporar a análise de produções científicas com ênfase em gênero, objetivando aprofundar, compreender e analisar a produção científica nacional por gênero.

REFERÊNCIAS

DIAS, Karla Cristina Oliveira; DE LIMA, Francisca Rosimere Alves. **Levantamento das produções sobre mulheres e relações de gênero nos artigos de periódicos em Ciência da Informação.** Pesquisa Brasileira em Ciência da Informação e Biblioteconomia, v. 10, n. 2, 2015.

DIAS, Thiago Magela Rodrigues. **Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes**. 2016. 181f. Tese (Doutorado em Modelagem Matemática e Computacional) - Programa em Pós-Graduação em Modelagem Matemática e Computacional, CEFETMG, Belo Horizonte, 2016.

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

MOROSINI, Marília Costa; FERNANDES, Cleoni Maria Barboza. Estado do Conhecimento: conceitos, finalidades e interlocuções. **Educação Por Escrito**, Porto Alegre, v. 5, n. 2, p.154-164, jul./dez. 2014. Disponível em: < http://revistaseletronicas.pucrs.br/ojs/index.php/porescrito/article/view/18875 >. Acesso em: 24 jun. 2018.

NALDI, Fulvio; LUZI, Daniela; VALENTE, Adriana; PARENTI, Ilaria Vannini. Scientific and technological performance by gender. In: **Handbook of quantitative science and technology research**. Springer, Dordrecht, 2004. p. 299-314.

OLINTO, G. **A inclusão das mulheres nas carreiras de ciência e tecnologia no Brasil.** Inclusão Social, Brasília, DF, v. 5 n. 1, p. 68-77, jul./dez. 2011.

SCOTT, Joan Wallach. **Gênero: uma categoria útil de análise histórica.** Educação e Realidade. Porto Alegre, v. 20, n. 2, jul./dez. 1995, p.71-99. Disponível em < https://repositorio.ufsc.br/bitstream/handle/123456789/1210/scott_gender2.pdf >. Acesso em: 10 jun. 2018.

SOUSA, Beatriz Alves; PERUCCHI, Valmira. Gênero na produção científica dos grupos de trabalho do ENANCIB: análise nos anais do XIII ENANCIB. In: ENANCIB 2013. **Anais eletrônico...** Santa Catarina: UFSC, 2013. Disponível em: <http://enancib.ibict.br/index.php/enancib/ xivenancib/paper/viewFile/4335/3458>. Acesso em: 25 abr. 2018.

VAN ARENSBERGEN, Pleun; VAN DER WEIJDEN, Inge; VAN DEN BESSELAAR, Peter. Gender differences in scientific productivity: a persisting phenomenon?. **Scientometrics**, v. 93, n. 3, p. 857-868, 2012.