Seleção de atributos para modelos de inferência sobre o desempenho científico de pesquisadores da área de conhecimento Odontologia

Renê Rodrigues Veloso
Doutor em Ciência da Computação pela Universidade Federal de Minas Gerais (UFMG) - Belo Horizonte, MG - Brasil. Professor da Universidade Estadual de Montes Claros (Unimontes) - MG - Brasil. Professor da Fundação Educacional Montes Claros (FEMC) - Montes Claros, MG - Brasil.
http://lattes.cnpq.br/5262545895128956
E-mail: rene.veloso@unimontes.br

Luís Antônio Guisso Lopes
Mestrado profissional em Modelagem Computacional e Sistemas pela Universidade Estadual de Montes Claros (Unimontes) - Montes Claros, MG - Brasil. Professor do Instituto Federal de Educação Ciência e Tecnologia do Norte de Minas Gerais (IFNMG) - Montes Claros, MG - Brasil.
http://lattes.cnpq.br/6181855641333903
E-mail: luis.guisso@ifnmg.edu.br

Hugo Andrei Mendes da Silva
Mestrado profissional em Modelagem Computacional e Sistemas pela Universidade Estadual de Montes Claros (Unimontes) – Montes Claros, MG - Brasil. Professor da Fundação Educacional Montes Claros (FEMC) - Montes Claros, MG - Brasil.
http://lattes.cnpq.br/2759869407852345
E-mail: hugoandrei@femc.edu.br

Romulo Barbosa Veloso
Doutor em Engenharia Florestal pela Universidade Federal de Lavras (UFLA) - Lavras, MG - Brasil. Professor da Universidade Estadual de Montes Claros (Unimontes) - Montes Claros, MG - Brasil.
http://lattes.cnpq.br/3087255051684123
E-mail: romulo.veloso@unimontes.br

Nilton Alves Maia
Doutor em Engenharia Elétrica pela Universidade Federal de Minas Gerais (UFMG) - Belo Horizonte, MG - Brasil. Professor da Universidade Estadual de Montes Claros (Unimontes) - Montes Claros, MG - Brasil.
http://lattes.cnpq.br/3101079034762740
E-mail: nilton.maia@unimontes.br

Data de submissão: 18/06/2018. Data de aprovação: 16/07/2018. Data de publicação: 21/12/2018.

RESUMO

Esforços diversos foram empreendidos para a elevação da produção científica do Brasil. No entanto, ainda há ações a serem tomadas para conduzir os níveis produtivos atuais a patamares mais elevados. Neste sentido, acredita-se que a avaliação dos pesquisadores, a partir de seus históricos profissionais, é uma etapa importante para a tomada de decisão futura e, consequentemente, permite que essas diversas ações sejam mais bem direcionadas, trazendo benefícios como: formação de equipes otimizadas para execução de projetos, aplicação adequada de recursos para financiamentos, aumento do prestígio das instituições a partir do incremento de produtividade de seus pesquisadores, dentre outros. Por tratar-se de tema de estudo não esgotado e com ganhos relevantes para toda a sociedade, neste trabalho investiga-se a extração dos principais atributos dos pesquisadores que indicam o potencial produtivo futuro a partir de dados de seus currículos cadastrados na Plataforma Lattes. O foco do estudo foi a área de conhecimento Odontologia, a partir da qual foram empregados métodos de descoberta de conhecimento, tendo como referência o respectivo documento de área. Como resultado, os principais atributos dos pesquisadores são apresentados de acordo com a sua relevância na determinação de produtividade futura dos pesquisadores.

Palavras-chave: Extração de atributos. Currículo Lattes. Potencial de pesquisa. Odontologia.

Attribute selection for inference models on the scientific performance of researchers in the Odontology field of knowledge

ABSTRACT

Many efforts have been made to increase the scientific production in Brazil. However, there are still actions to be taken to conduct the current production levels to the higher ones. It is believed that the evaluation of researchers and of their professional curriculum is na important step for future decision making and hence it enables these several actions are better applied with benefits as: organization of optimized teams for project execution, proper application of funds, elevation the prestige of the institutions with increasing productivity of its researchers, among others. To the better of our knowledge, this is a not fully studied theme but with relevant gains for the whole society, so, in this work, we investigate the extraction of the mainly attributes for predicting the potential of researchers from data registered in the Lattes Platform. This study focused the researchers of Odontology, from which we applied KDD methods based on the respective document of area. We present the mainly attributes according to their relevance in predicting the productivity of the researchers.

Keywords: Attributes extraction. Lattes curriculum. Potencial of research. Odontology.

Selección de atributos para modelos de inferencia sobre el desempeño científico de investigadores del área de conocimiento Odontología

RESUMEN

Esfuerzos diversos fueron emprendidos para la elevación de la producción científica de Brasil. Sin embargo, todavía hay acciones a tomar para conducir los niveles productivos actuales a niveles más elevados. En este sentido, se cree que la evaluación de los investigadores, a partir de sus históricos profesionales, es una etapa importante para la toma de decisión futura y, consecuentemente, permite que esas diversas acciones sean mejores dirigidas, trayendo beneficios tales como: formación de equipos optimizadas para ejecución de proyectos, aplicación adecuada de recursos para financiamientos, aumento del prestigio de las instituciones a partir del incremento de productividad de sus investigadores, entre otros. Por tratarse de un tema de estudio no agotado y con ganancias relevantes para toda la sociedad, en este trabajo, se investiga la extracción de los principales atributos de los investigadores que indican el potencial productivo futuro a partir de datos de sus currículos registrados en la Plataforma Lattes. El foco del estudio fue el área de conocimiento Odontología, a partir de la cual se emplearon métodos de descubrimiento de conocimiento, teniendo como referencia el respectivo documento de área. Los principales atributos de los investigadores se presentan de acuerdo con su relevancia en la determinación de la productividad futura de los investigadores.

***Palabras clave**: Extracción de atributos. Curriculum Lattes. Potencial de investigación. Odontología.*

INTRODUÇÃO

Nos últimos anos, variados esforços foram empreendidos para a elevação dos níveis de produção científica no país. No período de 2007 a 2010, por exemplo, ocorreram avanços bastante satisfatórios no fomento à pesquisa científica e à inovação tecnológica no Brasil com a implantação do Plano de Ação em Ciência, Tecnologia e Inovação para o Desenvolvimento Nacional (PACTI). Tal fato é devido à metodologia empregada, que contou com um plano concreto de ações, prioridades, institucionalidade, metas e orçamento, envolvendo a Federação e os estados, que resultou no incremento de publicações indexadas junto a bases internacionais (REZENDE, 2011; LETA, 2012).

Apesar dos esforços, no entanto, o Brasil apresenta índices produtivos ainda aquém do esperado. Como exemplo, exibe uma deficiência no quantitativo de pesquisadores em relação às existentes em grande parte dos países desenvolvidos. Adicionalmente, possui colaboração com pesquisadores internacionais abaixo da média do continente, classificando-se apenas acima da Venezuela, mas atrás do Chile, Argentina e Uruguai em números de registros de patentes (REZENDE, 2011; NOORDEN, 2014).

Abbasi, Altmann e Hossain (2011) afirmam que, para impulsionar o desenvolvimento científico no país, faz-se necessária a aplicação adequada de recursos para financiamento de pesquisas, a composição de equipes de trabalho competentes para a execução de projetos e a avaliação individual do perfil do pesquisador como suporte ao próprio crescimento profissional. Essa perspectiva conduz à percepção de que há um problema de cunho relevante a ser investigado, suscitando a busca de meios promotores do avanço da pesquisa cujo alvo é o pesquisador.

A necessidade por avanços na pesquisa científica torna a busca por diretrizes que guiem as tarefas para sua promoção em um campo aberto à exploração contemplado com estudos das mais variadas abordagens, mas que ainda não esgotaram o tema (MUGNAINI; JANUZZI; QUONIAM, 2004). Nesse contexto, há pesquisas que buscam prever o sucesso em publicações por meio da análise de citações (BRYNKO, 2010); investigam a predição de índices bibliométricos[1] futuros empregando dados dos currículos dos pesquisadores associados ao índice sob investigação (ACUNA; ALLESINA; KORDING, 2012); estudam as influências de gênero, língua, prestígio da instituição de formação e publicações na prolificidade científica de recém-doutores (LAURANCE et al., 2013); avaliam redes sociais a fim de se apontar seu efeito sobre o desempenho de pesquisadores com base nas citações recebidas (ABBASI; ALTMANN; HOSSAIN, 2011); examinam múltiplas redes sociais para prever seu impacto nas citações (CIMENLER; REEVES; SKVORETZ, 2014); analisam redes de coautorias em busca de indicadores de sucesso na publicação de artigos (SARIGÖL et al., 2014); e buscam projetar a perspectiva de liderança do pesquisador (DIJK; MANOR; CAREY, 2014).

Apesar das pesquisas citadas apresentarem variadas propostas de predição do potencial futuro de um cientista, até onde foi possível verificar, não foram encontrados estudos que tratem do problema em âmbito nacional e, particularmente, fazendo uso da principal fonte de dados sobre os pesquisadores brasileiros, o Currículo Lattes (CL).

Diante do exposto, este trabalho aborda o problema de encontrar os atributos mais relevantes dos pesquisadores, com base em seus currículos na Plataforma Lattes (PL), que possam subsidiar a construção de modelos de inferência de potencial produtivo futuro de pesquisadores. Este trabalho contextualiza-se na área de conhecimento Odontologia, tomando como base as regras estipuladas pelo CA-OD (Comitê de Assessoramento da área de conhecimento Odontologia). Contudo, os resultados aqui apresentados aplicam-se a outras áreas, uma vez que os principais atributos que caracterizam o potencial de um pesquisador de alto rendimento independem da sua área de atuação.

[1]Índices bibliométricos são resultados de técnicas quantitativas e estatísticas que permitem a avaliação da produção científica, importantes para o reconhecimento dos investigadores junto à comunidade científica.

As contribuições deste trabalho objetivam beneficiar: I) programas de pós-graduação no sentido de identificar o potencial de pesquisadores atuantes nos respectivos quadros; II) a formação profissional do pesquisador a fim de possibilitar uma projeção de seu potencial e conduzir a uma reflexão de pontos de melhoria em seu perfil; e III) agências de fomento de pesquisa na identificação de equipes com maior potencial de avanço científico para aplicação de recursos. Além disto, a pesquisa pretende contribuir com metodologias para análise da base de currículos Lattes, que é importante para a área de ciência da informação.

MATERIAIS E MÉTODOS

O ponto inicial deste trabalho foi a obtenção da base de dados de pesquisadores brasileiros, coletados de modo automático por meio de scripts em formato XML a partir da Plataforma Lattes (PL). De posse dos currículos de pesquisadores contemplados com bolsas de produtividades em pesquisa (PQ), partiu-se para as etapas do processo de descoberta de conhecimento (KDD - Knowledge Discovery in Databases).

Tais fasess são adaptadas e distribuídas para o fluxo do processo aplicado por esta pesquisa e são apresentados na figura 1. Nessa figura, destacam-se:

a) as raias A compreendem as tarefas de aquisição e de seleção dos dados, nas quais: currículos Lattes contêm o histórico de atuação profissional dos pesquisadores, sendo a fonte de dados primordial para alimentação do processo; o Document Type Definition[2] (DTD) é obtido com fins de fornecimento da estrutura dos currículos sob análise para posterior pré-seleção dos atributos de interesse; fatores de impacto dos artigos publicados e regras CA-OD são obtidos para arotulação dos dados; gênero dos pesquisadores são computados e conceitos das instituições de ensino superior junto à Capes e atuações em equipes de pesquisa como líderes são colhidos para aplicação na modelagem;

b) a raia B compreende o pré-processamento dos dados, na qual: atributos de interesse são pré-selecionados para contabilização dos dados históricos dos pesquisadores; anos de referências e projeções são definidos para alimentar as contagens de produções científicas e as rotulações dos perfis dos pesquisadores; a rotulação identifica se um pesquisador está apto a receber a bolsa PQ dado um ano de referência e uma projeção (entre 1 e 5 anos); contagens e rótulos são agregados e armazenados para aplicação na escolha de atributos relevantes e na mineração de dados;

c) a raia C compreende a seleção e a transformação dos atributos relevantes selecionados em quantidades variadas por meio de escores e ganho de informação, sendo validados pelos classificadores indicados.

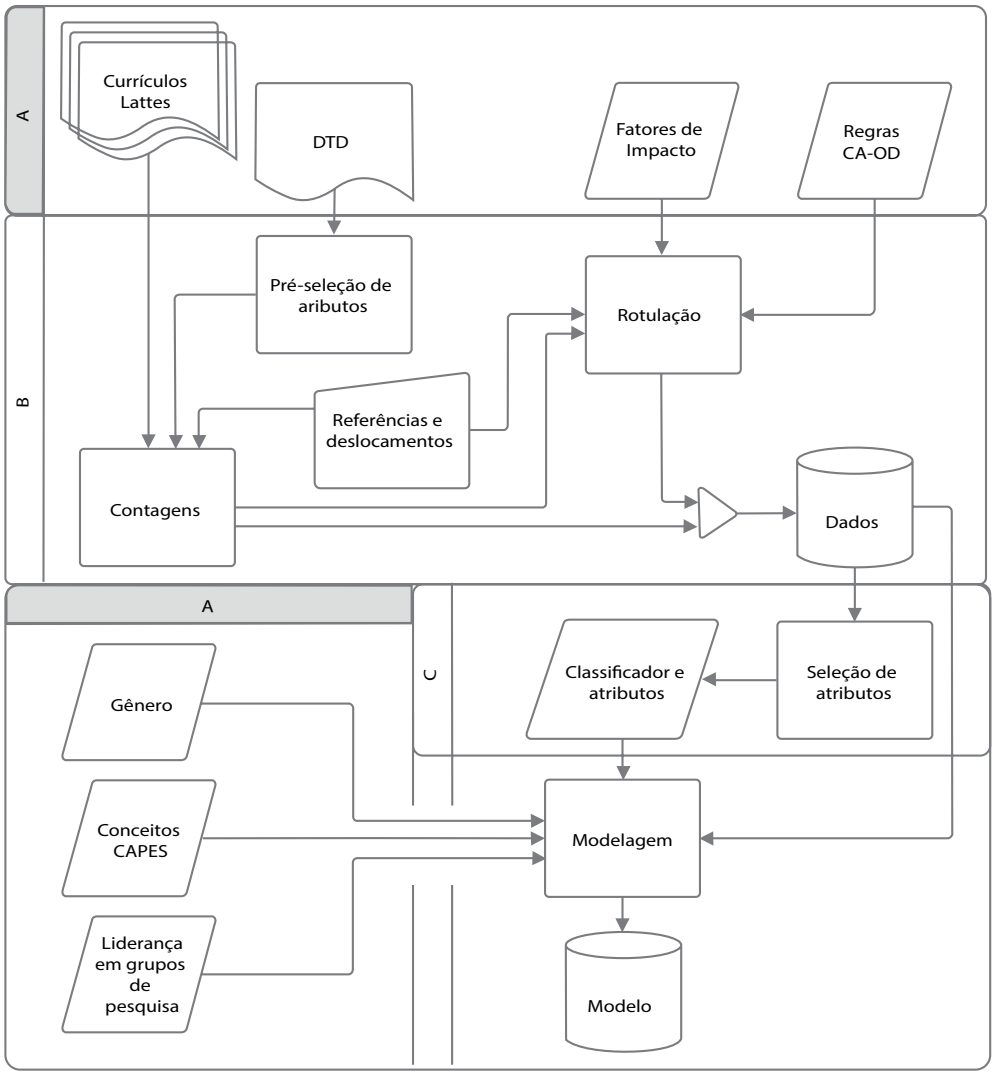
As seções seguintes detalham os pontos relevantes existentes em cada tarefa citada anteriormente.

AQUISIÇÃO E SELEÇÃO DE DADOS

A principal fonte dos dados a serem minerados nesta pesquisa é o CL. Os dados estão expostos por meio de um website da PL, mas também permite-se a descarga deles em seu formato original, o XML. Neste trabalho, adotou-se a extração dos dados a partir de seu formato nativo, sendo que as tarefas foram realizadas na seguinte ordem: i) coleta dos currículos a partir da PL e amostragem; ii) inserção de informações dos fatores de impacto (IFs – Impact Factors) dos periódicos em que os pesquisadores publicaram trabalhos; iii) acréscimo de dados sobre o gênero dos pesquisadores, os conceitos Capes dos programas de pós-graduação de origem de cada pesquisador e informações sobre liderança em equipes de pesquisa. Tais tarefas são detalhadas na sequência.

[2]O Document Type Definition define a estrutura, os atributos e as regras para valores permitidos emum arquivo de dados XML.

Figura 1 – Visão geral sobre o processo de construção e aplicação do modelo de inferência



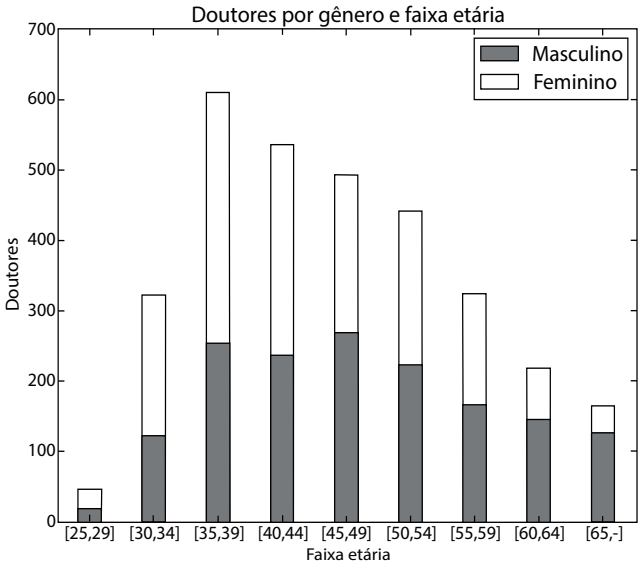
Fonte: Elaboração dos autores.

A PL disponibiliza números que descrevem os pesquisadores cadastrados no CL através da ferramenta Painel Lattes, os quais se tornam úteis para fins de validação da amostragem obtida. Para a área de conhecimento Odontologia, na data de coleta dos dados empregados nesta pesquisa (2015), havia 5.144 doutores com currículos cadastrados, dos quais 3.143 se declararam atuantes com pesquisa e ensino (P&E). A figura 2 mostra a distribuição de todos os currículos dos pesquisadores ligados à P&E no que tange à faixa etária e ao gênero. Quanto ao gênero, a distribuição é bastante equilibrada, sendo 1.589 (50,6%) currículos de pesquisadores do sexo feminino e 1.554 (49,4%) do sexo masculino.

Verifica-se que ocorre leve desequilíbrio na distribuição por faixa etária para o gênero feminino, havendo concentração proporcionalmente maior de currículos para indivíduos abaixo dos 45 anos. Isto sugere maior ingresso de mulheres no campo da pesquisa em períodos recentes, indicando resultado coerente com as políticas de incentivos governamentais, conforme apontado por Cavalcante et al. (2008), Guedes, Azevedo e Ferreira (2015). Para o conjunto masculino, ocorre distribuição mais equilibrada na quantidade de currículos após os 34 anos de idade, denotando certa homogeneidade do ingresso masculino no campo da P&E.

Dos 3.143 currículos, foram selecionados 1.499 currículos aleatoriamente (amostragem casual simples). A partir dessa seleção foram obtidos todos os doutores que atuavam com P&E no ano de 2015, resultando em amostragem de 1.079 currículos, 34,3% do total de pesquisadores apontados pela PL como atuantes na P&E. Considerando-se uma população finita, nível de confiança de 95% e a variabilidade máxima, a margem de erro proporcionada por esta coleta de dados é de ±2,42%. A proporção entre as classes amostradas foi de ≈23%, resultando em margem de erro ainda menor, 2,03%. Portanto, conclui-se que o tamanho amostral está adequado aos propósitos da investigação.

Figura 2 − Distribuição de currículos de doutores para a área Odontologia na PL



Fonte: CNPq (2015a), adaptado pelos autores.

Os dados amostrados se mostram coerentes com os valores globais apresentados pela PL, considerando outras áreas de conhecimento. Julga-se que o processo de coleta foi aleatório o suficiente para selecionar uma amostra representativa da população diante do que foi observado pelo número de nomes femininos e masculinos contados, 517 (47,9%) e 562 (52,1%), respectivamente.

Para compreender os dados que compõem a base de dados de currículos, obteve-se o DTD disponibilizado pelo PL, o qual define a estrutura e as regras para os atributos (ou campos) contidos nos arquivos XML. Verifica-se que a grande maioria dos campos é de preenchimento opcional e muitos deles ainda permitem que os dados sejam inseridos por digitação e não por meio de uma seleção entre opções predefinidas.

O potencial produtivo de um pesquisador é identificado de acordo com os critérios do CA-OD aplicado sobre os cômputos dos dados do CL e com suporte de dados que provêm de outra fonte. Os critérios, com vigência de 2015 a 2017, contemplam avaliações dos fatores de impacto (IFs) dos periódicos nos quais os pesquisadores publicaram seus trabalhos.

Assim, construiu-se um web *scraper/crawler* (i.e., um coletor de páginas web) para recuperar os IFs disponibilizados pelo CiteFactor (2015) e armazená-los para posterior uso durante a rotulação, considerando o conjunto de currículos obtidos. Amostras dos dados coletados foram verificadas manualmente junto aos periódicos e perante dados disponibilizados em fóruns do Research-Gate[3] validando a sua autenticidade.

É notório que a classificação de bolsistas PQ diante de seus pares levanta suspeitas do emprego de subjetividade. Foram identificados três potenciais aspectos subjetivos que podem ser considerados: o gênero, o conceito da instituição de formação e a liderança em equipes de pesquisa. Para tanto, realizou-se uma coleta dessas informações visando verificar os efeitos da adição destes na melhoria do modelo de inferência. Como tais dados não são fornecidos diretamente pelo CL do pesquisador, foram requeridos procedimentos adicionais para a sua obtenção:

a) o gênero de cada pesquisador foi inserido manualmente a partir do primeiro nome, uma vez que essa informação não consta nos currículos;

[3]Research Gate é uma rede social direcionada a pesquisadores com objetivo de facilitar o compartilhamento de informações, conhecimento e experiências. Possui cerca de 9 milhões de membros. Disponível em: <https://www.researchgate.net/>. Acesso em: 12/08/2015.

b) o indicador de qualidade da instituição de doutoramento dos pesquisadores foi obtido junto à plataforma Sucupira[4]. A figura 3 ilustra as distribuições de currículos coletados por conceito identificado;

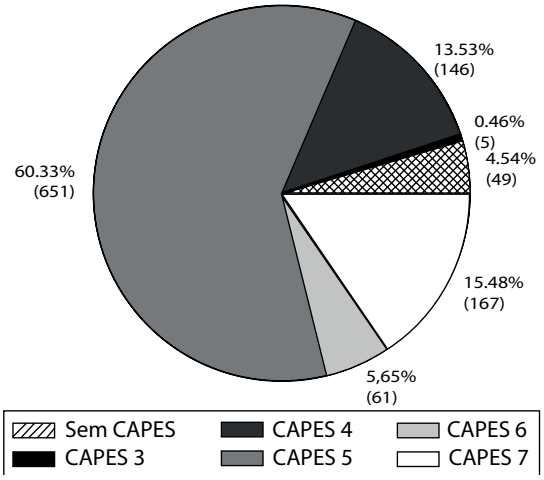
c) o Diretório de Grupos de Pesquisa (DGP) forneceu dados complementares quanto à liderança em equipes de pesquisa, os quais foram obtidos manualmente pelo acesso ao website do censo[5] e descarga dos arquivos censitários de 2010 a 2014. As participações globais, indicadas pelo Painel Lattes, encontram-se ilustradas pela figura 4a.

d) A figura 4b ilustra a participação coletada, sendo que as proporções apresentadas entre ambas as figuras reforçam a representatividade da amostragem. Quanto aos dados coletados, das 197 atuações encontradas, 128 ocorreram como primeiro líder (primeira barra) e 69 como segundo líder (segunda barra). Pela observação da figura 4b, é possível verificar que a atuação como primeiro líder não possui participações sobrepostas, ou seja, um dado pesquisador atuou em apenas um projeto com tal responsabilidade. O mesmo ocorre para o papel de segundo líder. No entanto, quando se verifica apenas a condição de liderança (terceira barra), ocorrem participações múltiplas em 16 projetos, indicando que, dos 1.079 currículos analisados, 16,77% atuaram em projetos de pesquisa registrados junto ao CNPq com tal condição.

PRÉ-PROCESSAMENTO

A primeira tarefa de pré-processamento correspondeu à identificação dos elementos relevantes no currículo. Após verificada a coerência da amostragem, foi necessário compreender a estrutura empregada na organização dos dados, por meio do DTD apresentado pela PL, a fim de realizar a pré-seleção dos elementos no currículo. Através da análise do DTD, estipulou-se quais campos possuíam potencial para aplicação na fase de mineração de dados, sendo identificados 87 elementos relevantes (não listados aqui por questões de espaço). Os elementos descrevem cada pesquisador contendo a evolução profissional, as produções bibliográficas, técnicas e artístico-culturais, as orientações concluídas e em andamento, as formações complementares, e as participações em bancas julgadoras, em eventos e congressos.

Figura 3 – Proporções e totais de currículos de pesquisadores obtidos agrupados por conceito da instituição de doutoramento

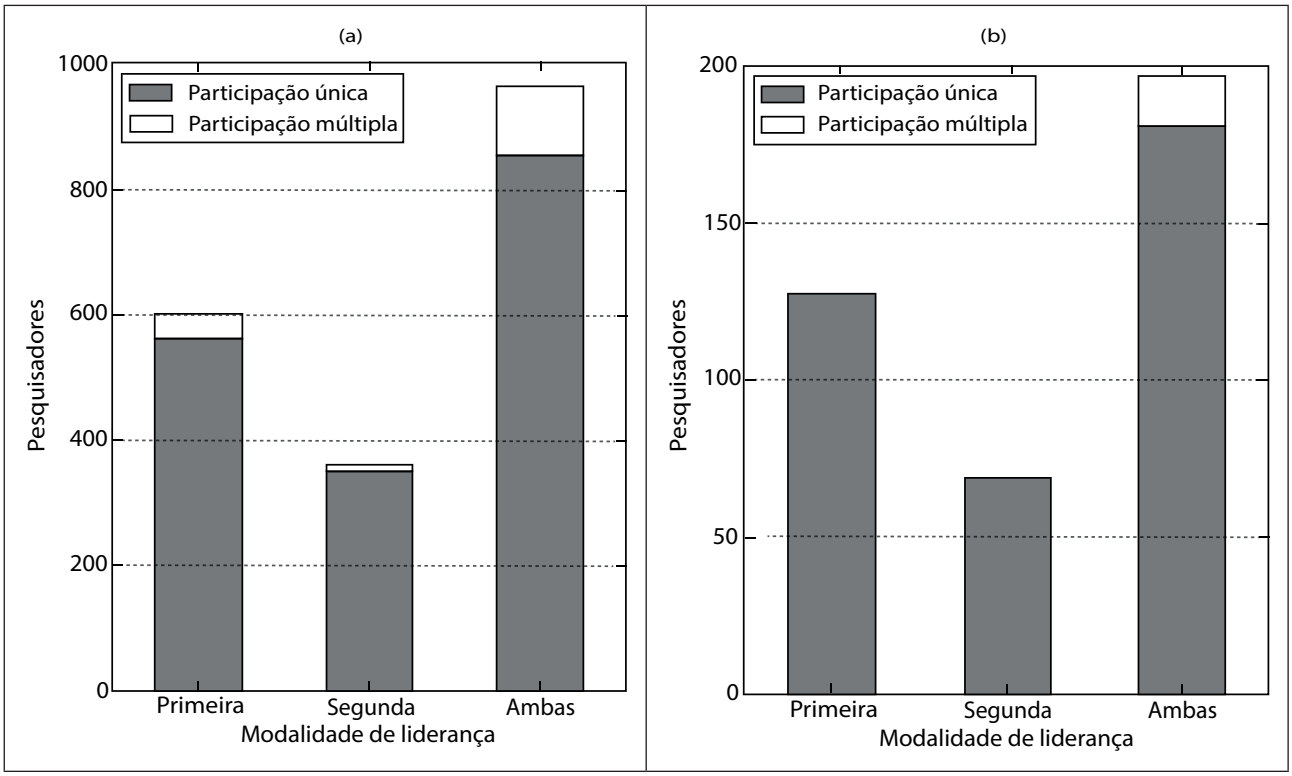


Fonte: Elaboração dos autores.

[4]Disponível em: <https://sucupira.capes.gov.br/sucupira/>. Acesso em: 20/02/2016.

[5]Disponível em: <http://lattes.cnpq.br/web/dgp/censos2>. Acesso em: 13/01/2016.

Figura 4 – Liderança em grupos de pesquisa registrados junto ao CNPq. (a) Todos os grupos de pesquisa registrados. (b) Grupos de pesquisa dos quais participam pesquisadores coletados



Fonte: (a) CNPq (2015a), adaptado pelo autor. (b) elaborado pelo autor.

Nota – A primeira barra representa a atuação de pesquisadores com papéis de líderes principais, e a segunda, como líderes secundários. A terceira barra apresenta a participação total, em ambas as modalidades. A "participação múltipla" representa a indicação de que o mesmo pesquisador atuou em vários projetos naquele papel de liderança.

A partir dos currículos colhidos, dos elementos pré-selecionados e da definição dos períodos de análise, foram realizados cômputos dos totais das atividades registradas. O período sob análise corresponde ao ano de referência da rotulação e à projeção em forma de deslocamento a período anterior ao ano de referência. Toma-se como exemplo a necessidade de verificação do estado do pesquisador projetado em 5 anos a partir do ano de 2010. O algoritmo assume 2015 como ano de referência para rotulação e inicia a contagem de atividades em 2001 com término em 2010, fixando o período de coleta em 10 anos. Assim, o procedimento consistiu em varrer os currículos para cada período estabelecido e contar os itens associados aos elementos escolhidos para análise. Assim, produziu-se um perfil da atuação científica do pesquisador que foi aplicado às etapas seguintes.

Destaca-se que não foi realizada a inserção de dados nos elementos identificados como úteis. O currículo é bastante flexível e possui poucos campos obrigatórios de preenchimento, o que impossibilita a percepção de que houve uma falha ou uma intenção em deixar um dado campo em branco. Assim, optou-se por não tratar "dados omissos" para não gerar prejuízos pela construção de perfis falsos, tais como a atribuição de certo número de publicações científicas a indivíduos que não as possuem de fato. Verificou-se que 57 (5,28%) pesquisadores não tiveram publicações registradas no período de 10 anos para os dados usados na projeção de 5 anos. A média para esta observação é de ≈26 artigos por pesquisador, bem como a mediana é 12. Logo, imputar algum valor a tais pesquisadores, de fato, geraria um perfil distorcido com impacto negativo na construção do modelo.

Na sequência, foi necessário efetuar a rotulação da classe a que cada pesquisador pertence considerando a produtividade científica e formação de recursos humanos. Como é uma tarefa deste trabalho a classificação dos pesquisadores, uma abordagem inerentemente supervisionada, exige-se que os dados sejam rotulados.

O indicador (rótulo) adequado do perfil de produtividade foi estabelecido como a conquista ou a potencial conquista de uma bolsa PQ. As bolsas são divididas por categorias e níveis cuja progressão, geralmente, se dá de forma sequencial e definidas pelo CA da área de atuação do pesquisador. Parte-se da categoria 2 e segue-se para 1D, 1C, 1B e 1A, sendo a última destinada a pesquisadores que demonstrem excelência na condução e consolidação de grupos de pesquisa, na formação de recursos humanos e na produção científica. A quantidade de bolsas PQ é limitada e os bolsistas de fato são selecionados: (a) pelo alcance de um conjunto mínimo requisitos; e (b) por uma classificação em relação aos seus pares. Havendo disponibilidade de bolsas, os candidatos aprovados em (a) são contemplados segundo sua ordem de classificação em (b). Ressalta-se, no entanto, que esse processo existe por causa da quantidade restrita de bolsas.

Na hipótese de haver quantidade ilimitada de bolsas, todos aqueles que cumprissem os requisitos apontados pelo CA-OD seriam categorizados como pesquisadores de alta produtividade científica. É com o suporte desses argumentos que o emprego do potencial de conquista da bolsa PQ, e não apenas da conquista de fato, se justifica no processo de rotulação das classes de pesquisadores a serem investigados.

O algoritmo de rotulação se encarrega de analisar o perfil do pesquisador mediante cada um dos requisitos do CA-OD. O processo envolve a verificação das contagens de produções tendo como referências os IFs recuperados; o exame das orientações no período sob análise; e, sob os critérios estabelecidos, o enquadramento em uma das situações previstas (quadro 1). É importante salientar que a categoria ou o nível da bolsa PQ não produzem uma classificação diferenciada, mas, tão somente, se há ou não o potencial para a sua conquista.

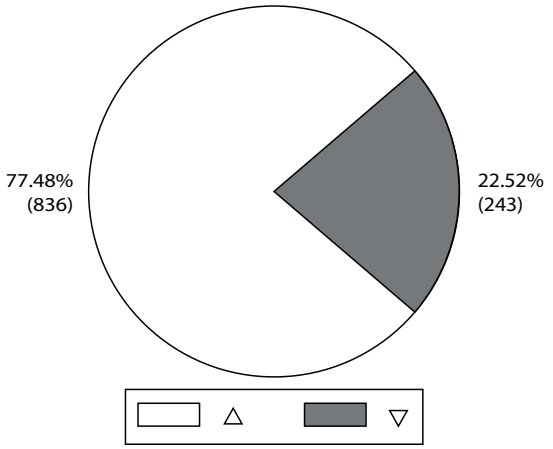
Portanto, ao fim, o rótulo será ▼, pesquisador sem potencial de bolsista PQ, ou △, pesquisador com potencial de bolsista PQ. Tais rótulos correspondem a um "pesquisador não produtivo" e a um "pesquisador produtivo", respectivamente. A figura 5 indica as proporções e quantidades das classes obtidas após o processamento da rotulação e, como é possível notar, ocorre desbalanceamento acentuado entre elas.

Quadro 1 − Requisitos mínimos para pleito de bolsas junto ao CA-OD.

Requisito	PQ 2	PQ 1D	PQ 1C/1B/1A
Publicações em periódicos	5 com IF >=1.0	20 com IF >= 1.0 sendo 5 com IF >=1.5	20 com IF >= 1.2 sendo 10 com IF >= 1.5
Orientações concluídas de mestres	1	2	3
Orientações concluídas de doutores	-	1	1
Orientações concluídas de pós-doutores	-	-	1
Orientações em andamento de mestre	Sim	Sim	Sim
Orientações em andamento de doutor	Sim	Sim	Sim
Orientações em andamento de pós-doutor	-	-	Sim

Fonte: CNPq (2015b), adaptado pelos autores.

Figura 5 − Desbalanceamento de classes entre currículos obtidos



Fonte: Elaboração dos autores.

As contagens expressando o perfil dos pesquisadores e os rótulos indicativos de seu potencial de conquista de bolsa são agrupados e armazenados. Estipulou-se que projeções para até 5 anos seriam plausíveis e que projeções superiores a este período seriam demasiadamente especulativas, já que o cômputo de dados envolve os últimos 10 anos de atuação do pesquisador. Com isto, inicialmente, foram criados 5 conjuntos de dados representando as projeções para 1, 2, 3, 4 e 5 anos, ou seja, com dados contabilizados para os anos de 2014, 2013, 2012, 2011 e 2010, respectivamente, cuja data de rotulação corresponde sempre a 2015.

Um procedimento adicional conferiu segmentação à contagem de artigos publicados gerando novos conjuntos de dados.O método anterior realizou o cômputo de artigos publicados sem discriminar o IF do periódico no qual ele foi publicado, gerando totalizações simples em 5 conjuntos de dados. Alternativamente, esta foi substituída por outras quatro contagens representando os cômputos dos artigos publicados diante dos IFs estabelecidos pelo CA-OD como identificadores do nível de produtividade do pesquisador.

Foram, assim, independentemente contabilizados os artigos de periódicos com IF: (a) inferior a 1,0; (b) de 1,0 a inferior a 1,2; (c) de 1,2 a inferior a 1,5; e (d) de 1,5 e acima.

A ideia é permitir uma comparação entre a produção maciça e a produção segmentada pela submissão dos conjuntos de atributos distintos à mineração de dados.

Do mesmo modo que o procedimento anterior, estes campos foram contabilizados em conjunto com os demais relevantes para períodos de projeções de 1, 2, 3, 4 e 5 anos e, por fim, armazenados para uso posterior. Assim, a quantidade de conjuntos de dados totais para submissão à análise é de 10 conjuntos.

Os perfis de pesquisadores com potencial para conquista de bolsas PQ na categoria 2 suscitam análise distinta. A categoria 1 apresenta mobilidade reduzida (GUEDES; AZEVEDO; FERREIRA, 2015), o que indica que um dado pesquisador que esteja naquela posição tende a se manter nela, possivelmente subindo de nível dentro da categoria no decorrer dos períodos. Ainda, apesar da transição da categoria 2 para a categoria 1 ocorrer, ela é limitada devido aos critérios de proporções estabelecidos pelo CNPq.

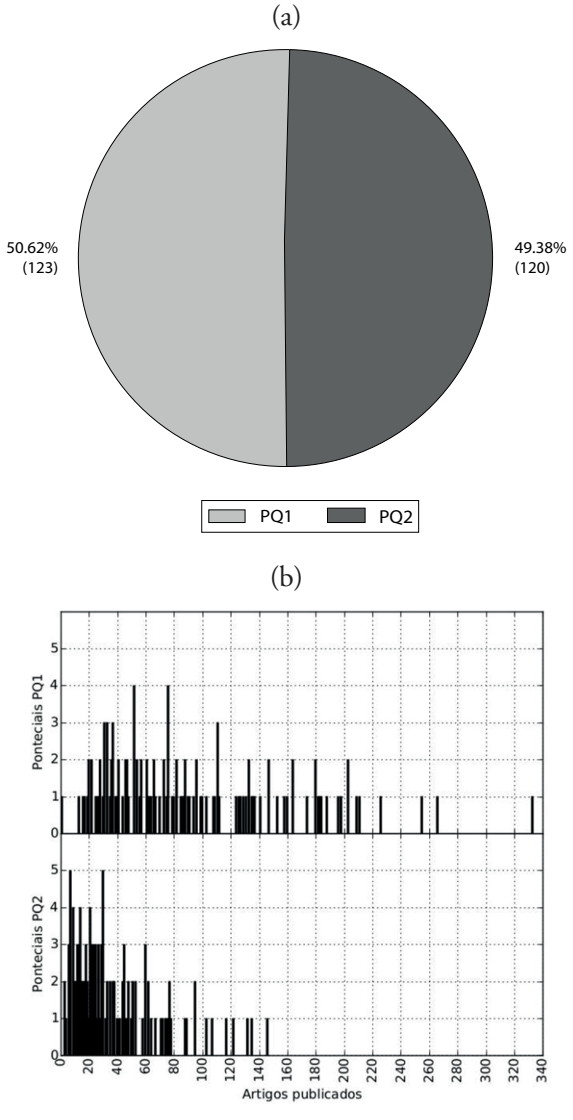
Portanto, analisar o perfil de potenciais PQ2, ou seja, daqueles que possuem produção científica relativamente reduzida em relação à categoria superior, pode apresentar resultados interessantes, seja para os que desejem fazer projeções para aumentar o desempenho próprio, especialmente para ingresso na categoria, ou mesmo para os gestores fazerem uso dos resultados para analisar tais perfis de forma mais pontual.

Serão portanto criados novos conjuntos de dados pela exclusão de potenciais bolsistas PQ1 do atual conjunto de dados, totalizando 20 conjuntos. Tomando como exemplos os dados gerados para projeções de 5 anos, a figura 6a ilustra a proporção entre os pesquisadores rotulados como PQ1 e PQ2, na qual é nítido o equilíbrio entre as classes.

A figura 6b ilustra a distribuição entre publicações de tais pesquisadores. Nesta, é possível perceber que os potenciais PQ2 se concentram em regiões com menor número de publicações totais, apesar de haver picos de produções por parte de alguns deles.

Já os potenciais PQ1 têm as produções distribuídas com maior uniformidade em faixa um pouco superior. Esta segmentação e as associações com os demais elementos pré-selecionados podem gerar modelos diferenciados para análises dos perfis PQ2 e foram, portanto, alvo de análises.

Figura 6 – Dados para análises de relações entre categorias. (a) Proporções entre totais de pesquisadores com potencial PQ1 e PQ2. (b) Distribuição de frequência da publicação de artigos acumulada em 10 anos para projeção de 5 anos. O gráfico superior ilustra pesquisadores com potencial PQ1 e o inferior de potencial PQ2



Fonte: Elaboração dos autores.

Assim, os dados apresentaram-se prontos para serem tratados pela próxima etapa.

SELEÇÃO DE ATRIBUTOS E TRANSFORMAÇÃO DE DADOS

Nesta etapa foram selecionados atributos relevantes para aplicação à construção do modelo de inferência. Em um primeiro passo, os atributos que possuíam baixa variância, aqui arbitrada como inferior a 0,1 e validada experimentalmente, foram excluídos. Tal procedimento se justifica devido ao fato de valores praticamente constantes para um dado atributo serem indiscriminantes dos perfis sob avaliação. Como resultado, dois conjuntos de atributos foram excluídos.

O primeiro conjunto considera dados de todos os pesquisadores, que contém: APRESENTACAO-DE-OBRA-ARTISTICA, APRESENTACAO-EM-RADIO-OU-TV, ARRANJO-MUSICAL, ARTES-CENICAS, ARTES-VISUAIS, CARTA-MAPA-OU-SIMILAR, COMPOSICAO-MUSICAL, DESENHO-INDUSTRIAL, MANUTENCAO-DE-OBRA-ARTISTICA, MAQUETE, MARCA, MBA, MUSICA, OBRA-DE-ARTES-VISUAIS, PARTICIPACAO-EM-EXPOSICAO, PARTICIPACAO-EM-FEIRA, PARTICIPACAO-EM-OLIMPIADA, PARTITURA-MUSICAL, PREFACIO-POSFACIO, SONOPLASTIA e TOPOGRAFIA-DE-CIRCUITO-INTEGRADO.

O segundo conjunto contém os atributos dos quais foram retirados pesquisadores com potencial identificado para PQ1, que são: ORIENTACAO-EM-ANDAMENTO-DE-POS-DOUTORADO e ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO. Diversos atributos possuem média e desvio padrão zero, bem como os demais estão bastante próximos a isso.

Dada a insignificância dos valores apresentados, há o indício de que eles não contribuem de forma direta com a indicação de produtividade científica de um pesquisador.

Em um segundo passo, atributos foram selecionados pela verificação de suas importâncias. Anteriormente, atributos irrelevantes foram excluídos por meio de um julgamento prévio do especialista e por apresentarem baixa variância. Agora, foram submetidos ao algoritmo de classificação para a definição de um ranking que representa o grau de importância de cada atributo. A definição de quais atributos são selecionados diante da quantidade especificada depende de suas respectivas importâncias individuais dentro do conjunto sob análise, as quais foram identificadas por teste estatístico univariado x 2 e por ganho de informação, resultando em duas baterias de testes aplicadas em cada um dos 20 conjuntos de dados especificados na tabela 1.

Esses conjuntos de dados agrupam os pesquisadores considerando: D1, todos os pesquisadores e sem segmentação de artigos por periódicos; D2, todos os pesquisadores e com segmentação de artigos por periódicos; D3, exceto os pesquisadores rotulados como PQ1 e sem segmentação de artigos por periódicos; e D4, exceto os pesquisadores rotulados como PQ1 e com segmentação de artigos por periódicos.

Um procedimento adicional seria a eliminação de atributos altamente correlacionados. Mas, cabe destacar, esse procedimento não foi identificado como uma boa prática por reduzir demasiadamente a quantidade de atributos disponíveis para observação pelo analista. Verificou-se que há situações nas quais dois ou três atributos seriam suficientes para um modelo computacional manter uma aproximação com os resultados mais bem caracterizados, mas tal quantidade seria insuficiente para um analista humano julgar um dado pesquisador. Como os resultados apresentaram complexidades pressupostas como adequadas, eles foram mantidos.

Tabela 1 - Totais de conjuntos de dados submetidos à classificação por projeção e cenário

Projeção	D1	D2	D3	D4
1 ano	68	71	67	70
2 anos	67	70	66	69
3 anos	66	69	65	68
4 anos	65	68	65	68
5 anos	66	69	65	68

Fonte: Elaboração dos autores.

RESULTADOS E DISCUSSÃO

Cada conjunto de dados foi submetido a um classificador que gera um ranking dos melhores atributos considerando cada uma das projeções. De forma mais específica, o classificador Random Forest foi utilizado, pois gera estimativas do grau de importância de cada atributo no desempenho da classificação. Os atributos selecionados são apresentados na tabela 2.

A tabela 2 separa os pesquisadores em dois grupos, contendo os melhores atributos considerando o potencial de todos os pesquisadores e apenas os pesquisadores com potencial para PQ2. Como exemplo, suponha que desejemos verificar quais atividades determinado pesquisador deve se empenhar caso pretenda concorrer a uma bolsa PQ daqui a 3 anos. Nesse contexto, o foco inicial poderia ser em certa quantidade de atributos, por exemplo, nos 10 atributos mais representativos. Assim, em ordem de importância, esses atributos seriam: ORIENTACOES-CONCLUIDAS-PARA-MESTRADO, ARTIGO-FATOR-IMPACTO-1-5, ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO, ARTIGO-FATOR-IMPACTO-1-2, PARTICIPACAO-EM-BANCA-DE-EXAME-QUALIFICACAO, ARTIGO-FATOR-IMPACTO-0, PARTICIPACAO-EM-BANCA-DE-DOUTORADO, ARTIGO-FATOR-IMPACTO-1-0, PARTICIPACAO-EM-BANCA-DE-MESTRADO e ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO. Os 10 atributos usados no exemplo são os que mais se destacaram no modelo preditivo, considerando as atividades de pesquisadores PQ em um período de 10 anos. Logo, representam uma meta a ser atingida por quem pretende ter condições de pleitear uma bolsa de produtividade em até 3 anos.

Tabela 2 − Ordem de relevância dos atributos selecionados por conjunto de dados por projeção para os modelos de simulação

Atributo	Todos os pesquisadores					Exceto potenciais PQ1				
	1 ano	2 anos	3 anos	4 anos	5 anos	1 ano	2 anos	3 anos	4 anos	5 anos
APRESENTACAO-DE-TRABALHO	-	-	-	-	-	-	2º	2º	28º	2º
ARTIGO-ACEITO-PARA-PUBLICACAO	-	-	-	27º	23º	-	-	-	-	-
ARTIGO-FATOR-IMPACTO-0	7º	9º	6º	4º	4º	6º	10º	10º	3º	11º
ARTIGO-FATOR-IMPACTO-1-0	6º	7º	8º	9º	7º	10º	23º	24º	9º	28º
ARTIGO-FATOR-IMPACTO-1-2	5º	4º	4º	3º	3º	3º	8º	8º	2º	9º
ARTIGO-FATOR-IMPACTO-1-5	4º	3º	2º	1º	1º	4º	9º	9º	1º	8º
BANCA-JULGADORA-PARA-CONCURSO-PUBLICO	15º	13º	13º	12º	12º	13º	-	-	16º	-
BANCA-JULGADORA-PARA-LIVRE-DOCENCIA	14º	12º	12º	11º	11º	19º	26º	27º	25º	24º
BANCA-JULGADORA-PARA-PROFESSOR-TITULAR	27º	29º	29º	28º	25º	-	-	-	-	-
CONSELHO-COMISSAO-E-CONSULTORIA	18º	14º	14º	13º	13º	15º	20º	19º	12º	22º
CURSO-DE-CURTA-DURACAO-MINISTRADO	-	-	-	-	-	-	13º	13º	-	13º
DEMAIS-TRABALHOS	-	-	-	-	-	-	28º	30º	-	21º
DIRECAO-E-ADMINISTRACAO	23º	21º	19º	20º	22º	21º	17º	18º	21º	19º
EDITORACAO	-	-	-	-	-	-	25º	26º	-	-
ENSINO	30º	27º	26º	29º	28º	28º	-	-	30º	-
EXTENSAO-UNIVERSITARIA	-	-	-	-	-	-	-	29º	-	27º
ORGANIZACAO-DE-EVENTO	-	-	-	-	-	27º	-	-	-	-
ORIENTACAO-EM-ANDAMENTO-DE-APERFEICOAMENTO-ESPECIALIZACAO	29º	28º	23º	22º	24º	-	-	-	20º	-
ORIENTACAO-EM-ANDAMENTO-DE-DOUTORADO	3º	5º	11º	16º	-	5º	18º	22º	27º	-
ORIENTACAO-EM-ANDAMENTO-DE-GRADUACAO	-	24º	21º	24º	27º	-	-	-	23º	-
ORIENTACAO-EM-ANDAMENTO-DE-INICIACAO-CIENTIFICA	19º	-	-	-	-	18º	-	-	-	-
ORIENTACAO-EM-ANDAMENTO-DE-MESTRADO	11º	18º	-	-	-	9º	-	-	-	-
ORIENTACAO-EM-ANDAMENTO-DE-POS-DOUTORADO	13º	-	-	-	-	-	-	-	-	-
ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO	2º	2º	3º	5º	5º	2º	12º	12º	8º	14º

(Continua)

Tabela 2 − Ordem de relevância dos atributos selecionados por conjunto de dados por projeção para os modelos de simulação (Conclusão)

Atributo	Todos os pesquisadores					Exceto potenciais PQ1				
	1 ano	2 anos	3 anos	4 anos	5 anos	1 ano	2 anos	3 anos	4 anos	5 anos
ORIENTACOES-CONCLUIDAS-PARA-MESTRADO	1º	1º	1º	2º	2º	1º	6º	6º	5º	4º
ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO	12º	11º	10º	10º	10º	-	-	-	-	-
OUTRA-PRODUCAO-BIBLIOGRAFICA	-	-	-	-	-	-	27º	28º	-	18º
OUTRA-PRODUCAO-TECNICA	-	-	-	-	-	-	19º	17º	-	15º
OUTRAS-BANCAS-JULGADORAS	21º	22º	20º	18º	16º	20º	14º	16º	24º	17º
OUTRAS-ORIENTACOES-CONCLUIDAS	16º	15º	15º	15º	14º	11º	11º	11º	14º	10º
OUTRAS-PARTICIPACOES-EM-BANCA	28º	26º	27º	25º	19º	24º	29º	25º	17º	26º
PARTICIPACAO-EM-BANCA-DE-DOUTORADO	8º	8º	7º	7º	8º	8º	5º	5º	7º	5º
PARTICIPACAO-EM-BANCA-DE-EXAME-QUALIFICACAO	9º	6º	5º	6º	6º	7º	4º	4º	4º	7º
PARTICIPACAO-EM-BANCA-DE-MESTRADO	10º	10º	9º	8º	9º	-	3º	3º	6º	3º
PARTICIPACAO-EM-CONGRESSO	-	-	-	-	-	-	21º	21º	-	25º
PARTICIPACAO-EM-ENCONTRO	-	-	-	-	-	-	30º	-	-	-
PARTICIPACAO-EM-PROJETO	24º	19º	18º	19º	18º	17º	15º	14º	13º	12º
PARTICIPACAO-EM-SEMINARIO	-	-	-	-	-	26º	-	-	-	29º
PARTICIPACAO-EM-SIMPOSIO	22º	23º	22º	21º	20º	25º	-	-	18º	-
PRODUTO-TECNOLOGICO	-	30º	-	-	-	30º	-	-	-	-
PROGRAMA-DE-RADIO-OU-TV	-	-	30º	30º	30º	-	-	-	-	-
RELATORIO-DE-PESQUISA	-	-	28º	-	29º	16º	16º	15º	15º	16º
SOFTWARE	-	-	-	-	-	-	-	-	26º	-
TEXTO-EM-JORNAL-OU-REVISTA	25º	25º	25º	23º	21º	22º	-	-	19º	-
TRABALHO-EM-EVENTOS	26º	20º	24º	26º	26º	23º	1º	1º	22º	1º
TRABALHO-TECNICO	20º	17º	17º	17º	17º	12º	7º	7º	10º	6º
TRADUCAO	-	-	-	-	-	29º	-	-	29º	30º

Fonte: Elaboração dos autores.

CONCLUSÕES

Apesar dos incentivos concedidos à evolução científica no país, ainda há ações a serem tomadas para conduzir os níveis produtivos atuais a patamares mais elevados. O desenvolvimento científico requer aplicações direcionadas de recursos destinados à pesquisa, bem como se beneficiaria pela composição de equipes de trabalho otimizadas para a execução exitosa de projetos. Adicionalmente, a avaliação individual do perfil do pesquisador pode ser empregada como suporte ao próprio crescimento profissional, com impactos diretos na qualidade da atuação em projetos. Trata-se de um tema de estudo não esgotado e com ganhos relevantes para toda a sociedade.

O presente trabalho teve o objetivo de investigar materiais e métodos para obter a inferência do potencial produtivo futuro de pesquisadores da área de conhecimento Odontologia.Para tal, foram empregados métodos de descoberta de conhecimento em dados sobre currículos de pesquisadores cadastrados na Plataforma Lattes. Nesse contexto, explorou-se um conjunto de técnicas de seleção de atributos e algoritmos de mineração de dados aplicados a dados rotulados a partir de critérios estabelecidos pelo Comitê de Assessoramento para concessão de bolsas de Produtividade em Pesquisa na área.

Foram determinadas as características mais relevantes para a inferência do nível produtivo, tendo elas papel de elevada importância na geração dos respectivos modelos diante dos períodos de projeções futuras. Verificou-se que os pesquisadores mais recentemente atuantes precisam desenvolver, em média, mais atividades distintas para obter destaque.

Como conclusão, foi possível deduzir que a produtividade de um pesquisador está associada a um conjunto de ações e não a um volume de atividades específicas. De fato, a quantidade de produção para determinação do nível de produtividade é definida por órgãos competentes, mas a variação de atividades produz um composto que resulta na elevação indireta dos itens avaliados e com resultados diretos na produtividade.

Com isto, há um conjunto de n atributos gerando 2^n perfis de pesquisadores, dado que cada atributo representa uma atividade que pode ou não ser executada. Por conseguinte, este trabalho contribui com o apontamento da relevância e de quais atividades específicas favorecem a determinação do perfil de desempenho dos pesquisadores, facilitando as reflexões sobre pontos de melhoria que podem ser empregados em direcionamentos para atuações futuras desses pesquisadores.

Como trabalhos futuros, pretende-se disponibilizar um sistema de geração de relatórios e inferências para estudos de produtividade de equipes de pesquisadores, bem como um ambiente para simulações de perfis de pesquisadores gratuito via web.

REFERÊNCIAS

ABBASI, A.; ALTMANN, J.; HOSSAIN, L. Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, v. 5, n. 4, p. 594 -607, 2011.

ACUNA, D. E.; ALLESINA, S.; KORDING, K. P. Future impact: predicting scientific success. *Nature*, v. 489, n. 7415, p. 201- 202, 2012.

BRYNKO, B. The science of predicting nobel prize winners. *Information Today*, v. 27, n. 10, p. 43- 43, 2010.

CAVALCANTE, R. A. et al. Perfil dos pesquisadores da área de odontologia no Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). *Revista Brasileira de Epidemiologia*, v. 11, p. 106 - 113, 2008.

CIMENLER, O.; REEVES, K. A.; SKVORETZ, J. A regression analysis of researchers' social network metrics on their citation performance in a college of engineering. *Journal of Informetrics*, v. 8, n. 3, p. 667 - 682, 2014.

CONSELHO NACIONAL DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. *Plataforma Lattes*. 2015. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 31 mar. 2015.

_____. *Critérios de Julgamento* - CA-OD: Vigência 2015 a 2017. 2015. Disponível em: <http://cnpq.br/web/guest/view/-/journal_content/56_INSTANCE_0oED/10157/49701>. Acesso em: 12 maio 2015.

DIJK, D.; MANOR, O.; CAREY, L. Publication metrics and success on the academic job market. *Current Biology*, v. 24, n. 11, p. R516 - R517, 2014.

GUEDES, M. de C.; AZEVEDO, N.; FERREIRA, L. O. A produtividade científica tem sexo? Um estudo sobre bolsistas de produtividade do CNPq. *Cadernos Pagu*, p. 367 - 399, 2015.

LAURANCE, W.F. et al. Predicting publication success for biologists. *BioScience*, v. 63, n. 10, 2013.

LETA, J. Brazilian growth in the mainstream science: the role of human resources and national journals. *Journal of Scientometric Research*, v. 1, n. 1, p. 44-52, 2012.

MUGNAINI, R.; JANUZZI, P.M.; QUONIAM, L. Indicadores bibliométricos da produção científica brasileira: uma análise a partir da base pascal. *Ciência da Informação*, v. 33, n. 2, p. 123 -131, 2004.

NOORDEN, R.V. *The impact gap*: South America by the numbers. 2014. Disponível em: <http://www.nature.com/news/the-impact-gap-south-america-by-the-numbers-1.15393 >. Acesso em: 16/04/2015.

REZENDE, S.M. Produção científica e tecnológica no brasil: conquistas recentes e desafios para a próxima década. *RAE*, v. 51, p. 202 - 2011.

SARIGÖL, E. et al. Predicting scientific success based on coauthorship networks. *EPJ Data Science,* v. 3, n. 1, p. 1-16, 2014.