

Identificação de Acadêmicos Promissores de Sucesso Científico através de Programação Linear Inteira

Antônio de Abreu Batista Júnior

Universidade Federal do Maranhão - UFMA
Departamento de Informática, São Luís, MA, Brasil
antonio.batista@ufma.br

Jesús Pascual Mena Chalco

Universidade Federal do ABC - UFABC
Centro de Matemática, Computação e Cognição, Santo André, SP, Brasil
jesus.mena@ufabc.edu.br

RESUMO

Embora quantificar a relevância atual de acadêmicos seja de interesse substancial, existe uma ampla variedade de situações em que a relevância futura é mais importante do que a atual (e.g., em orientar decisões de financiamento de pesquisa, contratação e premiação). Neste artigo, propomos um modelo para predição de acadêmicos de sucesso. Explorando o impacto científico a curto prazo de acadêmicos avaliado através de indicadores derivados do cálculo do número de citações, formulamos um modelo baseado no arcabouço da mochila inteira. Testamos o modelo usando dados de publicação e de citação disponíveis na base de dados da ArnetMiner, de 78230 acadêmicos cuja a primeira publicação foi no ano de 1998 ou 1999. Os resultados experimentais indicam que o modelo pode auxiliar a encontrar uma parcela grande de acadêmicos com sucesso científico.

PALAVRAS CHAVE. Relevância futura de acadêmicos, Programação linear inteira, Cientometria.

Outras aplicações em PO, Programação matemática, Otimização combinatória.

ABSTRACT

Although quantifying the current relevance of academics is of substantial interest, there is a wide variety of situations in which future significance is more important than the current one. In this article, we propose a model for the prediction of success' scholars. Exploring the short-term scientific impact of academics assessed through indicators derived from calculating citation counts, we formulated a model based on the integer knapsack framework. We tested the model using publication and citation data available in ArnetMiner database, from 78230 academics whose first publication was in the year 1998 or 1999. The experimental results indicate that the model allows finding a large portion of scholars with scientific success.

KEYWORDS. Future relevance of scholars, Integer linear programming, Scientometrics.

Other applications in OR, Mathematical programming, Combinatorial optimization.

1. Introdução

A necessidade de avaliar o potencial de impacto futuro de acadêmicos com pouco tempo de carreira ocorre em um número de aplicações, incluindo decisões de promoção, contratação de recursos humanos e apoio à pesquisa [Gogoglou e Manolopoulos, 2017]. Para essas aplicações, a tarefa de estimar a relevância futura de acadêmicos, em vez de suas realizações passadas (relevância atual), tem grande importância para tomadores de decisão acadêmicos e de agências de fomento à pesquisa. No entanto, a disponibilidade crescente de dados sobre os resultados de pesquisa somado ao pequeno tempo de carreira destes acadêmicos torna essa tarefa desafiadora.

Recentemente, um número elevado de trabalhos tem sido feito para quantificar o impacto científico de acadêmicos e identificar os mecanismos que influenciam sua evolução futura. Dong et al. [2016] identificaram fatores quantitativos que dirigem artigos publicados a aumentarem o índice-h futuro de seus autores. Por outro lado, Qi et al. [2017] mediram o impacto das colaborações na carreira futura de um pesquisador e destacaram o crescimento da carreira de jovens pesquisadores em colaborar com pesquisadores importantes.

Abrishami e Aliakbary [2019] sugeriram medir o potencial de impacto futuro de acadêmicos indiretamente através do potencial de impacto futuro de seus artigos (número de citações futuras) usando técnicas de aprendizado profundo. Por outro lado, Zhou et al. [2018] destacaram a atribuição de créditos para os coautores de publicações com muitos autores como um componente crítico para determinar o potencial de impacto futuro de cientistas.

Acuna et al. [2012] apresentaram uma fórmula para estimar o índice-h futuro de cientistas da ciência da vida usando métodos de regressão linear simples e Bütün e Kaya [2019] consideraram a predição de número de citações de cientistas como um problema de predição de *links*.

Recentemente, Shi et al. [2019] desenvolveram um método para identificação de artigos importantes baseado em citações precoces de descobridores¹. No entanto, uma importante contribuição desse trabalho é a descoberta de que os próprios descobridores são potenciais acadêmicos de sucesso uma vez que eles tem recebido citações acima da média. Outros trabalhos usaram métodos de aprendizado supervisionado [Ning et al., 2017; Weihs e Etzioni, 2017] e não supervisionado [Panagopoulos et al., 2017] a fim de desenvolver modelos para predição de acadêmicos de sucesso, e uma parcela grande de outros trabalhos explorou caminhos aleatórios em grafos de coautoria [Wijegunawardana et al., 2016; Ding et al., 2018].

E, apesar da ampla variedade de abordagens propostas para quantificar o impacto científico futuro de pesquisadores, um arcabouço computacional (uma estrutura de apoio essencial) para prever acadêmicos de sucesso a partir de diferentes indicadores bibliométricos ainda é pouco explorado [Gogoglou e Manolopoulos, 2017].

Neste artigo, propomos um modelo para predição de acadêmicos de sucesso formulado como o problema da mochila inteira com restrição de cardinalidade [Caprara et al., 2000]. O modelo permite identificar precocemente (i.e., pouco tempo de carreira) futuros acadêmicos de sucesso científico a partir de seus impactos científico a curto prazo avaliados por meio de indicadores derivados de citação. No melhor do nosso conhecimento, este é o primeiro modelo para predição de acadêmicos de sucesso usando o arcabouço da mochila, sendo assim, a principal contribuição deste trabalho.

O restante deste artigo está estruturado da seguinte forma: a Seção 2 apresenta as características preditoras de sucesso científico usadas no modelo proposto; a Seção 3 apresenta a

¹Descobridores são pesquisadores jovens que frequentemente citam artigos recém publicados que se tornam altamente citados no futuro.

formulação do modelo de otimização para o problema de predição de acadêmicos; a Seção 4 apresenta o conjunto de dados; a Seção 5 apresenta um estudo de caso e os resultados obtidos por meio do modelo de otimização proposto; finalmente, a Seção 6 apresenta as conclusões do trabalho e possíveis direções para futuras pesquisas.

2. Características preditoras

Um número de fatores pode determinar o impacto futuro de um acadêmico, variando da sua área de pesquisa aos locais de publicações diferentes onde ele tem publicado. Segundo [Gogoglou e Manolopoulos, 2017] os fatores preditores podem ser divididos em três categorias, segundo a origem do fator: (a) centrados em características pessoais; (b) centrados na publicação; e (c) centrados no veículo de publicação. Na Figura 1 apresentamos um resumo dos principais fatores. Sobre tais fatores pode-se especular que um cientista que tem publicado em vários periódicos diferentes é também provável ser alguém com treinamento amplo que contribui de muitas formas, assim como o número de publicações em periódicos importantes aumenta a visibilidade dos outros artigos de um acadêmico, publicados e ainda a ser publicados [Acuna et al., 2012].

Hoje, sabemos que um acadêmico bem conectado, com uma grande rede de colaboração, e pertencente a uma instituição respeitada terá mais oportunidades de divulgação do seu trabalho [Gogoglou e Manolopoulos, 2017]. Também, é bem entendido que para se publicar em periódicos científicos de grande importância em uma determinada área de pesquisa, acadêmicos precisam convencer revisores e editores muito críticos [Bornmann et al., 2014]. O alto impacto de periódicos científicos reflete a capacidade de publicação de um acadêmico enquanto que o alto impacto de um artigo reflete a sua utilidade para as pesquisas futuras. Sobre isso pode-se especular que o número de artigos publicados em periódicos científicos de alto impacto e o número de citações (utilidade) desses trabalhos de pesquisa são boas características preditoras de sucesso científico [Acuna et al., 2012].

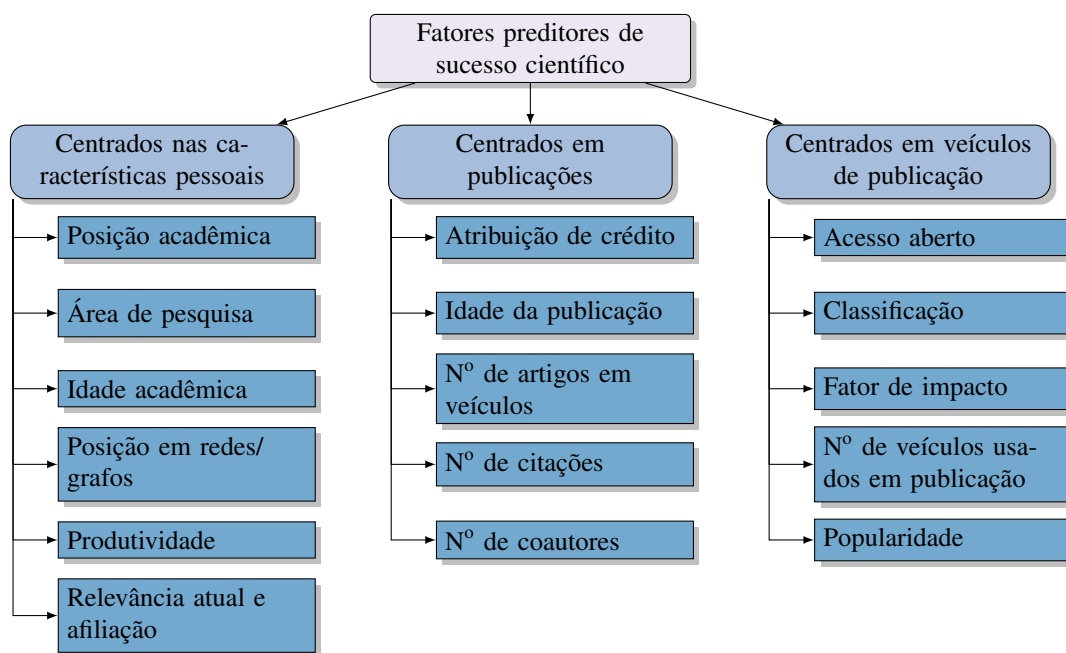


Figura 1: Categorias de fatores preditores de acadêmicos de sucesso.

Finalmente, deve-se notar que a quantificação da evolução temporal de resultados de pesquisa de um acadêmico (e.g., a taxa de crescimento do número de citações ano a ano, e a taxa de atividade dentro de uma janela de tempo) podem fornecer mais informação sobre o seu desempenho futuro do que a simples quantificação do seu desempenho de pesquisa atual (e.g., número total de artigos publicados) [Weihs e Etzioni, 2017]. Além disso, também é importante notar que é pouco conhecido atualmente os pontos fortes e fracos de indicadores preditores de sucesso científico do estado da arte [Penner et al., 2013].

3. Modelo baseado no arcabouço da mochila para predição de acadêmicos de sucesso

Neste seção, apresentamos a base do modelo que pretendemos utilizar para identificação de jovens acadêmicos promissores de sucesso científico. O modelo, dado um conjunto de pessoas $P = \{1, 2, \dots, n\}$ indexadas por j , em que cada uma tem uma idade acadêmica² a_j e uma relevância futura p_j , permitirá obter um subconjunto delas que reúne a maior relevância futura, supondo que existe um inteiro W e um inteiro k que restringe, respectivamente, a máxima soma de idades e o número de acadêmicos nesse subconjunto. Acreditamos que esse modelo possa auxiliar tomadores de decisão acadêmicos em decisões, como por exemplo, de contratação de recursos humanos, e de apoio ou fomento à pesquisa.

Neste contexto, denotamos como acadêmico promissor de sucesso científico uma pessoa, acadêmica jovem, com grande potencial mas com pouca informação bibliométrica atrelada à sua atuação científica devido à sua pouca idade acadêmica. Formalmente, o problema de identificação de jovens acadêmicos promissores de sucesso científico é formulado como o problema da mochila booleana usando programação linear inteira. Nesse sentido, as variáveis de decisão que podemos considerar são definidas como:

$$x_j = \begin{cases} 1 & , \text{ se } j \text{ é um acadêmico promissor de sucesso científico.} \\ 0 & , \text{ caso contrário.} \end{cases} \quad (1)$$

$$\text{E, o objetivo é maximizar } \sum_{j \in P} p_j x_j \quad (2)$$

$$\text{sujeito a } \sum_{j \in P} A_j x_j \leq W, \quad (3)$$

$$\sum_{j \in P} x_j \leq K, \text{ com } x_j \in \{0, 1\} \text{ e } j \in P. \quad (4)$$

A função objetivo, Eq. 2, busca pelo subconjunto de acadêmicos de máxima soma de saída de pesquisa futura (relevância futura). A restrição, Eq. 3, garante que a soma de idades acadêmicas dos membros do subconjunto não seja maior do que um certo valor inteiro W . As restrições, Eq. 4, impõem um número máximo de acadêmicos que podem fazer parte da solução e define o domínio das variáveis de decisão. Neste trabalho, consideramos que $\sum_{j \in P} A_j > W$ e $A_j \leq W$ para cada $j \in P$.

²Por idade acadêmica entendemos, o período de tempo decorrido desde o ano da primeira publicação científica de um pesquisador até o ano de referência, por exemplo o ano corrente. O ano da primeira publicação tem sido frequentemente considerado um *proxy* da idade do pesquisador [Nane et al., 2017].

O impacto científico de um acadêmico pode ser avaliado a partir de diferentes indicadores, como por exemplo o número de citações ou seu índice-h. A fim de compararmos diferentes acadêmicos com diferentes idades normalizamos os seus impactos científicos pela suas idades acadêmicas. Nesse contexto, inicialmente para esta proposta adotamos como impacto futuro de um acadêmico, o seu impacto a curto prazo modulado por sua idade:

$$\underbrace{\text{impacto científico futuro}}_{p_j} = \underbrace{\text{impacto a curto prazo}}_{\frac{c_j}{A_j} + \frac{h_j}{A_j}} = \frac{c_j + h_j}{A_j} \quad (5)$$

em que c_j é o número total de citações do acadêmico j de idade A_j e h_j é seu índice-h [Hirsch, 2005]. Matematicamente, um acadêmico com um conjunto de n artigos, em que o i -ésimo deles tem recebido exatamente X_i citações, tem índice-h = $\max\{i : X_i \geq i\}$.

Como um exemplo ilustrativo, para melhor entendimento da proposta, suponha que tenhamos um conjunto de sete acadêmicos em que foram coletadas as informações bibliométricas relacionadas com o índice-h, e número de citações. Na Tabela 1, apresentamos todos os valores iniciais para cada um dos sete. Para efeitos de simplificação, consideramos a idade acadêmica um valor inteiro que representa apenas o número total de anos a partir de sua primeira publicação. A partir do desempenho científico de cada um pode se notar que os acadêmicos 2 e 5 são aqueles com o maior impacto a curto prazo.

Tabela 1: Exemplo de características de um conjunto de sete acadêmicos.

Acadêmico	Índice-h	Citações (c)	Idade acadêmica (A)	Potencial (p)
1	2	12	2	7
2	3	15	2	9
3	1	2	1	3
4	2	6	2	4
5	1	8	1	9
6	2	2	1	4
7	4	20	3	8

Na Tabela 2, apresentamos todos os subconjuntos de tamanho no máximo 2. Notamos que, as combinações 13, 18 e 25 não são soluções viáveis devido à soma das idades dos acadêmicos ser superior a $W = 4$. Desconsiderando essas 3 combinações, todas as outras combinações podem ser consideradas soluções viáveis. Finalmente, notamos que, dentre as soluções viáveis, a combinação 16, composta pelos acadêmicos 2 e 5, é aquela com a maior soma de potenciais e portanto a solução que agrupa as pessoas com maior potencial de pesquisa. É importante destacar que podem ser identificadas mais de uma solução ótima, nesses casos uma análise dessas soluções seria interessante, por exemplo, identificar acadêmicos substituíveis entre soluções ou encontrar a solução de soma mínima de idades.

Dentre todas as sete pessoas, os acadêmicos 2 e 5 teriam os melhores desempenhos atuais. Eles também são preditos por terem os melhores desempenhos futuros. Na subseção 4, apresentamos um conjunto de dados de mais de 2 milhões de artigos associados a mais de 1 milhão de autores e evidenciamos que um bom desempenho acadêmico atual é uma forte característica preditora de sucesso científico.

Note que este modelo, de forma intuitiva permite calcular o valor da relevância futura de um acadêmico de muitas formas diferentes. Assim, novos modelos/métodos podem ser utilizados

Tabela 2: Todas as combinações de tamanho no máximo $K = 2$.

Combinação	Solução (S)	Soma de potenciais ($\sum_{j \in S} p_j$)	Soma de idades ($\sum_{j \in S} A_j$)	Tipo de Solução
1	{1}	7	2	viável
2	{2}	9	2	viável
3	{3}	3	1	viável
4	{4}	4	2	viável
5	{5}	9	1	viável
6	{6}	4	1	viável
7	{7}	8	3	viável
8	{1, 2}	16	4	viável
9	{1, 3}	10	3	viável
10	{1, 4}	11	4	viável
11	{1, 5}	16	3	viável
12	{1, 6}	11	3	viável
13	{1, 7}	15	5	não viável
14	{2, 3}	12	3	viável
15	{2, 4}	13	4	viável
16	{2, 5}	18	3	ótima
17	{2, 6}	13	3	viável
18	{2, 7}	17	5	não viável
19	{3, 4}	7	3	viável
20	{3, 5}	12	2	viável
21	{3, 6}	7	2	viável
22	{3, 7}	11	4	viável
23	{4, 5}	13	3	viável
24	{4, 6}	8	3	viável
25	{4, 7}	12	5	não viável
26	{5, 6}	13	2	viável
27	{5, 7}	17	4	viável
28	{6, 7}	12	4	viável

ou propostos a fim de obter indicadores que sirvam de insumo para o modelo apresentado nesta seção.

4. Conjunto de dados

Para todos os testes desta proposta foi considerado um conjunto de dados extraído da ArnetMiner [Tang et al., 2008]³. O conjunto de dados da ArnetMiner é uma coleção de publicações da área de computação coletada a partir do repositório bibliográfico de ciência da computação (DBLP⁴), da biblioteca digital da (ACM⁵) e do CiteSeer⁶. O conjunto de dados tem registros de publicações de 1936 até maio de 2014.

A base de dados tem sido usada em vários trabalhos que consideram como objeto a área da ciência da computação [Zhang et al., 2016]. Ela inclui informações de artigo, de citação, de autor e de colaboração científica. Ao todo, são 2.092.356 artigos registrados de 1936 à 2014, com 1.207.061 autores únicos, e 8.024.869 citações.

Neste trabalho, como estudo de caso, extraímos todos os artigos publicados entre os anos de 1995 e 2014 (20 anos), em um total de 1.723.791 deles publicados em mais de 150.000 veículos de publicação diferentes, e distribuídos anualmente conforme o gráfico da Figura 2. Estes dados

³<http://aminer.org/aminernetwork>, último acesso em abril de 2019.

⁴<http://dblp.uni-trier.de>, último acesso em abril de 2019.

⁵<http://dl.acm.org>, último acesso em abril de 2019.

⁶<http://csxstatic.ist.psu.edu>, último acesso em abril de 2019.

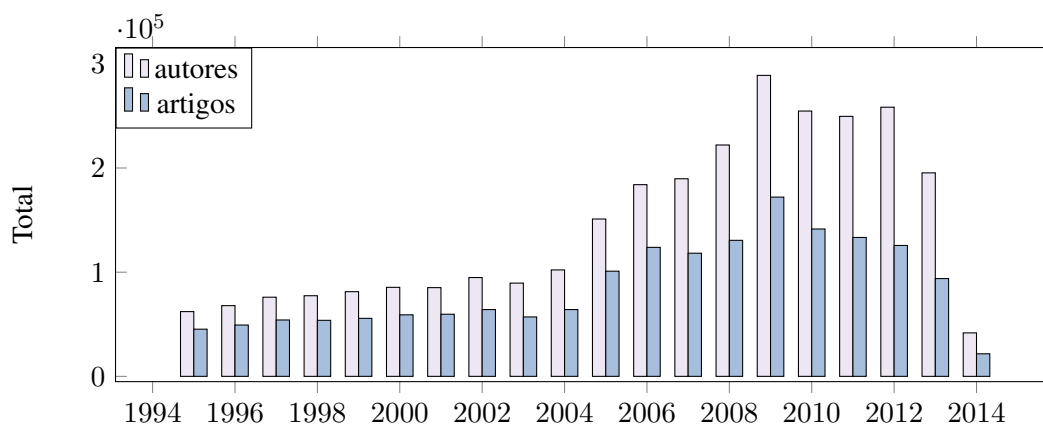


Figura 2: Distribuição do número de publicações e de autores, por ano, na base de dados da ArnetMiner. O recorte considerado compreende de 1995 a 2014 (20 anos).

são usados para dois propósitos neste trabalho: (i) quantificar o efeito do impacto científico a curto prazo da saída de pesquisa de um acadêmico em seu sucesso futuro, e (ii) quantificar e validar o impacto científico futuro de acadêmicos.

5. Resultados experimentais

Neste seção temos realizado dois tipos de experimentos no conjunto de dados da ArnetMiner. O primeiro trata da quantificação do efeito do impacto da saída de pesquisa de um acadêmico em seu sucesso futuro. Já o segundo experimento trata da validação do impacto científico futuro de um estrela de pesquisa⁷.

5.1. Quantificação do efeito do impacto da saída de pesquisa de um acadêmico em seu sucesso futuro

A hipótese H que consideramos é – *No ano da predição, jovens acadêmicos com totais de citações acima da média têm totais de citações futuras superiores à de jovens com totais de artigos publicados acima da média.*

Testou-se a hipótese H comparando as distribuições de totais de citações futuras nos conjuntos C e P. Em que C contém totais de citações futuras de jovens acadêmicos com totais de citações acima da média, no ano da predição, e P contém totais de citações futuras de jovens acadêmicos com totais de artigos publicados acima da média, no ano da predição.

A fim de comparar as distribuições de totais de citações futuras, aplicou-se um teste U de Mann-Whitney [Mann e Whitney, 1947]⁸. Testou-se a hipótese nula de que as distribuições de totais de citações futuras em C e P são as mesmas contra a hipótese alternativa de que os valores da distribuição em C são estocasticamente maiores do que em P.

Os p-valores do teste U de Mann-Whitney bem como as médias e variâncias de cada ano, nos dois conjuntos C e P são mostrados na Tabela 3. Para todos os anos considerados os p-valores estão bem abaixo do nível de significância de 0.01. Isso nos leva a seguramente rejeitar a hipótese nula, concluindo que os totais de citações futuras de pesquisadores em C são estocasticamente maiores do que em P.

⁷Rising stars [Panagopoulos et al., 2017], como são conhecidas as estrelas de pesquisa, são cientistas iniciantes com um número pequeno de artigos e de citações, mas que no futuro próximo, estes números elevados, do ponto de vista bibliométrico, os credenciarão a classe de cientistas de sucesso.

⁸O teste é usado para verificar se duas amostras independentes foram retiradas de populações com médias iguais.

Tabela 3: Os p-valores do teste U de *Mann-Whitney*. Também são mostradas as médias e variâncias dos totais de citações a longo prazo nos conjuntos C e P.

ano	p – valor	Média		Variância	
		C	P	C	P
1995	$3.796e^{-11}$	242.09	154.96	-	-
1996	$3.671e^{-10}$	234.09	169.97	65410.72	105982.6
1997	$3.298e^{-13}$	205.03	134.91	65778.47	80803.79
1998	$2.731e^{-14}$	184.73	122.18	69759.09	81496.24
1999	$1.756e^{-14}$	180.56	96.04	48788.62	25496.87

Esse resultado indica que o sucesso de citação a curto prazo é indicativo de sucesso de citação a longo prazo. Pode-se especular que o reconhecimento da comunidade científica, da importância dos trabalhos de pesquisa de um acadêmico, citando seus trabalhos, é um forte indício do potencial de impacto científico futuro desse acadêmico. Como indicam os resultados é preferível o uso de indicadores de qualidade derivados de citações em modelos de predição de acadêmicos de sucesso.

Neste contexto, Bornmann et al. [2014] têm argumentado que impacto a curto prazo é mais importante do que impacto a longo prazo uma vez que ele pode influenciar toda a trajetória de carreira de um acadêmico ou o destino de uma nova publicação. Como mostrado nesta seção nossos experimentos têm apresentado elementos que confirmam essa afirmação.

5.2. Validação do modelo usando indicadores de impacto científico a curto prazo

Neste segundo experimento, selecionamos 78230 acadêmicos que iniciaram suas carreiras no ano de 1998 ou 1999. Para cada um deles avaliamos seus impactos científicos no ano de 2001. Destes 78230, selecionamos 2444 acadêmicos com índices-h > 1, e resolvemos a seguinte instância do modelo:

$$\text{maximizar } \sum_{j=1}^{2444} \frac{h_j + c_j}{A_j} x_j \quad (6)$$

$$\text{sujeito à } \sum_{j=1}^{2444} \frac{1}{\sqrt[4]{A_j/2}} x_j \leq 40, \quad (7)$$

$$\sum_{j=1}^{2444} x_j \leq 49 \quad \text{e} \quad x_j \in \{0, 1\} \quad (8)$$

A função objetivo, Eq. 6, busca pelo subconjunto de acadêmicos iniciantes em 1998 ou 1999 mais relevantes no ano de 2001 (promissores de sucesso futuro). A Eq. 7 limita a soma máxima dos custos com acadêmicos em função de suas idades. A escolha da função $f : A_j \rightarrow \frac{1}{\sqrt[4]{A_j/2}}$ para o cálculo do custo de um acadêmico se deu devido à diferença entre os valores de f para dois acadêmicos ser menor do que o valor da diferença de suas idades. A restrição, Eq. 8, impõe que

no máximo 49 acadêmicos façam parte da solução. Já os valores de W e K foram configurados de forma arbitrária para 40 e 49 respectivamente. O modelo foi resolvido usando o pacote R, lpSolve⁹.

Neste trabalho, avaliamos o impacto científico futuro de acadêmicos contabilizando os seus totais de citações muitos anos depois do ano da predição, i.e., avaliando seus totais de citações futuras. A Tabela 4 mostra uma solução encontrada, na qual se destaca que todos os 49 acadêmicos desta solução apresentaram um crescimento expressivo em seus totais de citações futuras (2014) com relação ao momento da predição (2001), ao todo, considerou-se um período de 14 anos. O comprimento deste período deve-se a cobertura de dados disponível na base *ArnetMiner*.

Tabela 4: Acadêmicos com início de carreira em 1998 ou 1999 que foram preditos como acadêmicos de sucesso, e seus desempenhos individuais no momento da predição e futuro.

Acadêmico	1ª publicação	2001			2014		
		h	citações	publicações	h	citações	publicações
M. Aron	1998	5	66	9	7	333	9
J. Broch	1998	3	95	3	4	891	4
D. Brooks	1999	3	74	10	13	1104	84
M. Budiu	1998	4	66	9	11	606	46
R. Buyya	1999	3	46	19	14	642	232
S. E. Czerwinski	1999	3	57	6	5	621	8
Z. Dang	1999	4	43	14	7	174	61
E. Frank	1998	5	85	14	12	2219	57
D. Gibson	1998	5	125	7	10	721	14
A. Goy	1999	4	50	10	8	176	42
A. R. Greenwald	1999	4	40	12	12	352	46
A. Gupta	1999	4	59	21	13	507	28
A. Gujar	1998	4	81	9	6	322	9
G. He	1999	1	44	1	1	290	1
Y. Hu	1998	2	85	3	9	1170	44
J. Jetcheva	1998	2	86	3	4	742	6
M. Kawahito	1999	4	62	6	7	286	15
A. Keromytis	1998	5	70	17	15	706	166
A. Kleppe	1998	3	89	9	7	624	20
J. Knittel	1999	1	51	1	1	109	1
Y. Ko	1998	2	66	5	6	511	44
F. Kuhns	1999	5	40	8	6	124	14
A. LaMarca	1999	4	51	13	14	770	57
K. Leyton-Brown	1999	3	49	8	10	447	61
T. Lindholm	1999	1	287	1	1	825	2
X. Liu	1999	1	52	1	3	444	5
E. J. Marinissen	1998	6	140	15	16	754	73
O. Mitsuru	1998	1	83	1	1	297	1
S. Mungee	1998	3	64	4	3	152	4
K. Nigam	1998	6	123	12	12	1013	14
C. O’Ryan	1999	5	53	15	9	162	19
B. O’Sullivan	1999	1	46	1	1	151	1
L. Page	1998	2	152	2	2	1471	4
A. Pang	1998	2	77	2	2	275	2
G. Qu	1998	5	88	23	11	505	89
S. Rajagopalan	1998	2	66	2	2	320	2
R. Raskar	1998	4	110	18	15	920	156
E. M. Royer	1999	2	69	4	3	1009	4
J. Siméon	1998	4	97	9	11	509	26
T. Šimunić	1999	6	73	12	11	398	25
J. Shanmugasundaram	1998	5	92	10	13	1122	56
R. R. Taylor	1999	3	42	5	5	186	7
M. Takeuchi	1999	2	46	2	4	142	13
H. Zheng	1999	3	49	4	4	414	4
C. Zhang	1999	2	47	2	6	819	10
B. Y. Zhao	1999	4	71	10	14	1808	100
W. E. Walsh	1998	5	122	11	11	464	27
J. Whaley	1999	3	84	5	10	565	16
T. Yasue	1999	3	58	4	8	289	13

Finalmente, a fim de evidenciar o bom desempenho do modelo da mochila, comparamos os totais de citações em 2014 dos 49 acadêmicos promissores de sucesso preditos em 2001 pelo modelo da mochila e pelo modelo proposto por Zhou et al. [2018] que usa o algoritmo HITS¹⁰

⁹<http://CRAN.R-project.org/package=lpSolve>, último acesso em julho de 2019.

¹⁰O algoritmo HITS é um algoritmo de análise de links inicialmente proposto para classificar páginas web. Atualmente,

em redes de coautoria direcionadas a fim de identificar acadêmicos potenciais com curtas carreiras. Neste trabalho, por uma questão de clareza, chamaremos este último modelo de modelo HITS.

Na Figura 3 são apresentados três valores. O valor observado, em cinza claro, representa o total de acadêmicos classificados, segundo seus totais de citações futuras, por faixa observado na base de dados da *ArnetMiner*. O valor associado ao resultado após aplicação da mochila está em azul claro. Finalmente, o resultado da execução do modelo HITS está na cor azul escura. As citações foram agrupadas em cinco grupos, no primeiro grupo (citações entre 0 e 100) apenas acadêmicos preditos pelo modelo HITS foram classificados nesse grupo, para o segundo grupo (citações entre 100 e 500) temos acadêmicos tanto do modelo da mochila quanto do modelo HITS.

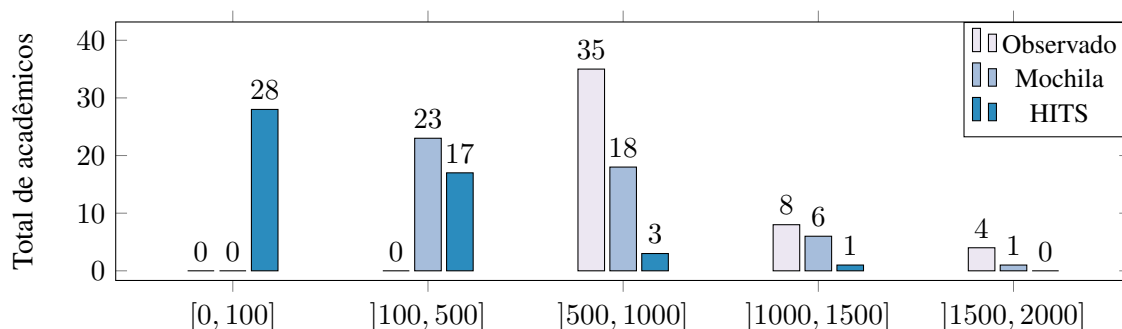


Figura 3: Distribuição das 49 estrelas de pesquisa preditas, na Tabela 4, e das 49 estrelas de pesquisa encontradas na base de dados da *Arnetminer*, por faixas de totais de citações futuras (em 2014).

Note que os resultados obtidos mostram que os 49 acadêmicos mais relevantes (no momento da predição), encontrados pelo modelo da mochila, estão entre os mais relevantes observados, no futuro. Eles têm uma média de totais de citações futuras bem superior ao outro modelo (HITS), em torno de 598 contra uma média de aproximadamente 160. Além disso, pode se observar que os 49 melhores acadêmicos de sucesso, de fato, na base de dados da *ArnetMiner* têm mais de 500 citações cada, enquanto existe 23 preditos pela modelo da mochila com no máximo 500 citações cada. Por um lado, isso indica que o modelo proposto tem margem para melhorar a sua eficácia seja considerando um tamanho de carreira maior ou incorporando outros fatores preditores de sucesso futuro como: a capacidade de um acadêmico de publicar em diferentes jornais [Acuna et al., 2012], a contribuição de cada coautor de publicações com muitos autores [Zhou et al., 2018], o efeito da colaboração de iniciantes com pesquisadores renomados [Qi et al., 2017], o prestígio da instituição formadora e a área de pesquisa.

Diferentemente do trabalho de Zhang et al. [2016] que considera somente acadêmicos ativos no momento da validação de impacto científico futuro, neste trabalho, consideramos acadêmicos que pararam de publicar ao longo de suas carreiras (como por exemplo, Larry Page o autor do algoritmo *Pagerank* [Brin e Page, 1998]), mas pelo menos um de seus artigos antigos continuam a receber citações até hoje, como acadêmicos de sucesso já que alguns de seus artigos foram e continuam sendo importantes para o avanço da ciência.

6. Conclusões

No contexto acadêmico, ter algumas informações sobre impacto ou relevância de um acadêmico, que retratem seu futuro provável é extremamente importante para tomadores de decisão quando se tratando de decisões de contratação de recursos humanos, de promoção e de alocação de recursos (apoio) para pesquisa.

essa medida é comumente utilizada para classificar publicações, cientistas e veículos de publicação. No nosso trabalho usamos para identificação precoce de acadêmicos de sucesso como sugerido por Zhou et al. [2018].

Neste trabalho, propomos um modelo para a predição de acadêmicos de sucesso baseado no arcabouço da mochila. Os experimentos sobre dados reais indicam que o nosso método pode encontrar uma parcela grande de futuros acadêmicos de sucesso com curtas carreiras científicas. Eles também indicam que o sucesso futuro de acadêmicos é guiado em uma parcela grande pelo reconhecimento pela comunidade científica, da importância dos seus trabalhos científicos. Ter seus trabalhos reconhecidos, no início da carreira, pela comunidade científica é um indicativo de que este acadêmico tende a continuar fazendo contribuições científicas importantes para a ciência no restante de sua carreira.

No entanto, apesar do bom resultado, nós mantemos a linha de raciocínio de Acuna et al. [2012], que a melhor forma de prever sucesso científico de um acadêmico é por pares. Contudo, os indicadores quantitativos de potencial futuro podem corrigir tendências enviesadas da avaliação por pares e facilitar a deliberação. Avaliadores não devem ceder à tentação de basear suas decisões apenas em números. Os indicadores não devem substituir o juízo informado como manda o princípio 1 do Manifesto de Leiden para métricas de pesquisa [Hicks et al., 2015].

E, apesar do nosso modelo identificar uma parcela grande dos acadêmicos de sucesso ainda no início de carreira, a pesquisa revelou o comportamento de arranque tardio¹¹ de uma parcela grande de acadêmicos que obtiveram sucesso científico futuro. Como continuação do trabalho, pretendemos aprofundar o entendimento sobre a trajetória de sucesso dessa classe de acadêmicos e incorporar esse conhecimento em nosso modelo a fim de melhorar nossas predições.

Por fim, é importante notar que no caso de instâncias grandes pode ser computacionalmente inviável encontrar a solução exata devido à característica NP-difícil do problema de programação linear inteira. No entanto, acreditamos que a relaxação do arcabouço da mochila inteira pode ser uma solução alternativa.

Apoio

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão - Brasil (FAPEMA) - Bolsa de Doutorado - # BD-08792/17.

Referências

- Abrishami, A. e Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, 13(2):485 – 499.
- Acuna, D. E., Allesina, S., e Kording, K. P. (2012). Predicting scientific success. *Nature*, 489(201):201–202.
- Bornmann, L., Leydesdorff, L., e Wang, J. (2014). How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics*, 8(1):175 – 180.
- Brin, S. e Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117.
- Bütün, E. e Kaya, M. (2019). Predicting citation count of scientists as a link prediction problem. *IEEE Transactions on Cybernetics*, p. 1–12.
- Caprara, A., Kellerer, H., Pferschy, U., e Pisinger, D. (2000). Approximation algorithms for knapsack problems with cardinality constraints. *European Journal of Operational Research*, 123(2):333 – 345.
- Ding, F., Liu, Y., Chen, X., e Chen, F. (2018). Rising star evaluation in heterogeneous social network. *IEEE Access*, 6:29436–29443.

¹¹O arranque tardio pode ser entendido como um fenômeno presente em algumas áreas mais teóricas como a de matemática ou de algumas áreas cujos resultados são difundidos ou comunicados após longos períodos de estudo.

- Dong, Y., Johnson, R. A., e Chawla, N. V. (2016). Can scientific impact be predicted? *IEEE Transactions on Big Data*, 2:18–30.
- Gogoglou, A. e Manolopoulos, Y. (2017). Predicting the evolution of scientific output. In Nguyen, N. T., Papadopoulos, G. A., Jkdrzejowicz, P., Trawinski, B., e Vossen, G., editors, *Computational Collective Intelligence*, p. 244–254, Cham. Springer International Publishing.
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., e Rafols, I. (2015). Bibliometrics: The leiden manifesto for research metrics. *Nature News*, 520(7548):429.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Science*, 102:16569–16572.
- Mann, H. B. e Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18(1):50–60.
- Nane, G. F., Larivière, V., e Costas, R. (2017). Predicting the age of researchers using bibliometric data. *Journal of Informetrics*, 11(3):713 – 729.
- Ning, Z., Liu, Y., e Kong, X. (2017). Social gene — a new method to find rising stars. In *2017 International Symposium on Networks, Computers and Communications (ISNCC)*, p. 1–6.
- Panagopoulos, G., Tsatsaronis, G., e Varlamis, I. (2017). Detecting rising stars in dynamic collaborative networks. *Journal of Informetrics*, 11(1):198 – 222.
- Penner, O., Petersen, A. M., Pan, R. K., e Fortunato, S. (2013). Commentary: The case for caution in predicting scientists' future impact. *Physics Today*, 66(4):8–9.
- Qi, M., Zeng, A., Li, M., Fan, Y., e Di, Z. (2017). Standing on the shoulders of giants: the effect of outstanding scientists on young collaborators' careers. *Scientometrics*, 111(3):1839–1850.
- Shi, G.-Y., Kong, Y.-X., Yuan, G.-H., Wu, R.-J., Zeng, A., e Medo, M. (2019). Discoverers in scientific citation data. *Journal of Informetrics*, 13(2):717 – 725.
- Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., e Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, p. 990–998, New York, NY, USA. ACM.
- Weihs, L. e Etzioni, O. (2017). Learning to predict citation-based impact measures. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, p. 1–10.
- Wijegunawardana, P., Mehrotra, K., e Mohan, C. (2016). Finding rising stars in heterogeneous social networks. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, p. 614–618.
- Zhang, J., Xia, F., Wang, W., Bai, X., Yu, S., Bekele, T. M., e Peng, Z. (2016). Cocarank: A collaboration caliber-based method for finding academic rising stars. In *Proceedings of the 25th International Conference Companion on World Wide Web*, p. 395–400.
- Zhou, J., Zeng, A., Fan, Y., e Di, Z. (2018). Identifying important scholars via directed scientific collaboration networks. *Scientometrics*, 114(3):1327–1343.