Instituto de Ciências Matemáticas e de Computação - USP

Relatório Científico

Predição de conexões em redes de colaboração científica

Número de Processo: 2020/11559-2

Período de Vigência: 01/04/2021 a 31/03/2022

Período Coberto pelo Relatório Científico: 01/04/2021 a 10/09/2021

Aluno: Rodrigo Bragato Piva

Orientador: Francisco Aparecido Rodrigues

Departamento de Matemática Aplicada e Estatística (SME)

Instituto de Ciências Matemáticas e Computação (ICMC)

Universidade de São Paulo (USP)

1 Resumo

Grande parte do desenvolvimento científico atual é feito através da colaboração entre cientistas e centros de pesquisa. Essa colaboração define uma rede complexa [1], onde cada cientista representa um vértice e dois cientistas são conectados se publicaram um artigo em conjunto [2]. Pesquisas anteriores mostraram que essas redes são sem escala, apresentam estrutura de comunidades e distância curta entre pesquisadores [3, 1].

A descrição da topologia tem auxiliado a entender como as colaborações emergem e influenciam no desenvolvimento científico [4].Uma questão fundamental que surge na análise dessas redes é a possibilidade de predizer novas colaboração e também entender as razões pelas quais dois pesquisadores iniciam um novo trabalho de pesquisa. Nesse projeto, abordaremos essas duas questões.

O nosso trabalho possui dois objetivos principais:

- **Predizer links em redes de colaboração científica**: vamos considerar métodos de aprendizado de máquinas para predizer os links. Nesse caso, o problema será mapeado como um problema de classificação binária, representando a presença ou ausência de conexões. Cada observação constitui uma aresta e os atributos considerados serão dados por medidas topológicas das arestas, como as descritas para predição de links em [5].

- **Identificar os atributos que favorecem a colaboração entre cientistas**: usando dados dos pesquisadores, tais como fator-h, área de atuação e número de citações, instituições de origem e país de origem dos pesquisadores e número de artigos publicados, vamos usar métodos de aprendizado de máquina, como o algoritmo florestas aleatórias, para determinar os atributos mais importantes para definir a colaboração entre cientistas.

2 Atividades desenvolvidas

2.1 Coleta de Dados do Pesquisador

A coleta dos dados foi feita em Python, usando-se a biblioteca Selenium (https://selenium-python.readthedocs.io/) com o webdriver do Opera. Para o *parse* dos elementos de HTML será usada a biblioteca BeatifulSoup (https://www.crummy.com/software/BeautifulSoup/).

O principal desafio para a coleta de dados é a identificação dos autores, pois é extremamente comum a abreviação de sobrenomes. Um exemplo é o coordenador deste projeto, que aparece em algumas publi-

cações como “Rodrigues, F.A.” ou “Rodrigues, Francisco A.”. Portanto, essa etapa exige que os autores sejam identificados corretamente em suas diferentes citações. Isto levou à escolha da plataforma Scopus, onde cada autor possui um ID, que define o link de sua página.

As figuras 1 e 2 são um exemplo da página de pesquisador do Scopus, e os dados presentes no site. É importante ressaltar que a maioria dos dados não é de fácil acesso, sendo necessário estar logado com uma conta vinculada ou acessar o site por uma rede universitária, como o eduroam, para ter acesso à todas as informações.

Figura 1: Informações presentes na página do autor.

Figura 2: Informações presentes nos detalhes do autor.

As informações disponíveis na plataforma Scopus que foram coletadas como dados do pesquisador são:

- O ID do autor, que está presente no link de sua página https://www.scopus.com/authid/detail.uri?authorId=10140494200
- Áreas de publicação do pesquisador

- Número de publicações
- Número de co-autores
- Universidade afiliada
- Número de citações
- Nome do autor
- Fator-h
- Cidade
- País

Para a coleta de dados, o navegador é aberto através da biblioteca Selenium e a página do autor é acessada. Então o código-fonte da página é obtido e enviado para a biblioteca BeatifulSoup que realisa o parse dos elementos de HTML. Para extrair os dados do HTML, foi necessário analisar manualmente o código fonte da página e encontrar os elementos que possuem os dados desejados.

Durante o desenvolvimento do projeto, mudanças ocorreram no site e o scrapper teve que ser alterado para encontrar os novos elementos que contém os dados. Dessa forma, é sempre necessário checar se os dados coletados estão corretos.

2.2 Coleta de dados de colaboração

A página do autor também possui uma lista de suas publicações, como na figura 3, e permite-nos coletar os dados de quais pesquisadores publicaram artigos com o autor sendo analisado e em quais artigos essa colaboração ocorreu. Além disso, os nomes dos autores são um link para suas respectivas páginas, o que nos permite coletar o ID dos co-autores e adicionar os não visitados à lista de próximos pesquisadores à serem analisados, expandindo a rede.

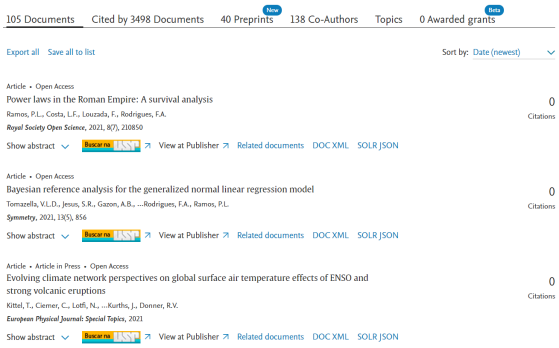

Figura 3: Informações sobre as publicações do autor.

O programa então percorre a lista de publicações conectando todos os pesquisadores que estão listados como autores do artigo. O site possui um máximo de 200 publicações por página, sendo necessário checar no código-fonte se há mais de uma página de publicações, e utilizar o Selenium para automaticamente alternar para as próximas páginas.

O código possui dois modos para conectar os autores:

- Múltiplas Arestas: Cada artigo é uma nova aresta entre os nodos (pesquisadores), com as propriedades de nome. Dessa forma geramos um multigrafo não dirigido.

- Arestas com Peso: O primeiro artigo entre dois autores cria uma aresta de peso 1, e cada publicação seguinte aumenta o peso em um. Dessa forma a aresta entre dois nodos tem como seu peso o número de artigos em que os dois pesquisadores trabalharam juntos. Neste modo geramos um grafo não dirigido.

Optou-se pelo segundo modo de operação, arestas com peso, pois as informações específicas das publicações não serão úteis no próximos passos de predição de link. Dessa forma o peso da conexão indica quão forte é a conexão entre dois pesquisadores.

2.3 Construção da Rede de Colaborações

Para criar o grafo, foi utilizada a biblioteca Networkx, com os dados coletados sendo armazenados como propriedades do nó. A rede final possui dois tipos de nós:

- Pesquisadores cujos dados foram coletados acessando a página do autor

- Pesquisadores encontrados na lista de co-autores, mas que não foram explorados pois o código chegou ao limite escolhido de autores. Estes nós apenas possuem seu ID e arestas criadas em outro autores.

Esta biblioteca facilita a criação de grafos e inserção de nós e arestas, permitindo diferentes propriedades, isto é útil pois quando um pesquisador que ainda não existe é encontrado na lista de co-autores, ele é criado e possui apenas seu ID e conexões com pesquisadores explorados.

A rede é então exportada no formato gml e este arquivo pode ser aberto pela biblioteca networkx, carregando o grafo para realizar análises. Foram geradas visualizações da rede utilizando o aplicativo Cytoscape (https://cytoscape.org), que pode ser vista na figura 4.

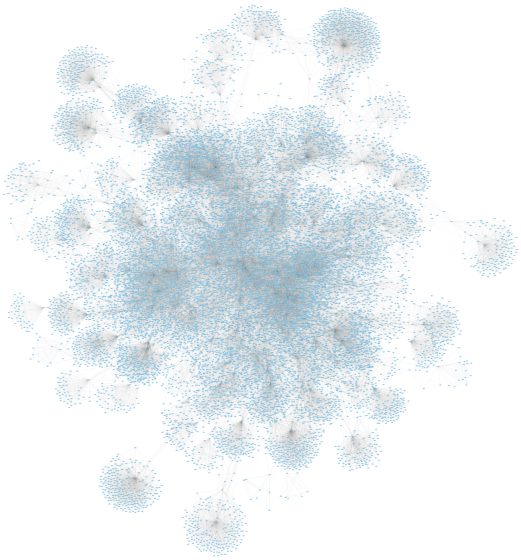

Figura 4: Visualização de uma rede colaborativa que obtivemos usando a metodologia descrita no texto.

Podemos ainda criar novas redes a partir dos dados de colaboração coletados, possibilitando análises além da predição de links. Dois exemplos disso foram:

- Uma rede de países, onde o peso de uma aresta entre dois países é o número de conexões de pesquisadores de um país com o outro (ver figura 5).

- Uma rede de países, onde o peso de uma aresta entre dois países é o número de conexões de pesquisadores de um país com o outro (ver figura 6).

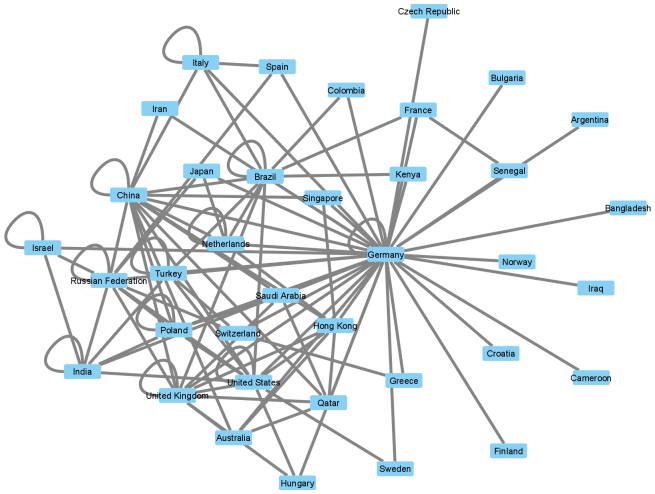

Figura 5: Rede colaborativa entre países.

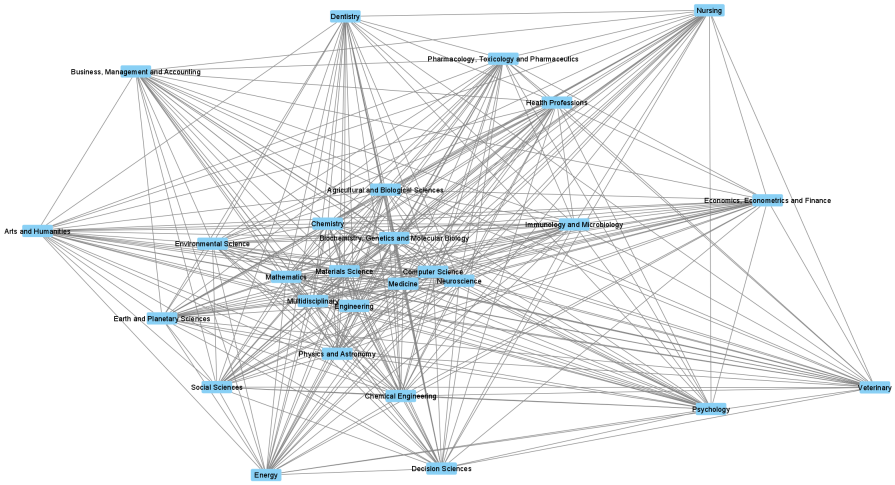

Figura 6: Rede entre áreas de pesquisa.

2.4 Predição de links

Para a predição de links, será criada uma rede neural em Python usando a biblioteca Keras (https://keras.io), tendo como entrada parâmetros de dois pesquisadores e a saída sendo uma predição se há uma colaboração entre os dois.

O primeiro passo é escolher os parâmetros de entrada e formatá-los de uma forma que a rede neural possa utilizar. As informações coletadas de cada pesquisador podem ser divididas em dados numéricos, informações que são um número, e dados categóricos, informações que indicam a que categoria o pesquisador pertence (Dos possíveis países, o pesquisador pertence à "Brasil"). Os dados que serão extraídos do grafo são:

- Dados Numéricos:
  - Número de publicações
  - Número de co-autores
  - Número de citações
  - Fator-h
- Dados Categóricos:
  - Áreas de publicação
  - Universidade afiliada
  - País
  - Cidade

Para a rede neural, é necessário que as entradas estejam entre 0 e 1, tornando necessária a normalização dos dados. A predição de link analisa dois pesquisadores, então para a normalização dos dados numéricos foi feita a razão dos valores.

Os dados categóricos foram transformados em dados binários, ou seja, se os pesquisadores são da mesma universidade o valor será 1, e se forem de universidades diferentes o valor será 0. Isto foi aplicado à universidade, país e cidade.

As áreas de publicação são uma lista, com cada pesquisador tendo um grupo de áreas na qual já publicou. Foi utilizado o índice de Jaccard para extrair um valor de similaridade normalizado dos dois conjuntos de dados. Sendo A e B os conjuntos de áreas, o índice é:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Um exemplo, com valores fictícios, do tratamento de dados:

	Autor A	Autor B	Resultado
Publicações	10	40	0.25
Co-autores	8	4	0.5
Citações	240	310	0.7742
Fator-h	4	6	0.6667
Universidade	Uni. X	Uni. Y	0
País	País X	País X	1
Cidade	Cid. X	Cid. Y	0
Áreas	{A,B,C,D,E,F,S}	{A,C,F,H,L,R,S}	0.4

Tabela 1: Exemplo de normalização

Além destes dados coletados pelo *web crawler*, também foi utilizado um valor calculado a partir do grafo, o coeficiente de Jaccard entre dois nós [6] que indica a semelhança entre os conjuntos de co-autores dos dois pesquisadores, quantos mais co-autores em comum, mais próximo de 1 o valor.

Foi criado um código que percorre os grafos gerados pelo *web crawler* analisando as conexões entre pesquisadores e armazenando os parâmetros em um arquivo no formato JSON. Após analisar todas as conexões existentes, o programa armazena dados de autores não conectados, obtendo o mesmo número de pares conectados e não conectados. O banco de dados obtido possui 294954 dados.

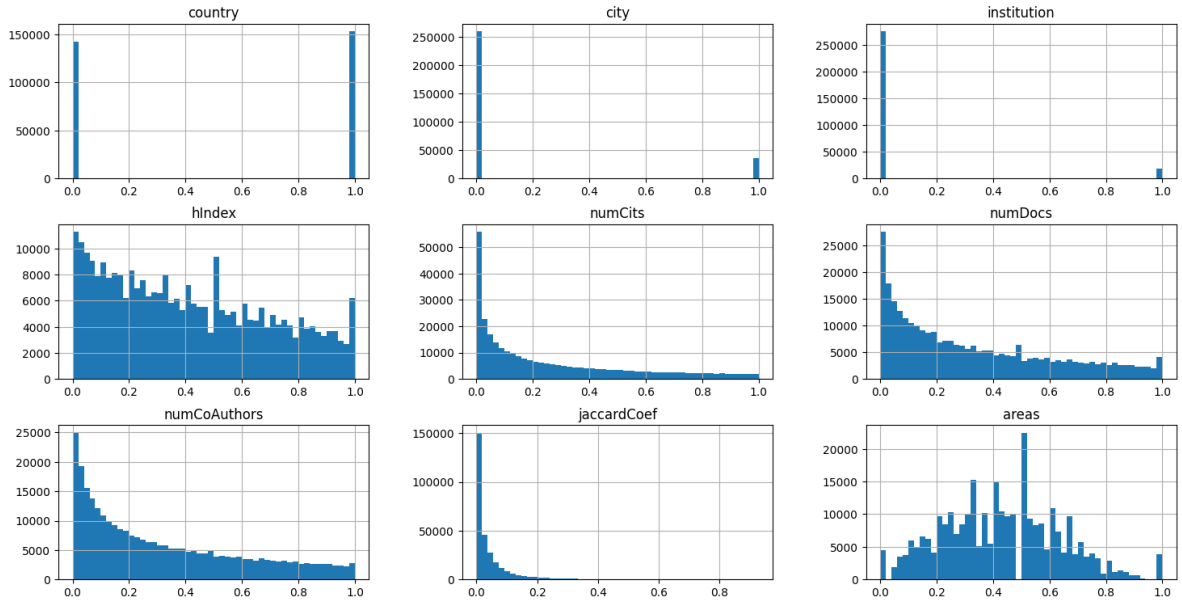

Figura 7: Histograma dos dados obtidos

Com os 9 parâmetros determinados, foi criada uma rede neural convolucional no seguinte formato:

Layer (type)	Output Shape	#Param
Dense	(None, 16)	160
Dense	(None, 24)	408
Dense	(None, 24)	600
Dense	(None, 16)	400
Dense	(None, 8)	136
Dense	(None, 4)	36
Dense	(None, 1)	5

Tabela 2: Resumo do modelo

Os banco foi separado em dados de teste e de treinamento. Após o treinamento a rede obteve uma acurácia de 0.86.

2.5 Importância dos atributos dos autores

Para avaliar a importância de cada parâmetro, foram criados 9 novos bancos de dados, cada um tendo um dado removido. A rede foi novamente treinada, tendo apenas o número de valores na entrada alterado, e a evolução do treinamento de cada um dos novos bancos foi registrada:

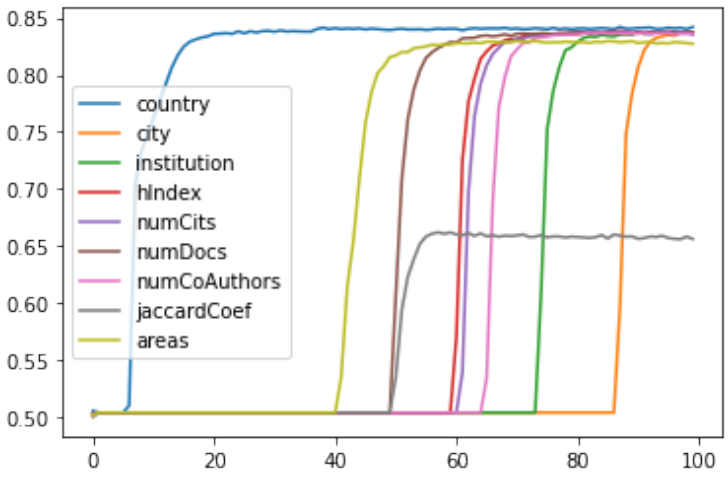

Figura 8: Acurácia das Redes

Podemos ver que o parâmetro mais importante é o coeficiente de Jaccard, pois a rede treinada sem ele obteve um desempenho significativamente pior, com as outras 8 tendo desempenhos parecidos.

Para visualizar melhor o impacto dos outros parâmetros, foram criados 8 novos bancos de dados, todos sem o coeficiente de Jaccard e com um dos outros parâmetros removidos.

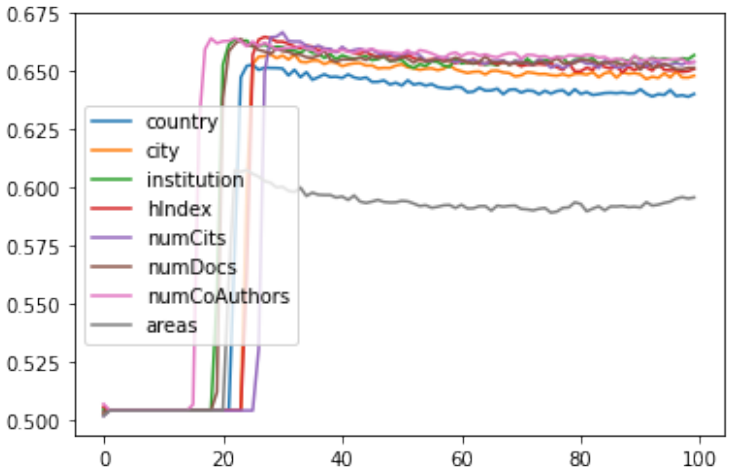

Figura 9: Acurácia das Redes

Novamente, vemos que um parâmetro afeta significativamente o desempenho da rede, neste caso as áreas de atuação do pesquisador. Além deste, o país dos pesquisadores também é um parâmetro que influencia na acurácia da rede, porém com um impacto menor.

3 Conclusão

Com os resultados obtidos pela rede neural, vemos que os parâmetros utilizados são suficientes para se obter uma certeza de 86% na predição de links e que os dois parâmetros mais importantes para a predição são as áreas e o coeficiente de Jaccard.

O coeficiente de Jaccard . Já o parâmetro de áreas indica a semelhança entre as áreas de publicação dos autores. Isto mostra que, para dois autores escrevem um artigo juntos, os principais fatores são o número de pessoas em comum com a qual já trabalharam e quão parecidas são suas áreas de atuação.

Referências

[1] Mark Newman. *Networks: an introduction*. Oxford University Press, Inc., 2010.

[2] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[3] Mark EJ Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101(suppl 1):5200–5205, 2004.

[4] Vicente P Guerrero Bote, Carlos Olmeda-Gómez, and Félix de Moya-Anegón. Quantifying the benefits of international scientific collaboration. *Journal of the American Society for Information Science and Technology*, 64(2):392–404, 2013.

[5] Víctor Martínez, Fernando Berzal, and Juan-Carlos Cubero. A survey of link prediction in complex networks. *ACM Computing Surveys (CSUR)*, 49(4):1–33, 2016.

[6] J Kleinberg D Liben-Nowell. The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.