SCYLAX
Documentação de Entrega de Funcionalidades

Anne Magály de P. Canuto [1], João C. Xavier-Júnior [2] Arthur Costa Gorgônio [1], Cephas A S Barreto [1], Mateus Firmino Barros [2], Song J M S Costa [2]

[1] Departamento de Informática e Matemática Aplicada (DIMAp)
Universidade Federal do Rio Grande do Norte - Natal, RN - Brasil

[2]Instituto Metrópole Digital (IMD)
Universidade Federal do Rio Grande do Norte - Natal, RN - Brasil

1. Introdução

A pesquisa brasileira tem boa parte de seus dados armazenados na plataforma Lattes (`http://lattes.cnpq.br/`), sistema em que pesquisadores podem armazenar informações sobre seus trabalhos científicos, artísticos e profissionais.

Tendo em vista a grande quantidade de dados armazenados na plataforma Lattes, foi iniciada uma pesquisa que tem por objetivo utilizar as informações contidas na plataforma Lattes para entender o estado atual e também o desenvolvimento da pesquisa brasileira. O principal resultado dessa pesquisa é a plataforma Scylax, um sistema que consome dados da plataforma Lattes, e fornece informações relevantes para avaliação das pesquisas realizadas no Brasil.

Apesar de já estar em funcionamento, há uma série de desafios relacionados à proposta do Scylax. Esses desafios vão desde a grande quantidade de dados de dados envolvidos em cada análise ou até a dificuldade de entender como realizar análises mais complexas, principalmente as que necessitam de comparações. Nesse sentido, o grupo de pesquisa com Aprendizado de Máquina da UFRN, composto por pesquisadores do Instituto Metrópole Digital (IMD) e do Departamento de Informática e matemática Aplicada (DIMAp), conduziram pesquisa que objetivou a construção de funcionalidades para o Scylax com uso de técnicas de aprendizado de Máquina.

Este documento apresenta, portanto, as funcionalidades construídas pelo grupo mencionado para o Scylax, que são: análise de pesquisadores com base em suas pesquisas; análise da produção científica de programas de pós-graduação; e análise da produção científica de redes de colaboração. Além da descrição geral de cada funcionalidade, também são detalhados os dados utilizados para sua construção, as informações mostradas, as capturas de tela do protótipo de sistema e observações pertinentes.

2. Análise de Produção Científica de Pesquisadores

2.1. Descrição geral

A análise de produção científica de pesquisador compreende um grupo de informações recuperadas sobre os dados de toda a pesquisa de um dado pesquisador. Para que fosse possível analisar cada pesquisador de forma comparativa, foram utilizadas técnicas não-supervisionadas de Aprendizado de Máquina que, ao final, retornam os dados de forma agrupada. Dessa forma, é possível localizar o pesquisador dentro de um grupo e, ao observar as características gerais do grupo, perceber como está colocado o pesquisador em questão.

2.2. Dados utilizados

Para construção da funcionalidade, foram utilizados os seguintes dados:

- h-index;
- número de colaborações internacionais;
- número de colaboradores internacionais;
- quantidade de publicações em para cada qualis, separadas por década;
- grupo gerado pela técnica de AM.

2.3. Funcionalidade exibida no sistema

A funcionalidade de análise de produção científica de pesquisadores foi dividida em duas partes: na primeira são exibidos recursos relacionados aos grupos e suas métricas gerais; já na segunda é possível escolher um pesquisador e observar algumas de suas métricas.

Para utilizar a funcionalidade, é necessário que o usuário escolha em quantos grupos quer dividir os dados. Essa personalização utiliza como grupos as saídas das técnicas de AM já mencionadas e possibilita a análise dos dados de forma mais ou menos segregada, de acordo com a necessidade do usuário. Após a escolha do número de grupos, são exibidas informações gerais sobre as publicações de cada grupo, inclusive de forma temporal. Na segunda parte, como já mencionado, é possível digitar o nome de um pesquisador e visualizar informações de sua pesquisa, também incluindo uma avaliação gráfica que considera o tempo.

2.3.1. Análise dos grupos - informações exibidas

a. número de publicações por década para cada grupo;
b. h-index médio para cada grupo;
c. número médio de colaborações internacionais para cada grupo;
d. média de pesquisadores em colaborações internacionais para cada grupo;
e. número de publicações por qualis em cada década ([1] 1981 a 1990, [2] 1991 a 2000, [3] 2001 a 2010, e [4] 2011 a 2020) por grupo.

2.3.2. Análise dos grupos - capturas de tela

Figura 1 – Gráfico do número de publicações por década para cada grupo

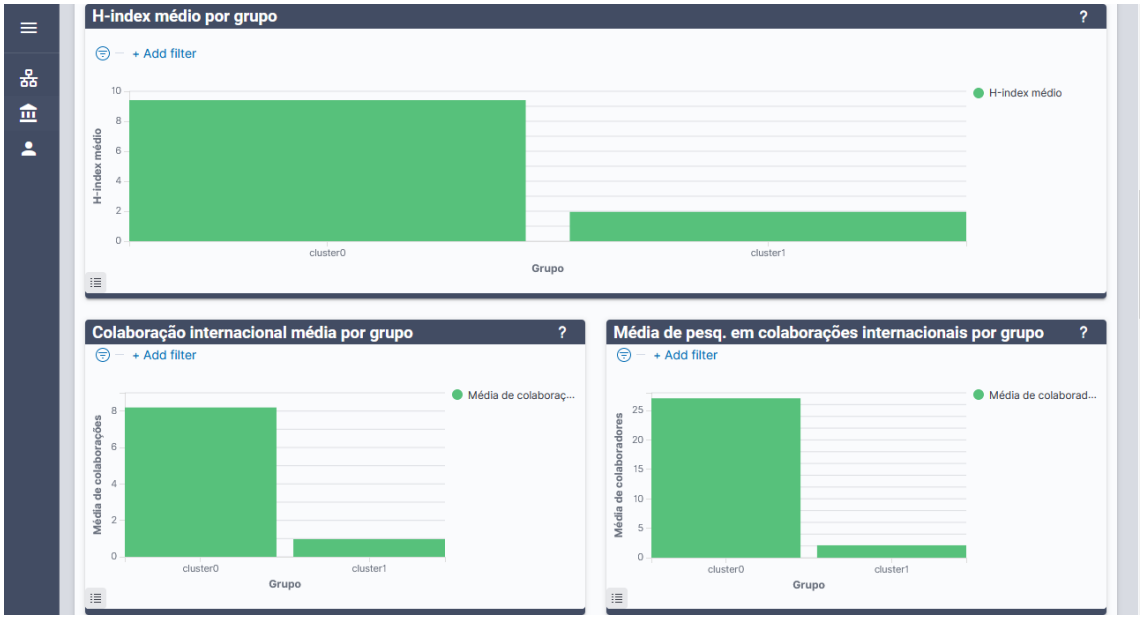

Figura 2 – Gráficos do h-index médio, colaborações internacionais médias e média de pesquisadores em colaborações internacionais por grupo

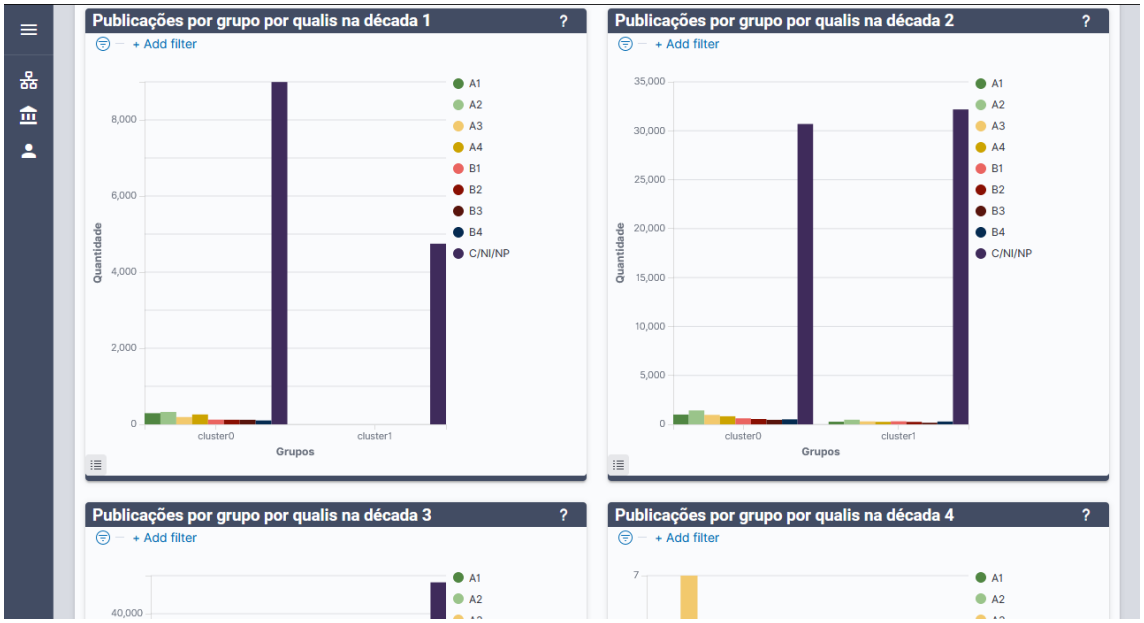

Figura 3 – Publicações por década e qualis (décadas [1] e [2])

Figura 4 – Publicações por década e qualis (décadas [3] e [4])

2.3.3. Análise do pesquisador - informações exibidas

a. h-index do pesquisador;
b. número de publicações por década para o pesquisador;
c. número de colaborações internacionais do pesquisador;
d. número de colaboradores internacionais que publicaram com o pesquisador;
e. número de publicações do pesquisador, separadas por qualis, em cada década ([1] 1981 a 1990, [2] 1991 a 2000, [3] 2001 a 2010, e [4] 2011 a 2020).

2.3.4. Análise do pesquisador - capturas de tela

Figura 5 – Publicações por década de um pesquisador

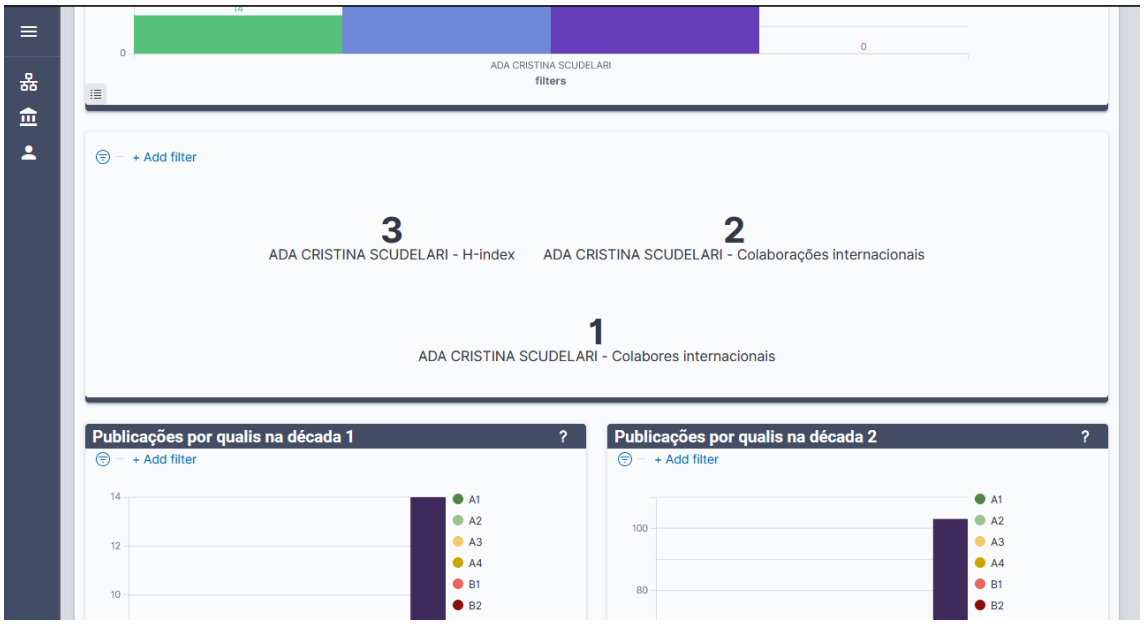

Figura 6 – H-index, colaborações internacionais do pesquisador e colaboradores internacionais nos trabalhos do pesquisador

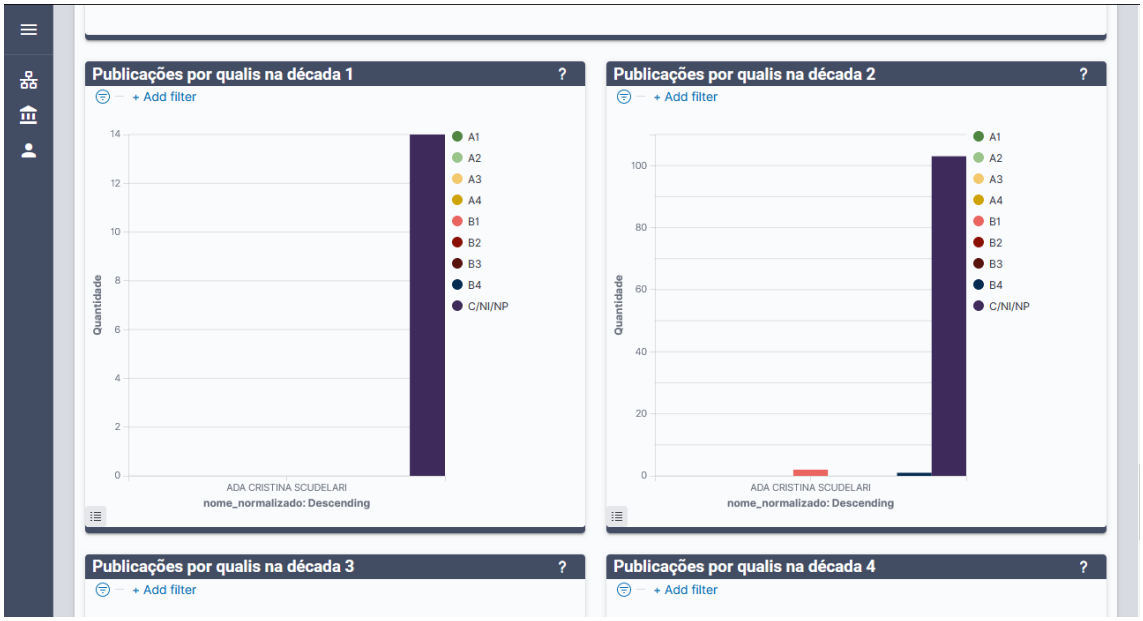

Figura 7 – Total de publicações do pesquisador por qualis (décadas [1] e [2])

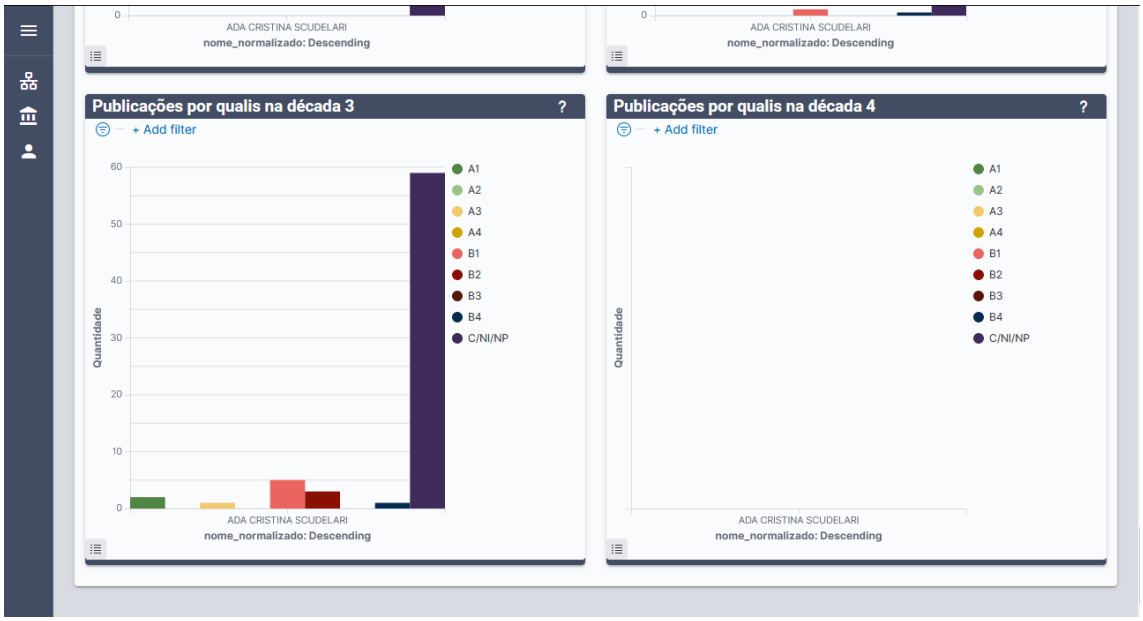

Figura 8 – Total de publicações do pesquisador por qualis (décadas [3] e [4])

3. Análise de Produção Científica de Programas de Pós-graduação

3.1. Descrição geral

De forma análoga à produção científica de um pesquisador, a produção de um programa de pós-graduação é composta por todas as publicações vinculadas a autores que pertencem a certo programa de pós-graduação. A base de dados utilizada para esta análise foi construída utilizando os dados de cada pesquisa científica, seus autores, métricas relacionadas e os programas aos quais a pesquisa em questão está associada. Sobre esses dados

foram executadas técnicas de agrupamento, e então, utilizados os grupos gerados como fonte de informação para a construção dos gráficos.

3.2. Dados utilizados

Para construção da funcionalidade, foram utilizados os seguintes dados referentes a cada produção:

- quantidade de autores;
- lista de autores;
- ano da produção;
- tipo da produção;
- citescore;
- quartil citescore;
- percentil;
- quartil percentil;
- qualis;
- quantidade de programas de pós-graduação envolvidos;
- lista de programas de pós-graduação envolvidos.

3.3. Funcionalidade exibida no sistema

Esta funcionalidade foi desenvolvida com o objetivo de analisar as produções científicas de um programa de pós-graduação. De forma similar à funcionalidade de análise de produção científica de pesquisadores, é necessário selecionar a "quantidade de grupos" (divisões) pela qual serão divididos os dados. Com o número de grupos escolhido, há a análise geral dos índices de cada grupo e também a possibilidade de comparar um programa alvo com outro(s) programa(s).

Ao gerar os gráficos, também há a divisão em duas partes: na primeira, é mostrada a análise dos grupos como um todo; já na segunda, são exibidas as informações da comparação entre um programa alvo e outros programas, todos escolhidos pelo usuário. A seguir, são apresentadas as métricas utilizadas para cada parte da funcionalidade bem como as capturas de tela do protótipo de sistema desenvolvido.

3.3.1. Análise dos grupos - informações exibidas

a. quantidade de programas por grupo;
b. quantidade de publicações por ano, em cada grupo;
c. citescore;
d. quartil citescore;
e. percentil médio;

f. quartil percentil médio;
g. quantidade de publicações separados por qualis;
h. quantidade de publicações separados pelo ranking.

3.3.2. Análise dos grupos - capturas de tela

Figura 9 – Seleção da quantidade de grupos e gráfico quantidade de produções por grupo

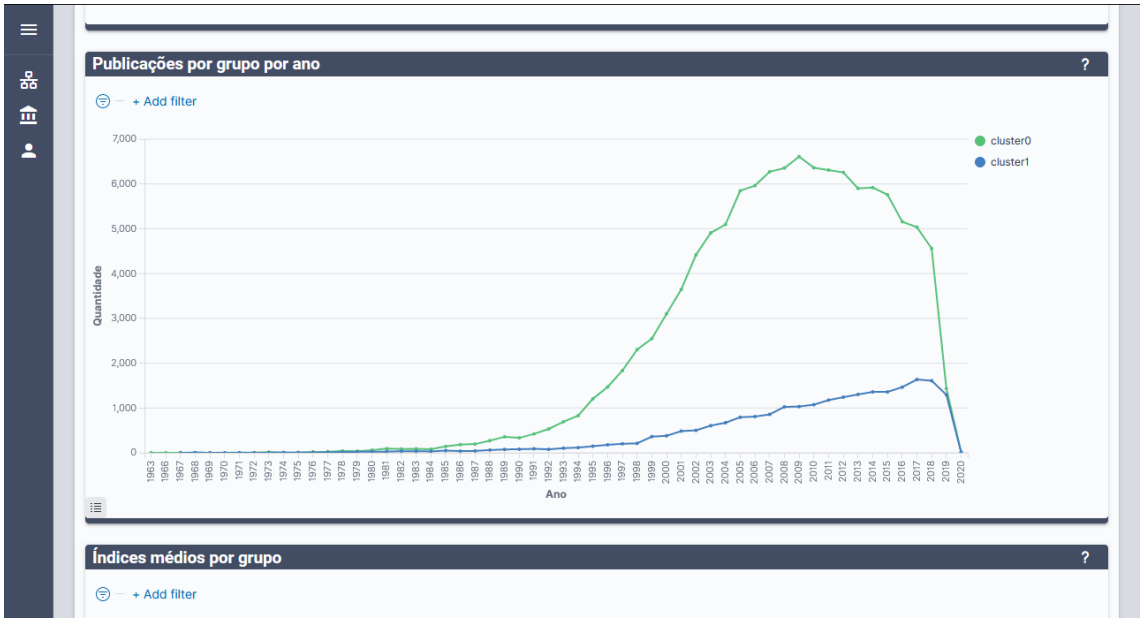

Figura 10 – Gráfico da quantidade de publicações para cada um dos grupos por ano

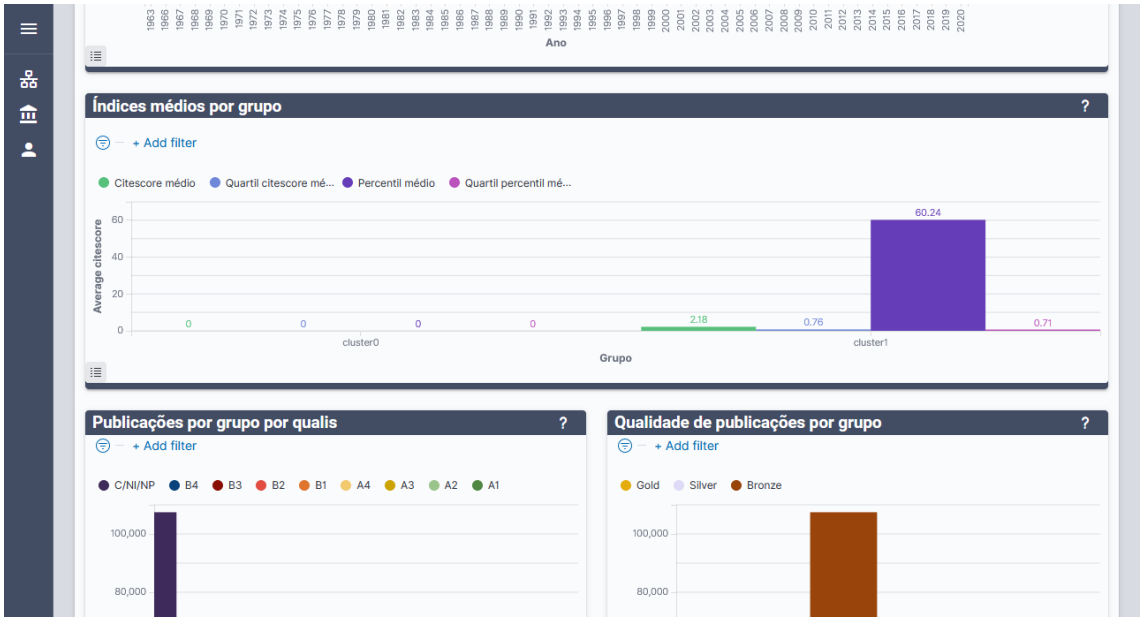

Figura 11 – Gráfico dos índices citescore, quartil citescore, percentil médio, quartil percentil médio por grupo

Figura 12 – Gráficos da quantidade das publicações por qualis e por ranking (ouro, prata, bronze)

3.3.3. Análise do programa de pós-graduação frente a outros - informações exibidas

a. citescore médio;
b. quartil citescore médio;
c. percentil médio;
d. quartil percentil médio;

e. quantidade de publicações por qualis;
f. quantidade de publicações por ranking;
g. distribuição das publicações dos programas em relação ao qualis.

3.3.4. Análise do programa de pós-graduação frente a outros - capturas de tela

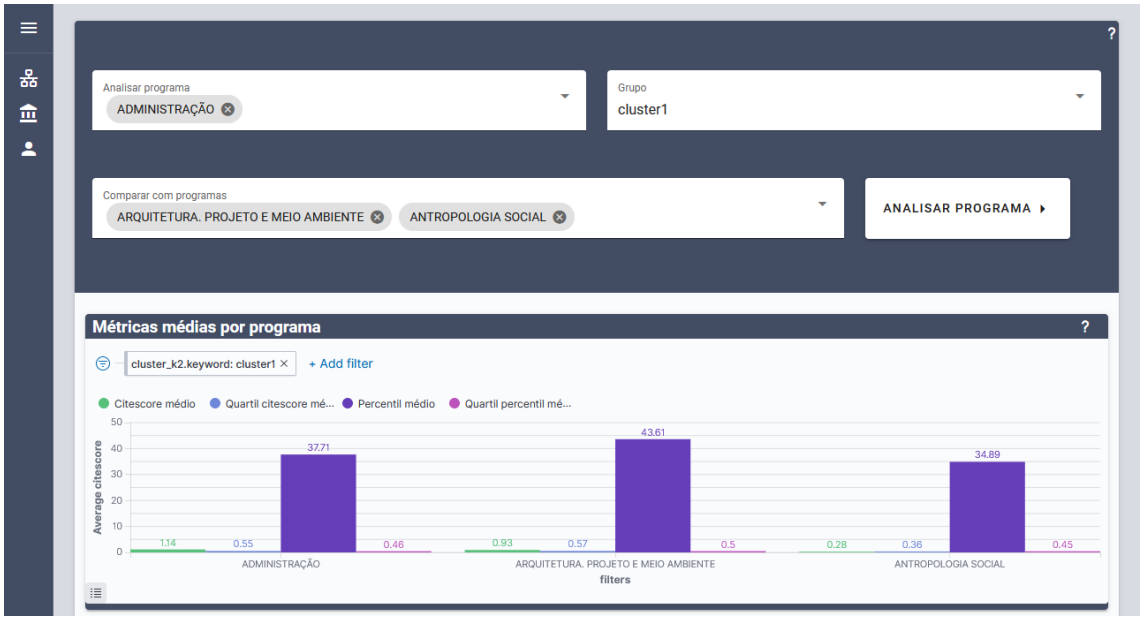

Figura 13 – Parte da funcionalidade com a seleção dos programas a serem analisados e gráficos com os índices dos programas escolhidos

Figura 14 – Quantidade de produções por qualis e de produções por ranking (ouro, prata e bronze) dos programas selecionados

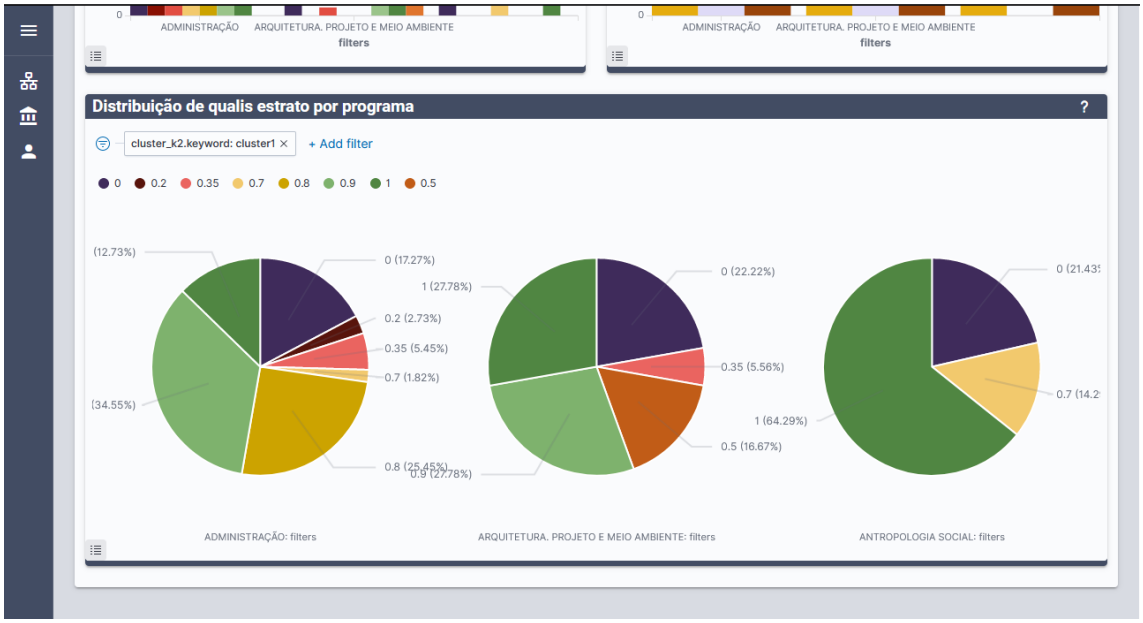

Figura 15 – Gráfico de pizza de cada programa referente a distribuição do qualis das publicações

4. Análise de Redes de Colaboração

4.1. Descrição geral

Entende-se, para fins da pesquisa realizada, que uma rede de colaboração é dada por um grupo de autores que possui pelo menos uma publicação em comum. Além desse conceito, também é utilizado o conceito de sub-rede, que compreende qualquer subconjunto entre os autores que compõem uma rede de colaboração.

Esta funcionalidade objetiva analisar qualitativamente uma rede de colaboração e também identificar seus pontos mais fortes e menos fortes. Isso é feito através dos índices dos autores que compõem uma rede e também através dos índices das publicações produzidas pela rede.

No sistema, a funcionalidade foi dividida em três partes: resumo - produção; resumo - colaboração; e detalhes. Na primeira, resumo - produção, são exibidos os dados gerais da rede como um todo. Na segunda, resumo - colaboração, são exibidas informações relacionadas à colaborações internacionais com participação da rede. Na terceira e última parte, detalhes, são detalhadas informações relativas às sub-redes possíveis para a rede de colaboração em questão.

4.2. Dados utilizados

Para a construção da funcionalidade, foi criada um *dataset* com foco nos pesquisadores e com os seguintes dados:

- o pesquisador é docente?;
- tipo da produção;
- subtipo da produção;
- número de produções do pesquisador;
- lista de produções do pesquisador;
- código da área do conhecimento;
- nome do programa de pós-graduação do pesquisador;
- h-index do pesquisador.

Com base na identificação dos autores que compõem uma certa rede, o *dataset* mencionado serve de primeira fonte para consulta e construção de uma base de dados dinâmica, referente à rede, e que contem os seguintes dados:

- lista de trabalhos realizados pela rede;
- índices dos autores;
- índices das produções da rede;
- mesmos dados para todas as sub-redes.

4.3. Funcionalidade exibida no sistema

Nessa funcionalidade, é necessário que o usuário informe quais pesquisadores serão analisados como uma rede de colaboração. Após indicar os pesquisadores, os trabalhos científicos em comum são recuperados dos dados e então são apresentados os gráficos.

Como já citado, as três partes da funcionalidade foram pensadas para fornecer um resumo da produção feita realizada pela rede (resumo - produção), dar uma visão geral sobre as colaborações internacionais realizadas pela rede (resumo - colaboração) e fornecer detalhes da rede, inclusive mostrando aspectos relacionados às sub-redes (detalhes).

A seguir serão apresentadas, de acordo com a divisão da funcionalidade, as informações mostradas e as capturas de tela do sistema protótipo desenvolvido.

4.3.1. Análise da rede de colaboração (resumo - produção) - informações exibidas

a. total de publicações da rede;
b. intervalo de atividade;
c. citescore médio dos trabalhos produzidos pela rede;
d. número de publicações por tipo;
e. número de publicações por qualis.

4.3.2. Análise da rede de colaboração (resumo - produção) - capturas de tela

Figura 16 – Início da funcionalidade - seleção dos pesquisadores que irão compor a rede

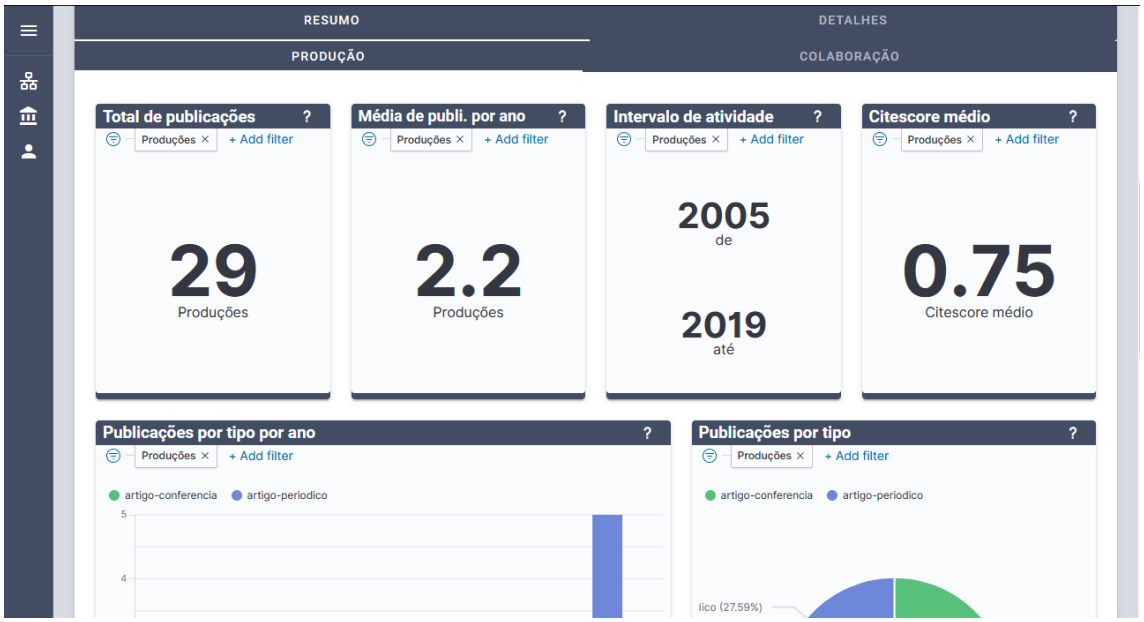

Figura 17 – Métricas sumarizadas da rede de colaboração

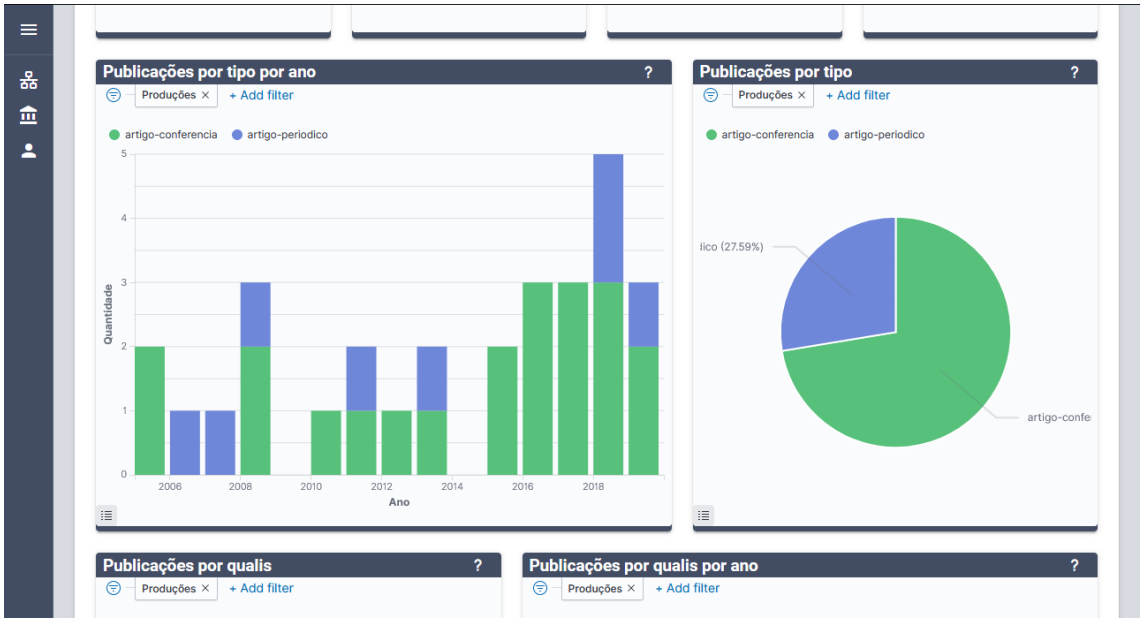

Figura 18 – Gráficos do quantidade de publicações por tipo no tempo e quantidade por tipo no geral

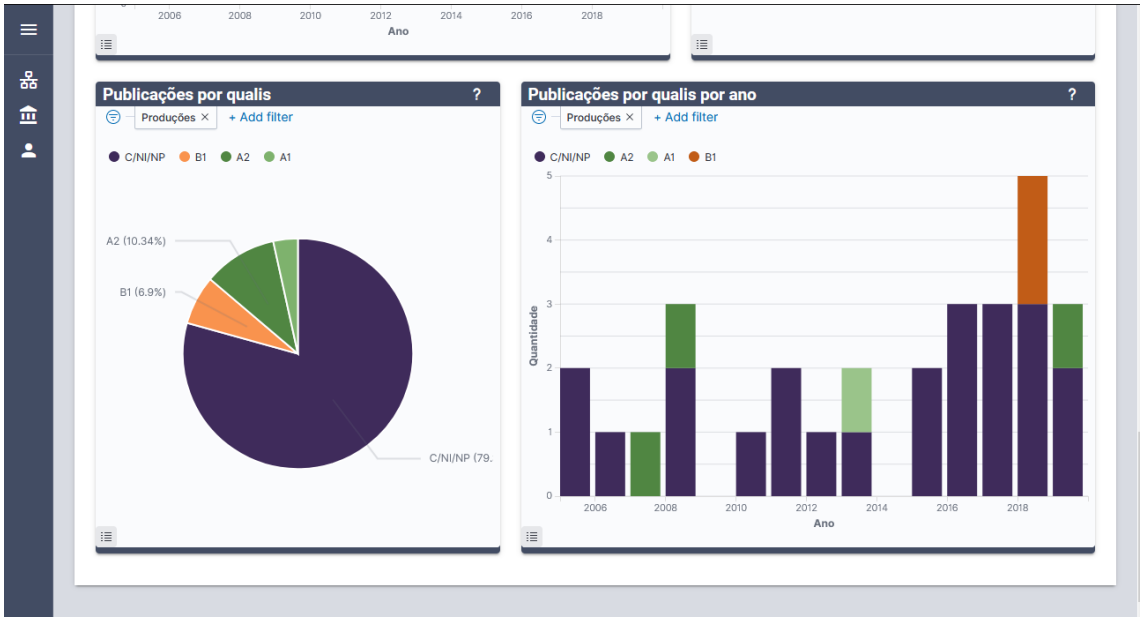

Figura 19 – Gráfico da publicação por qualis, geral e por período

4.3.3. Análise da rede de colaboração (resumo - colaboração) - informações exibidas

- número de colaborações internacionais para cada pesquisador;
- média de pesquisadores em colaborações internacionais para cada pesquisador.

4.3.4. Análise da rede de colaboração (resumo - colaboração) - capturas de tela

Figura 20 – Gráfico das colaborações internacionais dos pesquisadores da rede

Figura 21 – Gráfico dos pesquisadores em publicações internacionais da rede

4.3.5. Análise da rede de colaboração (detalhes) - informações exibidas

As informações exibidas nessa parte são repetidas para cada sub-rede. Por exemplo, se uma rede é composta pelos pesquisadores *A, B* e *C*, esta parte da funcionalidade irá exibir as informações para as sub-redes *A*-*B*; *A*-*C*; *B*-*C*; e *A*-*B*-*C*.

1. número de publicações por qualis;
2. número total de publicações;
3. citescore das publicações.

4.3.6. Análise da rede de colaboração (detalhes) - capturas de tela

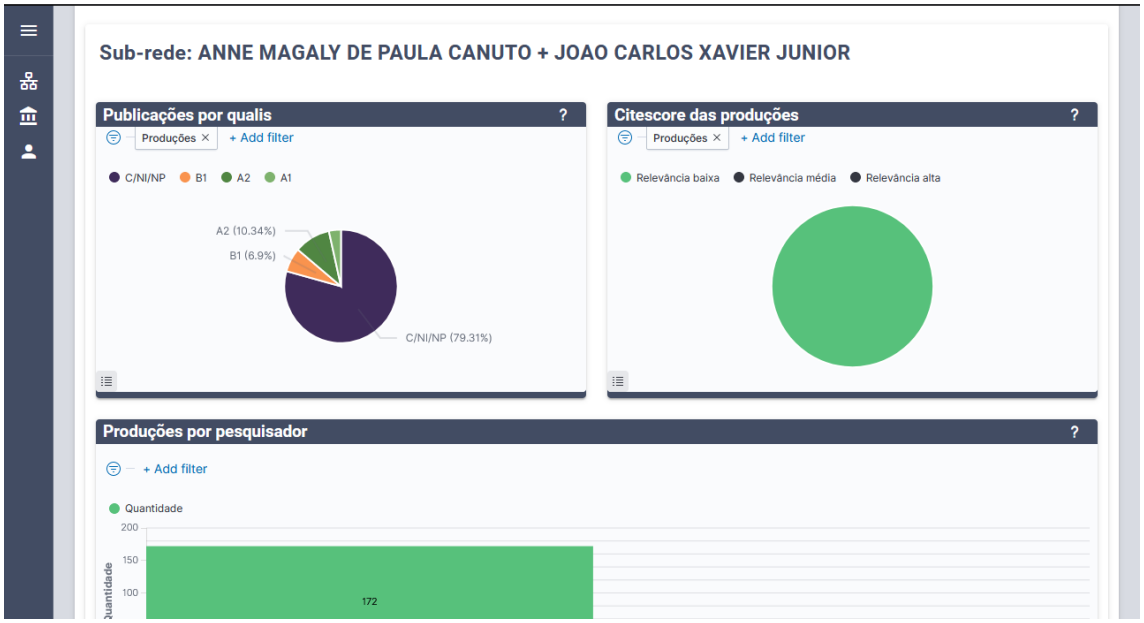

Figura 22 – gráficos das publicações por qualis da sub-rede analisada

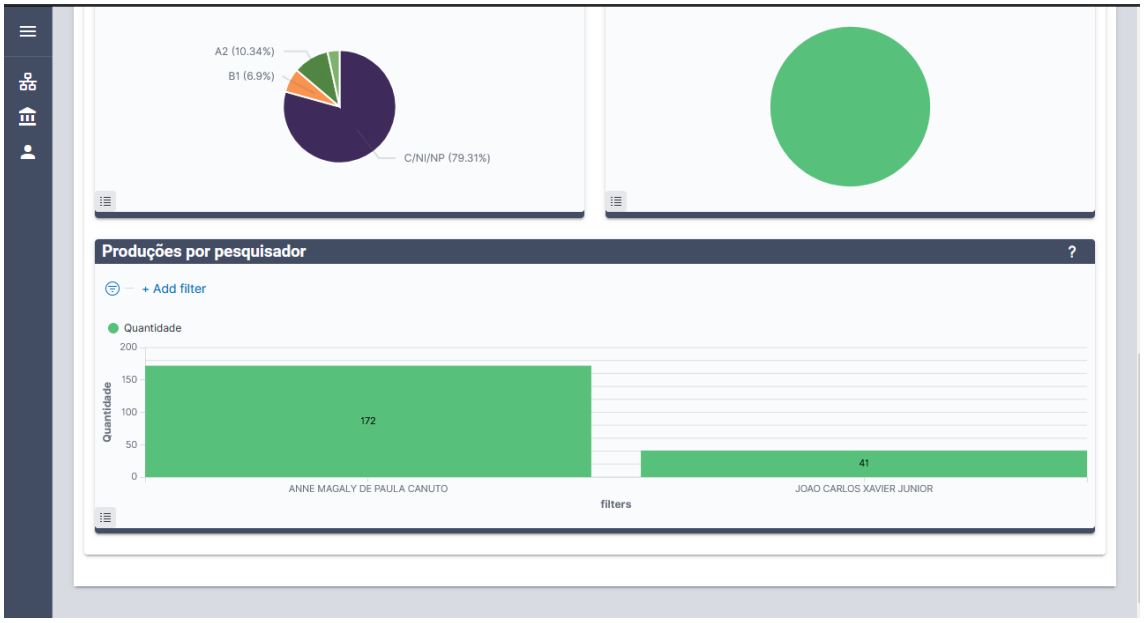

Figura 23 – gráficos das publicações por qualis da sub-rede analisada

5. Considerações Finais

O presente relatório apresentou o protótipo de sistema desenvolvido para demonstrar o funcionamento das seguintes funcionalidades: análise de pesquisadores com base em

suas pesquisas; análise da produção científica de programas de pós-graduação; e análise da produção científica de redes de colaboração. Além da descrição geral de cada funcionalidade, foram citados os dados usados para construção, informações que são exibidas e capturas de tela para cada funcionalidade, todos seguindo o fluxo visual presente no sistema desenvolvido.

De maneira geral, é possível afirmar que tratou-se de uma pesquisa desafiadora, dado que a análise de produção científica sempre está rodeada por questões relacionadas à inviabilidade de comparar possíveis entes de contextos diferentes, questões relacionadas ao tempo em que foi feita a pesquisa e etc. Contudo, a utilização de técnicas não-supervisionadas de Aprendizado de Máquina possibilitou a segregação adequada dos dados e, como consequência, a observação dos entes individuais frente aos grupos gerados, seja para a funcionalidade de análise de pesquisadores ou mesmo para a análise de programas de pós-graduação.

Em relação à ultima funcionalidade apresentada neste relatório, a análise de redes de colaboração, é necessário mencionar a dificuldade de tal análise devido à complexidade computacional presente no cruzamento de dados de produções de pesquisadores em um universo de busca tão grande. Essa dificuldade foi mitigada com a utilização de dados intermediários construídos sob demanda e mantidos a cada consulta.

Foram utilizadas as seguintes tecnologias em toda a pesquisa:

- **análise e exploração de dados**: Weka-API, Weka-GUI, Scikit-learn, Pandas, Numpy;
- **back-end, consultas e gráficos**: Elastic Search, Kibana;
- **front-end**: Vue-js.

É importante também citar que, mesmo no sistema protótipo, levamos em consideração a dificuldade de entendimento do funcionamento de cada funcionalidade e também das possibilidades de personalização dos gráficos em cada tela. Por esse motivo foram adicionados campos de ajuda, com explicações detalhadas sobre cada funcionalidade e também sobre os gráficos.

Por fim, todas as bases, códigos, este relatório e também o vídeo de apresentação da plataforma em utilização estão presentes no repositório `https://github.com/ml-imd/Sclx_research`. Nele os diversos artefados estão separados por pastas.