Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Marilsa de Assis

Buscador de Publicações e Colaborações Científicas

Curitiba
2019

Marilsa de Assis

Buscador de Publicações e Colaborações Científicas

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Marco Antonio Zanata Alves

Curitiba
2019

Buscador de Publicações e Colaborações Científicas

Marilsa de Assis

Universidade Federal do Paraná, Curitiba PR, Brasil

Resumo

Atualmente há diversas bases de dados de bibliotecas digitas e motores de busca de publicações científicas, o que facilita a divulgação e democratiza o acesso a esse conteúdo. Com o crescimento do incentivo à ciência em diversos países, as publicações científicas tem aumentado exponencialmente, o que tem gerado um grande volume de dados. Apesar das diversas opções para busca de artigos científicos, não há opções para a visualização de colaboração entre autores. Além disso, o cruzamento de informações geográficas para localizar eventuais colaboradores deve ser feita de maneira manual. Visando atender a facilitar a busca e análise dos dados das publicações de artigos científicos, nesse trabalho foi desenvolvido uma plataforma *on-line* que possibilita a busca por artigos, e permite a visualização da rede de colaboração dos autores e coautores em formato geográfico. A base dados utilizada pelo *website* foi a *Digital Bibliography Library Project (DBLP)*, que contêm artigos na área de ciência da computação, o mesmo foi desenvolvido utilizando a linguagem de programação *Python* 3.7 e o *Framework* para *Web Flask* e banco de dados *PostegreSQL*.
Palavras-chave: Rede de Colaboração, Web Scraping, DBLP, Bibliotecas Digitas

Abstract

Nowadays there are several digital libraries databases and scientific publications search engines, which facilitates the dissemination and democratizes access to this content. With the growth of science in many countries, scientific publications have increased exponentially, which has generated a great deal of data. Despite the many options for searching for scientific articles, there are no options for viewing collaboration between authors. In addition, cross-referencing geographic information to locate potential collaborators must be done manually. Aiming to facilitate the search and analysis of scientific articles in this work we built an online platform to enable the search for articles allowing the visualization of the authors and coauthors collaboration network using geographical visualization. The database used by our website was DBLP, which contain articles in the area of computer science. It was developed using the programming language Python 3.7 and the Framework for Web Flask and database PostegreSQL.
Keywords: Networking, Web Scraping, DBLP, digital libraries

1. Introdução

Desde o início da humanidade, o conhecimento e a sua transmissão as futuras gerações é um fato inerente a nossa própria existência, pois essa comunicação trouxe e traz melhorias na qualidade de vida e impulsiona a evolução da nossa sociedade.

A capacidade dos seres humanos de aprender e passar esse aprendizado adiante, compartilhar conhecimento garantiu a nossa sobrevivência diante dos perigos e adversidades que a natureza nos infligia nos primórdios da civilização. Hoje em dia, isso ainda garante a nossa sobrevivência, muito embora os perigos das sociedades modernas sejam outros. Mas o compartilhamento de conhecimento atualmente é feito por outros motivos além da sobrevivência, principalmente nessa fase da chamada Sociedade do Conhecimento, onde informação e o conhecimento podem significar poder, sucesso e tem um alto valor financeiro.

Ao longo dos anos as formas de compartilhar conhecimentos entre as sociedades, indivíduos e gerações, quer seja o saber comum, senso comum ou o saber científico mudaram, evoluíram. Se antes o senso comum era passado de forma oral dependendo da memória, com o surgimento da escrita foi revolucionado para outras formas, tais como tábuas de argila, pergami-

nhos, papiros, livros, jornais, *blogs*, *ebooks*, *podcasts*, vídeos e outros meios digitais atuais. O saber científico que antes estava contido no senso comum, após o surgimento da metodologia científica passou a ser compartilhado separadamente, chegando atualmente no formato de Artigos Científicos.

Independente da forma de compartilhamento do conhecimento, podemos afirmar a que divulgação, a colaboração entre diferentes pessoas de diferentes culturas produziu diversos avanços em todas as áreas, avanços sociais e tecnológicos.

Para o avanço da ciência, a comunicação é uma atividade indispensável e imprescindível, pois por meio dela ocorre a troca de informações e impulsionam o processo de desenvolvimento científico quando por exemplo, um autor desenvolve seu trabalho atual com base no trabalho de outro. Atualmente as publicações científicas são o principal meio pelo qual os pesquisadores divulgam os resultados de suas pesquisas, se destacam e se tornar visíveis academicamente, pois também é a através dessas publicações que eles conseguem auxílios financeiros e reconhecimento perante a sociedade.

No estudo da *National Science Foundation (NSF)* [**?**] divulgado em 2018, no ano de 2016 foram publicados 2.295.608 artigos entre todas as áreas do conhecimento, o primeiro lugar nesse ano foi a China com 426.165, seguida dos Estados Unidos com 408.985, o Brasil está em 12º lugar com 53.607.

A produção científica tem crescido exponencialmente, mas esse crescimento criou algumas dificuldades como por exemplo, a busca por publicações relevantes relacionadas aos nossos temas de interesse e na identificação de colaborações entre pesquisadores globais (ou seja, *networking* entre autores e coautores e suas afiliações).

1.1. Motivacao

A situação descrita acima gerou a motivação para o desenvolvimento deste trabalho de conclusão de curso ao estilo de uma Prova de Conceito (POC), pois não há uma forma fácil de busca de publicações e nem uma forma visual rápida para identificarmos autores e coautores, concentrações de publicações sobre os mesmos temas sem interação entre os autores e as universidades ou centros de pesquisa.

Existem vários *websites* de busca que estão gradativamente melhorando suas formas de encontrar publicações científicas como o Google Acadêmico e outros, que utilizam motores de busca para percorrer os sites indexados da internet e retornar os resultados. Esses motores de busca utilizam processos automatizados que extraem metadados dos documentos. Há também *websites* especializados como o DBLP, que possui um motor de busca, mas dependem da ação humana para atualização dos metadados referentes as publicações científicas, que prezam pela qualidade desses dados.

Porém nos exemplos citados acima ao pesquisarmos algum tema, os resultados da busca serão muitos e não há uma forma amigável para avaliarmos os metadados, como por exemplo, a rede de relacionamento dos autores e suas afiliações.

1.2. Objetivos

Com a formulação do problema acima e visando atender essa necessidade, surgiu a ideia da criação de um *website* onde seria possível pesquisar publicações, conforme o tema de interesse ou autor, visualizar num grafo a rede de relacionamento dos autores retornados na busca e as afiliações dos mesmos de forma gráfica e simples.

1.3. Objetivos Específicos

Sendo assim, podemos definir os objetivos específicos, limitando o escopo ao desenvolvimento de uma prova de conceito:

- Desenvolvimento de um *website* em *Python* para busca dos artigos por título e autor.
- Desenvolvimento um grafo da rede de cooperação entre os autores e coautores retornados da busca.
- Desenvolvimento de uns gráficos em forma de mapa com a localidade geográfica das afiliações dos autores identificadas na base de dados histórica do DBLP.

1.4. Escopo e Limitações

Contudo, considerando o objetivo global e os específicos, nesse trabalho nós definimos um escopo bem delimitado sobre o desenvolvimento do sistema:

- Utilizar as informações da base DBLP, pois ele já concentra informações de publicações de vários outros e de várias revistas, jornais e conferências. Outro ganho ao utilizar a base dados do DBLP é que o mesmo disponibiliza várias informações referentes aos autores e artigos.

- Utilizar somente as publicações denominadas como Artigos da base de dados DBLP. Somente artigos da área de Ciência da computação.
- Utilizar somente informações históricas, ou seja, não será desenvolvido um processo para atualizações online real time na *Application Programming Interface (API)* do DBLP.
- O funcionamento do *website* será *offline*, ou seja, na máquina local.

2. Trabalhos Correlatos

Há vários trabalhos acadêmicos correlatos visando a análises de redes de relacionamento entre autores e coautores. Alguns também utilizam a base de dados DBLP e outros usam outras fontes. Porém são análises pontuais e não oferecem a opção de busca de artigos.

Como por exemplo, artigo *Adoção da plataforma lattes* [**?**], que utiliza como fonte de dados para caracterização de redes científicas, que faz uma extração dessa base de dados e análise bibliométricas e da rede de colaboração científica da mesma.

Outro artigo que também faz análises de *Networking* e bibliométricas sobre o Google Acadêmico é o *Building and Analyzing a Global Co-Authorship Network Using Google Scholar Data* [**?**].

O Google Acadêmico [**?**] que é a ferramenta de busca de publicações acadêmicas da Google, possui várias opções de filtros para as buscas, inclusive uma área de métricas com algumas opções que permitem a avaliação da importância dos artigos através de citações. É uma ferramenta gratuita que tem impulsionado a divulgação de produções cientificas com grande abrangência nos resultados retornados pelo seu buscador, mas não oferece opções visuais para análises dos dados retornados.

O *website IBM Explore News* [**?**], utiliza a API do *IBM Watson* para construir redes de informações e possui um sistema de busca de notícias e artigos. Utiliza Processamento de Linguagem Natural (PLN) para analisar e apresentar as informações de forma visual. Essa plataforma é capaz de gerar grafos de relacionamento das notícias e entidades mencionadas, um gráfico de nuvem de palavras, um mapa mundial com as localidades e um gráfico de evolução das notícias retornadas na busca.

A ferramenta *DBL-Browser* [**?**] disponibilizada em 2005, oferecia recursos gráficos e a de busca de artigos no conteúdo do DBLP. Comparando ao trabalho aqui proposto, essa ferramenta oferecia mesmo recursos e vários outros, exceto a opção de mapa com localização geográfica, porém a última atualização foi em 2013 em versão Beta. Embora essa versão esteja disponível para ser baixada no *website Source Forge*, tal código não foi utilizado nesse trabalho.

3. Materiais e Métodos

Atualmente há várias bases de dados bibliográficos na *internet*, as bibliotecas digitais, que oferecem serviços de busca para diversas publicações científicas, podem ser específicas de uma área ou não, gratuitas ou não. Essa bibliotecas digitais tem democratizado o acesso aos trabalhos publicados nas áreas científicas. Como exemplos, podemos citar a Plataforma Lattes, DBLP, Sociedade Criacionista Brasileira (SCB), National Center for Biotechnology Information (NCBI), Elsevier's abstract and citation database (Scopus), Plataforma Sucupira e várias outras. No *site* da *Wikipedia* [**?**], há uma lista com 149 bases de dados acadêmicas e motores de busca de publicações científicas.

Porém, considerando a facilidade na obtenção dos dados em tempo hábil para a conclusão desta POC e a qualidade de metadados, foi delimitado como escopo neste trabalho utilizar somente a base de dados disponibilizada pelo DBLP. Contudo consideramos que o modelo aqui apresentado é factível de extensão para que sejam incluídas outras bases de dados.

O DBLP já faz um trabalho de busca e atualização nas principais bibliotecas digitais e editoras ligadas à ciência da computação, tais como: *Association for Computing Machinery (ACM), International Federation for Information Processing (IFIP), Institute of Electrical and Electronics Engineers (IEEE)*, entre outras instituições.

Também já fazem vários tratamentos nos dados e disponibilizam um arquivo em formato *eXtensible Markup Language (XML)* com todo o histórico para baixar sob a licença *Open Data Commons ODC-BY* 1.0.

3.1. DBLP

Digital Bibliography Library Project (DBLP) [**?**], é um repositório bibliográfico da área de ciência da computação desenvolvido por Michael Ley e mantido pela *Universität Trier*, na Alemanha. Esse projeto começou em 1993 como um subproduto da tese de doutorado recém-finalizada do autor e era uma coleção de índices de conteúdos de procedimentos e revistas das áreas de banco de dados e lógica de programação. Ao longo dos

anos evoluiu para repositório abrangente de trabalhos relevantes de pesquisa em ciência da computação.

Atualmente constam em sua base 4.646.295 publicações, 2.307.522 autores, 5.692 artigos em conferências e 1.599 artigos de jornais. O DBLP atualmente é um serviço e não um projeto de pesquisa [**?**], o serviço é operado com recursos limitados. Sua missão é fornecer acesso gratuito a metadados bibliográficos de alta qualidade para o benefício da comunidade internacional de pesquisa em ciência da computação. O DBLP fornece uma API de consulta XML simples para buscas em sua base de dados online. Fornecem também uma base de dados histórica em formato XML atualizado diariamente.

Todos os dados de sua base são liberados sob a licença *Open Data Commons ODC-BY* 1.0. O arquivo XML utiliza o estilo de formatação *BibTeX* [**?**] e possui também um arquivo com extensão Document Type Definition (DTD).

Um dos principais motivos que motivaram nossa escolha dessa base de dados, são os vários tratamentos que já são feitos nela, tais como:

- Tratamento sinônimos e homônimos para nome dos autores.
- Possui uma chave única para identificação dos artigos.
- Possui uma chave única para identificação da unicidade dos autores, mesmo quando eles fazem publicações com nomes diferentes, por exemplo: abreviaturas ou mesmo trocas de nomes.
- Atualizações constantes.
- Possui publicações das principais editoras e bibliotecas digitais na área da ciência da computação.

Devido ao tamanho da base de dados do DBLP, o *website* não utiliza Sistema de gerenciamento de banco de dados (SGBD), as informações são armazenadas em milhares de arquivos de sistema e são utilizados *scripts* escritos em *C, Perl* e *Java* para a sua manutenção e funcionamento. Essa base de dados contêm os elementos *Articles, Inproceedings, Proceedings, Book, Iincollection, Phdthesis, Mastersthesis* e *WWW*.

Delimitamos o escopo em utilizar somente as informações dos elementos *Articles* e *WWW*. O elemento *Articles* contém as informações dos artigos publicados em conferências ou *workshops*, periódicos ou revistas e o elemento *WWW* contém os dados pessoais dos autores.

Neste trabalho foram utilizados os seguintes arquivos do DBLP:
dblp.dtd: Arquivo DTD (Document type definition) contendo a estrutura de elementos e atributos válidos do XML.
dblp.xml.gz: Arquivo compactado que contêm a base de dados *DBLP.XML*.

O processo de *parse* do *XML* foi desenvolvido em *Python* 3.7 utilizando a API *Simple API for XML (SAX) LXML* e guardando no banco de dados *PostgreSQL*.

3.2. Parse do XML

Devido ao tamanho do arquivo *DBLP.XML*, foi necessário utilizar a metodologia SAX [**?**], que faz uma leitura serial do arquivo e é orientada a eventos. Os eventos utilizados foram de início de uma *tag* e fim da mesma.

O processamento de um XML com SAX nos permite iterar sobre o documento gerando um laço do início ao fim do mesmo e durante essa iteração são disparados os eventos de início e fim dos elementos/*tags*, assim podemos controlar o uso da memória para leitura de grandes arquivos.

Foram desenvolvidos dois procedimentos de leituras utilizando o a API SAX LXML em *Python* para leitura separada dos elementos *Articles* e *WWW*. O pacote LXML é uma biblioteca *Python* para tratamento de arquivos XML e HyperText Markup Language (HTML), é uma evolução da API *ElementTree*. O pacote *LXML* possui alta performance na leitura de grandes arquivos, fornecem o alto desempenho nas tarefas de análise, serialização do arquivo e transformação.

3.3. Banco de Dados

O banco de dados utilizado neste trabalho foi o PostgreSQL [**?**] devido a sua facilidade de acesso, instalação e alta performance, trata-se de um SGBD relacional *Open Source*.

O elemento *WWW* do *DBLP.XML* gerou duas tabelas que são relacionadas pelo campo *key_author*, são elas:
authors-names: Contém os diferentes nomes de um mesmo autor, pois em alguns artigos o nome pode estar com abreviaturas e em outro sem. O DBLP já faz esse tipo tratamento e inclui no elemento *author* uma lista com todas as versões de nomes por eles identificados e gera uma chave única para os diferentes nomes do mesmo autor.
authors: Contém os dados pessoais dos autores, contém uma chave para identificação da unicidade dos autores já disponibilizada pelo DBLP que é utilizada

como índice no próprio *website* para acesso à página de cada autor. Ou seja, essa chave não é um número, mas sim um endereço para localização dos autores no *website* do DBLP. Conforme documentação da base XML disponibilizada pelo website, o primeiro nome da lista de nomes é o mais atual e por isso é o nome utilizado nos gráficos e para pesquisa.

O elemento raiz *ARTICLE* do *DBLP.XML* gerou também duas tabelas que são relacionadas pelo campo *key_article* entre elas e pelo campo *key_author* com as duas citadas acima, são elas:

articles-authors: Como um mesmo artigo pode possuir um ou mais autores, foi necessário criar uma tabela para relacionar os artigos com os autores.

articles: Contém as informações dos artigos e possui uma chave para identificação da unicidade dos artigos já disponibilizada pelo DBLP que é utilizada no próprio *website*. Ou seja, essa chave não é um número, mas sim um endereço para localização dos artigos no *website* do DBLP.

Na Figura 1 apresentamos a modelagem das tabelas implementadas no banco da dados local.

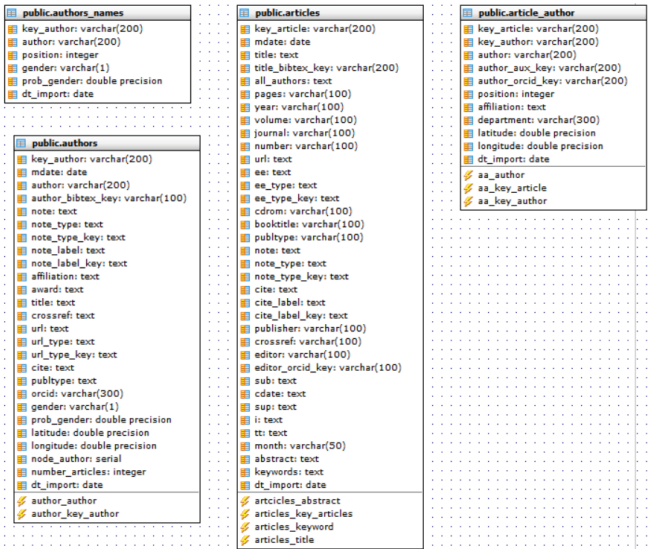

Figura 1: Tabelas geradas a partir dos elementos *Articles* e *WWW*.

3.4. *Website* para busca dos artigos e visualização dos gráficos

Como já citado anteriormente, nos objetivos, o *website* proposto nesse trabalho visa possibilitar a busca de artigos e gerar um grafo dos autores e um mapa das suas afiliações.

Para desenvolvimento do *website* foram utilizados a linguagem de programação *Python* 3.7, o *framework Flask* e a linguagem HTML. Para desenvolvimento do grafo de relacionamento entre os autores, foram utilizados os pacotes *NetworkX* e *Plotly*. Para desenvolvimento do mapa de afiliações foi utilizado o pacote *Plotly*. O *Framework* Flask é um micro *framework* para desenvolvimento *web* em *Python*, é chamado de micro, pois não possui a camada de abstração do banco de dados, porém é simples e rápido desenvolvimento, sendo ideal para websites pequenos.

O pacote *NetworkX* é uma biblioteca *Python* utilizada para a criação, manipulação e análise de grafos. É uma das mais populares, com uma comunidade ativa e vários exemplos na internet, possui várias opções de configurações. Utilizamos a opção *Graph* e *Spring Layout* que utiliza o algoritmo *Fruchterman Reingold Force Directed* para desenho e posicionamento dos nós, cuja principal finalidade é um desenho visualmente agradável e que o menor número possível de arestas se cruzem.

Neste projeto foi utilizado um único arquivo *Python*, chamado *app.py*, com as funções para busca no banco de dados, geração dos gráficos e renderização em um outro arquivo com extensão HTML. Assim, foram utilizados somente dois arquivos, um *.py* e outro *.html*, visando a simplificação do processo.

3.5. *Networking* entre os autores e coautores

A teoria dos grafos teve origem num artigo publicado por Leonhard Euler em 1736, sobre o problema das sete pontes de Königsberg [**?**]. É um ramo da matemática que estuda objetos combinatórios, ou seja, a relação entre os objetos de um determinado conjunto. Atualmente é muito utilizado para análises de redes de sociais.

Um grafo pode ser representado pela expressão $G(V, E)$, onde V é o número de vértices que também podem ser chamados de nó ou atores, E são as arestas que são pares do elemento V e representam as relações entre os atores. Então os autores são os nós, vértices e a relação autor com coautor quando publicam artigos juntos são as arestas, assim temos uma rede de relacionamento através das publicações em conjunto. Um autor pode ser co-autor em outras publicações e assim por diante. O tamanho do nó varia conforme a quantidade de artigos. Como o escopo foi delimitado em utilizar somente o elemento raiz *ARTICLE*, o conceito de comunidades não foi utilizado, pois o elemento re-

presenta o conjunto *G* da expressão do grafo. Obtemos acima seguinte expressão:

ARTICLE(autor/coautor, artigo)

Por exemplo, o artigo *"Analysing Social Networks Within Bibliographical Data"*, autores Stefan Klink, Patrick Reuther, Alexander Weber, Bernd Walter e Michael Ley, sendo que o autor é o primeiro da lista e os demais são os co-autores, temos a seguinte relação:

V={Stefan Klink, Patrick Reuther, Alexander Weber, Bernd Walter, Michael Ley}

A relação comum entre o autor e os coautores é o artigo, o **A** é *"Analysing Social Networks Within Bibliographical Data"*.

A = {{Stefan Klink, Patrick Reuther], {Stefan Klink, Alexander Weber}, {Stefan Klink, Bernd Walter}, {Stefan Klink, Michael Ley}}

Nesta análise foi utilizado uma rede não direcional de ligação simples, ou seja, a relação não possui peso, somente o nó possui peso que é a quantidade de artigos retornados como resultado da busca por título ou nome do autor.

3.6. Mapa de afiliações

Afiliações são as ligações, vínculos que os autores têm com as instituições que podem ser, universidades ou centros de pesquisas. Esse vínculo pode mudar com o tempo, por exemplo na época que um determinado artigo foi publicado o autor estava em uma determinada universidade e depois de um tempo em outra ou mesmo ter ligações com mais de uma instituição ao mesmo tempo.

Essa informação sobre filiações é importante para obtermos dados sobre a rede de relacionamento entre as instituições e seus autores, regionalismos e concentrações de publicações e outros.

Na base do DBLP há um campo de afiliação para cada autor, porém ele não está preenchido para a maioria dos autores. Foi então desenvolvido um processo de *Web Scraping* para a partir do endereço na *internet* onde está a publicação do artigo que normalmente são nos websites das editoras para extrair as informações de filiações. Mas esse processo de *Web scraping* é demorado e trabalhos, gerando também um grande volume de retrabalho, pois a cada *website* diferente o código para varredura do mesmo muda. *Web Scraping* resumidamente é a técnica de extração de dados utilizada para coleta de informações em *websites*, páginas HTML.

Outro problema enfrentado foi devido as políticas de segurança desses *websites* que limitam o acesso a eles no processo *scraping* dos mesmos. Também, devido à grande quantidade artigos o processo é lento, a cada artigo é necessário abrir a página que o hospeda, identificar as informações, tratá-las e salvar no banco de dados.

Então apesar de o código de *Web Scraping* ter sido desenvolvido utilizando a biblioteca *BeautifulSoup*, não foi possível buscar as informações de afiliação no prazo estabelecido para a entrega deste trabalho, pois o processo é demorado. A biblioteca *Python Request* é utilizada para solicitar ao servidor web baixar o conteúdo de uma página HTML e a biblioteca *BeautifulSoup* é utilizada para fazer a leitura do HTML de forma que seja possível extrair informações da mesma.

Porém para os casos em que a afiliação já estava preenchida, foi utilizado o pacote *GeoPy Geocodes* para identificar as coordenadas de la para plotagem num mapa mundial. O pacote *GeoPy Geocodes* é uma biblioteca *Python* utilizada para identificação de geocoordenadas gratuito.

Para tornar o grafo e o mapa visualmente interativo e agradável no *website*, foi utilizado a biblioteca *Plotly*.

Entretanto não foi possível identificar as coordenadas da maioria das afiliações, pois não há o endereço das mesmas. Mas mesmo assim, com os dados que foram conseguidos já é possível fazer algumas análises conforme apresentadas na próxima seção.

4. Resultados e Discussões

Para apresentação dos resultados, a base de dados foi atualizada com a posição com uma fotografia histórica até 08/06/2019. Nessa data de atualização o tamanho do arquivo *DBLP.XML* era de 2,42 GB. A Tabela abaixo apresenta a quantidade de registros encontrada na base.

Tabelas	Quantidade
articles	2.046.023
articles_authors	5.930.441
authors	2.300.374
authors_names	2.351.599

As tabelas *authors* e *authors_names* contém todos os autores, inclusive autores pertencentes a outros elementos que não seja o *Articles*, pois no processo de carga não havia como separá-los. Logo a quantidade de autores que escreveram um ou mais artigos é 1.434.878,

ou seja, temos uma média de 0,70 autores por artigos. A tabela de artigos possui publicações desde 1936 e como podemos ver na Figura 2, em 2018 forma publicados 175 milhões e entre Janeiro até Maio/2019 são 82 milhões.

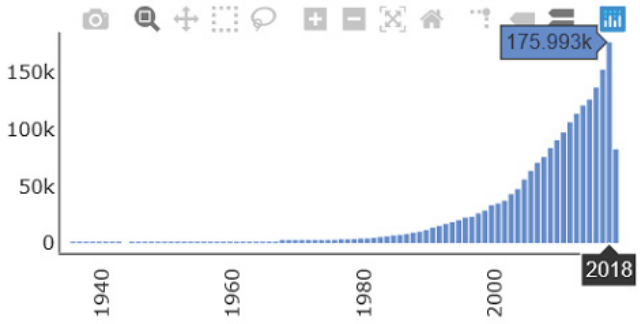

Figura 2: Quantidade de artigos por ano

Na Figura 3 apresenta o gráfico de nuvem de palavras com todos os títulos dos artigos na base de dados, onde podemos verificar que as palavras que mais aparecem são *Neural* e *Network*.

Figura 3: Nuvem de palavras de todos os títulos

A ferramenta desenvolvida neste trabalho possibilita a opção de busca por título de artigos, gera um grafo para a visualização da rede de relacionamento entre os autores e coautores, o mapa de filiações é gerado e a nuvem de palavras também, porém devido a limitações de recursos tais como, tempo, infraestrutura, arquitetura, sistemas e outros, apenas uma pequena parcela de usuários finais seriam beneficiados pela mesma caso ela fosse disponibilizada online com esse escopo reduzido.

Nas Figuras 4,5,6,7,8, Seguem as imagens do *website* desenvolvido demonstrando as funcionalidades de busca e geração dos gráficos.

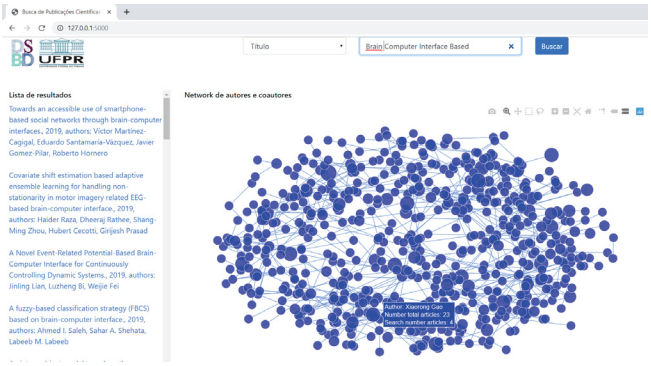

Figura 4: Grafo dos autores e coautores

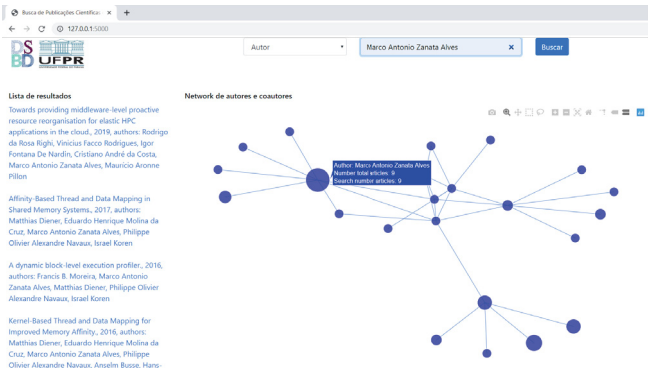

Figura 5: Grafo dos autores e coautores

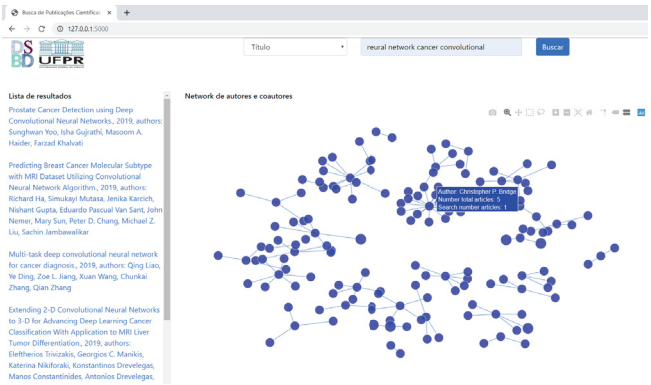

Figura 6: Grafo dos autores e coautores

Figura 7: Mapa das afiliações

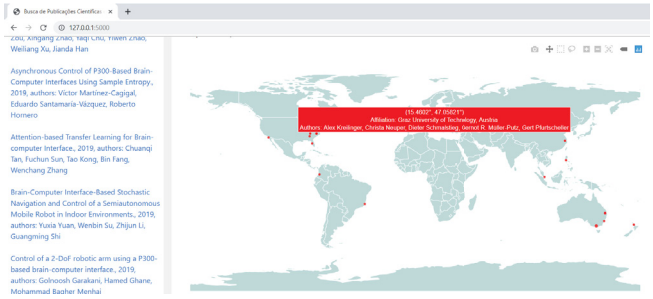

Figura 8: Mapa das afiliações

5. Conclusões e Trabalhos Futuros

Com o desenvolvimento deste trabalho, foi possível averiguar que apesar das diversas ferramentas disponíveis de busca de publicações científicas, as ferramentas que possuem opções visuais para análise não possuem capacidade de volume e diversidade de áreas e as que possuem volume e diversidade não oferecem opções visuais.

Um grande volume de dados não significa grande volume conhecimento, então apesar das ferramentas já existentes, ainda há uma deficiência e a necessidade por uma que nos permita a busca fácil e rápida, bem como opções visuais para facilitar a interação, divulgação e pesquisa das publicações científicas, autores e suas afiliações.

Encaramos nosso trabalho como uma prova de conceito, e podemos dizer que as expectativas foram atendidas. Além disso, ressaltamos a relevância de ferramentas para busca de material científico que unam performance, grande volume de dados, qualidade de metadados e as opções de visualizações gráficas.

Como possíveis trabalhos futuros, pensamos em diversas frentes.

- Incluir no banco de dados os demais elementos que constam na base de dados do DBLP, e melhorar o desempenho do banco de dados. Assim, podemos incluir outras opções de busca, como por exemplo, busca por resumo e palavras chaves.
- Aprofundar a análise exploratória dos dados através por exemplo de algoritmos de análise de similaridade dos resumos dos artigos. Também entendemos que sistemas de clusterização e recomendação de artigos podem ser facilmente implementados. Aprofundar também as análises de *networking* entre os autores e coautores, incluindo também a visão de comunidades e a inclusão de arestas no gráfico de mapas que seriam linhas conectando as afiliações.
- Expandir o *website* incluindo publicações de outras áreas além da Ciência da Computação e de incluir dados de outras bibliotecas digitais, como por exemplo da Plataforma Lattes.